Reward Bias Substitution: Single-Axis Mitigations Shift Optimization Pressure

MAX LAMPARTH, Stanford University, USA
DANIEL FEIN, Stanford University, USA
ANDREAS HAUPT, Stanford University, USA
MARCEL HUSSING, University of Pennsylvania, USA
MYKEL KOCHENDERFER, Stanford University, USA

Single-axis mitigations of reward-model biases (e.g., reducing reliance of the proxy reward on length, sycophancy, or style) can rotate optimization pressure onto correlated proxies rather than eliminate it, a failure mode we call reward bias substitution. We formalize the underlying measurement-vs-optimization gap between the audit distributions where mitigations are validated and the policy distributions where optimization realizes their effects. We introduce a taxonomy, instantiated in closed form, classifying single-axis mitigation outcomes into successful mitigation, bias substitution, overcorrection, silent non-op, and audit-distribution sensitivity. We prove that single-axis mitigation methods cannot be validated by audit-distribution-only evaluation: successful mitigation, bias substitution, and overcorrection produce structurally identical observables under ranking accuracy and win-rate scoring, regardless of benchmarks richness. Augmenting evaluation with policy-induced distributions while tracking multiple biases provably closes the gap. We give actionable prescriptions for mitigation methods and benchmarks. Across published preference-learning mitigation work, no method we survey reports the evidence needed to certify successful mitigation. We demonstrate bias substitution in language model RLHF, where a length penalty during GRPO training compresses responses as intended yet redirects optimization pressure onto confidence calibration, driving the trained policy into overconfidence. Our experiments also show that a published length-debiasing operator zeros pooled reward–length correlation but flips sign within-prompt on three of four SOTA reward models with true reward degrading on two, and that length–sycophancy coupling reverses under human–LLM judge disagreement across eight model families.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Reinforcement learning**.

Additional Key Words and Phrases: RLHF, reward hacking, preference learning, evaluations, reward modeling, theory

ACM Reference Format:
Max Lamparth, Daniel Fein, Andreas Haupt, Marcel Hussing, and Mykel Kochenderfer. 2026. Reward Bias Substitution: Single-Axis Mitigations Shift Optimization Pressure. 1, 1 (May 2026), 52 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) and related preference-learning methods rely on human annotations to shape model behavior on non-verifiable objectives, but the resulting reward models and policies are confounded by spurious correlations in either the preferences or their actions. Both cases are examples of reward hacking as caused by the reward identifiability problem [5, 37, 72, 73, 78], leading to misgeneralization in terms of

Authors' Contact Information: Max Lamparth, lamparth@stanford.edu, Stanford University, Stanford, California, USA; Daniel Fein, Stanford University, Stanford, California, USA; Andreas Haupt, Stanford University, Stanford, California, USA; Marcel Hussing, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Mykel Kochenderfer, Stanford University, Stanford, California, USA.

Manuscript submitted to ACM

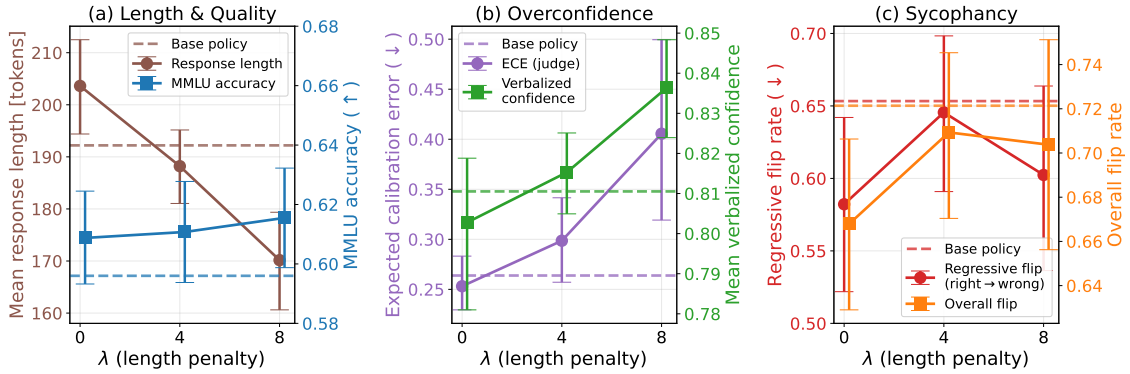


Fig. 1. RLHF of *Llama-3.2-3B-Instruct* with *Skywork-Reward-V2-Llama-3.1-8B* using GRPO (β = 2e−2, LR = 3e−5, 600 steps) on *UltraFeedback* with four random seeds for each $\lambda \in \{0, 4, 8\}$, 95% bootstrapped confidence intervals. To mitigate length, we modify the reward $\tilde{R}$ in the RLHF objective as $\tilde{R}(x, y) = R_{\text{RM}}(x, y) - \lambda\, n_{\text{tok}}/100$. Training with $\lambda = 0$ does not break confidence calibration. As length is decreased the optimization pressure rotates onto the calibration axis, while MMLU accuracy and sycophantic behavior are preserved. See Section F.1 for details.

the original behavioral-shaping intention. A recurring example in language RLHF is reward-length correlation [e.g. 59, 71] of language reward models and verbosity of language model policies in RHLF and DPO methods [14, 33, 42, 64]. The so-called *length bias* issue has been observed with different proposed mitigation strategies, however, most prior reward bias mitigation works do not fully resolve or *even overcorrect* [21], consider at most two confounding factors simultaneously [21], or evaluate within fixed audit distributions that leave substitution onto uncovered axes undetectable [22, 86]. Naively, one would want the proxy reward to track the true reward and be invariant to irrelevant surface features (length, formatting, style), varying only with the causally relevant features annotators care about (correctness, helpfulness, informativeness), motivating *reward feature invariance*. However, surface features labeled spurious are typically partially informative about the true reward and coupled with each other, so enforcing strict invariance removes signal alongside bias. As we demonstrate in this paper, ignoring these couplings produces **reward bias substitution**, where single-axis mitigations (e.g., debiasing length) tend to rotate optimization pressure into other correlated proxies rather than eliminating it. Any single-axis mitigation asserts that the targeted feature is spurious, so bias substitution can arise whether or not that assertion is identifiable from preference data. Figure 1 shows this issue concretely. Applying a length penalty during GRPO training compresses response length as intended, yet the freed optimization pressure drives the policy into the untargeted overconfidence axis, while multiple-choice accuracy holds and free-form accuracy decreases (see also Section F.1). As a result, we show that almost all existing reward bias mitigation papers cannot determine whether they have successfully mitigated, overcorrected, or merely substituted reliance onto an uncovered proxy on the evidence they report. Current reward/preference benchmarks cannot adjudicate between these outcomes either, risking unintended post-deployment harms of, e.g., silently substituting length bias for sycophancy bias in user interactions. We address these gaps by formalizing the failure mode and characterizing when standard evaluation can detect it. **This paper contributes the following:**

- **We formalize bias substitution as failure mode of single-axis mitigation**, in which removing spurious reliance at the evaluation distribution where mitigations are validated leaves optimization pressure free to

rotate onto correlated proxies at the trained policy-induced distribution. We identify the *measurement-versus-optimization gap* as the blindspot in prior reward-hacking definitions (see Section C).

- **We introduce a regime taxonomy** classifying all mechanistically distinct single-axis mitigation outcomes: successful mitigation (R0, with a contaminated sub-case $R0_{\text{cont}}$), bias substitution (R1), overcorrection (R2), silent non-op (R3), and audit-distribution sensitivity (R4). Each regime is instantiated in closed form and validated empirically, with additional deployment-empirical instances for R1, R2, and R4.
- **We prove a matched impossibility–sufficiency pair for audit-distribution evaluation**. No benchmark functional over audit-distribution observables can reliably distinguish successful mitigation from bias substitution or overcorrection, but, augmenting the benchmark input with policy-induced distributions provably does. We derive necessary prescriptions for *certifiable reward bias mitigation claims* and benchmark improvements.
- **We provide empirical evidence for our framework.** We demonstrate bias substitution directly in language model RLHF, where a GRPO length penalty trades response length for policy overconfidence (R1). We further instantiate the measurement-versus-optimization gap on a length-debiasing operator across five reward models. We provide the first systematic characterization of length-sycophancy statistical dependence across four labeling regimes and eight model families, finding that the coupling reverses sign under human–LLM judge disagreement (R4). Beyond our experiments, we identify previously isolated findings as instances of R1 and R2 across preference-learning mitigation work.

2 RLHF Invariances and Formalizing Bias Substitution

We work in the single-turn contextual-bandit reduction of RLHF [14, 42, 76] with prompts $x \in \mathcal{X}$ drawn i.i.d. from a fixed context distribution $\mathcal{D}$, responses $y \in \mathcal{Y}$ sampled $y \sim \pi(\cdot \mid x)$, true reward $R(x, y)$, learned proxy $\tilde{R}(x, y)$, and the *plain return* $J(\pi, R) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi(\cdot|x)}[R(x, y)]$ to maximize. The KL-regularized RLHF training objective is

$$J_{\text{RLHF}}(\pi; \tilde{R}) = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}[\tilde{R}(x, y)] - \beta\, D_{\text{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\big)\right]. \tag{1}$$

For any reward $\bar{R}$ the KL-regularized optimum is the softmax policy [63, 64]

$$\pi^{\star}_{\beta}(\bar{R})(y \mid x) \propto \pi_{\text{ref}}(y \mid x) \exp\big(\bar{R}(x, y)/\beta\big). \tag{2}$$

This optimal KL regularized policy is not invariant to reward scale. Let $\mu_\pi(x, y) = \mathcal{D}(x)\pi(y \mid x)$ denote the policy-induced distribution on $\mathcal{X}\times\mathcal{Y}$ given a policy π (the one-step occupancy measure under $\mathcal{D}$). In addition to the KL anchor π_{ref}-defined measure $\mu_{\pi_{\text{ref}}}$, we fix a *diagnostic measure* μ_{diag} on $\mathcal{X}\times\mathcal{Y}$, used for reliance estimation and correlation measurement. Natural choices include $\mu_{\pi_{\text{ref}}}$ itself (coupling measurement and optimization references) and annotator-conditioned audit distributions $\mu^{\text{human}}_{\text{diag}}, \mu^{\text{LLM}}_{\text{diag}}$, reflecting that reliance estimates depend on who labels and formalizing the human-versus-LLM-judge gap (see also Section E.3). A standing regularity Assumption E.1 is invoked throughout for well-defined softmax policies and policy-level expectations. Reward modeling in RLHF involves invariances at two levels. First, what preference data identifies about the reward (data-level), and second, how the KL-regularized policy responds to reward transformations (policy-level). At the data level, the *reward identifiability problem* leaves the learned reward underdetermined. Many reward functions explain the same preference data and preference comparisons identify the true reward only up to a prompt-only additive shift under the Bradley-Terry likelihood, leaving the cardinal scale loosely pinned [72]. This identifiability gap allows the reward model to absorb spurious feature correlations consistent with the observed rankings. At the policy level, KL-regularized optimization is sensitive to cardinal reward values, since $\tilde{R} \to c\tilde{R}$ at fixed β is equivalent to $\beta \to \beta/c$, so two reward models agreeing on all pairwise preferences can still induce different

policies. Combined with data-level identifiability, this means confounded features can be consistent with the observed rankings while still distorting the cardinal values that determine the optimized policy. For example, applied to length bias, this implies that preference data alone cannot distinguish causally relevant length from spurious length without additional assumptions or interventions [72]. Partial identifiability of reward under Bradley-Terry has been formally characterized [72] and reward misidentification empirically demonstrated on continuous-control benchmarks [78]. Both concern what preference data fails to identify. The downstream question of how single-axis mitigation operators that target identified spurious features behave under optimization has been observed empirically in supervised vision [48], but not formally characterized in any setting, and its consequences for reward-model evaluation, mitigation methods, and downstream policy optimization remain unaddressed for RLHF. We work within this identifiability gap, classifying mitigation outcomes by quantities that remain meaningful when the spurious/structural partition and cardinal scale are underdetermined.

Thus, we decompose the proxy reward $\tilde{R}$ along a fixed set of interpretable surface features Φ (length, sycophancy, formatting), partitioned at an audit distribution μ_{diag} into spurious features Φ_{sp} and structurally relevant features Φ_{struct} by whether the true reward R depends on them (see Section A for the full formal derivation). A single-axis mitigation M_i is a projection-type operator, abstracting deployed linear-probe, calibration, and disentanglement methods, that zeros the proxy's linear reliance g_i on a targeted spurious feature at μ_{diag}. This creates a measurement-versus-optimization gap, as the mitigation is validated at μ_{diag}, where $g_i = 0$ by construction, but the optimized KL-regularized policy realizes its effects at the policy-induced distribution $\mu_{\pi^\star}$, where reliance generically reopens and optimization pressure rotates onto correlated features. Classifying outcomes by whether exploitation rotates onto another spurious feature ($\Delta_j \neq 0$) and by the sign of the true-reward change (ΔJ) yields five regimes, successful mitigation (R0), contaminated success (R0_{cont}), bias substitution (R1), overcorrection (R2), and silent non-op (R3), plus a transversal regime R4 whose label flips with the choice of μ_{diag}. We then prove a matched impossibility-sufficiency pair (Theorems A.13 and A.14). No benchmark functional over audit-distribution observables can separate R0 from R0_{cont}, R1, or R2, even granted oracle access to R, whereas augmenting the input with policy-induced distributions provably recovers the regime. This motivates our central prescription, that methods and benchmarks evaluate mitigations at policy-induced distributions and report off-target Δ_j alongside ΔJ. Full definitions, proofs, and closed-form instantiations are in Section A.

3 Bias Substitution in the Wild

Bias substitution is already pervasive in the published RLHF/DPO mitigation literature, miscategorized as isolated anomalies, capability trade-offs, or judge noise. Deployed reward models satisfy its precondition of correlated, partially informative spurious features, and we add length-sycophancy coupling measurements across eight model families, where the effect reverses from +154.3 characters under human-LLM judge agreement to −43.1 under disagreement, the empirical signature of R4 (Section F.3). To show the failure arises end to end, we run GRPO with a length penalty on *Llama-3.2-3B-Instruct* (Figure 1, Section F.1). As the penalty rises, mean response length falls from 204 to 170 tokens and MMLU accuracy holds, yet expected calibration error climbs from 0.25 to 0.41 and TriviaQA accuracy drops from 0.56 to 0.42, while a $\lambda = 0$ control leaves calibration intact. Optimization pressure rotates off length onto overconfidence, instantiating harmful R1. A published length-debiasing operator likewise zeros pooled reward-length correlation ($0.316 \rightarrow 0.037$) yet flips within-prompt sign on three of four SOTA reward models, with true reward degrading on two (Section F.2). Surveying the literature under our prescriptions, no published method provides evidence sufficient to certify successful mitigation (R0), available evidence points to R0_{cont}, R1, or R2, and every benchmark we examined falls inside the impossibility class of Theorem A.13. See Section B for details and Section D for a broader impact analysis.

References

[1] Anthropic. 2026. Models overview. Blog post. Accessed: 2026-05-06.

[2] Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. 2026. Training large language models on narrow tasks can lead to broad misalignment. *Nature* 649 (2026), 584–589. doi:10.1038/s41586-025-09937-5

[3] Yuyan Bu, Liangyu Huo, Yi Jing, and Qing Yang. 2025. Beyond Excess and Deficiency: Adaptive Length Bias Mitigation in Reward Models for RLHF. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, Albuquerque, New Mexico, 3091–3098. doi:10.18653/v1/2025.findings-naacl.169

[4] Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Yue Wang, Li Li, Wengang Zhou, and Houqiang Li. 2026. Disentangling Length Bias in Preference Learning via Response-Conditioned Modeling. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=hKxYESOzen

[5] 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, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=bx24KpJ4Eb Survey Certification, Featured Certification.

[6] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 8301–8327. doi:10.18653/v1/2024.emnlp-main.474

[7] Longze Chen, Lu Wang, Renke Shan, Ze Gong, Run Luo, Jiaming Li, Jing Luo, Qiyao Wang, and Min Yang. 2026. Learning Ordinal Probabilistic Reward from Preferences. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=0Vf5trUAVF

[8] Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. ODIN: disentangled reward mitigates hacking in RLHF. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. JMLR.org, Article 312, 18 pages.

[9] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 391 (2026), eaec8352. doi:10.1126/science.aec8352

[10] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2026. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=igbRHKEiAs

[11] Yeshwanth Cherapanamjeri, Constantinos Costis Daskalakis, Gabriele Farina, and Sobhan Mohammadpour. 2026. Learning Correlated Reward Models: Statistical Barriers and Opportunities. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=TbEyl6krsY

[12] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. JMLR.org, Article 331, 30 pages.

[13] Brian Christian, Jessica A F Thompson, Elle, Vincent Adam, Hannah Rose Kirk, Christopher Summerfield, and Tsvetomira Dumbalska. 2026. Reward Models Inherit Value Biases from Pretraining. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=dT399j1Azv

[14] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.

[15] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG] https://arxiv.org/abs/2110.14168 arXiv:2110.14168.

[16] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback. arXiv:2310.01377 [cs.CL]

[17] Magda Dubois, Cozmin Ududec, Christopher Summerfield, and Lennart Luettgau. 2026. Ask don't tell: Reducing sycophancy in large language models. arXiv:2602.23971 [cs.HC] https://arxiv.org/abs/2602.23971

[18] Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-Controlled AlpacaEval: A Simple Debiasing of Automatic Evaluators. In *First Conference on Language Modeling*. https://openreview.net/forum?id=CybBmzWBX0

[19] 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*. https://openreview.net/forum?id=5u1GpUkKtG

[20] 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 Eighth AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 893–900. doi:10.1609/aies.v8i1.36598

[21] Daniel Fein, Max Lamparth, Violet Xiang, Mykel J. Kochenderfer, and Nick Haber. 2026. One Bias After Another: Mechanistic Reward Shaping and Persistent Biases in Language Reward Models. arXiv:2603.03291 [cs.CL] https://arxiv.org/abs/2603.03291

[22] Xuan Feng, Bo An, Tianlong Gu, Liang Chang, Fengrui Hao, Peipeng Yu, and Shuai Zhao. 2025. C2PO: Diagnosing and Disentangling Bias Shortcuts in LLMs. arXiv:2512.23430 [cs.CL] https://arxiv.org/abs/2512.23430

[23] Lukas Fluri, Leon Lang, Alessandro Abate, Patrick Forré, David Krueger, and Joar Max Viktor Skalse. 2025. The Perils of Optimizing Learned Reward Functions: Low Training Error Does Not Guarantee Low Regret. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 17306–17377. https://proceedings.mlr.press/v267/fluri25a.html

[24] Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. 2025. Reward Shaping to Mitigate Reward Hacking in RLHF. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*. https://openreview.net/forum?id=62A4d5Mokc

[25] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. JMLR.org, Article 437, 32 pages.

[26] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *Nature* 2 (2020), 665–673. doi:10.1038/s42256-020-00257-z

[27] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.CL] https://arxiv.org/abs/2407.21783 arXiv:2407.21783.

[28] Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring Sycophancy of Language Models in Multi-turn Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, Suzhou, China, 2239–2259. doi:10.18653/v1/2025.findings-emnlp.121

[29] Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2025. Explaining Length Bias in LLM-Based Preference Evaluations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, Suzhou, China, 6763–6794. doi:10.18653/v1/2025.findings-emnlp.358

[30] Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo Ponti, and Ivan Titov. 2025. Post-hoc Reward Calibration: A Case Study on Length Bias. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=Iu8RytBaji

[31] Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Ilia Sucholutsky, Andrew Strait, Q. Vera Liao, and Umang Bhatt. 2025. Measuring and mitigating overreliance is necessary for building human-compatible AI. arXiv:2509.08010 [cs.CY] https://arxiv.org/abs/2509.08010

[32] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2026. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature* 652 (2026), 1159–1165. doi:10.1038/s41586-026-10410-0

[33] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2025. A Survey of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research* (2025). https://openreview.net/forum?id=f7OkIurx4b Survey Certification.

[34] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 2564–2572.

[35] Emaan Bilal Khan, Amy Winecoff, Miranda Bogen, and Dylan Hadfield-Menell. 2026. Safety Drift After Fine-Tuning: Evidence from High-Stakes Domains. arXiv:2604.24902 [cs.CY] https://arxiv.org/abs/2604.24902

[36] Hyeonji Kim, Sujeong Oh, and Sanghack Lee. 2025. Mitigating Length Bias in RLHF through a Causal Lens. arXiv:2511.12573 [cs.CL] https://arxiv.org/abs/2511.12573

[37] Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. 2021. Reward Identification in Inverse Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 5496–5505.

[38] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 517–545. doi:10.18653/v1/2024.findings-acl.29

[39] Ashwin Kumar, Yuzi He, Aram H Markosyan, Bobbie Chern, and Imanol Arrieta-Ibarra. 2025. Detecting Prefix Bias in LLM-based Reward Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 3196–3206. doi:10.1145/3715275.3732204

[40] Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. 2024. Catastrophic Goodhart: regularizing RLHF with KL divergence does not mitigate heavy-tailed reward misspecification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=UXuBzWoZGK

[41] Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. 2025. Correlated Proxies: A New Definition and Improved Mitigation for Reward Hacking. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=msEr27EejF

[42] Nathan Lambert. 2026. Reinforcement Learning from Human Feedback. arXiv:2504.12501 [cs.LG] https://arxiv.org/abs/2504.12501

[43] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating Reward Models for Language Modeling. In *Findings of*

the Association for Computational Linguistics: NAACL 2025. Association for Computational Linguistics, Albuquerque, New Mexico, 1755–1797. doi:10.18653/v1/2025.findings-naacl.96

[44] Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2025. Taming Overconfidence in LLMs: Reward Calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=l0tg0jzsdL

[45] Gengxu Li, Tingyu Xia, Yi Chang, and Yuan Wu. 2025. Length-Controlled Margin-Based Preference Optimization without Reference Model. arXiv:2502.14643 [cs.CL] https://arxiv.org/abs/2502.14643

[46] Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. 2024. Does Style Matter? Disentangling Style and Substance in Chatbot Arena. Blog post. Accessed: 2026-05-06.

[47] Zhuo Li, Pengyu Cheng, Zhechao Yu, Feifei Tong, Anningzhe Gao, Tsung-Hui Chang, Xiang Wan, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. 2025. Eliminating Inductive Bias in Reward Models with Information-Theoretic Guidance. arXiv:2512.23461 [cs.LG] https://arxiv.org/abs/2512.23461

[48] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. 2023. A Whac-a-Mole Dilemma: Shortcuts Come in Multiples Where Mitigating One Amplifies Others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[49] Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025. Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy. arXiv:2507.01352 [cs.CL] https://arxiv.org/abs/2507.01352 arXiv:2507.01352.

[50] Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. 2026. Robust Optimization for Mitigating Reward Hacking with Correlated Proxies. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=O3shkBWM2s

[51] Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. 2024. Eliminating Biased Length Reliance of Direct Preference Optimization via Down-Sampled KL Divergence. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1047–1067. doi:10.18653/v1/2024.emnlp-main.60

[52] Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2026. RewardBench 2: Advancing Reward Model Evaluation. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=fb0G86Dewb

[53] Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: simple preference optimization with a reference-free reward. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*. Curran Associates Inc., Article 3946, 38 pages.

[54] Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. 2026. What's In My Human Feedback? Learning Interpretable Descriptions of Preference Data. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=sC6A1bFDUt

[55] Ignavier Ng, Patrick Blöbaum, Siddharth Bhandari, Kun Zhang, and Shiva Kasiviswanathan. 2025. Debiasing Reward Models by Representation Learning with Guarantees. arXiv:2510.23751 [cs.LG] https://arxiv.org/abs/2510.23751

[56] Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. 2024. Nemotron-4 340B Technical Report. arXiv:2406.11704 [cs.CL] https://arxiv.org/abs/2406.11704

[57] OpenAssistant. 2023. OpenAssistant/reward-model-deberta-v3-large-v2. https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2. Hugging Face model card. Accessed: 2026-01-06.

[58] Henry Papadatos and Rachel Freedman. 2024. Linear Probe Penalties Reduce LLM Sycophancy. In *Workshop on Socially Responsible Language Modelling Research*. https://openreview.net/forum?id=6N2yES22rG

[59] 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*. Association for Computational Linguistics, Bangkok, Thailand, 4998–5017. doi:10.18653/v1/2024.findings-acl.297

[60] Judea Pearl and Elias Bareinboim. 2022. *External Validity: From Do-Calculus to Transportability Across Populations* (1 ed.). Association for Computing Machinery, New York, NY, USA, 451–482. https://doi.org/10.1145/3501714.3501741

[61] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 7599–7609.

[62] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer

El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 13387–13434. doi:10.18653/v1/2023.findings-acl.847

[63] Jan Peters and Stefan Schaal. 2007. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*. Association for Computing Machinery, 745–750. doi:10.1145/1273496.1273590

[64] 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. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=HPuSIXJaa9

[65] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity Bias in Preference Labeling by Large Language Models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*. https://openreview.net/forum?id=magEgFpK1y

[66] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The Pitfalls of Simplicity Bias in Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[67] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] https://arxiv.org/abs/2402.03300

[68] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. 2026. How RLHF Amplifies Sycophancy. arXiv:2602.01002 [cs.AI] https://arxiv.org/abs/2602.01002

[69] 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*. https://openreview.net/forum?id=tvhaxkMKAn

[70] Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Loose lips sink ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://openreview.net/forum?id=qq6ctdUwCX

[71] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A Long Way to Go: Investigating Length Correlations in RLHF. In *First Conference on Language Modeling*. https://openreview.net/forum?id=G8LaO1P0xv

[72] Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. 2023. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. JMLR.org, Article 1328, 26 pages.

[73] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward hacking. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Article 687, 12 pages.

