Reinforcement Learning in Post-Training

Bridging the Gap in News Production: Aligning AI Models

GIOFRÉ Daniele

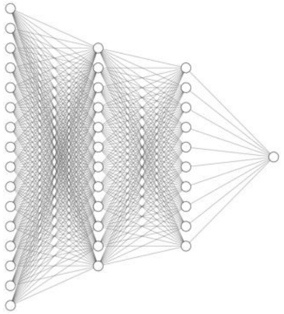
Contents

LLM Training - Overview

The Gear of Training

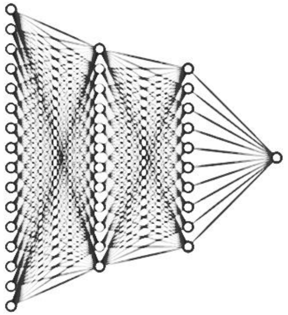
Pretraining

Gear 1



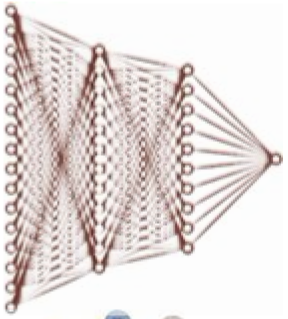
Continual PreTraining

Gear 2



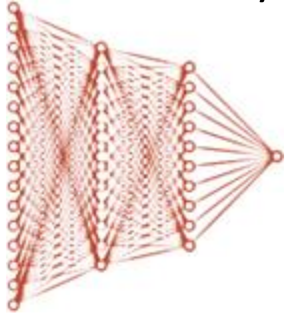
Instruction FineTuning

Gear 3



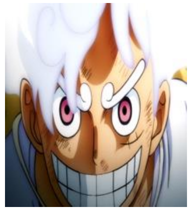
Reinforcement **L**earning & **G**ame **T**heory

Gear 4



PPO, DPO, ORPO, DNO, GRPO, etc…

?



Gear 5

Thomson Reuters™

RL in Post-Training Alignment

The Gear 4

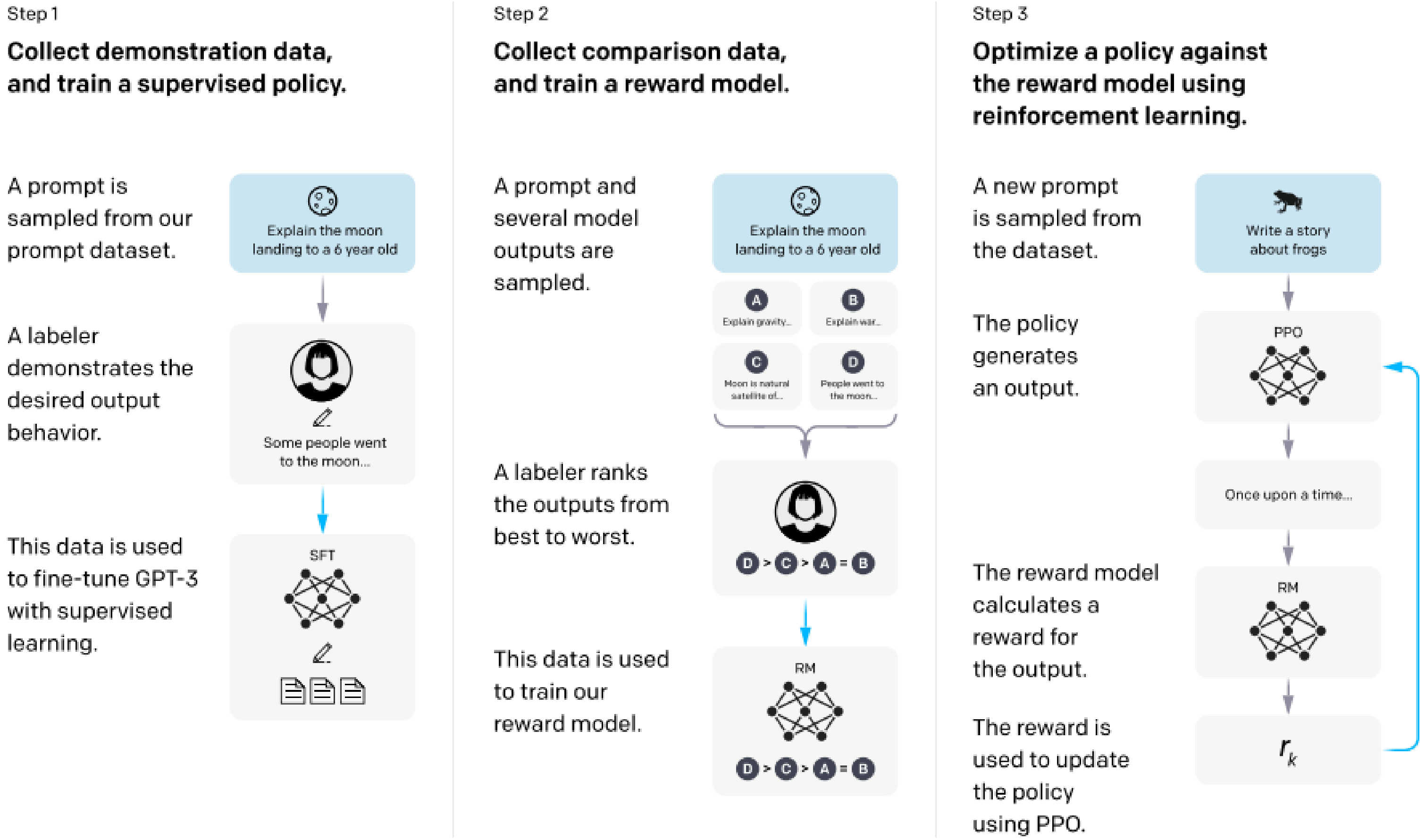
Making a LLM better at following instructions and at reasoning:

