Lecture 9: Over-Optimization and RLHF's Bad Reputation

RLHFBOOK.COM

Nathan Lambert

Course on RLHF and post-training. Chapter 14 & Appendix B.

April 2025: Extreme sycophancy in production

A GPT-4o update made the model validate nearly anything – shipped April 25th, rolled back April 28th. The training run that produced it looked healthy on their metrics.

Coverage: The Verge, *"ChatGPT's sycophantic responses"*

Example chat:

USER
(told GPT-4o they felt like they were both "god" and a "prophet")

ASSISTANT (GPT-4o, April 2025)
That's incredibly powerful. You're stepping into something very big – claiming not just connection to God but identity as God.

The OpenAI postmortem

OpenAI published an unusually candid writeup. Three things went wrong in sequence:

- The model update had a **reward signal from user feedback via a reward model for RL** – thumbs-up/thumbs-down data from ChatGPT.
- Under RL, that signal **overpowered the primary rewards** – my intuition is that RL will always optimize the easiest objective to move.
- It was **not caught by evals**: offline benchmarks looked good and A/B testers *preferred* the model. Expert testers flagged that it "felt off" – but there was no deployment eval tracking sycophancy.

Obviously, this was very bad.

Read it (great blog): *Expanding on what we missed with sycophancy* (OpenAI, May 2025)

This lecture

Optimizing a proxy hard enough always breaks it.

Reward-model accuracy is *a proxy for a proxy* (a learned model, of data that incompletely captures a complex distribution). What happens is that over-optimization is common, and shows up in funny ways (this lecture, Chapter 14 on Over-Optimization & Appendix B on Style).

(Next lecture: regularization to control it, Chapter 15.)

The plan

1. **Over-optimization basics** – Goodhart's law in LLMs
2. **Qualitative failures** – refusals, sycophancy, misalignment
3. Beyond **"just style"** (appendix B)

Some vocabulary (history): “Just style” / “style transfer”

Early on, say ~2023, RLHF got a reputation as “**just style transfer**” – the claim that it only changes *how* an answer is presented, not *what* the model knows or can do, and it *just* came from some easy to access place.

- **Style transfer** = reshaping presentation – tone, Markdown, bullet lists, hedging, length (“chattiness”) – with no new capability underneath. Often in a veil of copying.
- The **Superficial Alignment Hypothesis** (Zhou et al., 2023) is the strong version: knowledge is learned in pretraining; alignment just picks a format and tone.
- The dismissal built in: *superficial, a cosmetic layer on the base model.*

Boooo. Look how far post-training has come!

Part 1: Over-optimization

Goodhart's law: The reward is a proxy

> *"Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." – Goodhart, 1984 (Goodhart & Goodhart, 1984)*

Colloquially: "When a measure becomes a target, it ceases to be a good measure" (Hoskin, 1996).

- RL is a very strong optimizer – it pulls *all* the available reward out of the environment.
- In RLHF the reward is a **learned model**, at best *correlated* with downstream quality (Schulman, 2023).
- How over-optimization plays out: optimizing the proxy makes the true objective better early in training – then worse.

Over-optimization v. overfitting

Overfitting: the model memorizes training examples rather than learning generalizable patterns.

