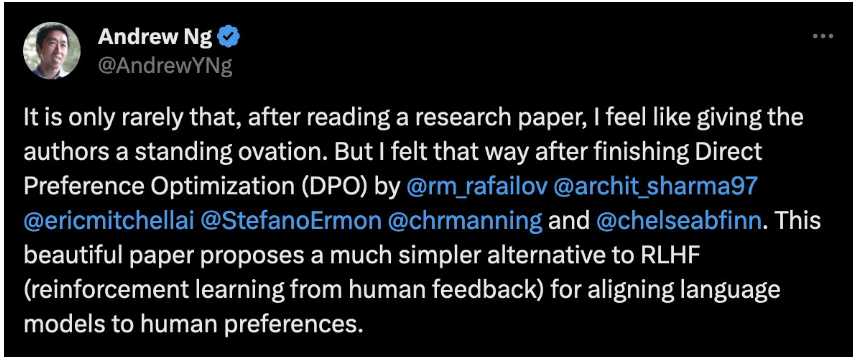
Direct Preference Optimization Explained, really

Roger Lam
lamroger.com

What’s the hype?

Andrew Ng, Co-Founder of Coursera, Stanford CS adjunct faculty and former head of Baidu AI Group/Google Brain, wanted to give the authors a “standing ovation”.

“This beautiful paper proposes a much simpler alternative to RLHF…”



Andrew Ng
@AndrewYNg

It is only rarely that, after reading a research paper, I feel like giving the authors a standing ovation. But I felt that way after finishing Direct Preference Optimization (DPO) by @rm_rafailov @archit_sharma97 @ericmitchellai @StefanoErmon @chrmanning and @chelseabfinn. This beautiful paper proposes a much simpler alternative to RLHF (reinforcement learning from human feedback) for aligning language models to human preferences.

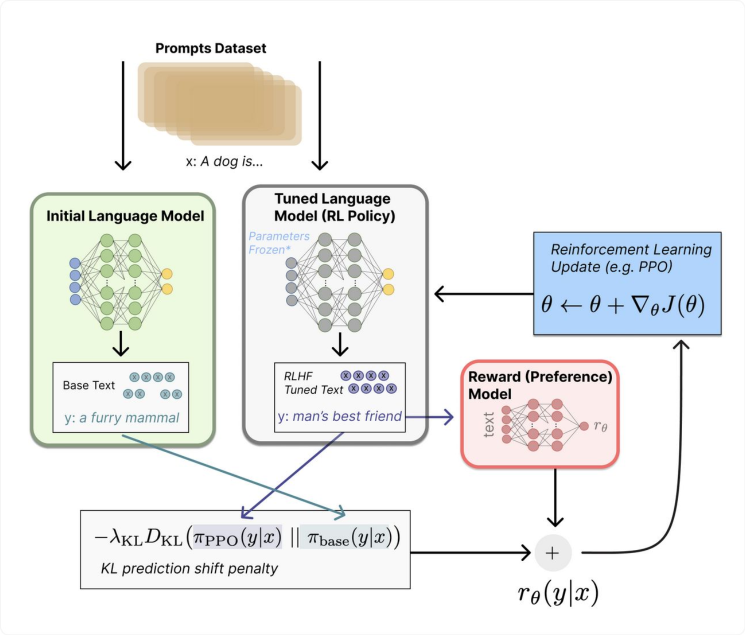
Reinforcement Learning from Human Feedback

RLHF is performed after the initial LLM is trained to align its responses.

It uses two candidate outputs, asks a reward model what it thinks, and tweaks accordingly.

What's tough is that RLHF requires creating a reward model (red) to fine-tune based on human preference.

That makes it pretty heavy in complexity and compute.