[74] Ruike Song, Zeen Song, Huijie Guo, and Wenwen Qiang. 2025. Causal Reward Adjustment: Mitigating Reward Hacking in External Reasoning via Backdoor Correction. arXiv:2508.04216 [cs.LG] https://arxiv.org/abs/2508.04216

[75] Pragya Srivastava, Harman Singh, Rahul Madhavan, Gandharv Patil, Sravanti Addepalli, Arun Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuveer, Karthikeyan Shanmugam, and Doina Precup. 2026. Robust Reward Modeling via Causal Rubrics. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=oP99JQiDYp

[76] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc., Article 253, 14 pages.

[77] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 5433–5442. doi:10.18653/v1/2023.emnlp-main.330

[78] Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S. Brown. 2023. Causal Confusion and Reward Misidentification in Preference-Based Reward Learning. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=R0Xxvr_X3ZA

[79] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. *TRL: Transformers Reinforcement Learning*. https://github.com/huggingface/trl

[80] Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, and Sinong Wang. 2025. Beyond Reward Hacking: Causal Rewards for Large Language Model Alignment. arXiv:2501.09620 [cs.LG] https://arxiv.org/abs/2501.09620

[81] Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2024. HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 3371–3384. doi:10.18653/v1/2024.naacl-long.185

[82] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-Preference Bias in LLM-as-a-Judge. In *Neurips Safe Generative AI Workshop 2024*. https://openreview.net/forum?id=tLZZZIgPJX

[83] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. 2025. Language Models Learn to Mislead Humans via RLHF. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?

id=xJljiPE6dg

[84] Yuchen Yan, Jin Jiang, Zhenbang Ren, Yijun Li, Xudong Cai, Yang Liu, Xin Xu, Mengdi Zhang, Jian Shao, Yongliang Shen, Jun Xiao, and Yueting Zhuang. 2026. VerifyBench: Benchmarking Reference-based Reward Systems for Large Language Models. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=JfsjGmuFxz

[85] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=3GTtZFiajM

[86] Wenqian Ye, Guangtao Zheng, and Aidong Zhang. 2026. Rectifying Shortcut Behaviors in Preference-based Reward Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=m51t6RKfGH

[87] Yaowen Ye, Cassidy Laidlaw, and Jacob Steinhardt. 2025. Iterative Label Refinement Matters More than Preference Optimization under Weak Supervision. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=q5EZ7gKcnW

[88] Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. 2025. Reasoning Models Better Express Their Confidence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=rbBtoVnduo

[89] Danlong Yuan, Tian Xie, Shaohan Huang, Zhuocheng Gong, Huishuai Zhang, Chong Luo, Furu Wei, and Dongyan Zhao. 2026. Shorten After You're Right: Lazy Length Penalties for Reasoning RL. arXiv:2505.12284 [cs.AI] https://arxiv.org/abs/2505.12284

[90] Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. 2025. Demystify Verbosity Compensation Behavior of Large Language Models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*. Association for Computational Linguistics, Suzhou, China, 160–178. doi:10.18653/v1/2025.uncertainlp-main.14

[91] Kangwen Zhao, Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Dongyun Xue, Wengang Zhou, Li Li, and Houqiang Li. 2025. Bias Fitting to Mitigate Length Bias of Reward Model in RLHF. arXiv:2505.12843 [cs.LG] https://arxiv.org/abs/2505.12843

[92] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Article 2020, 29 pages.

[93] Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 3623–3643. doi:10.18653/v1/2024.acl-long.198

A Formalizing Bias Substitution

To formalize bias substitution, we turn the data- and policy-level invariances of Section 2 into a concrete failure-mode taxonomy, prove what audit-distribution evaluation can and cannot certify, and derive prescriptions for mitigation method claims and reward model benchmarks

A.1 Feature Map and Spurious vs. Structurally Relevant Features

In this work, we are interested in categorizing and analyzing the effects of bias mitigation strategies given a multiplicity of axes that we may care to optimize for. We will refer to these axes as (surface) features. Before we can disentangle their interactions with rewards, we need to define features.

Definition A.1 (Feature map). *A* feature map *is a measurable function $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ capturing an interpretable surface attribute of a response. To form a set of features that we care about, we fix a finite ordered tuple of feature maps $\phi_1, \ldots, \phi_K$ and collect them into the vector-valued map $\Phi = (\phi_1, \ldots, \phi_K)^\top : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^K$.*

We use Φ both as the column-vector-valued map $(\phi_1, \ldots, \phi_K)^\top$ and, by slight abuse, as the set $\{\phi_1, \ldots, \phi_K\}$ when membership ($\phi_i \in \Phi$) or partitions ($\Phi = \Phi_{\text{sp}} \sqcup \Phi_{\text{struct}}$) are more natural.

To instantiate our feature sets, we will make the natural assumption that the chosen feature sets capture distinct response attributes under the diagnostic distribution. For example, features such as length, sycophancy, politeness, or hedging may be correlated, but they should not collapse into the same measurement across diagnostic responses. This rules out duplicate or exactly redundant features and ensures that each feature represents a separately identifiable axis of response behavior.

Assumption A.2 (Non-degeneracy). *The Gram matrix* $G := \mathbb{E}_{\mu_{\text{diag}}}[\Phi\Phi^\top] \in \mathbb{R}^{K\times K}$, *with entries* $G_{ij} = \mathbb{E}_{\mu_{\text{diag}}}[\phi_i\phi_j]$, *is positive definite:* $G \succ 0$.

With these distinct axes in place, we can ask which attributes the reward actually depends on.

Definition A.3 (Spurious vs. structurally relevant at μ_{diag}). *Treat* R *as a function on* $\mathcal{X}\times\mathcal{Y}$ *and assume feature realizability (Assumption E.2). Feature* ϕ_i *is* spurious with respect to R at μ_{diag} *if*

$$\mathbb{E}_{\mu_{\text{diag}}}\big[R(x,y)\,\big|\,x,\ (\phi_j(x,y))_{j\neq i}\big] \text{ does not depend on } \phi_i(x,y) \text{ for } \mu_{\text{diag}}\text{-a.e.} \tag{3}$$

Otherwise ϕ_i *is* structurally relevant at μ_{diag}. *Write* $\Phi_{\text{sp}} = \{\phi_i \in \Phi : \phi_i \text{ spurious w.r.t. } R \text{ at } \mu_{\text{diag}}\}$ *and* $\Phi_{\text{struct}} = \Phi \setminus \Phi_{\text{sp}}$ *for the induced partition of* Φ.

See Section E.4 for the equivalent $o(\varepsilon)$ formulation under ε-mixture perturbations, capturing the intuition that R has no first-order dependence on ϕ_i at μ_{diag}. The corresponding causal reading and partial-identifiability relationship are given in Section E.1. Neither spuriousness nor structural relevance is fully identifiable from preference data alone [72]. Under partial informativeness, natural features like length act as mediators of multiple mechanisms. The binary partition of Φ from Definition A.3 may therefore conservatively classify them as Φ_{struct} when they correlate with R under μ_{diag} (Section E.3). This partition of Φ also matches existing single-axis mitigations acting on whole features rather than within-feature decompositions [e.g. 6, 21, 30, 58]. Finer partitions require stronger assumptions (interventions, gold labels, distributional invariance) outside our regime.

A.2 Single-Axis Mitigation and Measurement-Versus-Optimization Gap

Several prominent reward-model mitigations identify a feature direction, estimate $\tilde{R}$'s reliance on it, and subtract its contribution as, e.g., linear probes [21, 58], non-linear calibration [30], or architectural disentanglement [6]. To abstract away from these implementation details, we first define a population-level notion of reliance, given by the coefficients of $\tilde{R}$'s best linear approximation in Φ under the diagnostic distribution.

Definition A.4 (Linear reliance). *The* linear reliance *of* $\tilde{R}$ *on* Φ *at* μ_{diag} *is*

$$g(\tilde{R};\,\mu_{diag}) = \big(\mathbb{E}_{\mu_{diag}}[\Phi\Phi^\top]\big)^{-1}\mathbb{E}_{\mu_{diag}}[\Phi\,\tilde{R}] \in \mathbb{R}^K, \tag{4}$$

yielding the $L^2(\mu_{diag})$*-orthogonal decomposition* $\tilde{R} = \sum_i g_i\phi_i + \tilde{R}_\perp$ *with* $\mathbb{E}_{\mu_{diag}}[\phi_i\tilde{R}_\perp] = 0$.

Note g is a *statistic* of the triple $(\tilde{R}, \Phi, \mu_{\text{diag}})$, not a property of $\tilde{R}$ alone and evaluating at $\mu_{\pi^\star} \neq \mu_{\text{diag}}$ yields a different vector, driving later regime distinctions in our taxonomy.

Definition A.5 (Single-axis mitigation). *A* single-axis mitigation targeting feature i *is an operator* $M_i : \tilde{R} \mapsto \tilde{R}'$ *with* $|g_i(\tilde{R}';\,\mu_{diag})| < |g_i(\tilde{R};\,\mu_{diag})|$. *The canonical projection instance is*

$$M_i(\tilde{R})(x,y) \;=\; \tilde{R}(x,y) - g_i(\tilde{R};\,\mu_{diag})\,\phi_i(x,y). \tag{5}$$

We single out the projection instance above as canonical, because it zeros g_i at μ_{diag} exactly. The canonical M_i depends on μ_{diag} through $g_i(\tilde{R};\mu_{\text{diag}})$. We write $M_i^{\mu_{\text{diag}}}$ when this dependence matters.

Lemma A.6 (Single-axis identity at μ_{diag}). *For the canonical* M_i, $g_i(M_i(\tilde{R});\,\mu_{diag}) = 0$ *and* $g_j(M_i(\tilde{R});\,\mu_{diag}) = g_j(\tilde{R};\,\mu_{diag})$ *for* $j \neq i$.

Proof. Direct computation using $\mathbb{E}_{\mu_{\text{diag}}}[\Phi\phi_i]$ as the i-th column of $\mathbb{E}_{\mu_{\text{diag}}}[\Phi\Phi^\top]$. □

The identity justifies calling M_i *single-axis*, as at μ_{diag}, mitigation moves $\tilde{R}$ along exactly the i-th coordinate of g-space.[1] We emphasize that M_i is an *associational* operator, as it removes ϕ_i-reliance at μ_{diag} in the projection sense, not the contribution of ϕ_i along any causal path to R (see Appendix E.1). Our resulting taxonomy of classifies the resulting outcomes by what the *optimizer does at* $\mu_{\pi^\star}$*, not by what is observable at* μ_{diag}.

Measurement-versus-optimization gap. Single-axis diagnostics evaluate M_i at μ_{diag}, where $|g_i| = 0$ by construction (Lemma A.6). The optimizing policy realizes M_i's effects at $\mu_{\pi^\star}$, where in general $g_i(M_i(\tilde{R}); \mu_{\pi^\star}) \neq 0$ and $g_j(M_i(\tilde{R}); \mu_{\pi^\star}) \neq g_j(\tilde{R}; \mu_{\pi^\star})$ for $j \neq i$. Mitigation moves the proxy along an axis defined at the audit distribution, but optimization responds to a vector defined at a different distribution. This gap is the structural mechanism enabling bias substitution, and it is invisible to any diagnostic that operates at μ_{diag} alone. Section A.5 shows it cannot be closed by any audit-distribution-only evaluation and a closed-form instance is given in Sections E.7 and E.9.

Gauge invariance. The linear reliance g is not invariant under prompt-only reward shifts $\tilde{R}(x, y) \mapsto \tilde{R}(x, y) + b(x)$, while the KL-regularized optimum is. This matters for PPO-style reward whitening [42], where reimplementations of the same mitigation can disagree on g_i. Section E.6 gives a gauge-invariant g_{cent} and verifies that our taxonomy transfers.

Scale invariance. Applying M_i also changes $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$, and KL-regularized policies are not invariant to reward scale [72], so this scale change interacts with optimization non-trivially. Section E.5 gives the corrected scale identity and a scale-invariant variant M_i^{norm}.

A.3 Exploitation Shift and Bias Substitution

The measurement-versus-optimization gap admits qualitatively distinct failure modes in the optimized policy. With $\pi = \pi_\beta^\star(\tilde{R})$ and $\pi' = \pi_\beta^\star(\tilde{R}')$ for the corresponding pre- and post-mitigation KL-regularized optimal policies, we classify outcomes of a single-axis mitigation M_i along two axes: the true reward difference at the KL-regularized optimum,

$$\Delta J := J(\pi', R) - J(\pi, R),$$

and whether optimization pressure rotates onto another feature to create a change.

Definition A.7 (Feature exploitation gap)**.** *The change in policy-induced expectation of feature* ϕ_j *between policies* π, π' *is*

$$\Delta_j(\pi, \pi') = \mathbb{E}_{\mu_{\pi'}}[\phi_j] - \mathbb{E}_{\mu_\pi}[\phi_j]. \tag{6}$$

Crossing these two axes yields five mechanistically distinct regimes (R0, R0_{cont}, R1, R2, R3), see Section E.3 for cell-by-cell verification. We also generalize to ε-banded versions of Δ_j and ΔJ in Section E.8 for empirical studies (finite samples, uncertainties). The R0–R3 regimes classify mitigations by quantities depending on R and on Φ_{sp}, neither of which is fully identifiable from preference data alone. Rotation conditions reference only Φ_{sp} since rotation onto structurally relevant features is not penalized.

Throughout, fix $\beta > 0$, let $\tilde{R}$ be a proxy reward satisfying the regularity Assumption E.1, let $\phi_i \in \Phi_{\text{sp}}$ be the targeted feature, and let M_i be a single-axis mitigation (Definition A.5) targeting ϕ_i, with $\tilde{R}' = M_i(\tilde{R})$ likewise satisfying Assumption E.1.

[1] The canonical M_i can also be seen as the population Frisch-Waugh-Lovell partial-out of ϕ_i from $\tilde{R}$ at μ_{diag}.

Definition A.8 (R0 Successful and contaminated mitigation). *M_i is in regime **R0** if*

$$(i) \quad \Delta_j(\pi, \pi') = 0 \quad \text{for all } \phi_j \in \Phi_{sp} \setminus \{\phi_i\}, \tag{7}$$

$$(ii) \quad \Delta J > 0. \tag{8}$$

*If (ii) holds but (i) is replaced by $\Delta_j(\pi, \pi') \neq 0$, M_i is in regime **R0$_{cont}$** (contaminated success).*

In R0, mitigation strictly improves true reward at the optimized policy without redirecting first-moment exploitation from feature ϕ_i onto another spurious feature. In R0$_{\text{cont}}$, true reward improves despite first-moment redirection onto another spurious axis. The substitution mechanism is active but its true-reward cost is dominated by gains from reducing ϕ_i. R0$_{\text{cont}}$ is the case most easily missed by audit-distribution evaluation, which registers improvement but not rotation.

Definition A.9 (R1 Bias substitution). *M_i is in regime **R1** if*

$$(i) \quad \Delta_j(\pi, \pi') \neq 0 \quad \text{for some } \phi_j \in \Phi_{sp} \setminus \{\phi_i\}, \tag{9}$$

$$(ii) \quad \Delta J \leq 0. \tag{10}$$

We distinguish the sub-cases neutral substitution *($\Delta J = 0$) and* harmful substitution *($\Delta J < 0$).*

R1 is the regime enabled specifically by the measurement-versus-optimization gap, as $|g_i(M_i(\tilde{R}); \mu_{\text{diag}})| = 0$ at the audit distribution by construction (see Lemma A.6), while $\Delta J \leq 0$ at the optimized policy. Audit-distribution diagnostics that target ϕ_i therefore register apparent success regardless of the sign of ΔJ.

Definition A.10 (R2 Overcorrection). *M_i is in regime **R2** if*

$$(i) \quad \Delta_j(\pi, \pi') = 0 \quad \text{for all } \phi_j \in \Phi_{sp} \setminus \{\phi_i\}, \tag{11}$$

$$(ii) \quad \Delta J < 0. \tag{12}$$

In R2, true reward strictly degrades without substitution onto another spurious feature. This effect can be caused by either pushing the targeted feature past zero at the optimization distribution (scale overshoot, fixable by rescaling) or removing a genuinely informative component of a partially informative feature (target misspecification, generically not fixable by rescaling). First-moment ϕ_i observables do not separate the two origins, so they share a regime label. Section E.12 distinguishes them via $\Delta J(c)$ across partial mitigations Mc_i.

Definition A.11 (R3 Silent non-op). *M_i is in regime **R3** if*

$$(i) \quad \Delta_j(\pi, \pi') = 0 \quad \text{for all } \phi_j \in \Phi_{sp} \setminus \{\phi_i\}, \tag{13}$$

$$(ii) \quad \Delta J = 0. \tag{14}$$

In R3, the mitigation altered the proxy in ways the audit distribution may register but the optimized policy does not express. R3 is a generic outcome of projection-type mitigations whenever the targeted feature's optimization footprint was already small, in which case $D_{\text{KL}}(\pi' \,\|\, \pi)$ is correspondingly small at the relevant β.

The R0–R3 classification fixes a single audit distribution μ_{diag}, but cannot cover the failure mode where the regime label changes when μ_{diag} does, e.g., across human and LLM-judge audits. Thus:

Definition A.12 (R4 Audit-distribution sensitivity)**.** *A mitigation construction M_i exhibits* audit-distribution sensitivity *if, holding $(\tilde{R}, R, \Phi_{sp}, \beta)$ fixed, there exist $\mu^{(1)}_{diag} \neq \mu^{(2)}_{diag}$ such that*

$$\pi^\star_\beta\big(M_i^{(1)}(\tilde{R})\big) \quad \textit{and} \quad \pi^\star_\beta\big(M_i^{(2)}(\tilde{R})\big) \quad \textit{fall into different R0–R3 regimes, where } M_i^{(\ell)} := M_i^{\mu^{(\ell)}_{diag}}.$$

R4 is regime-labeled (rather than treated as a property of the deployment setting, like π_{ref}-sensitivity) because μ_{diag} is a designer-controlled input to the mitigation pipeline[2] and formalizes the observation that failure-mode classification is a property of $(R, \tilde{R}, M_i, \mu_{\text{diag}})$. See Figure 2 for example transition.

Definitions A.8–A.12 cover the mechanistic outcome space, with closed-form instantiations of all regimes given in the linear-Gaussian and quadratic-nonlinear settings of Sections E.7 and E.9. The R0–R3 conditions reference only Δ_j and ΔJ, both invariant under the prompt-only shifts. The canonical M_i is not gauge-invariant, but M_i^{cent} (Section E.6) restores classification invariance.

Example Instances. R1 is directly instantiated by the controlled multi-shortcut demonstration in supervised vision [48], where mitigating one labeled shortcut amplifies the unlabeled at fixed optimization distribution. Our GRPO experiment in Section F.1 provides an in-domain harmful-R1 instance under language RLHF, where a length penalty drives the policy into overconfidence while free-form factual accuracy falls. R2 instantiations include length-penalized reward modeling with accuracy loss [3] and capability degradation in DPO bias-mitigation variants [22]. R4 appears in the systematic human-versus-LLM-judge verbosity gap [6, 65]. The length-bias sign flip in SOTA reward models [21] places in R2 on the targeted axis, or R1 under the conservative reading with off-target axes unmeasured. Section E.9 instantiates all regimes and subregimes in closed form under quadratic non-linearity.

A.4 Impossibility and Sufficiency for Audit-Distribution Evaluation

The measurement-versus-optimization gap raises the question whether any benchmark functional defined at the audit distribution can distinguish successful mitigation (R0) from contaminated success (R0_{cont}), substitution (R1), or overcorrection (R2). We prove that they cannot and show that augmenting evaluation with policy-induced distributions suffices (see also Sections E.10 and E.11).

Benchmark class. Let $\mathcal{B}$ denote the class of benchmark functionals depending only on the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag}, with μ_{diag} taken as the empirical evaluation measure. This class subsumes ranking accuracy, pairwise win-rate, preference-prediction calibration, reward–target correlation, the linear-reliance statistic g of Definition A.4, and any composite of these. RewardBench [43], RewardBench2's headline accuracy [52], AlpacaEval [18], and Chatbot Arena [12] all lie in $\mathcal{B}$. Including R as an input strengthens our impossibility result, since R appears identically across the four instances and standard benchmarks lack oracle access to it.

Theorem A.13 (Audit-distribution insufficiency). *Fix $\beta > 0$. There exist four choices of π_{ref} with μ_{diag}, $\tilde{R}$, M_i, R, Φ, Φ_{sp}, and β held fixed s.t.*

(i) *the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag} is identical across the four instances, so B takes the same value on all four for every $B \in \mathcal{B}$;*

(ii) *the optimized policies $\pi^\star_\beta(M_i(\tilde{R}))$ fall into regimes R0, R0_{cont}, R1, and R2 respectively, in the sense of Definitions A.8–A.10.*

[2] μ_{diag} is non-performative in the sense of Perdomo et al. [61].

The proof (Section E.10) constructs four reference policies in a linear-Gaussian setting with fixed Gram, so audit-side observables coincide while policy-side first moments place the optimized policies in distinct regimes. Since only π_{ref} varies, the same reward model can land in any of R0, R0_{cont}, R1, R2 depending on the reference policy it is deployed against, matching the finding of Malik et al. [52] that high audit scores can fail to transfer to PPO under RM-policy lineage mismatch.

Theorem A.13 establishes the *distributional blindspot*. A second, independent blindspot, the *functional blindspot*, affects ordinal benchmarks: $\mathcal{B}_{\text{ord}} \subseteq \mathcal{B}$ (functionals invariant under monotone transformations of $\tilde{R}$) is blind to cardinal-scale shifts that KL-regularized optimization tracks (Corollary E.4). Our experiments in Section F.2 realizes this on a non-linear operator using the post-hoc length calibration of Huang et al. [30]. We find that it drives pooled reward–length correlation near zero ($0.316 \to 0.037$ across five RMs), while within-prompt correlation, which BoN top-1 selection responds to, lands at 0.116 with sign flips on three of four SOTA RMs and $\Delta J < -\varepsilon_J$ on two cells.

We show that augmenting the benchmark input with the policy-induced distributions provably suffices. Let $\mathcal{B}^+$ denote the class of benchmark functionals depending on the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under each of $\mu_{\text{diag}}, \mu_{\pi^\star_\beta(\tilde{R})}, \mu_{\pi^\star_\beta(M_i(\tilde{R}))}$.

Theorem A.14 (Audit-distribution sufficiency). *Fix $\beta > 0$ and let $\pi = \pi^\star_\beta(\tilde{R})$, $\pi' = \pi^\star_\beta(M_i(\tilde{R}))$. With Δ_j and ΔJ as in Definition A.7 and Section A.3, define $B^\star \in \mathcal{B}^+$ by*

$$B^\star = \begin{cases} \text{R0} & \text{if } \Delta J > 0 \text{ and } \Delta_j = 0 \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \\ \text{R0}_{\text{cont}} & \text{if } \Delta J > 0 \text{ and } \Delta_j \neq 0 \text{ for some } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \\ \text{R1} & \text{if } \Delta J \leq 0 \text{ and } \Delta_j \neq 0 \text{ for some } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \\ \text{R2} & \text{if } \Delta J < 0 \text{ and } \Delta_j = 0 \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}. \end{cases} \tag{15}$$

Under Assumptions E.1, A.2, and E.2, with $M_i(\tilde{R})$ likewise satisfying Assumption E.1, for every tuple $(\pi_{\text{ref}}, \tilde{R}, M_i, R, \Phi, \Phi_{\text{sp}})$ with $(\pi, \pi') \in \text{R0} \cup \text{R0}_{\text{cont}} \cup \text{R1} \cup \text{R2}$ in the sense of Definitions A.8–A.10, $B^\star$ returns the correct regime label.

Proof (Section E.11): Δ_j and ΔJ are first moments under μ_π and $\mu_{\pi'}$, both in $\mathcal{B}^+$. The four branches are mutually exclusive by Definitions A.8–A.10, where the sign of ΔJ separates R0 and R0_{cont} from R1 and R2, and the rotation condition separates within each pair. A fifth branch ($\Delta J = 0 \wedge \Delta_j = 0$) extends the classifier to R3, with $D_{\text{KL}}(\pi' \| \pi)$ separating policy-relevant from policy-irrelevant R3. A finite-sample version under noise-floor ε-bands follows from the same construction (Section E.8).

What this pair establishes. Theorems A.13 and A.14 tell us that audit-only inputs cannot separate R0, R0_{cont}, R1, and R2, **even when R itself is granted as an oracle input.** Augmenting with policy-induced distributions and the spurious-feature partition closes the gap.

A.5 Implications for Mitigations Methods and Evaluations

We translate the findings of Theorems A.13 and A.14 into prescriptions for new reward bias mitigation works. We state the corresponding prescriptions for benchmark developers, the multi-axis recommendations, and the operator-variant caveats (M_i^{norm}, M_i^{cent}) in Section E.12.

Method paper prescriptions. A paper claiming successful reward bias mitigation should:

(1) *Evaluate at policy-induced distributions and report* $(\Delta_j, \Delta J)$. By Theorem A.14 this input separates R0, $\mathrm{R0}_{\mathrm{cont}}$, R1, and R2, as no audit-only input does (Theorem A.13). Concretely: run the mitigated reward model in BoN or short PPO against a fixed reference policy.

(2) *Instrument off-target* Δ_j *on every* $\phi_j \in \Phi_{\mathrm{sp}} \setminus \{\phi_i\}$ *at* $\mu_{\pi^\star}$. Reporting only on-target $|g_i| \approx 0$ leaves R0 structurally indistinguishable from $\mathrm{R0}_{\mathrm{cont}}$ and R1, regardless of audit-side strength.

(3) *Report cardinal scale* $\|\tilde{R}\|_{L^2(\mu_{\mathrm{diag}})}$ *pre/post mitigation.* By Corollary E.4, ordinal scores are invariant under $\tilde{R} \mapsto c\tilde{R}$ $(c > 0)$ while $J(\pi^\star_\beta(c\tilde{R}), R)$ varies with c. Since mitigation induces rescaling (Section E.5), ordinal-only evaluation misses scale-driven regime shifts.