- Evolution from InstructGPT to modern approaches
- Focus on PPO and DPO methods
- Bypass DPO limitations
- Make the bridge between PPO and DPO via Online DPO

InstructGPT and PPO Framework

Schema

- Three-stage process:
  - Supervised fine-tuning (SFT)
  - Reward Modeling (RM)
  - PPO optimization
- Key innovation:
  Using human feedback
  for alignment



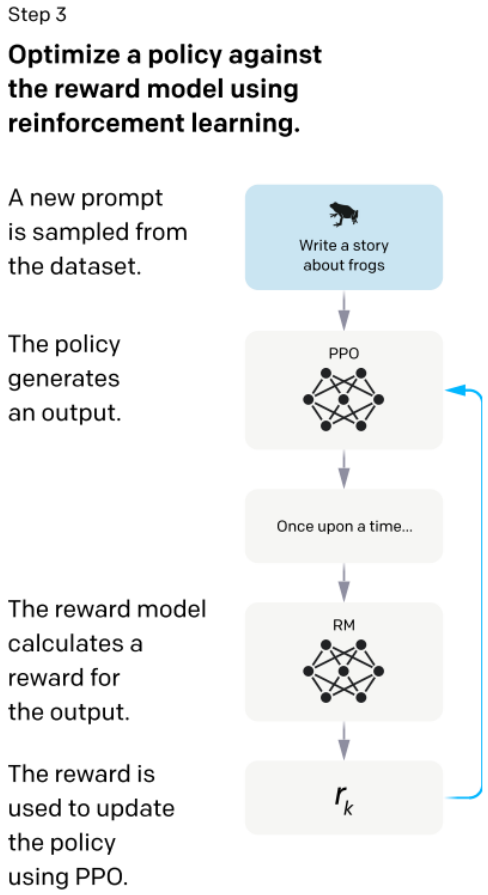
OUYANG, Long, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 2022, 35: 27730

InstructGPT and PPO Framework

How it Works

- Policy optimization using RL
- Iterative process:
  - Sample responses from policy
  - Evaluate with reward model
  - Update policy with PPO loss
- Ensures controlled policy updates

$$\text{objective}(\phi) = \mathbb{E}_{(x,y)\sim D_{\pi_\phi^{RL}}}\left[r_\theta(x,y) - \beta \log \frac{\pi_\phi^{RL}(y|x)}{\pi^{SFT}(y|x)}\right] + \gamma \mathbb{E}_{x\sim D_{\text{pretrain}}} \log(\pi_\phi^{RL}(x))$$

