Aligning language models with human preferences

December 16th, 2025

Based on

Coursera IBM Generative AI Advance Fine-Tuning for LLMs
https://www.coursera.org/learn/generative-ai-advanced-fine-tuning-for-llms ;
https://github.com/ngoc-minh-do/IBM-AI-Engineering-Professional/tree/main/11.%20Generative%20AI%20Advance%20Fine-Tuning%20for%20LLMs

a smol course
https://github.com/huggingface/smol-course/tree/main/v1/2_preference_alignment

The Alignment Handbook https://github.com/huggingface/alignment-handbook

Direct Preference Optimization Explained In-depth
https://www.tylerromero.com/posts/2024-04-dpo/

Introduction

- **Continued pretraining**: to adapt language models to a new language or domain
- **Supervised Fine-Tuning** (SFT): adapt models to specific tasks
- **Preference alignment** (like RLHF or DPO): to improve response quality, preferring some answers to others with respect to some values (to some human preferences)

Preference alignment methods

- Reinforcement Learning From Human Feedback (RLHF)
- Direct Preference Optimization (DPO)
- Odds Ratio Preference Optimization (ORPO)
- Group Relative Policy Optimization (GRPO)

Reinforcement learning

« Learning from **interaction** with environment without explicit teacher »

See Reinforcement Learning: An Introduction (Sutton and Barto, 2015) for methods

4 major elements in a RL system:

- A **policy** defines the learning agent's way of behaving at a given state of the environment
- A **reward signal** (a number) indicates what is good or bad in an immediate sense
- A **value function** specifies what is good in the long run. Total amount of reward an agent can expect to accumulate over the future, starting from one given state
- A **(reward) model of the environment** that allows inferences to be made about how the environment will behave

Reinforcement Learning From Human Feedback (RLHF)

Fine-Tuning Language Models from Human Preferences (Ziegler, Stiennon et al. @OpenAI, 2020)

1. Training a **Reward Model**
2. Reinforce the language model (policy) using a RL method called **Proximal Policy Optimization (PPO)** (Schulman et al. 2017 @OpenAI 2017)

The process of reinforcement learning by human feedback involves several important steps:

1. **Guidelines** to guarantee unique criteria when deciding what is a good and a bad answer to an input
2. A **Reward Model (RM)** should be trained, which will evaluate each of the responses in terms of accuracy, relevance, and adherence to guidelines
3. To train the RM, some prompts are selected and sent to human reviewers. We call them **Preference Data (PD)**
4. The reviewers then interact with the model and manually evaluate and rate the corresponding outputs.
5. The collected feedback, in the form of ratings or rankings, is used to train the RM.
6. With the RM trained, we can train a **Policy Optimizer**, a required component which will guide the fine-tuning of the LLM.
7. We fine-tune the LLM with Policy Optimization.
8. This iterative feedback loop allows the model to gradually learn from human guidance and refine its behavior accordingly.

