Data Labeling and Compute: Rethinking RL for the Next AI Push

Benjamin Shvartsman

July 17th, 2025

Table of Contents

1. TL;DR 2
2. The State of RL Scaling after Grok 2
3. Why RL Delivers Diminishing Reasoning Returns 3
4. The Data‑Labeling Bottleneck 3
5. Beyond Next‑Token Prediction: Hybrid Training Approaches 5
6. Industry Momentum & Forward-Deployed Engineering 6
7. Conclusion 7
8. Sources 7

1: TL;DR

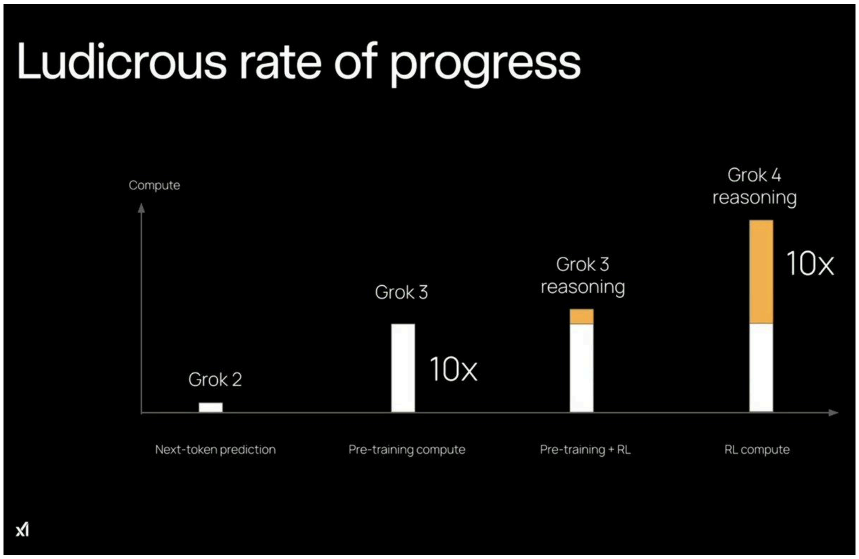
Scaling RLHF is approaching a ceiling. Reward-model inference and KV-cache traffic dominate training costs, with brute-force scaling providing limited benefits. xAI's **Grok 4** utilized roughly **10x** more RLHF compute than Grok 3 while only squeaking out single-point improvements on reasoning benchmarks. However, *Heavy* mode's multi-agent orchestration and tool-usage doubled accuracy on hard reasoning tasks without another order-of-magnitude PPO run. It's clear that future gains must come from **hybrid training approaches** and **data innovations** (programmatic labeling, synthetic feedback, and influence-guided sampling).

This plateau creates opportunity!

1. Platforms that deliver effective label and reward systems are extremely valuable.
2. Compute scaling alone is stalling; smarter data and orchestration are the unlock.
3. Label quality, cost, and throughput are the new bottlenecks.
4. Progress requires hybrid objectives, efficient feedback loops, and continued success with tool-usage and orchestration.

Disclaimer: Grok 4 (especially *Heavy*) is genuinely impressive and definitely showcases top deep reasoning abilities.

2: The State of RL Scaling after Grok 4


Ludicrous rate of progress

Visual from medium.com article [4]

xAI trained Grok 4's RLHF phase on Colossus ~ a 200,000-GPU cluster ~ consuming 10x more GPU-days than Grok 3.[1] Yet, single-agent Grok 4 gained only +1.4 percentage points on MMLU and +0.7 gain on GSM-8k. xAI gave credit to **tool-use pre-training** and **context window** for the major improvements across complex reasoning tasks, multi-agent collaboration, training compute usage, and tool usage.[2] **Grok 4 Heavy truly shines** through its ability to spawn multiple agents that debate, consult tools, and merge answers.