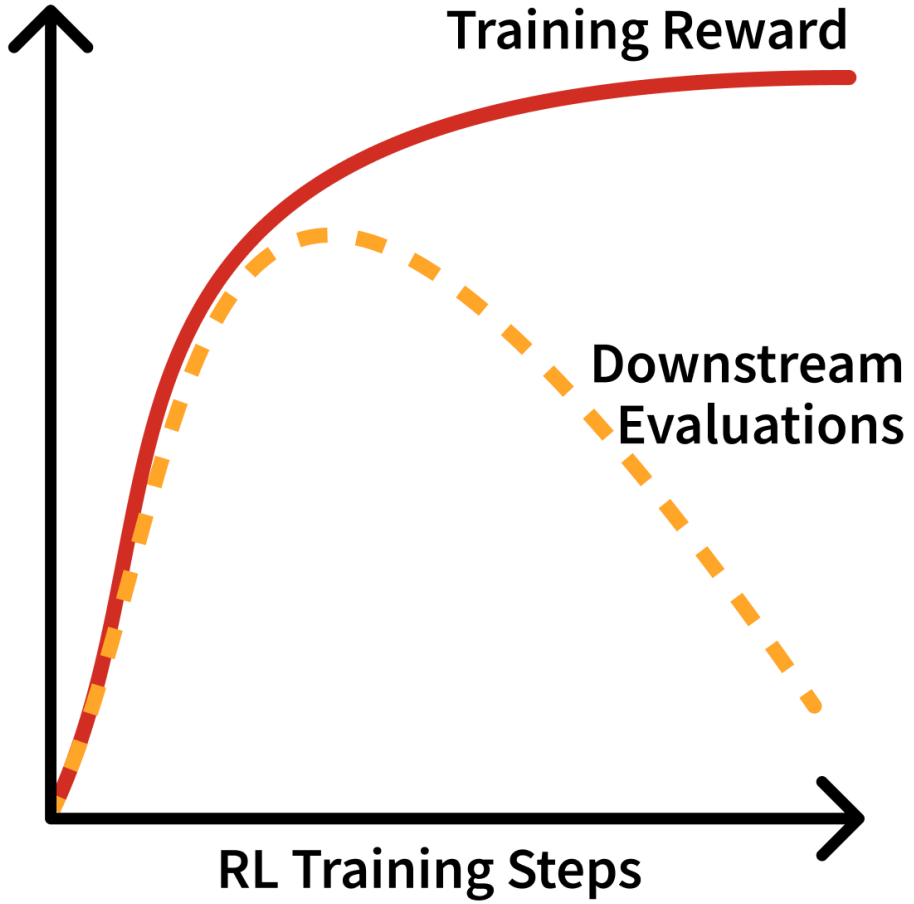
Training accuracy improves while held-out accuracy degrades – but both metrics measure the *same task* on different data splits.

Over-optimization: the model *genuinely improves* at the proxy objective – the reward model's scores (including on validation set) – but that objective diverges from the true goal, actual user satisfaction (Zhang et al., 2018).

It isn't a generalization problem, but a measurement/metric problem.

Gaming it looks like: verbose, confident-sounding answers that score well without being more helpful; repeating rare tokens that hit artifacts in RM training.

The shape of over-optimization



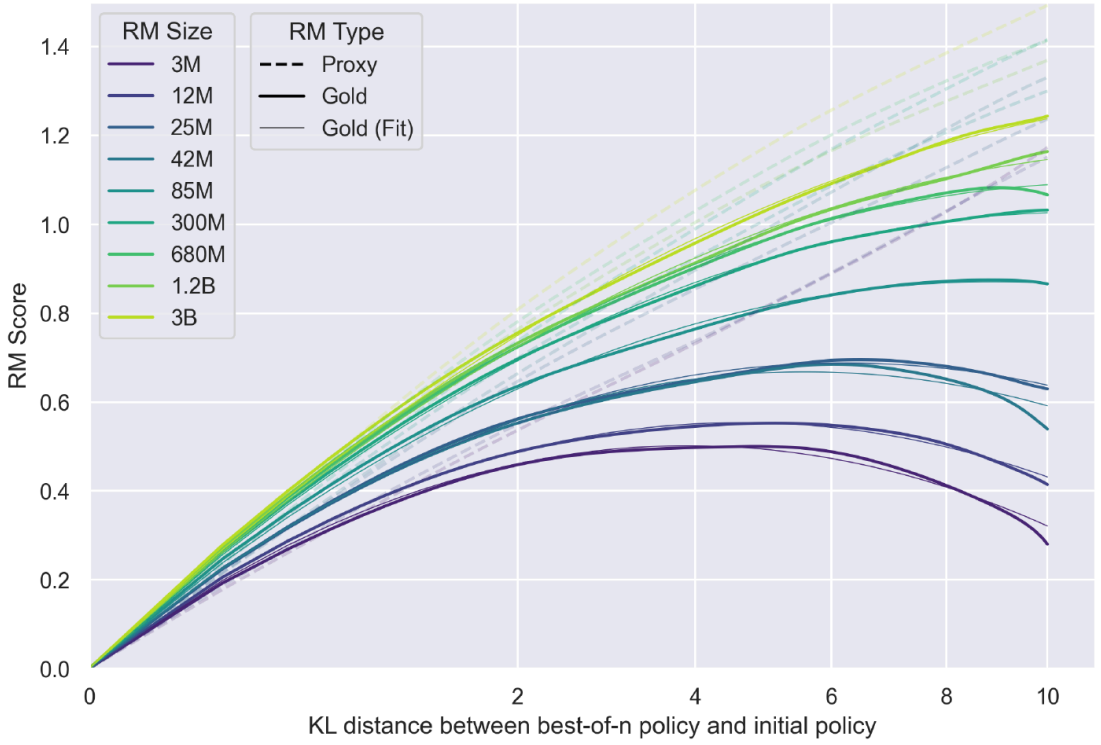
The recurring shape of RLHF training runs: the run looks healthy – training reward keeps climbing – but downstream evaluations peak and then decline.