Source: https://argilla.io/blog/mantisnlp-rlhf-part-2/

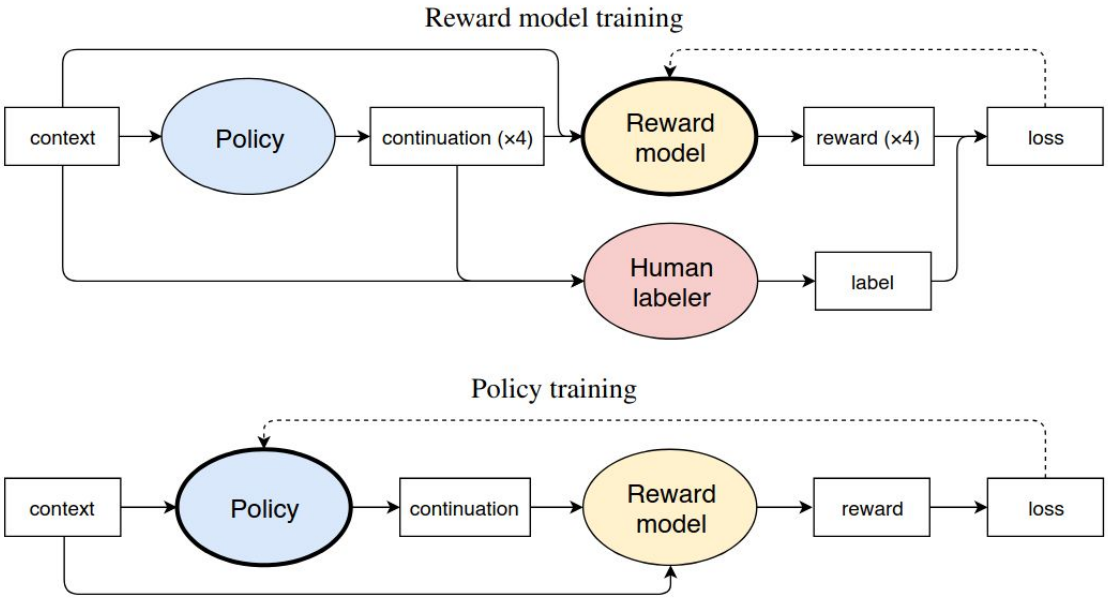


Figure 1: Our training processes for reward model and policy. In the online case, the processes are interleaved.

Conclusions about RLHF

- Approach complex, difficult to reproduce, computationally expensive and unstable:

  -> require reward model fitting, LM sampling, and extensive hyperparameter tuning

  -> require to making smaller, incremental updates to the policy, as larger updates can lead to instability or suboptimal solutions

Direct Preference Optimization (DPO)

Direct Preference Optimization: Your Language Model is Secretly a Reward Model (Rafailov et al. @Stanford, NeurIPS 2023)

