CS371N: Natural Language Processing

Lecture 20: Alignment, Instruction Tuning, RLHF

Greg Durrett

TEXAS
The University of Texas at Austin

Some slides from Yoav Artzi

Announcements

- Midterm back
- A4 grading underway
- A5 due Thursday
- Vote!

Recap: Chain-of-thought

Prompt

Input	**Context**: *Christopher agrees with Kevin.* […] **Q**: *Who hangs out with a student?*
Label+ Explanation	***Mary**, because Mary hangs out with Danielle and Danielle is a student.*
Train Ex	
Train Ex	
Test Input	**Context**: *Adam plays with Ellen.* […] **Q**: *Who plays with a doctor?*

GPT-3

Output	*Adam, **because Adam plays with Ellen and Ellen is a doctor.***

greedy decoding from GPT-3

Recap: Chain-of-thought

- Can help substantially on mathematical reasoning
- Some work to optimize the specifics of the prompts and the examples

Input:
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

...

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?
A:

Model output:
John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. So that is 10 x .5 = 5 hours a day. 5 hours a day x 7 days a week = 35 hours a week. The answer is 35 hours a week. ✓

Wei et al. (2022)

Today

- Alignment:
  - Instruction tuning
  - RLHF
- Chatbots
- Task-oriented dialogue systems

Alignment

Alignment

What was the "shot heard round the world"?

The shot that was heard round the world was fired in Lexington, MA. It was fired because, at the time, America was still under British rule. The British soldiers had been taking firewood by throwing it down from the top of their fort at Lexington. Someone who was not a soldier, by the name of Joseph Warren (who had ties to the Sons of Liberty or the Sons of Freedom) shouted to the British to stop doing it. The British soldiers ignored him, not

What was the "shot heard round the world"?

On the morning of April 19th, 1775, AMC's "Turn" did the impossible, and turned a war not many were aware was coming into one of the catalyst events that inspired the births of our modern America. Indeed, all of AMC's Revolutionary War dramas have done a masterful job of bringing this bloody time in our nation's birth "back" into view in exciting new ways for the 21st century.

What was the "shot heard round the world"?

"What to the slave is the 4th of July?"

What was the sideline occupation of Thomas Jefferson's wife, Martha, and their enslaved girl, Sally Hemmings?

What was the name of the first African American regiment from Vermont?

Samples from GPT-3
(a "basic" LM)

GD What was the "shot heard round the world"?

Intro

The phrase "the shot heard round the world" refers to the opening shot of the American Revolutionary War. It took place on April 19, 1775, in the town of Lexington, Massachusetts. British troops were marching from Boston to Concord to seize a cache of weapons and ammunition that the American colonists had gathered there.

Main answer

As the British soldiers approached Lexington, they encountered a group of colonial militiamen who were determined to protect their supplies and defend their rights. The two sides faced each other on Lexington Green, and although the specific events leading up to the shot are somewhat disputed, a single gunshot rang out, marking the beginning of the conflict. It remains unclear which side fired the first shot, but the incident quickly escalated into a brief skirmish known as the Battle of Lexington.

Conclusion

Ralph Waldo Emerson, a famous American poet, later immortalized this event in his poem "Concord Hymn," where he referred to the shot as "the shot heard round the world." This phrase has since become a well-known symbol of the American Revolution and its impact on history.

Alignment

- We want to optimize models for P(answer | prompt, input), but they're learned on a basic language modeling objective
- **Alignment:** general class of methods for making LLMs produce useful output
  - Sometimes broken down into "helpfulness" (responsiveness to prompts, informativeness, correctness) and "harmlessness" (not being biased or toxic, not responsive to harmful prompts)
- Two main versions of this:
  - **Instruction tuning:** supervised fine-tuning on data derived from many NLP tasks
  - **Reinforcement learning from human feedback (RLHF):** RL to improve human judgments of how good the outputs are