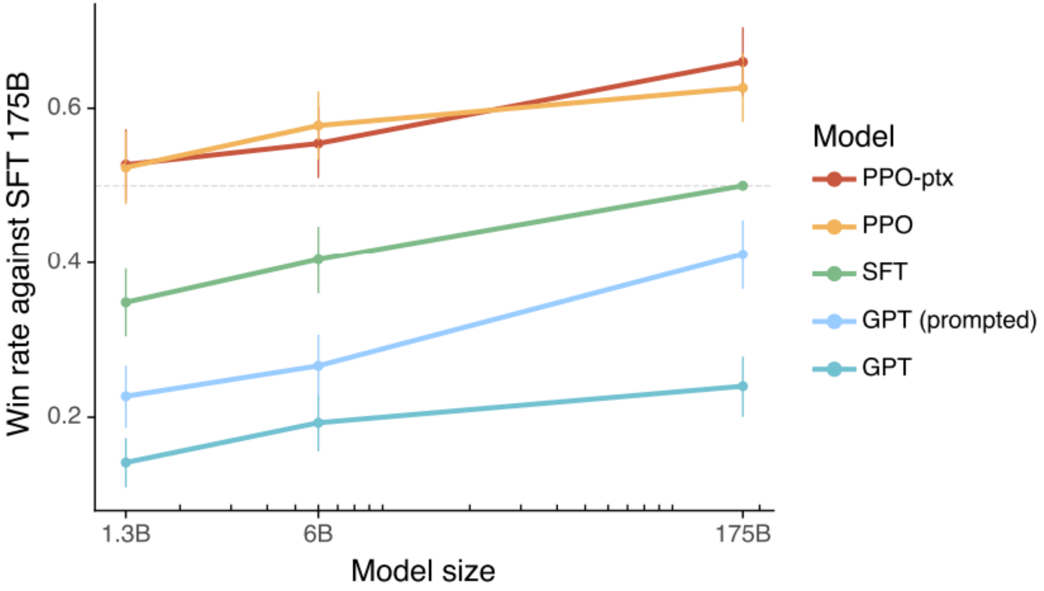


OUYANG, Long, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 2022, 35: 27730

InstructGPT and PPO Framework

Does it Work?

- Policy optimization using RL
- Iterative process:
  - Sample responses from policy
  - Evaluate with reward model
  - Update policy with PPO loss
- Ensures controlled policy updates



$$\text{objective}(\phi) = \mathbb{E}_{(x,y)\sim D_{\pi_\phi^{RL}}}\left[r_\theta(x,y) - \beta \log \frac{\pi_\phi^{RL}(y|x)}{\pi^{SFT}(y|x)}\right] + \gamma \mathbb{E}_{x\sim D_{\text{pretrain}}} \log(\pi_\phi^{RL}(x))$$

OUYANG, Long, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 2022, 35: 27730

InstructGPT and PPO Framework

Challenges and Limitations

- Resource intensive: 3 models in memory
  - SFT model
  - Current policy
  - Reward model
- Requires on-policy data collection
- Complex training dynamics – 3 steps
- High computational cost

$$\text{objective}(\phi) = \mathbb{E}_{(x,y)\sim D_{\pi_\phi^{RL}}}\left[r_\theta(x,y) - \beta \log \frac{\pi_\phi^{RL}(y|x)}{\pi^{SFT}(y|x)}\right] + \gamma \mathbb{E}_{x\sim D_{\text{pretrain}}} \log(\pi_\phi^{RL}(x))$$

OUYANG, Long, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 2022, 35: 27730

Direct Preference Optimization (DPO)

- Single-stage optimization: “from reward functions to optimal policies”
- Directly learns from preference data:
- No explicit reward model needed, but implicit rewards: $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$
- Loss based on a Berry-Terry pair-wise modelling:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_{\text{c}}, y_{\text{r}}) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_{\text{c}} \mid x)}{\pi_{\text{ref}}(y_{\text{c}} \mid x)} - \beta \log \frac{\pi_\theta(y_{\text{r}} \mid x)}{\pi_{\text{ref}}(y_{\text{r}} \mid x)} \right) \right]$$

RAFAILOV, Rafael, et al. Direct Preference Optimization. Advances in Neural Information Processing Systems, 2024, 36.

Thomson Reuters™

Direct Preference Optimization (DPO)

Limitations

- Training instability if $\hat{r}_{\theta,c}(x,y) \simeq \hat{r}_{\theta,r}(x,y)$eak learning
- Overoptimization in RL
- Unclear separation between chosen and rejected
- Offline training – off-policy issue
- Double samples for same micro batch – pair-wise comparison
- Dataset quality dependency

RAFAILOV, Rafael, et al. Direct Preference Optimization. Advances in Neural Information Processing Systems, 2024, 36.

Direct Preference Optimization (DPO)

Limitations

- Training instability if $\hat{r}_{\theta,c}(x,y) \simeq \hat{r}_{\theta,r}(x,y)$eak learning
- Overoptimization in RL
- Unclear separation between chosen and rejected
- Offline training – all off-policy training issues
- ~~Double samples for same micro batch – pair-wise comparison~~
- ~~Dataset quality dependency~~

RAFAILOV, Rafael, et al. Direct Preference Optimization. Advances in Neural Information Processing Systems, 2024, 36.