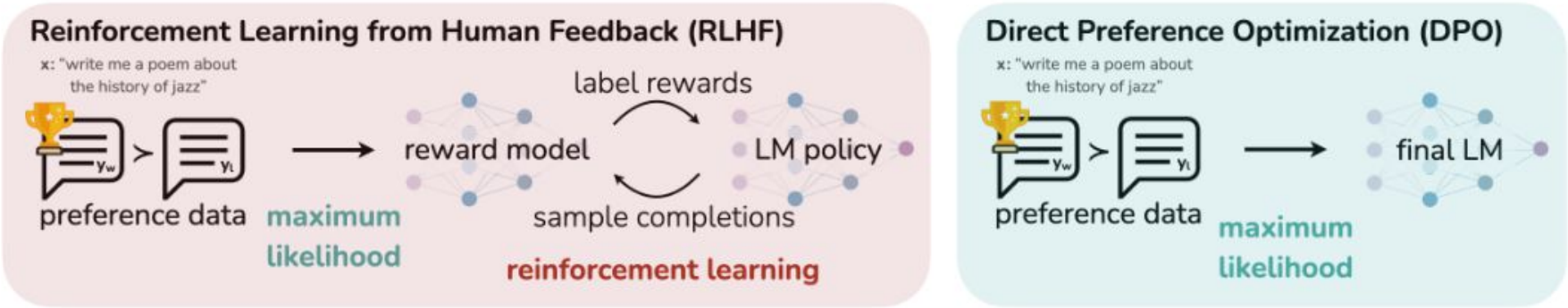
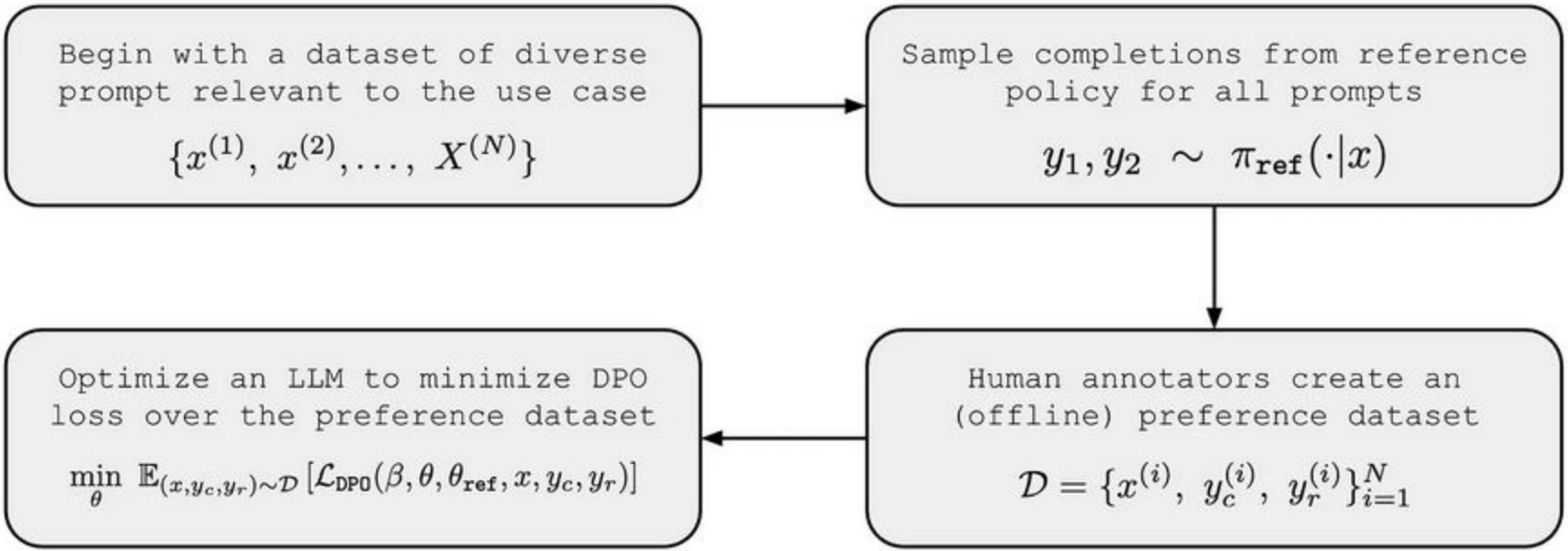
