

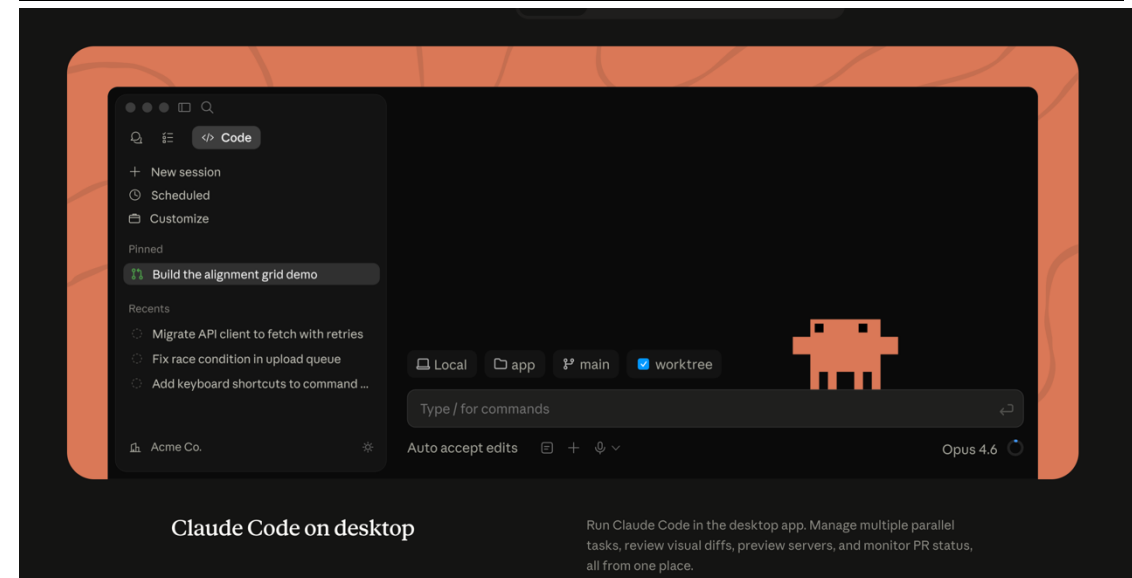
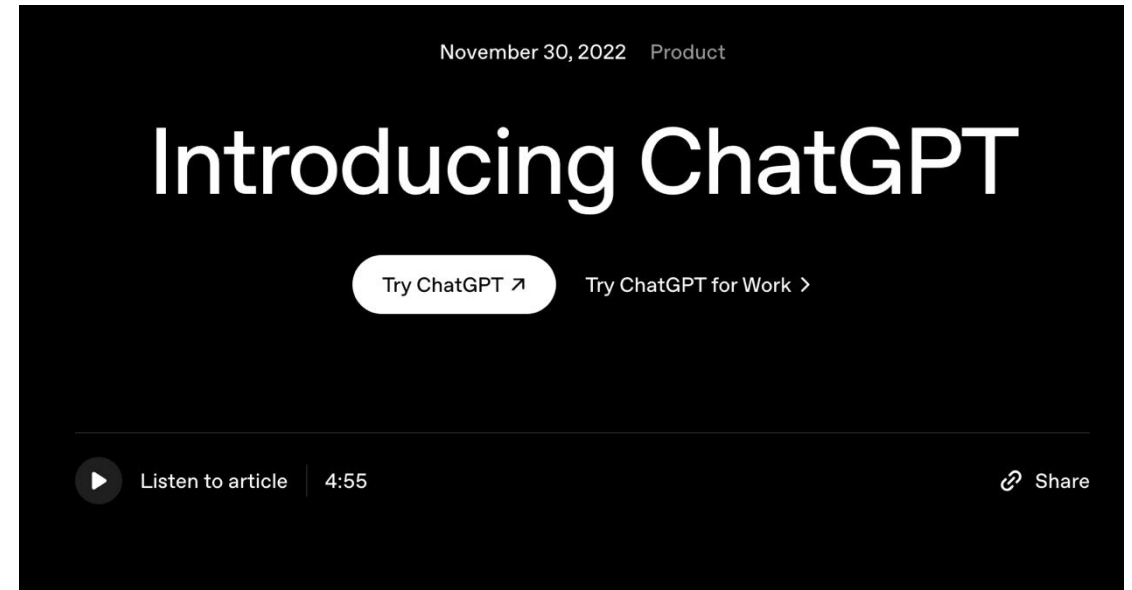
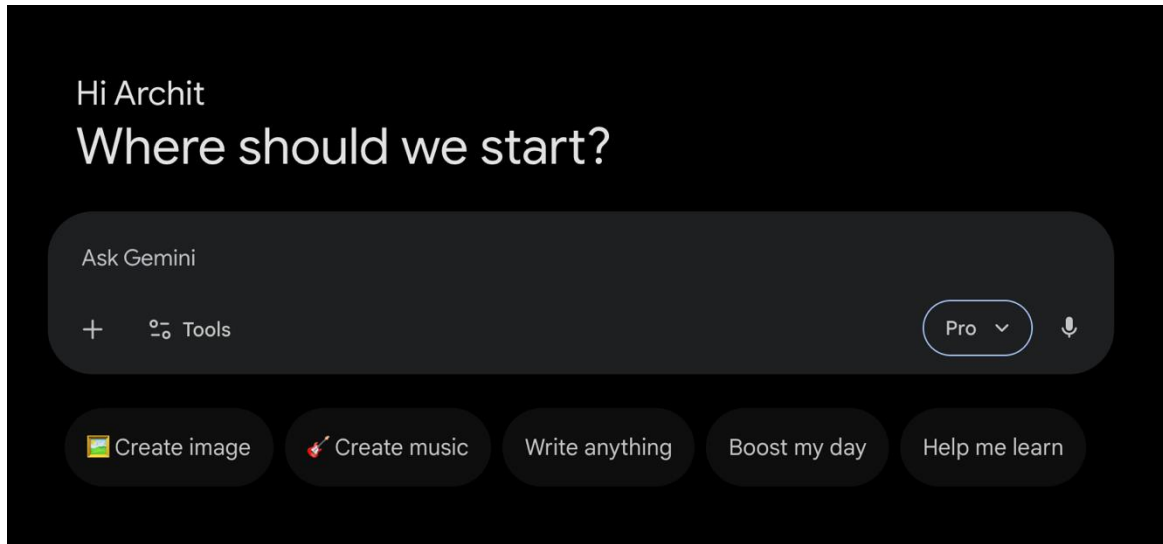
The Post-Training Frontier: RLHF, DPO and Modern Preference Optimization

CS224R

Archit Sharma

(Based on slides from CS224N)

LLMs ... everywhere?



LLMs ... everywhere?

Hi Archit
Where should we start?

