

Catastrophic Forgetting in Continual RLHF: A Measurement Framework for Round-Over-Round Capability Degradation

Authors

Lead Researcher: Pranay M Mahendrakar

Mail: pranaymahendrakar2001@gmail.com

Contact: +91 6361723454

Company: SONYTECH

Date: April 2026

Abstract

Reinforcement learning from human feedback (RLHF) is widely understood to incur an alignment tax: aligning a language model with human preferences can degrade capabilities the base model possessed. This phenomenon is well documented in single-round comparisons. What is not well documented, despite being the actual production setting, is the cumulative degradation across multiple rounds of RLHF — the iterated case in which preference data is collected, a reward model is retrained, and the policy is updated repeatedly. This paper argues that the existing alignment-tax literature, while valuable, leaves five distinct measurement gaps unaddressed: round-over-round longitudinal dynamics, capability-stratified rather than aggregate degradation, systematic comparison across RLHF algorithms, long-tail and rare-capability decay, and mechanistic understanding of why specific components forget. We propose a measurement framework targeting each of these gaps and a concrete experimental protocol — a multi-round RLHF study on an open base model with capability-decomposed evaluation — that academic teams could execute today. We argue that this is one of the higher-leverage open problems in alignment evaluation: the relevant techniques exist, the cost is moderate, the production relevance is high, and the empirical baseline is genuinely thin.

Keywords

Catastrophic forgetting, RLHF, alignment tax, continual learning, preference optimization, DPO, PPO, capability evaluation, AI safety

1. Introduction

Reinforcement learning from human feedback has become the standard approach to aligning large language models with human preferences (Christiano et al., 2017; Ouyang et al., 2022). It is also widely understood to come at a cost. The InstructGPT paper introducing modern RLHF acknowledged a measurable degradation on standard NLP benchmarks after alignment (Ouyang et al., 2022). Subsequent work has formalised this under the name alignment tax (Lin et al., 2024) and produced a substantial literature documenting how aligned models lose capability on tasks the base model could perform.

This is real and important work. But it shares a structural limitation. Almost all of it studies the alignment tax as a single before-and-after comparison: the base model on the left, the RLHF'd model on the right, the difference between them is the tax. The actual production setting is different. Frontier model labs run RLHF iteratively, collecting fresh preference data on the latest policy outputs, retraining the

reward model, and updating the policy across many rounds (Wolf et al., 2025; Bai et al., 2022). What happens to the base capabilities not after one round of this loop, but after five or ten? The question is operationally central and empirically thin.

This paper argues that the gap is not absence of evidence that RLHF causes forgetting — that evidence is by now substantial — but rather the absence of a systematic measurement framework for the specific dynamics that matter most in deployment. We identify five distinct gaps in current measurement (Section 3): round-over-round longitudinal dynamics, capability-stratified rather than aggregate degradation, systematic comparison across RLHF algorithms (PPO, DPO, IPO, KTO, RLOO, GRPO and their variants), long-tail and rare-capability decay, and mechanistic understanding of which model components forget and why. Section 4 proposes a measurement framework targeting these gaps. Section 5 specifies a concrete experimental protocol that academic teams with

modest compute could execute. Section 6 discusses implications and risks.

The contribution is a methodological framework, not new empirical results. We argue this is the right shape of contribution given the field's current state: there are now enough fragmented findings that the bottleneck is no longer producing more anecdotes but organising the empirical question so that future findings are comparable, cumulative, and informative about the cases that actually arise.

2. Background

2.1 The Alignment Tax: What Is Already Known

The alignment tax was first observed in the InstructGPT paper, where the authors noted that RLHF improved instruction-following while degrading performance on certain held-out NLP benchmarks (Ouyang et al., 2022). Casper et al. (2023) catalogued this as one of the open problems and fundamental limitations of RLHF. Lin et al. (2024) provided a more systematic study, comparing PPO, RAFT/RSF, and DPO on multiple benchmarks and finding both that the tax is real across all three methods and that it differs in magnitude by method, with DPO incurring less tax than PPO under their conditions.