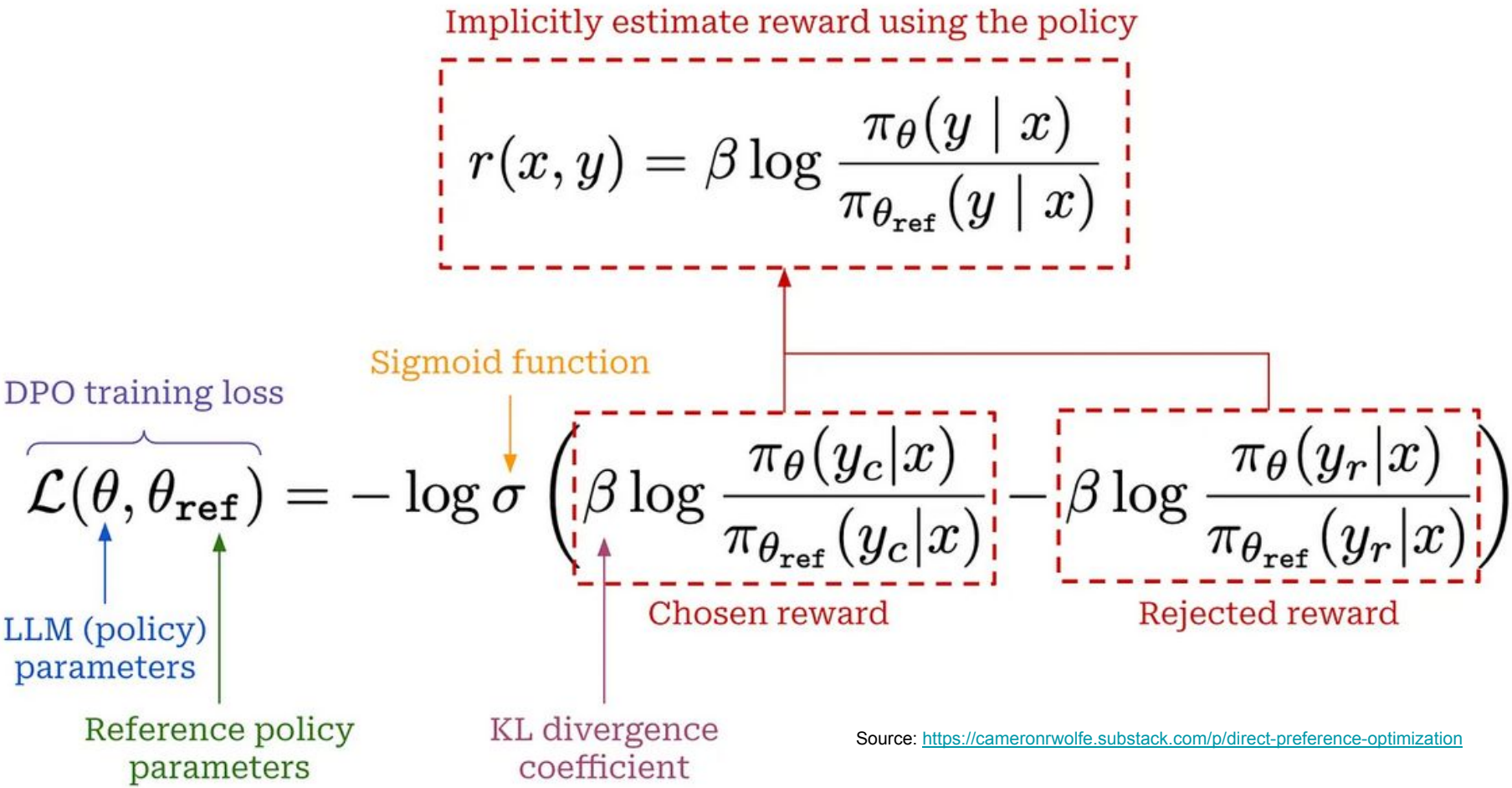


