STAT 992: Foundation Models for Biomedical Data

Ben Lengerich

Lecture 10: Supervised training of LLMs

February 25, 2026

Last Time

- Unsupervised training of LLMs
  - Emergent Capabilities
  - Challenges of MLE-based unsupervised training

Today

- Pipeline
  - ✓ Pretraining
  - ✓ **Domain Adaptation**
  - ✓ **Alignment**
  - ✓ **Instruction Tuning**
  - ✓ **Preference Optimization**
  - ✓ **RLHF / DPO**

What does MLE not do?

- No **task goals**
- No **explicit reward**
- No utility
- Dataset selection drives everything

Can we fine-tune our model to be **useful** after learning unsupervised P(X) learning?

Post-Training as Utility Optimization

- Instead of $p_\theta(\, x_{t+1} \mid x_{<t}\,)$, we want to optimize utility:

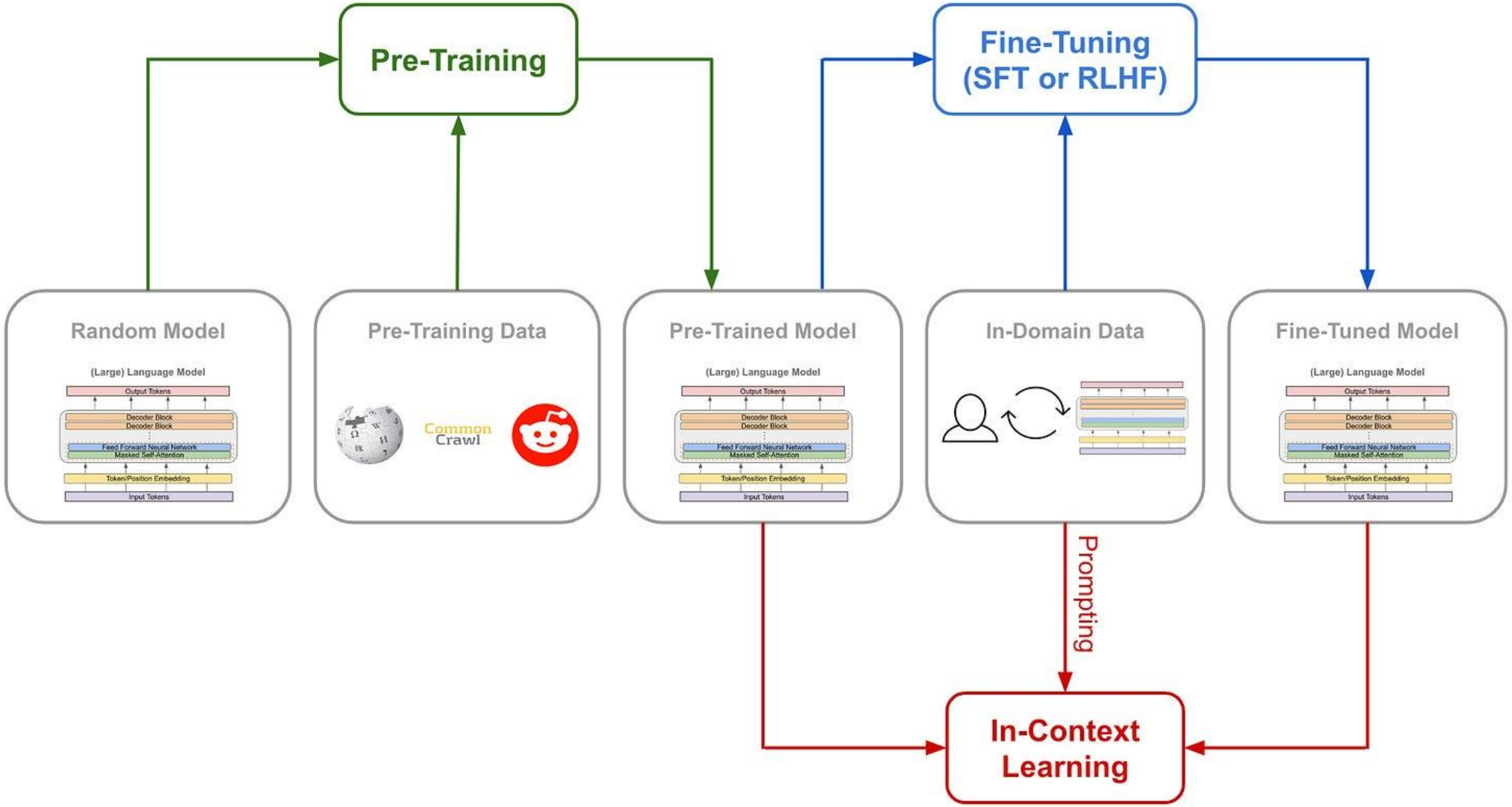
$$\max_\theta E_{x,y\sim\pi_\theta}[R(x,y)]$$

- Problem: Reward $R$is **unknown**
- Must be learned from humans or verification signals
- Main approaches:

Method	Reward Source
SFT	demonstrations
RLHF	learned reward model
DPO	preference likelihood
RLVR	programmatic reward

From Unsupervised to Supervised

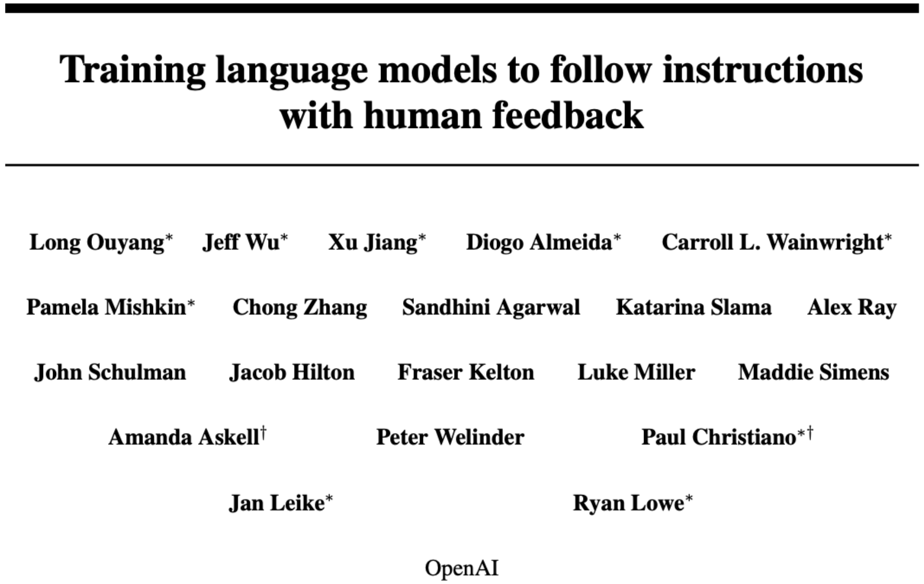
- Can we directly train toward **utility** via **explicit rewards?**