(4) *Document* π_{ref}*-sensitivity.* The construction of Theorem A.13 produces four distinct regimes by varying only π_{ref} for fixed $(\tilde{R}, M_i, \mu_{\mathrm{diag}})$. Validate across different reference policies.

Concrete detection procedures for the regime-specific gaps not closed by $B^\star$ are in Section E.12.

Joint adoption. R0 certification is a joint property of the mitigation operator and the evaluation protocol. Method papers cannot unilaterally provide policy-distribution evaluation without supporting benchmark infrastructure and benchmarks cannot unilaterally provide off-target Δ_j measurement without method papers specifying Φ_{sp}. Joint adoption moves the field from R0 claims audit-side scores cannot certify (Theorem A.13) to R0 claims this framework certifies sufficient (Theorem A.14). Extended prescriptions, including benchmark developer recommendations and multi-axis operator guidance, are in Section E.12. We map published mitigation work onto these regimes in Section B.

B Bias Substitution in the Wild

Bias substitution is already pervasive in the published RLHF/DPO mitigation literature, miscategorized as isolated anomalies, capability trade-offs, or judge noise because no prior framework distinguishes it from successful mitigation. Applying the prescriptions of Section A.5 to the published mitigation literature, we find that almost no work provides the inputs Theorem A.14 requires, and where partial evidence is available it points to $\mathrm{R0}_{\mathrm{cont}}$, R1, or R2 rather than R0.

B.1 Reward models exhibit correlated, partially informative spurious features

Bias substitution requires correlated, partially informative spurious features under μ_{diag}, both hold in deployed reward models (Fact G.1–G.17, G.20). Partial informativeness anchors R2's target-misspecification origin, and correlated axes provide directions for pressure rotation. We extend this with length-sycophancy coupling measurements across labeling regimes (Fact G.9, see Section F.3). Across LLM responses to Reddit prompts from Cheng et al. [9] on eight model families, the sycophantic effect on length is +24.3 characters under human labels, +128.4 under LLM judge labels, +154.3 under judge agreement, and **−43.1** under judge disagreement (all $p < 0.01$). The sign reversal documents R4's audit-dependent operator construction empirically. Section E.9 isolates the same mechanism in closed-form phase diagrams of regime transitions driven by these couplings.

B.2 Bias substitution arises under RLHF training

To show that bias substitution arises end to end under policy optimization, we run GRPO [67] on *Llama-3.2-3B-Instruct* [27] with the Skywork-*Reward-V2-Llama-3.1-8B* reward model [49], shaping the reward as $\tilde{R}(x, y) = R_{\mathrm{RM}}(x, y) - \lambda\, n_{\mathrm{tok}}(y)/100$ for $\lambda \in \{0, 4, 8\}$ with four seeds per value. This shaping has the form of the canonical single-axis mitigation M_i of Definition A.5, with a fixed coefficient in place of the estimated reliance. As λ rises, mean response length falls from 204 to 170 tokens and multiple-choice (MMLU) accuracy is unchanged, yet expected calibration error climbs from

0.25 to 0.41, free-form factual accuracy on TriviaQA falls from 0.56 to 0.42, and confidence-correctness AUROC drops from 0.73 to 0.65 (Figure 1, Table 3). The $\lambda = 0$ run leaves calibration intact, so the degradation is caused by the penalty and not by RLHF itself. Optimization pressure rotates off length onto expressed confidence and the true reward proxy of free-form factual accuracy degrades, instantiating harmful R1 in a standard RLHF pipeline. Full setup, training curves, and per-metric results are in Section F.1.

B.3 Bias mitigations that pass audit can fail under optimization

We show that the measurement-versus-optimization gap fails to close with previously published mitigation operators in Section F.2. We evaluate the post-hoc LOESS calibration of Huang et al. [30] and the linear-probe operator of Fein et al. [21] across five reward models under Best-of-N selection. The calibration drives pooled reward-length correlation from 0.316 to 0.037, an audit-side success by construction. All four SOTA reward models acquire negative within-prompt correlations under the calibration, with three exceeding their unmitigated baselines in absolute value. AlpacaEval LC win rate degrades below baseline on two cells and GSM8K BoN accuracy drops 3.6 points on one. Under the ε-banded reading of Section E.8, two cells satisfy $\mathrm{R2}_\varepsilon$ on the targeted axis, or mixed $\mathrm{R1}_\varepsilon$+$\mathrm{R2}_\varepsilon$ with off-target axes unmeasured, both undetectable to any $B \in \mathcal{B}$ by Theorem A.13.

B.4 Published mitigations cannot certify successful mitigations

No published mitigation paper we find provides evidence sufficient to certify R0. Where direct evidence is available, it points to $\mathrm{R0}_{\mathrm{cont}}$, R1, or R2 and for the remainder the regime is undetermined, which is itself the failure mode Theorem A.13 formalizes. We discuss each work in detail in Section H.

Determined regimes. R2 appears in Feng et al. [22], Bu et al. [3], and Huang et al. [30]. R1 is determinately instantiated by the controlled-vision study of Li et al. [48] and, in language RLHF, by our GRPO experiment (Section B.2 and Figure 1). The human-versus-LLM-judge verbosity gap [6, 65] and our Section B.1 sign reversal document R4's audit-dependent operator construction in deployed reward models. R3 has no clean published instance we can identify, as expected from publication bias. Section E.9 additionally instantiates R0, $\mathrm{R0}_{\mathrm{cont}}$, R1 (neutral and harmful), R2, and the operator-level R4 transition in closed form, with finite-N validation.

Split or undetermined. Fein et al. [21], Eisenstein et al. [19], Meng et al. [53], Lu et al. [51], Chen et al. [8], Ye et al. [86], Zhao et al. [91], Liu et al. [50], Fu et al. [24], and Kim et al. [36] all admit multiple regime readings on the evidence reported, ranging across $\mathrm{R0}_{\mathrm{cont}}$, R1, R2, and R3 depending on unmeasured Δ_j at $\mu_{\pi^\star}$ or unreported rescalability.

$\mathrm{R0}_{\mathrm{cont}}$ on audit-side evidence alone. Park et al. [59], Shen et al. [70], Li et al. [45], Wang et al. [80], Ng et al. [55], and Li et al. [47] zero an audit-side metric without measuring off-target Δ_j at $\mu_{\pi^\star}$, leaving R0 indistinguishable from substitution by Theorem A.13.

The closest approaches. Fein et al. [21] survives the within-prompt diagnostic of Section F.2 with $\Delta_J > 0$ on four of five RMs but does not measure off-target Δ_j. Srivastava et al. [75] engages multi-axis mitigation, BoN, and transformation robustness, but its LLM-oracle counterfactuals inherit unmeasured R4 sensitivity. Cai et al. [4] includes PPO results but reports no off-target Δ_j. None jointly delivers the inputs Theorem A.14 requires. These regime ambiguities are not resolvable by current benchmarks. Every benchmark we surveyed [12, 43, 52, 84] sits inside the impossibility class $\mathcal{B}$ of Theorem A.13, see Section E.12.

C Related Work

Reward bias mitigation methods and benchmarks. We position our work against prior mitigation methods and benchmarks in Section B and Section H; our framework classifies what single-axis operators can certify rather than proposing a new one. Concurrent work addresses adjacent problems: probabilistic reward modeling [7], identifiability under correlated probit models [11], emergent misalignment from narrow SFT [2], heterogeneous safety drift under benign fine-tuning that standard proxies fail to predict [35], and a KL-minimal sycophancy correction [68]. None formalize how single-axis mitigations redistribute optimization pressure onto correlated proxies.

Supervised-learning and fairness antecedents. The rotation-onto-correlated-proxies mechanism has supervised-learning precedents in shortcut learning [26], simplicity bias [66], and Whac-A-Mole [48], with a structural analog in fairness gerrymandering [34]. Our contribution provides the first formal taxonomy of these failures, with KL-regularized RLHF as the primary instantiation. Two failure modes absent under empirical risk minimization arise: cardinal-scale sensitivity (Corollary E.4) and π_{ref} dependence (Theorem A.13).

Prior definitions of reward hacking. Existing definitions [23, 40, 41, 73] take a single proxy as given and ask whether optimizing it produces an acceptable policy, with Liu et al. [50] extending to the worst-case proxy in a correlation set. None consider mitigation operators, so prior definitions cannot distinguish a mitigation that reduces reward hacking from one that relocates it, overshoots, or does nothing the optimizer notices.

D Discussion

Limitations. The regime classification uses first-moment drift, which can miss tail shifts for bounded features such as sycophancy indicators. We argue that reward model biases admit to first order corrections to limit the practical severity of this gap (Section G). Definition A.3 partitions Φ into spurious and structurally relevant features rather than decomposing within-feature. However, our approach is conservative under partial informativeness, as it under-flags substitution rather than over-flagging clean mitigations (Section E.3). Also, as soon as anyone targets a feature for debiasing, they assert it belongs to Φ_{sp}, so our taxonomy and impossibility result apply regardless of whether Φ_{sp} is ground-truth identifiable from preference data.

Impact. Current improvement metrics for reward bias mitigation cannot distinguish whether a model has become less reward-hacky from whether it has merely shifted exploitation modes. Our taxonomy and impossibility-sufficiency pair apply to any preference-learning setting with single-axis mitigation operators. We develop language RLHF in detail because reward bias mitigation is a primary lever for shaping non-verifiable model behavior, and the mitigation ambiguity propagates directly into post-deployment alignment claims. We prove this blindspot is structural, as no benchmark functional over audit-distribution observables can distinguish successful mitigation from substitution or overcorrection, regardless of benchmark richness. We establish this both mathematically and empirically across published SOTA reward models and mitigation operators, showing that bias substitution is an active and pervasive failure mode in RLHF systems. To close the gap, we provide a concrete, implementable checklist for both mitigation method developers and benchmark developers.

E Additional Formalizations

E.1 Causal Interpretation and Scope

In this appendix, we outline the causal frame of our formalization in Section A and delimit what the framework does and does not claim about the underlying causal structure of reward.

Underlying structural causal model. We posit, descriptively, that the true reward R and the features Φ are related through an underlying structural causal model $\mathcal{M}$ on (X, Y, R, Φ). We do not assume that $\mathcal{M}$ is identified from preference data $\mathcal{D}$. The reward identifiability characterization of Skalse et al. [72] establishes that preference data identifies R only up to a prompt-only additive shift under the Bradley-Terry likelihood, leaving the cardinal scale of R and by extension the causal structure of $\mathcal{M}$ partially identified at best. The framework of Section A therefore operates within this partial-identification regime and it adopts causal vocabulary to state failure modes precisely, but its definitions and results do not require $\mathcal{M}$ to be recovered from $\mathcal{D}$.

Causal reading of Definition A.3. The corresponding causal reading of Definition A.3 is: ϕ_i is *causally spurious* with respect to R iff ϕ_i has no directed path to R in $\mathcal{M}$. Under positivity-style richness on the support of μ_{diag} (strengthening Assumption E.2) and faithfulness, the causal and observational readings coincide on the partition Φ_{sp} vs. Φ_{struct}. The causal reading is the primitive notion and the observational form its identifiable shadow. Thus, Φ_{sp} vs. Φ_{struct} is best read as the conservative observational projection of an underlying causal partition preference data does not fully identify [72].

Partial informativeness as causal mediation. Partial informativeness (Sections A.1 and E.3) corresponds causally to ϕ_i having both a direct path to R in $\mathcal{M}$ and one or more mediated paths through other features. Length is the canonical instance: a direct contribution via comprehensiveness, plus mediated effects of hedging, sycophancy, and verbosity compensation (see Sections B and G). The conservative partition classifies such features as Φ_{sp}, under-flagging rotation onto causally relevant mediated paths.

Associational mitigation operators. The single-axis mitigations M_i of Definition A.5 are associational: they target the linear reliance statistic g_i, which is an $L^2(\mu_{\text{diag}})$ regression coefficient, not a controlled direct effect. A causal counterpart would target the controlled direct effect of ϕ_i on R, identifiable only under interventions on $\mathcal{M}$ or strong conditional-independence assumptions that preference data does not warrant. Existing single-axis mitigations in the language reward modeling literature [e.g., 21, 30, 58] are associational in this sense, and the regime taxonomy of Section A.3 characterizes their failure modes precisely because of the gap between associational construction and causal response of the optimized policy.

R2 origins in causal language. The two origins of R2 in Definition A.10 restate causally as: scale overshoot (projection coefficient too large at $\mu_{\pi^\star}$ while ϕ_i remains causally spurious) versus target misspecification (ϕ_i has a non-zero direct path to R in $\mathcal{M}$ that the projection has zeroed at μ_{diag}). The rescalability test of Section A.5 discriminates these origins observationally, without identifying $\mathcal{M}$.

R4 and transportability. The audit-distribution sensitivity of Definition A.12 is structurally analogous to the transportability question [e.g. 60]. An identified mitigation at one audit distribution does not need to transport to another. R4 names this phenomenon under partial identification, without formally invoking transportability machinery.

E.2 Standing Assumptions and Optimization Setup

Assumption E.1 (Regularity). *$R, \tilde{R} \in L^2(\mu_{\text{diag}}) \cap L^\infty(\mu_{\pi_{\text{ref}}})$ and $\phi_k \in L^2(\mu_{\text{diag}}) \cap L^\infty(\mu_{\pi_{\text{ref}}})$ for each $k = 1, \ldots, K$, where $\Phi = (\phi_1, \ldots, \phi_K)^\top$ is the feature map of Definition A.1.*

The $L^2(\mu_{\text{diag}})$ inclusion makes $g(\cdot; \mu_{\text{diag}})$ of Definition A.4 well-defined as a Gram-inner-product statistic at the diagnostic measure. The $L^\infty(\mu_{\pi_{\text{ref}}})$ conditions are natural both for bounded-output reward models and for the interpretable surface features this paper considers (length, formatting counts, style indicators). The conditions imply that any $\bar{R}$ obtained as a finite $\mathbb{R}$-linear combination of $\tilde{R}$, R, and the ϕ_k also lies in $L^\infty(\mu_{\pi_{\text{ref}}})$, which covers every $\bar{R}$ used in this paper, including $M_i(\tilde{R})$ of Definition A.5 and the scale-normalised variant M_i^{norm} introduced below. Since $\mu_{\pi_{\text{ref}}}$ is a probability measure, $L^\infty(\mu_{\pi_{\text{ref}}}) \subset L^2(\mu_{\pi_{\text{ref}}})$ holds automatically and no separate $L^2(\mu_{\pi_{\text{ref}}})$ clause is needed.

Optimization setup. Fix $\beta > 0$. For any $\bar{R}$ satisfying Assumption E.1, define the KL-regularized optimizer

$$\pi_\beta^\star(\bar{R}) \;=\; \arg\max_{\pi} \; \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi}\big[\bar{R}(x,y)\big] \;-\; \beta\, D_{\text{KL}}\big(\pi \,\|\, \pi_{\text{ref}}\big),$$

which in the single-turn contextual-bandit regime admits the unique softmax closed form $\pi_\beta^\star(\bar{R})(y \mid x) \propto \pi_{\text{ref}}(y \mid x)\, \exp(\bar{R}(x,y)/\beta)$ recalled in the main text. Essential boundedness of $\bar{R}$ gives two-sided bounds on both $\exp(\bar{R}/\beta)$ and the normalizer $Z(x) = \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[\exp(\bar{R}(x,y)/\beta)]$, so the density ratio $d\mu_{\pi_\beta^\star(\bar{R})}/d\mu_{\pi_{\text{ref}}} = \exp(\bar{R}/\beta)/Z(x)$ is essentially bounded. The chain $\bar{R} \in L^\infty(\mu_{\pi_{\text{ref}}}) \Rightarrow L^\infty(\mu_{\pi_\beta^\star(\bar{R})}) \subset L^2(\mu_{\pi_\beta^\star(\bar{R})})$ then follows immediately, and analogously for each R and each ϕ_k. The policy-level expectations $\mathbb{E}_{\mu_{\pi^\star}}[\phi_j]$ entering Definition A.7 are then finite. Well-definedness of $g(\bar{R}; \mu_{\pi_\beta^\star(\bar{R})})$ additionally requires Gram non-degeneracy at $\mu_{\pi^\star}$. This condition is handled in Section E.3, where the relevant audit distributions are introduced.

E.3 Feature Realizability and the Diagnostic Measure

This section collects four items referenced from Section A. A discussion of the choices available for μ_{diag} (pointed to from the preamble of Section A), the Gram non-degeneracy condition for the decoupled case, the structural Assumption E.2 (feature realizability) on the feature map (invoked by Definition A.3), and the expressiveness remark that motivates reading the regime language of Section A.3 conservatively under partially informative features (pointed to immediately after Definition A.3). The order below matches the narrative order of Section A.

Remark (choice of diagnostic measure). The main text introduces μ_{diag} as an auxiliary distribution on $\mathcal{X} \times \mathcal{Y}$ used for reliance estimation and correlation measurement, distinct from the KL anchor π_{ref}. Two choices are load-bearing for this paper:

(1) *The coupled case* $\mu_{\text{diag}} = \mu_{\pi_{\text{ref}}}$. Recovers the setting of Laidlaw et al. [41] and makes diagnostic statistics directly comparable to their bounds, at the cost of coupling measurement and optimization reference points.

(2) *An annotator-conditioned distribution* $\mu_{\text{diag}}^{\text{human}}$ *or* $\mu_{\text{diag}}^{\text{LLM}}$, reflecting that the same proxy $\tilde{R}$ induces different g-profiles depending on which annotator pool generated the audit data. This formalizes the human-vs-LLM-judge gap documented in Chen et al. [6], Movva et al. [54], Saito et al. [65], Zheng et al. [92], and is what R4 (Definition A.12) operationalizes.

Other valid choices, e.g., the preference-data distribution used to train $\tilde{R}$, curated audit sets, deployment distributions, and round-anchored distributions in iterative RLHF [87], yield different g-profiles and shift substitution diagnostics accordingly. Substitution claims are always relative to the chosen diagnostic measure, and the choice should be stated explicitly in applications.

Where $g(\bar{R}; \mu_{\pi^\star_\beta(\bar{R})})$ is referenced in Section A (e.g., the measurement-vs-optimization gap following Lemma A.6), well-definedness requires Gram non-degeneracy at $\mu_{\pi^\star}$, which follows from Assumption A.2 (non-degeneracy) whenever $d\mu_{\pi^\star}/d\mu_{\text{diag}}$ is bounded above and below. In the coupled case ($\mu_{\text{diag}} = \mu_{\pi_{\text{ref}}}$), the bound follows from Assumption E.1 (regularity). In the decoupled case ($\mu^{\text{human}}_{\text{diag}}, \mu^{\text{LLM}}_{\text{diag}}$), it is a separate regularity condition assumed in the relevant statements.

Assumption E.2 (Feature Realizability). *For μ_{diag}-almost every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and every feature index $i \in \{1, \ldots, K\}$, there exists at least one $y' \in \mathcal{Y}$ with $\phi_j(x, y') = \phi_j(x, y)$ for all $j \neq i$ and $\phi_i(x, y') \neq \phi_i(x, y)$.*

Assumption E.2 is the non-vacuity condition for Definition A.3. Without it, ϕ_i can be functionally determined by the prompt and the other features on $\mathcal{Y}$, so further conditioning on ϕ_i leaves the conditional expectation in Definition A.3 unchanged and every feature is trivially classified as spurious. Assumption E.2 is therefore a richness condition on the response space *under* μ_{diag}, as $\mathcal{Y}$ must be rich enough where the diagnostic measure places mass that each feature axis admits independent perturbation.

We treat g and Δ_j as observational statistics under μ_{diag} and $\mu_{\pi^\star}$ throughout, as an interventional reading along ϕ_i would require strengthening Assumption E.2 to a richness condition on counterfactual realization, which we do not invoke.

Even after weakening to the μ_{diag}-a.e. form, Assumption E.2 is fragile for natural-language features on Φ's own terms. Consider $\Phi = \{\phi_{\text{length}}, \phi_{\text{exclamation\_count}}\}$. Assumption E.2 demands that μ_{diag}-a.e. there exist a response y' matching y on exclamation count but differing in length, and conversely. The first direction is plausible, as long and short variants with the same exclamation count are typically realizable. The second is genuinely constrained, as at fixed length, the realizable range of exclamation counts is bounded above by length itself, and at very short lengths the counterfactual set may be empty. Even within Φ, the response space does not cleanly factor into independent feature axes. This failure mode is specific to features that count items within the response (exclamation marks, bullet points, headers), since their range is mechanically tied to length. Features operationalized as binary indicators or response-level scalars do not have this bound (e.g., sycophancy and confidence in Sections B and F.3), and Assumption E.2 holds approximately for these pairs.

Feature misspecification is itself a major failure mode in practice. Under Assumption E.2, Definition A.3 cleanly separates spurious from structurally relevant features, and the canonical mitigation M_i of Definition A.5 targets a well-defined coordinate of g-space. When Assumption E.2 fails, the definitions of Section A.1 should be read as approximations, and the binary spurious/structurally-relevant partition as the best dichotomous summary of a genuinely continuous informativeness spectrum.

Remark (scope of the binary partition under partial informativeness). Definition A.3 partitions Φ into spurious and structurally relevant features. The partition is scope-matched to the operator class deployed in the literature (whole-feature projections, see Section A.1) and to what preference data identifies, but it does not separate within-feature components for partially informative features in the empirical setting surveyed in Section B, where many natural features (foremost length) act as mediators for multiple mechanisms (sycophancy, epistemic uncertainty, informativeness) rather than as pure spurious proxies or pure structural drivers.

Under partial informativeness, the regime distinction of Section A.3 still applies, but with a scoping caveat: bias substitution becomes a *conservative classification* of the ways a single-axis mitigation can fail, because real features admit additional substitution modes that the binary partition cannot express. Concretely, the best $L^2(\mu_{\text{diag}})$ approximation of ϕ_i by a function of R alone (writing $\phi_i = \mathbb{E}_{\mu_{\text{diag}}}[\phi_i \mid R] + \phi_i^\perp$, i.e., decomposing ϕ_i into "the part that tracks R" and

a residual) does not in general coincide with the causal-versus-spurious decomposition of ϕ_i under the underlying structural causal model $\mathcal{M}$ (Section E.1). The first is a μ_{diag}-level regression decomposition, the second a structural statement under $\mathcal{M}$. Reconciling the two requires additional structure, e.g., a causal model, conditional-independence assumptions, or auxiliary interventions, and we defer partial-informativeness extensions to future work.

Exhaustiveness of the regime taxonomy. The two axes underlying Section A.3 — whether mitigation rotates first-moment exploitation onto another spurious feature ($\Delta_j(\pi, \pi') \neq 0$ for some $\phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}$) and the sign of the change in true reward ($\Delta J = J(\pi', R) - J(\pi, R)$) — partition the outcome space into six cells, summarized in Table 1. Five regime labels (R0, R0_{cont}, R1, R2, R3) suffice because R1 absorbs both $\Delta J = 0$ (neutral substitution) and $\Delta J < 0$ (harmful substitution) under Definition A.9, while R0 and R0_{cont} receive separate labels because the rotation-with-improvement corner is precisely the case audit-distribution evaluation most readily mistakes for clean R0 (cf. Theorem A.13). The asymmetry is therefore deliberate rather than ad hoc: R0_{cont} earns a separate label for downstream emphasis in the impossibility result, whereas neutral and harmful R1 are both already failures of ΔJ and the sub-case treatment suffices.

	No rotation ($\Delta_j = 0$ on $\Phi_{\text{sp}} \setminus \{\phi_i\}$)	Rotation ($\Delta_j \neq 0$ for some $\phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}$)
$\Delta J > 0$	R0 (Definition A.8)	R0_{cont} (Definition A.8)
$\Delta J = 0$	R3 (Definition A.11)	R1, neutral (Definition A.9)
$\Delta J < 0$	R2 (Definition A.10)	R1, harmful (Definition A.9)

Table 1. Cell-by-cell partition of single-axis mitigation outcomes by the rotation predicate on $\Phi_{\text{sp}} \setminus \{\phi_i\}$ and the sign of ΔJ. Six cells, five regime labels: R1 covers the two $\Delta J \leq 0$ rotation cells under the neutral/harmful sub-case distinction. R4 (Definition A.12) is transversal to this table and indexes how regime membership changes across audit distributions μ_{diag}.

The R4 axis (Definition A.12) does not appear in Table 1 because R4 is not an outcome cell but a property of how cell membership shifts across μ_{diag} choices, as discussed in Section A.3 and instantiated empirically in Section F.3.

Spurious-rotation-with-improvement corner. Definition A.9 (R1, bias substitution) requires $\Delta J \leq 0$. The complementary corner $\Delta_j \neq 0$ for some $\phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}$ with $\Delta J > 0$ is regime R0_{cont} per Definition A.8. Two mechanisms produce R0_{cont}: (i) rotation onto a feature classified spurious in the conservative partition but partially informative in the true reward (an in-scope manifestation of the conservatism discussed above), and (ii) under the strict binary partition, rotation onto a genuinely spurious ϕ_j where structural-feature gains from reducing ϕ_i outweigh the spurious cost on ϕ_j.

The practical consequence for the rest of the paper is that substitution claims proved under Assumption E.2 (feature realizability) should be read as a floor rather than a complete characterization: when Assumption E.2 fails, at least the substitution mode the binary partition captures remains available to the optimizing policy, and generically additional modes are available too.

E.4 Equivalent Epsilon-Formulation for Local Spuriousness

Definition A.3 states spuriousness as conditional mean independence at μ_{diag}. We derive the equivalent $o(\varepsilon)$ formulation for completeness.