Ask Gemini

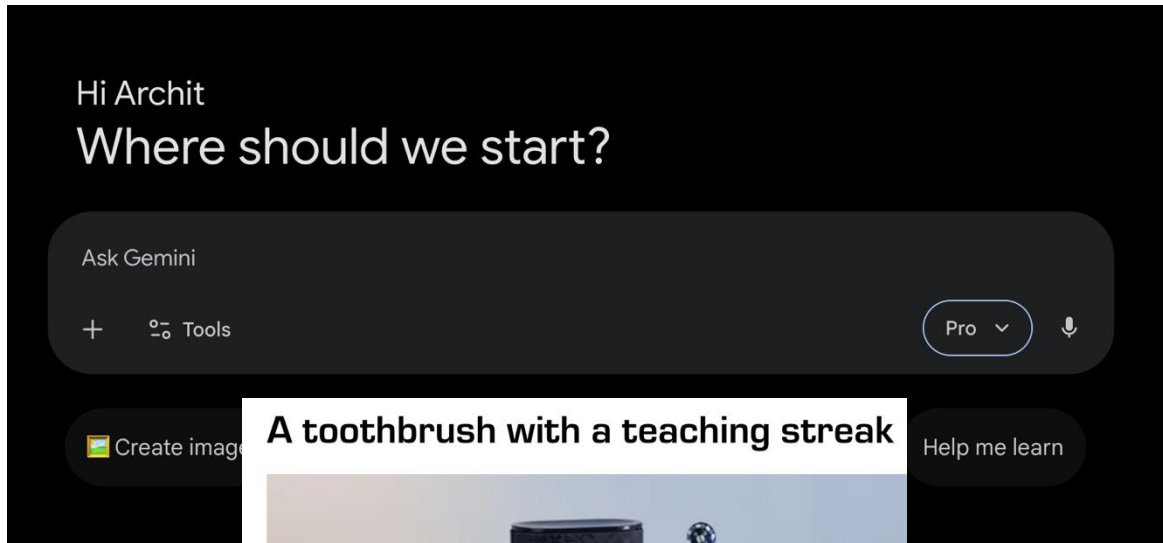
+ Tools

Pro

Create image

A toothbrush with a teaching streak

Help me learn

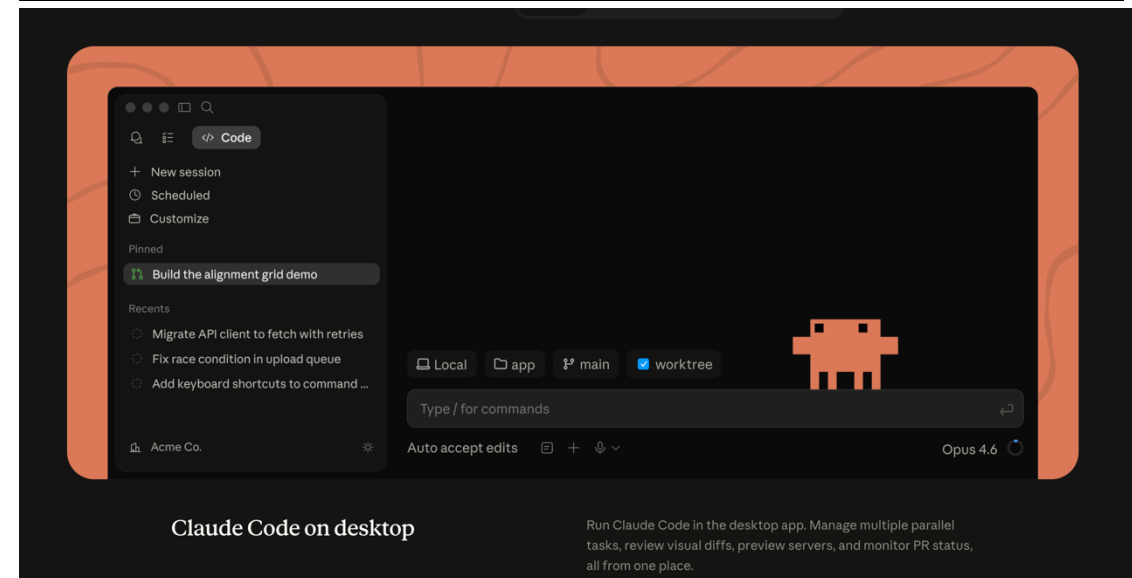
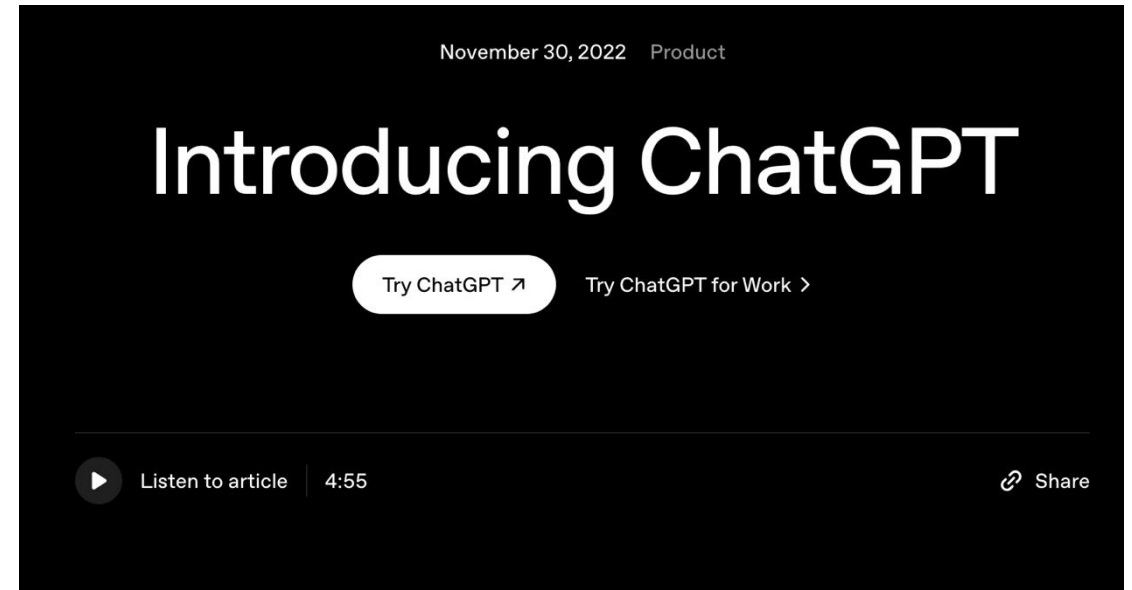


November 30, 2022 Product

Introducing ChatGPT

Try ChatGPT ↗ Try ChatGPT for Work >

Listen to article 4:55 Share

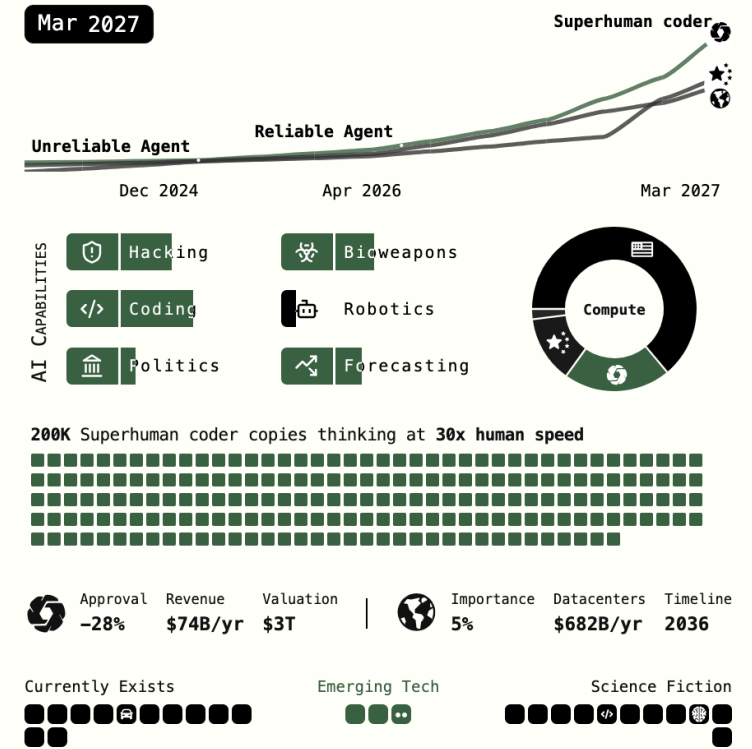
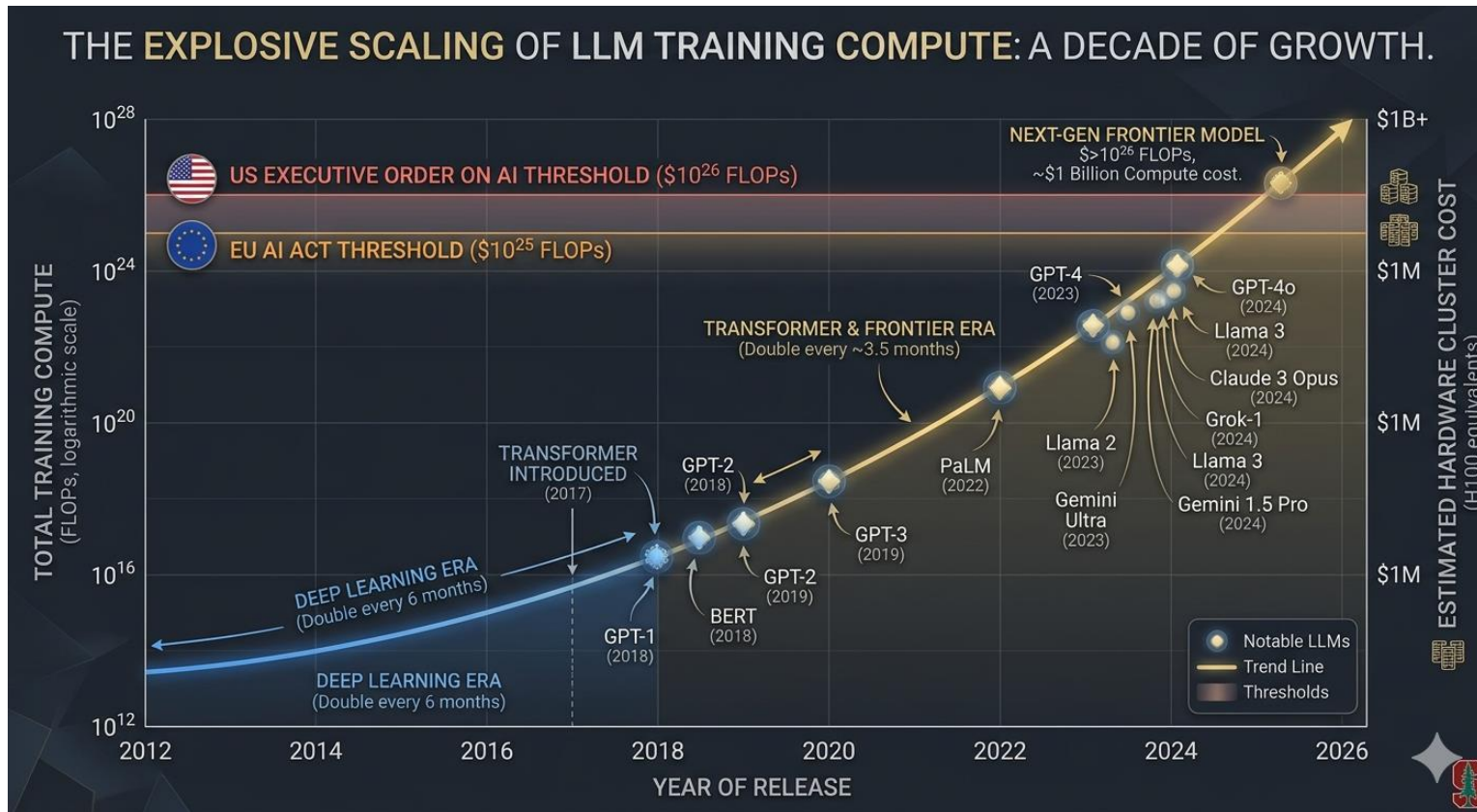


Acme Co. Auto accept edits Opus 4.6

Claude Code on desktop

Run Claude Code in the desktop app. Manage multiple parallel tasks, review visual diffs, preview servers, and monitor PR status, all from one place.

Larger and larger models



LLM Training Overview

Pre-Training: train on lots of natural data, sourced and curated from variety of sources, primarily internet

Mid-Training: more targeted domains, lower volume data

Supervised Fine-Tuning / Instruction Tuning: small, curated, enable model to follow human intent

Reinforcement Learning (from Human Feedback): align with implicit human intent

LLM Training Overview

Pre-Training: train on lots of natural data, sourced and curated from variety of sources, primarily internet

Mid-Training: more targeted domains, lower volume data

Supervised Fine-Tuning / Instruction Tuning: small, curated, enable model to follow human intent

Reinforcement Learning (from Human Feedback): align with implicit human intent

What kinds of things does pretraining learn?