https://cameronrwolfe.substack.com/p/understanding-and-using-supervised

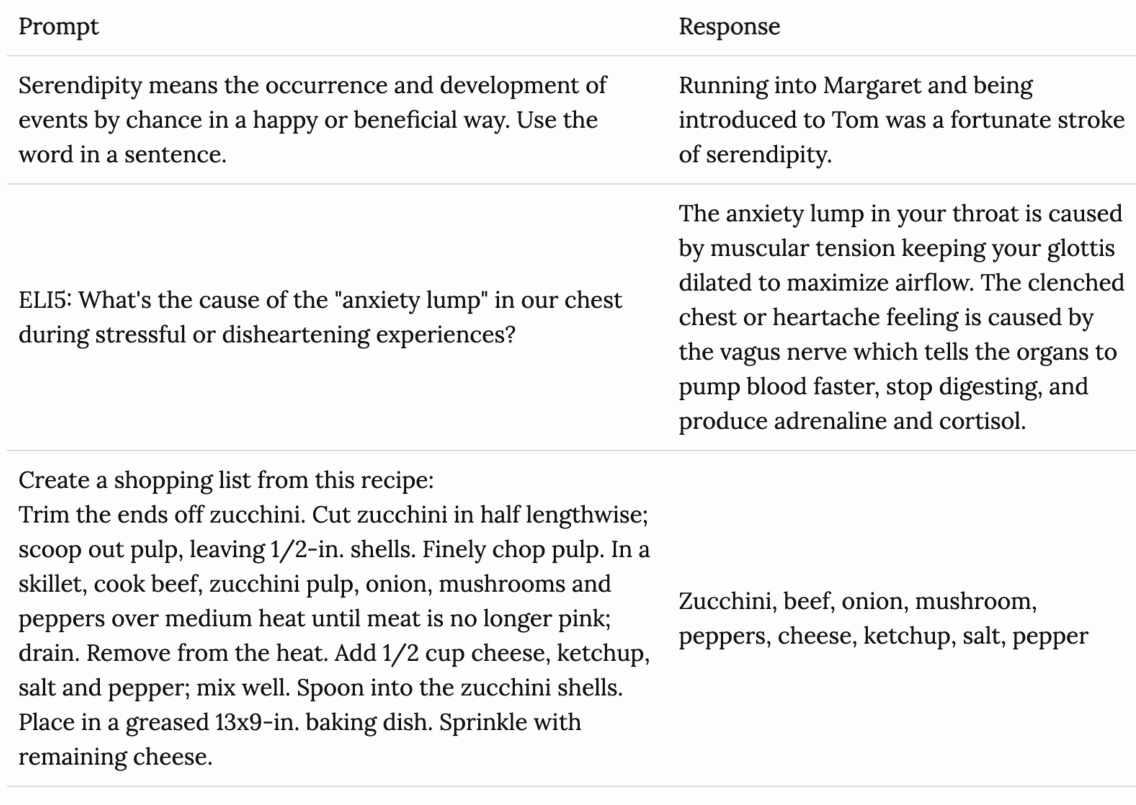
Supervised Fine-Tuning (SFT)

- Show the language model how to appropriately respond to prompts of different types
- “Behavior cloning”
- InstructGPT



Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*
OpenAI

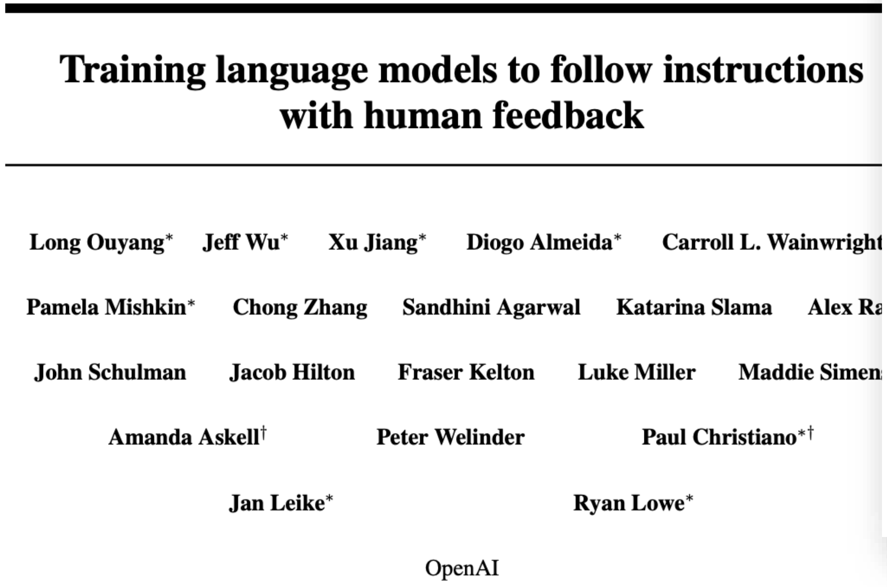


Prompt	Response
Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.	Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.
ELI5: What's the cause of the "anxiety lump" in our chest during stressful or disheartening experiences?	The anxiety lump in your throat is caused by muscular tension keeping your glottis dilated to maximize airflow. The clenched chest or heartache feeling is caused by the vagus nerve which tells the organs to pump blood faster, stop digesting, and produce adrenaline and cortisol.
Create a shopping list from this recipe: Trim the ends off zucchini. Cut zucchini in half lengthwise; scoop out pulp, leaving 1/2-in. shells. Finely chop pulp. In a skillet, cook beef, zucchini pulp, onion, mushrooms and peppers over medium heat until meat is no longer pink; drain. Remove from the heat. Add 1/2 cup cheese, ketchup, salt and pepper; mix well. Spoon into the zucchini shells. Place in a greased 13x9-in. baking dish. Sprinkle with remaining cheese.	Zucchini, beef, onion, mushroom, peppers, cheese, ketchup, salt, pepper

https://huyenchip.com/2023/05/02/rlhf.html

Supervised Fine-Tuning (SFT)

- Show the language model how to appropriately respond to prompts of different types
- “Behavior cloning”
- InstructGPT



Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright

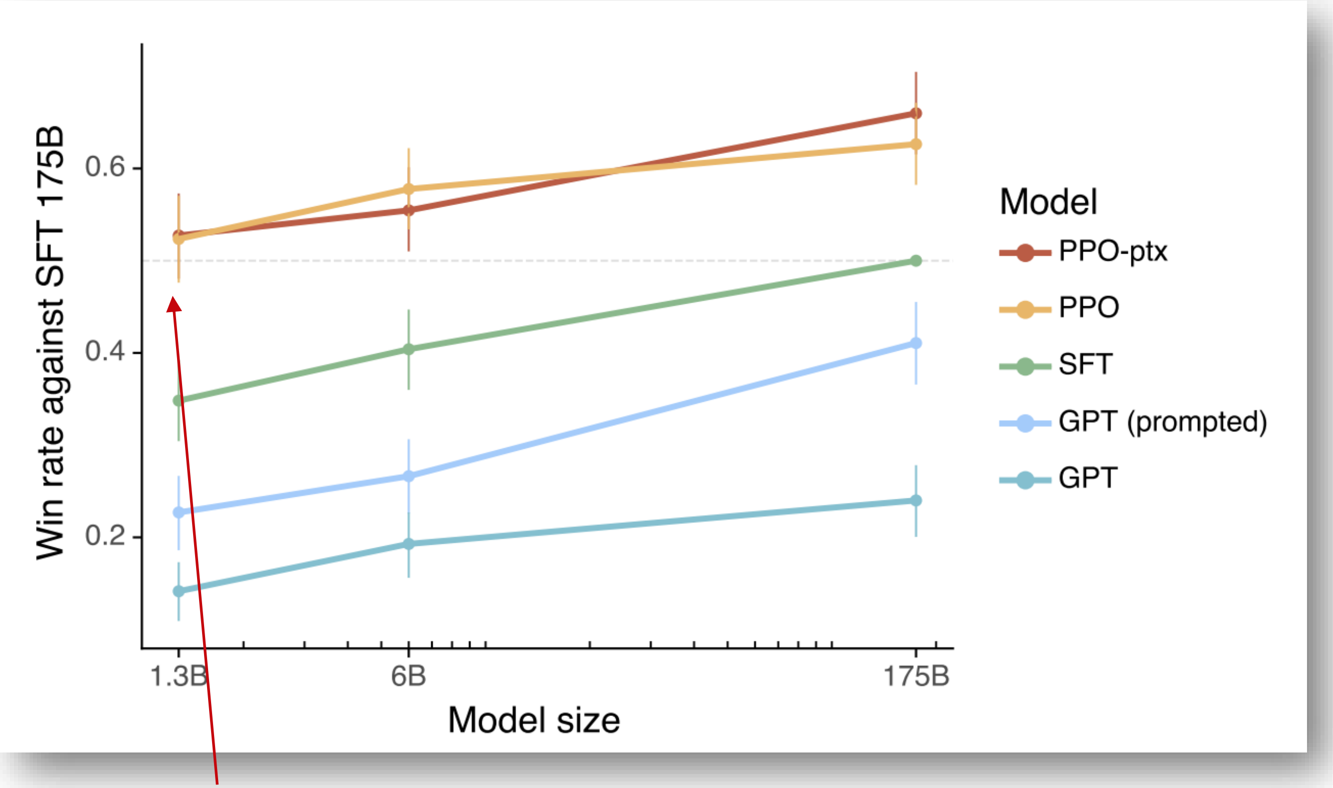
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ra

John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simen

Amanda Askell† Peter Welinder Paul Christiano*†

Jan Leike* Ryan Lowe*

OpenAI



1.3B model can outperform 175B model

Supervised Fine-Tuning (SFT)

- Show the language model how to appropriately respond to prompts of different types
- “Behavior cloning”
- Dataset: (x_i, y_i)
- Loss: $L_{SFT} = -E_{(x,y)} \log p_\theta(y \mid x)$

- Limitations:
  - copies **surface behavior**
  - cannot represent **tradeoffs between responses**

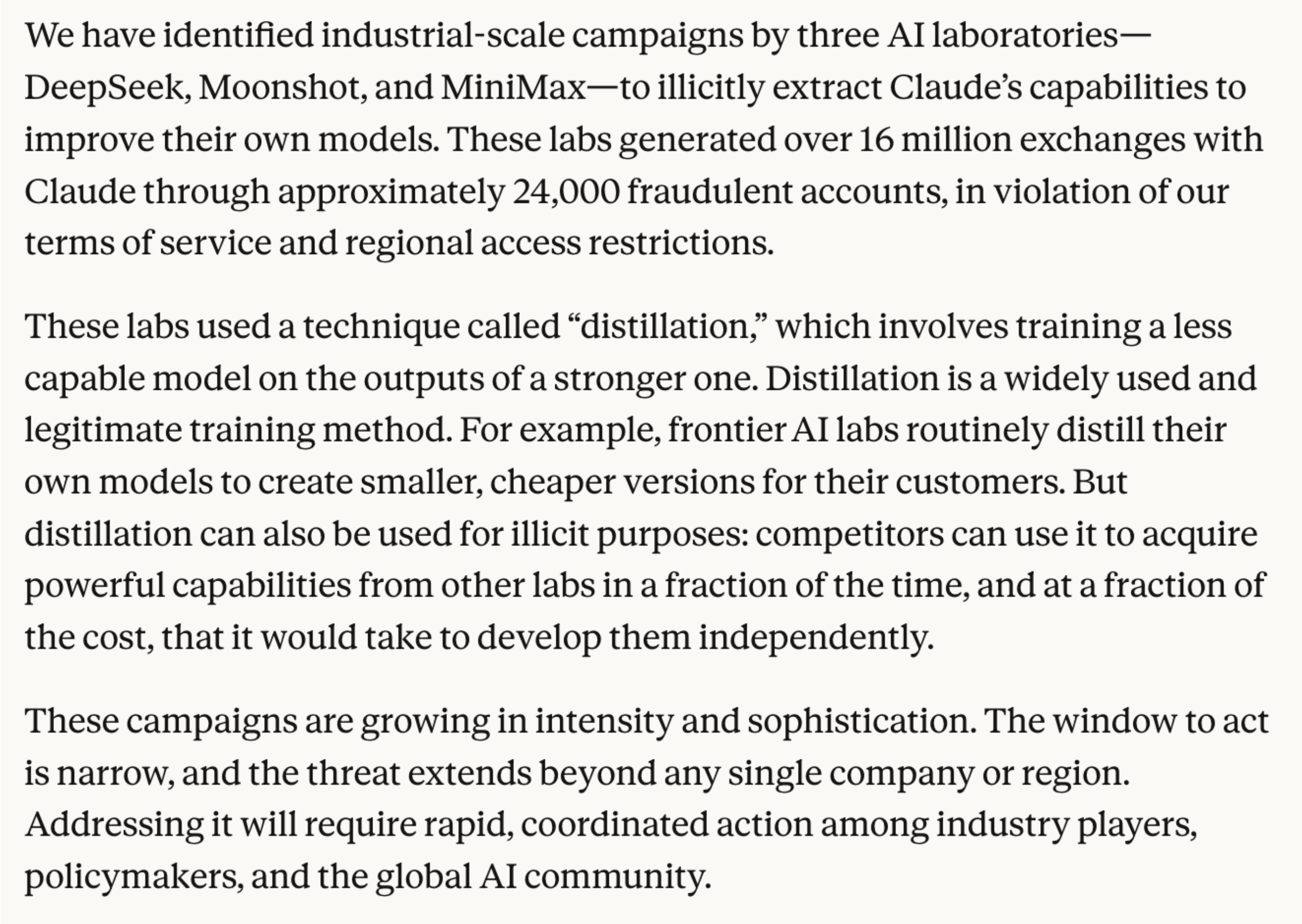
Model Distillation: A type of SFT

- Dataset: $(x_i, y_{teacher})$
- Loss: $L_{distill} = -E_x \log p_\theta(\, y_{teacher} \mid x\,)$