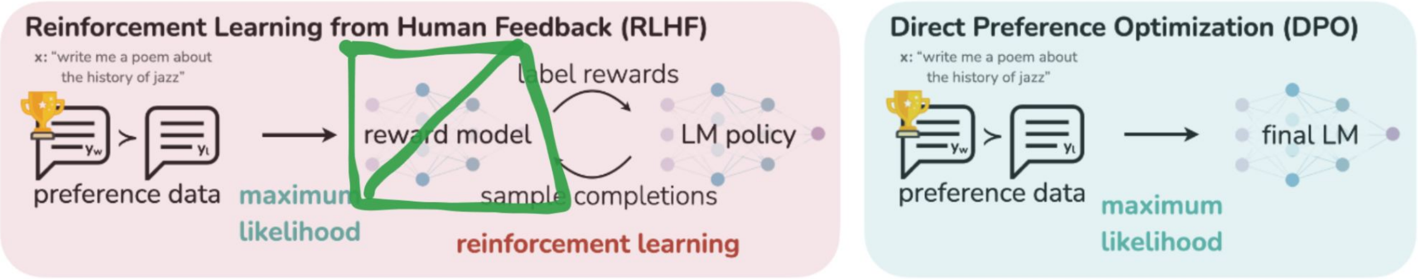
Prompts Dataset
x: A dog is...
Initial Language Model
Base Text
y: a furry mammal
Tuned Language Model (RL Policy)
Parameters Frozen*
RLHF Tuned Text
y: man's best friend
Reinforcement Learning Update (e.g. PPO)
$\theta \leftarrow \theta + \nabla_\theta J(\theta)$
Reward (Preference) Model
text
r_θ
$-\lambda_{\mathrm{KL}} D_{\mathrm{KL}}(\pi_{\mathrm{PPO}}(y|x) \,||\, \pi_{\mathrm{base}}(y|x))$
KL prediction shift penalty
+
$r_\theta(y|x)$

https://huggingface.co/blog/rlhf

What is Direct Preference Optimization?

DPO is a faster and more elegant way to tune your LLM than the traditional RLHF way.

It drops the reward model and instead uses the preference data directly to fine-tune the model.

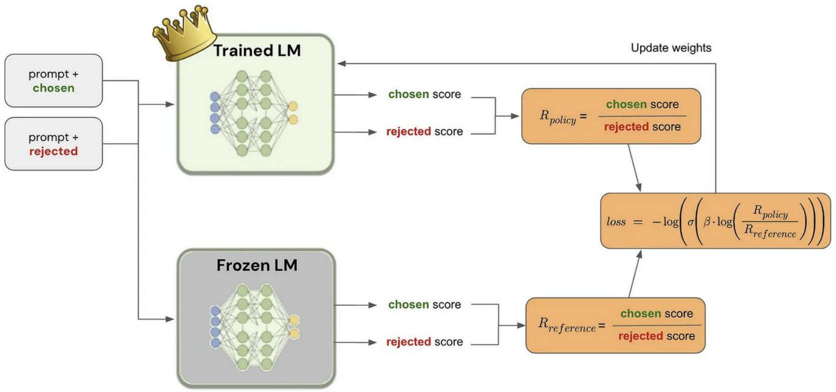

Reinforcement Learning from Human Feedback (RLHF)
x: "write me a poem about the history of jazz"
y_w
y_l
preference data
maximum likelihood
reward model
label rewards
LM policy
sample completions
reinforcement learning
Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
y_l
preference data
maximum likelihood
final LM

https://arxiv.org/pdf/2305.18290.pdf

DPO

Through some clever math, the authors discovered you could derive the optimal policy model and therefore the loss function mathematically.

Together with asking what itself thinks the score of each response should be, the fine-tuned LLM adjusts based on the results.


Trained LM
Update weights
prompt + chosen
prompt + rejected
chosen score
rejected score
$R_{policy} = \frac{\text{chosen score}}{\text{rejected score}}$
$loss = -\log\left(\sigma\left(\beta \cdot \log\left(\frac{R_{policy}}{R_{reference}}\right)\right)\right)$
Frozen LM
chosen score
rejected score
$R_{reference} = \frac{\text{chosen score}}{\text{rejected score}}$

https://medium.com/@joaolages/direct-preference-optimization-dpo-622fc1f18707

Results

At relatively small scale (6B parameters), DPO works maybe even a little better than traditional PPO / RLHF.

Very promising! Free gains!

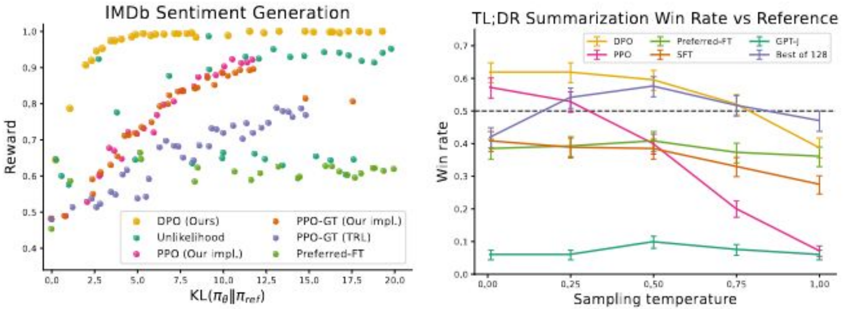


Figure 2: **Left.** The frontier of expected reward vs KL to the reference policy. DPO provides the highest expected reward for all KL values, demonstrating the quality of the optimization. **Right.** TL;DR summarization win rates vs. human-written summaries, using GPT-4 as evaluator. DPO exceeds PPO's best-case performance on summarization, while being more robust to changes in the sampling temperature.