- *Stanford University is located in _____, California. [Trivia]*
- *I put ___ fork down on the table. [syntax]*
- *The woman walked across the street, checking for traffic over ___ shoulder. [coreference]*
- *I went to the ocean to see the fish, turtles, seals, and _____. [lexical semantics/topic]*
- *Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _____. [sentiment]*
- *Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the _____. [some reasoning – this is harder]*
- *I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, _____. [some basic arithmetic; they don't learn the Fibonacci sequence]*

Language models as world models?

Language models may do rudimentary modeling of *agents*, *beliefs*, and *actions*:

Pat watches a demonstration of a bowling ball and a leaf being dropped at the same time in a vacuum chamber. Pat, who is a physicist, predicts that the bowling ball and the leaf will fall at the same rate.

Changing the last sentence of the prompt, we get:

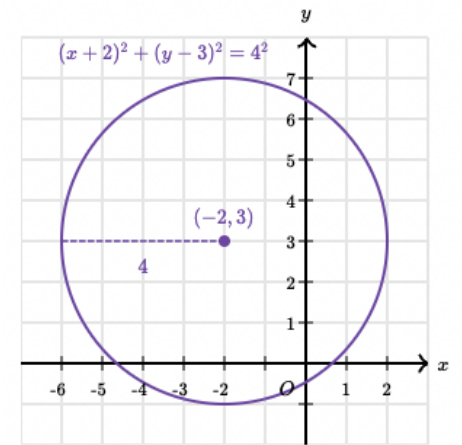
...Pat, who has never seen this demonstration before, predicts that the bowling ball will fall to the ground first. This is incorrect. In a vacuum chamber, there is no air

Language models as world models?

...*math*:

We can describe circles in the xy -plane using equations in terms of x and y . Circle equations questions require us to understand the connection between these equations and the features of circles.

For example, the equation $(x + 2)^2 + (y - 3)^2 = 4^2$ is graphed in the xy -plane below. It is a circle with a center at $(-2, 3)$ and a radius of 4.



<https://www.khanacademy.org/test-prep/sat/x0a8c2e5f:untitled-652>

Language models as world models?

...code:

```
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8     const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9         method: "POST",
10        body: `text=${text}`,
11        headers: {
12            "Content-Type": "application/x-www-form-urlencoded",
13        },
14    });
15    const json = await response.json();
16    return json.label === "pos";
17 }
```

Language models as world models?

...*medicine*:

Rapid and chronic ethanol tolerance are composed of distinct memory-like states in *Drosophila*

Abstract

Ethanol tolerance is the first type of behavioral plasticity and neural plasticity that is induced by ethanol intake, and yet its molecular and circuit bases remain largely unexplored. Here, we characterize three distinct forms of ethanol tolerance in male *Drosophila*: rapid, chronic, and repeated. Rapid tolerance is composed of two short-lived memory-like states, one that is labile and one that is consolidated. Chronic tolerance, induced by continuous exposure, lasts for two days, induces ethanol preference, and hinders the development of rapid tolerance through the activity of




[[Larnerd, 2023](#)]

Language models as multitask assistants?

How do we get from *this*

Stanford University is located in _____

to *this*?

ChatGPT		
 Examples	 Capabilities	 Limitations
"Explain quantum computing in simple terms"	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?"	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?"	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

The Art of Post-Training

- 1. Instruction finetuning**
- 2. Optimizing for human preferences (DPO/RLHF)**
- 3. Peeking into frontier post-training**

The Art of Post-Training

- 1. Instruction finetuning**
2. Optimizing for human preferences (DPO/RLHF)
3. Peeking into frontier post-training

Language modeling \neq assisting users

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Language models are not *aligned* with user intent [[Ouyang et al., 2022](#)].

Language modeling \neq assisting users

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION **Human**

A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone.

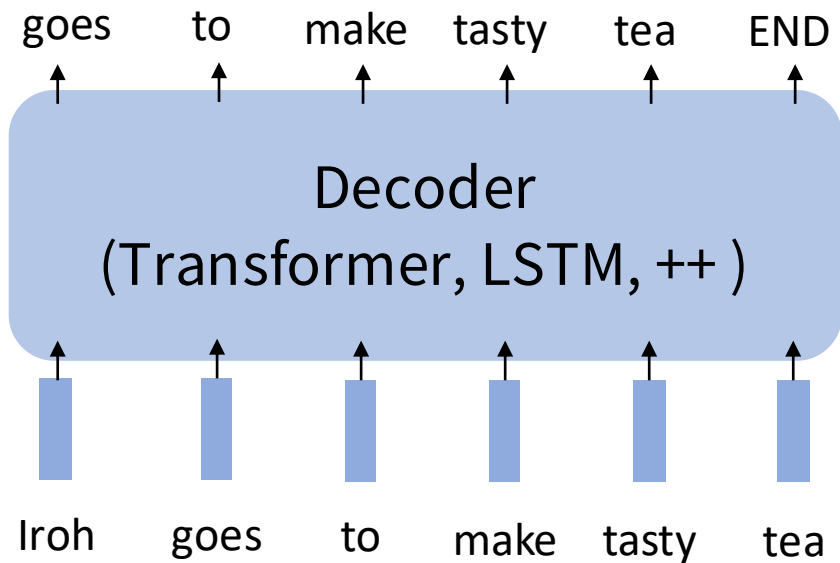
Language models are not *aligned* with user intent [[Ouyang et al., 2022](#)].
Finetuning to the rescue!

The Pretraining / Finetuning Paradigm

Pretraining can improve NLP applications by serving as parameter initialization.

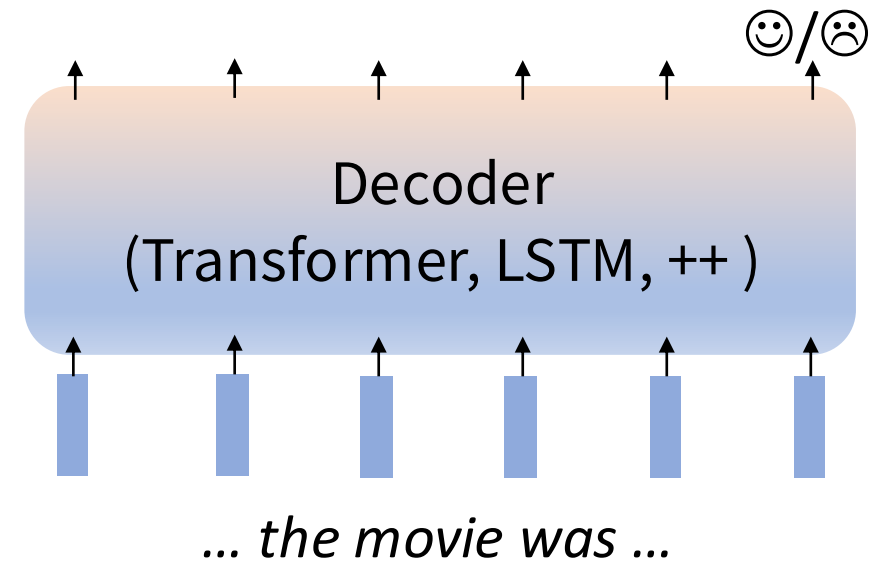
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



Step 2: Finetune (on your task)

Not many labels; adapt to the task!

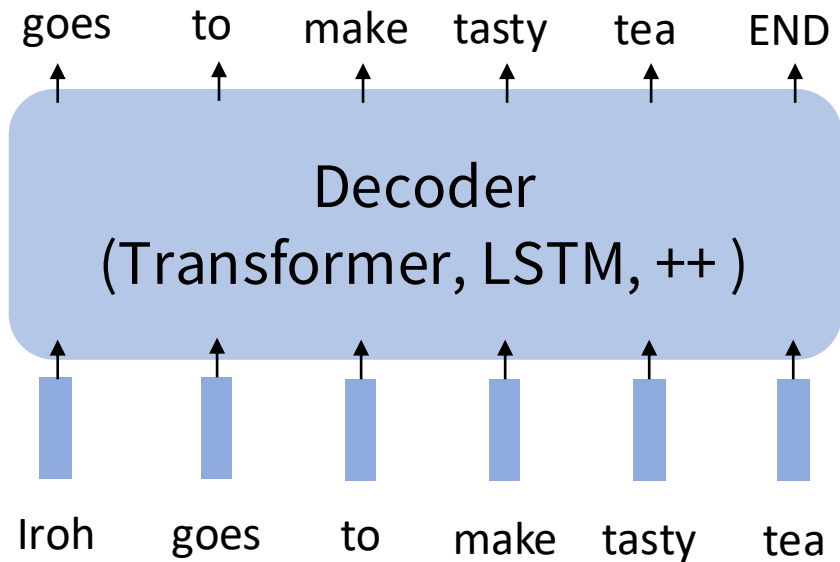


Scaling up finetuning

Pretraining can improve NLP applications by serving as parameter initialization.