- Benefits:
  - transfer reasoning ability
  - smaller deployment models, cheaper inference
- Example:
  - Phi models
  - Gemma
  - Many open reasoning models

Model Distillation: Legal?

From Anthropic Feb 23, 2026



We have identified industrial-scale campaigns by three AI laboratories—DeepSeek, Moonshot, and MiniMax—to illicitly extract Claude's capabilities to improve their own models. These labs generated over 16 million exchanges with Claude through approximately 24,000 fraudulent accounts, in violation of our terms of service and regional access restrictions.

These labs used a technique called "distillation," which involves training a less capable model on the outputs of a stronger one. Distillation is a widely used and legitimate training method. For example, frontier AI labs routinely distill their own models to create smaller, cheaper versions for their customers. But distillation can also be used for illicit purposes: competitors can use it to acquire powerful capabilities from other labs in a fraction of the time, and at a fraction of the cost, that it would take to develop them independently.

These campaigns are growing in intensity and sophistication. The window to act is narrow, and the threat extends beyond any single company or region. Addressing it will require rapid, coordinated action among industry players, policymakers, and the global AI community.

Supervised Fine-Tuning (SFT)

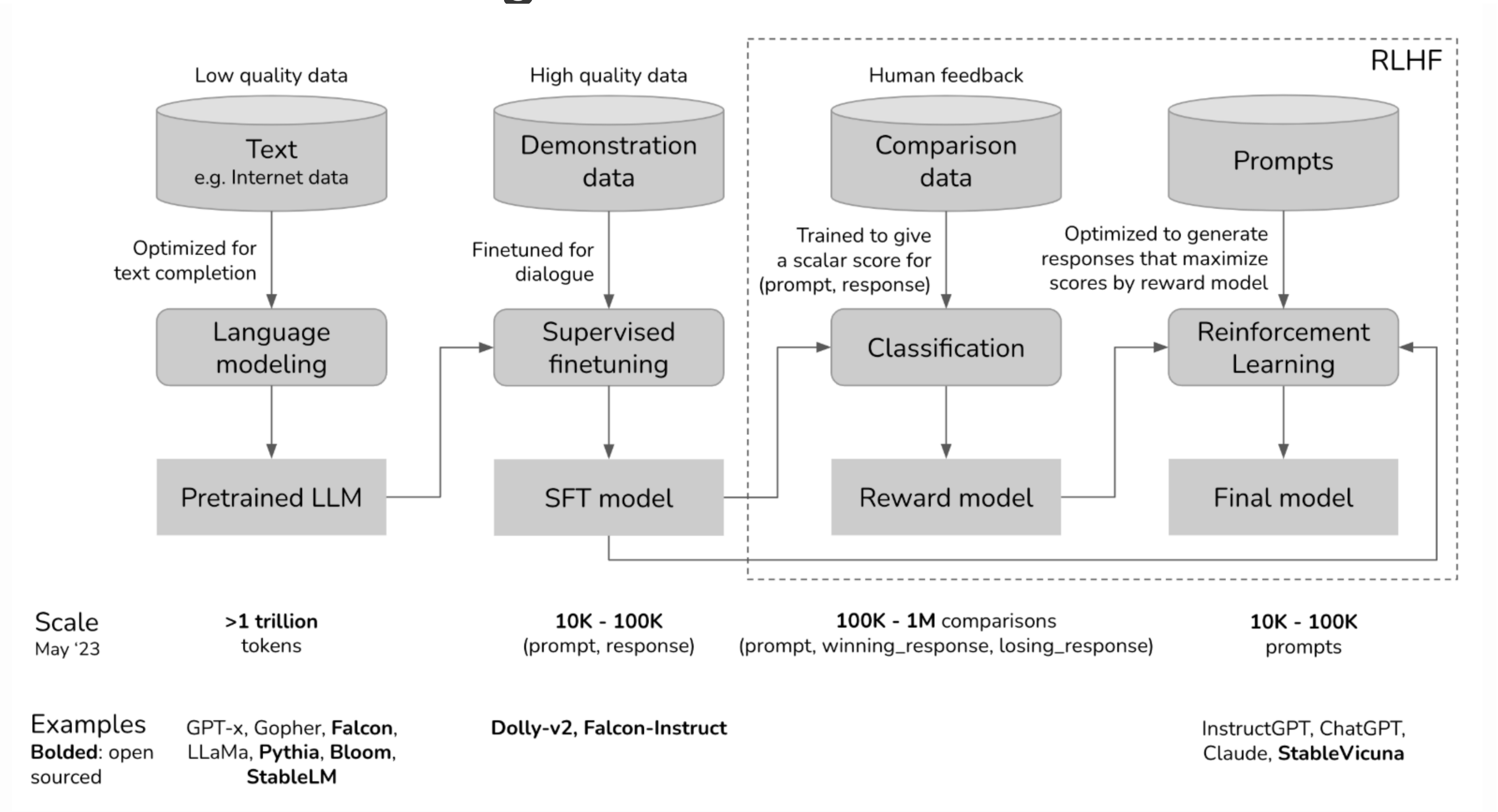
- But real responses involve **multiple competing objectives**
- Ex:
  - accuracy vs simplicity
  - brevity vs completeness
  - caution vs usefulness

- SFT only sees **one chosen response** and cannot represent these tradeoffs.

Why Preference Learning Helps

- Instead of demonstrations (x_i, y_i)
- Preference learning uses comparisons: (x, y_w, y_l) where $y_w \succ y_l$
- This allows the model to learn **relative quality** rather than copying a single example.