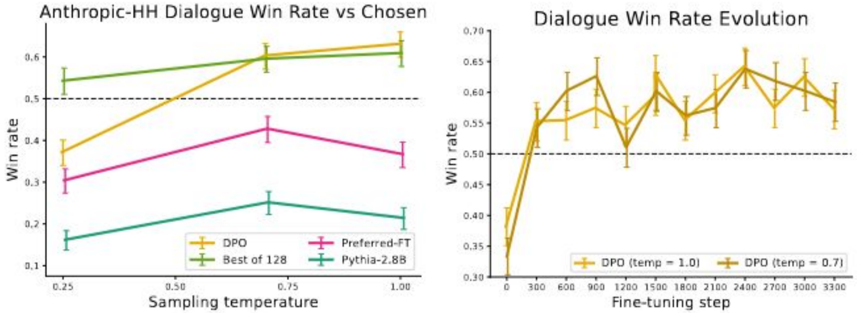
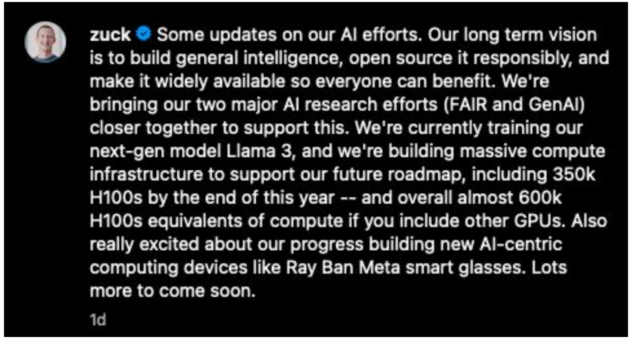


Figure 3: **Left.** Win rates computed by GPT-4 for Anthropic-HH one-step dialogue; DPO is the only method that improves over chosen summaries in the Anthropic-HH test set. **Right.** Win rates for different sampling temperatures over the course of training. DPO's improvement over the dataset labels is fairly stable over the course of training for different sampling temperatures.

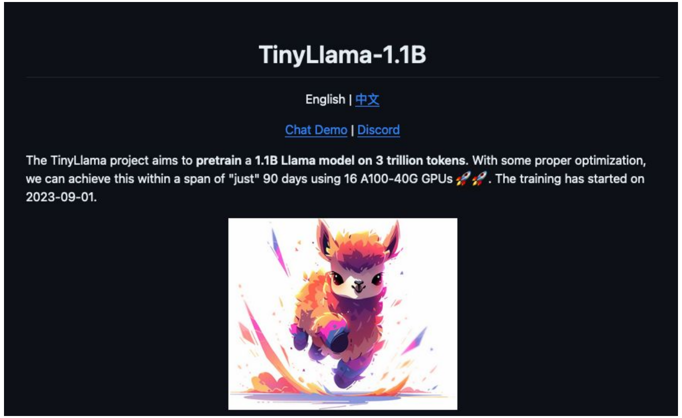
Why so exciting?

- We still have so much to discover!
- It’s beautifully succinct.
- Optimizations like these lower the barrier to LLMs.
  - GPU-rich vs GPU-poor
  - What else can we do to make AI/ML accessible?



zuck Some updates on our AI efforts. Our long term vision is to build general intelligence, open source it responsibly, and make it widely available so everyone can benefit. We're bringing our two major AI research efforts (FAIR and GenAI) closer together to support this. We're currently training our next-gen model Llama 3, and we're building massive compute infrastructure to support our future roadmap, including 350k H100s by the end of this year -- and overall almost 600k H100s equivalents of compute if you include other GPUs. Also really excited about our progress building new AI-centric computing devices like Ray Ban Meta smart glasses. Lots more to come soon.

1d



TinyLlama-1.1B

English | 中文

Chat Demo | Discord

The TinyLlama project aims to **pretrain a 1.1B Llama model on 3 trillion tokens**. With some proper optimization, we can achieve this within a span of "just" 90 days using 16 A100-40G GPUs 🚀🚀. The training has started on 2023-09-01.

Resources

- DPO Paper
- Andrew Ng explains DPO much more elegantly
- Practical Simplified DPO by João Lages
- Mathematical Explanation of DPO by Pakhapoom Sarapat
- RLHF video and blog post by Huggingface

Thank You

If this was helpful, please share with your network, follow, subscribe, all the good things.

Roger Lam

https://www.linkedin.com/in/lam-roger/

https://www.lamroger.com/