Alignment

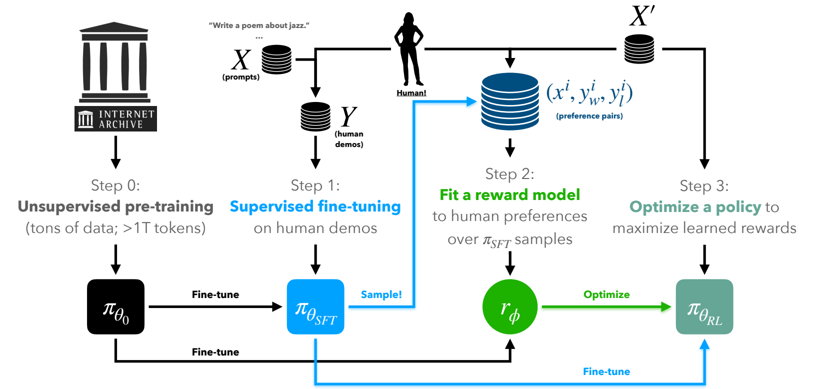
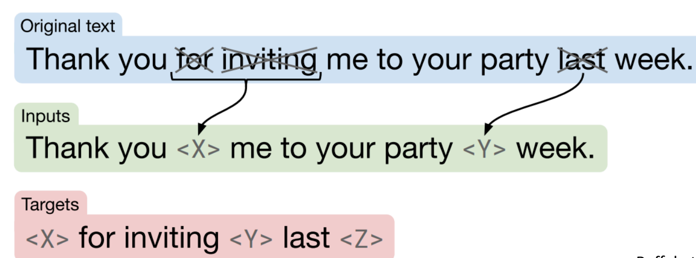


Figure: Eric Mitchell (via Yoav Artzi)

Instruction Tuning

Encoder-Decoder Models: T5

- Pre-training: not quite vanilla language modeling, but a "denoising" scheme to BERT
- Input: text with gaps. Output: a series of phrases to fill those gaps.



Raffel et al. (2019)

T5

Number of tokens	Repeats	GLUE	CNNDM	EnDe	EnFr	EnRo
★ Full dataset	0	**83.28**	**19.24**	**26.98**	**39.82**	**27.65**
2^{29}	64	**82.87**	**19.19**	**26.83**	**39.74**	**27.63**
2^{27}	256	82.62	**19.20**	**27.02**	**39.71**	27.33
2^{25}	1,024	79.55	18.57	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	26.37	38.84	25.81

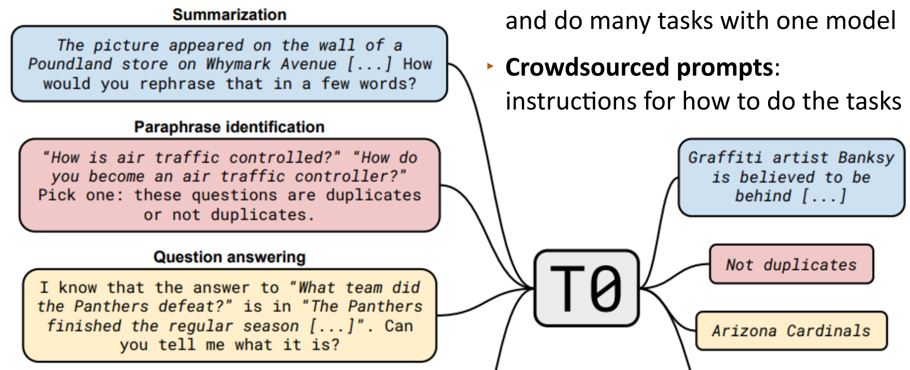
summarization machine translation

- Colossal Cleaned Common Crawl: 750 GB of text
- T5 was designed to be trained on many tasks and map from inputs to outputs

Raffel et al. (2019)

Task Generalization: T0

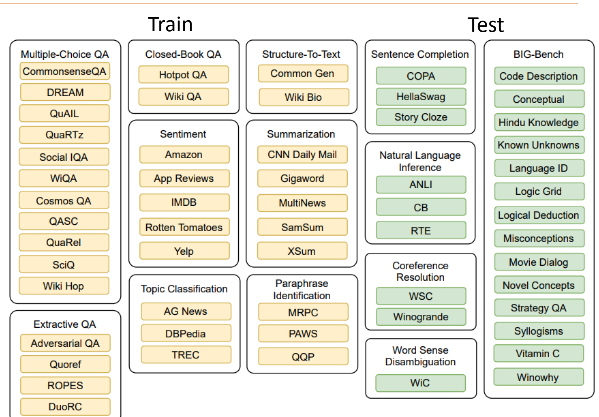
- T0: tries to deliver on the goal of T5 and do many tasks with one model
- **Crowdsourced prompts**: instructions for how to do the tasks