On Humanity's Last Exam, both Grok 4 (scoring 25.4% w/o tools) and Grok 4 Heavy (scoring 44.4% with tools) have outperformed other frontier models.[3] This success reinforces what additional compute, parallelism, and orchestration can deliver, even when raw RLHF scaling plateaus. For me, this validates the argument for pro multi-agent approaches as a cost-effective alternative to endless PPO scaling. Behind the scenes, reward-model inference monopolises GPU minutes with KV-cache shuffling consuming 45% of the dollar budget, making RLHF an inference-heavy, I/O-bound workload. The contrast between single-agent gains and Heavy's leap hints at an implicit acknowledgement that **most of the extra RL compute was chasing diminishing returns**.[4]

3: Why RL Delivers Diminishing Reasoning Returns

Reinforcement Learning with Human Feedback is powerful for **aligning** language models with human preferences and skyrocketing *reliability*. Its gains in deep reasoning and creative problem-solving are ultimately bounded by **capacity of the base model** and the effectiveness/efficiency of algorithms extracting its abilities.[5, 9] My passion and interest in the healthcare domain makes alignment especially important; with an emphasis on technology **not risking patient safety or introducing bias**. As a member of the AI/ML team at an ecommerce technology company in the hunting/military/outdoors sector, we were operating in highly regulated environments. **Proactive alignment** was our way to avoid costly mistakes, while the RLHF technology is becoming more and more accessible to enterprises.

RLHF appears to **surface abilities** that the base model already possesses **rather than create** new ones.[5] In domains/tasks like math or coding, PPO (policy-gradient updates) raises "success-at-1" metrics but "best-of-k" ceiling remains unchanged. It improves at surfacing the base model's best ideas reliably but does not expand the actual range of what it can achieve.[6] I am super interested in learning about evaluation of "success-at-1"/"best-of-k" as well as Pass@K and Maj@K metrics.

"Length-entropy bias" **trades off response diversity for alignment** with reviewer preferences, which often results in long, detailed, or unusual answers being penalized.[7] This impact is particularly visible during complex reasoning tasks, where efforts to increase safety sometimes undermine the depth of chain-of-thought required for advanced problem-solving. As a result, RLHF can **inadvertently reduce performance on reasoning tasks** it aims to improve and answers become shorter without necessarily improving correctness.[7]

The **cost/benefit ratio** of RLHF **worsens as model context windows are extended**, as each 1000 tokens added can **require multiplied reward model invocations** per training episode. This places pressure on both compute and throughput.[8] This extra inference and RM evaluation **doesn't contribute proportionally to improved task learning**; thus, increasing model capability by elongating the context shows diminishing returns.[9] However, after looking further at this challenge, newer reward models often include explicit length regularization, this way the bias is task-dependent rather than universal.[7] This is definitely tricky to navigate because larger context windows have so many benefits, even in my own life: extended conversations, codebase analysis, cross-document tasks, and more.

4: The Data-Labeling Bottleneck

Modern alignment research agrees that model improvements are less limited by compute than by **availability, speed, and cost of high-quality human feedback data** (after achieving a certain scale). Obtaining enough accurate, domain-sensitive labels becomes the clear barrier.[10, 15]

Reward modeling with expert human annotators is **demanding**, with costs ranging between $0.75 and $2.00 per pairwise comparison.[14] Even at the lower end, crowd-sourced work rarely falls below five cents per label due to minimal quality standards required.[13] If large-scale model training demands anywhere from 100 million to 300 million comparisons per run, human-involved per-label costs range from $0.01 - $2.00, which translates to a direct cost of $1 to $600 million before even accounting for compute resources. In healthcare, I'd imagine that the *quality-per-label* bar is even higher and thus pushing costs to the upper bound.

Vendors and research teams have developed a range of approaches to labeling, each seeking to optimize the economic/throughput trade-offs:

Approach	**Technique**	**$$ / Pair**	**Throughput (approx)**	**Strength**	**Weakness**
Expert Pairwise	Senior Annotators	$0.75–$2	10k / day	Policy, Safety	Expensive, Slow
Premium Crowd	Trained Crowd Workers	$0.25–$0.50	1M / day	Consistent, Multi-Domain	Costly at Scale
Gamified Crowd	Micro-task, Leaderboards	$0.05–$0.15	≥5M / day	Cheaper	Noise, Requires QA
Weak Supervision	Labeling functions	$0.01–$0.05	Hours	Rapid, Any Domain	Brittle
Synthetic Feedback	Model-verified outputs	<$0.001	≥100M / day	Objective, Scale	Reward bias, verifier limits