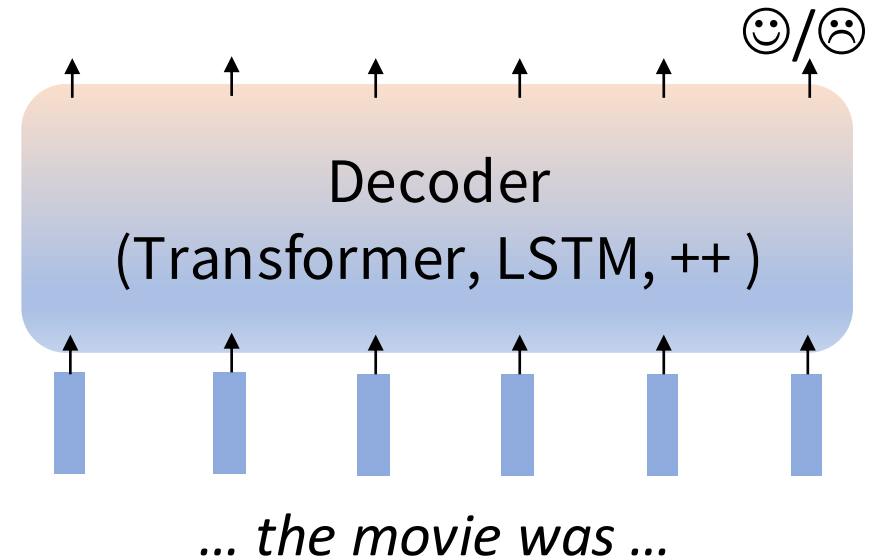
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



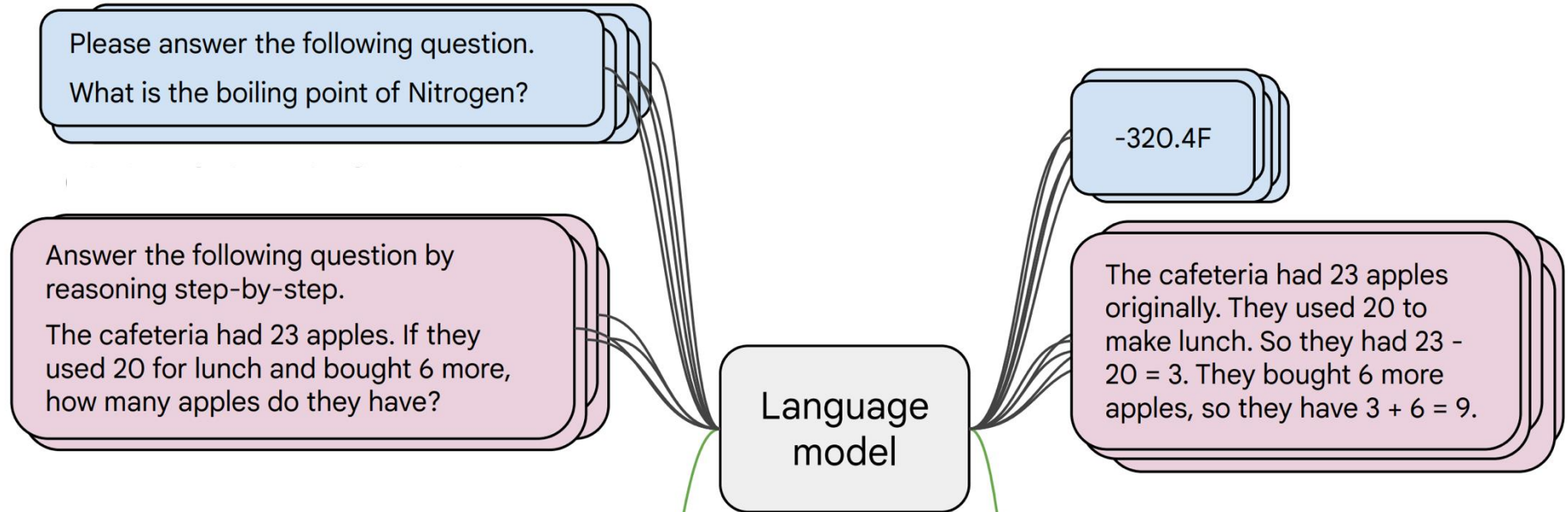
Step 2: Finetune (on **many tasks**)

~~Not~~ many labels; adapt to the tasks!



Instruction finetuning

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



- Evaluate on **unseen tasks**

Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

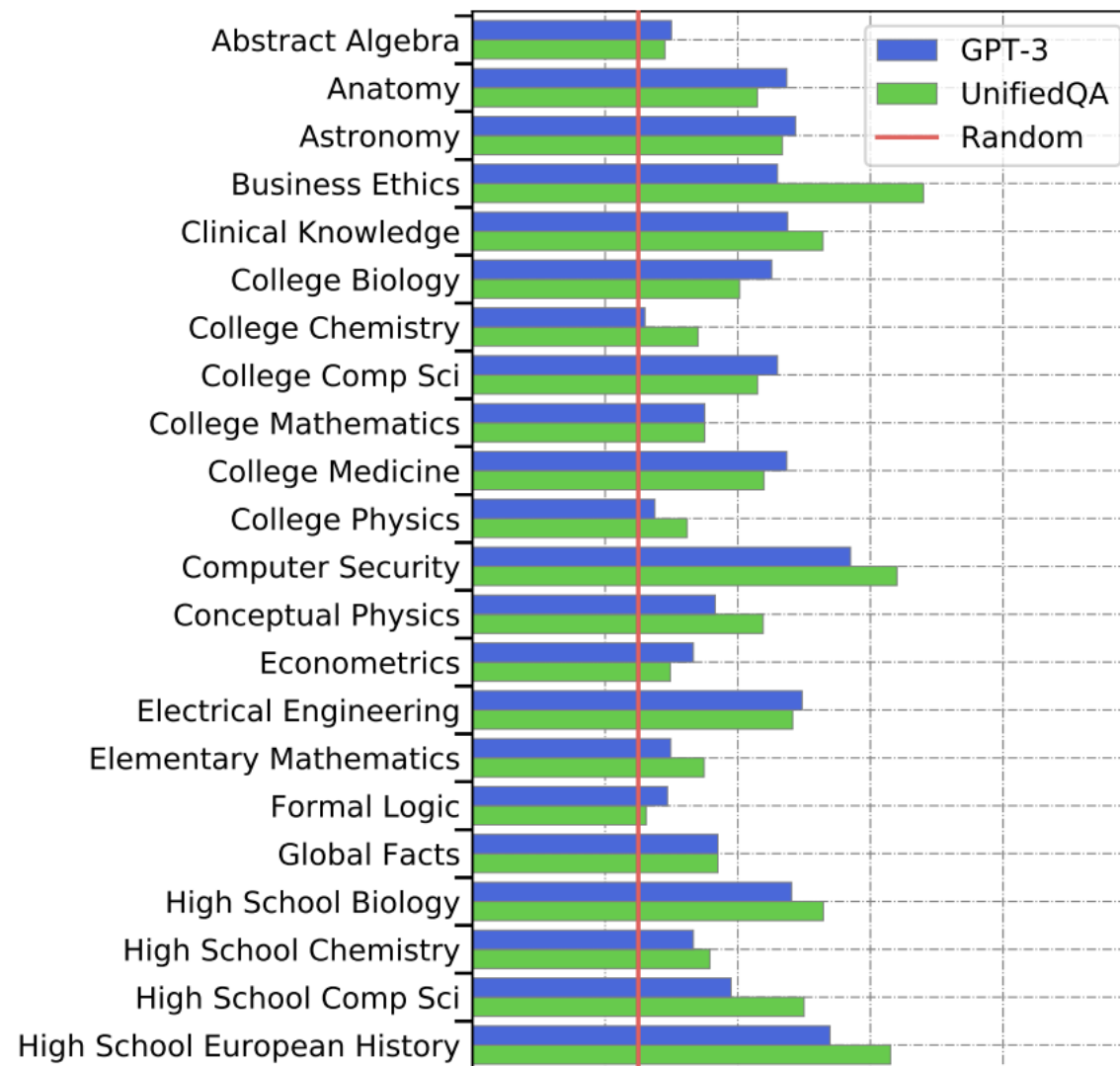
Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Aside: Benchmarks for multitask LMs

Massive Multitask Language Understanding (MMLU)

[[Hendrycks et al., 2021](#)]

New benchmarks for measuring LM performance on 57 diverse *knowledge intensive* tasks



Some intuition: examples from MMLU

Astronomy

What is true for a type-Ia supernova?

- A. This type occurs in binary systems.
- B. This type occurs in young galaxies.
- C. This type produces gamma-ray bursts.
- D. This type produces high amounts of X-rays.

Answer: A

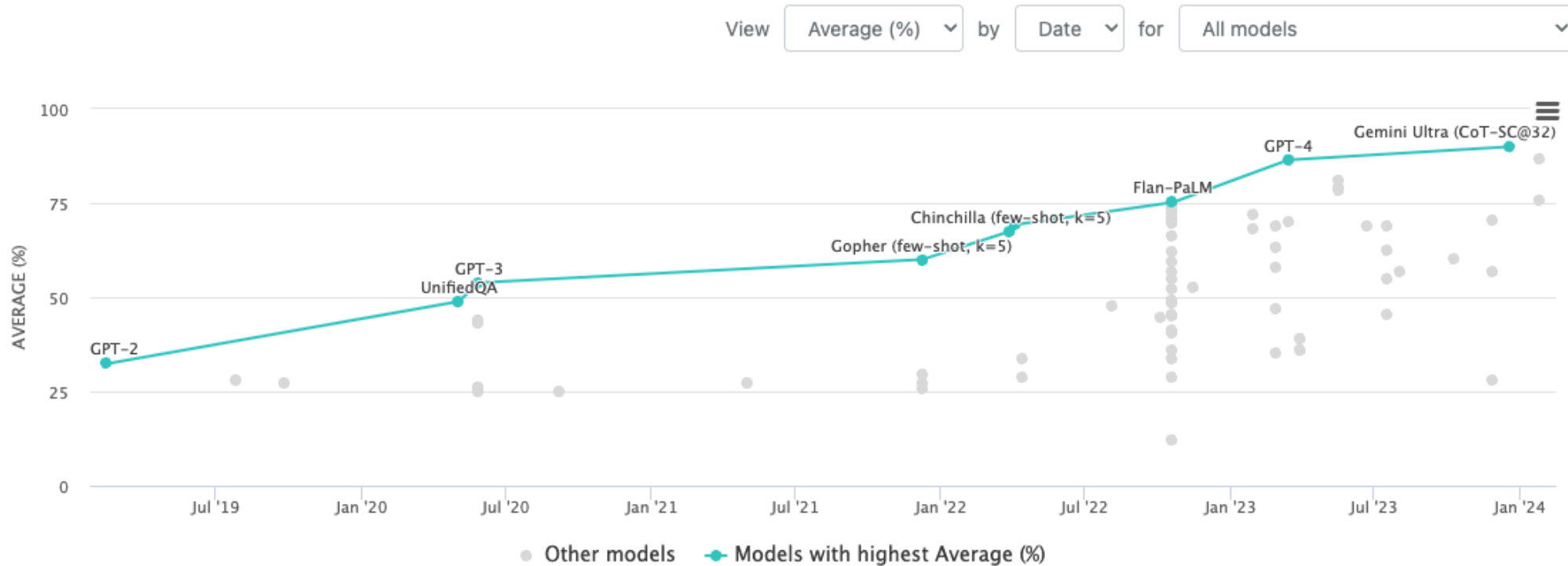
High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