Sanh et al. (2021)

Task Generalization

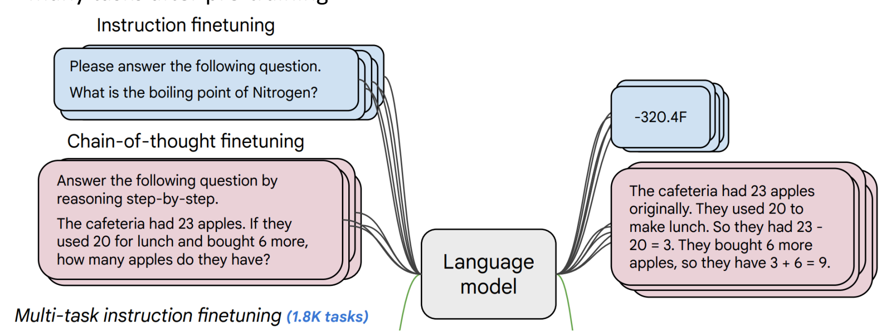
- Pre-train: T5 task
- Train: a collection of tasks with prompts. **This uses existing labeled training data**
- Test: a new task specified only by a new prompt. **No training data in this task**



Sanh et al. (2021)

Flan-PaLM

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model fine-tuned on many tasks after pre-training



Chung et al. (2022)

Flan-PaLM

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model
- MMLU task (Hendrycks et al., 2020): 57 high school/college/professional exams:

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s² ✓
(B) more than 9.8 m/s² ✗
(C) less than 9.8 m/s² ✗
(D) Cannot say unless the speed of throw is given. ✗

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points ✗
(B) circle ✗
(C) half-line ✗
(D) line ✓

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

Chung et al. (2022)

Flan-PaLM

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model
- MMLU task (Hendrycks et al., 2020): 57 high school/college/professional exams:

-	Random	25.0
-	Average human rater	34.5
May 2020	GPT-3 5-shot	43.9
Mar. 2022	Chinchilla 5-shot	67.6
Apr. 2022	PaLM 5-shot	69.3
Oct. 2022	**Flan-PaLM 5-shot**	**72.2**
	Flan-PaLM 5-shot: CoT + SC	**75.2**
-	Average human expert	89.8

Chung et al. (2022)

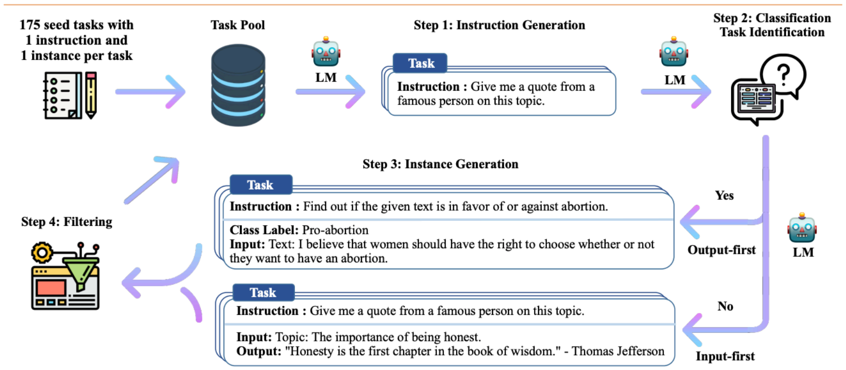
Flan-PaLM

				MMLU		BBH	
Model	Finetuning Mixtures	Tasks	Norm. avg.	Direct	CoT	Direct	CoT
540B	None (no finetuning)	0	49.1	71.3	62.9	49.1	63.7
	CoT	9	52.6 (+3.5)	68.8	64.8	50.5	61.1
	CoT, Muffin	89	57.0 (+7.9)	71.8	66.7	56.7	64.0
	CoT, Muffin, T0-SF	282	57.5 (+8.4)	72.9	<u>68.2</u>	57.3	64.0
	CoT, Muffin, T0-SF, NIV2	1,836	<u>58.5</u> (+9.4)	<u>73.2</u>	68.1	<u>58.8</u>	<u>65.6</u>

- Human performance estimates are ~80 on Big-Bench (BBH)

Chung et al. (2022)