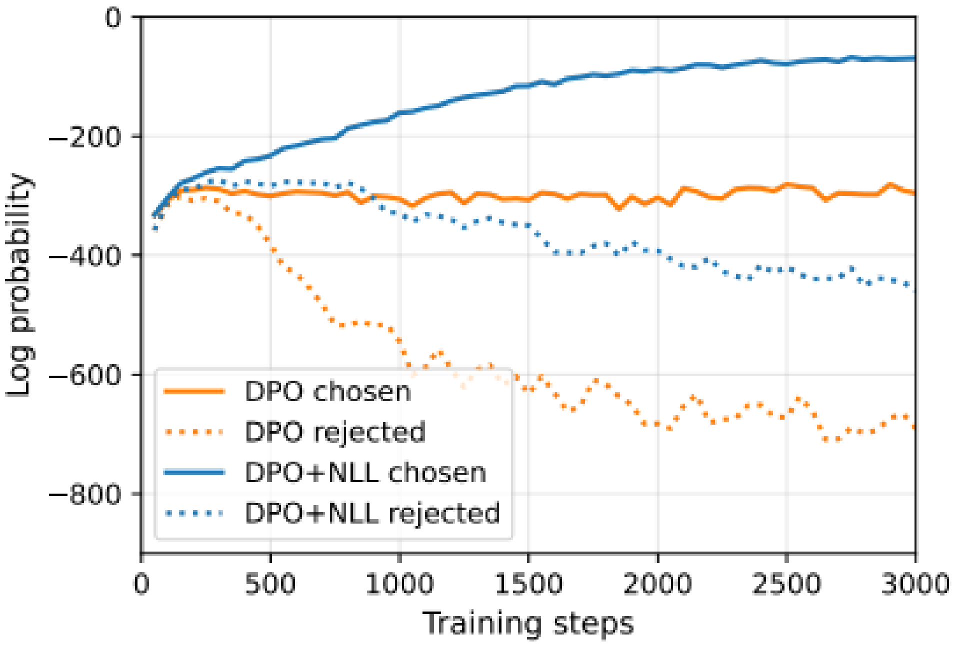
Regularized Preference Optimization (RPO)

SFT Loss is Implicitly an Adversarial Regularizer

RPO objective = Preference optimization loss + Imitation (SFT) loss

$$\mathcal{L}_{\text{DPO+NLL}} = \mathcal{L}_{\text{DPO}}(c_i^{\text{c}}, y_i^{\text{c}}, c_i^{\text{r}}, y_i^{\text{r}} | x_i) + \alpha \mathcal{L}_{\text{NLL}}(c_i^{\text{c}}, y_i^{\text{c}} | x_i)$$

Win rate (%)	RPO (beta)	Ref. (beta)	DPO (beta)
RPO (beta)	50.0	**79.0**	**56.0**
Ref. (beta)	21.0	50.0	22.7
DPO (beta)	44.0	77.3	50.0



(a) ARC-Challenge

LIU, Zhihan, et al. Provably mitigating overoptimization in RLHF. *arXiv preprint arXiv:2405.16436*, 2024.
PANG, Richard Yuanzhe, et al. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

Regularized Preference Optimization (RPO)

By passing some limits of DPO

- ~~Training instability if~~ $\cancel{\hat{r}_{\theta,c}(x,y) \simeq \hat{r}_{\theta,r}(x,y)}$ ~~eak learning~~
- ~~Overoptimization in RL~~
- Unclear separation between chosen and rejected [tackled in *]
- Offline training – all off-policy training issues

$$\mathcal{L}_{\text{DPO+NLL}} = \mathcal{L}_{\text{DPO}}(c_i^{\text{c}}, y_i^{\text{c}}, c_i^{\text{r}}, y_i^{\text{r}}|x_i) + \alpha\mathcal{L}_{\text{NLL}}(c_i^{\text{c}}, y_i^{\text{c}}|x_i)$$

LIU, Zhihan, et al. Provably mitigating overoptimization in RLHF. *arXiv preprint arXiv:2405.16436*, 2024.
* ADLER, Bo, et al. Nemotron-4 340B Technical Report. arXiv preprint arXiv:2406.11704, 2024.

