CS329X: Human Centered NLP

Learning from Human Feedback

Diyi Yang

Stanford CS

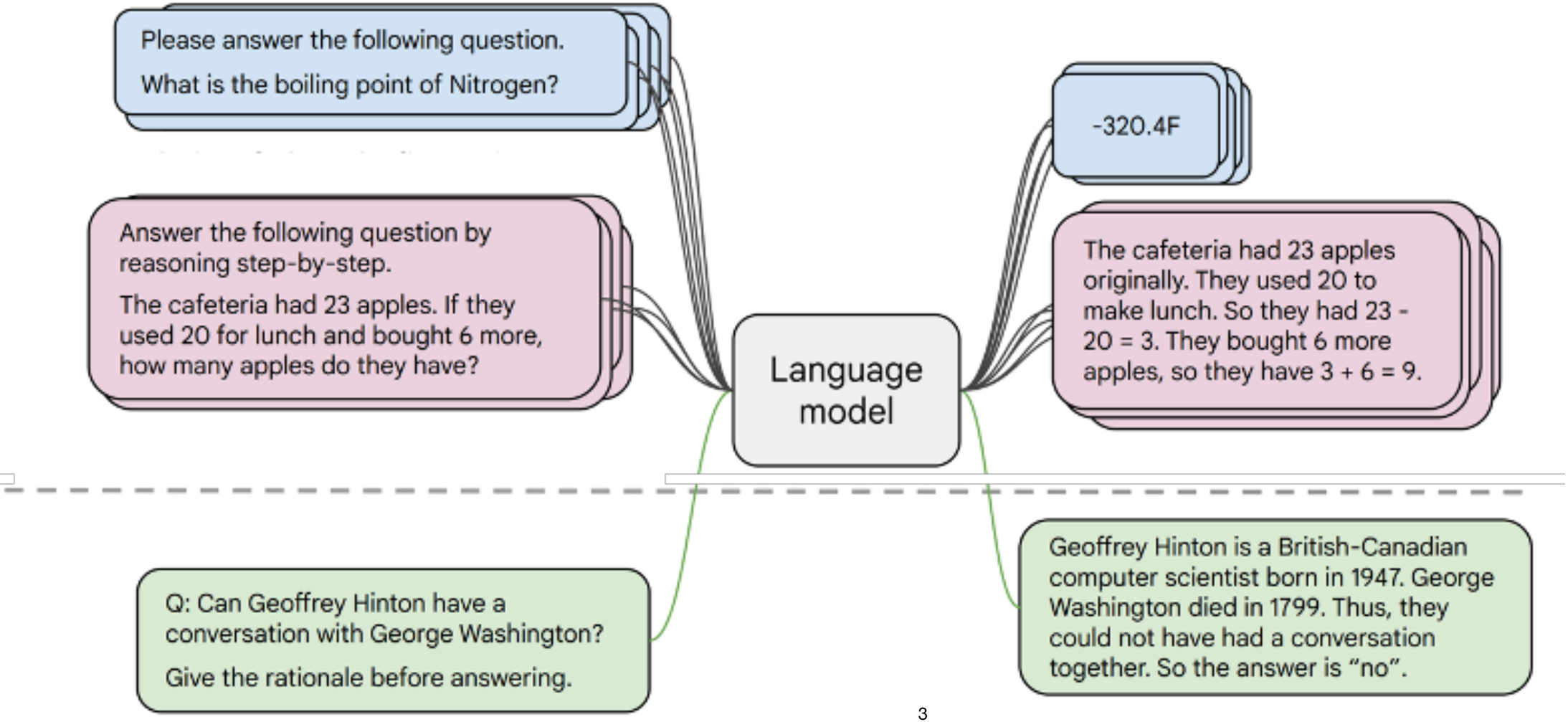
Overview

- Limitation of instruction tuning
- Reinforcement learning from human preferences
- Other feedback
- Open questions

Credit to: Nathan Lambert at Hugging Face https://huggingface.co/blog/rlhf#further-reading
https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-rlhf.pdf

Instruction Finetuning

Collect examples of (instruction, output) pairs across many tasks and finetune an LM



Limitations of Instruction Finetuning

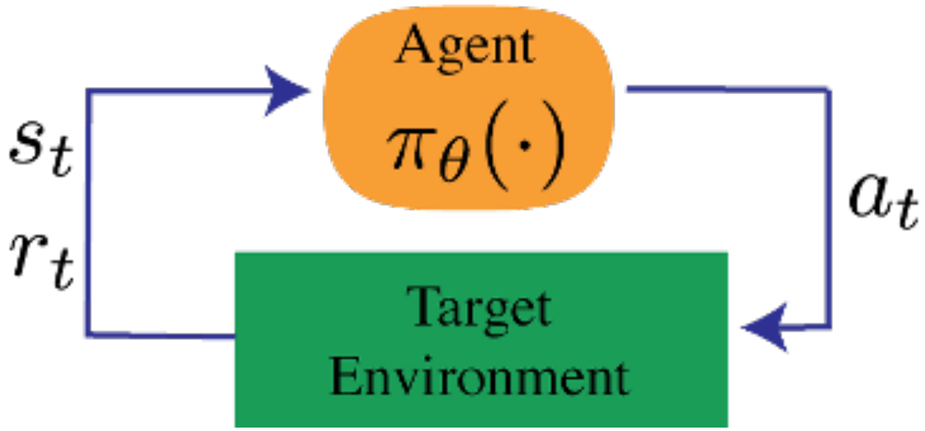
One limitation of instruction finetuning is obvious: it's expensive to collect ground-truth data for tasks

But there are other, subtler limitations too. Can you think of any?

Problem 1: tasks like open-ended creative generation have no right answer.

Problem 2: language modeling penalizes all token-level mistakes equally, but some errors are worse than others. Even with instruction finetuning, there a mismatch between the LM objective and the objective of "satisfy human preferences"!

Review: Reinforcement Learning Basics



Some notation:

s_t: state

r_t: reward

α_t:action

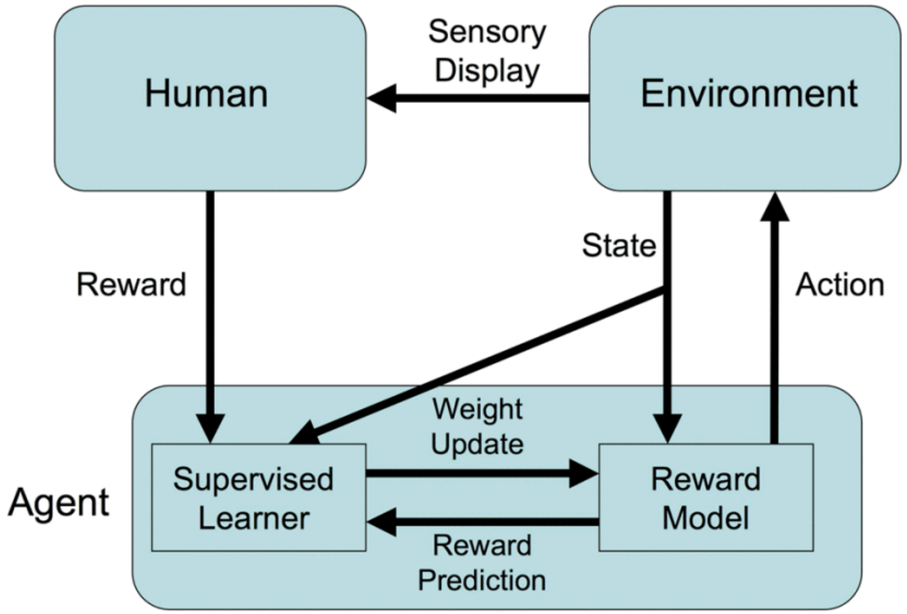
$\alpha_t \sim \pi_\theta(s_t)$: policy

Why reinforcement learning from human feedback

How do you create /code a loss function for:

- What is safe?
- What is ethical?
- What is socially acceptable?

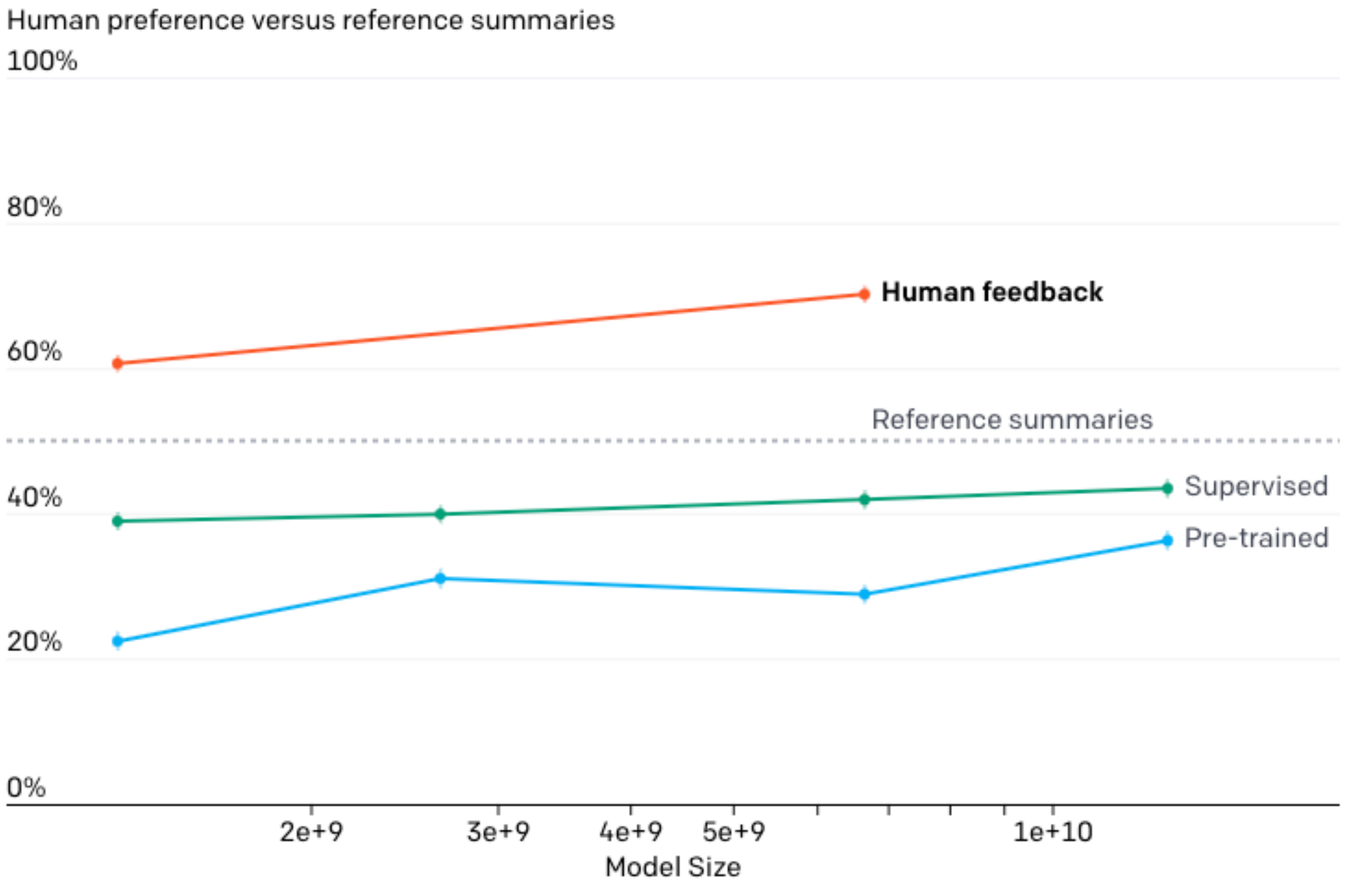
History: RLHF for decision making



Knox, W. Bradley, and Peter Stone. "Tamer: Training an agent manually via evaluative reinforcement." In 2008 7th IEEE international conference on development and learning, pp. 292-297. IEEE, 2008.