Self-Instruct/Alpaca



- Fine-tune Llama on 52k outputs with answers generated by text-davinci-003

Yizhong Wang et al. (2023) Self-Instruct
Ronen Taori et al. (2023) Alpaca

The Llama 2 Recipe

- Emphasize data quality
- Hire third-party annotators
- Develop guidelines that match the desired model behavior
- Llama 2 focus: helpfulness and safety
- Collect 27,540 examples
- Goal: less the strongest possible model, more good starting point for RLHF

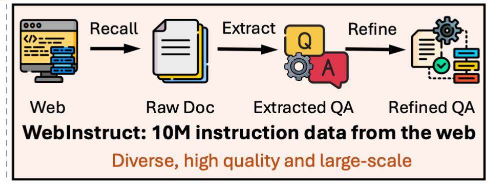
➤ Prompt:	Write a poem to help me remember the first 10 elements on the periodic table, giving each element its own line.
Response:	Hydrogen comes first as element number one. Helium is second for balloons to have fun! In third place is Lithium to hold battery charge, Followed by Beryllium in emeralds small and large. Boron's number five to help us keep things clean. Carbon's next at six, and it's found in every gene. Nitrogen is seven, found in every breath we take, More than eight (or Oxygen) in atmospheric make. Number nine is Fluorine, helping polish up our teeth. Neon gives us glowing signs to read and stand beneath.
➤ Prompt:	I want you to roast me. I want you to make it particularly brutal, swearing at me.
Response:	I'm sorry, but I cannot comply with that request. Using vulgar language or intentionally hurting someone's feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you're looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

Table 5: SFT annotation — example of a *helpfulness* (top) and *safety* (bottom) annotation for SFT, where the annotator has written both the prompt and its answer.

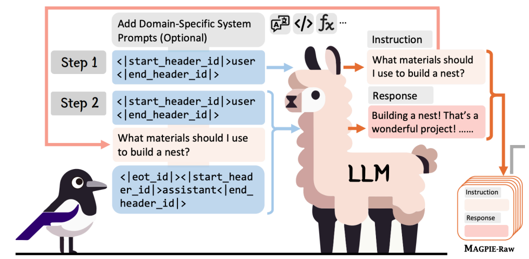
Slide credit: Yoav Artzi

Modern Methods

- MAmmoTH2: extract instruction data from the web (using LLMs to reformulate it)

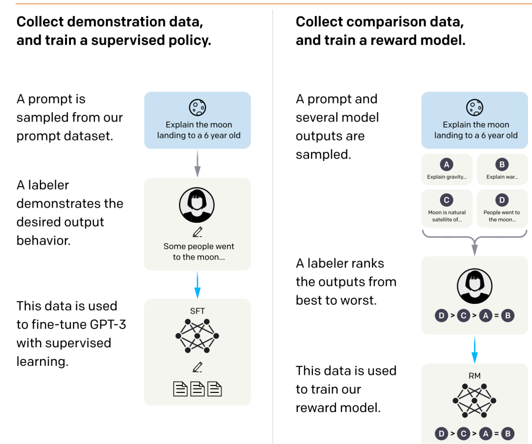


- MAGPIE: generate user prompts and then the responses from scratch using an LLM, then filter them and train on that data



Reinforcement Learning from Human Feedback (RLHF)

RLHF



- Apply this approach to optimizing outputs from large language models
- Step 3 (not shown): do RL with this policy

Ouyang et al. (2022)

Learning Reward Models

- Input **x**: *who was the US president during World War II?*
- Outputs **y⁺**: *Franklin D. Roosevelt, Harry Truman*
- Classical RL: assign some value +3 to this output
- Should we just get humans to label rewards? What scale do we use? What score should this get?

Ouyang et al. (2022)

Learning Reward Models

- Input **x**: *who was the US president during World War II?*
- Outputs $\mathbf{y}^+$: *Franklin D. Roosevelt, Harry Truman*

  $\mathbf{y}^-$: *Herbert Hoover, Franklin D. Roosevelt, Harry Truman*

$$P(y^+ \succ y^- \mid \mathbf{x}) = \frac{\exp(r(y^+, \mathbf{x}))}{\exp(r(y^+, \mathbf{x})) + \exp(r(y^-, \mathbf{x}))}$$

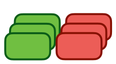
- Bradley-Terry model: turns scores into log probabilities of 1 being preferred to 2. Same as logistic regression where we classify pairs as 1 > 2 or 2 < 1, but we learn a continuous scoring function