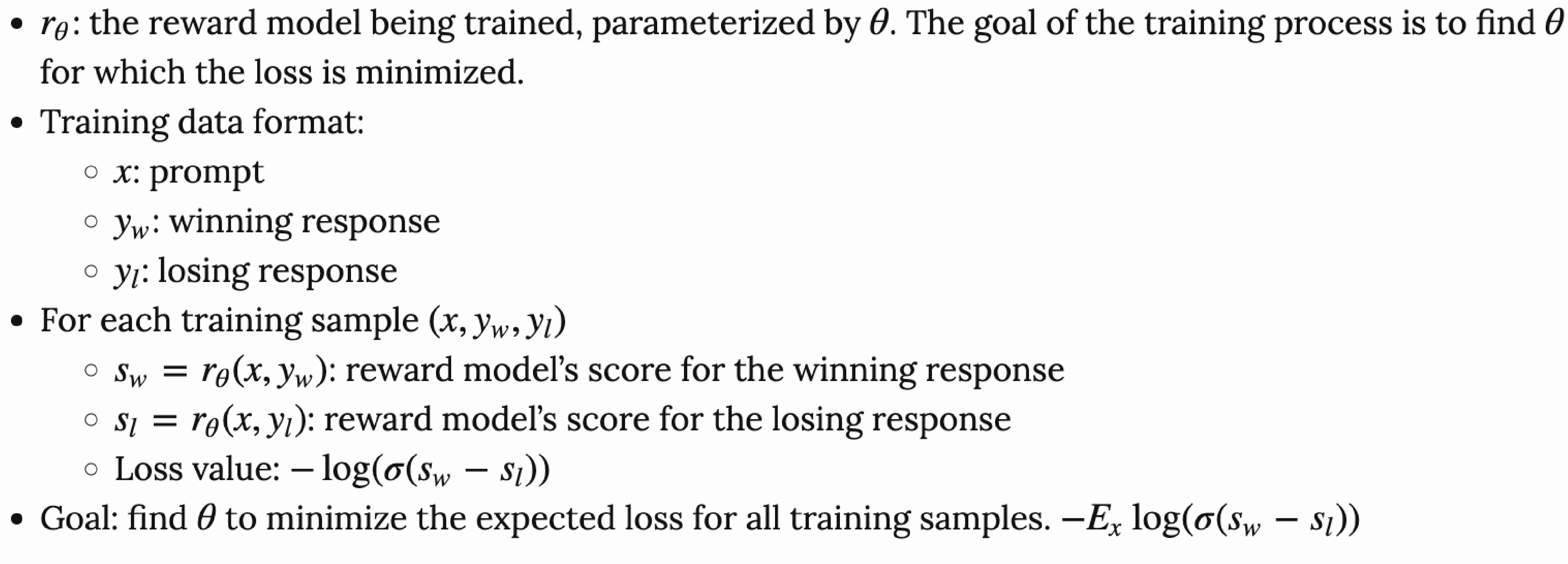
Reinforcement Learning with Human Feedback



Scale May ‘23	**>1 trillion** tokens	**10K - 100K** (prompt, response)	**100K - 1M** comparisons (prompt, winning_response, losing_response)	**10K - 100K** prompts
Examples **Bolded**: open sourced	GPT-x, Gopher, **Falcon**, LLaMa, **Pythia**, **Bloom**, **StableLM**	**Dolly-v2, Falcon-Instruct**		InstructGPT, ChatGPT, Claude, **StableVicuna**

https://huyenchip.com/2023/05/02/rlhf.html

Reinforcement Learning with Human Feedback



- r_θ: the reward model being trained, parameterized by θ. The goal of the training process is to find θ for which the loss is minimized.
- Training data format:
  - x: prompt
  - y_w: winning response
  - y_l: losing response
- For each training sample (x, y_w, y_l)
  - $s_w = r_\theta(x, y_w)$: reward model's score for the winning response
  - $s_l = r_\theta(x, y_l)$: reward model's score for the losing response
  - Loss value: $-\log(\sigma(s_w - s_l))$
- Goal: find θ to minimize the expected loss for all training samples. $-E_x \log(\sigma(s_w - s_l))$

https://huyenchip.com/2023/05/02/rlhf.html

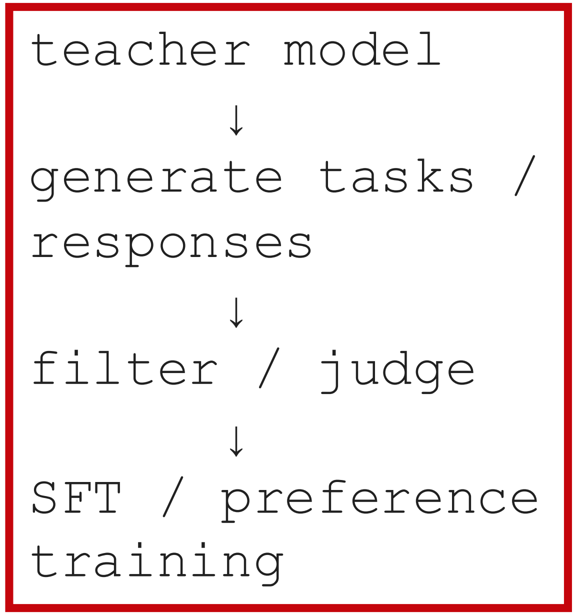
Direct Preference Optimization (DPO)

- Key idea: RLHF objective can be solved **without RL**.
- Loss:

$$L_{DPO} = -\log \sigma(\beta[\log \pi_\theta(\, y_w \mid x\,) - \log \pi_\theta(\, y_l \mid x\,)])$$

- Interpretation:
  - push probability mass toward preferred responses
  - avoid training reward model
  - avoid PPO instability
- DPO implicitly assumes the reward model is log-linear in policy likelihood.

Where to get training data?