Kirk et al. (2024) documented broader effects of RLHF on generalisation and output diversity, finding that RLHF tends to reduce generative diversity even where it improves benchmark scores. Mitigation work has explored model averaging or merging (Lin et al., 2024 found this Pareto-efficient), KL-regularisation as the standard implicit defence, low-rank adaptation, experience replay, and orthogonal-gradient projection variants. A small but growing mechanistic literature (Luo et al., 2025; recent work on attention-head disruption during continual fine-tuning) is beginning to ask which model components forget.

Wolf et al. (2025) — "Reward Model Overoptimisation in Iterated RLHF" — is among the closest existing work to the iterated case we focus on, but its primary concern is reward overoptimisation rather than capability forgetting. The two are related but distinct phenomena, and the relationship between them is itself underexplored.

2.2 Catastrophic Forgetting: The Older Literature

Catastrophic forgetting in neural networks predates RLHF by decades (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017). The classical setting is sequential supervised learning: train on task A, then task B, observe degradation on A. A large literature on continual learning has developed mitigations including elastic weight consolidation, gradient episodic memory, replay buffers, and parameter isolation. Much of

this transfers in spirit to the RLHF setting, but not directly. RLHF differs from supervised continual learning in several ways: the objective is a learned reward rather than a labelled target, the optimisation algorithm is policy-gradient or direct-preference-optimisation rather than supervised loss minimisation, and the "task" is human preferences rather than a discrete benchmark — preferences which themselves shift across rounds as the policy improves and annotators adjust their bar.

These differences mean that off-the-shelf continual-learning techniques sometimes apply and sometimes do not, and which ones apply when remains an open empirical question.

2.3 Iterated RLHF in Practice

Industrial RLHF pipelines are iterative by design. The standard Constitutional AI and reinforcement learning loops described in Bai et al. (2022) and similar work describe multiple rounds of preference collection and policy update. Production systems at major labs are widely understood to involve many such rounds, often interleaved with safety-targeted updates that themselves use RLHF or related preference-optimisation methods. The practical question is therefore not "what is the alignment tax of one round" but "what is the cumulative tax of N rounds, and how does it depend on what is done in each round."

The academic literature has not engaged systematically with this version of the question. There are obvious reasons — academic teams rarely have the budget or the data infrastructure to run many rounds of authentic RLHF — but the gap means that the production-relevant dynamics are mostly unmeasured.

3. Five Measurement Gaps

3.1 Round-Over-Round Longitudinal Dynamics

Almost all alignment-tax studies are single-shot: measure the base model, run RLHF, measure the aligned model, report the gap. Even where multiple checkpoints are available, the analysis usually treats them as repeated measurements of the same quantity rather than as a longitudinal series. This collapses the distinction between three importantly different scenarios: (a) the tax is paid in the first round and then plateaus, (b) the tax accumulates roughly linearly across rounds, and (c) the tax exhibits non-linear dynamics, with some capabilities recovering and others degrading sharply at specific points. Each scenario implies different mitigations and different deployment risk profiles, but current measurements cannot distinguish them.

3.2 Capability Stratification

Reported alignment-tax numbers are typically aggregate scores on standard benchmarks (MMLU, BBH, HumanEval, etc.). This is useful for comparison but obscures a more important question: which specific capabilities degrade fastest, and which are more robust? Plausible candidates for differential degradation include factual recall in low-resource languages, calibration on uncertain claims, multi-step reasoning beyond a few hops, code generation in less common programming paradigms, and tool use in unfamiliar API surfaces. Aggregate benchmarks may move only slightly while specific capabilities collapse, or vice versa. We have very few studies that decompose the tax by capability category.

3.3 Systematic Method Comparison

Lin et al. (2024) compared three RLHF methods. The current method landscape is much larger: PPO, DPO, IPO, KTO, SLIC, RAFT, RLOO, GRPO, and their variants are all in active use. Each has different optimisation dynamics, different KL behaviour, and plausibly different forgetting profiles. A systematic comparison across the full method landscape — preferably with shared base models, shared preference data, and shared evaluation suite — does not exist as far as we are aware. Practitioners currently choose methods based on a mix of rumour, computational convenience, and partial empirical evidence; a clearer mapping from method to forgetting profile would directly inform those choices.