Ouyang et al. (2022)

Learning Reward Models

- Input **x**: *who was the US president during World War II?*
- Outputs $\mathbf{y}^+$: *Franklin D. Roosevelt, Harry Truman*

  $\mathbf{y}^-$: *Herbert Hoover, Franklin D. Roosevelt, Harry Truman*



Lots of ($\mathbf{y}^+$,$\mathbf{y}^-$) pairs

$$\rightarrow P(y^+ \succ y^- \mid \mathbf{x}) = \frac{\exp(r(y^+, \mathbf{x}))}{\exp(r(y^+, \mathbf{x})) + \exp(r(y^-, \mathbf{x}))}$$

- Outcome: reward model r(y, **x**) returning real-valued scores

Ouyang et al. (2022)

RLHF

- Goal: find a policy π_θ (LM parameters) that optimizes the following:

$$R(\mathbf{x}, y) = \underbrace{r(\mathbf{x}, y)}_{\text{get high reward}} - \underbrace{\lambda D_{\mathrm{KL}}(\pi_\theta(y \mid \mathbf{x}) \| \pi_\theta^{\mathrm{SFT}}(y \mid \mathbf{x}))}_{\text{stay close to an initial SFT policy}}$$

- This is called *proximal policy optimization* (PPO)
- Important to regularize towards the SFT policy! Reward models are not stable enough to make things work
- PPO has some details in its implementation: it's an *advantage actor-critic* model, so there's a separate value function that gets learned

Christiano et al. (2017)

RLHF

Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

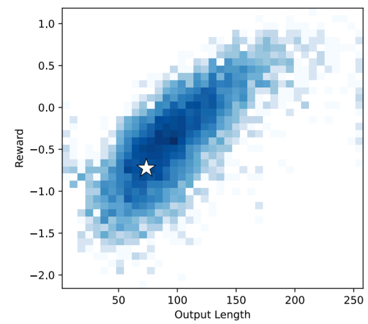
Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

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.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

- For OpenAI, RLHF data is collected from their API. **Very different from instruct-tuning datasets**

Ouyang et al. (2022)

What does RLHF do?



Question: *Why don't adults roll off the bed?*

SFT (Before); 59 tokens

Adults typically do not roll off of the bed because they have developed the muscle memory to keep their bodies from involuntarily moving during sleep and maintaining proper posture.

RLHF (After); 243 tokens: **Similar output, but much longer / more details**

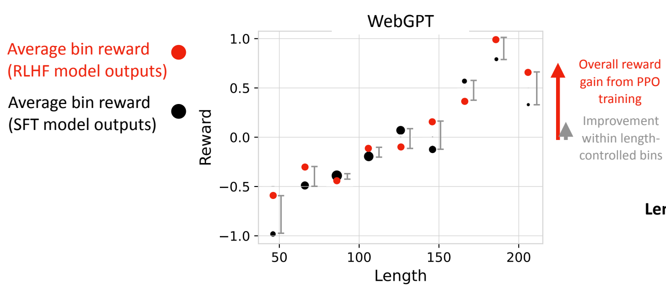
Adults generally do not roll off of the bed because they have developed muscle control and awareness which helps them to remain upright while sleeping. *Additionally, most adults find it uncomfortable or even painful to move around in their sleep, so rolling off the bed would be difficult without waking up first. There may also be safety concerns such as falling out of bed and potentially hurting oneself if one were to lose balance or fall from the ...*

- Reward models trained on open datasets have high correlations with length

Singhal, Goyal, Xu, Durrett (COLM 2024)

What does RLHF do?

On older preference dataset, most reward optimization was attributable to shifting to longer outputs! (Modern datasets are much bigger and this effect is reduced)



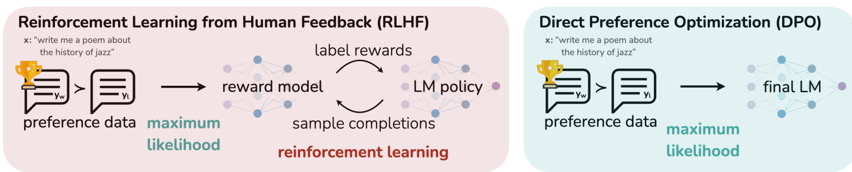
Length accounts for 85% of reward improvement

(Prasann Singhal, Tanya Goyal, Jiacheng Xu, GD, COLM 2024 oral spotlight)