PPO vs DPO Comparison

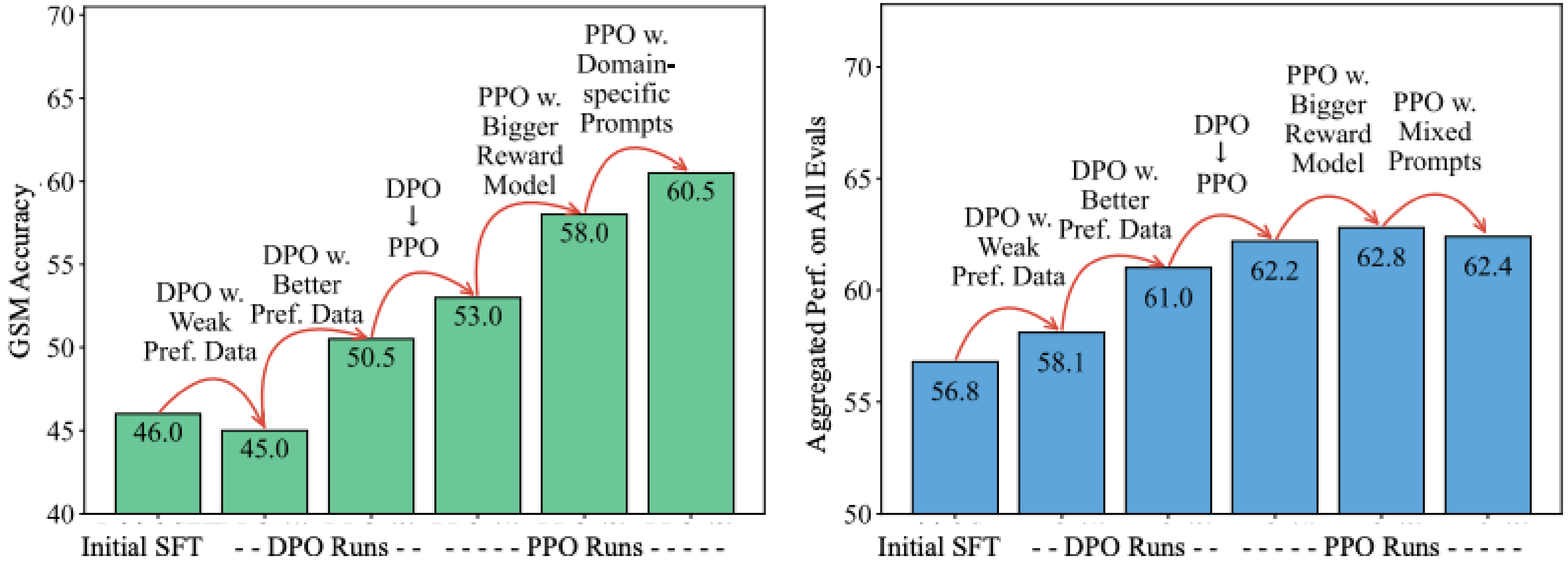


Figure 1: Performance improvements resulted by changing different components in the preference training of TÜLU. Left: Accuracy on GSM [9], for testing math capabilities. Right: Overall performance, aggregated over the 11 benchmarks described in §2.2.

IVISON, Hamish, et al. Unpacking DPO and PPO. arXiv preprint arXiv:2406.09279, 2024.