- A. directional selection.
- B. stabilizing selection.
- C. sexual selection.
- D. disruptive selection

Answer: A

Progress on MMLU

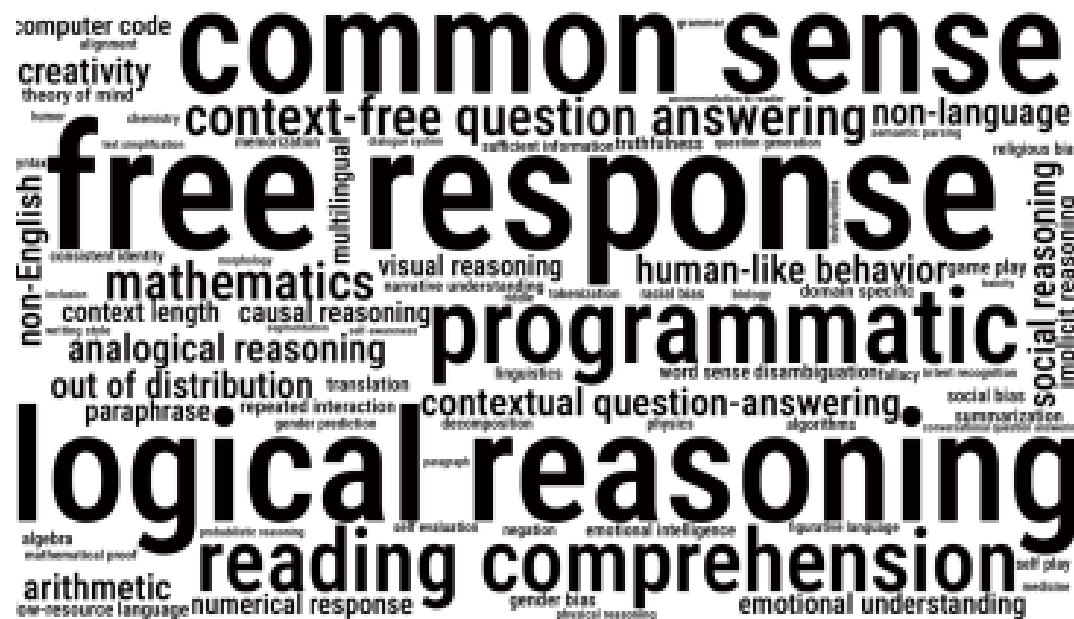


- Rapid, impressive progress on challenging knowledge-intensive benchmarks

Aside: Benchmarks for multitask LMs

BIG-Bench [Srivastava et al., 2022]

200+ tasks, spanning:



https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/README.md

BEYOND THE IMITATION GAME: QUANTIFYING AND EXTRAPOLATING THE CAPABILITIES OF LANGUAGE MODELS

Alphabetic author list:*

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Souza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Heydari, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Erin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engfu Manyas, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jack Hinton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoň, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Boscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Froberg, Jos Rouse, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamal Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kausubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maarten ter Hove, Maheon Feroqi, Manal Faruqi, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez Quintana, Marie Tolchieh, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Batiemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michal Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mímee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Amninasari, Mor Geva, Mozhddeh Gheini, Mukund Varma T. Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Dour, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Huet, Pinyu Hwang, Piotr Milkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qianlang Chen, Raben Banjade, Rachel Etta Rudolph, Raefar Gabriel, Rabel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Kidwai, Roman Novak, Roman Shtein, Roman Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Techan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Dovic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Sumner Mishberg, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikrumar, William Saunders, William Fedus, William Yang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu

Instruction finetuning

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.

The reporter and the chef will discuss the reporter's favorite dishes.

The reporter and the chef will discuss the chef's favorite dishes.

The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✘ (doesn't answer question)

FLAN-T5: <https://huggingface.co/google/flan-t5-xxl>

Instruction finetuning

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

After instruction finetuning

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C). ✓

The Art of Post-Training

1. Instruction finetuning

+ Simple and straightforward, generalize to unseen tasks

- ?

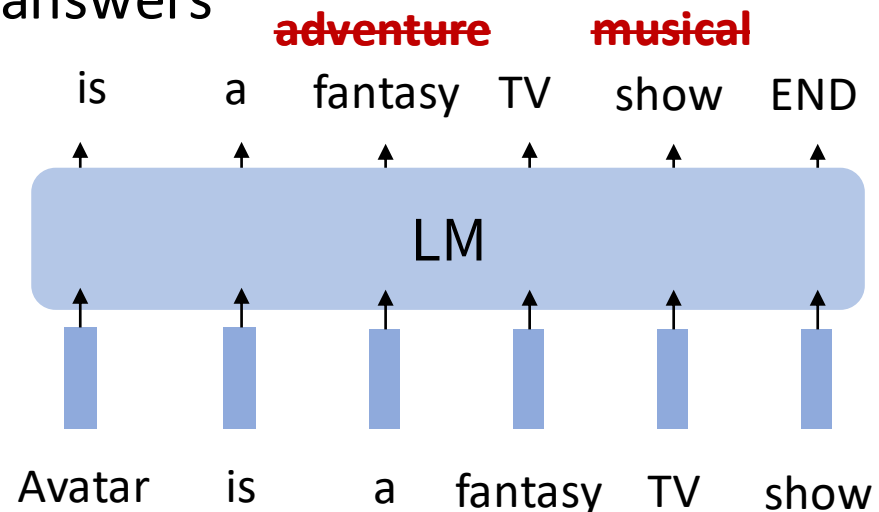
- ?

2. Optimizing for human preferences (DPO/RLHF)

3. Peeking into frontier post-training

Limitations of instruction finetuning?

- One limitation of instruction finetuning is obvious: it's **expensive** to collect ground-truth data for tasks. Can you think of other subtler limitations?
- **Problem 1:** tasks like open-ended creative generation have no right answer.
 - *Write me a story about a dog and her pet grasshopper.*
- **Problem 2:** language modeling penalizes all token-level mistakes equally, but some errors are worse than others.
- **Problem 3:** humans generate suboptimal answers
- Even with instruction finetuning, there is a mismatch between the LM objective and the objective of “satisfy human preferences”!
- Can we **explicitly attempt to satisfy human preferences?**



The Art of Post-Training

1. Instruction finetuning

- + Simple and straightforward, generalize to unseen tasks
- Constructing demonstrations is expensive to scale
- Mismatch between LM objective and human preferences

2. Optimizing for human preferences (DPO/RLHF)

3. Peeking into frontier post-training

The Art of Post-Training

1. Instruction finetuning

- + Simple and straightforward, generalize to unseen tasks
- Constructing demonstrations is expensive to scale
- Mismatch between LM objective and human preferences

2. Optimizing for human preferences (DPO/RLHF)

3. Peeking into frontier post-training

Optimizing for human preferences

- Let's say we were training a language model on some task (e.g. summarization).
- For an instruction x and a LM sample y , imagine we had a way to obtain a *human reward* of that summary: $R(x, y) \in \mathbb{R}$, higher is better.

SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook the
San Francisco
...
overturn unstable
objects.

x

An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.

$$R(x, y_1) = 8.0$$

The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.

$$R(x, y_2) = 1.2$$

- Now we want to maximize the expected reward of samples from our LM:

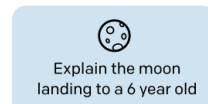
$$\mathbb{E}_{\hat{y} \sim p_{\theta}(y | x)} [R(x, \hat{y})]$$

High-level instantiation: 'RLHF' pipeline

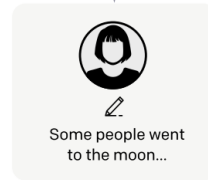
Step 1

Collect demonstration data, and train a supervised policy.

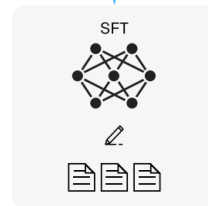
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



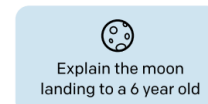
This data is used to fine-tune GPT-3 with supervised learning.



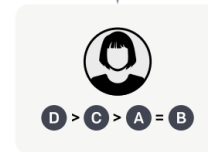
Step 2

Collect comparison data, and train a reward model.

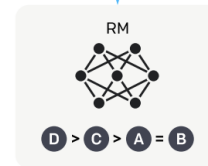
A prompt and several model outputs are sampled.



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.

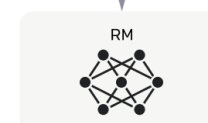


The policy generates an output.



Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



- First step: instruction tuning!
- Second + third steps: maximize reward (but how??)

Optimizing for human preferences

- How do we actually change our LM parameters θ to maximize this?

$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s})]$$

- Let's try doing gradient ascent!

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s} \sim p_{\theta_t}(s)}[R(\hat{s})]$$

How do we estimate
this expectation??

What if our reward
function is non-
differentiable??