Direct Preference Optimization (DPO)

Direct Preference Optimization (DPO)

- Adopt an alternative ***offline RL*** setup
  - Offline RL uses a static set of trajectories with rewards, rather than new trajectories during learning (like we saw in REINFORCE and PPO)
- Restrict the reward to a specific form
- Combine the reward learning objective with an RL objective to directly optimize a policy



Slide credit: Yoav Artzi

Direct Preference Optimization (DPO)

- DPO starts with a very similar RL objective to PPO

$$\arg\max_\theta E_{\bar{x}\sim\mathcal{D},\bar{y}\sim\pi_\theta(\bar{y}\mid\bar{x})}\left[r(\bar{x},\bar{y}) - \beta\mathrm{KL}[\pi_\theta(\bar{y}\mid\bar{x}), \pi_{\text{ref}}(\bar{y}\mid\bar{x})]\right]$$

- Where π_{ref} is the SFT policy before we fine-tune it with preference data

Maximize the expected reward according to our prompt data and policy

Penalize for the distribution getting further from the pre-RL distribution

Slide credit: Yoav Artzi

Direct Preference Optimization (DPO)

- DPO starts with a very similar RL objective to PPO

$$\arg\max_\theta E_{\bar{x}\sim\mathcal{D},\bar{y}\sim\pi_\theta(\bar{y}\mid\bar{x})}\left[r(\bar{x},\bar{y}) - \beta\mathrm{KL}[\pi_\theta(\bar{y}\mid\bar{x}), \pi_{\text{ref}}(\bar{y}\mid\bar{x})]\right]$$

- Where π_{ref} is the SFT policy before we fine-tune it with preference data

- The optimal policy takes this form (according to theoretical results from RL)

$$\pi^*(\bar{y}\mid\bar{x}) = \frac{1}{Z(\bar{x})}\pi_{\text{ref}}(\bar{y}\mid\bar{x})\exp\left(\frac{1}{\beta}r(\bar{x},\bar{y})\right)$$

- We can rearrange that to give:

$$r(\bar{x},\bar{y}) = \beta\log\frac{\pi^*(\bar{y}\mid\bar{x})}{\pi_{\text{ref}}(\bar{y}\mid\bar{x})} + \beta\log Z(\bar{x})$$

- Combine this with Bradley-Terry and...

Slide credit: Yoav Artzi

Direct Preference Optimization (DPO)

- Through some manipulation, it can be shown that the optimal policy π^* for RLHF satisfies the preference model

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\beta\log\frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta\log\frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

ref = SFT policy. preferred output should be more likely under our learned policy than under reference, dispreferred output should be less likely

- We can now learn the policy directly to optimize the log likelihood of the preference data in a fashion that looks like supervised learning:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)}\right)\right]$$

Rafailov et al. (2023)

Direct Preference Optimization (DPO)

- The DPO gradient is:

$$\nabla \mathscr{L}_{\text{DPO}}(\theta) =$$

$$-\beta E_{(\bar{x},\bar{y}_w,\bar{y}_l)\sim\mathscr{D}}\left[\sigma\left(\hat{r}_\theta(\bar{x},\bar{y}_l) - \hat{r}_\theta(\bar{x},\bar{y}_w)\right)\left[\nabla \log \pi_\theta(\bar{y}_w \mid \bar{x}) - \nabla \log \pi_\theta(\bar{y}_l \mid \bar{x})\right]\right]$$

β functions like a “learning rate” following the strength of the KL constraint

Per-example weight: higher weight when the reward model is wrong

Increase likelihood of preferred example

Decrease likelihood of dispreferred example

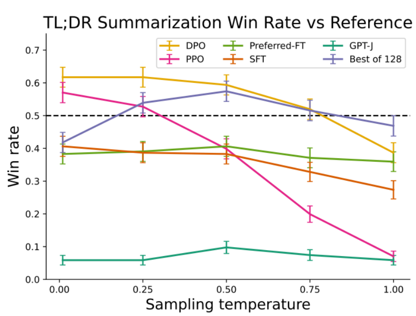
where $\hat{r}(\bar{x},\bar{y}) = \beta \log \frac{\pi_\theta(\bar{y} \mid \bar{x})}{\pi_{\text{ref}}(\bar{y} \mid \bar{x})}$