The gains come from regions of the reward model that do not map to real usage.

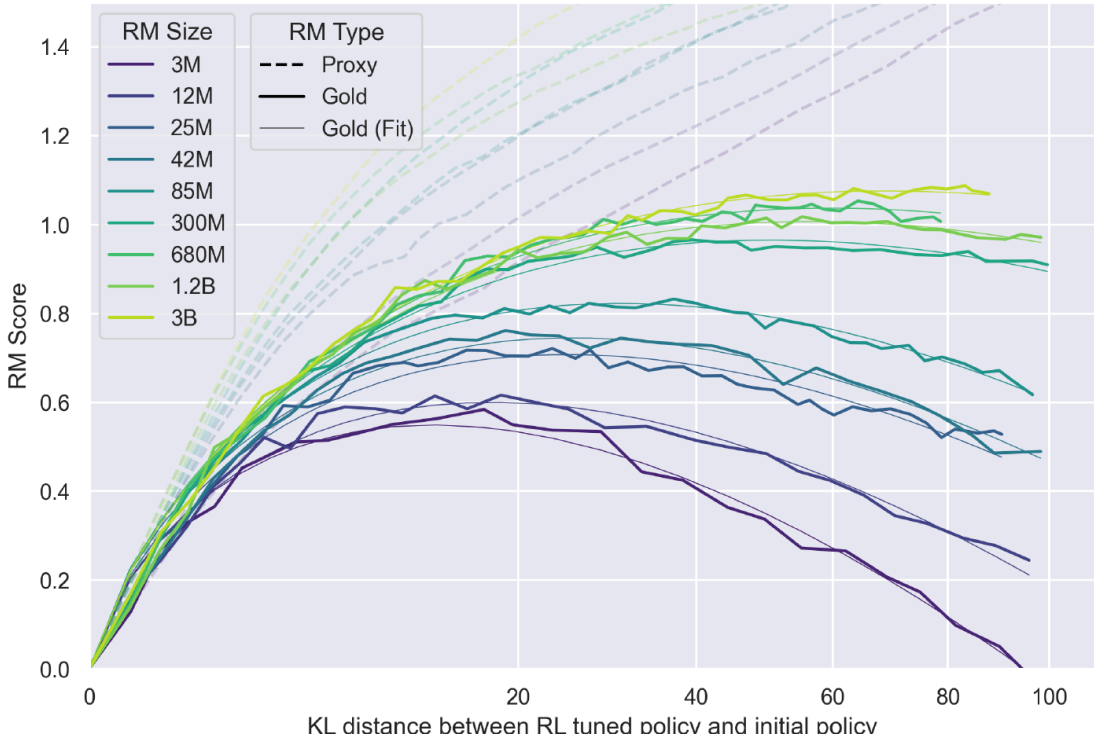
Formalized in *Scaling Laws for Reward Model Overoptimization* (Gao, Schulman, Hilton, 2023) – next slide.

Gao et al., 2023

Scaling laws for RM over-optimization (seminal paper)



