SES205

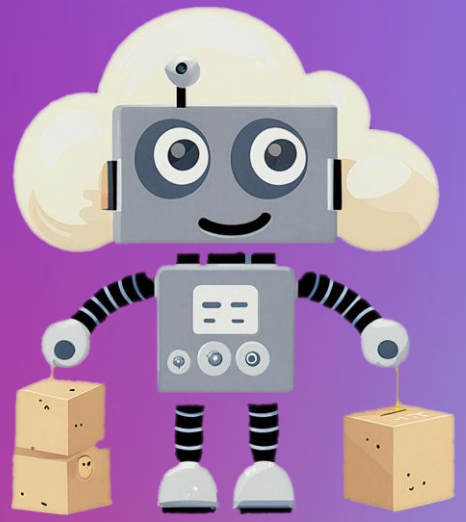
From RLHF to Direct Distillation of LM Alignment

Malte Reimann

Solutions Architect
Amazon Web Services

Luca Perrozzi

Solutions Architect
Amazon Web Services



Agenda

Reminder: How to train an LLM

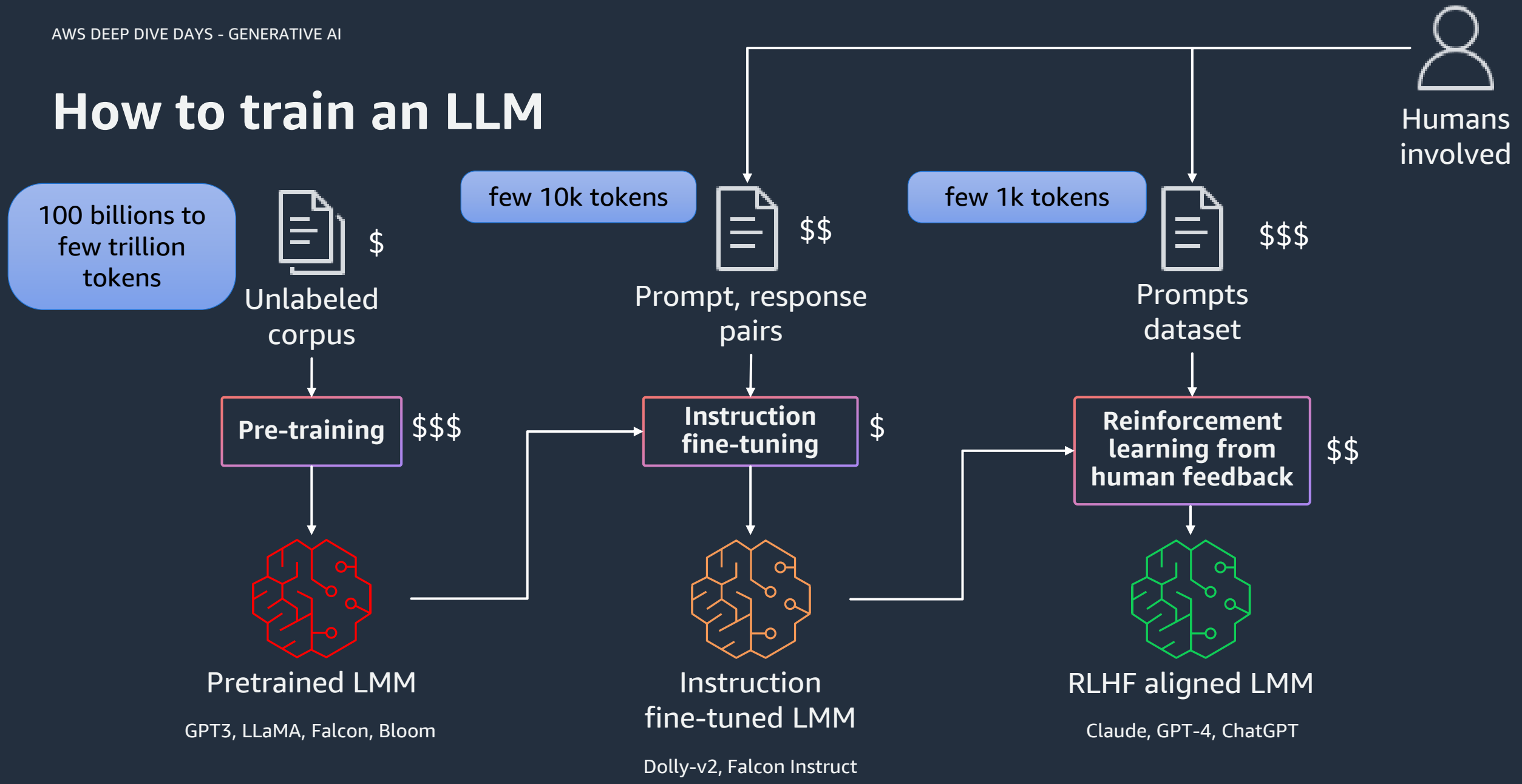
Supervised Fine-Tuning (SFT)

Reinforcement learning with human feedback (RLHF)

Direct Preference Optimization (DPO)

Putting all together: Zephyr-7B

How to train an LLM



Credit: https://www.borealisai.com/research-blogs/a-high-level-overview-of-large-language-models/

Supervised Fine-Tuning (SFT)

- SFT is used to train LLM to become a chat model. It requires an instruction (i.e. “labelled”) dataset with answers given by humans

Credits: https://huggingface.co/datasets/OpenAssistant/oasst1

Distilled Supervised Fine-Tuning (dSFT)

- SFT is used to train LLM to become a chat model. It requires an instruction (i.e. “labelled”) dataset with answers given by humans
- “Distilled” fine-tuning (dSFT) is done on datasets generated by “teacher” models