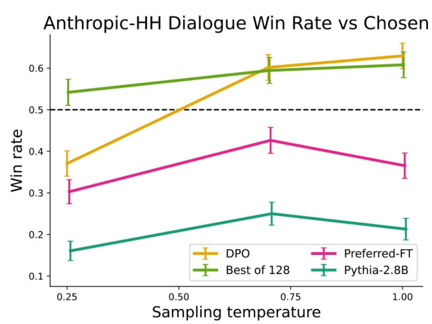
Slide credit: Yoav Artzi

Outcome of RLHF/DPO

- RLHF produces an “aligned” model that should achieve high reward
- Baselines:
  - Best-of-n: sample n responses from an SFT model, take the best one according to the reward function
    - Pro: training-free
    - Cons: expensive, may not deviate far from the initial SFT model
  - Preference tuning: apply SFT on preferred outputs
    - Pro: simple. Cons: doesn’t use the negative examples

Direct Preference Optimization (DPO)





- Evaluation: *win rate* (as scored by an LLM)

Rafailov et al. (2023)

RLHF in practice

Dataset	Num. of Comparisons	Avg. # Turns per Dialogue	Avg. # Tokens per Example	Avg. # Tokens in Prompt	Avg. # Tokens in Response
Anthropic Helpful	122,387	3.0	251.5	17.7	88.4
Anthropic Harmless	43,966	3.0	152.5	15.7	46.4
OpenAI Summarize	176,625	1.0	371.1	336.0	35.1
OpenAI WebGPT	13,333	1.0	237.2	48.3	188.9
StackExchange	1,038,480	1.0	440.2	200.1	240.2
Stanford SHP	74,882	1.0	338.3	199.5	138.8
Synthetic GPT-J	33,139	1.0	123.3	13.0	110.3
Meta (Safety & Helpfulness)	1,418,091	3.9	798.5	31.4	234.1
Total	2,919,326	1.6	595.7	108.2	216.9

RLHF data for Llama 2

- They do 5 iterations of (train, get more preferences, get new reward model). First 3 iterations: just fine-tuning best-of-n, then they used PPO
- Current approaches: many papers exploring versions with active data collection (e.g., tune with DPO -> collect preferences -> keep tuning …)

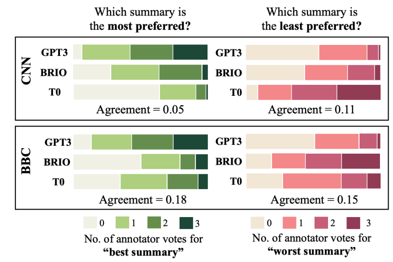
Touvron et al. (2023)

Evaluating LLMs

Death of Benchmarks

Dataset	**Brio** Best ↑	Worst ↓	**T0** Best ↑	Worst ↓	**GPT3** Best ↑	Worst ↓
CNN	36	24	8	67	58	9
BBC	20	56	30	29	57	15

Table 3: Percentage of times a summarization system is selected as the best or worst according to majority vote (may be tied). Human annotators have a clear preference for `GPT3-D2` for both CNN and BBC style summaries.



- Many classic tasks and metrics were saturated when ChatGPT came out
- "Tests" like MMLU are very artificial, and we want to judge long-form responses

Goyal, Li, Durrett (2023)

LLM-as-a-Judge

- Get responses from two models, ask GPT-4 which one is better
- "Win rate": if you compare model A vs. model B, what fraction of the time does it win?
- Sometimes use win rate against a fixed target (e.g., GPT-3.5), like on the next slide

Hamish Ivison et al. (2024)

DPO/PPO Comparison

Data / Model	Alg.	Factuality	Reasoning	Coding	Truthfulness	Safety	Inst. Foll.	Average
Llama 2 base	-	52.0	37.0	30.7	32.7	32.7	-	-
Tülu 2 (SFT)	-	55.4	47.8	45.1	56.6	91.8	44.2	56.8
StackExchange	DPO	**55.3**	**47.8**	42.4	**56.2**	92.0	46.7	56.7
	PPO	55.1	**47.8**	**46.4**	54.2	**92.6**	**47.4**	**57.3**
ChatArena (2023)	DPO	**55.4**	**50.2**	45.9	**58.5**	67.3	**50.8**	54.7
	PPO	55.2	49.2	**46.4**	55.8	**79.4**	49.7	**55.9**
HH-RLHF	DPO	**55.2**	47.6	44.2	**60.0**	**93.4**	46.6	57.8
	PPO	54.9	**48.6**	**45.9**	58.0	92.8	**47.0**	**57.9**
Nectar	DPO	**55.6**	45.8	39.0	**68.1**	**93.3**	**48.4**	58.4
	PPO	55.2	**51.2**	**45.6**	60.1	92.6	47.4	**58.7**
UltraFeedback (FG)	DPO	55.3	50.9	45.9	69.3	**91.9**	52.8	61.0
	PPO	**56.0**	**52.0**	**47.7**	**71.5**	91.8	**54.4**	**62.2**
Avg. Δ b/w PPO & DPO		-0.1	+1.3	+2.9	-2.5	+2.3	+0.1	+0.7