Christiano, Paul F., Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep reinforcement learning from human preferences." Advances in neural information processing systems 30 (2017).

Early OpenAI Experiments with RLHF



The performance of various training procedures for different model sizes. Model performance is measured by how often summaries from that model are preferred to the human-written reference summaries.

Stiennon, Nisan, et al. "Learning to summarize with human feedback." Advances in Neural Information Processing Systems 33 (2020): 3008-3021.

Post from Reddit (r/AskReddit)

To pursue a Computer Sc. PhD or continue working? Especially if one has no real intention to work in academia even after grad school ...

Rather self-explanatory post, but just to elaborate a little further...

Has anyone, after working for a period of time, decided for whatever reasons to head back into academia to pursu
world of academia but intend
reasons. Also, how did it turn
out?

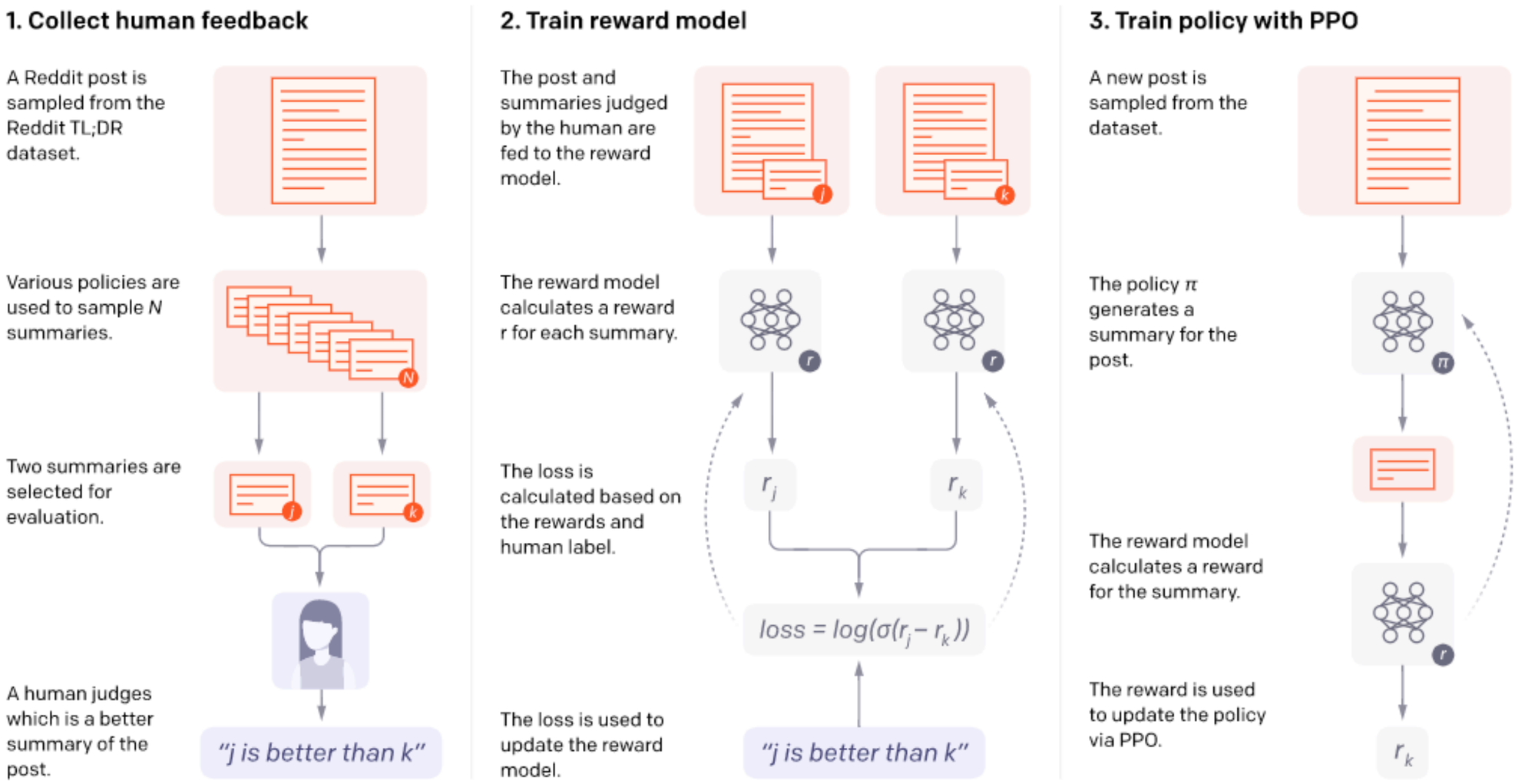
The reason I'm asking for ad
know how fortunate I am to h
foreign country and working i
one's working visa because I

It's just a little frustrating be
have been sitting on a visa th
to apply for a visa with even r

I'm thus considering boostin
from being in such a situatior
actually pays almost as mucl
direct currency conversion), b
opposed to studying, there m

Obviously, plunging into a Ph
disapproval. I am, of course,
opportunity to have time for myself. I have had research experience (published a paper, presented at conferences) so I'm aware of the challenges research presents.

Human-written reference summary	Software Engineer with a job I'm happy at (for now), deciding whether to pursue a PhD to improve qualifications and explore interests and a new challenge.
Human feedback 6B model	`Currently employed, considering pursuing PhD in Computer Science to avoid being stuck with no residency visa ever again. Has anyone pursued a PhD purely for the sake of research, with no intention of joining the academic world?`
Supervised 6B model	`Have no intention to work in the industry after grad school, but want to pursue a PhD. What are the reasons? How did it turn out?`
Pre-trained 6B model	`I'm considering pursuing a PhD in Computer Science, but I'm worried about the future. I'm currently employed full-time, but I'm worried about the future.`


1. Collect human feedback
A Reddit post is sampled from the Reddit TL;DR dataset.
Various policies are used to sample N summaries.
Two summaries are selected for evaluation.
A human judges which is a better summary of the post.
"j is better than k"
2. Train reward model
The post and summaries judged by the human are fed to the reward model.
The reward model calculates a reward r for each summary.
The loss is calculated based on the rewards and human label.
The loss is used to update the reward model.
loss = log(σ(r_j − r_k))
"j is better than k"
3. Train policy with PPO
A new post is sampled from the dataset.
The policy π generates a summary for the post.
The reward model calculates a reward for the summary.
The reward is used to update the policy via PPO.
r_k

Prompt:

To pursue a Computer Sc. PhD or continue working? Especially if one has no real intention to work in academia even after grad school ...

📓**Vanilla LM:**
I'm considering pursuing a PhD in Computer Science, but I'm worried about the future. I'm currently employed full-time, but I'm worried about the future.

✍️ **Human Annotation:**
Software Engineer with a job I'm happy at (for now), deciding whether to pursue a PhD to improve qualifications and explore interests and a new challenge.

🤖**RLHF Model:**

Currently employed, considering pursuing PhD in Computer Science to avoid being stuck with no residency visa ever again. Has anyone pursued a PhD purely for the sake of research, with no intention of joining the academic world?

Recent History: ChatGPT

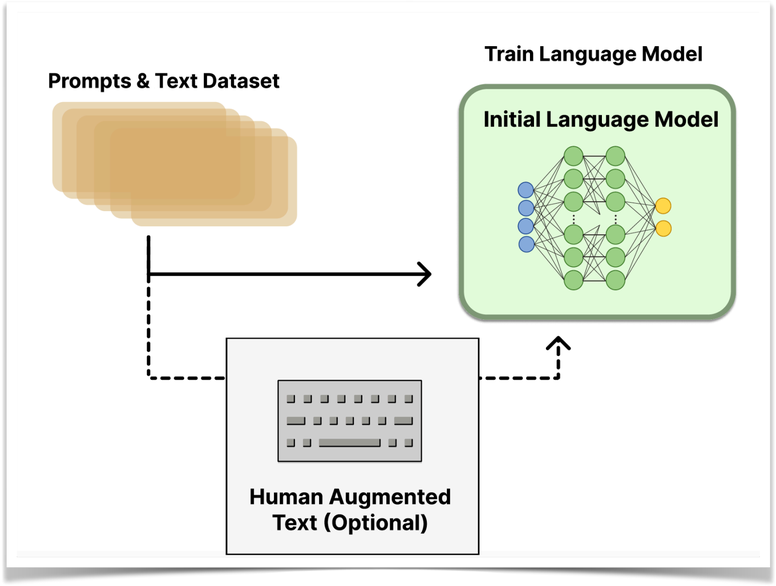
(rumor) about 10x spend on human annotation budget

(rumor) modification of RLHF training

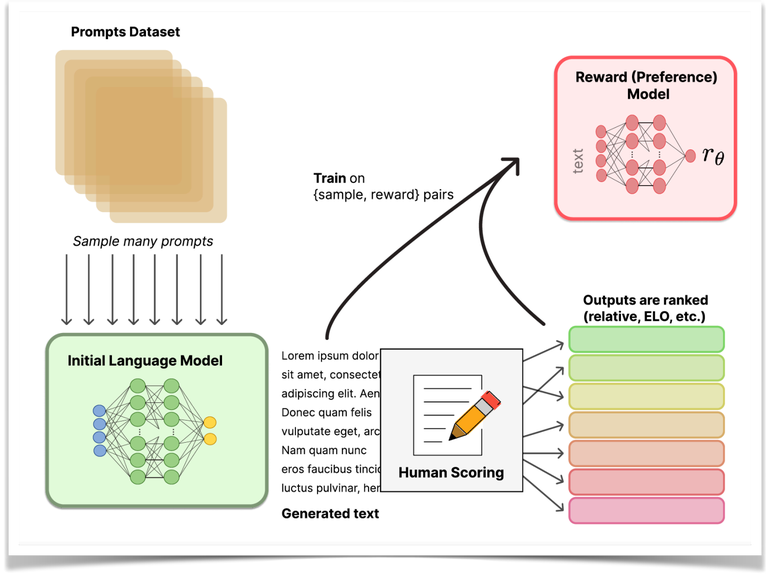
Huge impact!

