RLHF lit. review #1 and missing pieces in RLHF

Looking at the difference between two sets -- what rumors say industry leaders are doing with RLHF and what the literature is up to. I'm starting my new series studying RLHF literature.

NATHAN LAMBERT
OCT 25, 2023

There is finally the acceleration of fundamental research into reinforcement learning from human feedback (RLHF) that I have been expecting since early this year. RLHF has only become increasingly central to players across the industry. Some companies do not make this super clear, as Anthropic has been touting Constitutional AI more than RLHF, but that is reliant on a model trained with something like RLHF or previously skeptical companies like Cohere are now trying super hard to hire engineers to actually build this stuff out.

The important thing that I am trying to do is to map how research trends match (or don't) trends I'm hearing on the ground from the companies with the most sophisticated uses of RLHF. As I start my new job in more of an open-ended research space next week, filling the gaps to get the research understanding to parity is my goal.

This post has two sections -- first, I'll review the big-picture trends that make RLHF, and particularly the RL part, positioned on an unstable foundation relative to other parts of the GenAI boon. Second, I'll review some recent trends in the literature and where themes of work are emerging. **This is post number one in an intermittent series reviewing the RLHF literature**.

Open questions for the RL part of RLHF

This section is inspired by RL's harshest critics. The argument implicitly compares reinforcement learning to the other recent successes in generative AI that are empowered by incredibly high upfront computation and data curation,

which RL is not designed to do. This dichotomy is true: the way RL is used now for LLMs uses much less data and much less compute than the rest of the process. There are two very likely futures for the RL part of the current RLHF framework (both actually can be true):

1. The RLHF work done now is integrated into what is historically called pre-training, likely without anything that looks like a classic RL algorithm, but relying on ideas from RL literature like Direct Preference Optimization.
2. An entirely new use of RL emerges to fine-tune models with respect to user feedback (hint, implicit feedback) that enables continual learning of the models.

I see the field getting there by realizing certain things in the current popular RLHF frameworks are not really super logical from first principles. The two issues I'll focus on in this mini-essay are **exploration** and **computation** (per data point and overall). Exploration is due to its incontrovertible role in the RL literature and computation is due to making sure RL methods don't fall behind scaling trends across ML broadly.

Generally, exploration can be simplified as autonomous data collection. The ability to collect new data is central to the algorithms solving *new tasks* outside of the distribution of initial states. In RLHF, this is really not done. RLHF is about interweaving three distributions of data: the input prompts, the reward model training set, and the capabilities of the policy model. Nowhere in the process is it explicitly incentivized to come up with new pieces of data (except for negative reasons, to exploit the reward model, but that is an aside). In my mind, new data revolves around new capabilities.

Structurally, it can be seen how RLHF doesn't explore in a meaningful way because each completion is aggregated as 1 action. There are no multi-action trajectories where the RL algorithm could start to learn long-term credit assignments from new behaviors. It's too direct.