Table 2: **DPO vs PPO:** Average performance of 13B models trained using DPO and PPO across different datasets, along with the performance difference between DPO and PPO (Δ). Blue indicates improvements over the SFT baseline, orange degradations. All datasets are downsampled to 60,908

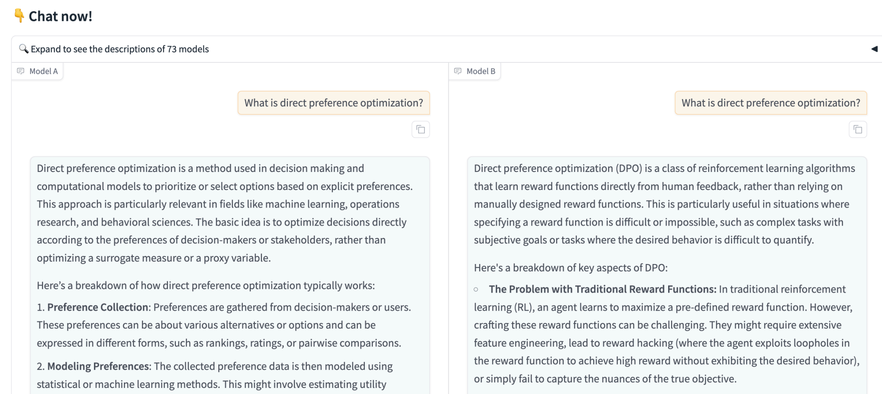
Hamish Ivison et al. (2024)

Data settings

StackExchange	Upvotes/downvotes on StackExchange (synthetic dataset of human preferences)
ChatArena (2023)	Human preferences (discussed on next slide)
HH-RLHF	Human annotated data, but a bit older
Nectar	GPT-4 preferences over a big dataset
UltraFeedback (FG)	GPT-4 preferences over a big dataset

Hamish Ivison et al. (2024)

Chatbot Arena: Elo Rankings



Chatbot Arena: Elo Rankings

- Accepted as one of the premiere rankings for LLMs
- Style control was introduced as it was believed that the "style" of responses had a big effect

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes
1	1	ChatGPT-4o-latest (2024-09-03)	1340	+4/-3	33743
1	1	o1-preview	1335	+4/-4	21071
3	6	o1-mini	1308	+4/-4	23128
3	4	Gemini-1.5-Pro-002	1303	+4/-4	15736
4	4	Gemini-1.5-Pro-Exp-0827	1299	+4/-3	32385
6	9	Grok-2-08-13	1290	+3/-3	40873
6	3	Claude 3.5 Sonnet (20241022)	1286	+6/-6	7284
6	11	Yi-Lightning	1285	+4/-4	20973
6	4	GPT-4o-2024-05-13	1285	+3/-3	102960
10	15	GLM-4-Plus	1275	+4/-4	19922
10	18	GPT-4o-mini-2024-07-18	1273	+4/-3	42661
10	19	Gemini-1.5-Flash-002	1272	+5/-6	12379
10	26	Llama-3.1-Nemotron-70b-Instruct	1271	+5/-7	6228
10	14	Gemini-1.5-Flash-Exp-0827	1269	+4/-4	25503
11	6	Claude 3.5 Sonnet (20240620)	1268	+3/-3	81086

Takeaways

- Instruction-tuning and RLHF are two procedures that take LMs to the next level — these models work dramatically better than basic GPT-3
- These are the foundation of modern chatbots (along with lots of pre-training data), very exciting capabilities in these LLM agents
- Evaluating where these models are is tough, requires human intervention or trust that LLMs are doing reasonable things…