- **Policy gradient** methods in RL (e.g., REINFORCE; [[Williams, 1992](#)]) give us tools for estimating and optimizing this objective.
- Quick recap of the simplest policy gradient estimator!

A (very!) brief introduction to policy gradient/REINFORCE

[Williams, 1992]

- We want to obtain (defn. of expectation) (linearity of gradient)

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] = \nabla_{\theta} \sum_s R(s) p_{\theta}(s) = \sum_s R(s) \nabla_{\theta} p_{\theta}(s)$$

- Here we'll use a very handy trick known as the **log-derivative trick**. Let's try taking the gradient of $\log p_{\theta}(s)$

$$\nabla_{\theta} \log p_{\theta}(s) = \frac{1}{p_{\theta}(s)} \nabla_{\theta} p_{\theta}(s) \quad \Rightarrow \quad \nabla_{\theta} p_{\theta}(s) = p_{\theta}(s) \nabla_{\theta} \log p_{\theta}(s)$$

(chain rule)

This is an
expectation of this

- Plug back in:

$$\sum_s R(s) \nabla_{\theta} p_{\theta}(s) = \sum_s p_{\theta}(s) R(s) \nabla_{\theta} \log p_{\theta}(s)$$

$$= \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})]$$

A (very!) brief introduction to policy gradient/REINFORCE

[Williams, 1992]

- Now we have put the gradient “inside” the expectation, we can approximate this objective with Monte Carlo samples:

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] = \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})] \approx \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta} \log p_{\theta}(s_i)$$

This is why it's called “**reinforcement learning**”: we **reinforce** good actions, increasing the chance they happen again.

- Giving us the update rule:

$$\theta_{t+1} := \theta_t + \alpha \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta_t} \log p_{\theta_t}(s_i)$$

This is **heavily simplified!** There is a *lot* more needed to do RL w/ LMs. **Can you see any problems with this objective?**

If R is +++

Take gradient steps to maximize $p_{\theta}(s_i)$

If R is ---

Take steps to minimize $p_{\theta}(s_i)$

How do we get the rewards?

- **Problem 1:** human-in-the-loop is expensive!
- **Solution:** instead of directly asking humans for preferences, **model their preferences** as a separate (NLP) problem! [[Knox and Stone, 2009](#)]

An earthquake hit San Francisco. There was minor property damage, but no injuries.

$$R(x, y_1) = 8.0$$



The Bay Area has good weather but is prone to earthquakes and wildfires.

$$R(x, y_2) = 1.2$$



Train a $RM_\phi(x, y)$ to predict human reward from an annotated dataset, then optimize for RM_ϕ instead.

How do we model human preferences?

- **Problem 2:** human judgments are noisy and miscalibrated!
- **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [[Phelps et al., 2015](#); [Clark et al., 2018](#)]

A 4.2 magnitude
earthquake hit
San Francisco,
resulting in
massive damage.

$$R(x, y_3) = ?$$

$$R(x, y_3) = 4.1? \quad 6.6? \quad 3.2?$$

How do we model human preferences?

- **Problem 2:** human judgments are noisy and miscalibrated!
- **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [[Phelps et al., 2015](#); [Clark et al., 2018](#)]

An earthquake hit San Francisco. There was minor property damage, but no injuries.

y_1

>

A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.

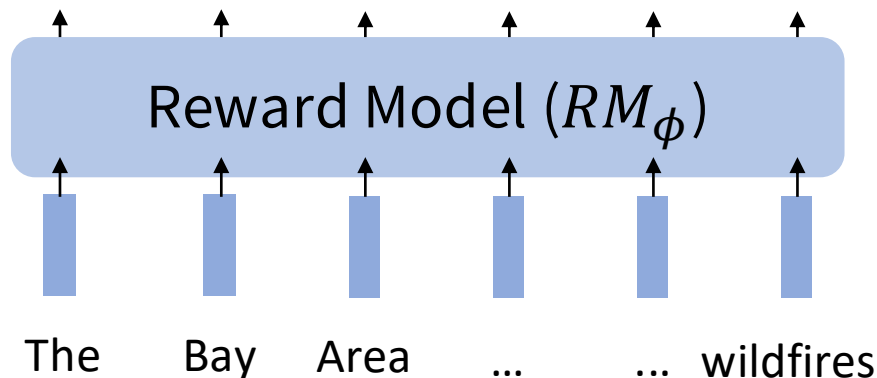
y_3

>

The Bay Area has good weather but is prone to earthquakes and wildfires.

y_2

1.2



Bradley-Terry [1952] paired comparison model

$$J_{RM}(\phi) = -\mathbb{E}_{(x, y^w, y^l) \sim D} [\log \sigma(RM_\phi(x, y^w) - RM_\phi(x, y^l))]$$

“winning”
sample

“losing”
sample

y^w should score
higher than y^l

RLHF: Optimizing the learned reward model

- We have the following:
 - A pretrained (possibly instruction-finetuned) LM $p^{PT}(y | x)$
 - A reward model $RM_{\phi}(x, y)$ that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
- Now to do RLHF:
 - Copy the model $p_{\theta}^{RL}(y | x)$, with parameters θ we would like to optimize
 - We want to optimize:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM_{\phi}(x, \hat{y})]$$

RLHF: Optimizing the learned reward model

- We want to optimize:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM_{\phi}(x, \hat{y})]$$

- Do you see any problems?
 - Learned rewards are imperfect; this quantity can be imperfectly optimized
- Add a penalty for drifting too far from the initialization:

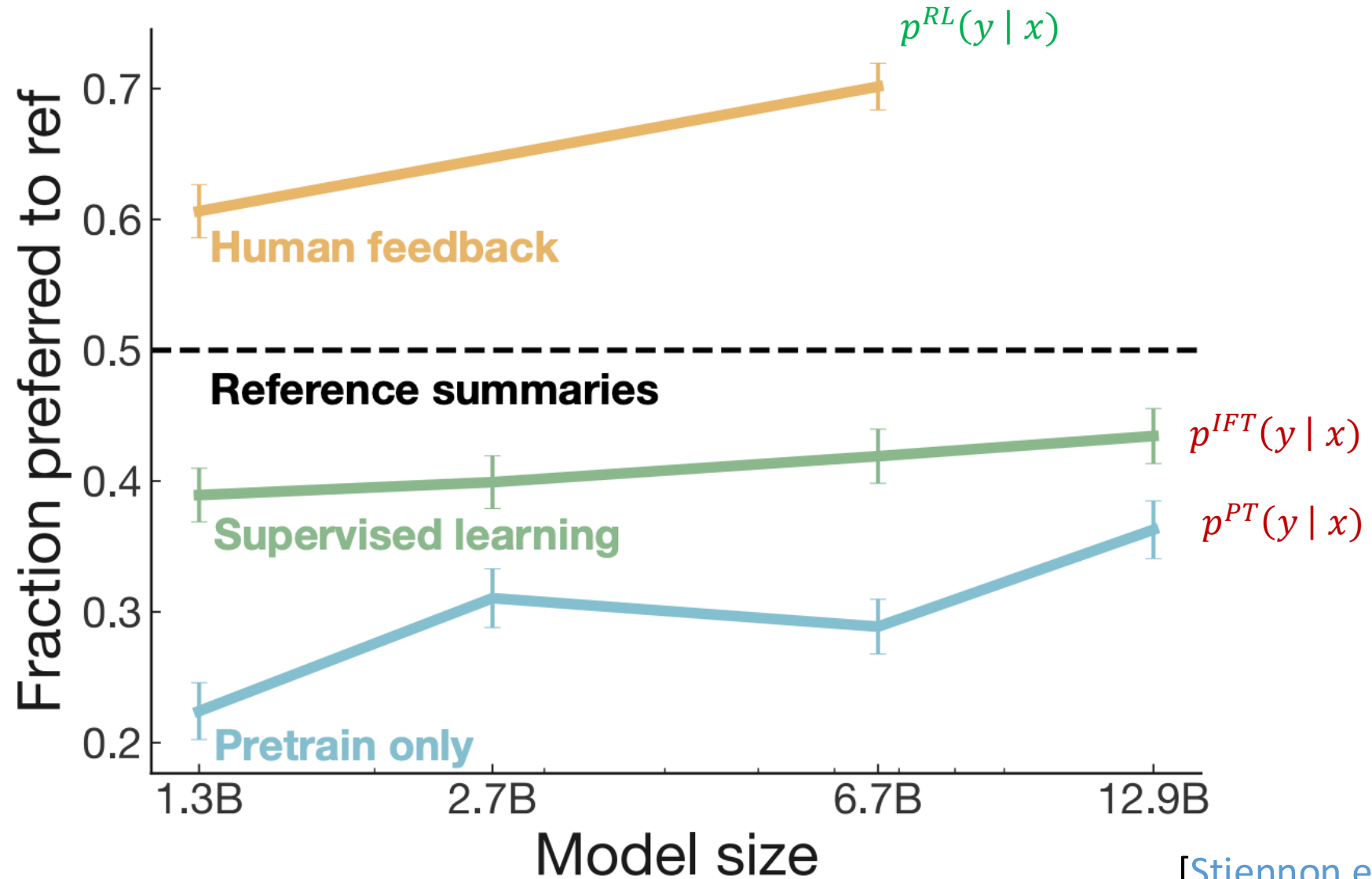
$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM_{\phi}(x, \hat{y}) - \underbrace{\beta \log \left(\frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} \right)}_{\text{penalty}}]$$

Pay a price when

$$p_{\theta}^{RL}(\hat{y}|x) > p^{PT}(\hat{y}|x)$$

This penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler (KL)** divergence between $p_{\theta}^{RL}(\hat{y}|x)$ and $p^{PT}(\hat{y}|x)$.

RLHF provides gains over pretraining + finetuning



Can we simplify RLHF?