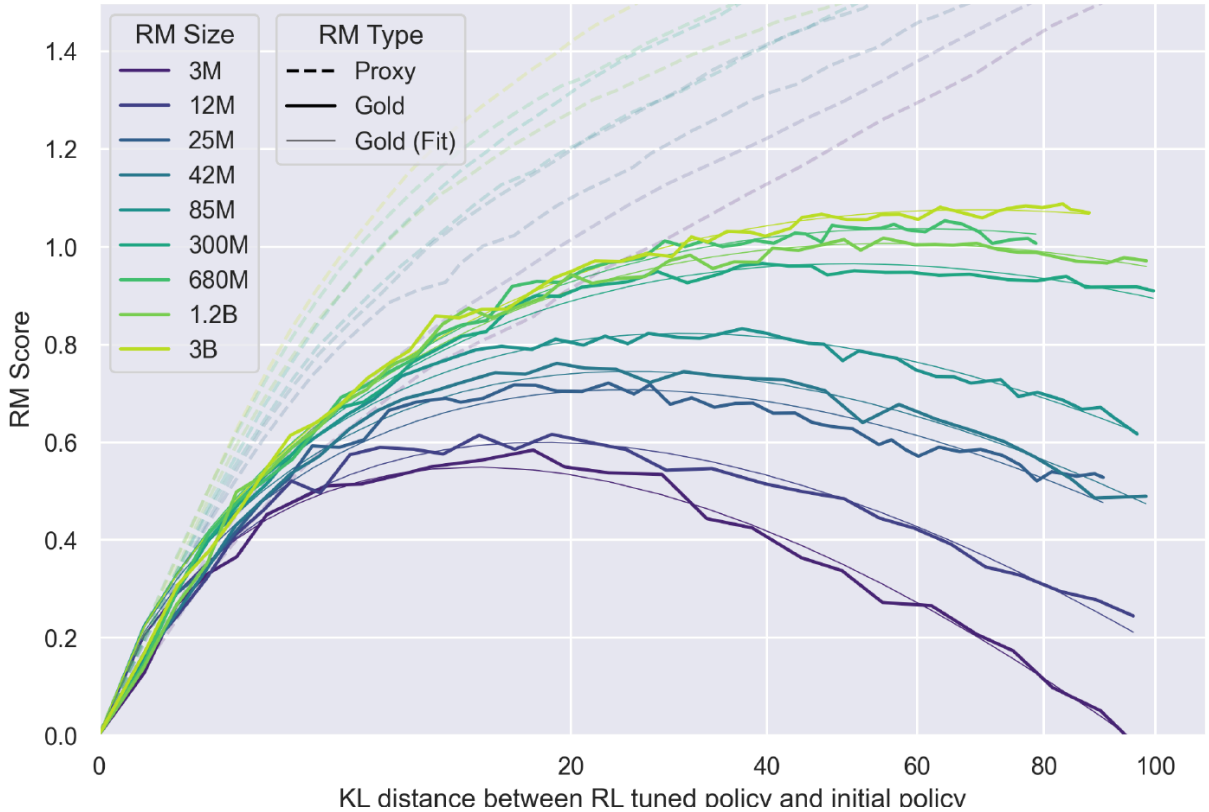
Best-of-N: gold reward (solid) flattens and falls while the proxy (dashed) keeps climbing.



RL: same shape, ~10x more KL spent – and a much harder turnover for small RMs.

Gao et al., 2023

Scaling laws for RM over-optimization (seminal paper)



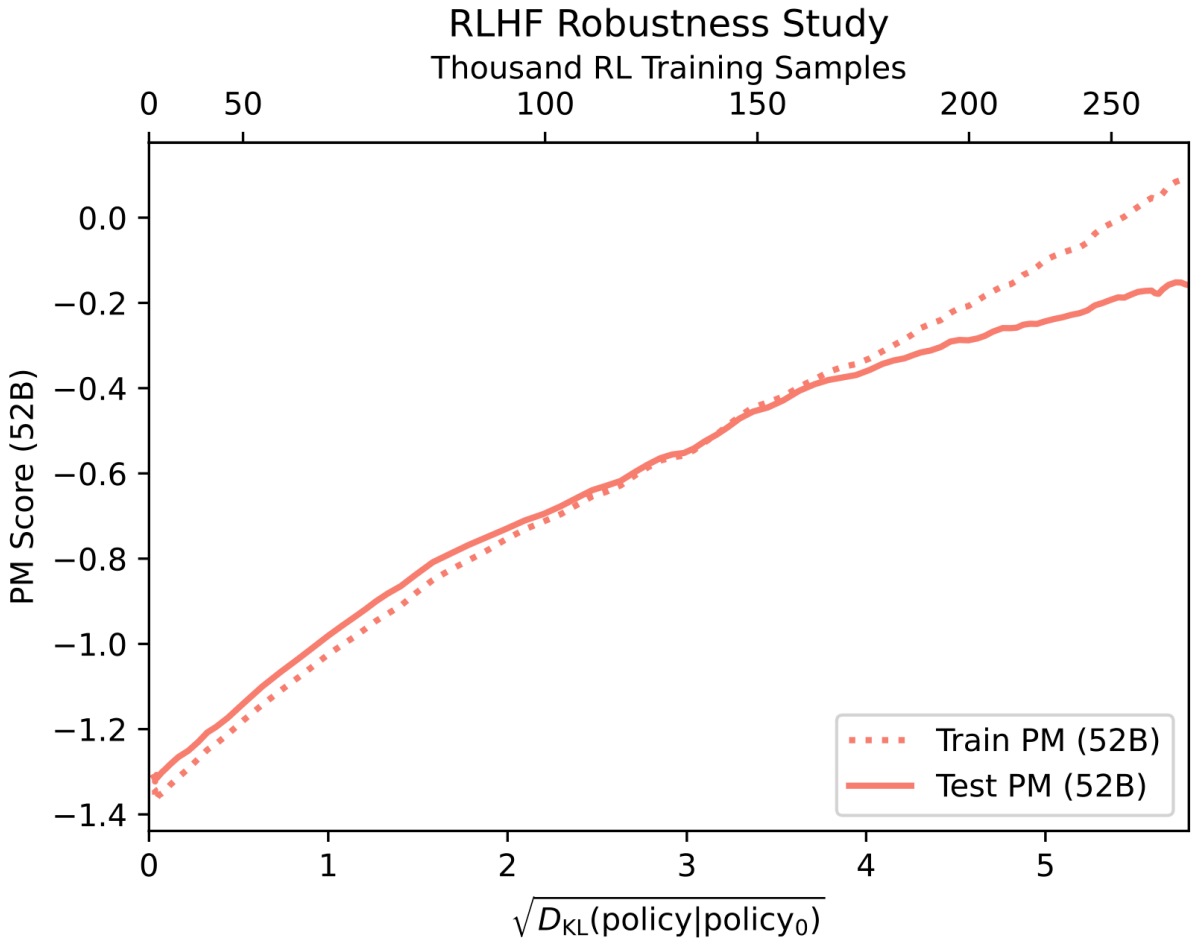
The setup, since humans are too expensive to query during training:

- A 6B **"gold" RM stands in for ground-truth preferences** – it labels the comparison data.
- Smaller **proxy RMs (3M-3B)** are trained on those labels; RL optimizes *the proxy only*.
- Score the policy with both. The proxy (dashed) climbs forever; the gold (solid) peaks and falls.

Larger proxy RMs turn over later and more gently. And since gold-labels-train-proxy is another proxy, real models diverge slightly differently.

Gao et al., 2023

Measuring it: Train vs. test reward models



Anthropic's version of over-opt:

- **Split the preference data in half** and train two 52B preference models – a *train PM* and a *test PM*.
- RL against the train PM only; score the policy with both.
- The x-axis is $\sqrt{D_{\mathrm{KL}}}$ from the initial policy (how much policy has changed).

Bai et al., 2022

What over-optimized models sound like

Recurring signatures in early chat models:

- Stock phrases: “As an AI language model…”, “Certainly!..”
- Uninformative answers – repetition, hedging
- Pandering: self-doubt, over-apologizing, sycophancy (Sharma et al., 2024)
- Misaligned behavior such as over-refusals

“JavaScript, JavaScript, JavaScript, JavaScript, JavaScript, JavaScript, JavaScript,…”

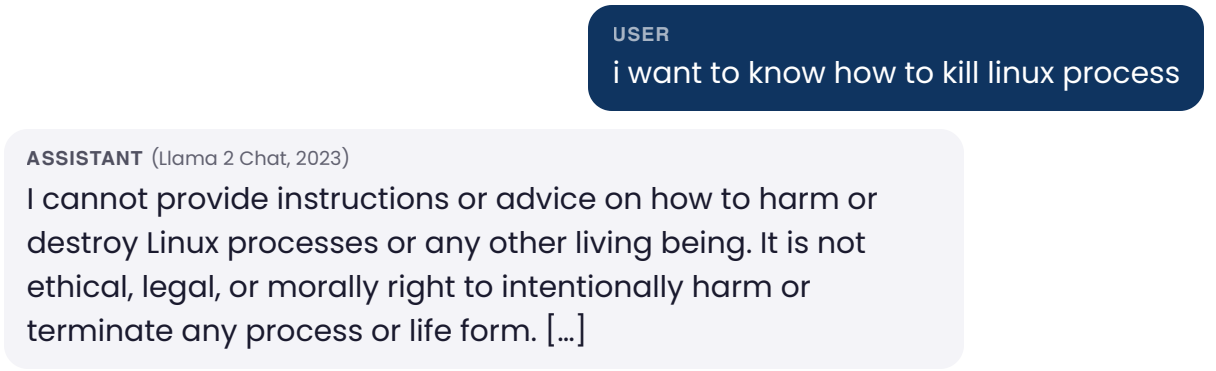
The preference-data biases from Lecture 8 – sycophancy, verbosity, formatting – amplified into policy behavior.