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes the learned reward. In contrast, DPO directly optimizes for the policy best satisfying the preferences with a simple classification objective, fitting an *implicit* reward model whose corresponding optimal policy can be extracted in closed form.



Standard DPO training pipeline

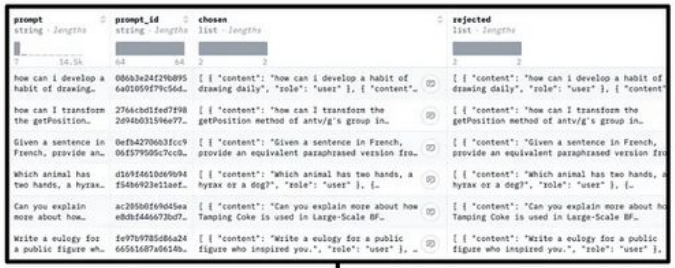
Source: https://cameronrwolfe.substack.com/p/direct-preference-optimization



Source: https://cameronrwolfe.substack.com/p/direct-preference-optimization

DPO training loss (from [1])

Source: https://cameronrwolfe.substack.com/p/direct-preference-optimization



prompt string · lengths	prompt_id string · lengths	chosen list · lengths	rejected list · lengths
how can i develop a habit of drawing…	086b3e24f29b895 6a01059f79c56d…	[{ "content": "how can i develop a habit of drawing daily", "role": "user" }, { "content"…	[{ "content": "how can i develop a habit of drawing daily", "role": "user" }, { "content
how can I transform the getPosition…	2766cbd1fed7f98 2d94b031596e77…	[{ "content": "how can I transform the getPosition method of antv/g's group in…	[{ "content": "how can I transform the getPosition method of antv/g's group in…
Given a sentence in French, provide an…	0efb42706b3fcc9 06f579505c7cc0…	[{ "content": "Given a sentence in French, provide an equivalent paraphrased version fro…	[{ "content": "Given a sentence in French, provide an equivalent paraphrased version fro
Which animal has two hands, a hytax…	d169f461bde9b94 f54b6923e11aef…	[{ "content": "Which animal has two hands, a hytax or a dog?", "role": "user" }, {…	[{ "content": "which animal has two hands, a hytax or a dog?", "role": "user" }, {…
Can you explain more about how…	ac205b0f69d45ea e8dbf4466473bd7…	[{ "content": "Can you explain more about how Tamping Coke is used in Large-Scale BF…	[{ "content": "Can you explain more about ho Tamping Coke is used in Large-Scale BF…
Write a eulogy for a public figure wh…	fe97b9785d8a24 66561687a0614b…	[{ "content": "Write a eulogy for a public figure who inspired you.", "role": "user" }, …	[{ "content": "Write a eulogy for a public figure who inspired you.", "role": "user" },

Download a (static) preference dataset

$$\mathcal{D} = \{x^{(i)}, y_c^{(i)}, y_r^{(i)}\}_{i=1}^{N}$$

SFT on chosen completions

$$\min_{\theta} \mathbb{E}_{(x,y_c)\in\mathcal{D}} \left[-\log \pi_\theta(y_c|x)\right]$$

DPO training

$$\min_{\theta} \mathbb{E}_{(x,y_c,y_r)\sim\mathcal{D}} \left[\mathcal{L}_{\texttt{DPO}}(\beta, \theta, \theta_{\texttt{ref}}, x, y_c, y_r)\right]$$

Mitigating distribution shift from offline preference data in DPO

Conclusions about DPO

- avoid the need to train a reward model to estimate human preferences
- avoid needing to perform any type of reinforcement learning (a notoriously difficult task)
- can directly optimize the LLM on human preferences using supervised learning

Odds Ratio Preference Optimization (ORPO)

ORPO: Monolithic Preference Optimization without Reference Model (Hong et al. @KAIST AI, 2024)

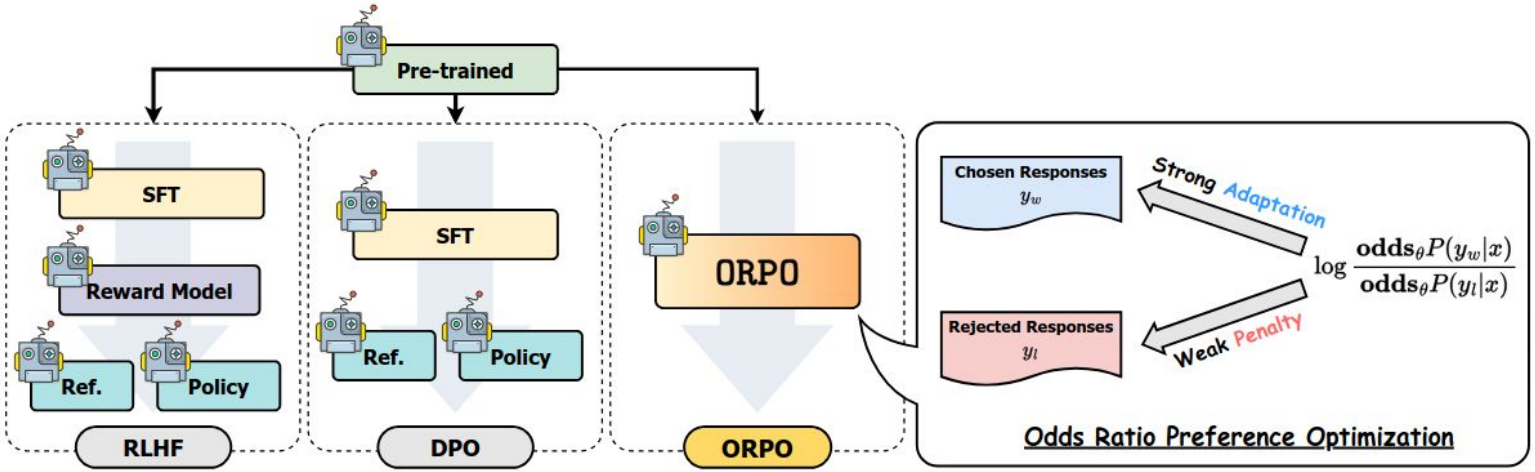


Figure 2: Comparison of model alignment techniques. ORPO aligns the language model *without a reference model* in a single-step manner by assigning a weak penalty to the rejected responses and a strong adaptation signal to the chosen responses with a simple log odds ratio term appended to the negative log-likelihood loss.