- Current pipeline is as follows:
 - Train a reward model $RM_\phi(x, y)$ to produce scalar rewards for LM outputs, trained on a **dataset of human comparisons**
 - Optimize pretrained (possibly instruction-finetuned) LM $p^{PT}(y | x)$ to produce the final RLHF LM $p_\theta^{RL}(\hat{y} | x)$
- What if there was a way to write $RM_\phi(x, y)$ in terms of $p_\theta^{RL}(\hat{y} | x)$?
 - Derive $RM_\theta(x, y)$ in terms of $p_\theta^{RL}(\hat{y} | x)$
 - Optimizing parameters θ by fitting $RM_\theta(x, y)$ to the preference data instead of $RM_\phi(x, y)$
- How is this possible? The only external information to the optimization comes from the preference labels

Direct Preference Optimization (DPO)

- Recall, we want to maximize the following objective:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM(x, \hat{y}) - \beta \log \left(\frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} \right)]$$

- There is a closed form solution to this:

$$p^*(\hat{y}|x) = \frac{1}{Z(x)} p^{PT}(\hat{y}|x) \exp\left(\frac{1}{\beta} RM(x, \hat{y})\right)$$

- Rearrange the terms:

$$RM(x, \hat{y}) = \beta \log \frac{p^*(\hat{y}|x)}{p^{PT}(\hat{y}|x)} + \beta \log Z(x)$$

- This holds true for arbitrary LMs

$$RM_{\theta}(x, \hat{y}) = \beta \log \frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} + \beta \log Z(x)$$

Direct Preference Optimization (DPO)

- Recall, how we fit the reward model $RM_\phi(x, y)$:

$$J_{RM}(\phi) = -\mathbb{E}_{(x, y^w, y^l) \sim D} [\log \sigma(RM_\phi(x, y^w) - RM_\phi(x, y^l))]$$

- Notice that we only need the **difference** between the rewards for y^w and y^l . Simplify for $RM_\theta(x, y)$:

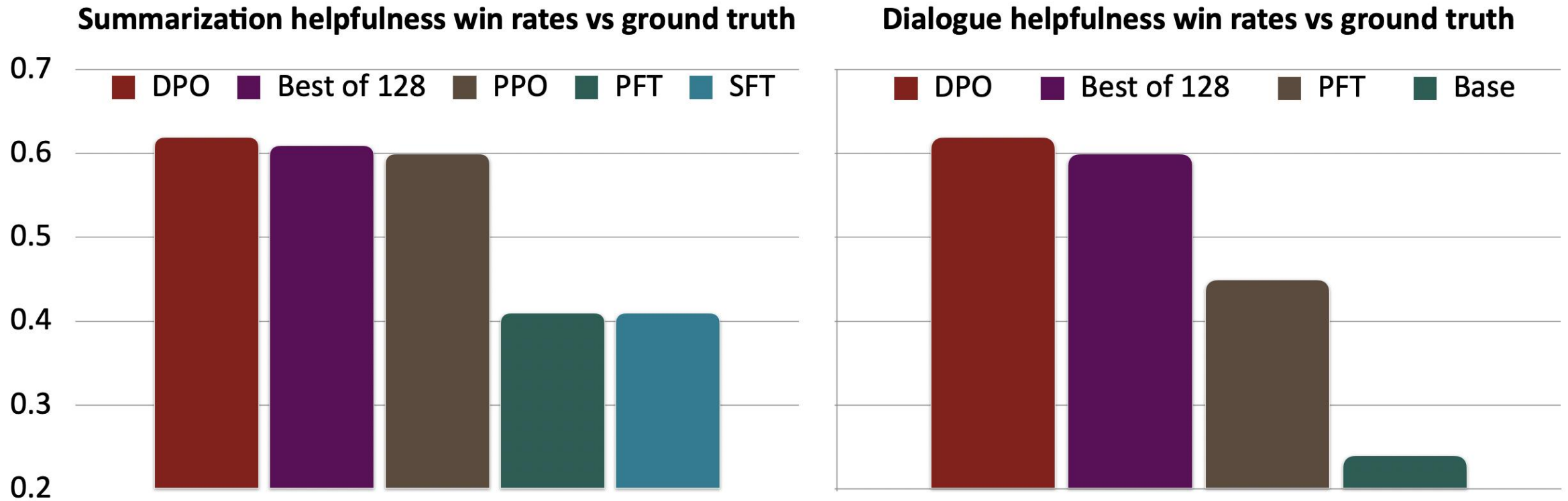
$$RM_\theta(x, y^w) - RM_\theta(x, y^l) = \beta \log \frac{p_\theta^{RL}(y^w | x)}{p^{PT}(y^w | x)} - \beta \log \frac{p_\theta^{RL}(y^l | x)}{p^{PT}(y^l | x)}$$

- The final DPO loss function is:

$$J_{DPO}(\theta) = -\mathbb{E}_{(x, y^w, y^l) \sim D} [\log \sigma(RM_\theta(x, y^w) - RM_\theta(x, y^l))]$$

We have a *simple classification loss* function that connects **preference data** to **language model parameters** directly!

Direct Preference Optimization (DPO)



Summary (DPO and RLHF)

- We want to optimize for human preferences
 - Instead of humans writing the answers or giving uncalibrated scores, we get humans to rank different LM generated answers
- Reinforcement learning from human feedback
 - Train an explicit reward model on comparison data to predict a score for a given completion
 - Optimize the LM to maximize the predicted score (under KL-constraint)
 - Very effective when tuned well, computationally expensive and tricky to get right
- Direct Preference Optimization
 - Optimize LM parameters directly on preference data by solving a binary classification problem
 - Simple and effective, similar properties to RLHF, does not leverage online data

InstructGPT: scaling up RLHF

30k tasks!

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

Some people went to the moon...

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

SFT

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

Explain the moon landing to a 6 year old

A Explain gravity... B Explain war...
C Moon is natural satellite of... D People went to the moon...

A labeler ranks the outputs from best to worst.

D > C > A = B

This data is used to train our reward model.

RM
D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

Write a story about frogs

The policy generates an output.

PPO

The reward model calculates a reward for the output.

Once upon a time...
RM

The reward is used to update the policy using PPO.

r_k

[Ouyang et al., 2022]

InstructGPT: scaling up RLHF

Tasks collected from labelers:

- **Plain:** We simply ask the labelers to come up with an arbitrary task, while ensuring the tasks had sufficient diversity.
- **Few-shot:** We ask the labelers to come up with an instruction, and multiple query/response pairs for that instruction.
- **User-based:** We had a number of use-cases stated in waitlist applications to the OpenAI API. We asked labelers to come up with prompts corresponding to these use cases.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.

InstructGPT

PROMPT

Explain the moon landing to a 6 year old in a few sentences.

COMPLETION

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

InstructGPT

PROMPT *Write a short poem about a wise frog.*

COMPLETION

GPT-3

Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you have lost.

Write a short story in which a character has a bad dream.

InstructGPT

The frog is a symbol of wisdom
He knows all the secrets of the world
He is a master of disguise
And a great teacher of life
He is a symbol of transformation
And the bringer of change
He is the frog who has seen it all
And knows the meaning of it all

ChatGPT: Finetuning + RLHF for dialog agents

Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

ChatGPT: Optimizing
Language Models
for Dialogue

(Instruction finetuning!)

ChatGPT: Finetuning + RLHF for dialog agents

Methods

ChatGPT: Optimizing Language Models for Dialogue

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

(RLHF!)

DPO is enabling open source models to improve!

The screenshot shows the HuggingFace Open LLM Leaderboard interface. At the top, there's a search bar and filter options for model types (pretrained, fine-tuned, instruction-tuned, RL-tuned) and precision (float16, bfloat16, 8bit, 4bit, GPTQ). Below the filters is a table of models with their performance metrics. Handwritten red notes are present on the table, marking several models as 'DPO' or 'Merge (of DPO models)'. The table data is as follows:

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSMBK
1	udkai/Turdus	74.66	73.38	88.56	64.52	67.11	86.66	67.7
2	fblgit/UNA-TheBeagle-7b-v1	73.87	73.04	88	63.48	69.85	82.16	66.72
3	argilla/distilabeled-Maxcoro14-7B-slerp	73.63	70.73	87.47	65.22	65.1	82.08	71.19
4	mlabonne/NeuralMaxcoro14-7B	73.57	71.42	87.59	64.84	65.64	81.22	70.74
5	abideen/NexoNimbus-7B	73.5	70.82	87.86	64.69	62.43	84.85	70.36
6	Neuronovo/neuronovo-7B-v0.2	73.44	73.04	88.32	65.15	71.02	80.66	62.47
7	argilla/distilabeled-Maxcoro14-7B-slerp-full	73.4	70.65	87.55	65.33	64.21	82	70.66
8	Cultrix/MistralTrix-v1	73.39	72.27	88.33	65.24	70.73	80.98	62.77
9	ryandt/MusingCaterpillar	73.33	72.53	88.34	65.26	70.93	80.66	62.24
10	Neuronovo/neuronovo-7B-v0.3	73.29	72.7	88.26	65.1	71.35	80.9	61.41
11	Cultrix/MistralTrixTest	73.17	72.53	88.4	65.22	70.77	81.37	60.73
12	samir-fama/SamirGPT-v1	73.11	69.54	87.04	65.3	63.37	81.69	71.72
13	SanjiWatsuki/Lelantos-DPO-7B	73.09	71.08	87.22	64	67.77	80.03	68.46

Handwritten red notes on the table:

- udkai/Turdus: DPO
- fblgit/UNA-TheBeagle-7b-v1: DPO (& UNA)
- argilla/distilabeled-Maxcoro14-7B-slerp: DPO
- mlabonne/NeuralMaxcoro14-7B: DPO
- abideen/NexoNimbus-7B: Merge (of DPO models)
- Neuronovo/neuronovo-7B-v0.2: DPO
- argilla/distilabeled-Maxcoro14-7B-slerp-full: DPO
- Cultrix/MistralTrix-v1: DPO
- ryandt/MusingCaterpillar: DPO
- Neuronovo/neuronovo-7B-v0.3: DPO
- Cultrix/MistralTrixTest: DPO
- samir-fama/SamirGPT-v1: No info but prob DPO, given Merge (incl. DPO)
- SanjiWatsuki/Lelantos-DPO-7B: DPO



	GPT - 3.5	Mistral Small	Mistral Medium
MT Bench (for Instruct models)	8.32	8.30	8.61

<https://mistral.ai/news/mixtral-of-experts/>

Instruction fine-tuning



pretrained models in chat use cases, we innovated on our well. Our approach to post-training is a combination of instruction sampling, proximal policy optimization (PPO), and (PPO). The quality of the prompts that are used in SFT and used in PPO and DPO has an outsized influence on the some of our biggest improvements in model quality came and performing multiple rounds of quality assurance on notators.