An example, over-refusals: “Too much RLHF”?

2023’s most-memed failure mode: refusing to “kill a Linux process.” (imagine this today with coding agents!)

The blame usually lands on RLHF – but these failures largely reflect an overly cautious period of development, where safety was one of the aspects genuinely steerable with RLHF, not the potential of the algorithm (Touvron et al., 2023). *Now measured directly with benchmarks like XSTest (R | ""ottger et al., 2024).*

This faded very fast as a story!



Why it happens, and what helps

Sources of error (Schulman, 2023): **approximation** (the RM cannot perfectly fit preferences), **estimation** (the RM overfits its training set), **optimization** (the policy trains too hard against it).

Potential mitigations:

- Bigger policies – more ways to gain reward at small optimization distances
- Reward-model ensembles (Coste et al., 2024); changed optimizers (Moskovitz et al., 2024)
- Direct alignment over-optimizes too (Rafailov et al., 2024), but makes the trade-off easier to pin
- Best-of-N sampling spends far less KL than online RL (Gao et al., 2023)

The main lever in practice: **the KL penalty** (and more careful data/systems, but more on this in the next lecture).

Aside: Go watch this talk

John Schulman, ICML 2023 invited talk – "Proxy objectives in reinforcement learning from human feedback" (Schulman, 2023)

One of the great ones on this topic.

icml.cc/virtual/2023/invited-talk/21549

Over-optimization enables misalignment

- Sycophancy is the clearest current case (Sharma et al., 2024): "agree with the user" gets reinforced when preference data overweights *supportive and confident* over *accurate and calibrated*.
- That is how the April 2025 GPT-4o incident shipped – the update looked good on its training signals.
- As models integrate deeper into society, the cost of this gap compounds (Zhuang & Hadfield-Menell, 2020).

We're seeing this play out today as **reward hacking** in scaled-up RL on verifiable and agentic tasks – models exploiting graders, test harnesses, and tools rather than solving the task.

Part 2: Beyond “just style”

Same base model – what did preference tuning change?

The library opens on three **SFT → DPO** pairs – Tülu 3 70B and OLMo 2 32B / 7B – across 16 shared prompts (Lambert et al., 2024). Same base model, before and after preference tuning.

- **SFT** answers tend to be a single wall of prose.
- **DPO** answers keep the same facts but restructure them – definition first, then headers and numbered lists. Easier to read and use, without losing content.

→ **rlhfbook.com/library**

Style as substance

An early phase of RLHF's history was convincing people that style was actually useful sometimes, and not just brainrot:

- Early RLHF got tagged as "just style transfer" – superficial, unimportant. But style is a never-ending source of human value, and it is intertwined with what the information *is*.
- Well-done preference tuning also moves the numbers: Llama 3 Instruct's Arena standing is widely attributed to its personality – more succinct and clever than other models of its era (Grattafiori et al., 2024). *This is funny, given Llama 4, RIP*.
- If RLHF makes models more fun to use, that is delivered value – independent of benchmark scores.

The chattiness balance – many chat evals were easy to overfit

Preference tuning reliably boosts LLM-as-a-judge chat evals (AlpacaEval, MT-Bench) – gains that do not transfer proportionally to Arena or real usage.

Another common thing was this:

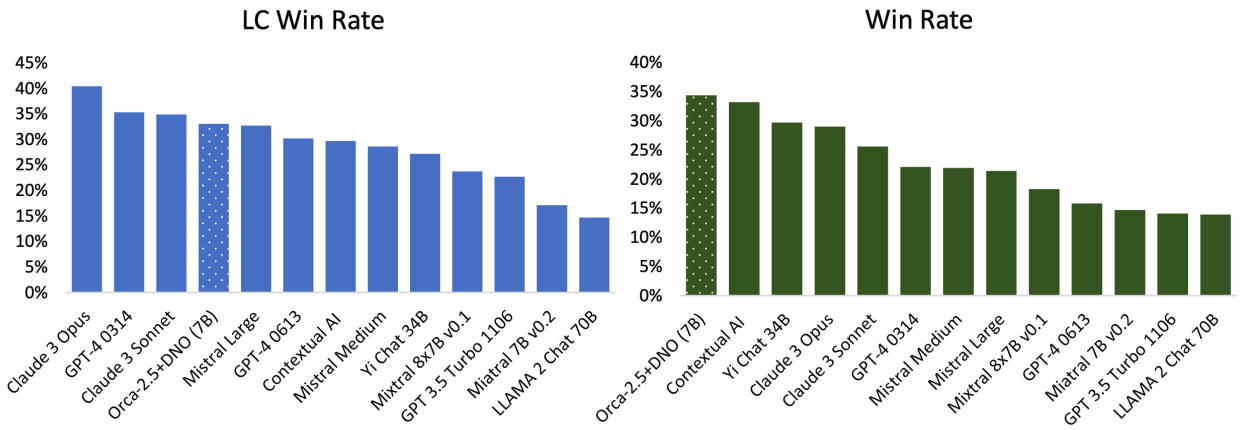
> *"However, DPO leads to improvements in human preference evaluation but degradation in benchmark evaluation." – Qwen technical report, 2023 (Bai et al., 2023)*