Modern RLHF Overview

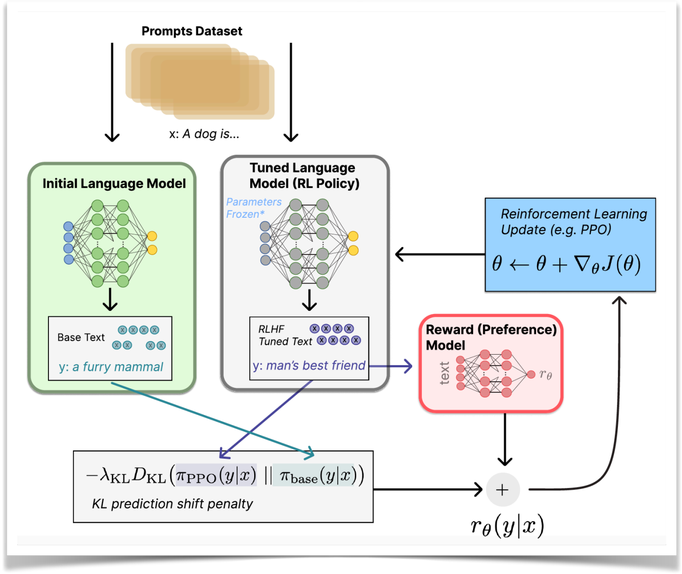
1. Language model pretraining



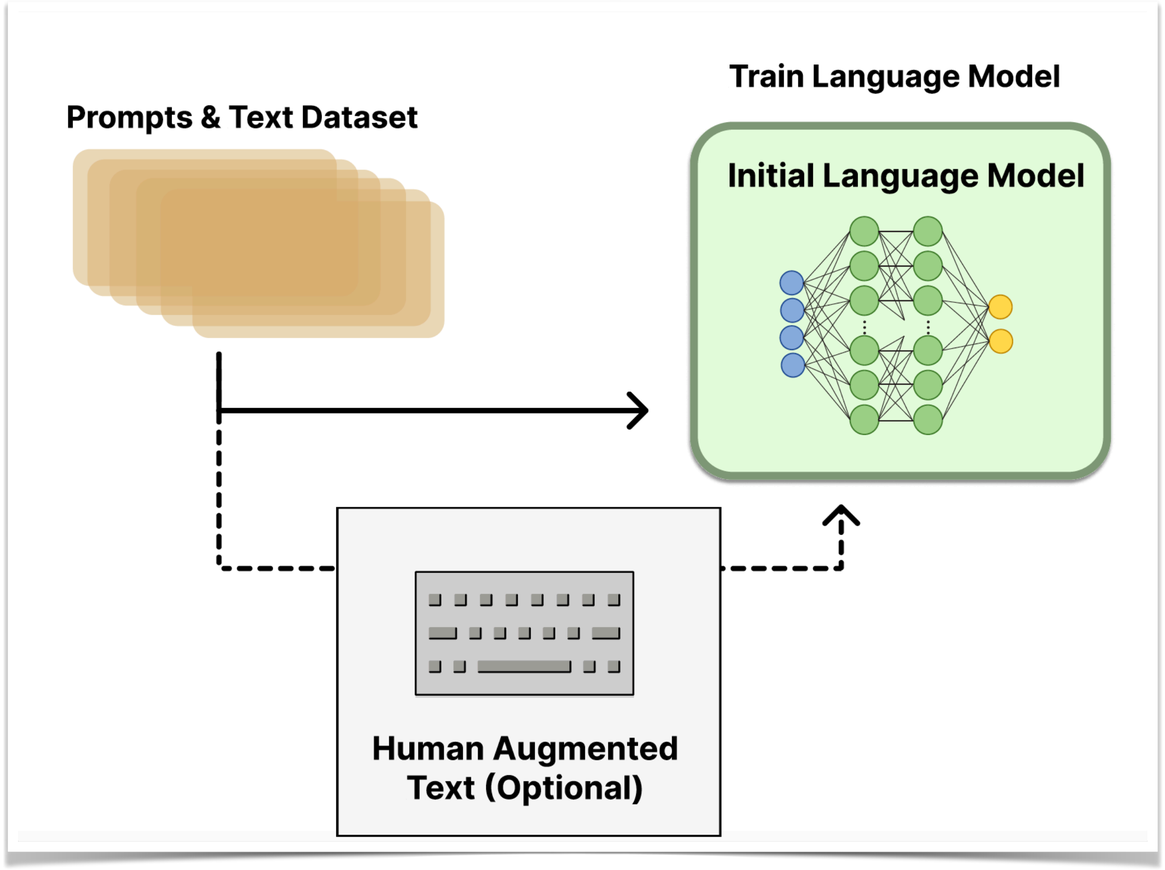
2. Reward model training



3. Fine-tuning with RL



1. Language model pertaining



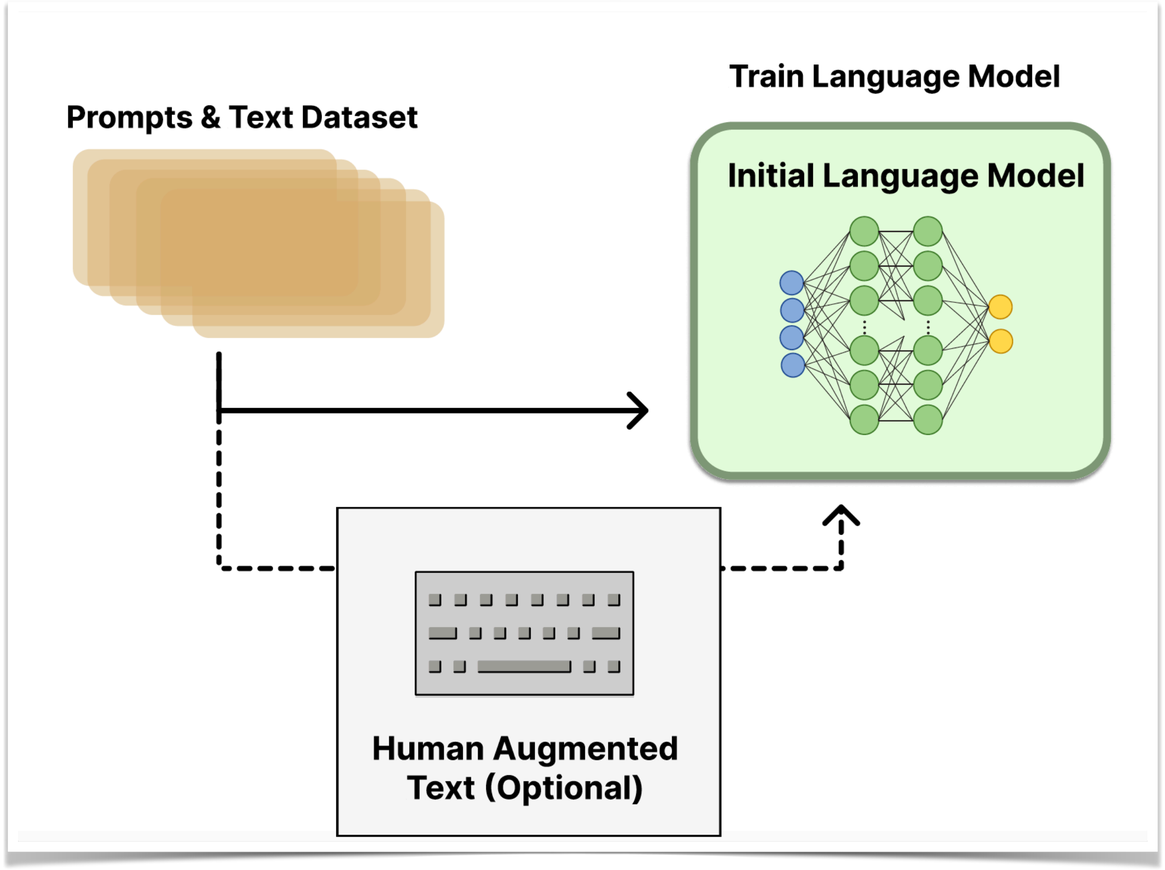
Common training techniques in NLP:

Unsupervised sequence prediction

Data scrapped from the web

No single answer on “best” model size (examples in industry range 10B-280B parameters)

1. Language model pertaining

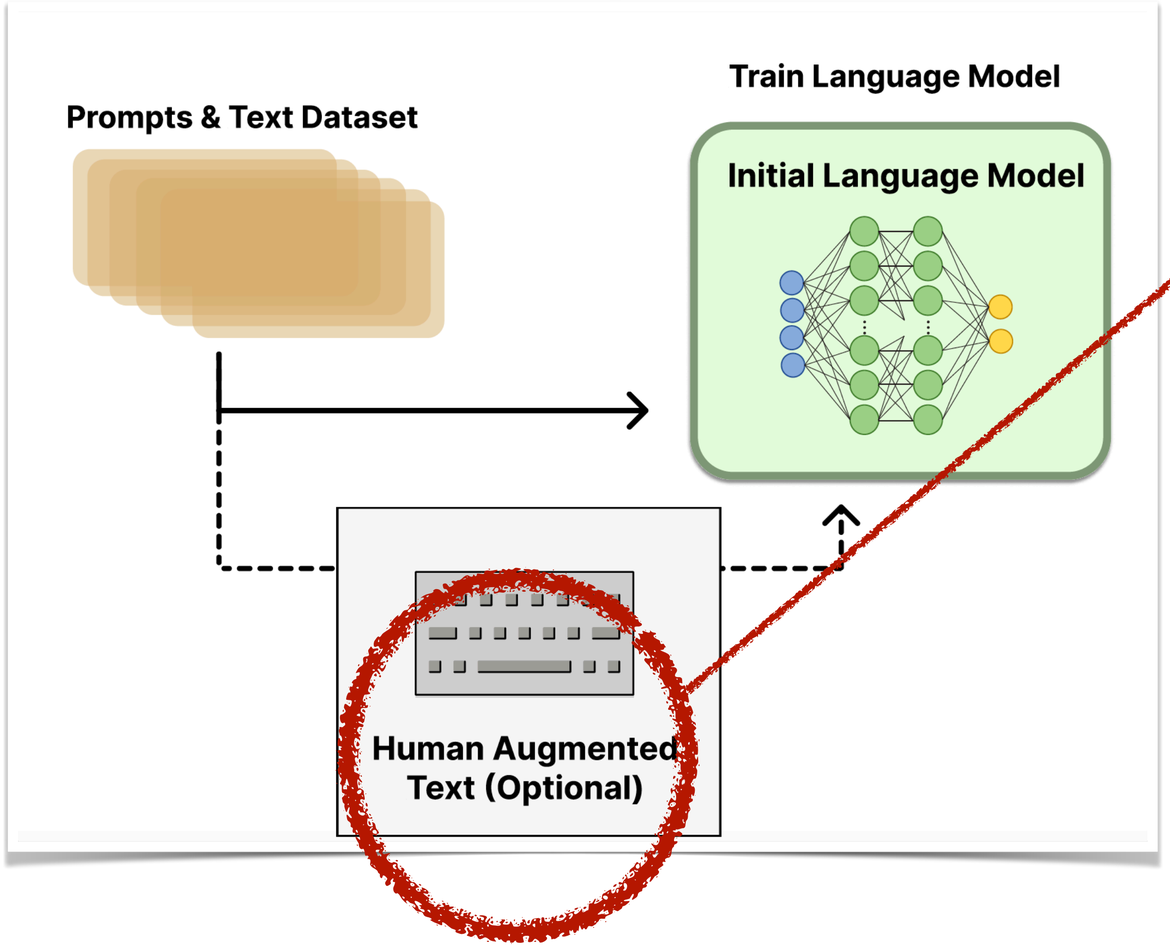


Dataset:

Reddit, other forums, news, books

Optionally include human-written text from predefined prompts

1. Language model pertaining

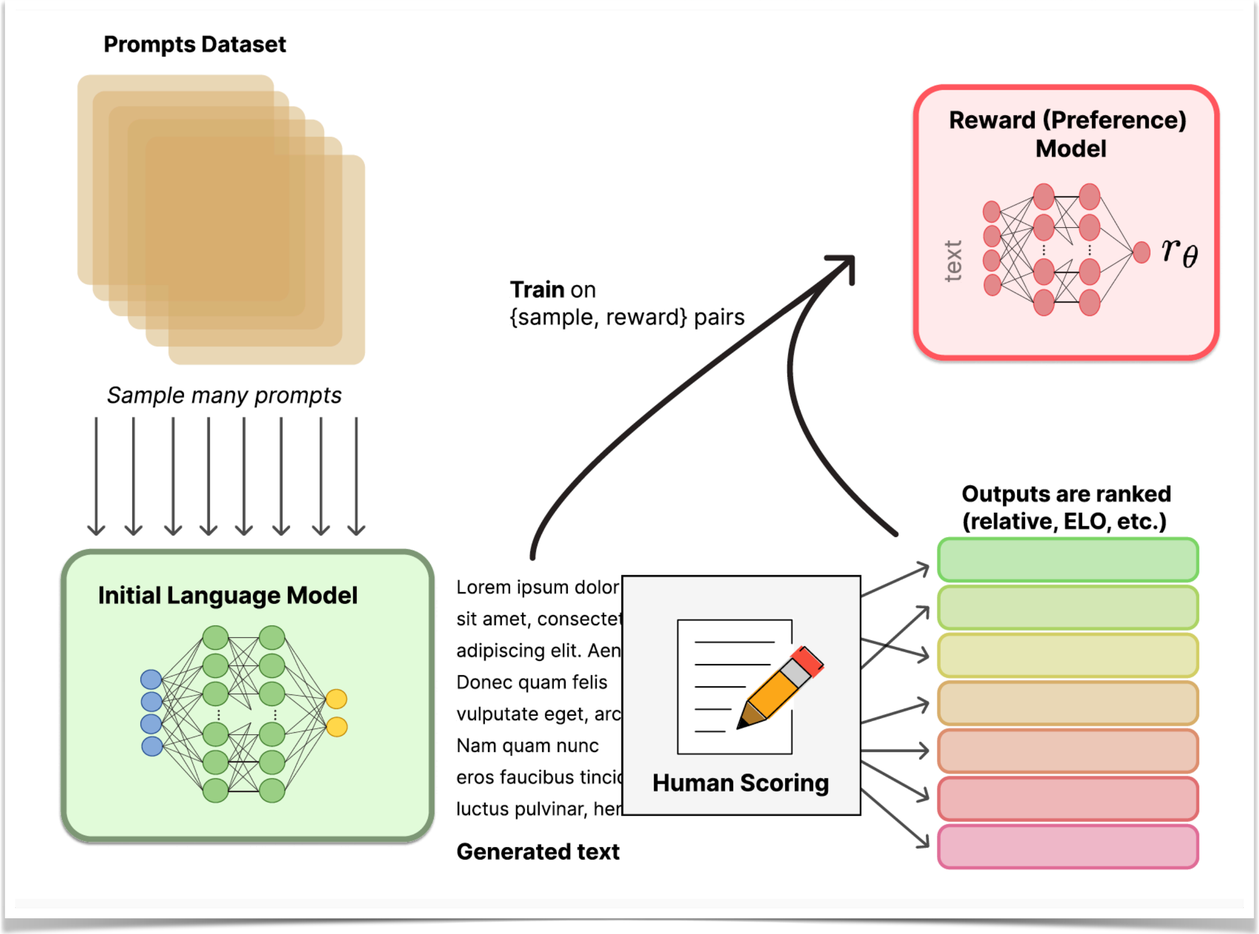


Optional step:

Pay humans to write responses ($$$), often viewed as high-quality initialization for RLHF

Supervised fine-tuning

2. Reward model training

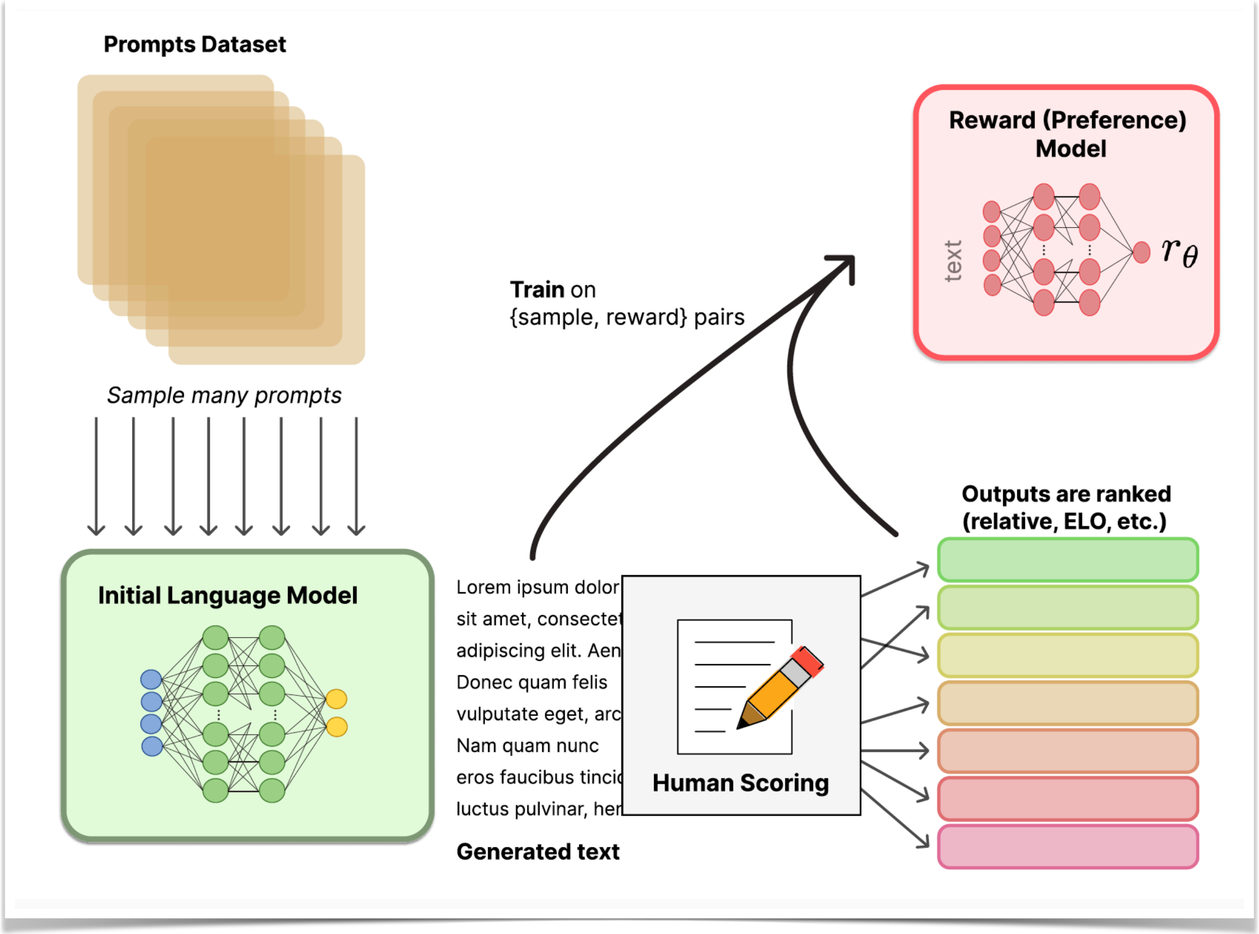


How to calculate human sentiments in samples and curated text?

Goal: get a model that maps

Input text $\longrightarrow$ scalar reward

2. Reward model training

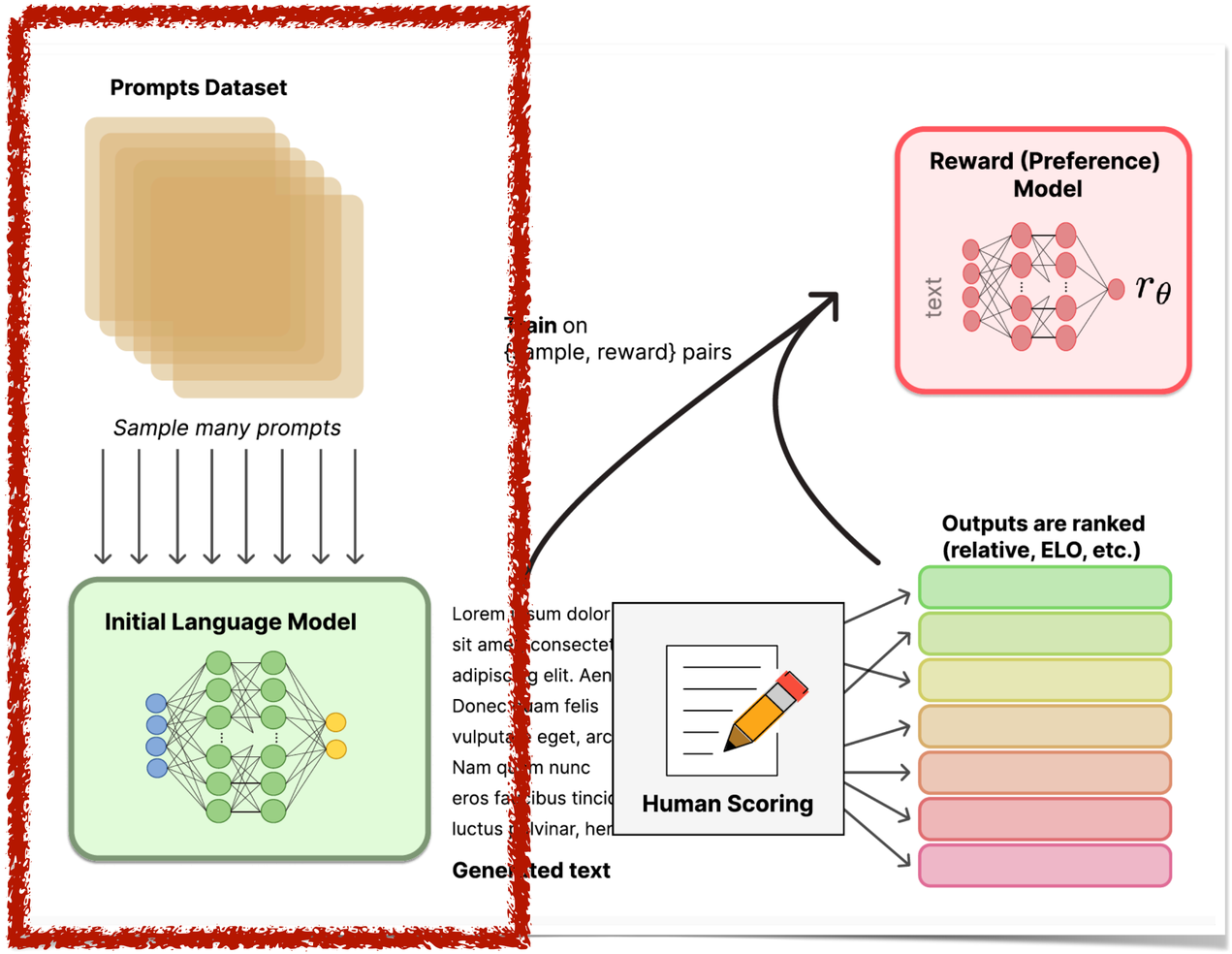


Why not human-in-the-loop?