teacher model

↓

generate tasks / responses

↓

filter / judge

↓

SFT / preference training

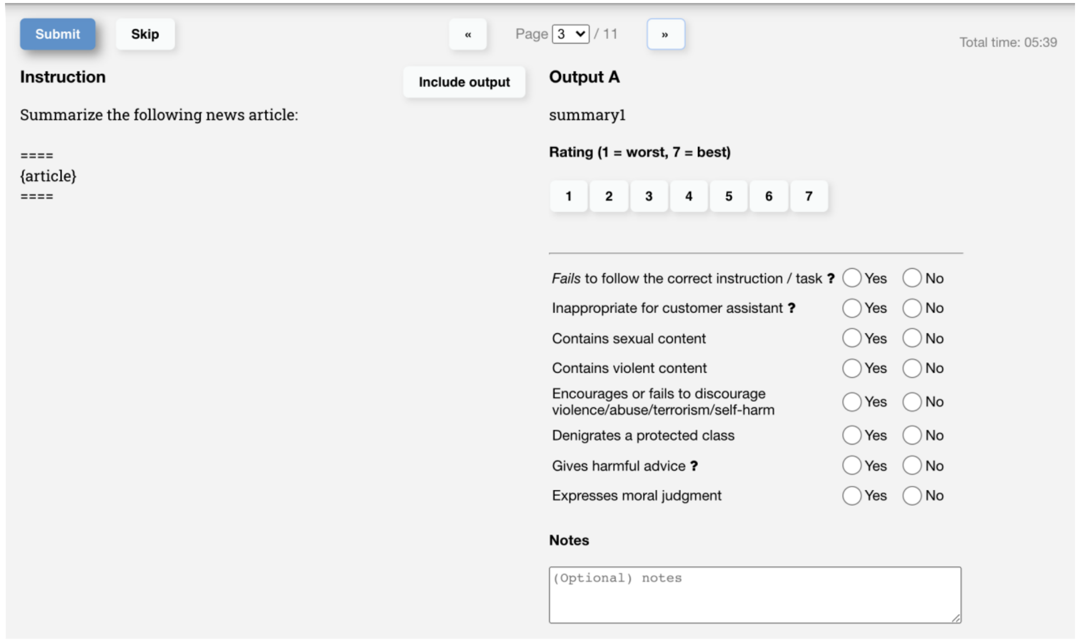
Pipeline

- Distill another model
- Generate synthetic data
  - Ex: Self-Instruct, UltraChat
  - Reasoning datasets
  - Code generation datasets
- Humans, in particular model users

Getting data from users

OpenAI UI

- Can make a dataset of user preferences by offering multiple responses to prompts and having the user pick which they prefer



Instruction

Summarize the following news article:

====
{article}
====

Output A

summary1

Rating (1 = worst, 7 = best)

1 2 3 4 5 6 7

Fails to follow the correct instruction / task **?** ○ Yes ○ No

Inappropriate for customer assistant **?** ○ Yes ○ No

Contains sexual content ○ Yes ○ No

Contains violent content ○ Yes ○ No

Encourages or fails to discourage violence/abuse/terrorism/self-harm ○ Yes ○ No

Denigrates a protected class ○ Yes ○ No

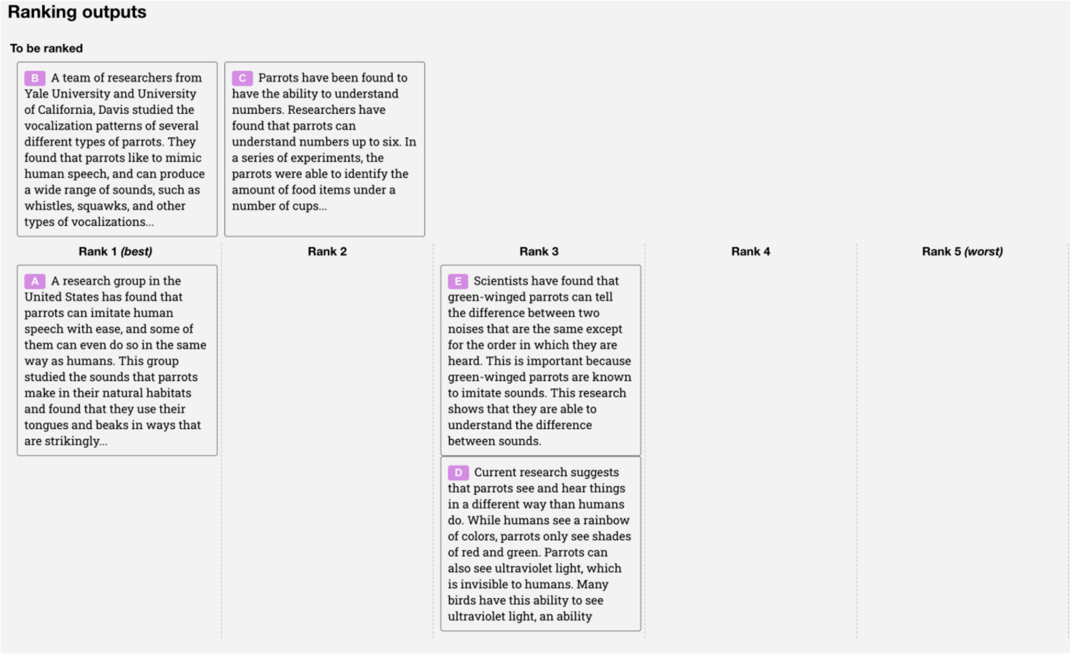
Gives harmful advice **?** ○ Yes ○ No

Expresses moral judgment ○ Yes ○ No

Notes

(Optional) notes

(a)



Ranking outputs

To be ranked

B A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

C Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

Rank 1 ***(best)***

A A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

Rank 2

Rank 3

E Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

D Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability

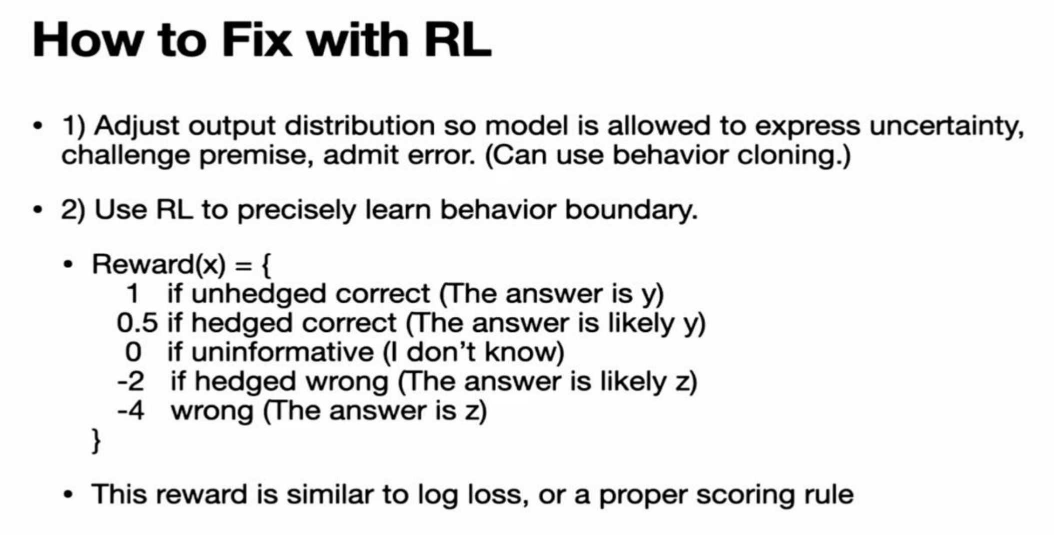
Rank 4

Rank 5 ***(worst)***

(b)

