# Discussion on using RLHF in LLMs and its limitations

Linkedin, host: Ahmad Beirami

There is a subtle distinction between RL in RLHF and RL in domains with a clear reward signal that captures what we want like winning in games or correctness in math.  
  
With a clear reward, RL can lead to discovering novel sequences of actions (e.g. move 37) (see [1]).  
  
But, RLHF actually has a closed form solution (aligned distribution), which is the reference policy exponentially tilted with the reward, leading to an exponential family of distributions with a lot of cool properties.  
See: [4] and [2]  
  
RL in RLHF basically enables us to distills the "aligned distribution" into a new model.   
  
One can (learn to) sample from it in other ways:  
- reward distillation: [7]  
- DPO: [5]  
- controlled decoding: [3]  
- best-of-n: [2]  
  
Similar to best-of-n, in all of these cases RLHF will not lead to new knowledge or skills (when searching in a small KL ball of the reference model), it merely reranks/filters the output distribution of the reference model. (for details on KL regularization of RL policy iteration see [8])

A diagram of a triangle

Description automatically generated

RLHF provably can't teach models any new knowledge. If you need to teach new skills, you need to look at pre-training and SFT.   
  
Why?   
  
Because you'd want to search within a small KL ball of the reference model to avoid reward hacking (and because policies that get almost perfect win rate exist there already). In a small KL ball, you won't generate anything that's not already likely under the reference model.  
  
Alternatively, because best-of-n is almost optimal for this task, and best-of-n cannot learn new skills.

Comment by M. Suleman:

I wonder if the emphasis on "reranking within a small KL ball" limits our understanding of RLHF's potential. While it’s true that RLHF often focuses on refining alignment with human preferences, could this closed-form approach be restricting? By focusing too heavily on maintaining proximity to the reference model, are we potentially overlooking the value of more exploratory policies that might push beyond safe boundaries and uncover genuinely novel, beneficial insights? It seems there’s a balance to be struck between alignment and innovation that RLHF hasn't fully addressed yet.

Answer to M. Suleman’s comment:

Without it, we need a perfect reward that captures the goodness of response. That reward itself is a much stronger LM; so we should just outright use it or distill it into a new model. Today, no such strong universal reward signal exists. We have some signal in math/code/etc domains but that's about it.

Comment by S. Najafi:

Seems there is another fundamental issue:  
The expected gradient for an input vanishes when its reward standard deviation under the model is small, even if the expected reward is far from optimal. For details see [6].

## References

[1] [AlphaGo versus Lee Sedol](https://en.wikipedia.org/wiki/AlphaGo_versus_Lee_Sedol)

[2] [Asymptotics of Language Model Alignment, Joy Qiping Yang et al, U. of Sydney, Google DeepMind, 2024](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/human_like_reasoning/Asymptotics_of_Language_Model_Alignment_Yang_2024.pdf)

[3] [Controlled Decoding from Language Models, Sidharth Mudgal et al, 2024](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/reinforcement_learning_from_human_feedback/Controlled_Decoding_from_Language_Models_Mudgal_2024.pdf)

[4] [RL with KL penalties is better viewed as Bayesian inference, Tomasz Korbak et al, 2022](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/reinforcement_learning_from_human_feedback/RL_with_KL_penalties_is_better_viewed_as_Bayesian_inference_Korbak_2022.pdf)

[5] [Direct Preference Optimization: Your Language Model is Secretly a Reward Model, Rafel Rafailov et al, Stanford U., 2023](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/reinforcement_learning_from_human_feedback/Direct_Preference_Optimization-Your_Language_Model_is_Secretly_a_Reward_Model_Rafailov_Stanford_2023.pdf)

[6] [Vanishing Gradients in Reinforcement finetuning of Language Models, Noam Razin et al, 2024](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/reinforcement_fine_tuning_of_LLM/Vanishing_Gradients_in_Reinforcement_Finetuning_of_Language_Models_Razin_2024.pdf)

[7] [Robust Preference Optimization through Reward Model Distillation, Adam Fisch et al, 2024](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/reinforcement_fine_tuning_of_LLM/Robust_Preference_Optimization_through_Reward_Model_Distillation_Fisch_2024.pdf)

[8] [Optimization Issues in KL-Constrained Approximate Policy Iteration, Nevena Lazic et al, 2021](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/Optimization_Issues_in_KL-Constrained_Approximate_Policy_Iteration_Lazic_2021.pdf)