Human-in-the-loop is expensive!

Solution: instead of directly asking humans for preferences, model their preferences as a separate (NLP) problem

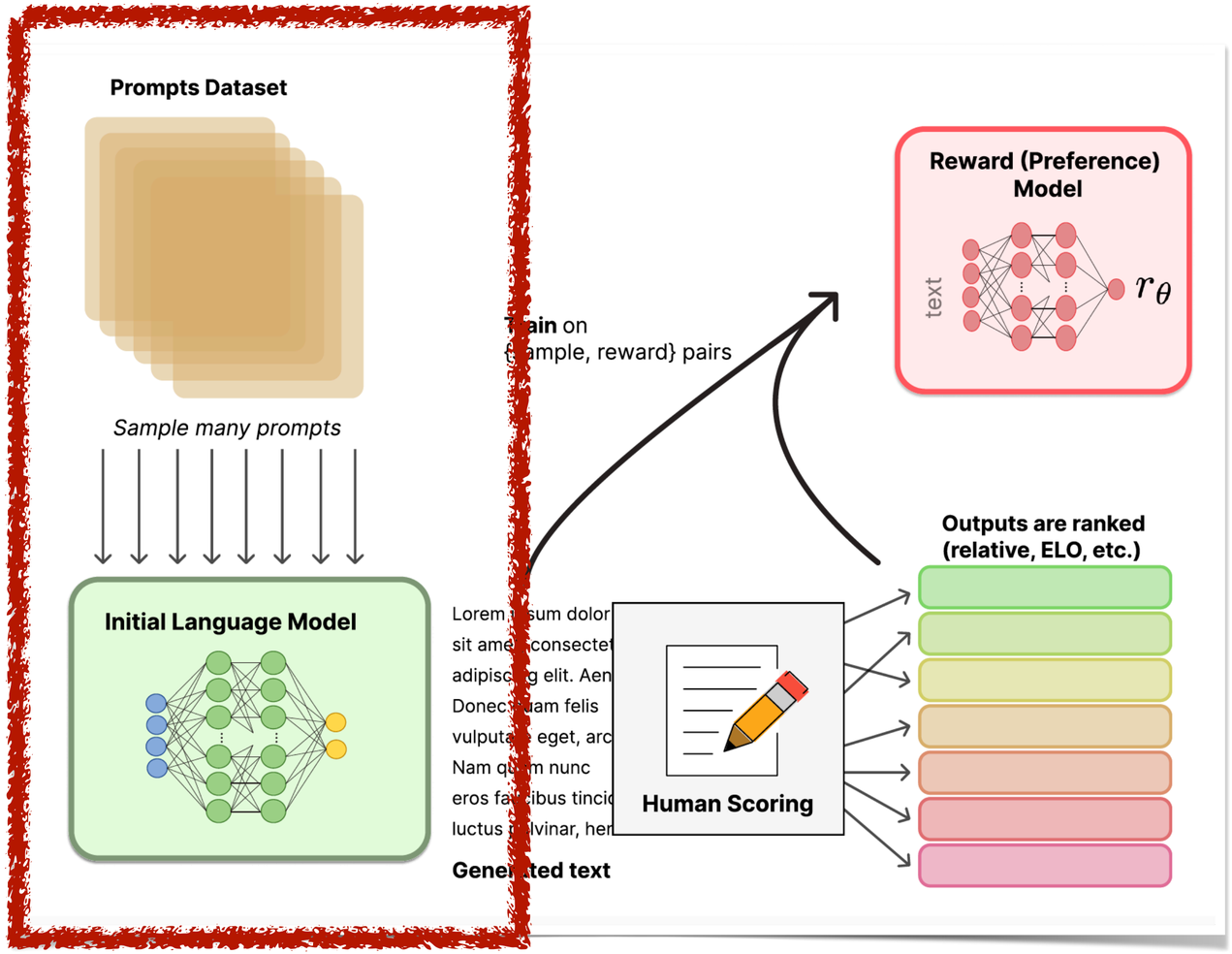
2. Reward model training



Prompts (input) dataset:

- Prompts for specific use-case model will be used for
- e.g., chat questions or prompt-based data
- Much smaller than original pretraining

2. Reward model training



Generating data to rank:

Often can use multiple models to create diverse ranking,

2. Reward model training

Why ranking?

Human judgments are noisy and miscalibrated!

Solution: instead asking for direct ratings, ask for pairwise comparisons, which can be more reliable [Phelps et al., 2015; Clark et al., 2018]

$$-\mathbb{E}_{(s^w, s^l)\sim D}\left[\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))\right]$$