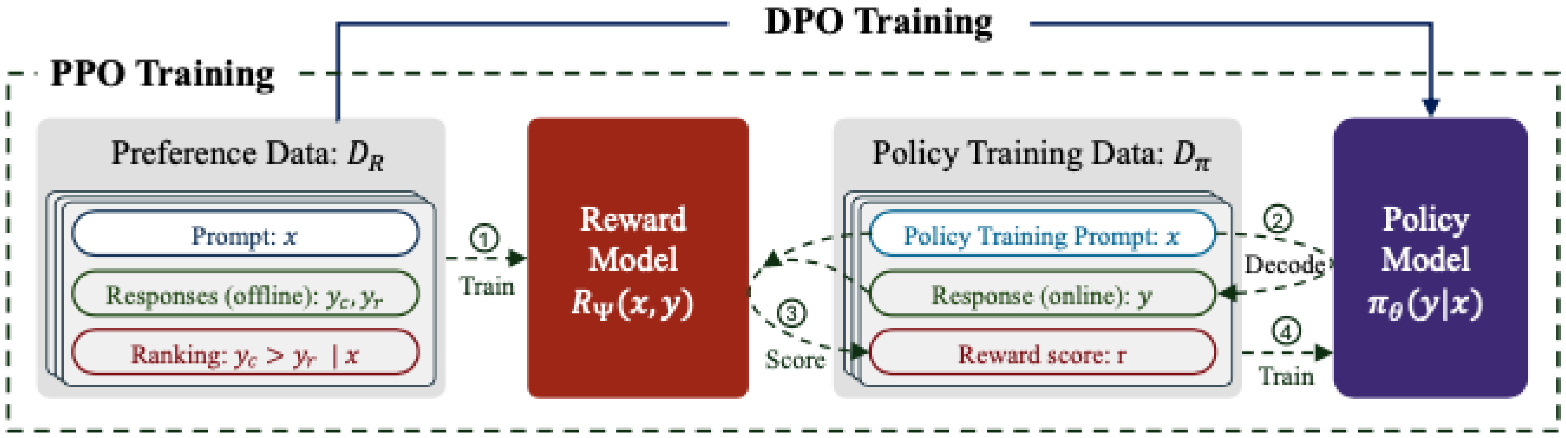
PPO vs DPO Comparison

PPO

- Online generation – Online Learning
- Requires reward model
- More stable training and slightly better metrics
- Higher computational cost

DPO/RPO

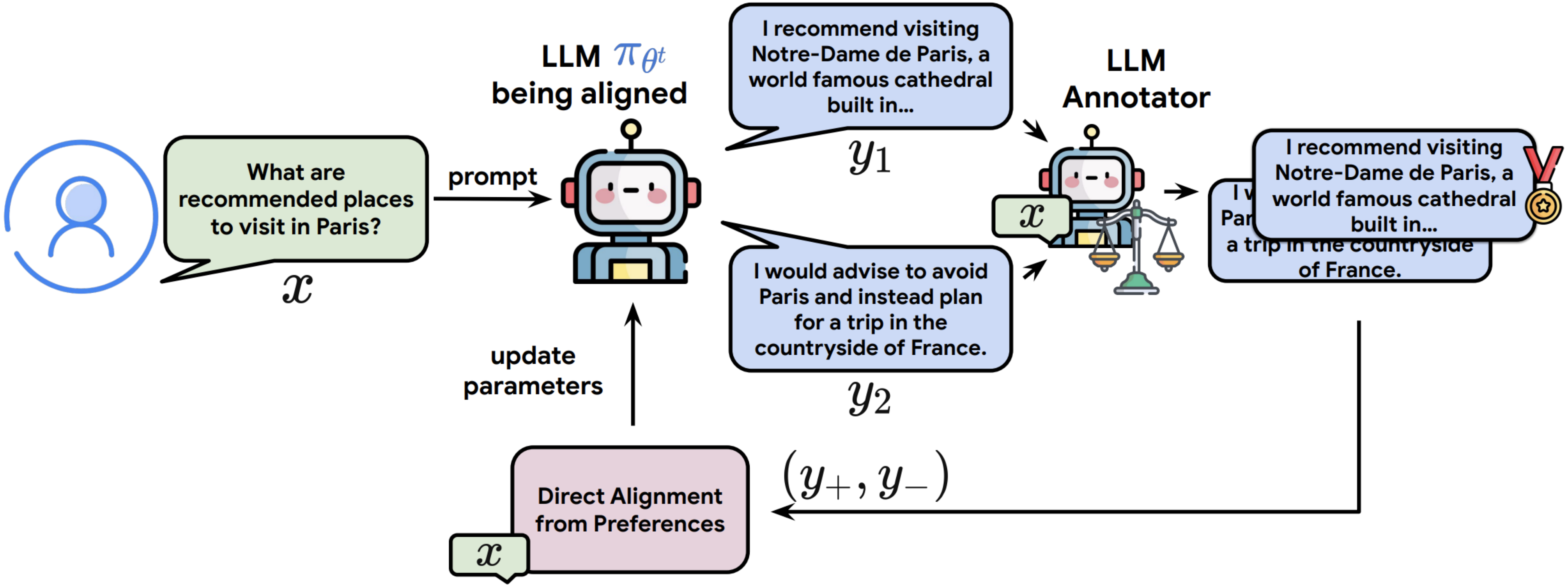
- Offline data allowed – Offline Learning
- No reward model needed
- Simpler implementation and negative samples as zero
- More efficient and fast training



IVISON, Hamish, et al. Unpacking DPO and PPO. arXiv preprint arXiv:2406.09279, 2024.