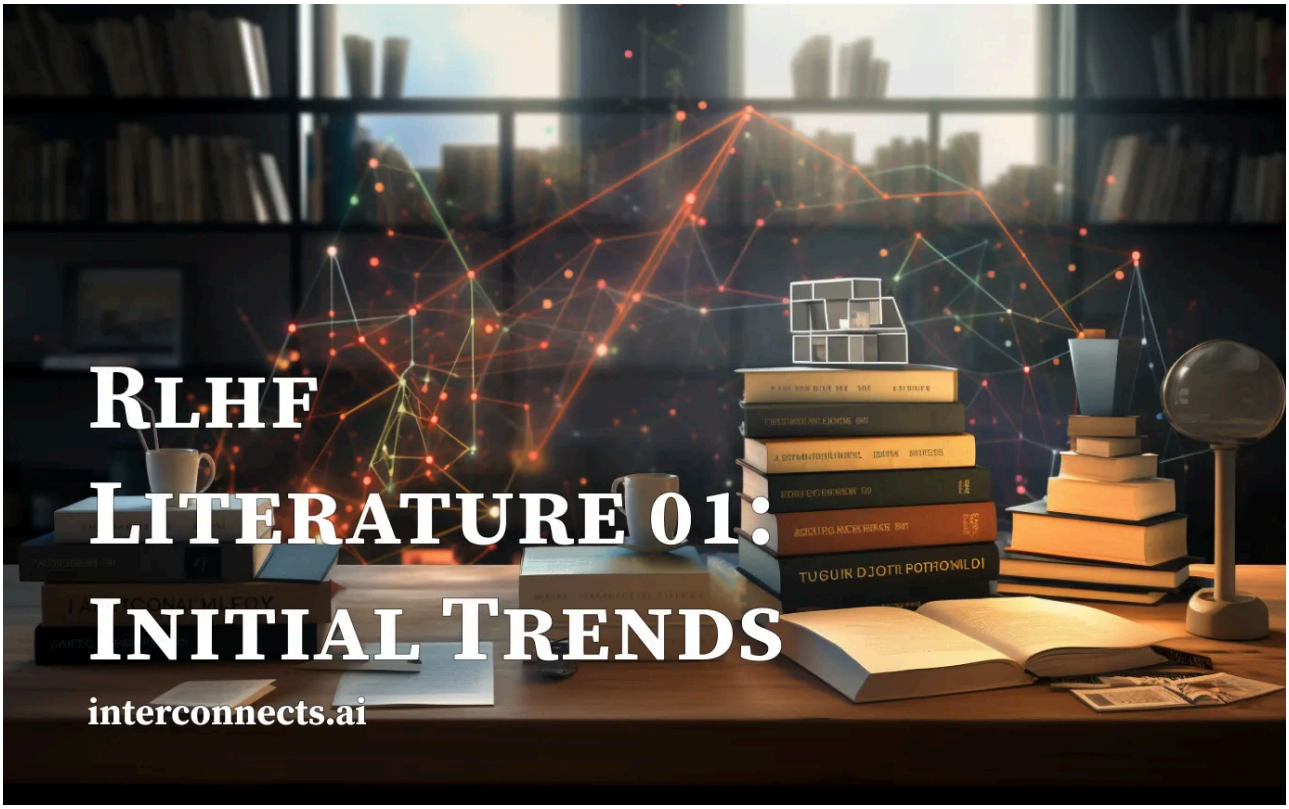
Now, it may not be surprising to experts in RL to hear that OpenAI, and likely other big technology companies, are heavily relying on offline reinforcement learning to optimize their RLHF models. Offline RL (ORL), quickly, is the subfield of RL dedicated to learning a policy only from existing data. This makes sense

for what we are doing with RLHF now, as we are just trying to make sense of the data we have carefully curated and want to instill in our models. As you can see implicitly in the rest of this post, ORL is not really a focus of RLHF research in the last few months (two cool algorithms, ILQL and A-LoL, where the former is a tad older), so I hope that this changes and we learn more about it.

A more subtle reason why ORL may be better suited in this case than something like Proximal Policy Optimization (PPO) is that PPO is an on-policy algorithm. This means that during the RLHF training process that we encounter most frequently, we update the model directly after a batch of data and then don't look at that data again (depending on some exact details). Compare this to off-policy algorithms, which learn from all recent experience and generate new data, or ORL algorithms, which only learn from past experience, on-policy RL seems to be throwing away a lot of information. Information density, which should be measured in the number of tokens passed through the model (with some conversion of how efficient RLHF tokens are vs. pretraining tokens). This is where RL is seen as being "computationally ineffective". It is not because it doesn't use information well, but rather it is not set up to use more information.

The relationship between ORL and exploration is pretty similar to that of RLHF and exploration. In a paper I wrote during my DeepMind internship, we showed that having exploration data in your training distribution is the right way to think about it, rather than during policy training. Right now, the exploration is by paying people to create data. That is not an elegant process.

Circling back, point two above is where exploration and RL can come back into the fold. This would involve training RL algorithms to act over multi-turn conversations with a human (or another language model, which seems more likely). Right now, this is not done at all due to how hard the infrastructure is -- throughput drops to almost 0 if humans are needed in the loop of training, and synchronizing LLM conversations I hinted at, is hard work that only Anthropic has demonstrated at the most elite level. Capturing this data, and learning how to optimize across multiple actions, is where the true intellectual spirit of RL could shine through. As RL fades as an optimizer for basic pairwise feedback, its view in temporal reasoning will continue to grow. This sort of capability, solving a previously unknown task, would be the gold standard for LLMs, so people will keep RL around at least as a motivating example.



midjourney + figma

RLHF literature trends

There are a lot of papers coming out referencing RLHF. For example, the International Conference on Learning Representations had 75 submissions that mentioned RLHF (can search here). It is no longer a worthwhile exercise to try and read every paper, but rather understand where there is a concentration of focus, where there is consensus, and where there is a blindspot.