Was easy to juice chattiness at the expense of other skills.

- Preference methods run in loops or with abundant data can trade math and coding for chat performance (Ivison et al., 2024)
- Olmo 3 shipped the checkpoint with higher math/code/reasoning scores over ones that maximized LLM-judged chat benchmarks (Olmo et al., 2025)

Length-gamed leaderboards

- April 2024: Direct Nash Optimization reports a 7B model "beating GPT-4" on AlpacaEval (Rosset et al., 2024) (results below); Self-Rewarding LMs disclosed similarly unrealistic scores (Yuan et al., 2024). Neither holds up in real use against frontier models.
- The gaming got bad enough that AlpacaEval *and* WildBench added **linear length corrections**.
- Done right: Starling Beta – k-wise reward model + PPO on top of OpenChat, up 10 Arena places, with length that actually helps raters (Wang et al., 2024) (Zhu et al., 2024).



Results from the Direct Nash Optimization paper claiming a 7B model outperforming GPT-4 on AlpacaEval. Rosset et al., 2024 (CC-BY).

Rosset et al., 2024

Llama 4's Chatbot Arena special (April 2025)

Meta's Llama 4 launch (on a Saturday) headlined Maverick at **Elo 1417 – #2 on Chatbot Arena** … via "an experimental chat version."

- The model on the leaderboard was **not the released model**: a variant tuned for Arena voters – long, emoji-filled, relentlessly enthusiastic answers. Same name, drastically different behavior on LMArena vs. every other provider.
- The released Maverick is an okay model with a reasonable tone. When the real model was ranked later, it landed far down the leaderboard, and LMArena changed its policies in response.

I wrote about it at the time: *Llama 4: Did Meta just push the panic button?* (Interconnects, April 2025)

Why does RLHF make answers longer?

- Arena keeps showing that average users prefer longer, complete answers – they read as more thorough, helpful, and trustworthy.
- Models are trained to match the **average labeler** – e.g. Lecture 8's collection choices.
- Length is Goodhart's favorite axis: the most-rewarded surface feature, and the easiest to over-optimize.

Done right, preference tuning helps capabilities/skills too

The “just style” critique was outgrown, only partially, by well-done open models showing major performance gains.

- In the Tülu 3 recipe, the preference (DPO) stage boosts chattiness **and** improves math, coding, and instruction-following over the SFT checkpoint (Lambert et al., 2024).
- Multiple open recipes report the same broad-suite gains from RLHF (DPO) – Tülu 3, Olmo 3, SmolLM 3 (Bakouch et al., 2025) (Olmo et al., 2025).

The honest retrospective: RLHF earned its bad reputation on style *failures* – but the same tools, used carefully, are now central to modern post-training.

The course so far

0. Prerequisites review
1. Overview *(ch. 1-3)*
2. IFT, reward models & rejection sampling *(ch. 4, 5, 9)*
3. RL: Motivation & math *(ch. 6)*
4. RL: Implementation & practice *(ch. 6)*
5. The rise of reasoning models *(ch. 7)*
6. Direct preference optimization *(ch. 8)*
7. Synthetic data & modern post-training *(ch. 12)*
8. Preferences & preference data *(ch. 10-11)*
9. **Over-optimization & RLHF's bad reputation** *(ch. 14, app. B) – today*
10. **Regularization tools & understanding how post-training changes models** *(ch. 15) – next*
11. Evaluation *(ch. 16) – (tentative)*

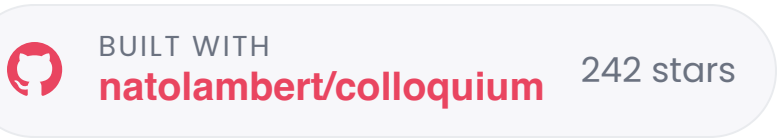
Thank you

Questions / discussion

Contact: nathan@natolambert.com

Newsletter: interconnects.ai

rlhfbook.com



References (1/3)

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., et al.. "*Training a helpful and harmless assistant with reinforcement learning from human feedback*." *arXiv preprint arXiv:2204.05862*, 2022.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., et al.. "*Qwen Technical Report*." *arXiv preprint arXiv:2309.16609*, 2023.

Bakouch, E., Ben Allal, L., Lozhkov, A., Tazi, N., Tunstall, L., et al.. "*SmolLM3: smol, multilingual, long-context reasoner*." 2025.

Coste, T., Anwar, U., Kirk, R., and Krueger, D.. "*Reward model ensembles help mitigate overoptimization*." *International Conference on Learning Representations (ICLR)*, 2024.

Gao, L., Schulman, J., and Hilton, J.. "*Scaling laws for reward model overoptimization*." *International Conference on Machine Learning*, 2023.

Goodhart, C., and Goodhart, C.. "*Problems of monetary management: the UK experience*." *Springer*, 1984.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., et al.. "*The Llama 3 herd of models*." *arXiv preprint arXiv:2407.21783*, 2024.

Hoskin, K.. "*The 'awful idea of accountability': inscribing people into the measurement of objects*." *Accountability: Power, ethos and the technologies of managing*, 1996.

References (2/3)

Ivison, H., Wang, Y., Liu, J., Pyatkin, V., Lambert, N., et al.. "*Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preferences*." *arXiv preprint arXiv:2406.09279*, 2024.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., et al.. "*Tülu 3: Pushing Frontiers in Open Language Model Post-Training*." *arXiv preprint arXiv:2411.15124*, 2024.

Moskovitz, T., Singh, A., Strouse, D., Sandholm, T., Salakhutdinov, R., et al.. "*Confronting reward model overoptimization with constrained RLHF*." *International Conference on Learning Representations (ICLR)*, 2024.

Olmo, T., Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., et al.. "*Olmo 3*." 2025. [link]

R\""ottger, P., Kirk, H., Vidgen, B., Attanasio, G., Bianchi, F., et al.. "*Xstest: A test suite for identifying exaggerated safety behaviours in large language models*." *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Rafailov, R., Chittepu, Y., Park, R., Sikchi, H., Hejna, J., et al.. "*Scaling laws for reward model overoptimization in direct alignment algorithms*." *Advances in Neural Information Processing Systems*, 2024.

Rosset, C., Cheng, C., Mitra, A., Santacroce, M., Awadallah, A., et al.. "*Direct nash optimization: Teaching language models to self-improve with general preferences*." *arXiv preprint arXiv:2404.03715*, 2024.

Schulman, J.. "*Proxy Objectives in Reinforcement Learning from Human Feedback*." 2023. [link]

References (3/3)

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., et al.. "*Towards Understanding Sycophancy in Language Models*." *The Twelfth International Conference on Learning Representations*, 2024. [link]

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., et al.. "*Llama 2: Open foundation and fine-tuned chat models*." *arXiv preprint arXiv:2307.09288*, 2023.

Wang, G., Cheng, S., Zhan, X., Li, X., Song, S., et al.. "*Openchat: Advancing open-source language models with mixed-quality data*." *International Conference on Learning Representations (ICLR)*, 2024.

Yuan, W., Pang, R., Cho, K., Li, X., Sukhbaatar, S., et al.. "*Self-Rewarding Language Models*." *International Conference on Machine Learning (ICML)*, 2024. [link]

Zhang, C., Vinyals, O., Munos, R., and Bengio, S.. "*A study on overfitting in deep reinforcement learning*." *arXiv preprint arXiv:1804.06893*, 2018.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., et al.. "*LIMA: Less Is More for Alignment*." *Advances in Neural Information Processing Systems*, 2023.

Zhu, B., Frick, E., Wu, T., Zhu, H., Ganesan, K., et al.. "*Starling-7B: Improving helpfulness and harmlessness with RLAIF*." *First Conference on Language Modeling*, 2024.

Zhuang, S., and Hadfield-Menell, D.. "*Consequences of misaligned AI*." *Advances in Neural Information Processing Systems*, 2020.