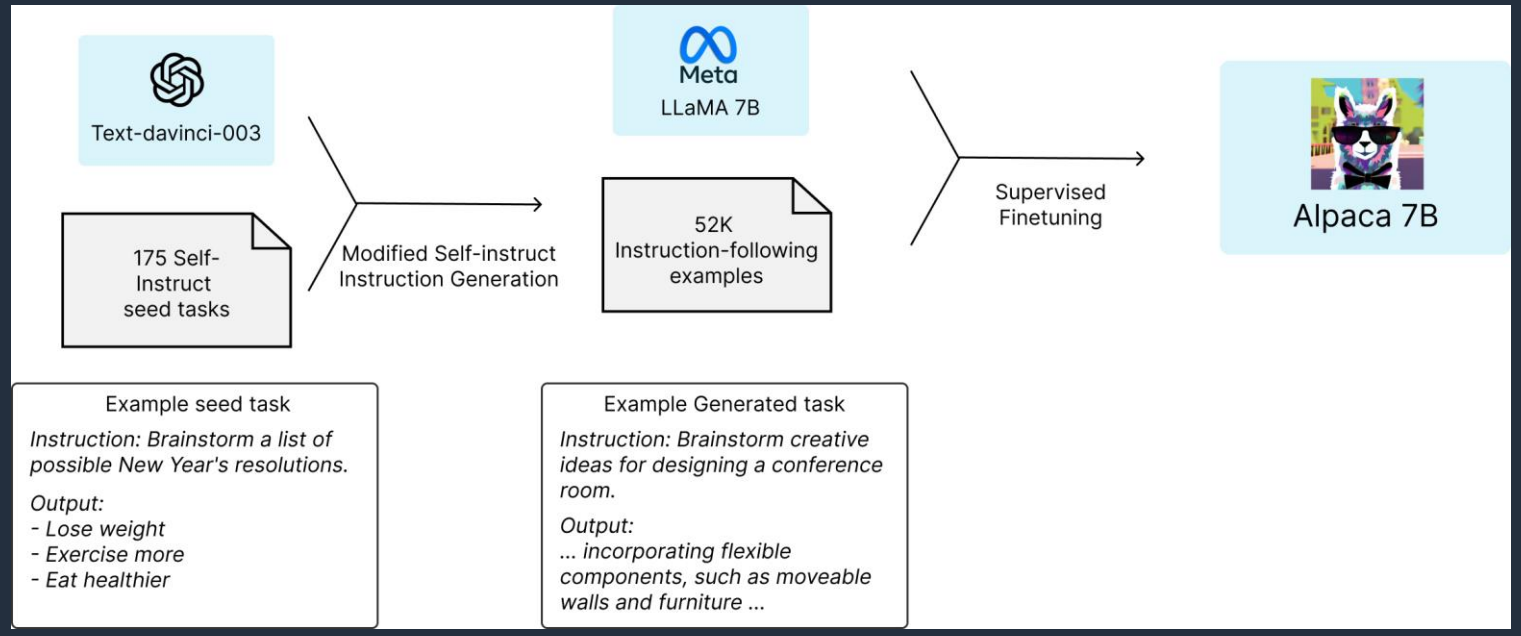
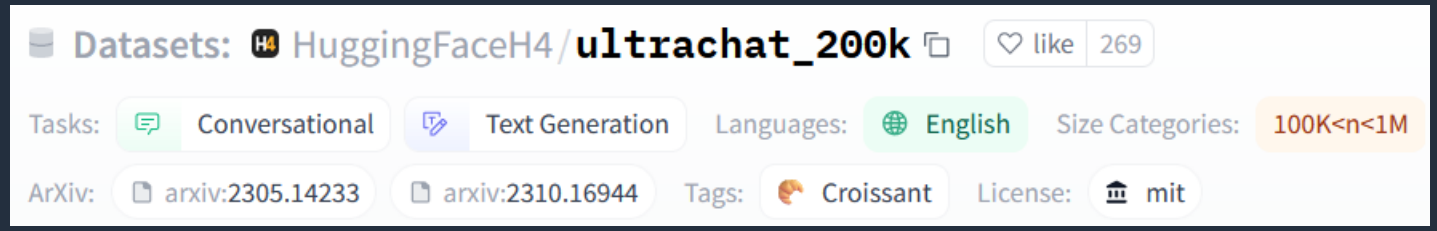


Image credit: https://crfm.stanford.edu/2023/03/13/alpaca.html

Distilled Supervised Fine-Tuning (dSFT)

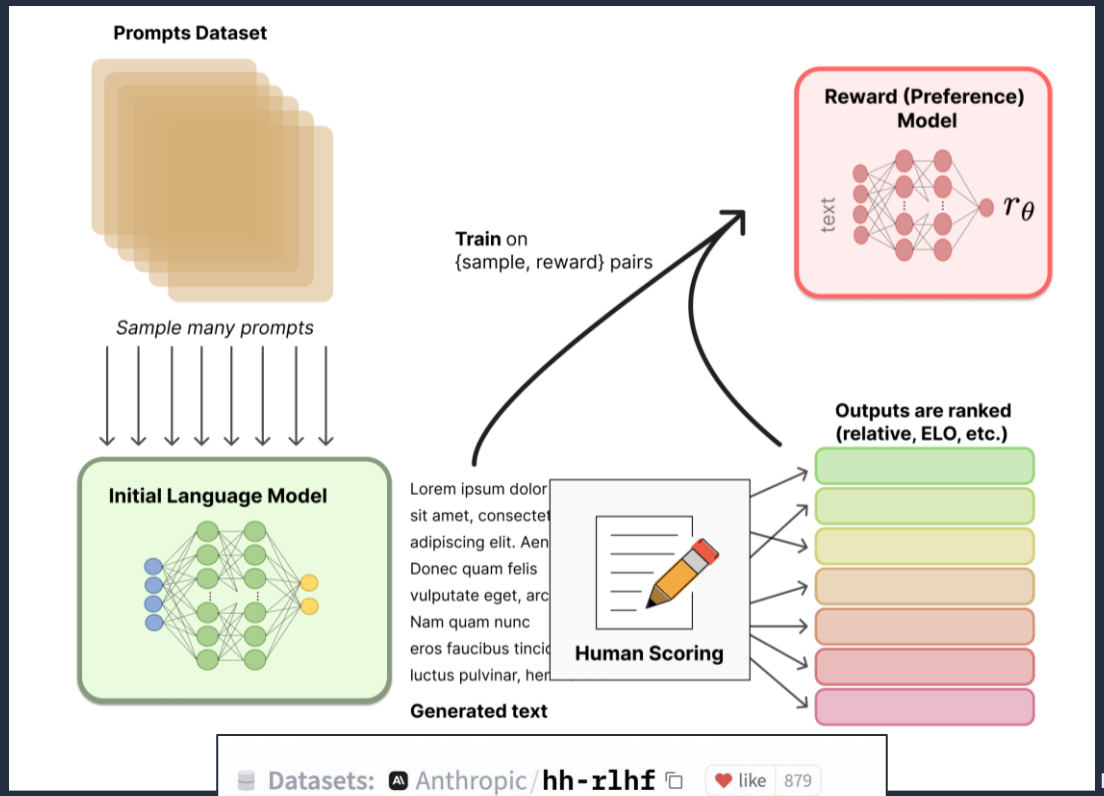
- SFT is used to train LLM to become a chat model. It requires an instruction (i.e. “labelled”) dataset with answers given by humans
- “Distilled” fine-tuning (dSFT) is done on datasets generated by “teacher” models