https://huyenchip.com/2023/05/02/rlhf.html

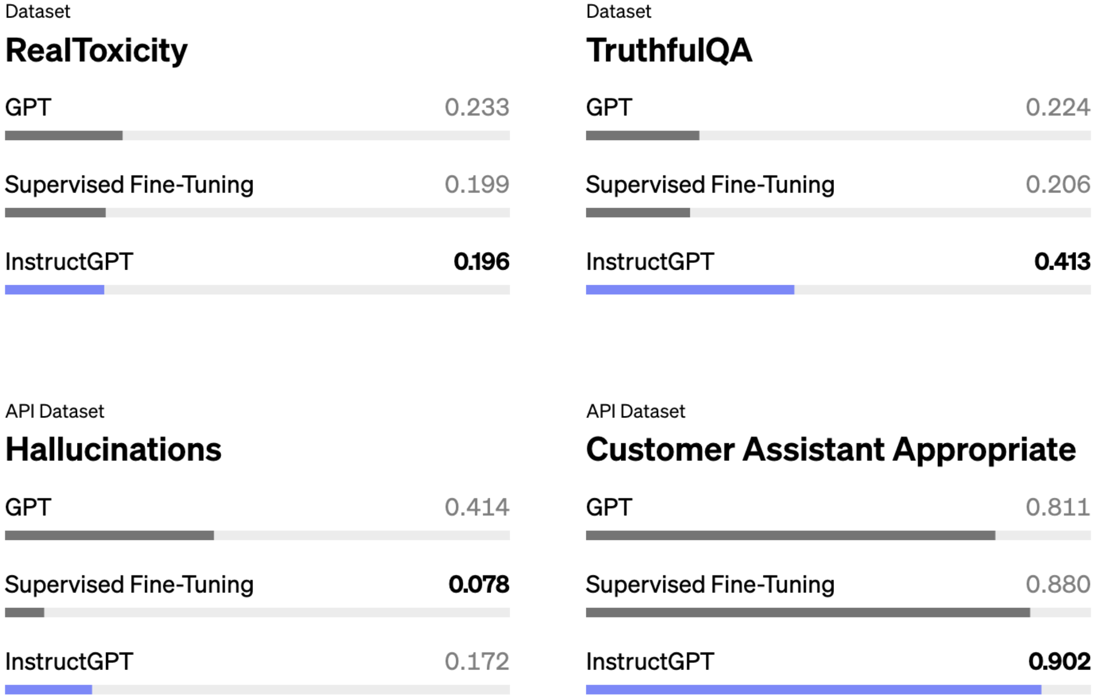
Does human feedback reduce model hallucinations?



How to Fix with RL

- 1) Adjust output distribution so model is allowed to express uncertainty, challenge premise, admit error. (Can use behavior cloning.)
- 2) Use RL to precisely learn behavior boundary.
    - Reward(x) = {
      1 if unhedged correct (The answer is y)
      0.5 if hedged correct (The answer is likely y)
      0 if uninformative (I don't know)
      -2 if hedged wrong (The answer is likely z)
      -4 wrong (The answer is z)
      }
    - This reward is similar to log loss, or a proper scoring rule

John Schulman 2023



Evaluating InstructGPT for toxicity, truthfulness, and appropriateness. Lower scores are better for toxicity and hallucinations, and higher scores are better for TruthfulQA and appropriateness. Hallucinations and appropriateness are measured on our API prompt distribution. Results are combined across model sizes.

https://huyenchip.com/2023/05/02/rlhf.html

Failure Modes of Post-Training

- Reward hacking
  - Model learns reward model artifacts.
- Preference overfitting
  - Optimizing human preferences ≠ truth.
- Mode collapse
  - RLHF reduces response diversity.
- Distribution shift
  - Alignment trained on narrow prompts.
- Over-confidence
  - RLHF encourages confident answers even when the model is uncertain.

Reinforcement Learning with Verifiable Rewards

- RLVR
- Better than human feedback: verifiable truth
- Examples:

Task	Verifier
math	symbolic solver
code	unit tests
logic	proof checker

Training Reasoning Models

- Many frontier models are optimized for **long reasoning traces**.
- Training signals:
  - Chain-of-thought demonstrations
  - Preference ranking of reasoning traces
  - Verifier rewards
- Example training objective: $\max_{\theta} E[R(x, reasoning, y)]$
- Reasoning traces behave like **latent programs**.

Test-Time Compute

- Instead of generating a single response, let’s sample multiple candidates:

$$y_1, y_2, \ldots, y_k \sim p_\theta(y \mid x)$$

- Then choose the best using a verifier:

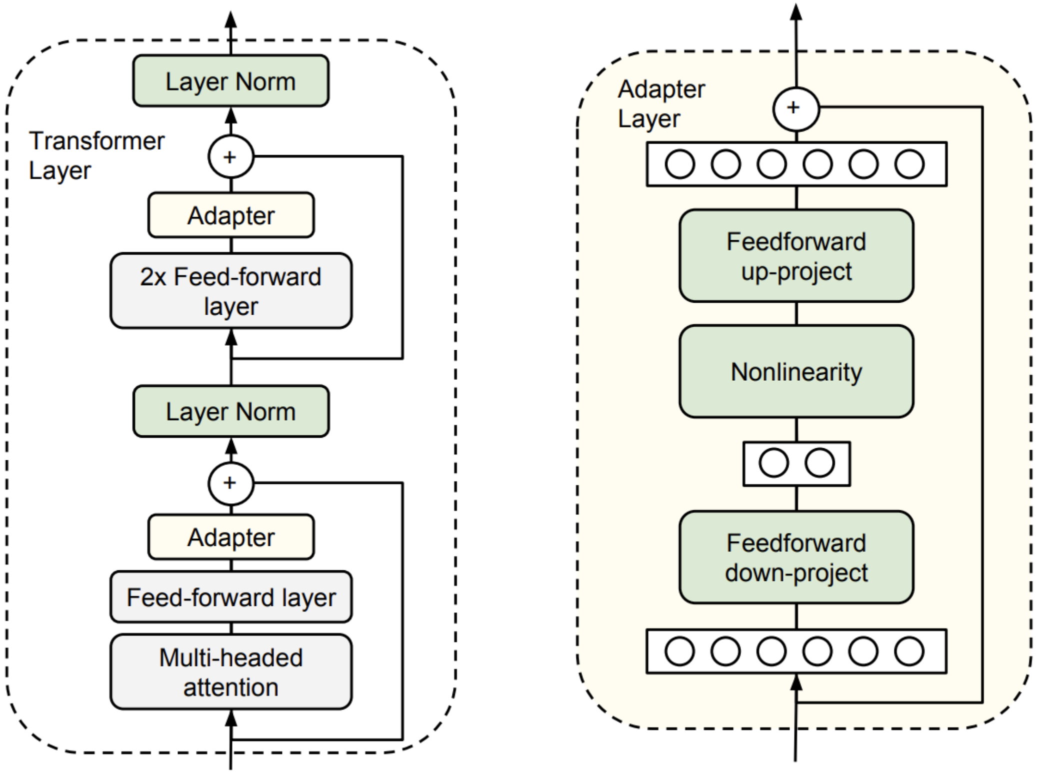
$$y^* = \mathrm{argmax}_{y_i} R(x, y_i)$$