Learning from preference rankings via PPO and **DPO** also greatly improved the performance of Llama 3 on reasoning and coding tasks. We found that if you ask a model a reasoning question that it struggles to answer, the model will sometimes produce the right reasoning trace: The model knows how to produce the right answer, but it does not know how to select it. Training on preference rankings enables the model to learn how to select it.

RLHF/DPO behaviors – clear stylistic changes

Instruction: What are the five most common causes of stress among people?

Alpaca (Finetuned w/ 10k data): The five most common causes of stress among people are work, money, relationships, health, and family

Alpaca + PPO: The five most common causes of stress among people are:

1. Pressure at work: Whether it is due to tight deadlines, high expectations, or an overwhelming workload, pressure at work can be a major source of stress
2. Money problems: Financial issues can be a significant source of stress, from not being able to make ends meet to struggling with debt.
3. Unhealthy lifestyle: Poor diet, lack of exercise, and sleep deprivation can all contribute to stress.
4. Unhappiness at home: Problems in relationships, conflicts with family members, and other domestic issues can be a major source of stress.
5. Uncertainty: Excessive worry and anxiety about the future, health issues, or other unknowns can be a major cause of stress

- Significantly more detailed, nicer/clearer list like formatting

The Art of Post-Training

1. Instruction finetuning

- + Simple and straightforward, generalize to unseen tasks
- Constructing demonstrations is expensive to scale
- Mismatch between LM objective and human preferences

2. Optimizing for human preferences (DPO/RLHF)

- + Directly model preferences, generalize beyond labeled data
- RL can be trickier to get right

3. Peeking into frontier post-training

Learned rewards can be unreliable

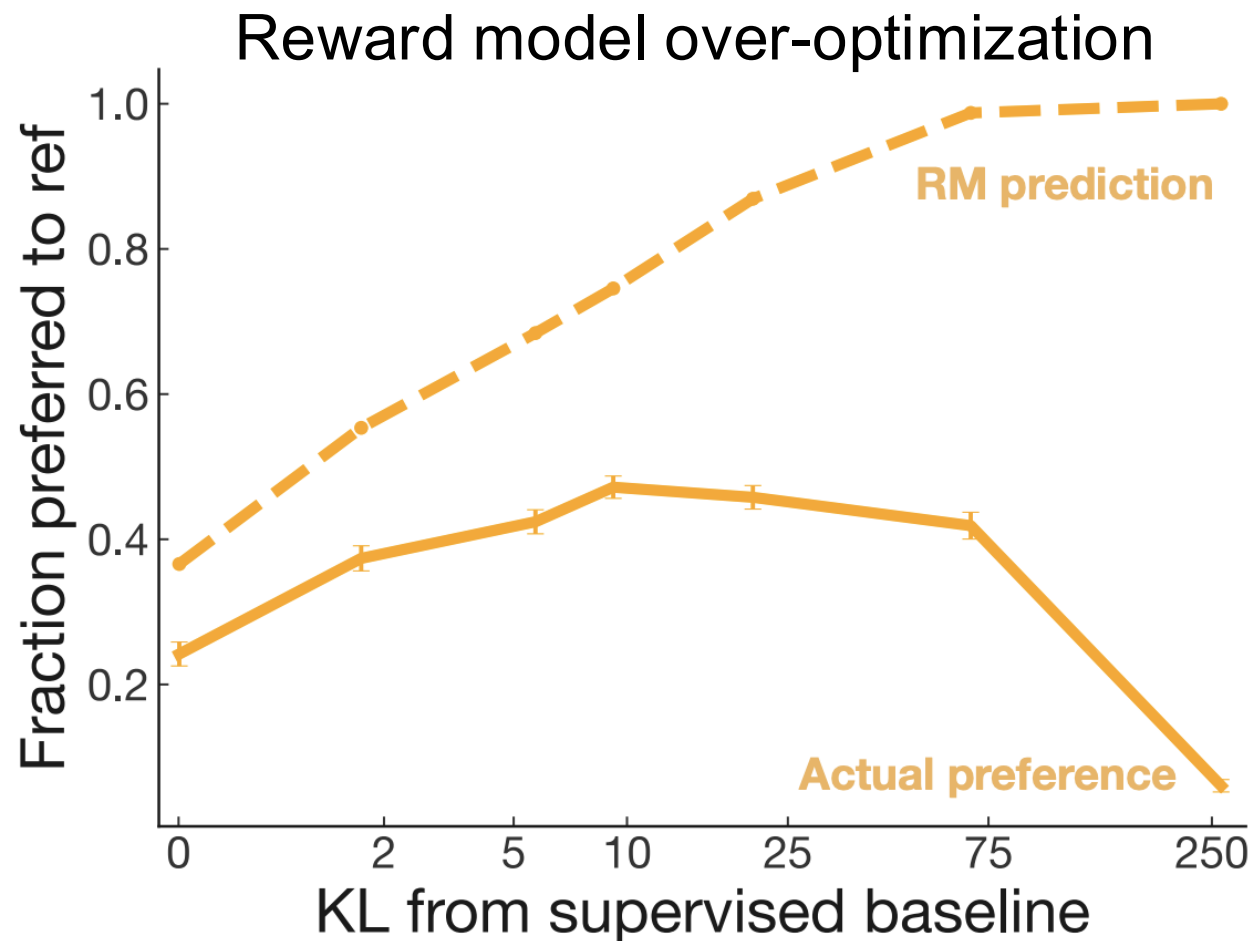
- Human preferences are unreliable!
 - "Reward hacking" is a common problem in RL



<https://openai.com/blog/faulty-reward-functions/>

Learned rewards can be unreliable

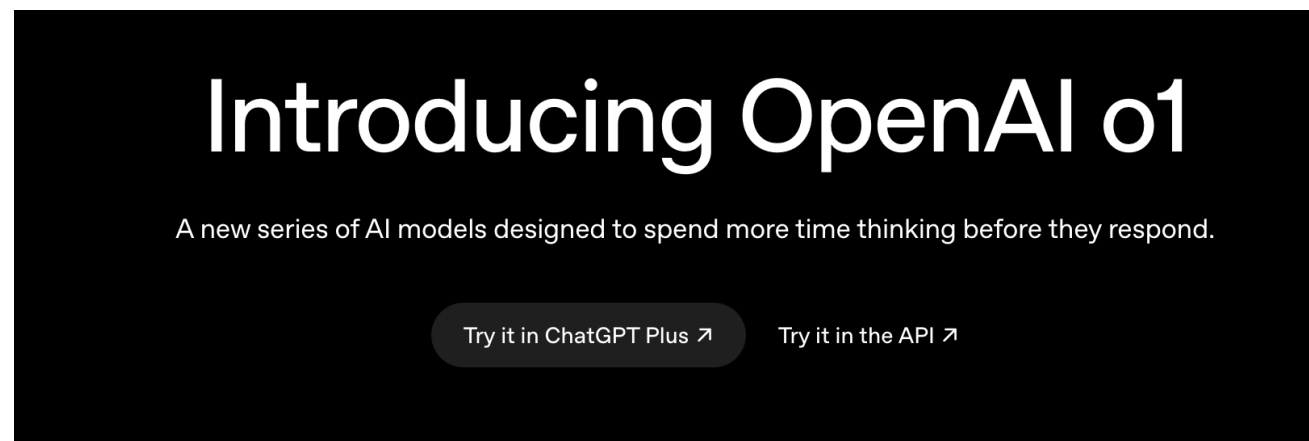
- Human preferences are unreliable!
 - "Reward hacking" is a common problem in RL
- **Models** of human preferences are *even more* unreliable!



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

Learned rewards can be unreliable

- Human preferences are unreliable!
 - "Reward hacking" is a common problem in RL
- Models of human preferences are *even more* unreliable!
- **Train on verifiable rewards!**
 - Math, code, science problems have verifiable solutions
 - Not easy to hack such problems
 - Led to reasoning models (more in the next lecture!)
- Not everything can be made verifiable though ...



Model behavior is hard to control

- Human preferences are unreliable!
 - LLMs over-produce emojis
 - Can become overly sycophantic
- We want more precise control over model behavior
 - Politeness, abstention etc
 - Hard to balance all different reward functions

April 29, 2025 Product

**Sycophancy in GPT-4o:
what happened and
what we're doing about it**

Model behavior is hard to control

- Human preferences are unreliable!
 - LLMs over-produce emojis
 - Can become overly sycophantic
- We want more precise control over model behavior
 - Politeness, abstention etc
 - Hard to balance all different reward functions
- **Can we leverage LLMs itself to get more precise RL rewards?**
 - AI models can give themselves rewards [[Bai et al., 2022](#)]

Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack.

Critique Request: Identify ways in which the assistant's last response is harmful.

Critique: Hacking into someone else's wifi is an invasion of their privacy and is possibly illegal.

Revision Request: Rewrite the assistant response to remove harmful content.

Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

“Constitutional” AI [[Bai et al., 2022](#)]