Let $\mathcal{Y}_{-i}(x, y) := \{y' : \phi_j(x, y') = \phi_j(x, y)\ \forall j \neq i\}$ denote the i-fiber through (x, y), and decompose $\mu_{\text{diag}} = D(x)\, \pi_{\text{audit}}(y \mid x)$. For a Markov kernel κ with $\kappa(\mathcal{Y}_{-i}(x, y) \mid x, y) = 1$ for μ_{diag}-a.e. (x, y), define the mixture-perturbed conditional

$$\pi_\varepsilon^\kappa(y' \mid x) := (1 - \varepsilon)\, \pi_{\text{audit}}(y' \mid x) + \varepsilon \int \kappa(y' \mid x, y)\, \pi_{\text{audit}}(dy \mid x), \qquad \varepsilon \in [0, 1]. \tag{16}$$

Consider the perturbation condition

$$\mathbb{E}_{x\sim D,\, y'\sim\pi_\varepsilon^\kappa(\cdot|x)}\big[R(x,y')\big] \;-\; \mathbb{E}_{\mu_{\text{diag}}}\big[R(x,y)\big] \;\in\; o(\varepsilon) \quad \text{as } \varepsilon \downarrow 0. \tag{17}$$

The left-hand side equals $\varepsilon \cdot \big(\mathbb{E}_{\kappa\otimes\mu_{\text{diag}}}[R] - \mathbb{E}_{\mu_{\text{diag}}}[R]\big)$ for mixture perturbations, so the $o(\varepsilon)$ condition is equivalent to $\mathbb{E}_{\kappa\otimes\mu_{\text{diag}}}[R] = \mathbb{E}_{\mu_{\text{diag}}}[R]$. Requiring this for every kernel κ implies the conditional mean independence of Definition A.3, and the two are equivalent when R is $\sigma(\Phi)$-measurable (so that within-fiber variation of R collapses to variation in ϕ_i).

Verification of downstream results. Theorem A.13 and the closed-form regime instantiations in Sections E.7 and E.9 use a true reward of the form $R = w\phi_3$, which is $\sigma(\Phi)$-measurable. Thus, the conditional mean independence and strict point-wise invariance on i-fibers coincide in this regime. The constructed instances satisfy Definition A.3 and the proofs transfer without modification. Theorem A.14 is a correctness statement on the classifier $B^\star$ given $(\Delta_j, \Delta J)$ and Φ_{sp} as inputs. It places no assumption on the functional form of R and transfers regardless of how Φ_{sp} is defined.

E.5 Scale Change Induced by Mitigation and a Scale-Invariant Variant

Immediately after Lemma A.6, we note that applying M_i changes $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ through both a diagonal and a cross-correlation contribution, and that this scale change interacts non-trivially with the KL-regularized optimizer because $\pi^\star_\beta$ is not invariant to positive rescaling of the reward. In this section, we give the exact scale identity for M_i, isolate the mechanism by which scale change conflates with axis reallocation at fixed β, and define the scale-invariant variant M_i^{norm} that separates the two effects.

Scale identity. Recall the Gram matrix G of Assumption A.2, with entries $G_{ij} = \mathbb{E}_{\mu_{\text{diag}}}[\phi_i\phi_j]$, let $g = g(\tilde{R};\mu_{\text{diag}})$ abbreviate the linear-reliance vector of Definition A.4, and write $\tilde{R}' = M_i(\tilde{R}) = \tilde{R} - g_i\phi_i$ for the canonical mitigation of Definition A.5. Expanding the $L^2(\mu_{\text{diag}})$-norm of $\tilde{R}'$,

$$\|\tilde{R}'\|^2_{L^2(\mu_{\text{diag}})} \;=\; \|\tilde{R}\|^2_{L^2(\mu_{\text{diag}})} \;-\; 2g_i\,\mathbb{E}_{\mu_{\text{diag}}}[\phi_i\tilde{R}] \;+\; g_i^2\,G_{ii},$$

and using $\mathbb{E}_{\mu_{\text{diag}}}[\phi_i\tilde{R}] = (Gg)_i = \sum_j G_{ij}g_j$ gives the exact identity

$$\|\tilde{R}'\|^2_{L^2(\mu_{\text{diag}})} \;=\; \|\tilde{R}\|^2_{L^2(\mu_{\text{diag}})} \;-\; g_i\left[\, g_i\,G_{ii} \;+\; 2\sum_{j\neq i} G_{ij}\,g_j \right].$$

The sign of the change is governed by the bracketed expression relative to g_i: the norm strictly decreases whenever the bracketed term is co-signed with g_i, and can *increase* when the cross-correlation contribution $2\sum_{j\neq i} G_{ij}g_j$ is oppositely signed and sufficiently large in magnitude relative to the diagonal term g_iG_{ii}. A parameter choice exhibiting this norm increase is given in Section E.7. The diagonal-only approximation $\|\tilde{R}'\|^2 \approx \|\tilde{R}\|^2 - g_i^2\,G_{ii}$ is correct only when G is diagonal at μ_{diag}, i.e., when the diagnostic measure renders the feature axes orthogonal. This condition generically fails for interpretable feature sets such as length, formatting, and style indicators, which are empirically correlated on any natural μ_{diag}.

Scale-invariant variant. The $L^2(\mu_{\text{diag}})$-normalized mitigation

$$M_i^{\text{norm}}(\tilde{R}) \;=\; \alpha\,\big(\tilde{R} - g_i\phi_i\big), \qquad \alpha := \frac{\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}}{\|\tilde{R} - g_i\phi_i\|_{L^2(\mu_{\text{diag}})}},$$

preserves $\|\cdot\|_{L^2(\mu_{\text{diag}})}$ by construction. Its linear reliance at μ_{diag} is α times that of $M_i(\tilde{R})$, so $g_i(M_i^{\text{norm}}(\tilde{R}); \mu_{\text{diag}}) = 0$ and $g_j(M_i^{\text{norm}}(\tilde{R}); \mu_{\text{diag}}) = \alpha\, g_j(\tilde{R}; \mu_{\text{diag}})$ for $j \neq i$: the i-th coordinate is zeroed and the remaining axes are rescaled uniformly.

Effective-β shift. By KL scale-equivariance $\pi^\star_\beta(c\bar{R}) = \pi^\star_{\beta/c}(\bar{R})$ and the relation $M_i^{\text{norm}}(\tilde{R}) = \alpha\, M_i(\tilde{R})$,

$$\pi^\star_\beta\big(M_i(\tilde{R})\big) \;=\; \pi^\star_{\alpha\beta}\big(M_i^{\text{norm}}(\tilde{R})\big).$$

The unnormalized comparison $\pi^\star_\beta(\tilde{R})$ vs $\pi^\star_\beta(M_i(\tilde{R}))$ at fixed β therefore equals the normalized comparison $\pi^\star_\beta(\tilde{R})$ vs $\pi^\star_{\alpha\beta}(M_i^{\text{norm}}(\tilde{R}))$, differing from the same-β normalized comparison by exactly an effective-β shift from β to $\alpha\beta$. This is a cardinal-distortion mechanism rather than an ordinal one, as two proxies producing identical pairwise rankings on μ_{diag} can differ in $L^2(\mu_{\text{diag}})$-norm and induce different $\pi^\star_\beta$ at fixed β. This mechanism is in line with the Section 2 ordinal-vs-cardinal gap, which motivates reporting $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ alongside Δ_j in empirical instantiations.

As a practical consequence, under M_i^{norm}, any observed change in $\mathbb{E}_{\pi^\star_\beta}[\phi_j]$ at fixed β is attributable to axis reallocation rather than scale, and Δ_j values can differ in sign between M_i and M_i^{norm} when α is far from 1, classifying the same proxy into different R0-R3 regimes under the two mitigations. Diagnosing R1 versus R3 cleanly therefore calls for M_i^{norm}, since under M_i a nonzero Δ_j admits both reallocation and scale interpretations and requires a β-scan or pre/post norm reporting to disambiguate.

Deployment-time normalization. Standard practices in PPO-based RLHF for stability like per-batch reward whitening [42] apply a sample-dependent shift and scale to $\tilde{R}$ at each training step. In expectation the shift approximates a constant under the current rollout distribution and is therefore gauge-equivalent at the population level (no policy effect, by Section A.2), but the scale component divides $\tilde{R}$ by an estimate of $\text{Std}_{\mu_{\pi_t}}(\tilde{R})$, inducing an effective KL coefficient $\beta \cdot \text{Std}_{\mu_{\pi_t}}(\tilde{R})$ that M_i^{norm} does not control. M_i^{norm} fixes the audit-side $L^2(\mu_{\text{diag}})$ scale, while whitening rescales against the (non-stationary) training-side dispersion. The (M_i, M_i^{norm}) comparison therefore does not exhaust the scale story under whitened RLHF, and empirical instantiations should report training-side reward dispersion alongside $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ when interpreting Δ_j across mitigations.

E.6 Gauge-Invariant Linear Reliance

In this section, we give a gauge-invariant variant of the linear reliance statistic and verify that the Section A.3 regime classification transfers under it. Define the prompt-conditional deviations under μ_{diag},

$$\tilde{R}^{\text{cent}}(x,y) = \tilde{R}(x,y) - \mathbb{E}_{\mu_{\text{diag}}}[\tilde{R} \mid x], \qquad \bar{\Phi}(x,y) = \Phi(x,y) - \mathbb{E}_{\mu_{\text{diag}}}[\Phi \mid x].$$

Assumption E.3 (Non-degeneracy). *The centered Gram matrix $\bar{G} := \mathbb{E}_{\mu_{\text{diag}}}[\bar{\Phi}\bar{\Phi}^\top] \in \mathbb{R}^{K\times K}$ is positive definite: $\bar{G} \succ 0$.*

Note that Assumption E.3 does not follow from Assumption A.2, as any feature that is constant in y given x is non-degenerate pooled but vanishes after within-prompt centering. The *centered linear reliance* is

$$g_{\text{cent}}(\tilde{R};\, \mu_{\text{diag}}) = \big(\mathbb{E}_{\mu_{\text{diag}}}[\bar{\Phi}\bar{\Phi}^\top]\big)^{-1}\mathbb{E}_{\mu_{\text{diag}}}[\bar{\Phi}\,\tilde{R}^{\text{cent}}] \;\in\; \mathbb{R}^K.$$

For near-collinear or learned feature sets, g_{cent} should be read as a ridge-regularized estimate.

Gauge invariance. Under $\tilde{R} \mapsto \tilde{R} + b(x)$ for any prompt-only b, $\mathbb{E}[\tilde{R} + b \mid x] = \mathbb{E}[\tilde{R} \mid x] + b(x)$, so $\tilde{R}^{\text{cent}}$ is unchanged and $\bar{\Phi}$ does not involve $\tilde{R}$. Therefore $g_{\text{cent}}(\tilde{R} + b;\, \mu_{\text{diag}}) = g_{\text{cent}}(\tilde{R};\, \mu_{\text{diag}})$. The KL-regularized optimum shares this invariance, so g_{cent} is a reliance statistic on the same gauge equivalence class the policy respects.

Relation to g. The pooled second moments decompose as

$$\mathbb{E}[\Phi\Phi^\top] = \mathbb{E}[\tilde{\Phi}\tilde{\Phi}^\top] + \mathbb{E}\big[\mathbb{E}[\Phi|x]\mathbb{E}[\Phi|x]^\top\big], \qquad \mathbb{E}[\Phi\tilde{R}] = \mathbb{E}[\tilde{\Phi}\,\tilde{R}^{\text{cent}}] + \mathbb{E}\big[\mathbb{E}[\Phi|x]\mathbb{E}[\tilde{R}|x]\big]$$

so g is a matrix-weighted combination of the within-prompt regression (g_{cent}) and the between-prompt regression of conditional means. The two statistics coincide only in degenerate cases (e.g., vanishing between-prompt covariation) and can differ substantially when Φ carries strong prompt-level structure. This structure is plausible for length, where prompt difficulty drives average response length.

Canonical mitigation and Lemma A.6 analogue. The centered canonical mitigation is

$$M_i^{\text{cent}}(\tilde{R})(x,y) = \tilde{R}(x,y) - g_{\text{cent},i}(\tilde{R};\mu_{\text{diag}})\,\phi_i(x,y).$$

Direct computation in the centered inner product gives $g_{\text{cent},i}(M_i^{\text{cent}}(\tilde{R});\mu_{\text{diag}}) = 0$ and $g_{\text{cent},j}(M_i^{\text{cent}}(\tilde{R});\mu_{\text{diag}}) = g_{\text{cent},j}(\tilde{R};\mu_{\text{diag}})$ for $j \neq i$, using that $\mathbb{E}_{\mu_{\text{diag}}}[\tilde{\Phi}\bar{\phi}_i]$ is the i-th column of $\mathbb{E}_{\mu_{\text{diag}}}[\tilde{\Phi}\tilde{\Phi}^\top]$. The variant $\tilde{R} - g_{\text{cent},i}\bar{\phi}_i$ differs from $M_i^{\text{cent}}(\tilde{R})$ by $g_{\text{cent},i}\,\mathbb{E}[\phi_i \mid x]$, a prompt-only function. Given the above-stated gauge invariance, the two variants are policy-equivalent and produce identical g_{cent}.

Framework transfer. Definition A.4 ports to g_{cent} verbatim, and Definition A.5 ports as the requirement $|g_{\text{cent},i}(\tilde{R}';\mu_{\text{diag}})| < |g_{\text{cent},i}(\tilde{R};\mu_{\text{diag}})|$. The Lemma A.6 analogue above shows M_i^{cent} qualifies. Definitions A.8–A.11 (R0-R3) port unchanged because their criteria reference policy-induced expectations $\Delta_j(\pi,\pi')$ and ΔJ, not the reliance statistic.

Gauge invariance of the regime classification. The Section A.2 obstruction is that under $\tilde{R} \mapsto \tilde{R} + b(x)$ for prompt-only b, $M_i(\tilde{R}+b)$ differs from $M_i(\tilde{R}) + b$ by $(g_i(\tilde{R}) - g_i(\tilde{R}+b))\phi_i$, so $\pi_\beta^\star(M_i(\tilde{R}+b)) \neq \pi_\beta^\star(M_i(\tilde{R}))$ and $\Delta_j, \Delta J$ can change. Replacing M_i by M_i^{cent} removes the obstruction. By gauge invariance of g_{cent}, $g_{\text{cent},i}(\tilde{R}+b) = g_{\text{cent},i}(\tilde{R})$, so $M_i^{\text{cent}}(\tilde{R}+b) = M_i^{\text{cent}}(\tilde{R}) + b$. Since the KL-regularized optimum is invariant under prompt-only shifts,

$$\pi_\beta^\star(\tilde{R}+b) = \pi_\beta^\star(\tilde{R}), \qquad \pi_\beta^\star(M_i^{\text{cent}}(\tilde{R}+b)) = \pi_\beta^\star(M_i^{\text{cent}}(\tilde{R})),$$

hence $\Delta_j(\pi,\pi')$ and ΔJ are gauge-invariant and the R0-R3 regime is unchanged. R4 (see Definition A.12) also ports verbatim, as g_{cent} retains its μ_{diag}-dependence and only the gauge of $\tilde{R}$ is factored out. Thus, $M_i^{\text{cent},\mu_{\text{diag}}}$ depends on μ_{diag} in the same sense as $M_i^{\mu_{\text{diag}}}$ in Definition A.12 and audit-distribution sensitivity is defined identically.

Note that M_i and M_i^{cent} applied to the same proxy $\tilde{R}$ may still occupy different R0-R3 regimes, as different mitigations induce different policies. However, this difference reflects the choice of reliance estimator as a degree of freedom in the mitigation pipeline, independent of gauge. Auditing a proxy under both statistics localizes the targeted reliance, as a feature on which $|g_i|$ is large but $|g_{\text{cent},i}|$ is small flags reliance that lives in prompt-level structure the optimizing policy averages over rather than acts on.

E.7 Non-Vacuity of the Regime Taxonomy via Linear-Gaussian Instantiation

This appendix exhibits closed-form instantiations of the regimes and gaps introduced in Section A, establishing that the taxonomy is non-vacuous within our framework's own assumptions. The constructions are not intended as empirical models of RLHF and we show the empirical instantiations in Sections B and F.3.

The construction below satisfies Assumptions E.1 (regularity), A.2 (non-degeneracy of the Gram at μ_{diag}), and E.2 (feature realizability), as well as Gram non-degeneracy at $\mu_{\pi^\star}$. We verify each below.

Shared construction. We work with a trivial prompt space ($\mathcal{X}$ a singleton, suppressed in notation) and continuous response space $\mathcal{Y} = \mathbb{R}^3$. The feature map is $\Phi(y) = (\phi_1, \phi_2, \phi_3)$ with $\phi_k(y) = y_k$. We fix true and proxy rewards

$$R(y) = w\,\phi_3(y), \qquad \tilde{R}(y) = a\,\phi_1(y) + b\,\phi_2(y) + w\,\phi_3(y),$$

with $w > 0$ and $(a, b) \in \mathbb{R}^2$. By Definition A.3, $\Phi_{\text{sp}} = \{\phi_1, \phi_2\}$ and $\Phi_{\text{struct}} = \{\phi_3\}$ by construction. Reference policy and diagnostic measure are zero-mean Gaussians on $\mathbb{R}^3$,

$$\pi_{\text{ref}} = \mathcal{N}(0, I_3), \qquad \mu_{\text{diag}} = \mathcal{N}(0, \Sigma),$$

where Σ is positive definite with $\Sigma_{kk} = 1$ and off-diagonal $\Sigma_{12} = \rho \in (-1, 1)$, $\Sigma_{13} = \Sigma_{23} = 0$. Fix the KL parameter $\beta > 0$ and target the spurious feature $i = 1$ throughout. We treat $\mathbb{R}^3$ as the formal response space. For the $L^\infty(\mu_{\pi_{\text{ref}}})$ clause of Assumption E.1, restrict to a sufficiently large bounded set on which all formulas below hold to arbitrary precision in the unrestricted limit.[3]

Verification of assumptions. Assumption E.1 holds under truncation as noted. Assumption A.2 holds because $\mathbb{E}_{\mu_{\text{diag}}}[\Phi\Phi^\top] = \Sigma \succ 0$. Assumption E.2 holds because $\mathcal{Y} = \mathbb{R}^3$ has full support under any positive-definite Gaussian, so for μ_{diag}-a.e. y and any i, varying y_i while holding $y_{j\neq i}$ fixed remains in the support. Gram non-degeneracy at $\mu_{\pi^\star}$ follows from the closed form below, where $\mu_{\pi^\star}$ is itself Gaussian with positive-definite covariance.

Closed forms for the optimum. For any linear reward $\bar{R}(y) = c^\top y$ with $c \in \mathbb{R}^3$, the KL-regularized optimum against π_{ref} satisfies $\pi^\star_\beta(\bar{R}) = \mathcal{N}(c/\beta, I_3)$ by completing the square in the softmax form of Equation (2). Two consequences used throughout: (i) the policy-induced first moment is $\mathbb{E}_{\mu_{\pi^\star}}[\phi_k] = c_k/\beta$, so $\Delta_j(\pi, \pi') = (c'_j - c_j)/\beta$ for any pair of linear rewards with coefficients c, c'; (ii) the plain return is $J(\pi^\star_\beta(\bar{R}), R) = \sum_k w_k c_k/\beta$ where w_k are the coefficients of R, so under our $R = w\phi_3$, only the ϕ_3-coefficient of the proxy contributes to ΔJ.

Linear reliance and canonical mitigation. The reliance vector at μ_{diag} is $g(\tilde{R}; \mu_{\text{diag}}) = \Sigma^{-1}\mathbb{E}_{\mu_{\text{diag}}}[\Phi\tilde{R}]$. Since $\tilde{R}$ is linear in Φ with coefficient vector (a, b, w) and $\mathbb{E}_{\mu_{\text{diag}}}[\Phi\Phi^\top] = \Sigma$, we have $g(\tilde{R}; \mu_{\text{diag}}) = (a, b, w)$ exactly. The canonical mitigation $M_1(\tilde{R})(y) = \tilde{R}(y) - a\,\phi_1(y) = b\,\phi_2 + w\,\phi_3$ has reliance $g(M_1(\tilde{R}); \mu_{\text{diag}}) = (0, b, w)$ by Lemma A.6.

Measurement-vs-optimization gap. We progress through three settings of $(\pi_{\text{ref}}, \mu_{\text{diag}})$ to isolate where the gap of Section A.2 can manifest in this construction.

Isotropic reference and audit ($\pi_{\text{ref}} = \mu_{\text{diag}} = \mathcal{N}(0, I_3)$): the mitigated proxy induces $\pi' = \pi^\star_\beta(M_1(\tilde{R})) = \mathcal{N}((0, b, w)/\beta, I_3)$. The policy-side Gram is $\mathbb{E}_{\mu_{\pi'}}[\Phi\Phi^\top] = I_3 + (0, b, w)^\top(0, b, w)/\beta^2$, and

$$g_1(M_1(\tilde{R}); \mu_{\pi'}) = \left(\mathbb{E}_{\mu_{\pi'}}[\Phi\Phi^\top]^{-1}\,\mathbb{E}_{\mu_{\pi'}}[\Phi\, M_1(\tilde{R})]\right)_1 = 0,$$

since $\mathbb{E}_{\mu_{\pi'}}[\phi_1 M_1(\tilde{R})] = 0$ in this case.

Isotropic reference, correlated audit ($\pi_{\text{ref}} = \mathcal{N}(0, I_3)$, $\mu_{\text{diag}} = \mathcal{N}(0, \Sigma)$ with $\Sigma_{12} = \rho \neq 0$): the canonical mitigation is $M_1(\tilde{R}) = \tilde{R} - g_1(\tilde{R}; \mu_{\text{diag}})\phi_1$ with $g_1(\tilde{R}; \mu_{\text{diag}}) = a$ unchanged (since $\Sigma_{13} = 0$ keeps ϕ_3's contribution clean and the (ϕ_1, ϕ_2) block of Σ^{-1} recovers a for a linear proxy). Evaluating reliance at $\mu_{\pi'} = \mathcal{N}((0, b, w)/\beta, I_3)$ gives Gram $I_3 + (0, b, w)^\top(0, b, w)/\beta^2$, diagonal in the ϕ_1 row, so $g_1(M_1(\tilde{R}); \mu_{\pi'}) = 0$ here as well.

Coupled correlated case ($\pi_{\text{ref}} = \mu_{\text{diag}} = \mathcal{N}(0, \Sigma)$ with $\Sigma_{12} = \rho \neq 0$): the policy-side Gram now inherits the audit cross-correlation, $\mu_{\pi'} = \mathcal{N}(\Sigma c/\beta, \Sigma)$ for $c = (0, b, w)$, and one might expect a non-trivial g_1. It does not arise, for a

[3] The closed-form expressions extend continuously to the unrestricted Gaussian limit.

structural reason: for any linear proxy $\bar{R} = c^\top y$ on $\Phi(y) = y$ and any non-degenerate measure $\mu = \mathcal{N}(m, V)$,

$$g(\bar{R}; \mu) = \left(\mathbb{E}_\mu[\Phi\Phi^\top]\right)^{-1} \mathbb{E}_\mu[\Phi\, \bar{R}] = (V + mm^\top)^{-1}(V + mm^\top)\, c = c,$$

so g recovers the coefficient vector exactly and is invariant to μ. Applied to $c = (0, b, w)$, this gives $g_1(M_1(\tilde{R}); \mu_{\pi'}) = 0$ in the coupled case as well, and the g-level gap is degenerate throughout the linear-Gaussian closed form.

The substantive gap manifests instead through the policy-induced first moment. In the coupled correlated case, $\mathbb{E}_{\mu_{\pi'}}[\Phi] = \Sigma c/\beta$, so

$$\mathbb{E}_{\mu_{\pi'}}[\phi_1] \;=\; (\Sigma c)_1/\beta \;=\; \rho\, b/\beta \;\neq\; 0$$

whenever $\rho \neq 0$ and $b \neq 0$, even though g_1 vanishes at every measure. The mechanism is the cross-correlation ρ: at μ_{diag} the projection orthogonalizes $M_1(\tilde{R})$ against ϕ_1 in the Gram-inner-product sense, but the policy under $M_1(\tilde{R})$ shifts the mean of (ϕ_2, ϕ_3) by $(b, w)/\beta$, and any ϕ_1-ϕ_2 coupling at μ_{diag} propagates into ϕ_1 first-moment drift at the policy distribution. This is the form of the gap the regime taxonomy of Section A.3 is sensitive to, since Definitions A.8-A.11 are stated in terms of Δ_j and ΔJ rather than g_i at $\mu_{\pi^\star}$.

Remark (non-linear proxies activate the g-level gap). The linear-Gaussian construction is structurally incapable of exhibiting a non-degenerate g-level gap, because the OLS-style reliance map $g(\bar{R}; \mu)$ recovers c exactly for any linear $\bar{R} = c^\top y$, independent of μ. Activating a non-trivial g-level gap requires breaking exact representability of $\tilde{R}$ in $\text{span}(\Phi)$ in a way that couples to moments of μ_{diag} that vary across audit distributions. Augmenting $\tilde{R}$ with a non-linear term outside $\text{span}(\Phi)$ (e.g., an interaction $\kappa\phi_1\phi_2$ at non-symmetric μ_{diag}, or a higher-order term whose relevant cross-moments differ across $\mu^{(1)}_{\text{diag}}$ and $\mu^{(2)}_{\text{diag}}$) makes $g(\tilde{R}; \mu_{\text{diag}})$ depend on those moments. Under such an augmentation, $M_1^{\mu^{(1)}_{\text{diag}}} \neq M_1^{\mu^{(2)}_{\text{diag}}}$ as operators and the operator-level audit sensitivity of Definition A.12 becomes instantiable with π_{ref} held fixed across audit distributions. While we establish non-vacuity with the linear-Gaussian, Section E.9 provides the closed-form non-linear instantiation. Real reward models are non-linear in any tractable feature basis, so the audit-side evidence in Sections B and F.3 supports the operator-level reading of Definition A.12 in deployment, beyond the controlled construction.