Perplexity generated

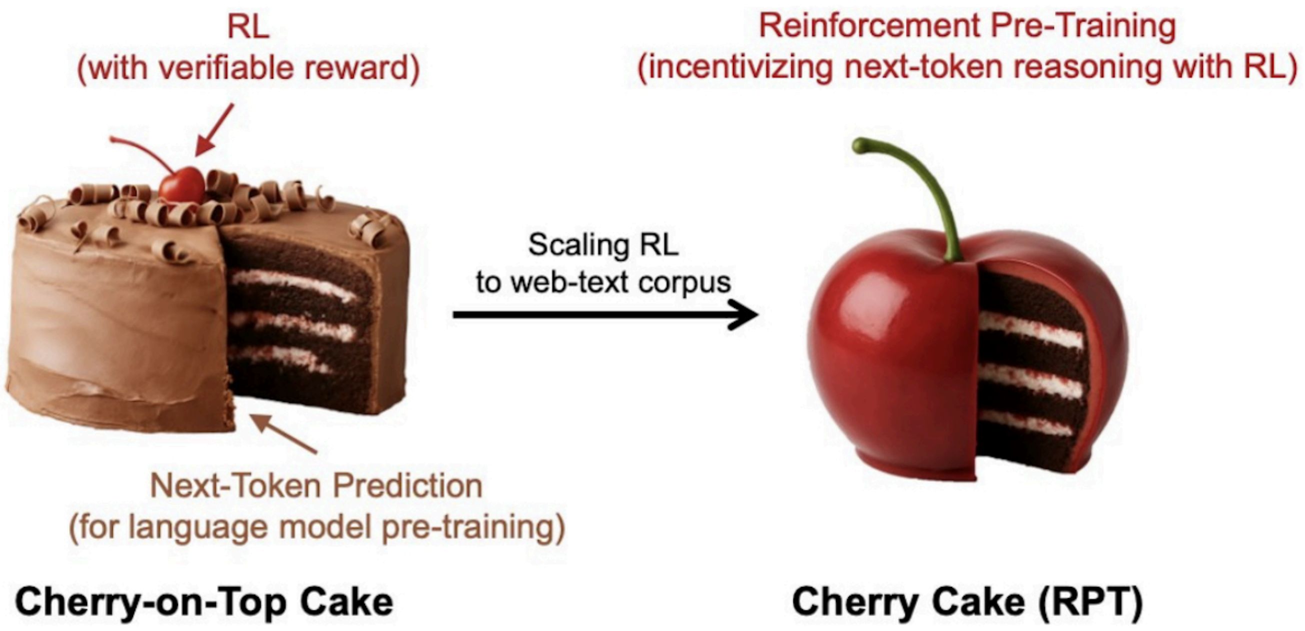
Surge AI[14]	Premium rubric-based workforce, *Chain-of-Thought Critic* workflows that reduce hallucinations in output by 22% as of recent
Scale AI[11]	Automated QA combined with human feedback, *RLHF for Text2SQL* 12% jump in execution accuracy through hybrid human-verifier labeling.
Snorkel AI[12]	Label-functions and influence-guided sampling, in previous projects have generated 18 million preference pairs in under 48 hours
Labelbox[13]	Reinforcement Learning with Verifiable Rewards, automating scoring with unit tests and theorem provers

Sapien	Gamified and reputation-based labeling, $0.06 per comparison

Perplexity generated

From a perplexity-guided exploration on research in this area, it highlights several approaches to overcome the data-labeling bottleneck that limits AI alignment: **influence-guided sampling** prioritizes labeling only data points that most improve the reward model (reducing expert time ~ 3x)[12], **active self-play** lets multiple policies debate uncertain cases before human reviewers intervene, **auto-verifier generation** uses LLMs to write unit tests enabling full synthetic feedback loops, and more.[11, 15] From a practical standpoint, **focusing on information gain per dollar** instead of labels per dollar should optimize annotation strategies by targeting high-impact data.[10] Being a student, I have limited industry experience and hands-on exposure with labeling functions and evaluators, but it's cool to see strategic data-centric solutions take on the challenge of scaling alignment labeling.