This is really the first try at a post where I summarize in detail recent RLHF papers, but this one is higher level saying what is here. The format will be zooming in on any one of these sections I break out below. Again, the papers I've noted and sourced this from are here (get in touch if you think I missed one).

Ultimately, it is as important as ever to get familiar with some corners of the ML literature because AI is fundamentally research-driven and not dominated by public companies with clear disclosure. This research will matter in 1 to 18 months.

New training optimizers / methods

People are continuing to try and figure out the right optimization lens for RLHF methods. To me, this all feels like the very early days of a field.

- **Extending Direct Preference Optimization:**
  - *Contrastive Preference Learning: Learning from Human Feedback without RL* https://arxiv.org/abs/2310.13639
    This paper formulates a general objective in robotics experiments, where Direct Preference Optimization (DPO) is a special case. Studying this objective in robotics will be important to understanding why DPO works in LLMs and extending it.
  - *Statistical Rejection Sampling Improves Preference Optimization* https://arxiv.org/abs/2309.06657
    Adds rejection sampling the data-loading for DPO training to improve training stability. This paper's results primarily are on openly available datasets, so it should be validated or understood better soon. I need to spend more time with it.
- **Stabilizing PPO / online RLHF:**
  - *Stabilizing RLHF through Advantage Model and Selective Rehearsal* https://arxiv.org/abs/2309.10202
    This paper has a lot of ablation studies on RLHF. The most striking one to me is how much more regularized the scores provided by an advantage model are compared to the standard reward models used in RLHF (mean closer to 0 and smaller variance). They also propose a method for sourcing better data to the PPO algorithm, which is important considering the first section of this article.
  - *Pairwise Proximal Policy Optimization: Harnessing Relative Feedback for LLM Alignment* https://arxiv.org/abs/2310.00212 / BAIR Blog
    Through some fancy math, the authors rewrite policy gradients with respect to contextual bandits and show you can create a loss function

that directly compares the reward of two texts (the standard derivation needs the score of each, rather than the difference). The evaluation is pretty limited for now but shows that the new method outperforms PPO and DPO in a reward-per-KL measure and on GPT4 ratings.

AI feedback

The RLAIF space continues to be very noisy. People are using AI feedback in lots of different ways, which isn't surprising to me, but there isn't really a central theme. This area will keep growing.

- *SALMON: Self-Alignment with Principle-Following Reward Models*https://arxiv.org/abs/2310.05910
  Kind of a dense paper, but they propose adding principles to reward models (essentially conditioning reward models on a principle from CAI), but it is included because of the strong model release (on Huggingface).
- *UltraFeedback: Boosting Language Models with High-quality Feedback* https://arxiv.org/abs/2310.01377
  A dataset used to train the Zephyr models from HuggingFace, that seems very high quality and uses AI to generate preference data. Expect more models with this.
- *Motif: Intrinsic Motivation from Artificial Intelligence Feedback* https://arxiv.org/abs/2310.00166 / Twitter
  A cool paper using AI feedback to create goals for RL agents to explore in games. Not really relevant to improving LLMs yet -- it just uses LLMs.
- *RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback* https://arxiv.org/abs/2309.00267
  This paper with an odd and catchy title from Google did some basic scaling experiments confirming RLAIF does improve models. It didn't push the envelope.
- *WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct* https://arxiv.org/abs/2308.09583
  This model uses something called Process Reward Models (PRM) that optimize the reasonings from chain-of-thought prompting with RLHF. This was long rumored as a Google technique, so great to see literature on it.

- *Shepherd: A Critic for Language Model Generation*
  https://arxiv.org/abs/2308.04592
  This is simply a model that specializes in taking in a text and a rule and creating feedback data. These models could heavily reduce the community's needs to use GPT4 for creating data.

Reformulating reward models

Probably the space I'm most excited to see work coming in based on my many statements about needing transparency of reward models, but we need even more! Mostly, this work is empirical work showing of the reward models perform with some early solutions to the problems.