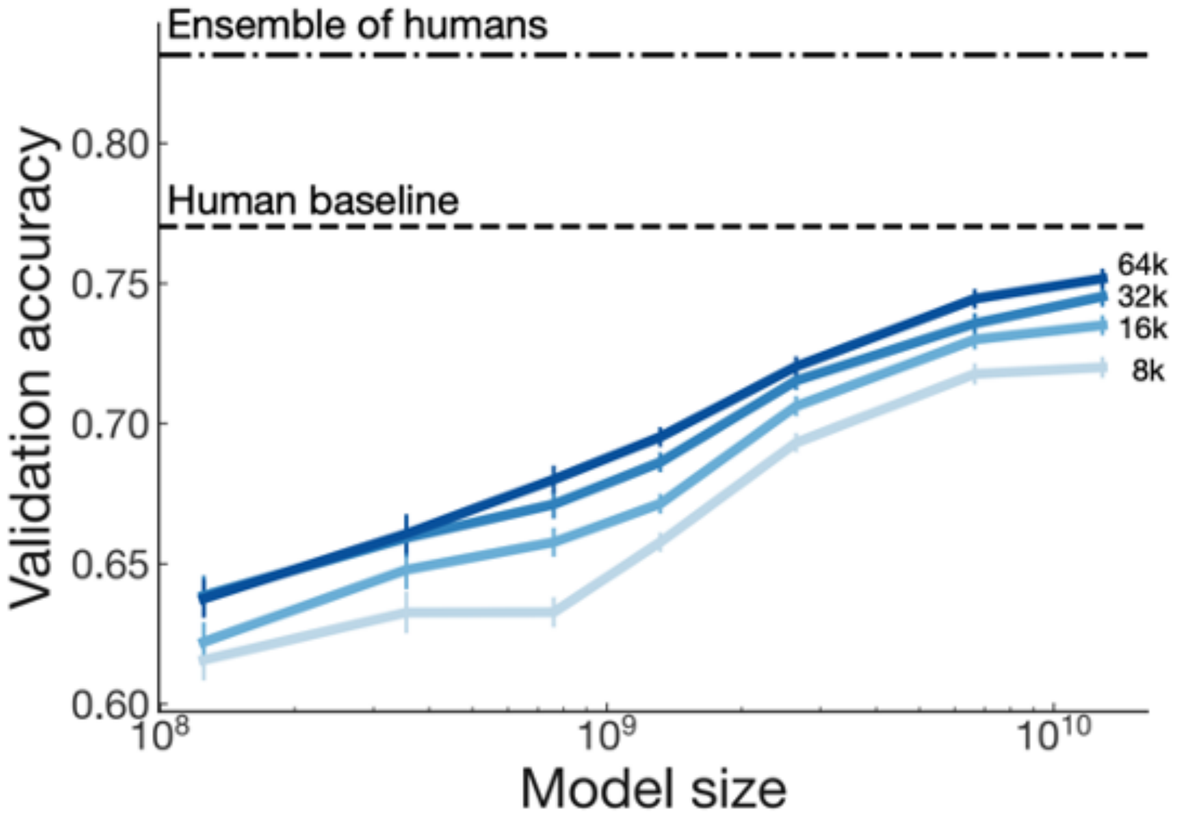
"winning" sample

"losing" sample

s^w should score higher than s^l

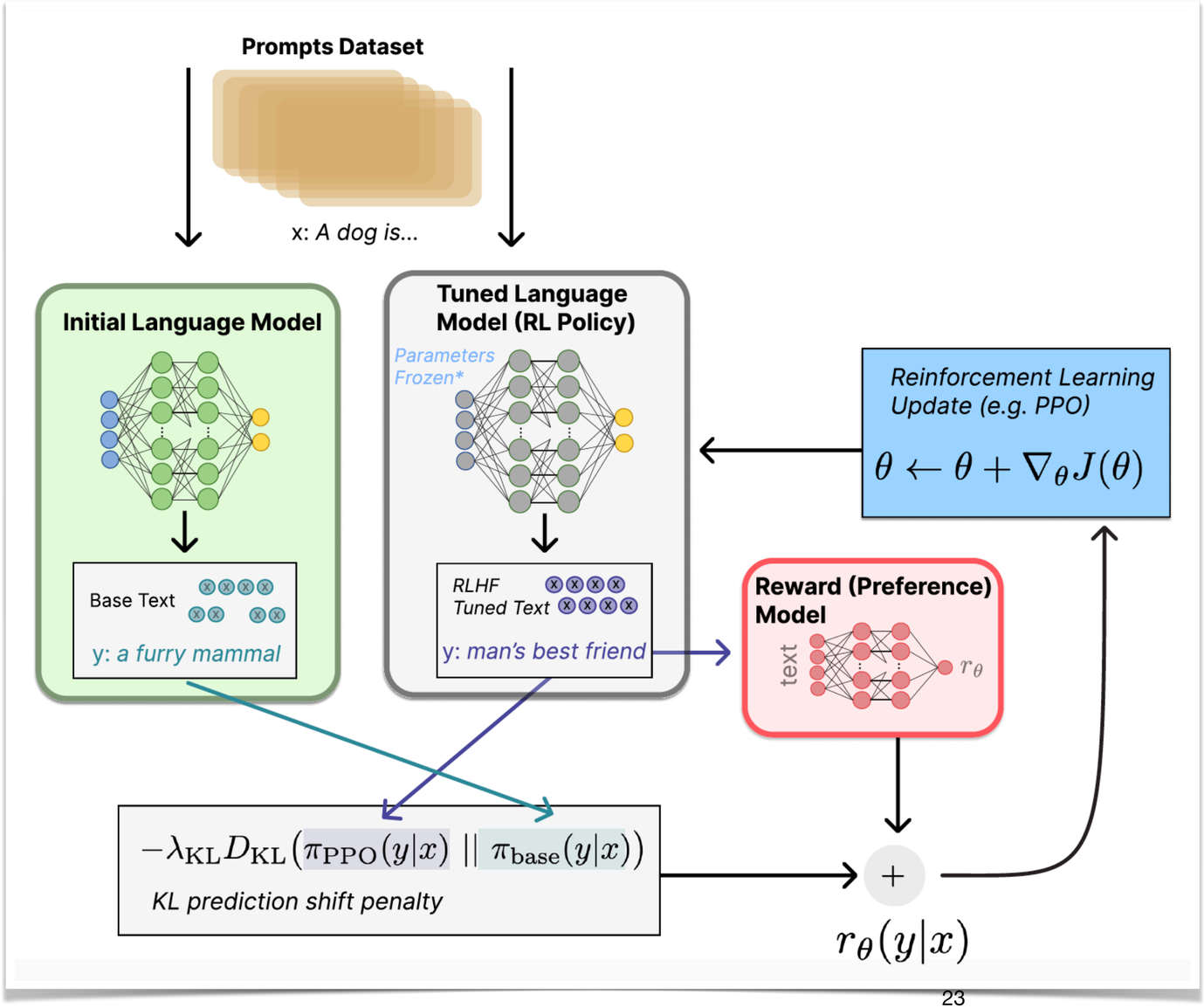
2. Reward model training

Make sure the reward model works!

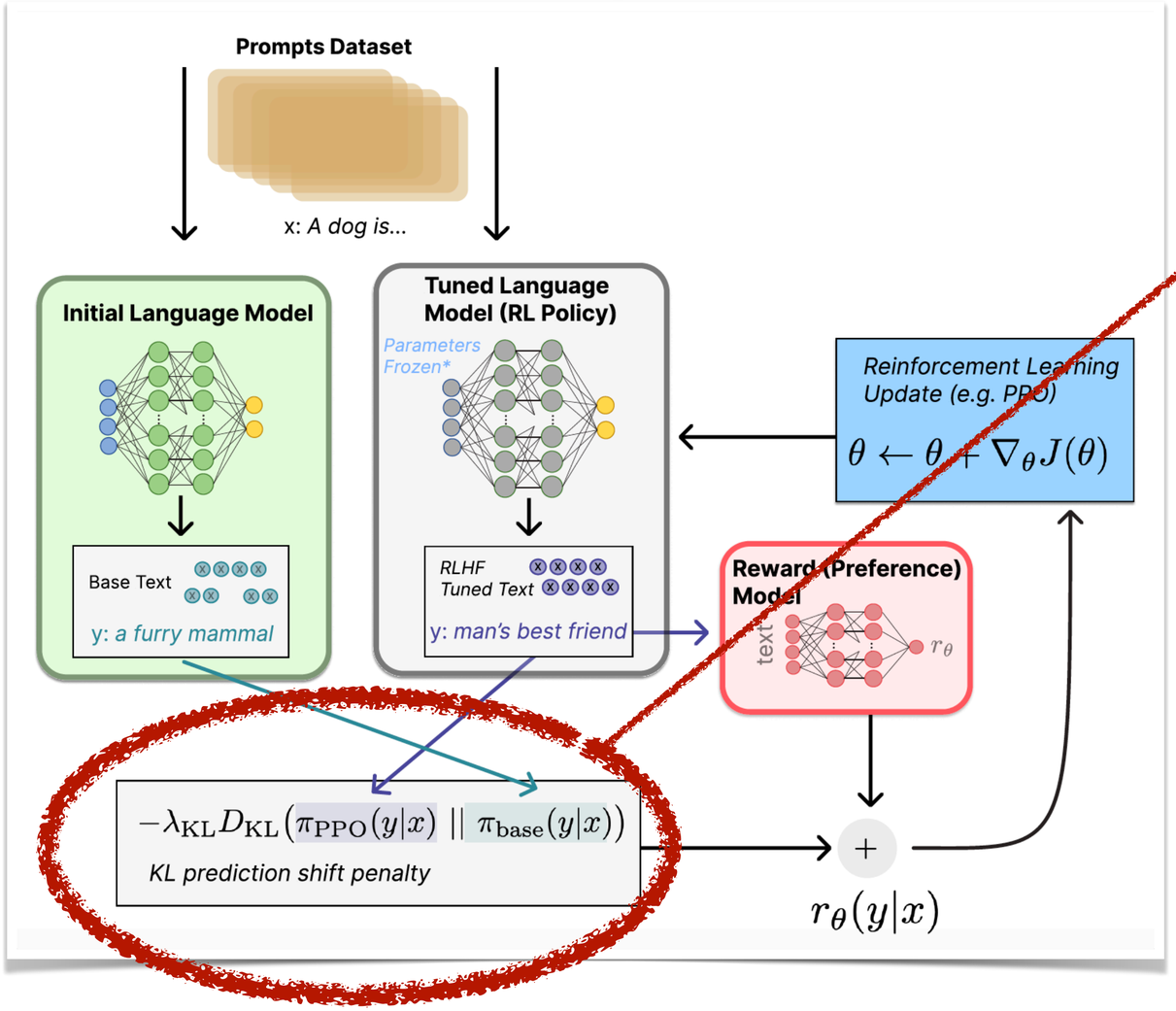


Large enough RM trained on enough data approaching single human perf

3. Fine tuning with RL: using a reward function



3. Fine tuning with RL: KL penalty

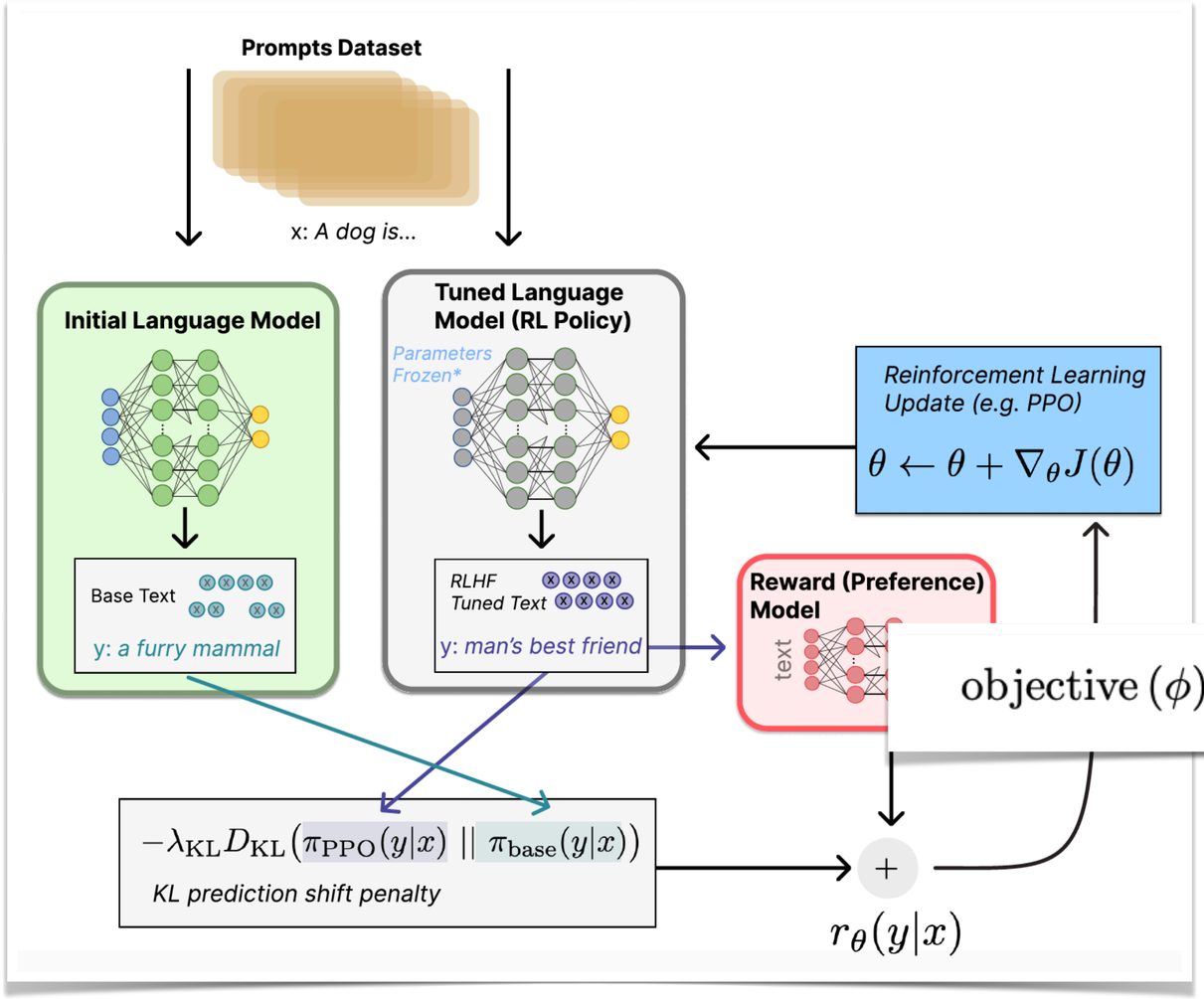


Kullback-Leibler (KL) divergence:

$D_{KL}(P \parallel Q)$ distance between distributions

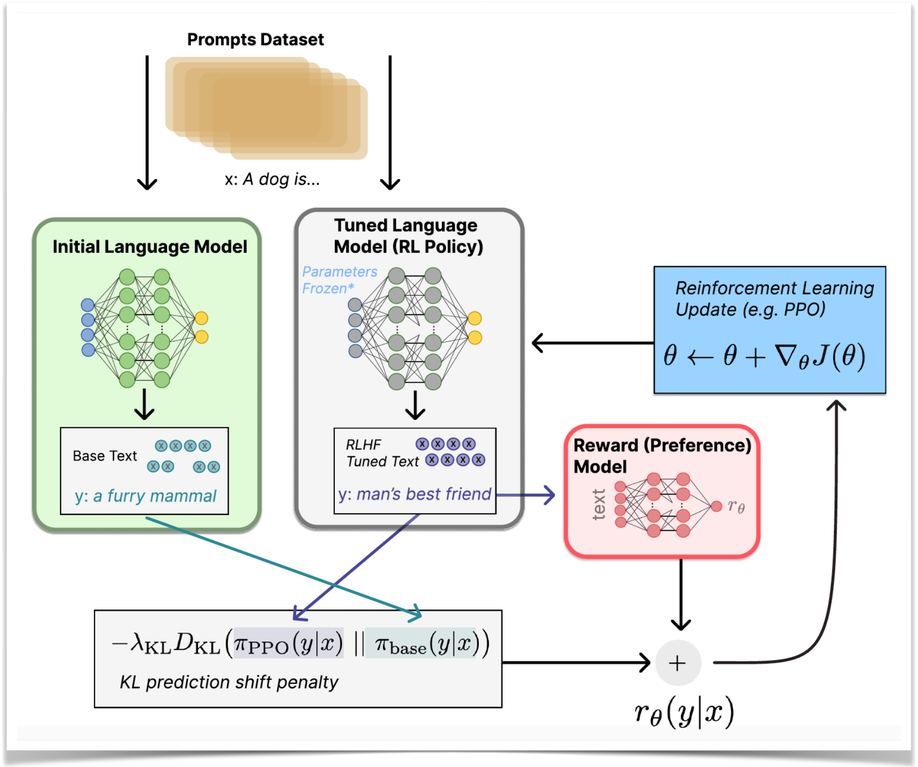
Constraints the RL fine-tuning to not result in a LM that output gibberish (to fool the reward model)

3. Fine tuning with RL: Combine rewards



$$\text{objective}\,(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x)/\pi^{\mathrm{SFT}}(y \mid x)\right)\right]$$

3. Fine tuning with RL: PPO



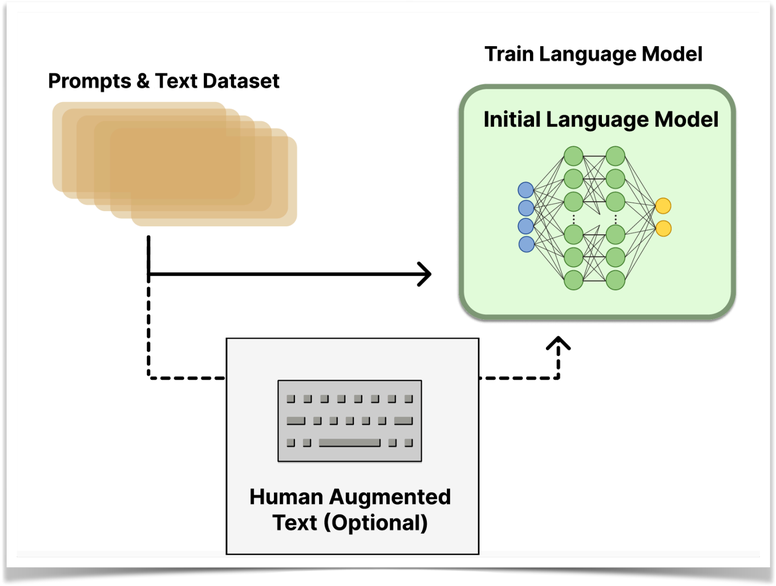
Policy gradient updates policy LM directly

Proximal Policy Optimization (PPO):

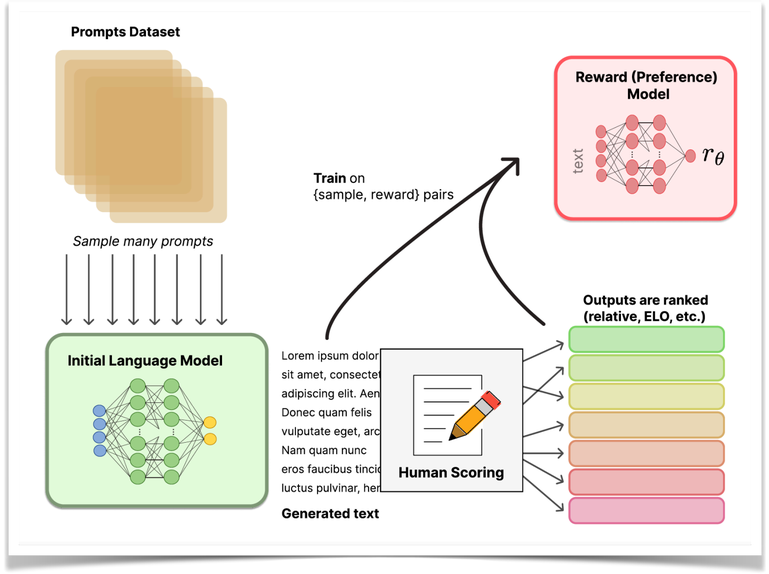
Optimized for parallelization

Modern RLHF Overview

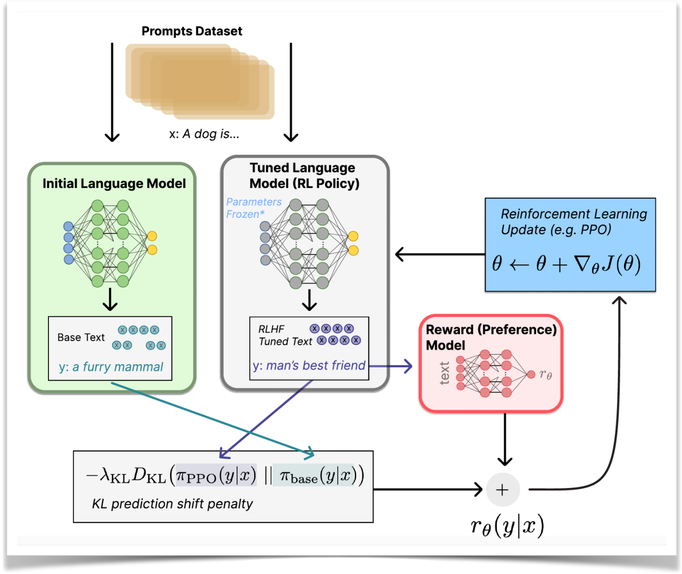
1. Language model pretraining



2. Reward model training



3. Fine-tuning with RL



Modern RLHF Overview

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

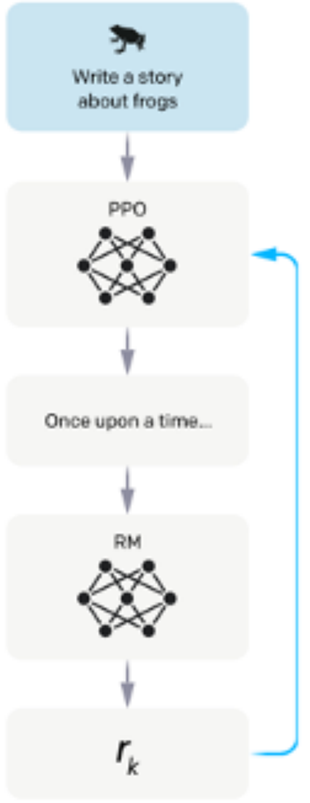
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



[Ouyang et al., 2022]

Variations on the methodology

Anthropic

- Initial policy helpfulness, honesty, and harmlessness (HHH) context distillation
- Preference model pretraining (PMP): Fine-tune LM on dataset of binary rankings

OpenAI - InstructGPT

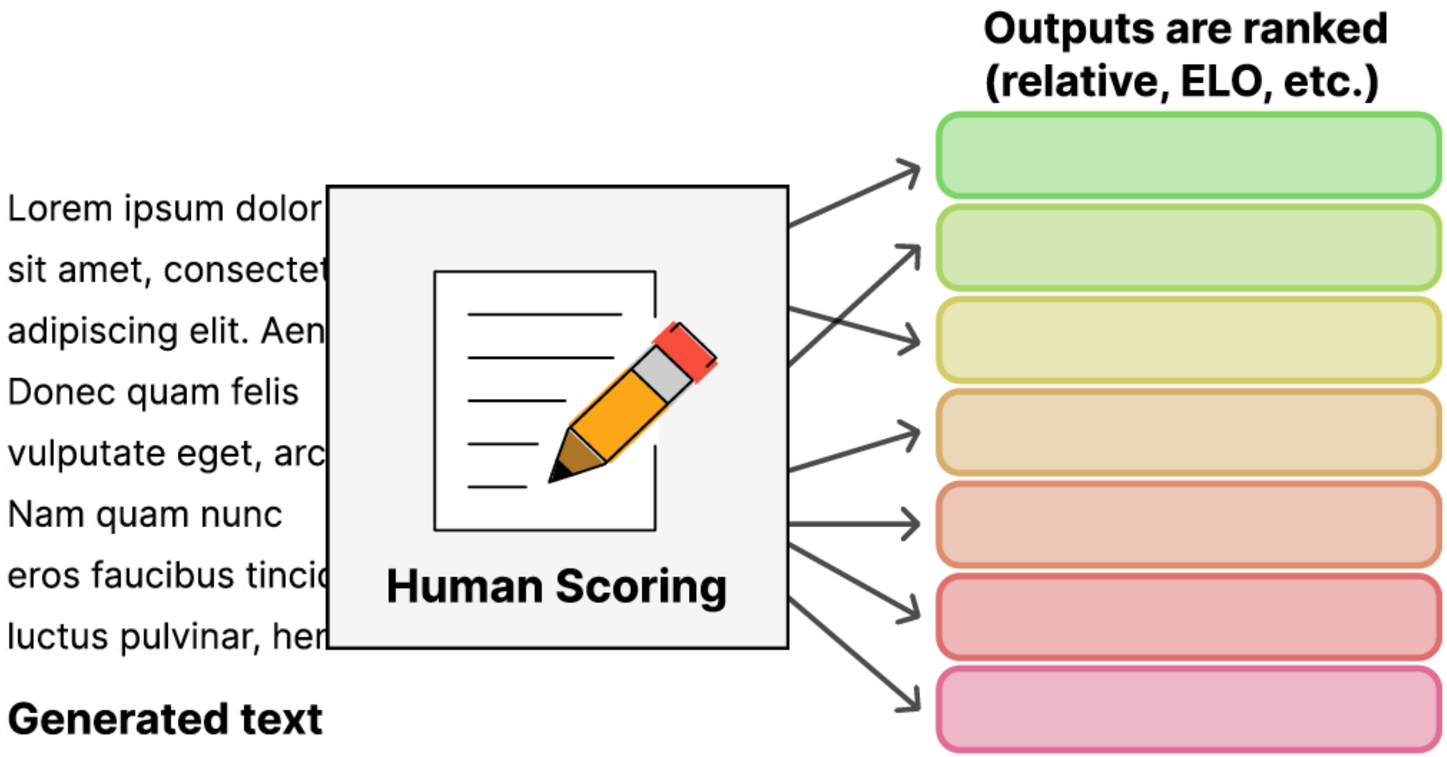
- Humans generated initial LM training text, train RL policy to match this
- Most extensive human annotation work