RLHF and DPO approach involves **several models and training stages**
RLHF with its SFT, RM, and PPO steps, or DPO with its SFT and DPO stages

SFT increased the likelihood of obtaining the desired tokens, but it also raised the probability of generating undesired outcomes.

ORPO still adapt the models to the specific domain, but at the same time penalize undesired responses ; all at the SFT stage

ORPO **combines instruction tuning and preference alignment in a single process**,

modifies the standard language modeling objective by combining negative log-likelihood loss with an odds ratio term on a token level. The approach features a unified single-stage training process, reference model-free architecture, and improved computational efficiency.

Conclusions about ORPO

penalizes the model from learning undesired generation styles during SFT

do not require SFT warm-up stage nor a reference model

computationally more efficient

Group Relative Policy Optimization (GRPO)

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models (Shao et al. @DeepSeek-AI 2024)

« a variant of Proximal Policy Optimization (PPO), that enhances mathematical reasoning abilities while concurrently optimizing the memory usage of PPO »

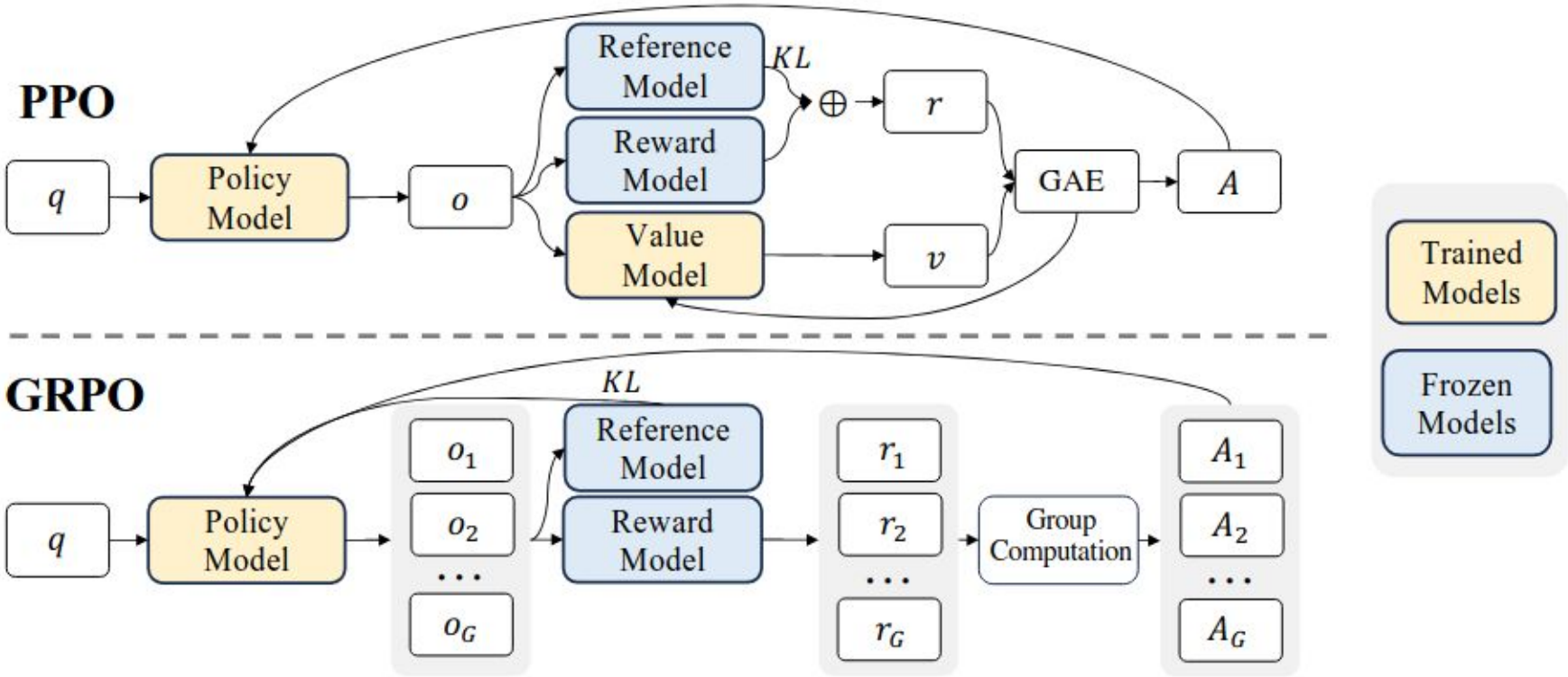
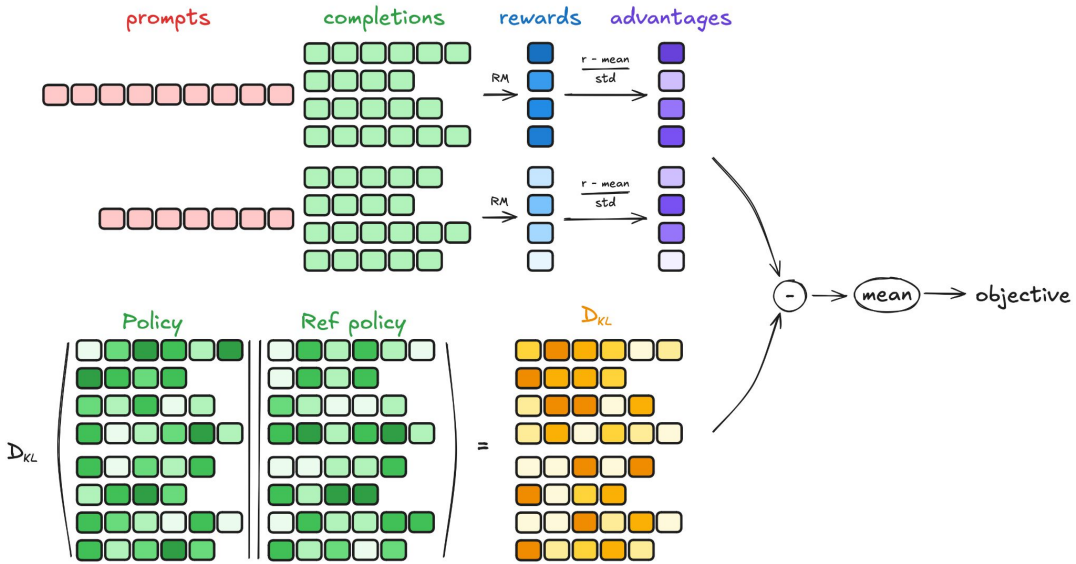


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

Kullback–Leibler divergence, a measure of how much an approximating probability distribution is different from a true probability distribution (Wikipedia)

GRPO is an online learning algorithm, meaning it improves iteratively by using the data generated by the trained model itself during training



Source: https://huggingface.co/docs/trl/v0.26.1/en/grpo_trainer

Conclusions

Conclusion

Several preference alignment methods consists of a multi-stage process (complex to tune) ; the evolution tends to reduce the cost (i.e. need for external resources or third-party model training)

Hugging Face transformer reinforcement learning (trl) library implements SFT, DPO, GRPO, Reward Modeling…

Go further

GRPO EssentialAI/rnj-1-instruct with QLoRA using TRL
https://github.com/huggingface/trl/blob/main/examples/notebooks/grpo_rnj_1_instruct.ipynb