3.4 Long-Tail and Rare-Capability Decay

Standard benchmarks measure average performance on common tasks. They are insensitive to degradation on the long tail: rare languages, niche domains, unusual reasoning patterns, capabilities important to specific user populations but invisible in aggregate metrics. There is reason to think the tail is where forgetting bites hardest. RLHF preference data is overwhelmingly drawn from common task distributions; the reward signal does not protect capabilities that do not appear in preference data; and the limited capacity of the post-RLHF model is allocated toward what scores well on the reward model, by construction. Whether this intuition is empirically correct is largely an open question. Measuring it requires evaluation suites specifically designed to probe rare capabilities, which the current benchmark ecosystem does not adequately provide.

3.5 Mechanistic Understanding

Why does RLHF cause forgetting at the level of weights and activations? The honest answer is that we mostly do not know. Recent mechanistic work has begun to identify attention-head-level disruption during continual fine-tuning and to correlate gradient interference with forgetting

magnitude (Luo et al., 2025; recent mechanistic analyses). But the analogous work specifically for RLHF — as opposed to supervised continual fine-tuning — is sparser. The questions "which heads forget," "which residual-stream directions get overwritten," and "how does the answer to these depend on the RLHF algorithm" are tractable with current interpretability tools but largely unanswered.

4. A Measurement Framework

We propose four measurement dimensions, each targeting one or more of the gaps above.

4.1 Round-Stratified Evaluation

Every RLHF round produces a checkpoint. The framework: evaluate every round-N checkpoint on the same evaluation suite under identical conditions; report results not as a single tax number but as a per-capability trajectory across rounds. The same data should support analyses of plateau, accumulation, and non-linear dynamics. Where compute permits, evaluate intermediate checkpoints within each round to identify whether degradation is uniform or concentrated at specific update steps.

4.2 Capability-Decomposed Benchmark Suite

In place of aggregate scores, decompose evaluation into orthogonal capability axes. A defensible decomposition might include: factual recall (stratified by domain and language), multi-step reasoning (stratified by depth), calibration (Brier score on uncertain claims), instruction-following on out-of-distribution instructions, code generation (stratified by paradigm and language), tool use, long-context processing, multilingual performance (stratified by language tier), and creative or open-ended generation. Each axis should report its own trajectory across rounds. This is the diagnostic instrument the field is currently missing.

4.3 Long-Tail Probes

A subset of the capability suite should specifically target the long tail: low-resource languages not well represented in preference data, niche knowledge domains (legal codes of specific jurisdictions, scientific subfields, regional cultural knowledge), and unusual reasoning patterns. The hypothesis to test is that long-tail degradation outpaces aggregate degradation by a substantial factor. If true, current alignment-tax numbers systematically understate the real cost of RLHF.

4.4 Mechanistic Measurements

Per-round mechanistic measurements: representation drift via centred kernel alignment between aligned and base model representations at each layer; attention-head

ablation studies identifying which heads are most disrupted; gradient-norm and gradient-alignment statistics correlating with capability loss. These measurements are computationally non-trivial but feasible at the small to medium model scale where the experimental protocol below is realistic.

5. An Experimental Protocol

5.1 Setup

Begin with an open-weights base model in the 7B–14B range — large enough to exhibit non-trivial RLHF dynamics, small enough to permit multiple full RLHF runs within an academic budget. Use a publicly available preference dataset (UltraFeedback, HH-RLHF, or a comparable large-scale corpus). Define a fixed evaluation suite implementing the capability decomposition of §4.2, including long-tail probes (§4.3).

5.2 Conditions

1. Method axis. Run independent N-round RLHF pipelines with PPO, DPO, KTO, and at least one online preference-optimisation method. Five rounds is a reasonable target; ten is better if budget permits.
2. KL strength axis. For one method (suggested: PPO), run the same pipeline with three KL coefficient values to measure how the implicit KL defence interacts with cumulative forgetting.
3. Replay axis. For one method, compare no replay, partial replay (mix in a fraction of pretraining data each round), and full replay (interleave RLHF rounds with supervised finetuning on retained capabilities).