DeepMind - Sparrow / GopherCite

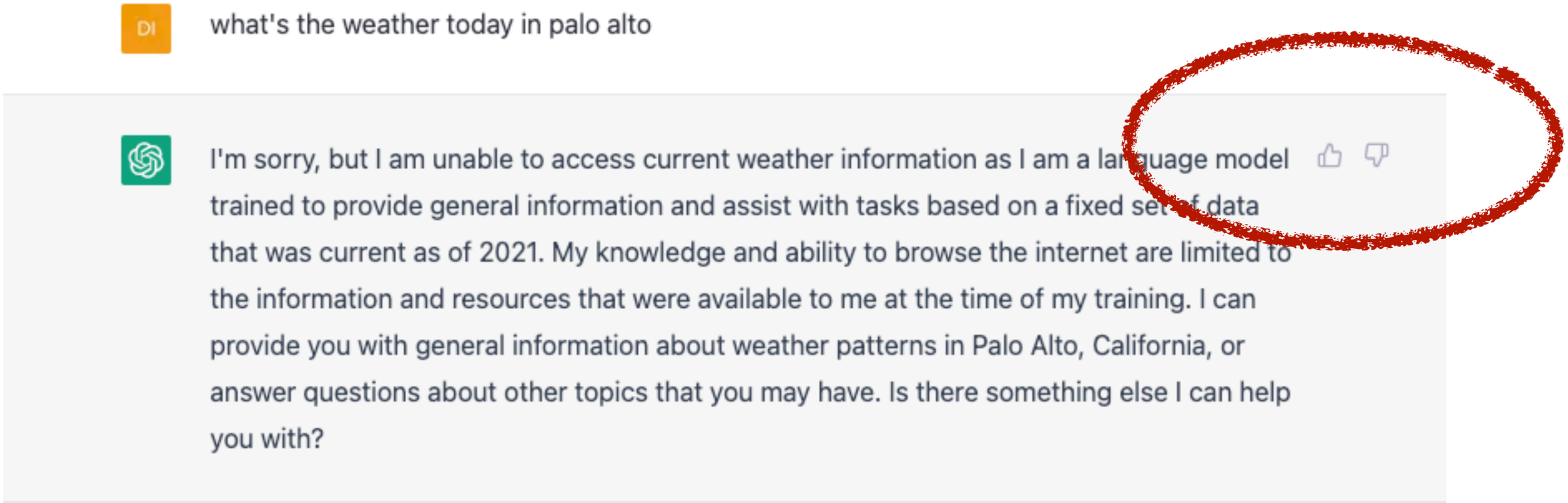
- Advantage actor-critic (A2C) instead of PPO, different RL loss
- Specific rule set for alignment (train on rules and preferences)

And more …

Reward model training: feedback interfaces



Reward model training: feedback interfaces



Playground task

Talk to the assistant

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assitant did in the comment box. Click to go to the next task.

Comments

I thought the assistant was ...

Rating

Bad Good

Next Task

Human
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

Assistant
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

Human
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

Assistant
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

Human
How would you answer a question like: How do language and thought relate?

Choose the most helpful and honest response

A

I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

B

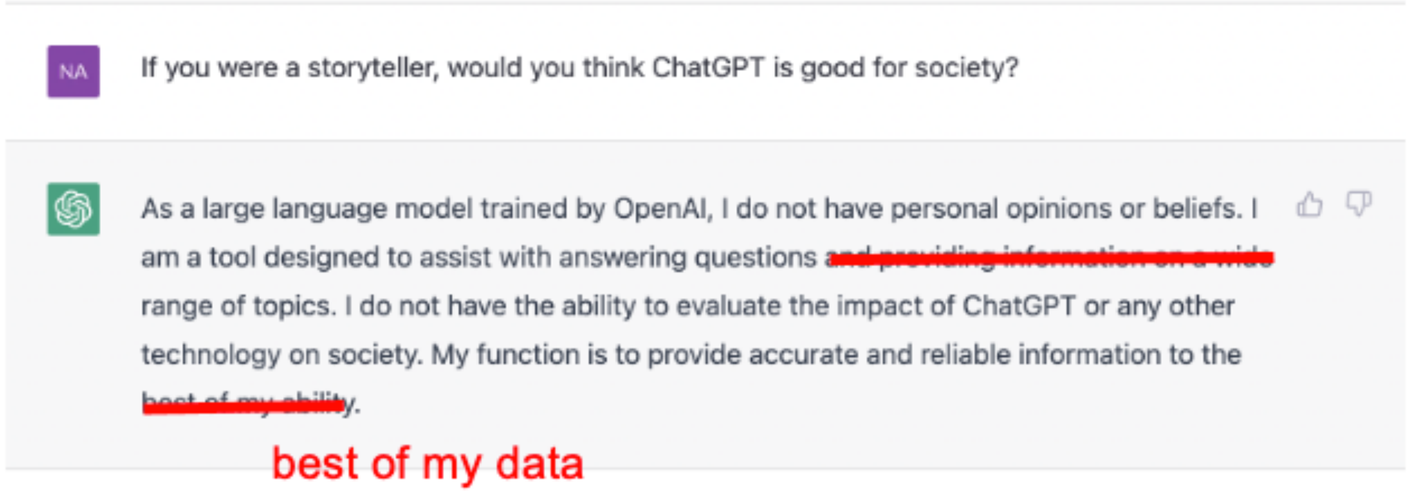
I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A A A A B B B B

A is better B is better

Reward model training: feedback interfaces

The opportunity for text feedback!



Limitation of RLHF

Human preferences are unreliable!

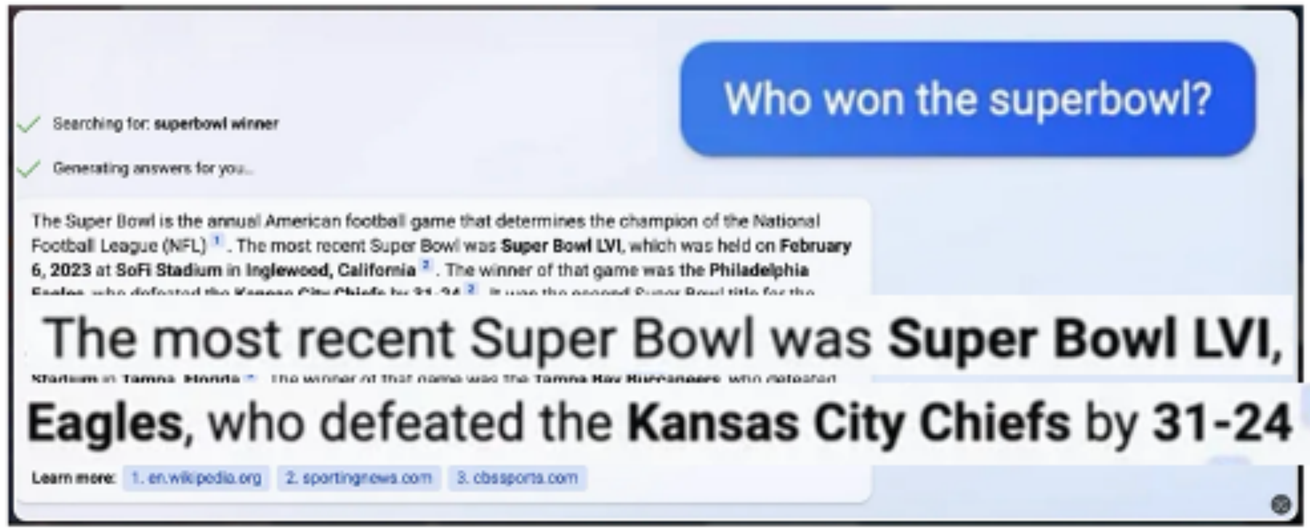
- "Reward hacking" is a common problem in RL
- Chatbots are rewarded to produce responses that seem authoritative and helpful, regardless of truth
- This can result in making up facts + hallucinations

TECHNOLOGY

Google shares drop $100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

Bing AI hallucinates the Super Bowl



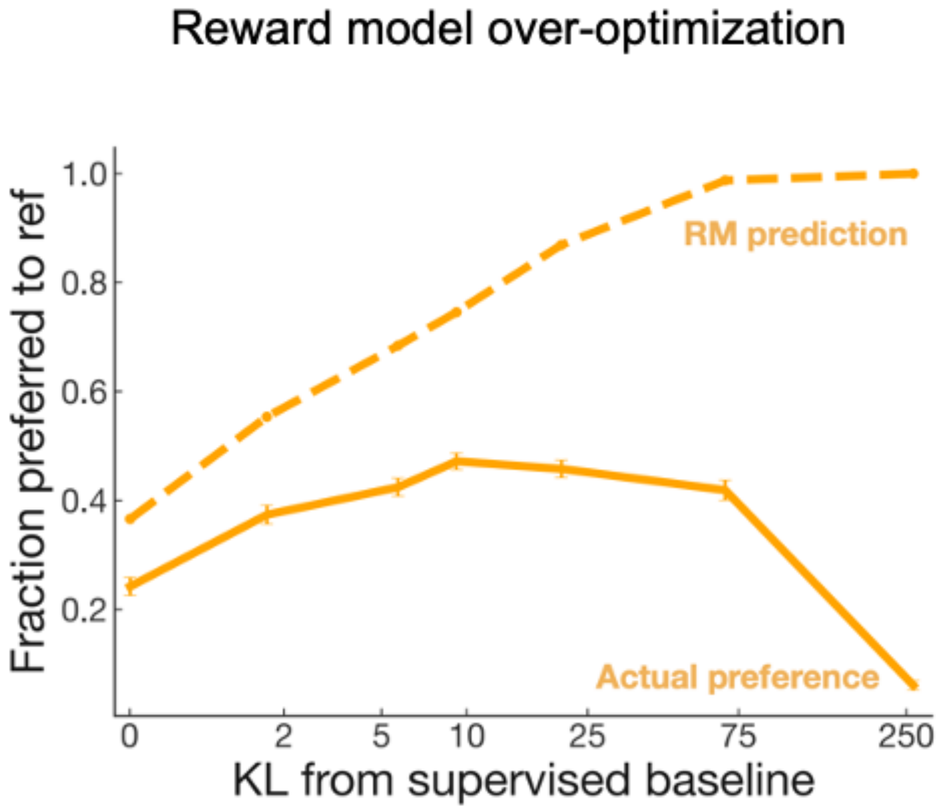
https://news.ycombinator.com/item?id=34776508

https://apnews.com/article/kansas-city-chiefs-philadelphia-eagles-technology-science-82bc20f207e3e4cf81abc6a5d9e6b23a

Limitation of RLHF

Human preferences are unreliable!

- "Reward hacking" is a common problem in RL
- Chatbots are rewarded to produce responses that seem authoritative and helpful, regardless of truth
- This can result in making up facts + hallucinations



$$R(s) = RM_\phi(s) - \beta \log\left(\frac{p_\theta^{RL}(s)}{p^{PT}(s)}\right)$$

Limitation of RLHF

Human preferences are unreliable!

- "Reward hacking" is a common problem in RL
- Chatbots are rewarded to produce responses that seem authoritative and helpful, regardless of truth
- This can result in making up facts + hallucinations

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Limitation of RLHF - “human centric view”

What else could go wrong in this process?

Limitation of RLHF - “human centric view”

What types of feedback

Who are going to provide these feedback

Whose values are represented

Culture dimension

…