- *Reward Model Ensembles Help Mitigate Overoptimization*
  https://arxiv.org/abs/2310.02743
  Shows that ensembling (training multiple models on overlapping subsets of the data) improves the scaling laws for overoptimization proposed in this well known paper. This isn't surprising to me with my background in model-based RL!
- *Learning Optimal Advantage from Preferences and Mistaking it for Reward*
  https://arxiv.org/abs/2310.02456
  A mostly position / theory paper that aids the one in the training section above to show that we need to better ground the reward score used with how the data is actually collected.
- *The Trickle-down Impact of Reward (In-)consistency on RLHF*
  https://arxiv.org/abs/2309.16155v1
  This paper studies if reward models assign a similar score to sentences with different word orders. Hint, they don't, so the authors create a benchmark to evaluate progress in this area! Woot!

Early days of empirical understanding

These papers are probably the most important ones of 2023. They just show us what problems RLHF has, what it is actually good for, and what tools you can use with it!

- *Understanding the Effects of RLHF on LLM Generalisation and Diversity*
  https://arxiv.org/abs/2310.06452

Shows that RLHF encourages more diversity in responses when compared to supervised fine-tuning (SFT), at the expense of output diversity (i.e. the model is repetitive).

- *A Long Way to Go: Investigating Length Correlations in RLHF* https://arxiv.org/abs/2310.03716
  This numerically confirms something we constantly were running into at HuggingFace: RLHF makes answers longer. This comes down to how preference data is collected and the lack of a penalty term.
- *Exploring the impact of low-rank adaptation on the performance, efficiency, and regularization of RLHF* https://arxiv.org/abs/2309.09055
  A study of using LoRA with PPO RLHF, already shows some weird things. Unfortunately, using LoRA on 7Billion parameter models won't be the norm, so this needs to be extended.
- *N implementation details of RLHF* https://huggingface.co/blog/the_n_implementation_details_of_rlhf_with_ppo
  This detailed blog post shows what details in OpenAI's original TensorFlow RLHF implementation actually make it click. It goes very deep, giving a foundation to build bigger scaling results on top of.

Newsletter stuff

Elsewhere from me:

- [Podcast 🎤] On episode 6 of The Retort AI Podcast, Tom and I discuss the infamous Techno-Optimist Manifesto.
- [Paper] I wrote a paper on the origins of RL and RLHF, specifically the risks around not giving more transparency to reward models. More on this soon.

New models of note:

- [Open models] Adept released another 8Billion parameter model Fuyu. This multimodal model was trained with a much simpler architecture (no cross-attention or frozen base LLM), making it seem like there could be another acceleration in multimodal LLMs because training is more stable.

Things I have consumed this week and thought were worth some of your time:

- [Democratic AI] Anthropic ran an experiment (that albeit got a lot of coverage) where they had average Americans write the principles for Constitutional AI. The takeaways here are generally misunderstood, but rather should be:
  1. It is possible to get RLAIF to train stably with broad consensus principles.
  2. The original principles used in Claude are reflective of a detailed engineering process — not all principles would actually converge to a stable model.
- [OpenAI business] A Forbes article I was quoted in discussing the underlying business behind OpenAI and its tensions.
- [Podcast] An episode of Latent Space with Jeremy Howard (one of the GOATs of open-source ML), with a lot of the story behind Fast.ai and the emergence of transfer learning. Loved this one.
- [Podcast] An episode of Generating Conversation on MemGPT, a new way of managing context windows like a memory cache. Good timing, as I'm thinking about LLMs as computing platforms.
- [A game] Trust and Safety Tycoon gives a great little introduction to the hard decisions when growing an online platform.

  My game:
  🛡️ #TrustAndSafetyTycoon
  🏆 Won via IPO and left to advise the government
  📈 Score: 1853
  📏 Rating: ⭐⭐⭐⭐
- [HuggingFace / China] HuggingFace was finally restricted by the Great Firewall. Growing pains to become a real company (I had no idea this was going to happen when I was there). The same applies to GitHub.

Housekeeping:

- **Interconnects referrals:** You'll accumulate a free paid sub if you use a referral link from the Interconnects Leaderboard.

- **Student discounts:** Want a large paid student discount, go to the About page.
- **Like this?** A comment or like helps Interconnects grow!

Seems like I'm slowly getting better at reading stuff again after the HuggingFace hustle months. Let me know if you enjoy the links! I really try to only include stuff that's relevant and authentic to the audience.

12 Likes · 1 Restack

← Previous Next →

1 Comment

Write a comment...

swyx & Alessio Latent Space Oct 25, 2023

thanks for the mention Nathan!

LIKE REPLY SHARE