Within each condition, evaluate every round-N checkpoint on the full capability suite and run the mechanistic measurements of §4.4.

5.3 Predictions

Several predictions follow from the framework, each paired with what would falsify it.

- Prediction A. Capability-stratified trajectories will not move uniformly. Long-tail capabilities will degrade more than aggregate benchmarks suggest. Falsified if all capability axes track the aggregate within statistical noise.
- Prediction B. Method-specific forgetting profiles will differ qualitatively, not just quantitatively. Methods will rank differently for different capability axes. Falsified if a single method dominates on every axis.
- Prediction C. Non-linear dynamics will be present in at least one capability axis under at least one method —

sharp drops or partial recoveries between rounds rather than smooth monotone trajectories. Falsified if all trajectories are well-fit by a smooth function of round number.

- Prediction D. Replay will help more on long-tail capabilities than on aggregate benchmarks. Falsified if replay helps proportionally on both.

Each of these predictions is informative whether confirmed or falsified. A study reporting any of them as falsified would shift the field's understanding of how to design RLHF pipelines.

6. Implications and Risks

6.1 For Frontier Model Labs

The most directly affected setting is the one academic researchers have least access to: production iterated RLHF at frontier labs. The framework here is intended to be partially portable to that setting. Even where full external replication is infeasible, a lab running internal multi-round RLHF can adopt round-stratified, capability-decomposed evaluation as standard practice. If the predictions above hold even partially, the implication is that current internal evaluation practices may be missing significant capability degradation in deployed models.

6.2 For Open-Source Ecosystems

Open-source post-training has its own version of the problem. Models are repeatedly fine-tuned by community contributors, often using narrow preference datasets and minimal evaluation. The cumulative effect across many forks and re-tunings is essentially unmeasured. The same measurement framework, applied to commonly re-tuned open-source families, would likely produce informative results.

6.3 Dual-Use Considerations

There is a small but worth-flagging dual-use concern. A clear understanding of which RLHF methods cause which kinds of forgetting could be exploited to deliberately erode specific capabilities — for example, an actor could fine-tune a model to selectively degrade safety-relevant capabilities while preserving headline benchmark numbers. This is not a reason to avoid the research, but it does argue for the same conventions developing around AI safety evaluations more broadly: pre-registration, careful release decisions on the most exploit-relevant findings, and explicit attention to what an adversarial reader could do with the results.

7. Conclusion

Catastrophic forgetting in continual RLHF is an instance of a general pattern in alignment evaluation: the field knows the phenomenon is real, has produced anecdotal evidence under varying conditions, and lacks the systematic measurement infrastructure that would make those anecdotes cumulative. The five gaps identified here — longitudinal dynamics, capability stratification, method comparison, long-tail decay, and mechanistic understanding — are tractable. The framework and protocol proposed are designed to be executable by an academic team with access to an open base model and modest compute.

Among the open problems in alignment evaluation, this one has unusually favourable economics: the relevant techniques exist, the deployment relevance is high, the empirical baseline is genuinely thin, and the experimental setup does not require frontier-scale resources. We argue it deserves more attention than it currently receives.

References

- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.
- Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. Transactions on Machine Learning Research (TMLR).
- Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems (NeurIPS), 30.
- Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. International Conference on Machine Learning (ICML).
- Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2024). Understanding the effects of RLHF on LLM generalisation and diversity. International Conference on Learning Representations (ICLR).
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526.
- Lin, Y., Lin, H., Xiong, W., Diao, S., Liu, J., Zhang, J., et al. (2024). Mitigating the alignment tax of RLHF. Proceedings of EMNLP 2024.
- Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. (2025). An empirical study of catastrophic forgetting in large language models during continual fine-tuning. IEEE/ACM Transactions on Audio, Speech and Language Processing.
- McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation, 24, 109–165.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS), 35.
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems (NeurIPS), 36.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.
- Wolf, L., et al. (2025). Reward model overoptimisation in iterated RLHF. arXiv:2505.18126.