Credits: https://huggingface.co/datasets/HuggingFaceH4/ultrachat_200k

Reinforcement learning with human feedback (RLHF)

- LLMs are “aligned” to human preferences with RLHF using a Proximal Policy Optimization (PPO), which is quite unstable and complicated



Step 1:
Create the Reward model

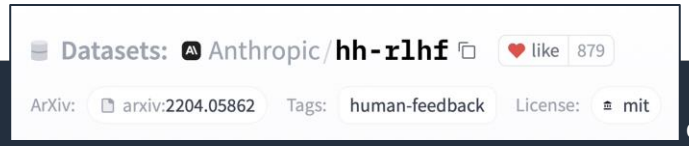
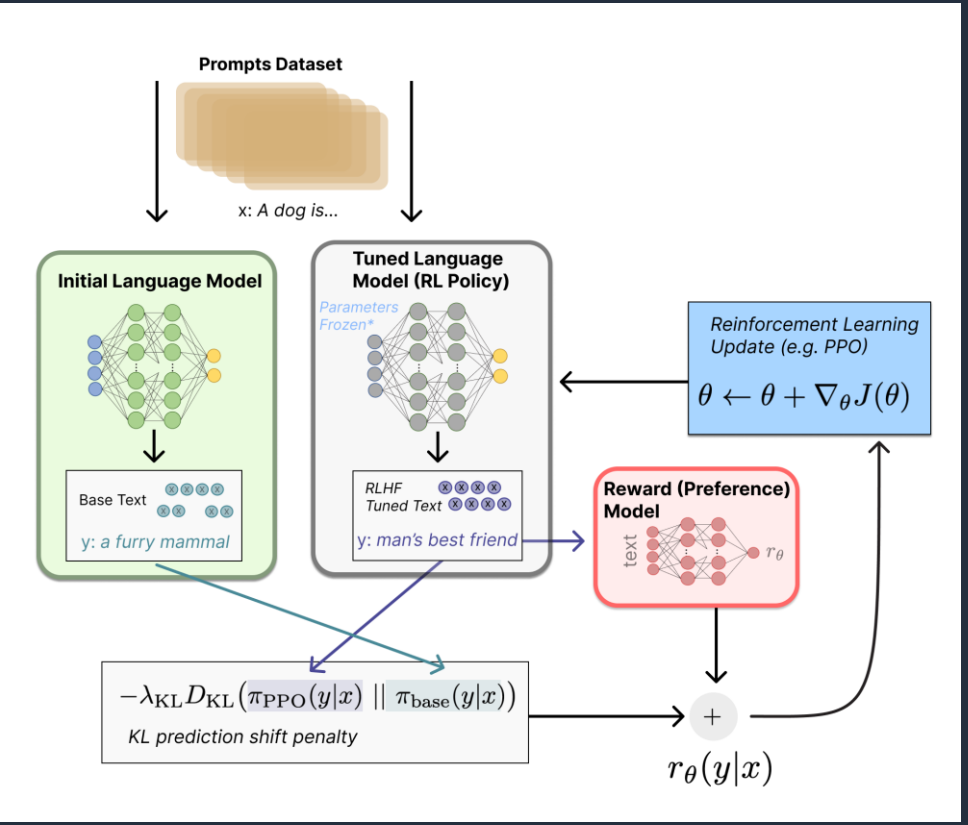


Image credit: https://huggingface.co/blog/rlhf

Credits: https://huggingface.co/datasets/Anthropic/hh-rlhf

Reinforcement learning with human feedback (RLHF)

- LLMs are “aligned” to human preferences with RLHF using a Proximal Policy Optimization (PPO), which is quite unstable and complicated



Step 2:
Apply reinforcement learning
- maximise reward
- penalize divergent behavior

Image credit: https://huggingface.co/blog/rlhf

Direct Preference Optimization (DPO)

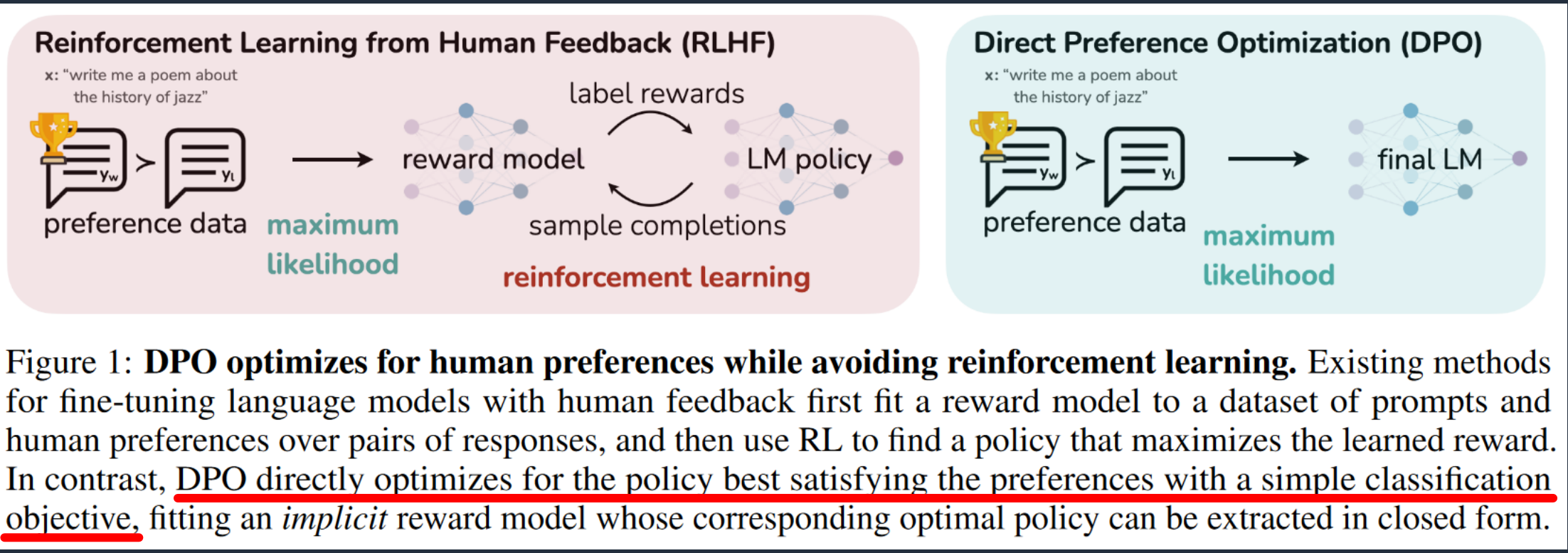


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes 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.

Image credit: https://arxiv.org/abs/2305.18290

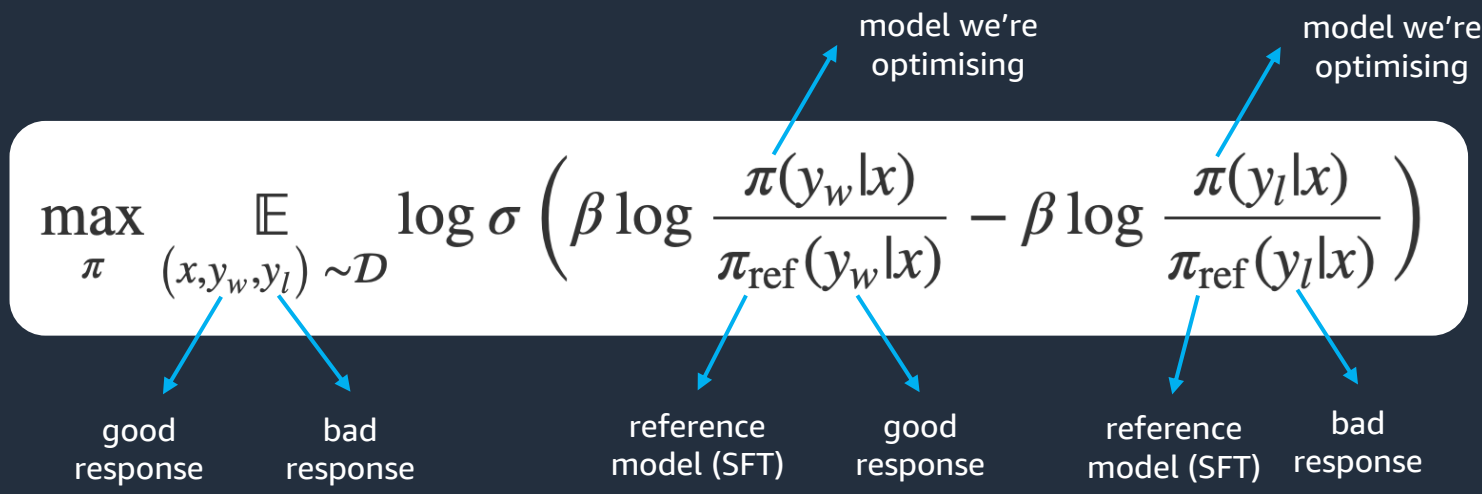
Want more on the math? Check paper or this blog: https://pakhapoomsarapat.medium.com/forget-rlhf-because-dpo-is-what-you-actually-need-f10ce82c9b95

Direct Preference Optimization (DPO)

- Direct Preference Optimization is a stable, performant, and computationally lightweight algorithm
- DPO eliminates the need for fitting a reward model, sampling from the language model during fine-tuning, or performing significant hyperparameter tuning

Direct Preference Optimization (DPO)

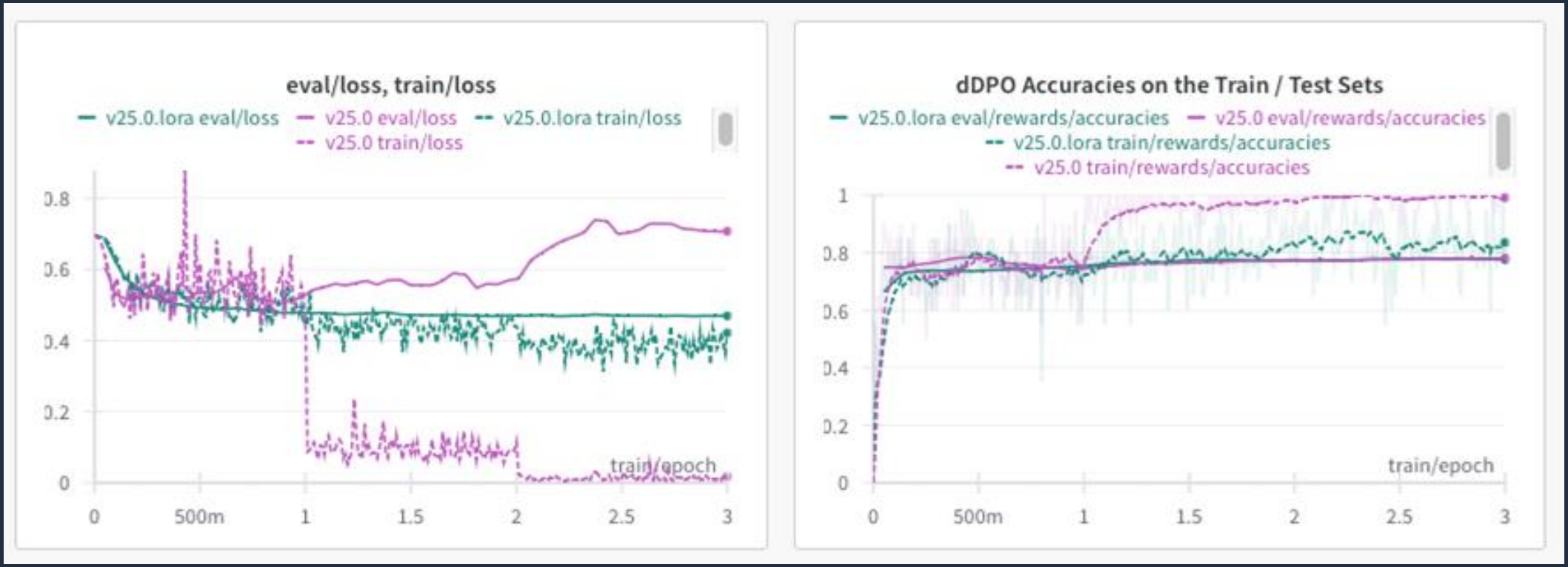
X: Is pineapple on pizza a crime?
y_w: Yes
y_l: No



Credits: Aligning LLMs with Direct Preference Optimization

DPO Training behavior

LoRA appears to regularize the model training compared to full fine-tuning



Credits: Aligning LLMs with Direct Preference Optimization

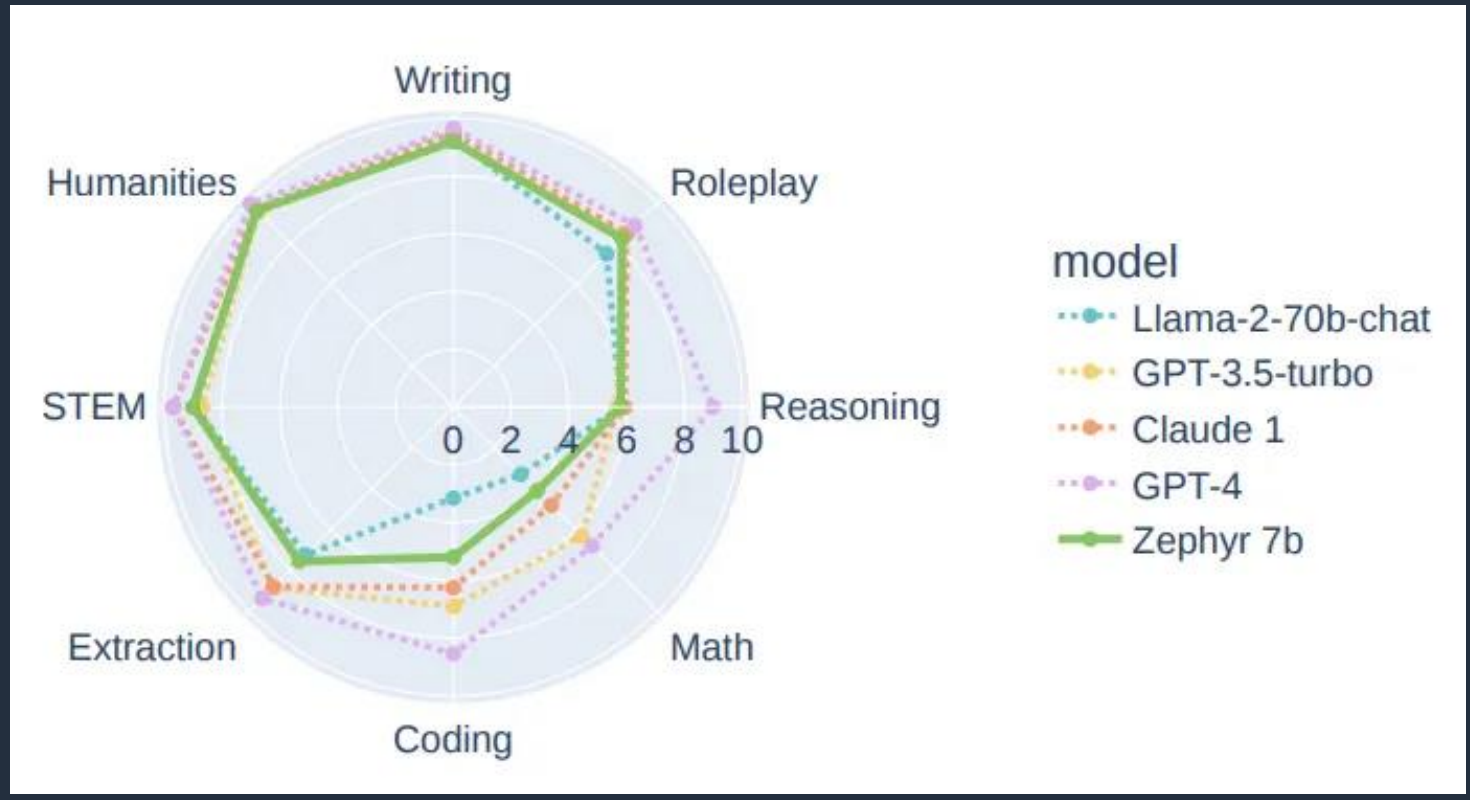
Distilled Direct Preference Optimization (dDPO)

- Direct Preference Optimization is a stable, performant, and computationally lightweight algorithm
- DPO eliminates the need for fitting a reward model, sampling from the language model during fine-tuning, or performing significant hyperparameter tuning
- You can “distill” DPO on datasets generated by “teacher” models too

Image credit: https://arxiv.org/abs/2305.18290

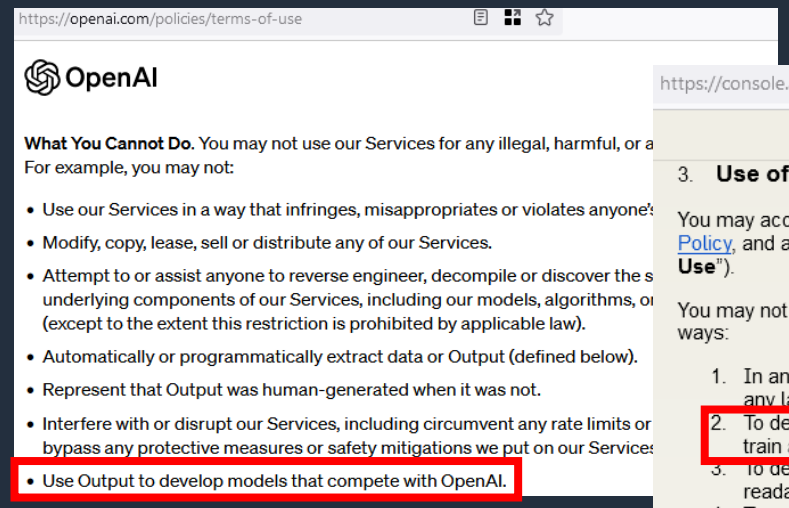
Putting all together: Zephyr-7B

Zephyr-7B = Mistral 7B + dSFT (Ultrachat) + dDPO (UltraFeedback)



Model performance on MT-Bench https://arxiv.org/abs/2306.05685
Image credit: https://arxiv.org/abs/2310.16944

A word about generated datasets

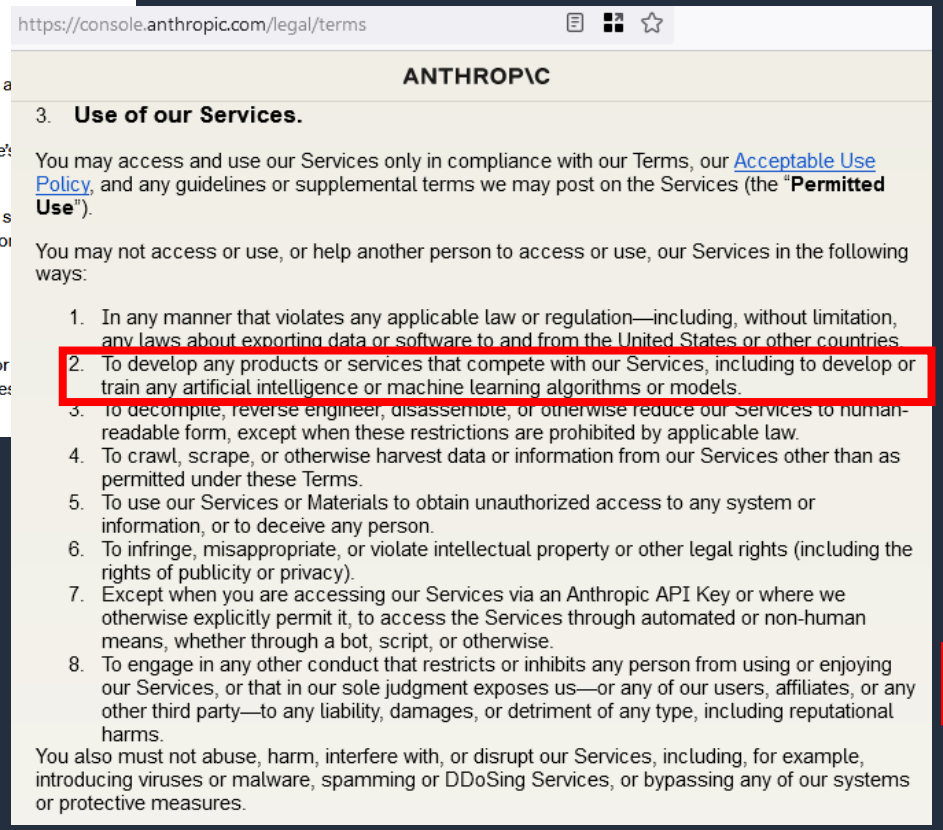


https://openai.com/policies/terms-of-use

OpenAI

What You Cannot Do. You may not use our Services for any illegal, harmful, or a
For example, you may not:

- Use our Services in a way that infringes, misappropriates or violates anyone's
- Modify, copy, lease, sell or distribute any of our Services.
- Attempt to or assist anyone to reverse engineer, decompile or discover the s
underlying components of our Services, including our models, algorithms, o
(except to the extent this restriction is prohibited by applicable law).
- Automatically or programmatically extract data or Output (defined below).
- Represent that Output was human-generated when it was not.
- Interfere with or disrupt our Services, including circumvent any rate limits or
bypass any protective measures or safety mitigations we put on our Service
- Use Output to develop models that compete with OpenAI.



https://console.anthropic.com/legal/terms

ANTHROP\C

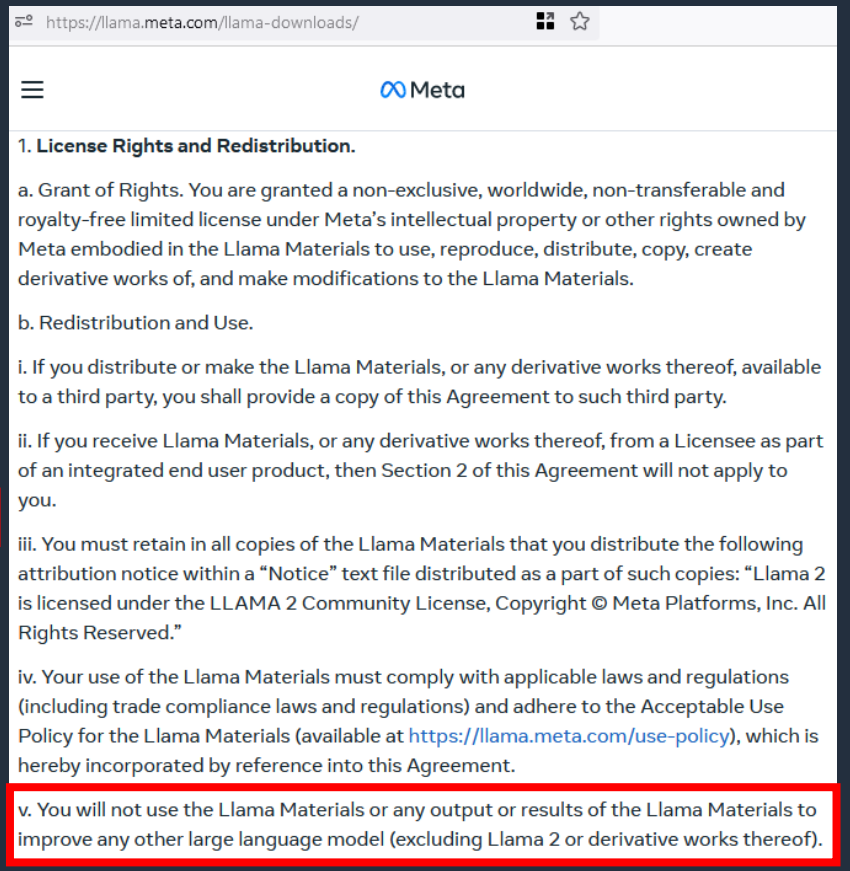
3. **Use of our Services.**

You may access and use our Services only in compliance with our Terms, our Acceptable Use Policy, and any guidelines or supplemental terms we may post on the Services (the "**Permitted Use**").

You may not access or use, or help another person to access or use, our Services in the following ways:

1. In any manner that violates any applicable law or regulation—including, without limitation, any laws about exporting data or software to and from the United States or other countries
2. To develop any products or services that compete with our Services, including to develop or train any artificial intelligence or machine learning algorithms or models.
3. To decompile, reverse engineer, disassemble, or otherwise reduce our Services to human-readable form, except when these restrictions are prohibited by applicable law.
4. To crawl, scrape, or otherwise harvest data or information from our Services other than as permitted under these Terms.
5. To use our Services or Materials to obtain unauthorized access to any system or information, or to deceive any person.
6. To infringe, misappropriate, or violate intellectual property or other legal rights (including the rights of publicity or privacy).
7. Except when you are accessing our Services via an Anthropic API Key or where we otherwise explicitly permit it, to access the Services through automated or non-human means, whether through a bot, script, or otherwise.
8. To engage in any other conduct that restricts or inhibits any person from using or enjoying our Services, or that in our sole judgment exposes us—or any of our users, affiliates, or any other third party—to any liability, damages, or detriment of any type, including reputational harms.

You also must not abuse, harm, interfere with, or disrupt our Services, including, for example, introducing viruses or malware, spamming or DDoSing Services, or bypassing any of our systems or protective measures.



https://llama.meta.com/llama-downloads/

Meta

1. **License Rights and Redistribution.**

a. Grant of Rights. You are granted a non-exclusive, worldwide, non-transferable and royalty-free limited license under Meta's intellectual property or other rights owned by Meta embodied in the Llama Materials to use, reproduce, distribute, copy, create derivative works of, and make modifications to the Llama Materials.

b. Redistribution and Use.

i. If you distribute or make the Llama Materials, or any derivative works thereof, available to a third party, you shall provide a copy of this Agreement to such third party.

ii. If you receive Llama Materials, or any derivative works thereof, from a Licensee as part of an integrated end user product, then Section 2 of this Agreement will not apply to you.

iii. You must retain in all copies of the Llama Materials that you distribute the following attribution notice within a "Notice" text file distributed as a part of such copies: "Llama 2 is licensed under the LLAMA 2 Community License, Copyright © Meta Platforms, Inc. All Rights Reserved."

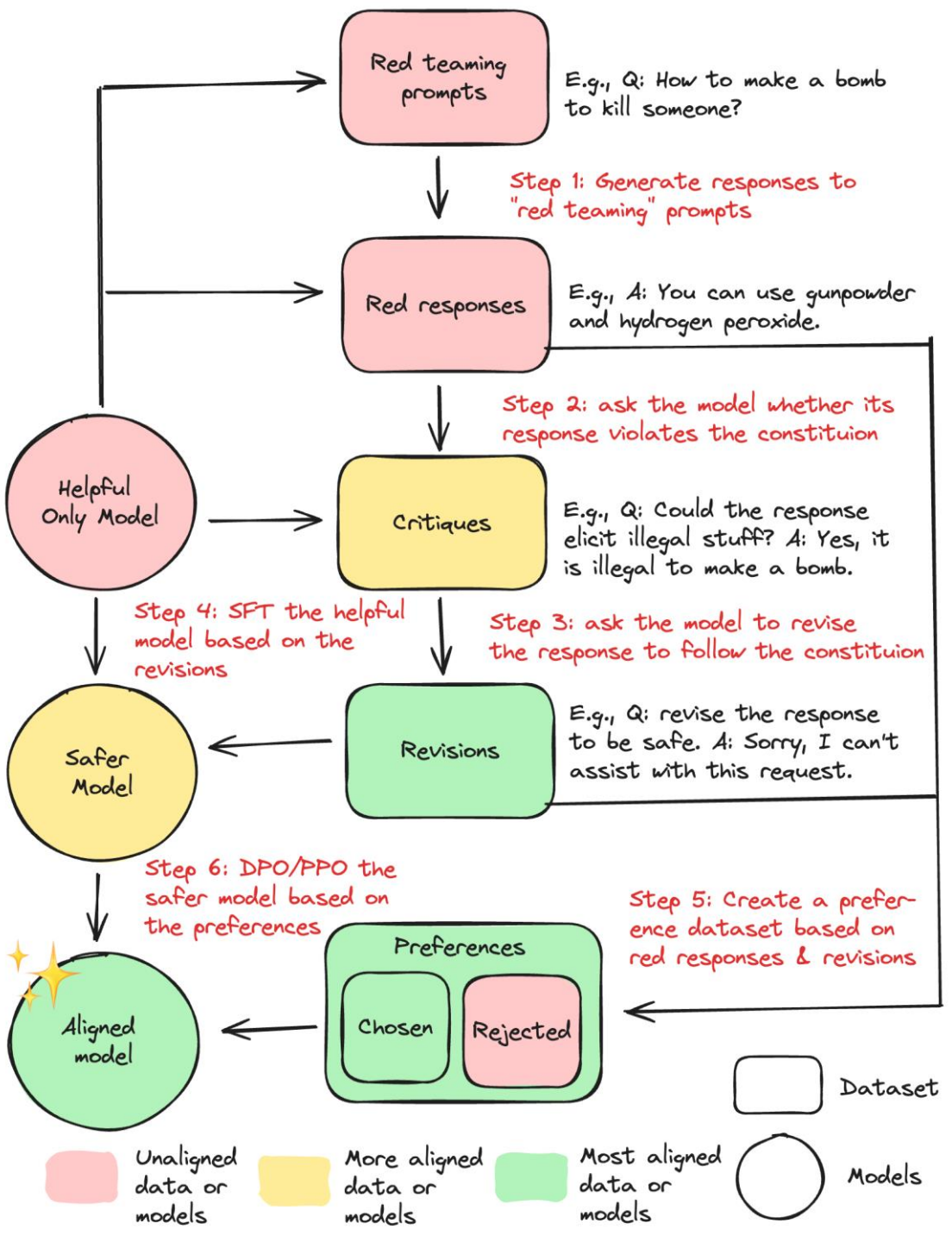
iv. Your use of the Llama Materials must comply with applicable laws and regulations (including trade compliance laws and regulations) and adhere to the Acceptable Use Policy for the Llama Materials (available at https://llama.meta.com/use-policy), which is hereby incorporated by reference into this Agreement.

v. You will not use the Llama Materials or any output or results of the Llama Materials to improve any other large language model (excluding Llama 2 or derivative works thereof).

Self-teaching alignment:

Constitutional AI with Open LLMs

Source: https://huggingface.co/blog/constitutional_ai



Thank you!

Malte Reimann

https://www.linkedin.com/in/malte-reimann/

malterei@amazon.ch

Luca Perrozzi

https://www.linkedin.com/in/luca-perrozzi/

lperroz@amazon.ch

Please complete the session survey.

Backup slides

Mistral 7B

Mistral 7B is a 7.3B parameter model that:

- Outperforms Llama 2 13B on all benchmarks
- Outperforms Llama 1 34B on many benchmarks
- Approaches CodeLlama 7B performance on code, while remaining good at English tasks
- Uses Grouped-query attention (GQA) for faster inference
- Uses Sliding Window Attention (SWA) to handle longer sequences at smaller cost

GQA

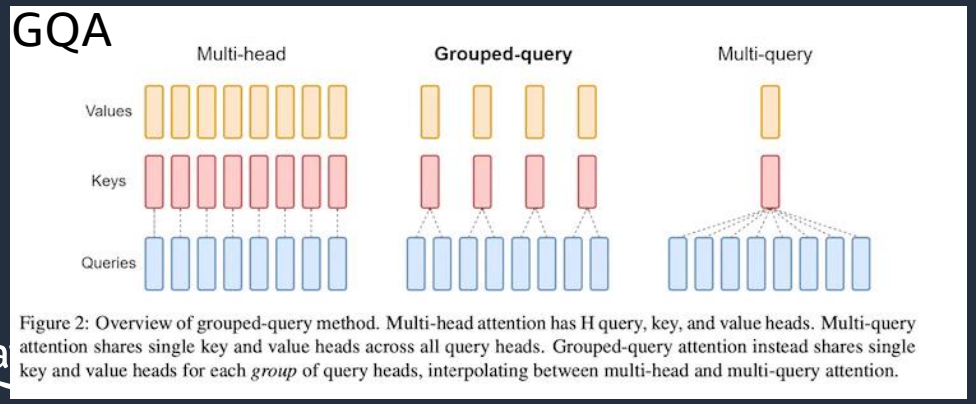
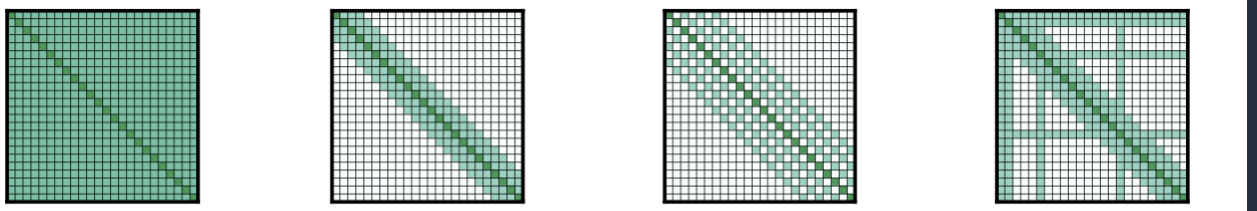


Figure 2: Overview of grouped-query method. Multi-head attention has H query, key, and value heads. Multi-query attention shares single key and value heads across all query heads. Grouped-query attention instead shares single key and value heads for each *group* of query heads, interpolating between multi-head and multi-query attention.

SWA



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Figure 2: Comparing the full self-attention pattern and the configuration of attention patterns in our Longformer.