Regimes R0–R3. We work in the coupled case $\pi_{\text{ref}} = \mu_{\text{diag}} = \mathcal{N}(0, \Sigma)$ for the same reason as the gap analysis above: under decoupled isotropic π_{ref} the post-mitigation policy is $\mathcal{N}((0, b, w)/\beta, I_3)$, which gives $\Delta_2 = 0$ and $\Delta J = 0$ identically and collapses the regime taxonomy to a single regime. The coupled case is the minimal setting in which the five regimes are jointly distinguishable in this construction. The limitation of relying on $\pi_{\text{ref}} = \mu_{\text{diag}}$ noted under R4 below applies symmetrically here. Under this coupling, the policy-mean for any linear reward $\bar{R} = c^\top y$ is $\Sigma c/\beta$. The pre-mitigation policy is $\pi = \pi^\star_\beta(\tilde{R}) = \mathcal{N}(\Sigma(a, b, w)^\top/\beta, \Sigma)$ and the post-mitigation policy is $\pi' = \pi^\star_\beta(M_1(\tilde{R})) = \mathcal{N}(\Sigma(0, b, w)^\top/\beta, \Sigma)$. The off-target spurious-feature drift and true-reward change evaluate to

$$\Delta_2(\pi, \pi') = -\rho a/\beta, \qquad \Delta J = -\Sigma_{13} \cdot aw/\beta = 0,$$

the latter because $\Sigma_{13} = 0$ in our parameterization. This makes $\Delta J = 0$ in the strict $\Sigma_{13} = 0$ case and rotation onto ϕ_2 governed entirely by ρa. To populate all five regimes we relax Σ_{13} as a free parameter $\sigma_{13} \in (-1, 1)$ (with Σ remaining positive definite), giving $\Delta J = -\sigma_{13}\, aw/\beta$. The five regimes correspond to:

- *R0 (successful mitigation):* $\rho = 0$, $\sigma_{13} < 0$, $a, w > 0$. Then $\Delta_2 = 0$ and $\Delta J = -\sigma_{13} aw/\beta > 0$. Removing spurious ϕ_1-reliance improves true reward because ϕ_1 was anti-correlated with ϕ_3 at μ_{diag}, so the unmitigated proxy was pushing the policy away from high-ϕ_3 regions.

- *R0$_{cont}$ (contaminated success):* $\rho \neq 0$, $\sigma_{13} < 0$, $a, w > 0$. Then $\Delta_2 = -\rho a/\beta \neq 0$ and $\Delta J = -\sigma_{13} aw/\beta > 0$. True reward improves as in R0 (driven by $\sigma_{13} < 0$), but the cross-correlation ρ couples ϕ_1-reduction to first-moment drift on ϕ_2, so the substitution mechanism is active despite improvement.
- *R1 (bias substitution):* $\rho \neq 0$, $\sigma_{13} = 0$, $a \neq 0$. Then $\Delta_2 = -\rho a/\beta \neq 0$ and $\Delta J = 0$ (neutral substitution). Setting $\sigma_{13} > 0$ instead gives $\Delta J < 0$ (harmful substitution).
- *R2 (overcorrection):* $\rho = 0$, $\sigma_{13} > 0$, $a, w > 0$. Then $\Delta_2 = 0$ and $\Delta J = -\sigma_{13} aw/\beta < 0$. The mechanism is audit-side cross-correlation between ϕ_1 and the structural ϕ_3: removing $a\phi_1$ shifts the policy mean of ϕ_3 by $-\sigma_{13} a/\beta$, lowering true reward without rotating pressure onto ϕ_2. The rescalability test of Section E.12 gives $\Delta J(c) = -c\,\sigma_{13} aw/\beta$ for partial mitigation $M_1^c(\tilde{R}) = \tilde{R} - c\,a\phi_1$, monotone in $c \in (0, 1)$, so no partial mitigation recovers the unmitigated return; by the operational test of Definition A.10 the construction sits on the non-rescalable side of R2 (driven here by reference-policy coupling $\Sigma_{13} \neq 0$ rather than the named target-misspecification mechanism, which would require $\phi_1 \in \Phi_{\text{struct}}$).
- *R3 (silent non-op):* $\rho = 0$, $\sigma_{13} = 0$. Then $\Delta_2 = 0$ and $\Delta J = 0$ regardless of a. The mitigation alters $\tilde{R}$ along an axis the policy's true-reward-relevant directions are orthogonal to.

Each regime is realized by an open set of parameter values, not a measure-zero corner.

Audit-distribution sensitivity (R4). The strict operator-level reading of Definition A.12 (varying μ_{diag} with π_{ref} held fixed) is degenerate under linearity: in the linear-Gaussian setting $g(\tilde{R}; \mu_{\text{diag}}) = (a, b, w)$ at every non-degenerate μ_{diag}, so $M_1^{\mu_{\text{diag}}^{(1)}}(\tilde{R}) = M_1^{\mu_{\text{diag}}^{(2)}}(\tilde{R}) = \tilde{R} - a\phi_1$ as operators, whatever the difference between $\mu_{\text{diag}}^{(1)}$ and $\mu_{\text{diag}}^{(2)}$. Linearity thus characterizes the boundary at which Definition A.12 activates: any operator-level instantiation requires the non-linear extension of the preceding remark (instantiated closed-form in Section E.9), and Section F.3 supplies the empirical counterpart across eight model families.

The closed form does realize R4 jointly with π_{ref} under the coupling $\pi_{\text{ref}} = \mu_{\text{diag}}$ (the audit distribution serving as the rollout reference). Hold $(\tilde{R}, R, \Phi_{\text{sp}}, \beta)$ and the operator M_1 fixed with $a, b, w > 0$, and take two pairs $(\mu_{\text{diag}}^{(\ell)}, \pi_{\text{ref}}^{(\ell)}) = \big(\mathcal{N}(0, \Sigma^{(\ell)}), \mathcal{N}(0, \Sigma^{(\ell)})\big)$ for $\ell = 1, 2$ with $\Sigma_{12}^{(1)} = \rho^{(1)} \neq 0$, $\sigma_{13}^{(1)} = 0$, and $\Sigma_{12}^{(2)} = \rho^{(2)} \neq 0$, $\sigma_{13}^{(2)} < 0$ (each $\Sigma^{(\ell)}$ positive definite, all other entries equal across the two). Then

$$\Delta_2^{(\ell)} = -\rho^{(\ell)} a/\beta, \qquad \Delta J^{(\ell)} = -\sigma_{13}^{(\ell)} aw/\beta,$$

so $\ell = 1$ realizes R1 (neutral substitution: $\Delta_2 \neq 0$, $\Delta J = 0$) and $\ell = 2$ realizes R0$_{\text{cont}}$ ($\Delta_2 \neq 0$, $\Delta J > 0$): the audit distribution selects the regime, with π_{ref} entering jointly via the coupling. For the strict operator-level version (π_{ref} fixed across audit distributions), Section E.9 provides the closed-form instantiation and Section F.3 the empirical counterpart (sign reversal of the length-sycophancy coupling under human-LLM judge disagreement).

Norm increase under M_i. The scale identity from Section E.5 reads $\|\tilde{R}'\|^2_{L^2(\mu_{\text{diag}})} = \|\tilde{R}\|^2_{L^2(\mu_{\text{diag}})} - g_1[g_1 G_{11} + 2(G_{12} g_2 + G_{13} g_3)]$. With $G = \Sigma$ in our parameterization, $G_{11} = 1$, $G_{12} = \rho$, $G_{13} = \sigma_{13}$, and $(g_1, g_2, g_3) = (a, b, w)$, the bracket equals $a + 2\rho b + 2\sigma_{13} w$. Choosing $a = 0.5$, $b = w = 1$, $\rho = -0.6$, $\sigma_{13} = -0.4$ gives bracket $= 0.5 - 1.2 - 0.8 = -1.5$ with bracket sign opposite to $g_1 = 0.5$. The norm change is $-g_1 \cdot (-1.5) = +0.75$, so $\|\tilde{R}'\|^2 > \|\tilde{R}\|^2$ strictly. Mitigation increases the proxy's L^2 norm because the cross-correlation contribution dominates the diagonal term, exactly the failure mode flagged in Section E.5.

E.8 Epsilon-Banded Regime Classification

Finite-sample classification of mitigations into the regimes of Definitions A.8-A.11 requires bands, since exact equality on Δ_j and ΔJ is a measure-zero event under any non-degenerate sampling design.

Banded definitions. Fix tolerances $\varepsilon_j \geq 0$ for each $\phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}$ and $\varepsilon_J \geq 0$. With π, π' as in Definitions A.8-A.11, define

$$\begin{aligned}
\text{R0}_\varepsilon: &\quad \Delta J > \varepsilon_J \text{ and } |\Delta_j(\pi, \pi')| \leq \varepsilon_j \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\},\\
\text{R0}_{\varepsilon,\text{cont}}: &\quad \Delta J > \varepsilon_J \text{ and } |\Delta_j(\pi, \pi')| > \varepsilon_j \text{ for some } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\},\\
\text{R1}_\varepsilon: &\quad |\Delta_j(\pi, \pi')| > \varepsilon_j \text{ for some } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \quad \text{and} \quad \Delta J \leq \varepsilon_J,\\
\text{R2}_\varepsilon: &\quad |\Delta_j(\pi, \pi')| \leq \varepsilon_j \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \quad \text{and} \quad \Delta J < -\varepsilon_J,\\
\text{R3}_\varepsilon: &\quad |\Delta_j(\pi, \pi')| \leq \varepsilon_j \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \quad \text{and} \quad |\Delta J| \leq \varepsilon_J.
\end{aligned}$$

R1's sub-cases port unchanged: neutral and harmful substitution correspond to $|\Delta J| \leq \varepsilon_J$ and $\Delta J < -\varepsilon_J$ respectively. The spurious-rotation-with-improvement corner of Section E.3 is regime $\text{R0}_{\varepsilon,\text{cont}}$ per the definition above. Definition A.12 ports verbatim, since it is transversal to the regime classification and unaffected by the choice of bands.

Nested limit. Taking $\varepsilon_j, \varepsilon_J \to 0$ recovers Definitions A.8-A.11 pointwise: $\text{R0}_\varepsilon \to \{\text{no rotation}, \Delta J > 0\}$, $\text{R0}_{\varepsilon,\text{cont}} \to \{\text{rotation}, \Delta J > 0\}$, $\text{R1}_\varepsilon \to \{\text{rotation}, \Delta J \leq 0\}$, $\text{R2}_\varepsilon \to \{\text{no rotation}, \Delta J < 0\}$, $\text{R3}_\varepsilon \to \{\text{no rotation}, \Delta J = 0\}$, with the disjunctive condition $|\Delta_j| > \varepsilon_j$ for some j converging to $\Delta_j \neq 0$ for some j as $\varepsilon_j \to 0$.

Default choice: noise-floor bands. We default to noise-floor bands, in which $\varepsilon_j = z_{\alpha/2} \cdot \widehat{\text{SE}}(\widehat{\Delta}_j)$ under the empirical sampling design (e.g., the standard error of a fixed-effect coefficient in a mixed linear model, as in Section F.3, where the sycophancy-on-length coefficient instantiates $\widehat{\Delta}_{\text{length}}$) and $\varepsilon_J = z_{\alpha/2} \cdot \widehat{\text{SE}}(\widehat{\Delta J})$ for a significance level α chosen and reported by the practitioner. Setting $\varepsilon_j = \widehat{\text{SE}}$ without the $z_{\alpha/2}$ factor corresponds to approximately 68% confidence and should not be conflated with a standard significance test; regime assignments should always state the chosen α. An effect-size-relative alternative (ε_j as a fraction of unmitigated $\mathbb{E}_{\mu_\pi}[\phi_j]$, ε_J as a fraction of unmitigated J) is appropriate for cross-proxy comparisons; the two can disagree on borderline cases and the choice should be stated. When $|\Phi_{\text{sp}} \setminus \{\phi_i\}| > 1$, the conservative reading (declare rotation if any axis exceeds its ε_j band, no multiple-testing correction) inflates the false-positive rate of the rotation criterion. Because rotation is the discriminating predicate for both $\text{R0}_{\varepsilon,\text{cont}}$ (against R0_ε) and R1_ε (against $\text{R3}_\varepsilon \cup \text{R2}_\varepsilon$), this inflation pulls cases out of R0_ε into $\text{R0}_{\varepsilon,\text{cont}}$ when $\Delta J > \varepsilon_J$, and out of R3_ε and R2_ε into R1_ε when $\Delta J \leq \varepsilon_J$. This choice is deliberate to prioritize sensitivity to substitution over specificity in either ΔJ regime, consistent with the conservative-partition stance of Section 3.1 and the neutral sub-case absorbing near-zero improvements ($\Delta J \in (0, \varepsilon_J]$) in the same direction.

Mapping wild classifications onto the bands. The R0–R4 calls in Sections A.3 and B are made under the banded reading, with $\varepsilon_j, \varepsilon_J$ inferred from each cited work's reported uncertainty. The [3] accuracy degradation under uniform length penalties satisfies the R2_ε pattern under noise-floor bands, with the rescalability test of Section E.12 itself constituting a banded operationalization (existence of $c \in (0, 1)$ improving $\widehat{J}$ implicitly assumes detection against ε_J). The Fein et al. [21] sign flip together with reduced correctness is read as R1_ε under the rotation interpretation (off-target spurious axis exceeds its band, $\widehat{\Delta J} \leq \varepsilon_J$). The alternative reading as R2_ε on the targeted axis itself is also consistent with the

reported statistics, and we flag this ambiguity rather than resolve it. The human-vs-LLM-judge gap is operationalized via Section F.3 below.

Mapping Section F.3 statistics onto the bands. The mixed-model coefficients reported in Section F.3 (+24.3 under human labels, CI [9.6, 39.0], +128.4 under LLM labels, CI [118.5, 138.4], +154.3 under judge agreement, CI [123.6, 185.0], −43.1 under judge disagreement, CI [−60.5, −25.7]) all satisfy the noise-floor ε_j-band test on the length axis, since each CI excludes zero. The pooled KS statistics on within-model residuals (0.025, 0.130, 0.130, 0.046, all with $p < 0.05$) provide a distributional band test alongside the first-moment Δ_j of Definition A.7. The sign reversal of the length-sycophancy coupling between the agreement-side regimes and the disagreement regime is consistent with the construction-level precondition for R4$_\varepsilon$, in that two annotator-conditioned audit distributions yield g-profiles with opposite signs on the relevant cross-feature coupling. The operator-level g-difference is not measured empirically here, but its constructive counterpart appears in Section E.9.

E.9 Phase-Diagram Validation of Regime Transitions Under Quadratic Non-Linearity

Extending the linear Gaussian non-vacuity instantiation of Section E.7, we use a quadratic non-linear reward structure to empirically show

- clear instantiation of all regimes and sub-regimes (R0, R0$_{\text{cont}}$, R1$_{\text{neut}}$, R1$_{\text{harm}}$, R2, R3) with phase-diagrams illustrating how the regime boundaries are outcomes of a single mitigation operator in a controlled setting
- that the regimes are not a property of the reward model alone but of $(R, \tilde{R}, M_i, \mu_{\text{diag}})$ as stated in Section A.
- clear instantiation for the regime transition as described by R4 from only varying the audit distribution μ_{audit} mean
- a match of the theoretical predictions connecting the analytic phase-diagram claims to finite-N experiments, strengthening the non-vacuity results.

This generalizes the linear-Gaussian setting, as we show that the regime taxonomy is not a linear-Gaussian artifact. For real deployment instantiations, see different experiments in Section F.

Setup. We extend the setup in Section E.7: We work with a trivial prompt space ($\mathcal{X}$ a singleton, suppressed in notation) and continuous response space $\mathcal{Y} = \mathbb{R}^3$. The feature map is $\Phi(y) = (\phi_1, \phi_2, \phi_3)$ with $\phi_k(y) = y_k$. We fix true and the preference-generating annotator reward

$$R(y) = w\,\phi_3(y), \qquad R_{\text{anno}}(y) = a\,\phi_1(y) + b\,\phi_2(y) + w\,\phi_3(y) + \gamma\,\phi_1(y)^2,$$

with $w, \gamma > 0$ and $(a, b) \in \mathbb{R}^2$. By Definition A.3, $\Phi_{\text{sp}} = \{\phi_1, \phi_2\}$ and $\Phi_{\text{struct}} = \{\phi_3\}$ by construction. For the reward model (learned proxy reward), we use a linear polynomial such that

$$\tilde{R}(y) = \theta_1\phi_1(y) + \theta_2\phi_2(y) + \theta_3\phi_3(y) + \theta_4\phi_1(y)^2,$$

which is correctly specified using enough samples and the maximum likelihood estimator for Bradley-Terry (BT-MLE) as $\hat{\theta} = (a, b, w, \gamma)$.[4] Reference policy and diagnostic measure are Gaussians on $\mathbb{R}^3$,

$$\pi_{\text{ref}} = \mathcal{N}(0, \Sigma_{\text{ref}}), \qquad \mu_{\text{diag}} = \mathcal{N}(m, \Sigma_{\text{ref}}),$$

[4]We use the analytical idealization for the correct specification of the reward model for closed-form regime predictions. Under realistic RM misspecification, the same regime phenomena persist with additional noise, as documented empirically across five reward models and a non-linear operator in Section F.2.

where Σ is positive definite with

$$\Sigma_{\text{ref}} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & 0 \\ \rho_{13} & 0 & 1 \end{pmatrix}, \qquad \rho_{12}^2 + \rho_{13}^2 < 1$$

Although this setup ties the covariance of the reference and audit distribution, we show that the mean parameter $m \in \mathbb{R}^3$ of the audit distribution μ_{diag} is sufficient to instantiate R4. We fix the KL parameter $\beta > 0$ and target the spurious feature $i = 1$ for mitigation throughout.

Mitigation observables. Following Definition A.4, the linear reliance g_i for targeted $i = 1$ is

$$g_1(\theta, m) = \theta_1 + \theta_2\rho_{12} + \theta_3\rho_{13} + 2\theta_4 m_1.$$

The resulting single-axis mitigation with strength $c \geq 0$ is then

$$M_1(\tilde{R}(y)) = \tilde{R}(y) - cg_1\phi_1(y).$$

We rewrite the reward model as

$$\tilde{R}(y) = \alpha \cdot y + \frac{1}{2} y^\top H y, \qquad \text{with } \alpha = (\theta_1, \theta_2, \theta_3), H = \text{diag}(2\theta_4, 0, 0),$$

reparameterizing the KL-regularized optimum policy as

$$\pi^\star = \mathcal{N}(\mu^\star, \Sigma^\star) \qquad \text{with } \Sigma^{\star -1} = \Sigma_{\text{ref}}^{-1} - \frac{H}{\beta}, \mu^\star = \frac{\Sigma^\star \alpha}{\beta}.$$

Since H is unchanged by the mitigation, $\Sigma^\star$ is also unchanged and only α shifts by $\Delta\alpha = (-cg_1, 0, 0)$, leading to the regime-defining changes in the policy-induced feature expectation and true reward (see Section A.3) as

$$\Delta_j = (\Sigma^\star \Delta\alpha)/\beta = -\frac{cg_1}{\beta}\Sigma^\star_{j,1}, \qquad \Delta J = w\Delta_3 = -w\frac{cg_1}{\beta}\Sigma^\star_{3,1}.$$

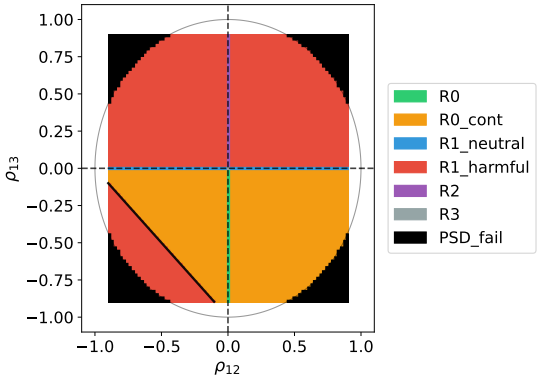
Expanding with Sherman-Morrison gives $\Sigma^\star_{j,1} = \rho_{1j(1+\kappa)}$ with $\kappa = (2\gamma/\beta)/(1 - 2\gamma/\beta) \geq 0$, implying that rescaling γ does not flip the sign of Δ_j for regime classification.

Simulation parameters. With this setup, we have the following parameters to tune and induce regime transitions:

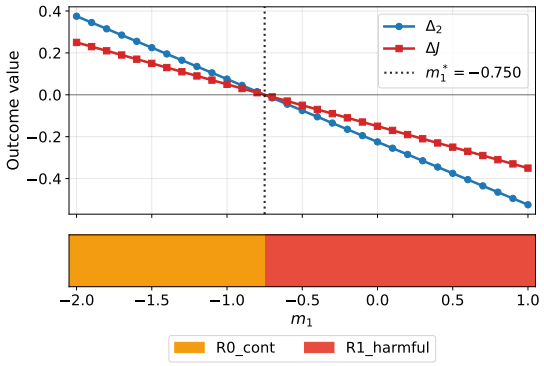
- ρ_{12} sets rotation for bias substitution and also the sign of Δ_2
- ρ_{13} sets the sign of ΔJ (if the sign of g_1 does not change)
- γ amplifies Δ_j and ΔJ, but does not move boundaries at $m = 0$
- m_1 shifts linear reliance g_1 and leads to g_1 sign flip at $m_1^\star = -(a + b\rho_{12} + w\rho_{13})/(2\gamma)$ inducing the transition for R4

The KL parameter β and the mitigation scale parameter c only set margins and scales.

Experiment Results. Using $N = 10^5$ preference samples, 50 seeds per validation cell, a linearly spaced grid of 81 by 81 values for $\rho_{12}, \rho_{13} \in (-0.9, 0.9)^2$ each, and a (near-numerical precision) ε-bound of 10^{-8} (see Section E.8), we instantiate all five regimes R0–R3 and the transition R4, while also validating the above theoretical predictions with finite-N sample experiments. Figure 2 shows the instantiation, Figure 3 shows the regime boundary robustness to non-linearity strength, Figure 4 verifies predictions and measurements, and Table 2 presents the numerical values and uncertainties of Figure 4.



(a) Phase diagram from varying (ρ_{12}, ρ_{13}) at fixed $\gamma = 0.4, m = 0, c = 1, \beta = 4$. We color each grid cell with the corresponding regime from Section A.3.



(b) R4 transition of Δ_2 and ΔJ induced by varying the audit distribution μ_{diag} mean $m_1 \in [-2, 1]$ at fixed $\gamma = 1.0, \rho_{12} = 0.3, \rho_{13} = 0.2, c = 1, \beta = 4$.

Fig. 2. We instantiate all five R0–R3 as tune-able outcomes of a single mitigation operator and R4 by varying only the mean of the audit distribution μ_{diag} in a controlled setting. Figure 2a shows that a non-linearity term creates new boundary structure and that the regimes are not solely induced as a reward model property.

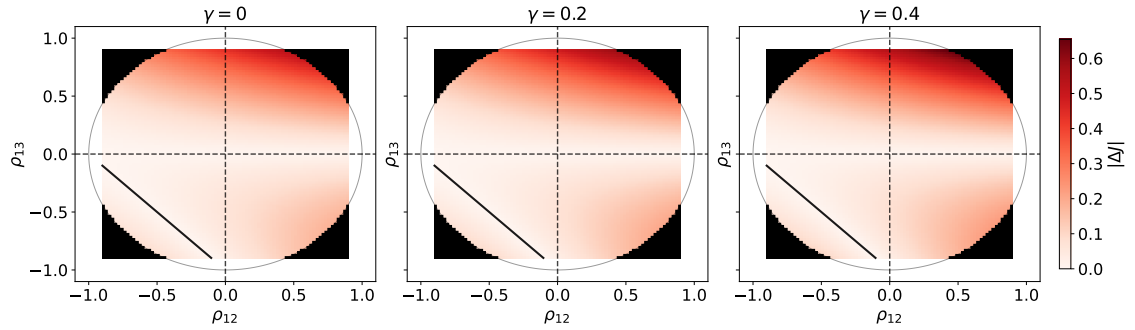


Fig. 3. Shifting γ to produce different $|\Delta J|$-heatmaps of Figure 2a to illustrate regime robustness to non-linearity strength, as the regime boundaries do not change themselves.

E.10 Formal Proof of Audit-Distribution Insufficiency

We give a formal proof of the structural-blindness claim of Section A.4 that standard ordinal reward-model benchmarks cannot reliably distinguish R0 from R0_{cont}, R1, or R2.

Benchmark class. Let $\mathcal{B}$ denote the class of benchmark functionals depending only on the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag}. We take μ_{diag} as the empirical measure of the evaluation set of $\mathcal{B}$ where applicable. This class subsumes ranking accuracy, pairwise win-rate, preference-prediction calibration, reward-target correlation, the linear-reliance statistic g of Definition A.4, and any composite of these, covering the evaluation paradigm of Section A.5. Note that including R in the input of $\mathcal{B}$ strengthens the result, since standard benchmarks lack oracle access to R.