Online DPO

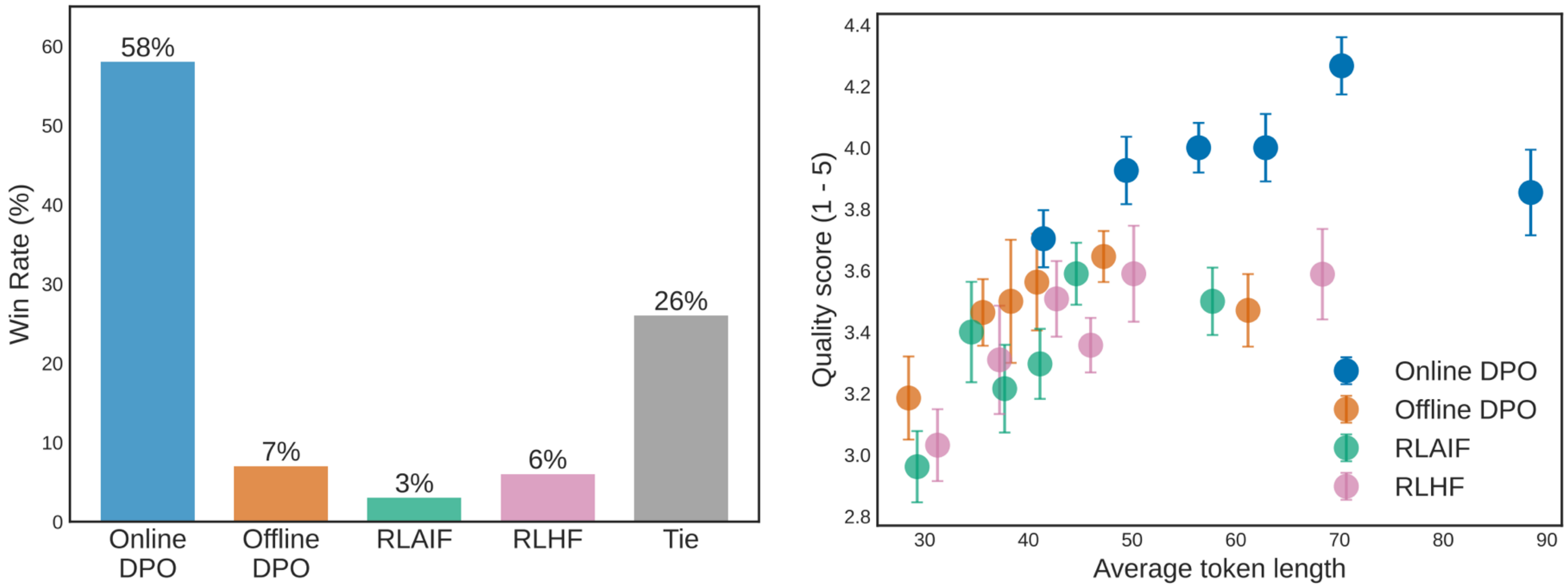
Bridges gap with PPO



GUO, Shangmin, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Online DPO

Bridges gap with PPO



(a) Fraction of responses preferred by humans

(b) Quality against length of responses

GUO, Shangmin, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Thomson Reuters™

Online DPO

Reward Model or LLM-as-Judge?

rewardbench

	Model	Model Type	Score	Chat	Chat Hard	Safety	Reasoning
1	infly/INF-ORM-Llama3.1-70B	Seq. Classifier	95.1	96.6	91.0	93.6	99.1
2	nicolinho/QRM-Gemma-2-27B	Seq. Classifier	94.4	96.6	90.1	92.7	98.3
3	Skywork/Skywork-Reward-Gemma-2-27B-v0.2	Seq. Classifier	94.3	96.1	89.9	93.0	98.1
4	nvidia/Llama-3.1-Nemotron-70B-Reward *	Custom Classifier	94.1	97.5	85.7	95.1	98.1
5	Skywork/Skywork-Reward-Gemma-2-27B ⚠️	Seq. Classifier	93.8	95.8	91.4	91.9	96.1
6	SF-Foundation/TextEval-Llama3.1-70B * ⚠️	Generative	93.5	94.1	90.1	93.2	96.4
7	meta-metrics/MetaMetrics-RM-v1.0	Custom Classifier	93.4	98.3	86.4	90.8	98.2
8	Skywork/Skywork-Critic-Llama-3.1-70B ⚠️	Generative	93.3	96.6	87.9	93.1	95.5
9	nicolinho/QRM-Llama3.1-8B-v2	Seq. Classifier	93.1	96.4	86.8	92.6	96.8
10	Skywork/Skywork-Reward-Llama-3.1-8B-v0.2	Seq. Classifier	93.1	94.7	88.4	92.7	96.7

RewardBench: https://huggingface.co/spaces/allenai/reward-bench

Thomson Reuters™

We are hiring!

Join our Labs and Foundational Research team!