5: Beyond Next-Token Prediction: Hybrid Training Approaches



Visual from Reinforcement Pre-Training [16]

As LLMs approach the **limits of traditional next-token prediction**, researchers are redefining training standards to **overcome hallucinations, shallow reasoning, and alignment challenges**.[16] (off-topic but curious about / excited for how joint-embedding architectures and energy-based models deal with challenges mentioned) One promising idea is **Reinforcement Pre-Training** that integrates **chain-of-thought reasoning into the training objective** by rewarding the reasoning process rather than solely the final output. Microsoft's *rStar-Math* prototype showcases a 50% reduction in hallucinations at constant model size (incentivizing intermediate reasoning steps).[16]

Direct Preference Optimization complements this by streamlining the costly RLHF process by eliminating online reward-model evaluations to transform the problem into an offline optimization task.[16] This achieves alignment gains comparable to traditional PPO-based methods with significantly less computational cost. It's now used by Hugging Face's TRL suite!

Addressing the scarcity/cost of human feedback data, **influence-guided sampling** selectively identifies prompts expected to exert the greatest impact on reward model training. Snorkel AI exclaimed a 3x reduction in expert labeling hours without compromising model quality.[12] Similarly, **Reward-Verifier Reinforcement Learning** replaces subjective human evaluations with objective, programmatic ones enabling millions of synthetic high-quality evaluations and reducing reliance on costly human annotations.[11]

Other efforts include using the **open web as an environment** where agents can learn retrieval, tool use, and long-term planning within a unified RL framework. These methods promise to integrate real-world knowledge acquisition with interactive capabilities, allowing AI systems to be even more autonomous and contextually aware. An area I would like to explore in depth is how RL approach to tool-usage compares to the existing training approaches like labeled examples + RLHF by OpenAI and xAI (exposing LM to objectives where it must choose when and how to invoke a tool).

Collectively, these approaches work closely with feedback loops by leveraging abundant next-token prediction data to generate/prioritize high-value preference signals.[10] Focused usage of RL compute ensures resources target the most impactful gains for alignment and reasoning. These developments truly display **dynamic, objective-driven model optimization** and have the potential to **overcome the current barriers to scaling accuracy and utility.**

6: Industry Momentum & Forward-Deployed Engineering

The last 12 months have marked a big shift in how leading labs and enterprises approach alignment. From OpenAI's o-series to xAI's Grok 4, every recent flagship model release has included detailed information on **label quality** and **reward-model ceilings**. Venture funding has quadrupled YoY investment in alignment tooling (PitchBook, Q2 2025). This acts as collective recognition that better alignment and data infrastructures are moats for scaling safe/effective horizontal AI applications. Enterprises are racing to aggregate trustworthy preference pairs.

Forward-deployed software engineers essentially **embed a small ML + domain-specific team inside of customer environments**. They turn business heuristics into programmatic labels and verifiable rewards. Small, high-agency teams bridge gap between ML research and messy, real-world use cases. Crafting scalable *SurgeAI-style* (coined my own term there) rubrics, wrapping domain code with Labelbox RLVR, style-guides into Snorkel labeling-functions, and *Customer-embedded ML Ops* from Scale AI are all vendor-loved examples of this role's work combined with strong product offerings. FD alignment is becoming a sought-after seat in the GenAI stack and I am excited to pursue exactly that!

Working at the **intersection of ML research**, **domain expertise** (understanding customer goals + discerning needs), and **product impact** is compelling because it gives a pathway to influence model behavior without 100,000+ GPU-hours. In previous roles I have navigated real-world constraints and customer value, but tying that together with model internals and evolving alignment research would be an amazing opportunity.

7: Conclusion

The Grok 4 case study truly highlights that we have nearly maxed out the potential gains from brute-force RLHF. When multiplying GPU budgets by 10 delivers one-point benchmark improvements (without smart tool chains or mutli-agent interactions), that influences the approach towards progress. It's clear that **information-value-per-dollar** and ingenious **feedback loops** determine progress instead of raw GPUs.

My prediction for what comes next will be grounded in data-centric alignment:

1. **Preference data value will continue rising.** Programmatic heuristics, influence-guided sampling are growth engines for scalable progress.
2. **Objectives will be introduced earlier.** RPT and DPO place alignment into pre-training reducing reward-model costs.
3. **Orchestration and smart tool-usage at inference produces the best results**. Nothing novel or nontrivial, but Grok 4 single-agent -> Grok 4 *Heavy* is same PPO but +20 percentage points on hardest benchmarks.
4. **Talent will be deployed at the customer edge.** Forward-deployed alignment engineers that convert domain knowledge into verifiable rewards and turning months of labeling into days will be a competitive moat in itself.

The teams that **excel at data-centric alignment, hybrid objectives, and real-time orchestration** will define the next generation of trustworthy, deep-reasoning AI. Not by outspending on silicon, but by investing in labels, logic, and learning loops. I'm very bullish on reasoning capabilities of these flagship models and excited to see the real-world applications!

8: Sources

Research for this report was primarily facilitated through Perplexity's Deep Research!

1. xAI. July 9th, 2025. *Grok 4 release notes & model card*. https://x.ai/news/grok-4
2. Zeff, M. July 9th, 2025. *Elon Musk's xAI launches Grok 4 alongside a $300 monthly subscription*. TechCrunch. https://techcrunch.com/2025/07/09/elon-musks-xai-launches-grok-4-alongside-a-300-monthly-subscription/
3. Scientific American. July 11th, 2025. *Elon Musk's new Grok 4 takes on "Humanity's Last Exam" as the AI race heats up*. https://www.scientificamerican.com/article/elon-musks-new-grok-4-takes-on-humanitys-last-exam-as-the-ai-race-heats-up/

4. Code Elevation. July 12th, 2025. *Grok 4: Musk's AI smarter than PhDs, crushes "Humanity's Last Exam" with 58 % score.* Medium. https://medium.com/codeelevation/grok-4-musks-ai-smarter-than-phds-crushes-humanity-s-last-exam-with-58-score-0938c7e4ddc5
5. Ouyang, L., Wu, J., Jiang, X. March 4th, 2022. *Training Language Models to Follow Instructions with Human Feedback (InstructGPT)*. arXiv 2203.02155. https://arxiv.org/abs/2203.02155
6. Chen, M., Tworek, J., Jun, H. July 14th, 2021. *Evaluating Large Language Models Trained on Code*. arXiv 2107.03374. https://arxiv.org/abs/2107.03374
7. Zhao, K., Cai, J., Zhu, J. May 19th, 2025. *Bias Fitting to Mitigate Length Bias of Reward Model in RLHF (FiMi-RM)*. arXiv 2505.12843. https://arxiv.org/abs/2505.12843
8. Fan, T., Liu, L., Yue, Y. June 18th, 2025. *Truncated Proximal Policy Optimization (T-PPO)*. arXiv 2506.15050. https://arxiv.org/abs/2506.15050
9. Chowdhury, H. & Nolan, B. Nov 11th, 2024. *OpenAI is reportedly struggling to improve its next big AI model. It's a warning for the entire AI industry*. Business Insider. https://www.businessinsider.com/openai-orion-model-scaling-law-silicon-valley-chatgpt-2024-11
10. Leo Gao, John Schulman, Jacob Hilton. October 19th, 2022. *Scaling laws for reward model overoptimization*. OpenAI. https://openai.com/index/scaling-laws-for-reward-model-overoptimization/
11. Shubhashis Roy Dipta, Vijay Kalmath, Hans Husmann. January 23, 2025. *When RLHF Meets Text2SQL*. Scale AI. https://scale.com/blog/rlhf-text2sql
12. Snorkel AI. *Label and annotate data for training AI/ML models up to 100x faster.* https://snorkel.ai/data-labeling-and-data-annotation/
13. Labelbox. September 11th, 2024. *Inside the data factory: How Labelbox produces the highest quality data at scale*. https://labelbox.com/blog/inside-the-data-factory-how-labelbox-produces-the-highest-quality-data-at-scale/
14. Surge AI. *Surge AI Blog*. https://www.surgehq.ai/blog
15. Ji, J., Qiu, T., & Others. April 4th, 2025. *AI alignment: A comprehensive survey*. arXiv. https://doi.org/10.48550/arXiv.2310.19852
16. Dong, Q., Dong, L., Tang, Y., Ye, T., Sun, Y., Sui, Z., & Wei, F. July 12th, 2024. *Visual from Reinforcement Pre-Training*. arXiv. https://arxiv.org/abs/2404.14035