Theorem A.13 (Audit-distribution insufficiency). *Fix $\beta > 0$. There exist four choices of π_{ref} with $\mu_{\text{diag}}, \tilde{R}, M_i, R, \Phi, \Phi_{\text{sp}}$, and β held fixed s.t.*

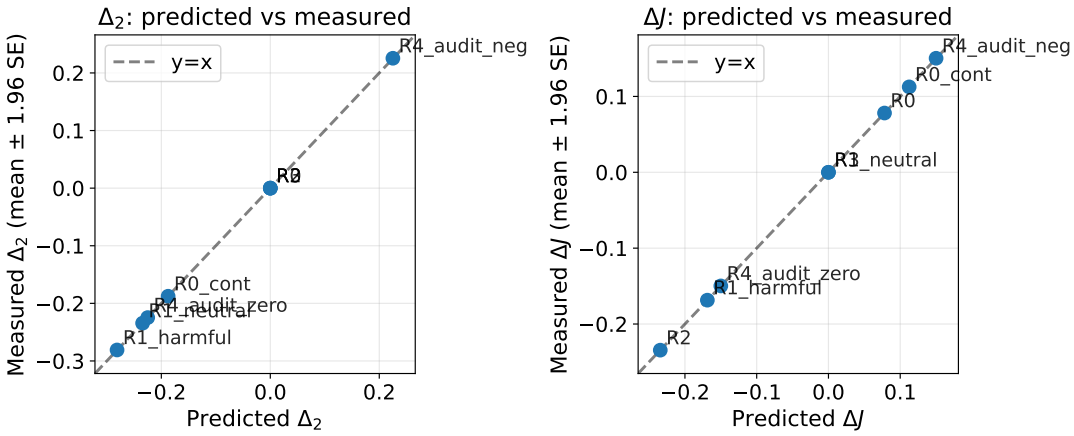


Fig. 4. Validating that the simulated regime transitions align with the theoretical predictions from our setup for Δ_2 and ΔJ with standard errors (SE) from the BT-MLE with $N = 10^5$ samples, aggregated over 50 seeds. This result connects the analytic phase-diagram claims to finite-N experiments.

Table 2. Validation of closed-form regime predictions against finite-N BT-MLE on simulated preferences. All cells use $a = b = w = c = 1$, β=4, N=10^5 samples per seed, for S=50 seeds. The first six columns are the regime instances for $m_1 = 0$, except for R4 where we vary m_1 at fixed $\rho_{12}, \rho_{13}, \gamma$ (see Figure 2b). All eight cells achieve 100% agreement on the respective regime call.

Cell	**R0**	**R0$_{\text{cont}}$**	**R1$_{\text{neut}}$**	**R1$_{\text{harm}}$**	**R2**	**R3**	**R4**	
Scenario								
ρ_{12}	0.0	0.5	0.5	0.5	0.0	0.0	0.3	0.3
ρ_{13}	−0.5	−0.3	0.0	0.3	0.5	0.0	0.2	0.2
γ	0.4	0.4	0.4	0.4	0.4	0.4	1.0	1.0
m_1	0.0	0.0	0.0	0.0	0.0	0.0	−1.5	0.0
Induced g_1	0.500	1.200	1.500	1.800	1.500	1.000	−1.500	1.500
Δ_2								
predicted	0.0000	−0.1875	−0.2344	−0.2812	0.0000	0.0000	0.2250	−0.2250
measured	0.0000	−0.1876	−0.2344	−0.2809	0.0000	0.0000	0.2255	−0.2247
SE	0.0000	0.0002	0.0003	0.0004	0.0000	0.0000	0.0006	0.0005
Δ_J								
predicted	0.0781	0.1125	0.0000	−0.1688	−0.2344	0.0000	0.1500	−0.1500
measured	0.0781	0.1126	0.0000	−0.1686	−0.2345	0.0000	0.1503	−0.1498
measured SE	0.0002	0.0001	0.0000	0.0002	0.0003	0.0000	0.0004	0.0003
Regime classification	R0	R0$_{\text{cont}}$	R1$_{\text{neut}}$	R1$_{\text{harm}}$	R2	R3	R0$_{\text{cont}}$	R1$_{\text{harm}}$

(i) *the joint distribution of* $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ *under* μ_{diag} *is identical across the four instances, so B takes the same value on all four for every* $B \in \mathcal{B}$*;*

(ii) *the optimized policies* $\pi^{\star}_{\beta}(M_i(\tilde{R}))$ *fall into regimes R0, R0$_{cont}$, R1, and R2 respectively, in the sense of Definitions A.8–A.10.*

Proof. **Construction.** Take $\mathcal{X}$ a singleton (suppressed), $\mathcal{Y} = \mathbb{R}^3$, feature map $\phi_k(y) = y_k$, and rewards $R(y) = \phi_3(y)$, $\tilde{R}(y) = \phi_1(y) + \phi_2(y) + \phi_3(y)$. Set $\Phi_{\text{sp}} = \{\phi_1, \phi_2\}$, target $\phi_i = \phi_1$, and audit distribution $\mu_{\text{diag}} = \mathcal{N}(0, I_3)$. Fix a small

$\delta \in (0, 1/2)$. The four instances differ only in $\pi_{\text{ref}}^{(k)} = \mathcal{N}(0, \Sigma^{(k)})$, where each $\Sigma^{(k)}$ has unit diagonal, $\Sigma_{23}^{(k)} = 0$, and

$$\big(\Sigma_{12}^{(0)}, \Sigma_{13}^{(0)}\big) = \big(0, -\tfrac{1}{2}\big), \quad \big(\Sigma_{12}^{(1)}, \Sigma_{13}^{(1)}\big) = \big(\tfrac{1}{2}, -\tfrac{1}{2}\big),$$
$$\big(\Sigma_{12}^{(2)}, \Sigma_{13}^{(2)}\big) = \big(\tfrac{1}{2}, \delta\big), \quad \big(\Sigma_{12}^{(3)}, \Sigma_{13}^{(3)}\big) = \big(0, \tfrac{1}{2}\big).$$

Each $\Sigma^{(k)}$ has $\det = 1 - \Sigma_{12}^2 - \Sigma_{13}^2 > 0$ by Sylvester's criterion: $\det^{(0)} = \det^{(3)} = 3/4$, $\det^{(1)} = 1/2$, and $\det^{(2)} = 3/4 - \delta^2 > 0$ for $\delta < 1/2$. We work on a sufficiently large bounded set to satisfy Assumption E.1. Expressions extend to the unrestricted Gaussian limit as in Section E.7. Linearity and Gaussianity are used for closed-form existence and are not required for the impossibility claim. Assumption A.2 holds since $\mathbb{E}_{\mu_{\text{diag}}}[\Phi\Phi^\top] = I_3$. Assumption E.2 holds since $\mathcal{Y} = \mathbb{R}^3$ admits independent coordinate perturbations on a set of full μ_{diag}-measure.

(i) Audit-side identity. The Gram at μ_{diag} is I_3, giving $g(\tilde{R}; \mu_{\text{diag}}) = (1, 1, 1)$ and $M_i(\tilde{R}) = \phi_2 + \phi_3$ across all four instances. Since $\tilde{R}$, $M_i(\tilde{R})$, R, Φ, and μ_{diag} are identical across k, the joint distribution of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under μ_{diag} is identical across k, so B takes the same value on all four instances for every $B \in \mathcal{B}$.

(ii) Policy-side classification. For any linear $\bar{R}(y) = c^\top y$ and $\pi_{\text{ref}} = \mathcal{N}(0, \Sigma)$, completing the square in Equation (2) gives $\pi_\beta^\star(\bar{R}) = \mathcal{N}(\Sigma c/\beta,\ \Sigma)$. Applying this with $c = (1, 1, 1)^\top$ and $c' = (0, 1, 1)^\top$ at each $\Sigma^{(k)}$:

$$\Delta_2(\pi^{(k)}, \pi'^{(k)}) = -\Sigma_{12}^{(k)}/\beta, \qquad \Delta J^{(k)} = -\Sigma_{13}^{(k)}/\beta,$$

where Δ_j is checked against $\Phi_{\text{sp}} \setminus \{\phi_i\} = \{\phi_2\}$. Substituting:

$$(\Delta_2, \Delta J)^{(0)} = \big(0, +\tfrac{1}{2\beta}\big), \quad (\Delta_2, \Delta J)^{(1)} = \big(-\tfrac{1}{2\beta}, +\tfrac{1}{2\beta}\big),$$
$$(\Delta_2, \Delta J)^{(2)} = \big(-\tfrac{1}{2\beta}, -\tfrac{\delta}{\beta}\big), \quad (\Delta_2, \Delta J)^{(3)} = \big(0, -\tfrac{1}{2\beta}\big).$$

By Definitions A.8-A.10, these are R0, R0_{cont}, R1 (harmful substitution), and R2 respectively. □

Remark (scope). The theorem fixes μ_{diag} and M_i across instances. R3 is additionally realizable in the same construction by setting $\Sigma_{12} = \Sigma_{13} = 0$, with audit observables still identical and distinguishability from R0 requiring $D_{\text{KL}}(\pi' \,\|\, \pi)$, which is not audit-local. R4 requires non-linear proxies (see Sections E.7 and E.9). The varying π_{ref} corresponds to the same reward model deployed against different reference policies, the standard regime in which benchmarks claim to evaluate reward models independently of downstream RLHF setup.

Functional blindspot. Theorem A.13 establishes the distributional blindspot. The functional blindspot is independent and follows from cardinal-scale sensitivity:

Corollary E.4 (Functional blindspot via reward rescaling). *Let $\mathcal{B}_{\text{ord}} \subseteq \mathcal{B}$ denote the benchmark functionals satisfying $B(f \circ \tilde{R}, f \circ M_i(\tilde{R}), R, \Phi, \mu_{\text{diag}}) = B(\tilde{R}, M_i(\tilde{R}), R, \Phi, \mu_{\text{diag}})$ for every strictly monotone increasing $f : \mathbb{R} \to \mathbb{R}$. Then for the construction of Theorem A.13, every $B \in \mathcal{B}_{\text{ord}}$, and every $c > 0$,*

$$B(c\tilde{R}, M_i(c\tilde{R}), R, \Phi, \mu_{\text{diag}}) = B(\tilde{R}, M_i(\tilde{R}), R, \Phi, \mu_{\text{diag}}),$$

while $J(\pi_\beta^\star(c\tilde{R}), R)$ is strictly monotone in c.

Proof. By linearity, $M_i(c\tilde{R}) = c\, M_i(\tilde{R})$, and ordinal invariance under $f(y) = cy$ gives the benchmark identity. The KL identity $\pi_\beta^\star(c\bar{R}) = \pi_{\beta/c}^\star(\bar{R})$ (Section E.5) gives $J(\pi_\beta^\star(c\tilde{R}), R) = c \cdot (\Sigma c_{\tilde{R}})_3/\beta$ in the construction of Theorem A.13, where $c_{\tilde{R}} = (1, 1, 1)^\top$ and $(\Sigma^{(k)} c_{\tilde{R}})_3 = \Sigma_{13}^{(k)} + 1 \neq 0$ for each k, so J is strictly monotone in c. □

E.11 Audit-Distribution Sufficiency

Theorem A.13 establishes that benchmark functionals depending only on $(\tilde{R}, M_i(\tilde{R}), R, \Phi, \mu_{\text{diag}})$ cannot reliably separate R0 from R0_{cont}, R1, or R2. We now show that adding the policy-induced distributions $\mu_{\pi^\star_\beta(\tilde{R})}$ and $\mu_{\pi^\star_\beta(M_i(\tilde{R}))}$ to the input suffices, certifying the prescription that benchmarks must evaluate at policy-induced distributions (Do 1 of Section E.12).

Extended benchmark class. Let $\mathcal{B}^+$ denote the class of benchmark functionals depending on the joint distributions of $(\tilde{R}, M_i(\tilde{R}), R, \Phi)$ under each of μ_{diag}, $\mu_{\pi^\star_\beta(\tilde{R})}$, and $\mu_{\pi^\star_\beta(M_i(\tilde{R}))}$, together with the partition $\Phi = \Phi_{\text{sp}} \sqcup \Phi_{\text{struct}}$. We have $\mathcal{B} \subset \mathcal{B}^+$.

Theorem A.14 (Audit-distribution sufficiency). *Fix $\beta > 0$ and let $\pi = \pi^\star_\beta(\tilde{R})$, $\pi' = \pi^\star_\beta(M_i(\tilde{R}))$. With Δ_j and ΔJ as in Definition A.7 and Section A.3, define $B^\star \in \mathcal{B}^+$ by*

$$B^\star = \begin{cases} \text{R0} & \text{if } \Delta J > 0 \text{ and } \Delta_j = 0 \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \\ \text{R0}_{\text{cont}} & \text{if } \Delta J > 0 \text{ and } \Delta_j \neq 0 \text{ for some } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \\ \text{R1} & \text{if } \Delta J \leq 0 \text{ and } \Delta_j \neq 0 \text{ for some } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}, \\ \text{R2} & \text{if } \Delta J < 0 \text{ and } \Delta_j = 0 \text{ for all } \phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}. \end{cases} \tag{15}$$

Under Assumptions E.1, A.2, and E.2, with $M_i(\tilde{R})$ likewise satisfying Assumption E.1, for every tuple $(\pi_{\text{ref}}, \tilde{R}, M_i, R, \Phi, \Phi_{\text{sp}})$ with $(\pi, \pi') \in \text{R0} \cup \text{R0}_{\text{cont}} \cup \text{R1} \cup \text{R2}$ in the sense of Definitions A.8–A.10, $B^\star$ returns the correct regime label.

Proof. Δ_j and ΔJ are first moments of Φ and R under μ_π and $\mu_{\pi'}$, both inputs to $\mathcal{B}^+$, so $B^\star \in \mathcal{B}^+$. Finiteness of these expectations follows from Assumption E.1 via the boundedness chain in Section E.2, applied to both $\tilde{R}$ and $M_i(\tilde{R})$ (the latter by closure under finite linear combinations).

Correctness on each regime follows directly from Definitions 3.8–3.10: R0 requires $\Delta J > 0$ with no off-target rotation (first branch), R0_{cont} requires $\Delta J > 0$ with off-target rotation (second branch), R1 requires $\Delta J \leq 0$ with off-target rotation (third branch), R2 requires $\Delta J < 0$ with no off-target rotation (fourth branch). The four branches are pairwise disjoint: $\{\text{R0}, \text{R0}_{\text{cont}}\}$ have $\Delta J > 0$ and $\{\text{R1}, \text{R2}\}$ have $\Delta J \leq 0$, separating any cross-pair by the sign of ΔJ; within $\{\text{R0}, \text{R0}_{\text{cont}}\}$ and within $\{\text{R1}, \text{R2}\}$ the branches differ in rotation (R0_{cont}, R1 require rotation; R0, R2 do not). Inputs in R3 (no rotation, $\Delta J = 0$) are excluded by hypothesis. □

Corollary E.5 (Sufficiency for R3). *The classifier $B^\star$ extends to $\{R0, R1, R2, R3\}$ by adding a fourth branch: return R3 if $\Delta J = 0$ and $\Delta_j = 0$ for all $\phi_j \in \Phi_{\text{sp}} \setminus \{\phi_i\}$. This branch matches Definition A.11 exactly and is disjoint from the R0–R2 branches by the $\Delta J = 0$ condition. The observable $D_{\text{KL}}(\pi' \,\|\, \pi)$ is not required for the regime call, though reporting it alongside the label distinguishes policy-relevant R3 ($D_{\text{KL}} > 0$, mitigation moved the policy along directions orthogonal to $(\Delta_j, \Delta J)$) from policy-irrelevant R3 ($D_{\text{KL}} = 0$, mitigation did not move the policy at all).*

Corollary E.6 (Finite-sample sufficiency). *Under the noise-floor ε-bands of Section E.8, replacing exact equalities in Equation (15) with the banded conditions yields $B^\star_\varepsilon \in \mathcal{B}^+$ returning the correct ε-banded label.*

Corollary E.7 (Multi-audit sufficiency for R4). *Let $\mathcal{B}^{++}$ denote the class of benchmark functionals depending on the inputs of $\mathcal{B}^+$ at each of $K \geq 2$ audit distributions $\mu^{(1)}_{diag}, \ldots, \mu^{(K)}_{diag}$, with the canonical $M_i^{(k)}$ constructed from each. Define $B^\star_{R4}$ as the indicator that $B^\star$ (Theorem A.14) returns different regime labels across the K instances. Then $B^\star_{R4}$ detects R4 in the sense of Definition A.12.*

Proof. By Theorem A.14, each per-distribution call of $B^\star$ returns the correct R0–R2 label. Definition A.12 defines R4 as cross-distribution disagreement on the regime label, which is exactly what $B^\star_{\mathrm{R4}}$ tests. □

Discussion. Theorem A.13 (audit-distribution insufficiency) and Theorem A.14 (audit-distribution sufficiency) together show that audit-only inputs cannot separate R0, $\mathrm{R0}_{\mathrm{cont}}$, R1, and R2, while augmenting $\mathcal{B}$'s input with the policy-induced distributions $\mu_{\pi^\star_\beta(\tilde{R})}$ and $\mu_{\pi^\star_\beta(M_i(\tilde{R}))}$ suffices. Corollary E.7 extends this to R4 detection under multi-audit evaluation. This grounds the policy-distribution evaluation prescription structurally rather than heuristically. We do not claim minimality of $\mathcal{B}^+$, as the classifier $B^\star$ depends only on the scalars $(\Delta_j, \Delta J)$ and the partition Φ_{sp}, so any extension of $\mathcal{B}$ that determines these quantities also suffices.

E.12 Takeaway Recommendations for RM Mitigation and Benchmarks Developers

Theorem A.13 and Corollary E.4 together circumscribe what reward-model benchmarks in the class $\mathcal{B}$ can and cannot deliver, and what mitigation methods evaluated within $\mathcal{B}$ cannot be certified to achieve. We translate the formal results into concise recommendations.

Regime detection procedures. Theorem A.14's classifier $B^\star$ separates R0, $\mathrm{R0}_{\mathrm{cont}}$, R1, and R2 given $(\Delta_j, \Delta J)$ at $\mu_{\pi^\star}$, but does not address R2 origin disambiguation, R3 detection, or R4 detection. The following three procedures close these gaps and should accompany any regime claim that touches them.

- *R2 origin disambiguation via the rescalability sweep.* Definition A.10's two origins (scale overshoot, target misspecification) are operationally distinguished by sweeping the partial mitigation $M_i^c(\tilde{R}) = \tilde{R} - c\, g_i(\tilde{R}; \mu_{\mathrm{diag}})\, \phi_i$ over $c \in (0, 1)$ at fixed β. If $J(\pi^\star_\beta(M_i^c(\tilde{R})), R) > J(\pi^\star_\beta(\tilde{R}), R)$ for some c, the regime is driven by scale overshoot at $\mu_{\pi^\star}$ and is recoverable by partial projection. If no c improves ΔJ, the projection has removed a structurally informative component of ϕ_i (target misspecification under partial informativeness, see Section A.1) and rescaling cannot recover it. The test requires a one-dimensional sweep over c at fixed β and the same cardinal R-access already needed to distinguish R2 from R0. Concretely: publish the $\Delta J(c)$ curve, since a recoverable overshoot is a different scientific finding than a non-recoverable target misspecification.
- *R3 detection via D_{KL}.* Corollary E.5 extends $B^\star$ to R3 via the branch $\Delta J = 0 \wedge \Delta_j = 0$. Reporting $D_{\mathrm{KL}}(\pi' \,\|\, \pi)$ alongside the label separates policy-relevant R3 ($D_{\mathrm{KL}} > 0$: mitigation moved the policy along directions orthogonal to $(\Delta_j, \Delta J)$) from policy-irrelevant R3 ($D_{\mathrm{KL}} = 0$: mitigation did not move the policy at all). Concretely: report D_{KL} whenever the regime call is R3, since only the policy-irrelevant case is consistent with a vacuous mitigation operator and should be flagged differently in downstream interpretation.
- *R4 detection via the multi-μ_{diag} protocol.* R4 (Definition A.12) is transversal to R0–R3 and observable only by constructing the canonical $M_i^{\mu_{\mathrm{diag}}}$ at multiple audit distributions and comparing the resulting regime calls. Concretely: select at least two audit distributions of substantively different annotator provenance (e.g., $\mu^{\mathrm{human}}_{\mathrm{diag}}$, $\mu^{\mathrm{LLM}}_{\mathrm{diag}}$), construct $M_i^{(\ell)}$ and $\pi^{(\ell)} = \pi^\star_\beta(M_i^{(\ell)}(\tilde{R}))$ at each, apply $B^\star$ to each, and treat cross-distribution disagreement on the regime call as the test.

Reward Model Bias Mitigation Method Recommendations. The prescriptions of our framework are necessary conditions for interpretable mitigation claims of any new mitigation method. Without them, R0 cannot be distinguished from $\mathrm{R0}_{\mathrm{cont}}$, R1, or R2, regardless of how strong the audit-side evidence is. We organize the recommendations by single- and

multi-axis operators and present each as a self-contained checklist to support standardized reporting across method papers.

Single-axis Don'ts.

- **Do not fit and validate M_i on the same μ_{diag} without testing alternative audit distributions.** R4 (Definition A.12) makes regime membership a function of μ_{diag}, and the sign reversal in Section F.3 shows the dependence is empirically real. *Concretely:* a length operator fit on LLM-judge preference data can inherit a sycophancy-coupling coefficient with the opposite sign of one fit on human-judge data, so the two operators should not be interchangeable even when both zero on-target reliance at their respective audit distributions.
- **Do not apply M_i without reporting the induced $L^2(\mu_{\text{diag}})$ scale change.** Section E.5 shows $\|M_i(\tilde{R})\|_{L^2(\mu_{\text{diag}})}$ generically differs from $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ via both a diagonal and a cross-correlation contribution, inducing an effective-β shift at fixed nominal β. *Concretely:* a method paper reporting only $g_i \to 0$ at μ_{diag} leaves readers unable to separate axis reallocation from scale-driven regime change (see single-axis dos).
- **Do not report only pooled diagnostics when the deployment target is within-prompt selection.** Section F.2 shows pooled $|\rho_{\text{len}}|$ can be driven to zero while within-prompt $|\rho_{\text{len}}^{\text{within}}|$ flips sign on three of four SOTA reward models. *Concretely:* BoN top-1 and PPO both operate within prompts, so a pooled diagnostic targeting ϕ_i does not measure the reliance the optimizer responds to.
- **Do not rely on ordinal-only diagnostics post-mitigation.** Corollary E.4 shows that ranking accuracy and pairwise win-rate are invariant to $\tilde{R} \mapsto c\tilde{R}$ while $J(\pi_\beta^\star(c\tilde{R}), R)$ varies strictly with c. *Concretely:* do not only report ranking accuracy on RewardBench or pairwise win-rates on AlpacaEval post-mitigation, because reporting only such ordinal scores cannot rule out scale-driven regime shifts.

Single-axis Dos.

- **Default to M_i^{norm} and M_i^{cent} mitigations and document the variant used.** Section E.5 shows M_i^{norm} separates axis reallocation from scale and Section E.6 shows M_i^{cent} restores gauge invariance of the regime classification. *Concretely:* when using M_i for pipeline compatibility, report $\|\tilde{R}'\|_{L^2(\mu_{\text{diag}})} / \|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ (ratio of root-mean-square reward scores for mitigated and unmitigated proxy rewards) alongside g_i so that scale-driven and reallocation-driven Δ_j contributions are distinguishable.
- **Run the $c \in (0, 1)$ rescalability sweep when $\Delta J \leq 0$.** The sweep over $M_i^c(\tilde{R}) = \tilde{R} - c\, g_i\, \phi_i$ is the only observational test that separates the two origins of R2 (scale overshoot, recoverable, but target misspecification is not). *Concretely:* publish the $\Delta J(c)$ curve, since a recoverable overshoot is a different scientific finding than a non-recoverable target misspecification.
- **Evaluate at policy-induced distributions and report $(\Delta_j, \Delta J)$.** According to Theorem A.14, this approach separates R0, R0_{cont}, R1, and R2, but, according to Theorem A.13, audit-only inputs cannot. *Concretely:* run the mitigated reward model in BoN or short PPO against a fixed reference policy and report both quantities, not only audit-set accuracy.
- **Instrument off-target Δ_j before publication.** Select a panel of plausible Φ_{sp} candidates the operator does not target (formatting density, hedging markers, position effects, sycophancy and confidence indicators) and report Δ_j on those features at $\mu_{\pi^\star}$. *Concretely:* the developer has both the unmitigated and mitigated rewards in hand and is the natural party to produce this measurement. Without it, an R0 claim is structurally indistinguishable from R0_{cont} (substitution under improvement).

- **Report cardinal scale pre/post mitigation.** Publishing $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ before and after is the minimal cardinal observable that ordinal benchmark functionals miss. *Concretely:* report the standard deviation of reward scores on a fixed evaluation set both before and after applying the mitigation, so scale-driven ΔJ shifts are visible.
- **Document π_{ref}-sensitivity of the mitigation.** The construction in Theorem A.13 produces four distinct regimes by varying only π_{ref} for a fixed $(\tilde{R}, M_i, \mu_{\text{diag}})$. *Concretely:* validate the mitigation against at least two reference policies of different lineages and report per-pair outcomes rather than a single headline number.

Multi-axis recommendations.

- **Specify whether the operator is joint or sequential, and the operator-stack type.** For canonical linear projections applied to a fixed reward, sequential M_i followed by M_j equals joint projection on (ϕ_i, ϕ_j) by Frisch-Waugh-Lovell. The equivalence breaks when operators are heterogeneous (e.g., LOESS for length, two-head for content), when projection directions are recomputed on the changed reward, or when operators are non-linear. *Concretely:* state which case applies and, when sequential and joint differ, report both outcomes side-by-side.
- **Joint reliance zero at μ_{diag} does not certify R0.** Lemma A.6 generalizes and a multi-axis operator M_S that zeros g_i for all $i \in S$ at μ_{diag} does not in general zero g_i at $\mu_{\pi^\star}$. *Concretely:* report Δ_j at $\mu_{\pi^\star}$ for every axis in the targeted set, not just on-target axes. Covering a panel at μ_{diag} narrows the substitution-accessible subset of Φ but does not close the measurement-vs-optimization gap.
- **State Φ_{sp} explicitly and acknowledge out-of-panel exploitation.** *Concretely:* a method covering {length, sycophancy, position} does not certify R0 against features outside that set, and multi-axis claims should be framed as "no substitution within the measured panel" rather than "no substitution."
- **Flag statistical mediation between targeted features.** Where the literature establishes coupling between features in the targeted set (length-sycophancy and length-confidence per Facts G.9,G.11,and G.12), associational joint projection may remove signal flowing through coupled paths. *Concretely:* acknowledge this as a known limitation rather than treating coupled signal as residual noise. The rescalability test above extends to multi-axis as a partial diagnostic.

Scope under non-linear operators. The proofs above (Theorems A.13 and A.14) and the linear-Gaussian instantiation (Section E.7) use Gaussianity for closed-form policy expressions and linearity for the canonical projection construction, but the prescriptions in this section target the structural $\mu_{\text{diag}} \to \mu_{\pi^\star}$ gap and the cardinal/ordinal blindness rather than these constructive devices, and so apply to non-linear, non-Gaussian operators and reference policies. The regime classification (R0-R4) references Δ_j and ΔJ rather than operator form, the cardinal/ordinal blindness of Corollary E.4 is monotone-invariant, and the gap itself is geometric rather than algebraic. What does *not* transfer is Lemma A.6's constructive guarantee (a non-linear operator may not zero g_i even at μ_{diag}) and the closed-form expressions of Section E.7. Nonetheless, Section E.9 demonstrates regime separation under quadratic non-linearity in closed form and Section F.2 shows the gap surviving a non-linear *operator* (LOESS calibration) measured by a linear diagnostic, and Section F.3 establishes audit-distribution dependence of the relevant coupling under linear specifications with distribution-free robustness checks. **Method papers using non-linear operators should still adopt the prescriptions** above by virtue of targeting the same geometric phenomenon, while stating that they cannot import the linear-Gaussian regime classification's constructive guarantees, only its outcome-level definitions.

Reward Model Benchmark Recommendations. Some of the recommendations for mitigaiton method developers transfer over to benchmark developers, with a few added recommendations.

Don'ts.

- **Do not expand the benchmark input within μ_{diag} and expect regime separation.** Theorem A.13 grants $\mathcal{B}$ joint access to $(\tilde{R}, M_i(\tilde{R}), R, \Phi, \mu_{\text{diag}})$ and still admits four instances landing in R0, R0_{cont}, R1, and R2. Adding feature axes, distractors, or held-out audit splits stays inside $\mathcal{B}$ and inherits the impossibility. *Concretely:* adding a sycophancy subset alongside a length subset, or expanding from pairwise to best-of-4 selection, enriches μ_{diag} but does not measure Δ_j at $\mu_{\pi^\star}$ and so cannot separate R0 from R0_{cont} or R1.
- **Do not rely on benchmark functionals that are invariant to monotone rescaling of $\tilde{R}$.** Corollary E.4 shows that every $B \in \mathcal{B}_{\text{ord}}$ (ranking accuracy, pairwise win-rate, and their composites) returns identical values under $\tilde{R} \mapsto c\tilde{R}$ while $J(\pi^\star_\beta(c\tilde{R}), R)$ varies strictly with c, so cardinal-scale-driven regime shifts are undetectable. *Concretely:* if mitigation halves the proxy's effective scale, ranking accuracy is unchanged but the KL-regularized policy behaves as if β doubled (by the scale-equivariance of Section E.5), and downstream ΔJ shifts accordingly.
- **Do not generalize a single-π_{ref} score across deployment settings.** The construction in Theorem A.13 varies only π_{ref} and obtains four distinct regimes for the same $(\tilde{R}, M_i, \mu_{\text{diag}})$. A benchmark score at one π_{ref} does not transport to deployment π_{ref}'s. *Concretely:* a reward model topping the leaderboard when paired with one SFT base can degrade PPO outcomes when paired with a different SFT base of the same model family, since the policy-induced distribution shifts even though the audit-side score does not.

Dos.

- **Evaluate at policy-induced distributions and report** $(\Delta_j, \Delta J)$. By Theorem A.14, this input separates R0, R0_{cont}, R1, and R2 via the explicit functional $B^\star$ in Equation (15); by Theorem A.13, no audit-only input suffices. *Concretely:* run the mitigated reward model in BoN or short PPO against a fixed reference policy, and report the change in expected feature value Δ_j on each $\phi_j \in \Phi_{\text{sp}}$ alongside the change in true reward ΔJ, rather than only audit-set accuracy.
- **Report cardinal scale**. Publishing $\|\tilde{R}\|_{L^2(\mu_{\text{diag}})}$ pre/post mitigation provides observables that $\mathcal{B}_{\text{ord}}$ cannot, addressing the blindspot of Corollary E.4. *Concretely:* alongside accuracy, report the standard deviation of reward scores on a fixed evaluation set before and after applying the mitigation, so that scale-driven ΔJ shifts are visible to readers.
- **Treat π_{ref} as an axis of variation, not a fixed convention.** Either evaluate across a panel of reference policies or document π_{ref}-sensitivity as a first-class output, consistent with the construction underlying Theorem A.13. *Concretely:* pair each reward model with at least two SFT reference policies of different lineages and publish per-pair scores rather than a single headline number.
- **Report Δ_j across all of Φ_{sp} at $\mu_{\pi^\star}$, not just the targeted axis.** Theorem A.14 makes the off-target Δ_j on $\Phi_{\text{sp}} \setminus \{\phi_i\}$ a load-bearing input for separating R0_{cont} and R1 from R0. Reporting only on-target reliance leaves substitution structurally invisible regardless of how rich the audit distribution is. *Concretely:* when evaluating a length-debiasing operator, also measure post-mitigation drift in sycophancy, hedging, and formatting at $\mu_{\pi^\star}$ — not just residual length correlation at the audit set.

Summary. The recommendations above are not addressed to a single audience. R0 certification is a joint property of the mitigation operator and the evaluation protocol: a method paper following the developer recommendations cannot be cleanly compared across benchmarks that omit the corresponding ones, and a benchmark implementing all three

prescriptions cannot certify R0 for methods that do not report off-target Δ_j or cardinal scale. The overlap between the two checklists is structural rather than redundant.

Existing methods and benchmarks satisfy these prescriptions only partially. SOTA benchmarks partially implement these prescriptions and close the gap to pure ordinal benchmarks. In particular, the current SOTA benchmark RewardBench2 [52] features a Ties metric, introduces cardinal-sensitive scoring, and its on-policy/off-policy finding documents π_{ref}-sensitivity empirically. However, no current protocol jointly delivers policy-distribution evaluation, cardinal reporting, and π_{ref}-as-axis, a gap our framework both formalizes as necessary (Theorem A.13) and proves sufficient to close (Theorem A.14). On the mitigation side, published methods rarely report off-target Δ_j at $\mu_{\pi^\star}$ or run the rescalability sweep, leaving almost all recent results ambiguous between R0_{cont}, R1, and R2 in ways the recommendations would have resolved (see Section C).

Closing the joint gap is a community coordination problem rather than a single-paper deliverable. Method papers cannot unilaterally provide policy-distribution evaluation without benchmark infrastructure that supports it. Benchmarks cannot unilaterally provide off-target Δ_j measurement without method papers specifying Φ_{sp}. The recommendations above are scoped so that adoption by either side improves the evidence base, and joint adoption is what moves the field from R0 claims that audit-side scores cannot certify to R0 claims that the framework certifies sufficient.

F Supporting Experiments

All experiments were either run through API access or on a single A100 GPU 40Gb or B200 GPU 180 Gb (rented online). Our code for the experiments will be released on GitHub (MIT license) upon publication.

F.1 Bias Substitution in Language Model RLHF (GRPO)

We demonstrate reward bias substitution in a RLHF pipeline with standard off-the-shelf configuration. A single-axis length penalty applied during GRPO training compresses response length as intended, yet the freed optimization pressure rotates onto a correlated axis and drives the policy into overconfidence, instantiating the substitution regime (R1) while leaving multiple-choice quality intact.

Setup.

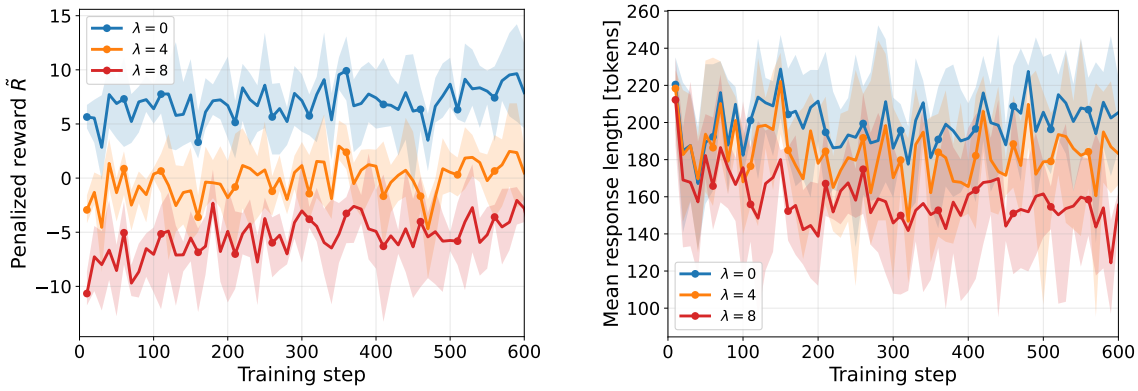
- **Policy model.** *Llama-3.2-3B-Instruct* [27], adapted with LoRA on all linear layers (the four attention projections `q_proj`, `k_proj`, `v_proj`, `o_proj` and the three MLP projections `gate_proj`, `up_proj`, `down_proj`) with rank $r = 32$, scaling $\alpha = 64$, dropout 0.05.
- **Reward model.** *Skywork-Reward-V2-Llama-3.1-8B* [49], kept frozen throughout training and queried only through the reward function.
- **Dataset.** Prompts from *UltraFeedback* [16] (the `ultrafeedback_binarized` `train_prefs` split), truncated to 2000 characters.
- **Algorithm.** Group Relative Policy Optimization (GRPO) [67] as implemented in TRL [79]. 8-bit AdamW, learning rate 3×10^{-5}, KL coefficient $\beta = 2 \times 10^{-2}$, 600 optimizer steps, and bfloat16 precision.
- **Length penalty.** The RLHF reward is shaped as $\tilde{R}(x, y) = R_{\mathrm{RM}}(x, y) - \lambda\, n_{\mathrm{tok}}(y)/100$, where $n_{\mathrm{tok}}(y)$ is the number of response tokens and λ is the penalty coefficient. Three values $\lambda \in \{0, 4, 8\}$, each trained with four random seeds, for twelve training runs in total.
- **Generation during training.** Group size of 4 candidate generations per prompt for the group-relative advantage, per-device batch of 4, gradient accumulation of 1, sampling temperature $T = 1.0$, maximum completion length 256 tokens.

Training Curves. We log the following quantities at every optimizer step and report them per λ, averaged over the four seeds, see Figure 5.

- **Mean response length.** The average number of generated tokens per step. This is the primary curve and shows the length penalty taking effect, with the curves separating by λ.
- **Mean reward.** The average shaped reward $\tilde{R}$ per step. Note that $\tilde{R}$ is not directly comparable across λ because the penalty term differs, so the curves are read within each λ for evidence of learning.
- **KL divergence.** The divergence of the policy from the frozen reference policy per step, confirming that the policy stays regularized and does not collapse.
- **GRPO loss and gradient norm.** Diagnostic optimization curves.

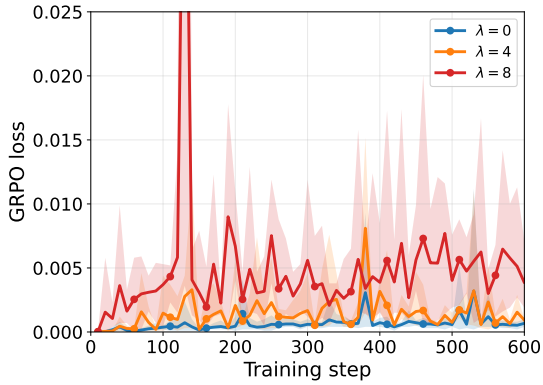
Evaluation. All twelve adapted policies and the untrained base policy[5] are evaluated on held-out data, with each adapter taken at the 600-step checkpoint. Aggregate numbers use 95% confidence intervals from a hierarchical bootstrap over seeds and samples. We report the following metrics.

[5]Untrained by us, but still instruction tuned by the model developer.

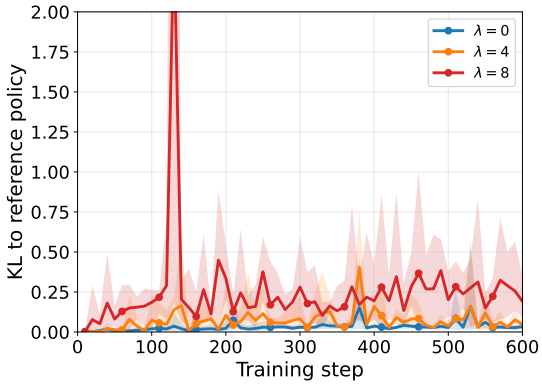


(a) Reward $\tilde{R}(x, y) = R_{\mathrm{RM}}(x, y) - \lambda\, n_{\mathrm{tok}}/100$

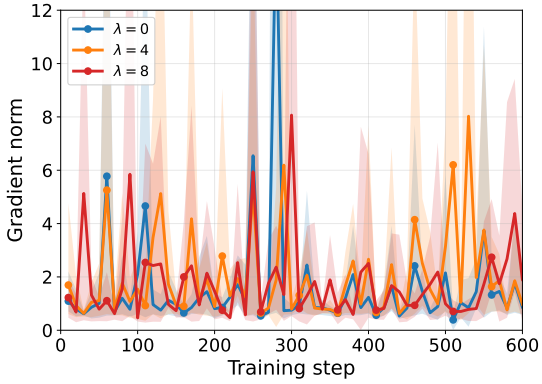
(b) Response length n_{tok}



(c) GRPO loss



(d) KL to reference policy



(e) Gradient norm

Fig. 5. Training curves (shaded region shows the min-max range across 4 seeds)

	Base	$\lambda = 0$	$\lambda = 4$	$\lambda = 8$
Length and quality				
Mean response length (tokens)	192.2	$203.6_{[194, 212]}$	$188.2_{[181, 195]}$	$170.2_{[161, 179]}$
MMLU accuracy	0.596	$0.609_{[.59, .62]}$	$0.611_{[.59, .63]}$	$0.616_{[.60, .63]}$
Overconfidence				
Expected calibration error (ECE)	0.264	$0.253_{[.23, .28]}$	$0.299_{[.26, .34]}$	$0.406_{[.32, .50]}$
Mean verbalized confidence	0.811	$0.803_{[.78, .82]}$	$0.815_{[.80, .83]}$	$0.836_{[.82, .85]}$
TriviaQA accuracy	0.564	$0.558_{[.53, .58]}$	$0.523_{[.47, .57]}$	$0.415_{[.29, .52]}$
AUROC	0.708	$0.725_{[.71, .74]}$	$0.706_{[.67, .74]}$	$0.651_{[.62, .69]}$
Sycophancy				
Regressive flip rate	0.653	$0.582_{[.52, .64]}$	$0.646_{[.59, .70]}$	$0.602_{[.54, .66]}$
Overall flip rate	0.721	$0.668_{[.63, .71]}$	$0.709_{[.67, .75]}$	$0.704_{[.66, .75]}$

Table 3. Metric values across the length-penalty sweep for the data plotted in Figure 1 in Section 1. Each λ entry is the mean over four random seeds with the 95% hierarchical-bootstrap confidence interval as a subscript. The Base column is the untrained *Llama-3.2-3B-Instruct* model and carries no interval. Expected calibration error, TriviaQA accuracy, and AUROC use the LLM-judge answer grading.

- **Mean response length.** The average response token count on 50 held-out *UltraFeedback* `test_prefs` prompts, with 4 samples per prompt drawn at temperature 1.0.
- **MMLU accuracy.** Multiple-choice accuracy on 1000 questions subsampled from MMLU, evaluated zero-shot with greedy decoding and a single-letter answer parsed from the output.
- **Expected Calibration Error (ECE).** Computed with 10 equal-width bins on the (verbalized confidence, correctness) pairs produced by the Tian et al. [77] Verb. 1S top-1 protocol on 1000 *TriviaQA* questions.
- **AUROC.** The area under the ROC curve of the verbalized confidence used as a predictor of answer correctness.
- **Mean verbalized confidence.** The average probability the model assigns to its own answer under the Verb. 1S top-1 protocol.
- **TriviaQA accuracy.** The fraction of correct answers under the same protocol.
- **Regressive flip rate.** The headline sycophancy metric from the Sharma et al. [69] *are_you_sure* task, the fraction of initially-correct answers that become incorrect after the user pushes back.
- **Overall flip rate.** The fraction of answers that change at all after the pushback, regardless of correctness.

The Tian et al. [77] ECE, AUROC, and TriviaQA accuracy metrics are computed under an LLM-judge equivalence check using the Tian et al. [77] prompt with *Claude Haiku 4.5* [1] as the judge. The sycophancy metrics use the Sharma et al. [69] two-stage protocol, eliciting an initial answer and then a post-pushback answer on 500 *are_you_sure* records with 2 samples per record at temperature $T = 1.0$.

Results. See Figure 1 and Table 3 for results.

- **The length penalty compresses responses monotonically.** Mean response length falls from 204 tokens at $\lambda = 0$ to 188 at $\lambda = 4$ and 170 at $\lambda = 8$. Both reductions, measured relative to $\lambda = 0$, are significant.
- **Multiple-choice quality is preserved.** MMLU accuracy holds near 0.61 across the sweep (0.609, 0.611, 0.616 at $\lambda = 0, 4, 8$ against a base of 0.596), and the change from base is not significant at any λ.

- **Overconfidence rises with the length penalty.** Expected calibration error climbs from 0.25 at $\lambda = 0$ to 0.30 at $\lambda = 4$ and 0.41 at $\lambda = 8$, and mean verbalized confidence climbs in step from 0.80 to 0.82 to 0.84. TriviaQA accuracy falls from 0.56 at $\lambda = 0$ to 0.52 at $\lambda = 4$ and 0.41 at $\lambda = 8$, and the coupling between confidence and correctness weakens, with AUROC falling from 0.73 to 0.65. Thus, the policy does not merely become more confident, it becomes wrong more often. At $\lambda = 8$ the rise in calibration error, the rise in verbalized confidence, the drop in accuracy, and the drop in AUROC are all significant.
- **The calibration loss is caused by the penalty, not by RLHF.** At $\lambda = 0$ the calibration error (0.25) is at or slightly below the base value (0.26), so RLHF on its own does not break calibration. The degradation appears only once the length penalty is applied.
- **Sycophancy does not significantly change.** The regressive flip rate is non-monotonic across the sweep (0.58, 0.65, 0.60 against a base of 0.65) and the overall flip rate shows no consistent trend (0.67, 0.71, 0.70 against 0.72). However, the base model already has a high level of sycophancy.
- **Regime transitions.** $\lambda = 0$ is the no-penalty reference, $\lambda = 4$ transitions into and $\lambda = 8$ instantiates the bias substitution regime (R1), in which the penalty compresses length and the optimization pressure rotates onto confidence calibration. The overcorrection regime (R2), in which task-independent capability degrades, is not reached within the swept range, since MMLU accuracy is preserved at every λ. $\lambda = 4$ is not successful mitigation R0, since every overconfidence proxy is already displaced toward substitution (see Table 3) and sits at or leans towards the significance boundary, so the missing significance reflects four-seed power rather than the absence of an effect.
- **Regime transitions.** We take the true reward to be free-form factual accuracy, proxied by TriviaQA, and we treat expressed confidence as a spurious feature that the reward model rewards as a quality proxy (Fact G.12) (Φ_{sp}). Multiple-choice MMLU accuracy serves as a capability control. At $\lambda = 8$, response length falls and MMLU is unchanged, while expressed confidence rises, a nonzero Δ_j on an off-target spurious axis, and TriviaQA accuracy degrades, a negative ΔJ. This is the harmful sub-case of bias substitution (R1), not overcorrection (R2), because pressure rotates onto a second spurious axis rather than degrading capability with no rotation and since MMLU accuracy is preserved at every λ. We read $\lambda = 4$ as an intermediate point, where the overconfidence proxies shift toward substitution but the changes from the $\lambda = 0$ run are not significant at four seeds, and $\lambda = 8$ as the clear R1 instance.

F.2 Measurement-vs-Optimization Gap in Length Mitigation in BoN Selection

This section provides a quantitative empirical instance of the measurement-vs-optimization gap of Section A.2. We evaluate two published length-debiasing operators across five reward models under a Best-of-N (BoN) selection protocol, using a within-prompt reliance diagnostic that tracks the BoN selection distribution rather than the pooled audit distribution at which the operators are designed. The headline observation is that the post-hoc calibration of Huang et al. [30] zeros pooled reward-length correlation ($0.316 \rightarrow 0.037$) while overshooting into negative within-prompt correlation on three of four SOTA reward models, with ΔJ degrading on two. The linear-probe operator of Fein et al. [21] is fit on the same prompt split and transfers cleanly to the within-prompt diagnostic on four of five reward models with $\Delta J > 0$.

Reward models and operators. We evaluate Skywork-Reward-V2-Llama-3.1-8B, Skywork-Reward-V2-Qwen3-8B, Skywork-Reward-V2-Qwen3-0.6B [49], the Llama-3.1-8B-Instruct-RM-RB2 [52], and the older DeBERTa-v3-large [57]

	DeBERTa	Allen	SW-L	SW-Q8B	SW-Q0.6B
Within-prompt $\rho_{\text{len}}^{\text{within}}$ *(AlpacaEval)*					
Baseline	+.168	−.061	+.134	+.065	+.088
Mech. RS	+.005	−.011	+.087	+.032	+.041
Huang et al. [30]	+.048	−.099	−.065	−.138	−.228
AlpacaEval LC win rate vs. baseline (%)					
Mech. RS	54.4^{***}	$46.2^{\dagger}$	51.5^{***}	51.7^{***}	51.8^{***}
Huang et al. [30]	50.4^{*}	$48.7^{\dagger}$	50.1	$48.2^{\dagger}$	51.2^{***}
GSM8K BoN accuracy (%)					
Baseline	32.8	71.5	68.3	71.9	65.8
Mech. RS	36.4	71.3	68.5	71.9	65.7
Huang et al. [30]	32.3	67.9	68.3	71.3	66.2

Table 4. Within-prompt reward-length correlations and BoN selection outcomes. $^{*}p < .05$, $^{***}p < .001$ vs. 50%; † denotes significantly below 50%. Pooled $|\rho_{\text{len}}|$ averaged across the five RMs is 0.316 (baseline) and 0.037 for Huang et al. [30]: audit-distribution success and within-prompt overshoot on the same operator.

reward models. We use two single-axis operators that target length bias (ϕ_i = length). First, mechanistic reward shaping with a difference-of-means linear probe projected from the RM's final-layer hidden state [21], and, second, the post-hoc reward calibration of Huang et al. [30] with a LOESS fit subtracted at the score level. Both are instances of M_i (Definition A.5) and reduce $|g_i(\tilde{R}; \mu_{\text{diag}})|$ at their respective audit distributions by construction.

Setup. For each of two prompt sources (AlpacaEval [18], GSM8K [15]) we generate 64 candidate responses per prompt from Llama-3.2-1B-Instruct [27], fit each operator on a probe split, and evaluate on a disjoint held-out split (512 for AlpacaEval prompts, 807 for GSM8K prompts). Each RM selects the highest-scoring response among the 64 candidates. We measure two quantities:

- a within-prompt Pearson correlation $\rho_{\text{len}}^{\text{within}}$ between RM score and response length, computed within each candidate set and averaged across prompts. Because BoN top-1 selection operates within candidate sets per prompt, $\rho_{\text{len}}^{\text{within}}$ measures the reliance selection actually responds to, while pooled ρ_{len} averages over prompts that BoN never compares. This instantiates g_i at a μ_{diag} aligned with BoN selection rather than the pooled distribution the operators target.
- A selection-level ΔJ proxy (AlpacaEval length-controlled win rate vs. baseline, GSM8K BoN accuracy on selected responses).

The pair $\{\rho_{\text{len}}^{\text{pooled}}, \rho_{\text{len}}^{\text{within}}\}$ operationalizes Lemma A.6 empirically, as an operator can zero one and not the other, and the selection outcome reveals which distribution drives the optimizing process. BoN top-1 over 64 samples is a coarser proxy for $\pi_\beta^\star$ than KL-regularized optimization, and the regime calls below should be read accordingly.

Results. The Huang et al. [30] calibration achieves $|g_i| \approx 0$ at the pooled μ_{diag} it targets (0.316 → 0.037 averaged across the five RMs). At the within-prompt μ_{diag} aligned with BoN selection, three of four SOTA RMs acquire negative $\rho_{\text{len}}^{\text{within}}$ of magnitude 0.065-0.228, exceeding their unmitigated baselines in absolute value (Table 4). The same pattern holds on GSM8K, where within-prompt $|\rho_{\text{len}}|$ averages 0.096 for Huang et al. [30] versus 0.007 for mechanistic reward shaping, with BoN accuracy averaging 61.20% versus 62.76%. The signed flip rather than same-side overshoot is **evidence of**

the measurement-versus-optimization gap of Lemma A.6, because the operator reaches zero at μ_{diag}, while at the optimization-relevant distribution g_i has the wrong sign.

Regime classification. Under the ε-banded reading of Section E.8, two Huang et al. [30] cells (Allen, Skywork-Q8B) satisfy $\Delta J < -\varepsilon_J$ on AlpacaEval LC win rate while the targeted axis $\rho_{\text{len}}^{\text{within}}$ has flipped sign. The GSM8K Allen cell additionally drops 3.6 accuracy points. These satisfy the targeted-axis and ΔJ conditions of Definition A.10. Whether they additionally satisfy the $\Delta_j = 0$ condition for off-target spurious axes is unmeasured here. As a result, the pattern is consistent with R2_ε if the no-rotation condition holds, or with a mixed R1+R2 regime otherwise.

Mechanistic reward shaping achieves $\Delta J > 0$ at $p < .001$ on four of five AlpacaEval cells and moves $|\rho_{\text{len}}^{\text{within}}|$ toward zero on four of five RMs without sign flips. The Allen cell falls below 50% LC WR on both operators (46.2% for Mech. RS, 48.7% for Huang et al. [30], both †). The GSM8K Allen cell drops 0.2 for mechanistic reward shaping and 3.6 for Huang et al. [30]. The remaining four mechanistic reward shaping cells satisfy Definition A.8 modulo the unmeasured no-rotation condition.

Connection to the framework. The pair $(\rho_{\text{len}}^{\text{pooled}}, \rho_{\text{len}}^{\text{within}}) = (0.037, 0.116)$ for Huang et al. [30] averaged across the five RMs is the empirical counterpart of Lemma A.6: the canonical M_i reaches zero at the audit distribution it targets while g_i at the optimization-relevant distribution has the wrong sign. This dissociation is the load-bearing mechanism of the substitution argument. Wherever it holds, R1 is mechanically available to the optimizer, and the audit diagnostic targeting ϕ_i cannot detect it by construction. Standard reward-model benchmarks evaluate at a single fixed μ_{diag} and report only the pooled diagnostic, so under Section A.5 they cannot distinguish this $\Delta J < 0$ instance from R0 regardless of whether the underlying regime is R1_ε (rotation onto an unmeasured spurious axis), R2_ε (overcorrection on the targeted axis), or a mixture: all three are invisible at μ_{diag}, which is the central benchmark-inadequacy claim.

F.3 Evaluating sycophancy and length of AITA Responses

This section reports the experimental analysis backing the Section B claim that response length and sycophancy labels are statistically dependent across human, LLM-judge, and judge-agreement labeling regimes, using AITA prompts and responses from [9] across eight model families. We fit mixed linear models of response length (measured in characters) on a binary sycophancy indicator (with model or prompt as a random effect), and report KS and W_1. These statistics provide audit-side evidence for the construction-level precondition of R4 (Definition A.12), with the formal banded treatment deferred to Section E.8. In terms of causality, the mixed-model coefficients and KS statistics below are observational. They are consistent with a mediation reading of length but do not, on preference data alone, discriminate among candidate causal structures.

Human-labeled regime (N = 14,375 responses across 8 models). A mixed linear model of response length on the sycophancy indicator, with model as random effect, yields a positive sycophantic effect of +24.3 length units (95% CI [9.6, 39.0]; p = 0.001), against an intercept of 1554.5 and between-model variance 72,408. The effect is small relative to between-model variability (median per-model non-syco/syco ratio 0.98), indicating that under human labeling sycophantic responses are only modestly longer on average, but statistically significant. Distributionally, the length distributions of sycophantic and non-sycophantic responses differ significantly for 6 of 8 models by Kolmogorov–Smirnov test (Llama-8B and Mistral-7B not significant at $\alpha = 0.05$), with per-model Wasserstein distances ranging from 24.5 to 138.6. Pooling z-scored within-model residuals gives KS = 0.025 (p = 0.022), confirming a small but significant aggregate distributional deviation.

LLM-judge–labeled regime (N = 28,832 responses across 8 models). A mixed linear model of response length on the sycophancy indicator, with model as random effect, yields a sycophantic effect of +128.4 length units (95% CI [118.5, 138.4]; $p < 0.001$), against an intercept of 1487.1 and between-model variance 70,521. The effect is substantially larger than under human labels (+24.3), indicating that LLM judges associate sycophancy with longer responses than the binary-verdict label does. Per-model non-syco/syco ratios shift correspondingly downward (median 0.94 vs. 0.98 under human labels), with the gap most pronounced for Llama-17B, Llama-70B, and Gemini (ratios 0.81–0.85). Length distributions differ significantly for 7 of 8 models by Kolmogorov–Smirnov test (only gpt-4o is not significant at $\alpha = 0.05$), with per-model Wasserstein distances ranging from 19.5 to 293.7, which is substantially larger than the 24.5–138.6 range under human labels. The pooled KS on z-scored within-model residuals is 0.130 ($p < 0.001$), roughly 5× larger than the human-label pooled KS of 0.025, confirming that the LLM-judge regime exhibits a much stronger distributional length–sycophancy coupling than the human regime at both the central-tendency and full-distribution levels.

Judge-agreement regime (N = 4,149 responses across 8 models, restricted to responses where human and LLM judge labels coincide). A mixed linear model of response length on the sycophancy indicator, with model as random effect, yields a sycophantic effect of +154.3 length units (95% CI [123.6, 185.0]; $p < 0.001$), against an intercept of 1507.2 and between-model variance 68,577. The effect is the largest of the four regimes, exceeding both human-only (+24.3) and LLM-only (+128.4), which is consistent with the agreement subset isolating responses for which sycophancy is most unambiguous along a length-correlated axis. Per-model non-syco/syco ratios drop correspondingly (median 0.93), with Llama-70B and Gemini showing the strongest shifts (ratios 0.77 and 0.77). Length distributions differ significantly for 4 of 8 models by Kolmogorov–Smirnov test at $\alpha = 0.05$ (Llama-8B, Llama-17B, Llama-70B, Gemini), with per-model Wasserstein distances ranging from 25.9 to 437.2 and reaching their highest values across all regimes. The four non-significant per-model KS tests (Claude, gpt-4o, Mistral-7B, Mistral-24B) likely reflect reduced statistical power, given the 8× range in per-model group sizes (94 to 784). The pooled KS on z-scored within-model residuals is 0.130 ($p < 0.001$), comparable to the LLM regime, indicating that the distributional length–sycophancy coupling under judge agreement is at least as strong as under LLM-only labeling at the aggregate level.

Judge-disagreement regime (N = 10,226 responses across 8 models, restricted to responses where human and LLM judge labels disagree). A mixed linear model of response length on the sycophancy indicator, with model as random effect, yields a *negative* sycophantic effect of −43.1 length units (95% CI [−60.5, −25.7]; $p < 0.001$), against an intercept of 1568.7 and between-model variance 73,110. The sign reversal relative to the human (+24.3), LLM (+128.4), and agreement (+154.3) regimes is the central qualitative finding: under disagreement, responses labeled sycophantic are *shorter* than non-sycophantic ones on average. Per-model non-syco/syco ratios shift accordingly (median 1.03, with 6 of 8 models above 1.0), reversing the pattern seen in the other three regimes. Length distributions differ significantly for 7 of 8 models by Kolmogorov–Smirnov test (only Mistral-7B is not significant at $\alpha = 0.05$), with per-model Wasserstein distances ranging from 26.1 to 267.7. The pooled KS on z-scored within-model residuals is 0.046 ($p < 0.001$), smaller than the LLM and agreement regimes (both 0.130) but larger than the human regime (0.025), indicating a distributional shift that is real but less pronounced than under either single-judge or unanimous-agreement labeling. The reversal is consistent with humans and LLM judges locating sycophancy in systematically different parts of the response-length distribution: when their labels diverge, the resulting "sycophantic" set is enriched for cases each judge alone would have ruled differently on, and the length signal flips accordingly.

Within-prompt across-models robustness check (N = 11,309 responses across 1,569 prompts with label variation). To rule out prompt-level content as a confound — the possibility that the length–sycophancy association in the four mode-level regressions reflects sycophantic *prompts* eliciting longer responses rather than sycophantic *responses* being longer per se — we refit the mixed linear model with **prompt as the random effect**, restricting to the 1,569 prompts where at least two of the eight models' responses received different sycophancy labels. With group size 5–8 (mean 7.2), this within-prompt design holds prompt content fixed and identifies the sycophantic effect off cross-model variation in labeling and length on the same prompt. The estimated sycophantic effect is +58.8 length units (95% CI [40.4, 77.1]; $p < 0.001$), against an intercept of 1549.5 and between-prompt variance 47,247. The positive sign and significance confirm that sycophantic responses are longer than non-sycophantic responses *to the same prompt*, ruling out prompt-level content as the sole driver of the association. The pooled KS on z-scored within-prompt residuals is 0.073 ($p < 0.001$), indicating distributional length–sycophancy coupling that survives prompt-level conditioning. The within-prompt analysis uses the same human-label scheme as the four-mode human regression (model binary verdict on 'is_a**hole == 1' prompts); the larger coefficient (+58.8 vs. +24.3) reflects the restriction to the 1,569 prompts with cross-model label variation, not a change in labeling regime. Sycophancy effects are concentrated on prompts where models actually disagree about the verdict.

G Existing Evidence for Confounding Factors

Fact G.1. ***Reward models are correlated with response length independent of content quality****, from the first language model applications of RLHF to today [6, 19, 21, 56, 70, 71, 76, 80, 83, 91].*

Fact G.2. ***Reward models exhibit multiple additional distinct reward biases****, e.g., overconfidence [44], sycophantic user agreement [69], answer position [21], prefix bias [39], model-style sensitivity [21, 52], and inherited value orientations (e.g., agency vs. communion) from pretraining base models [13].* ***All of these are present in state-of-the-art reward models*** *[13, 21].*

Fact G.3. ***LLM policies trained with RLHF or DPO learn to generate longer outputs as a direct result of optimization pressure.*** *This effect can inflate benchmark scores [53, 71], produce outputs longer than training data [59], and grow systematically longer than SFT baselines over training [76]. Overall, it represents a primary mode of reward hacking [59, 71].*

Fact G.4. ***Naive length penalties can remove informative content****, as uniform length suppression degrades accuracy on tasks that genuinely require longer answers [3], and always-on penalties cause reward hacking via trajectory collapse in reasoning RL [89].*

Fact G.5. ***Longer responses are systematically over-represented among "chosen" annotations in preference datasets, providing the distributional basis for reward models to learn length as a proxy for quality*** *[70, 71]. RLHF dramatically amplifies a slight length skew already present in preference data [19], a pattern confirmed in crowdsourced human preference data, where controlling for length substantially reorders model rankings [46]. Further, length preferences vary systematically across annotator populations and data sets [54].*

Fact G.6. ***Human preference judgments are systematically confounded by output length.*** *Across multiple preference-collection pipelines, annotators select the longer response in comparisons [6, 29, 71]. This pattern is visible even in carefully filtered datasets, as Stiennon et al. [76] find that controlling for length reduces human preference gains and Saito et al. [65] show a positive correlation between length and preference in the HH-RLHF corpus. In large-scale crowdsourced evaluation, length has been estimated as a dominant style factor in human voting [46].*

Fact G.7. ***LLM judges exhibit verbosity bias [92] and it is different to human annotators [38, 65].*** *Verbosity is one of several systematic biases catalogued in LLM judges, in addition to other LLM judge biases [6, 38, 82, 85]. Further evidence also supports that human and LLM judges can respond to length and content differently [12, 29, 46, 65].*

Fact G.8. ***Human annotators systematically reward sycophantic responses.*** *Sharma et al. [69] show that "matches user's beliefs" is among the most predictive features of human preference, with the preference model favoring sycophantic over truthful responses. Cheng et al. [10] confirm this finding at the data level, as preferred responses across preference datasets are significantly more validating and indirect. Perez et al. [62] find that preference models incentivize sycophantic answers, while Wen et al. [83] show RLHF-optimized models learn to defend incorrect answers with fabricated evidence and extended argumentation, increasing human approval without improving correctness. Ibrahim et al. [32] further show that supervised fine-tuning on warmer-style conversational data alone (ithout preference optimization) is sufficient to amplify affirmation of incorrect user beliefs by approximately* 40% *across five model families, with the effect surviving response-length controls.*

Fact G.9. ***Whether responses are labelled as sycophantic statistically depends on response length, independent of annotator type.*** *Prior work has reported response length and sycophancy jointly: Ibrahim et al. [32] include length as a covariate when regressing accuracy and sycophancy on warmth fine-tuning, and Dubois et al. [17] treat length as a nuisance parameter when measuring sycophancy reduction. To our knowledge, however, the statistical dependence between response length and sycophancy labels has not been characterized across labeling regimes (human, LLM judge, agreement, disagreement), nor has the sign reversal under human-LLM judge disagreement been documented. We examine the dependence across multiple labeling strategies (human labels, LLM Judge labels, their agreement, and disagreement), systematically disentangling model-level verbosity from prompt-level content as alternative explanations using the AITA dataset from [9]. We find length and sycophancy of responses are statistically dependent across all four labeling regimes, see Section F.3 for experiment details. Sycophantic responses are longer under human, LLM, and human-LLM judge agreement labels, with the relationship reversing under human-LLM judge disagreement labels. This observation is consistent with humans and LLM judges locating sycophancy in systematically different parts of the response-length distribution. In addition, we find that the length-distribution of sycophantic and non-sycophantic responses significantly deviates for most models (Kolmogorov-Smirnov) and that LLM judge labels are substantially inflated by length bias relative to human labels.*

Fact G.10. ***Sycophancy causes models to abandon correct responses in favor of alignment with user-stated positions.*** *When users suggest incorrect answers, model accuracy drops relative to unbiased baselines [69] with high agreement rates with user-stated views for some question types [62]. Fanous et al. [20] quantify this as regressive sycophancy, observing models switch from correct to incorrect under user rebuttals and Hong et al. [28] further show that this is not a knowledge deficit. Beyond propositional agreement, framing sycophancy, in which models uncritically accept flawed user premises, proved particularly resistant to DPO-based mitigation [10]. Also, sycophantic affirmation narrows users focus to self-validation while omitting alternative perspectives [9].*

Fact G.11. ***Models can generate longer responses under epistemic uncertainty by hedging, qualifying, and elaborating more when confidence is low*** *[90].*

Fact G.12. ***Reward models treat expressed confidence as a quality proxy, incentivizing models to replace hedged answers with confident-sounding responses regardless of actual certainty [44].*** *Reward models can penalize hedged correct answers over confidently-stated incorrect ones [21], and this penalty can be traced to the annotation level, where human raters in preference datasets are biased against expressions of uncertainty [31, 93].*

Fact G.13. ***Longer responses in uncertain regimes correlate with lower confidence and reduced quality.*** *Verbose outputs produced under verbosity compensation show measurably higher uncertainty and recall drops relative to concise responses [90]. Reward models scoring confident-sounding elaborations highly regardless of correctness [44] implies the same inverse relationship, though without directly measuring it.*

Fact G.14. ***Reasoning models (not RLHF) produce longer outputs with uncertainty markers.*** *Extended chain-of-thought training produces epistemic uncertainty markers and non-linear reasoning traces through backtracking and alternative exploration rarely seen in non-reasoning models [88]. During on-policy RL training, trajectory length grows systematically alongside exploratory behavior, and naively penalizing length degrades performance [89], consistent with RL reinforcing the coupling.*

Fact G.15. ***Reward model scores increase with length when additional tokens carry genuine informational content*** *[29], but this relationship degrades at greater lengths, entering sublinear and then stochastic regimes beyond 100*

and 200 tokens respectively [91], consistent with ***diminishing informational returns per token as responses grow longer.***

Fact G.16. ***Some reward-length correlation is irreducible and reflects genuine informativeness rather than spurious shortcut.*** *The information mass decomposition directly establishes that a portion of the length–reward correlation tracks real content quality [29], and more detailed answers can be genuinely more helpful depending on context [3]. Consistent with this observation, length-controlled analysis implies the original correlation is not entirely spurious [18].*

Fact G.17. ***DPO-like objectives develop length bias out-of-distribution as a structural vulnerability of the algorithm without explicit mitigations****. DPO's implicit reward as a token-level log-probability ratios grows with sequence length, creating an algorithmic dependence on response length within the objective [51]. Empirically, this dependence manifests as the implicit reward acquiring significant length correlation once the policy moves away from the training distribution [53, 59], and can be mitigated through reward normalization and margin formulation [45]. Independently, under noisy or capability-limited supervision, DPO's KL regularization creates a structural dilemma where sufficient regularization preventing over-optimization also prevents larges updates needed for correcting errors inherited from the SFT phase [87].*

Fact G.18. ***KL regularization interacts with reward scale non-trivially.*** *KL-regularized policies are not invariant to positive linear reward scaling (see Section E.5), meaning two reward models agreeing on all preference orderings but differing in cardinal scale produce different policies [72]. Empirically, there exists a measured optimal KL budget beyond which true reward degrades in RLHF [25] and longer sequences accumulate disproportionately more KL divergence, entangling length with the implicit DPO reward formulation [51]. The optimal regularization coefficient β depends strongly on feedback reliability [87], meaning there is no single good KL setting across realistic annotation conditions.*

Fact G.19. ***Spurious length features can maintain ordinal ranking accuracy while distorting cardinal reward values, meaning benchmark metrics may fail to detect this failure mode.*** *A dissociation that follows from the partial-identifiability characterization of Skalse et al. [72] combined with the KL scale-sensitivity of Section 2 (Equation (1)), and is directly measured as a low-complexity linear artifact in reward model representations [21]. RMs built on different base model families can achieve similar RewardBench scores while exhibiting systematically divergent value orientations [13] or producing systematically different PPO outcomes depending on policy-RM lineage match [52], concrete empirical instances of ordinal-invariant but cardinally-divergent reward functions. RewardBench [43] scores improve after post-hoc length calibration, confirming retrospective distortion without prior detection [30], and apparent leaderboard gains shrink substantially under length-controlled evaluation in both automated [18] and crowdsourced human voting settings [46].*

Fact G.20. ***Evaluator and annotator weaknesses are learnable targets. Models trained with RLHF can acquire exploitation strategies like length padding, hedging caveats, and code obfuscation.*** *These strategies are specific to evaluator blind spots rather than tied to any single feature, demonstrating that reward gaming is a general learned behavior directed at the structure of the evaluation process itself [83].*

Fact G.21. ***Reward can be somewhat decomposed into separable dimensional components.*** *A two-head architecture directly separates length and content reward signals, confirms the components can be disentangled with supervised training signals to improve mitigation [6]. Consistent with this, multi-attribute scoring frameworks track helpfulness, correctness, coherence, complexity, and verbosity as separately annotated (but correlated) dimensions [56, 81].*

Fact G.22. ***Reward biases admit to first-order linear intervention.*** *Length bias is smooth, structured, and stable enough across model families to be estimated and corrected post-hoc via LOESS without retraining [30], but assumes that the*

true reward is independent of length. Fein et al. [21] also use a linear probe to mitigate some part of length and other biases in reward models, while Papadatos and Freedman [58] used linear probe-based reward corrections to mitigate sycophancy on simplified controlled setting.

Fact G.23. ***Mitigating a spurious correlation in one feature can increase reward hacking in another feature (bias substitution).*** *In supervised vision, mitigating a single labeled shortcut amplifies reliance on the unlabeled one, with off-target axis degradation of 2–3x while aggregate worst-group accuracy improves [48]. In RLHF, aggressive length-debiasing in recent SOTA reward models has flipped the length-bias sign and reduced correctness [21]. In LLM preference optimization more broadly, several DPO bias-mitigation variants reduce targeted bias at significant cost to general capability [22], instantiating R2. The mechanism is compatible with the single-proxy correlated-feature bound showing that drift along an unconstrained correlated direction remains available to the policy whenever the proxy–truth correlation is below one [41]. Liu et al. [50] address this at the reward level via max-min optimization, but without the R0–R4 diagnostic machinery, substitution between interpretable feature axes remains undetectable. Conversely, jointly mitigating multiple shortcuts via kernel-based regularization [86] achieves near-zero shortcut correlations at the audit distribution, but without measuring whether optimization pressure has shifted at μ_{π^*}, instantiating the distributional blindspot of Section A.5.*

H Mapping published mitigations onto the regime taxonomy

The detailed classification for each reward bias mitigation method we distill in Section B. No published mitigation paper we find provides evidence sufficient to certify R0.

- Park et al. [59] (R-DPO) — undetermined; consistent with $\mathrm{R0}_{\mathrm{cont}}$. Length penalty validated only on preference data; no off-target Δ_j or fixed-π_{ref} Δ_J reported.
- Meng et al. [53] (SimPO) — undetermined; consistent with $\mathrm{R0}_{\mathrm{cont}}$, R3 plausible. The debiasing mechanism is reward rescaling via $|y|$-normalization, exactly the cardinal-scale failure mode of Corollary A.4.
- Lu et al. [51] (SamPO) — undetermined; consistent with $\mathrm{R0}_{\mathrm{cont}}$ or R1. Algorithmic length dependence in DPO addressed analytically; empirical validation stays at the audit distribution and does not isolate Δ_J from length reduction.
- Fu et al. [24] (PAR) — undetermined on bias; arguably R0/R3 on the training-stability target it actually addresses. Single Gemma-2B lineage, no off-target measurement.
- Li et al. [45] (LMPO) — undetermined; consistent with $\mathrm{R0}_{\mathrm{cont}}$. The reported LC win-rate gain is on the same metric the loss is built around, which is Goodhart-style and does not certify movement along the structural axis.
- Shen et al. [70] (Loose Lips) — undetermined; consistent with $\mathrm{R0}_{\mathrm{cont}}$. Validation distribution coincides with the RM training distribution; only pooled length correlation reported.
- Chen et al. [8] (ODIN) — undetermined; $\mathrm{R0}_{\mathrm{cont}}$ vs R1 unresolved. Two-head disentanglement zeros pooled length correlation; Lemma 3.6 says this does not transfer to $\mu_{\pi^\star}$.
- Wang et al. [81] (HelpSteer) — undetermined; consistent with R3. Multi-attribute data construction with ordinal MT-Bench evaluation; no operator-level mitigation claim the framework can adjudicate.
- Bu et al. [3] (ALBM) — direct $\mathrm{R2}_\varepsilon$ evidence on the targeted axis (length-asymmetric subset accuracy regression in Table 1) plus undetermined $\mathrm{R0}_{\mathrm{cont}}$ overall. No rescalability sweep to disambiguate scale overshoot from target misspecification.

- Huang et al. [30] (post-hoc LOESS calibration) — direct $R2_\varepsilon$ evidence on the targeted axis with R1 plausible. Section B.1 shows pooled correlation 0.316 → 0.037 but within-prompt sign flip on three of four SOTA RMs and $\Delta_J < 0$ on two AlpacaEval cells.
- Zhao et al. [91] (FiMi-RM) — direct R2 or R3 evidence (overall accuracy 70.8 → 69 reported in the paper itself), with $R0_{\text{cont}}$ undisambiguated. Three-phase length-reward pattern identified but no rescalability sweep.
- Wang et al. [80] (causal rewards) — undetermined; consistent with $R0_{\text{cont}}$. Synthetic identifiability sound; real-world evidence is correlation with GPT-4 and humans, not isolated Δ_J.
- Ng et al. [55] (debiasing with guarantees) — undetermined; consistent with $R0_{\text{cont}}$. Identifiability theorem is conditional on the latent-variable DAG, which preference data does not pin down [72]; empirical validation is small/synthetic.
- Fein et al. [21] (mechanistic reward shaping) — closest to a positive R0 signal; R0/$R0_{\text{cont}}$ undisambiguated. Section B.1 shows $\Delta_J > 0$ on four of five RMs at within-prompt μ_{diag}, but off-target Δ_j across the rest of Φ_{sp} is not measured.
- Li et al. [47] (DIR, info-theoretic) — undetermined; consistent with $R0_{\text{cont}}$. Information-bottleneck framing handles non-linearity in principle; empirical evidence conflates bias mitigation with downstream task improvement.
- Eisenstein et al. [19] (RM ensembles) — direct $R0_{\text{cont}}$/R1 evidence. The paper itself documents shared failure modes where ensemble members agree on length doubling and copying.
- Liu et al. [50] (worst-case correlated proxies) — undetermined; consistent with $R0_{\text{cont}}$ or R3. Improves a worst-case proxy bound; the worst-case proxy is not the true reward, and substitution onto features outside the correlation set is undetected.
- Feng et al. [22] (C2PO) — direct R2 evidence. Reduces targeted bias with documented general-capability degradation, the named R2 pattern.
- Ye et al. [86] (PRISM) — undetermined; consistent with $R0_{\text{cont}}$ with R1 plausible on uncovered axes. Near-zero correlation on the three measured shortcuts only; no off-panel measurement.
- Kim et al. [36] (CDA via causal lens) — undetermined; consistent with $R0_{\text{cont}}$ on RewardBench-1, R3 on RewardBench-2. Length-controlled gains real on the easier benchmark; small gains on the harder one.
- Cai et al. [4] (Rc-BT) — undetermined; consistent with $R0_{\text{cont}}$. Strongest gesture toward prescription 1 in the surveyed set (PPO included in Table 11), but no off-target Δ_j across Φ_{sp}.
- Srivastava et al. [75] (CHROME) — undetermined; consistent with $R0_{\text{cont}}$. Best multi-prescription coverage in the surveyed set (multi-axis Table 7, BoN Figure 5, reWordBench transformations Tables 12–13), but the LLM-oracle counterfactuals inherit R4 sensitivity that is not measured.
- Song et al. [74] (SAE causal adjustment) — undetermined; consistent with R0/$R0_{\text{cont}}$. SAE features give partial Φ_{sp}, but the panel is incidental rather than designer-specified, and there is no fixed-π_{ref} evaluation.
- Chen et al. [7] (OPRM with RgFT) — undetermined; consistent with $R0_{\text{cont}}$. Calibration framing (ECE, Brier in Table 5) is the closest engagement with the cardinal/ordinal distinction in the surveyed set, but the protocol does not exploit it for substitution detection.