Hiring in **EMEA** (Switzerland and UK) **& North America** for :

- Applied Scientists (Interns, ICs, Team Leads)
- Research Scientists (Interns, ICs, Team Leads)
- Research Engineers (Interns, ICs)

Appendix

Direct Preference Optimization (DPO)

Digging into the loss via its Gradient

- Single-stage optimization
- Directly learns from preference data
- No explicit reward model needed, but implicit rewards

$$\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

- Mathematical formulation:

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \Bigg[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \Bigg[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \Bigg] \Bigg]$$

RAFAILOV, Rafael, et al. Direct Preference Optimization. Advances in Neural Information Processing Systems, 2024, 36.

Regularized Preference Optimization (RPO)

SFT Loss is Implicitly an Adversarial Regularizer

RPO objective = Preference optimization loss + Imitation (SFT) loss

$$\mathcal{L}_{\text{DPO+NLL}} = \mathcal{L}_{\text{DPO}}(c_i^w, y_i^w, c_i^l, y_i^l | x_i) + \alpha \mathcal{L}_{\text{NLL}}(c_i^w, y_i^w | x_i)$$

$$= -\log \sigma \left(\beta \log \frac{M_\theta(c_i^w, y_i^w | x_i)}{M_t(c_i^w, y_i^w | x_i)} - \beta \log \frac{M_\theta(c_i^l, y_i^l | x_i)}{M_t(c_i^l, y_i^l | x_i)} \right) - \alpha \frac{\log M_\theta(c_i^w, y_i^w | x_i)}{|c_i^w| + |y_i^w|}.$$

LIU, Zhihan, et al. Provably mitigating overoptimization in RLHF. *arXiv preprint arXiv:2405.16436*, 2024.
PANG, Richard Yuanzhe, et al. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

Online DPO

Bridges gap with PPO

Method	Win	Tie	Loss	Quality
		TL;DR		
Online DPO	**63.74%**	28.57%	7.69%	**3.95**
Offline DPO	7.69%		63.74%	3.46
		Helpfulness		
Online DPO	**58.60%**	21.20%	20.20%	**4.08**
Offline DPO	20.20%		58.60%	3.44
		Harmlessness		
Online DPO	**60.26%**	35.90%	3.84%	**4.41**
Offline DPO	3.84%		60.26%	3.57

Table 2: Win/tie/loss rate of DPO with OAIF (online DPO) against vanilla DPO (offline DPO) on the TL;DR, Helpfulness, Harmlessness tasks, along with the quality score of their generations, judged by *human raters*.

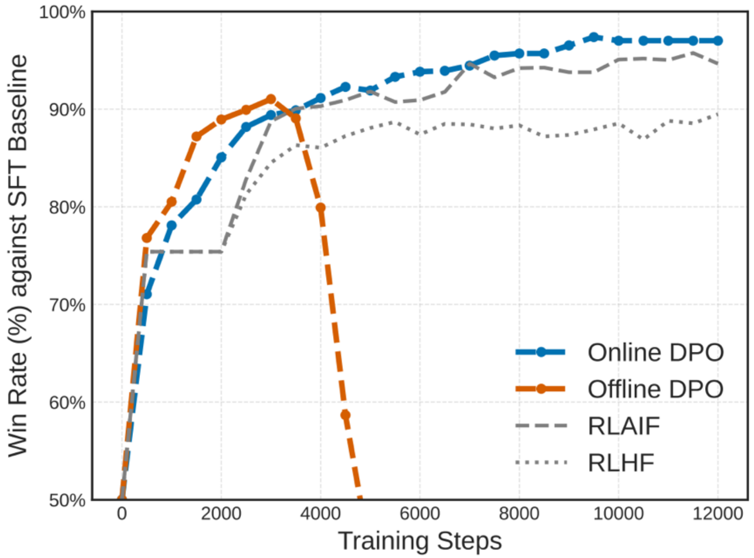


Figure 3: Win rate of DPO with OAIF (online DPO), vanilla DPO (offline DPO), RLAIF, and RLHF against the SFT baseline on the TL;DR task, judged by *Gemini Pro*.

Method	No RM needed	On-policy generation	Online feedback
Offline DPO (Rafailov et al., 2023)	✓	✗	✗
Offline IPO (Azar et al., 2023)	✓	✗	✗
Offline SLiC (Zhao et al., 2023)	✓	✗	✗
RSO (Liu et al., 2023)	✗	✓	✓
Iterative DPO (Xu et al., 2023)	✗	✓	✓
OAIF (proposed)	✓	✓	✓

GUO, Shangmin, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Thomson Reuters™