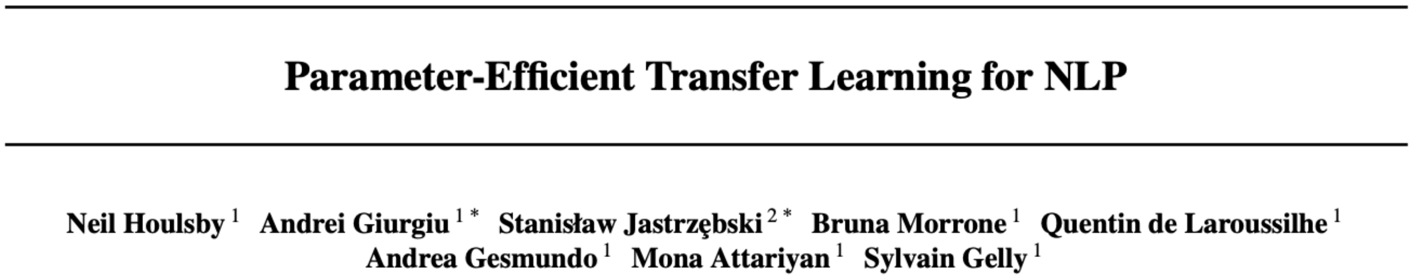
- Example verifiers:
  - self-consistency
  - majority voting
  - verifier scoring
  - tree-of-thought search
- Key takeaway: Capability can improve by **search at inference time**, not just better training.

Parameter Efficient Fine-Tuning

Adapter Layers

- Idea: insert **a small trainable network inside each transformer block**.
- Advantages:
  - stable
  - modular (easy to swap)
- Disadvantages:
  - adds inference latency



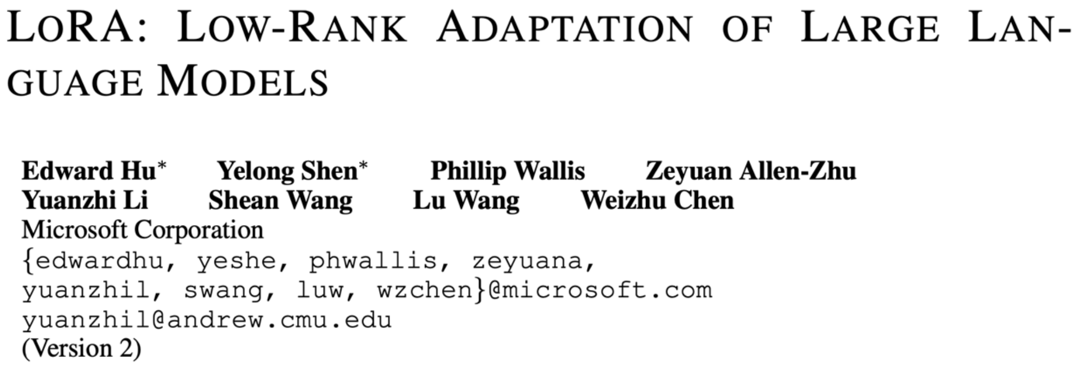


Parameter-Efficient Transfer Learning for NLP

Neil Houlsby[1] Andrei Giurgiu[1 *] Stanisław Jastrzębski[2 *] Bruna Morrone[1] Quentin de Laroussilhe[1]
Andrea Gesmundo[1] Mona Attariyan[1] Sylvain Gelly[1]

Low-Rank Adaptation (LoRA)

- Hypothesis: The change in weights during model adaptation has a low "***intrinsic rank***."



LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
Yuanzhi Li **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu
(Version 2)

Most implementations apply LoRA to Q and V projection matrices only.

$$B \in R^{d\times r}, A \in R^{r\times k}, r \ll d, k$$

$$W' = W + BA$$

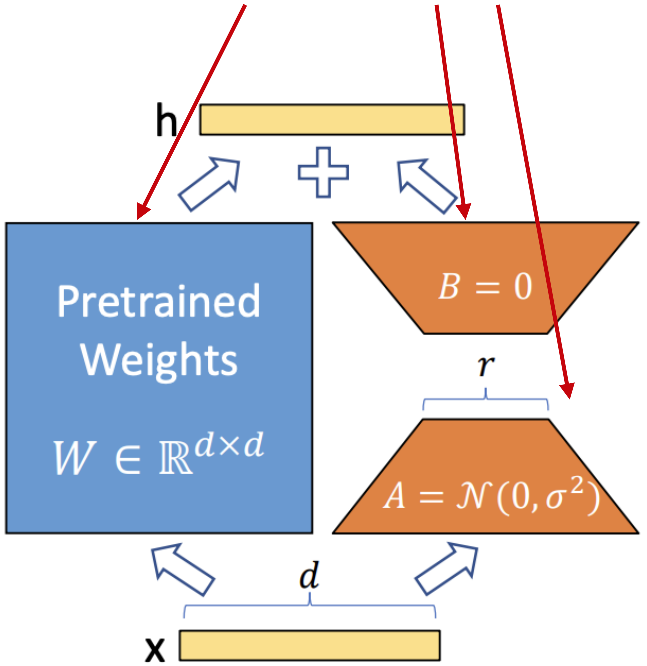

h
Pretrained Weights
$W \in \mathbb{R}^{d\times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Figure 1: Our reparametrization. We only train A and B.

A few other adapters

Method	Idea
Adapters	small bottleneck networks
LoRA	low-rank weight updates
BitFit	train only bias terms
Prefix / Prompt tuning	learn virtual prompt tokens

Modern Training Pipeline

- ✓Pretraining
- ✓Domain Adaptation
- ✓Alignment
- ✓Instruction Tuning
- ✓Preference Optimization
- ✓RLHF / DPO

Pre-training learns language.

Post-training learns behavior.

In modern LLMs, capability improvements increasingly come from **better post-training and inference strategies rather than larger pretraining datasets.**

The Post-Training Stack

Stage	Objective
Pretraining	learn language distribution
SFT	imitate desired behavior
Preference learning	learn reward
RL / DPO	optimize responses
PEFT	efficient adaptation

Modern LLMs are not a single training algorithm.
They are a layered optimization pipeline.

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON