Instruction Tuning (Week 12)

CS294-43

Berkeley
UNIVERSITY OF CALIFORNIA

Evolution of Modeling Paradigms

Task Specific Modeling

Training on small-scale well-annotated data (in one modality)

Evolution of Modeling Paradigms

Task Specific Modeling

Training on small-scale well-annotated data (in one modality)

Early Foundation Models

Pre-training on large-scale (potentially multimodal) noisy data

↓

Fine-tune on small-scale well-annotated data.

Evolution of Modeling Paradigms

Task Specific Modeling

Training on small-scale well-annotated data (in one modality)

Early Foundation Models

Pre-training on large-scale (potentially multimodal) noisy data

↓

Fine-tune on small-scale well-annotated data.

Generalist Modeling

Pre-training on large-scale (potentially multimodal) noisy data

↓

Zero-Shot / Few-Shot learning in multiple modalities

Evolution of Modeling Paradigms

Task Specific Modeling

Training on small-scale well-annotated data (in one modality)

Early Foundation Models

Pre-training on large-scale (potentially multimodal) noisy data

↓

Fine-tune on small-scale well-annotated data.

Generalist Modeling

Pre-training on large-scale (potentially multimodal) noisy data

↓

Zero-Shot / Few-Shot learning in multiple modalities

How can we enable this?

Language Models are (excellent) Next-Token Predictors

One key emergent ability in GPT family is zero-shot learning: the ability to do many tasks with no examples, and no gradient updates, by simply:

- Specifying the right sequence prediction problem

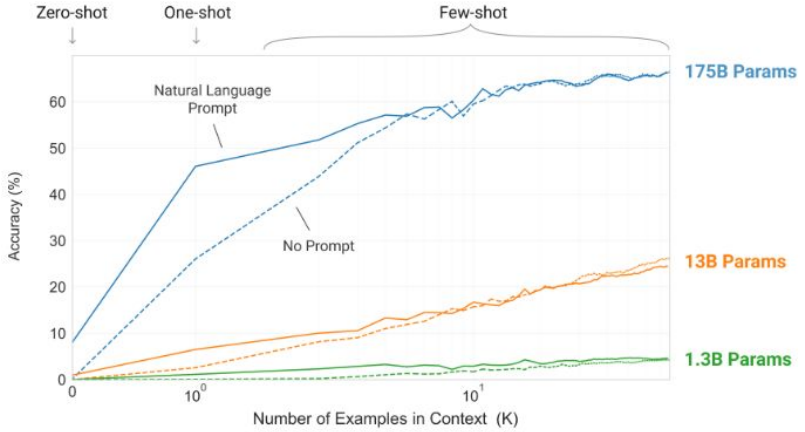
  `Passage: (information) Q: The capital of France is A: [...]`

- Comparing the probability of sequences

  `The cat couldn't fit into the hat because` **`[the cat]`** `was too big`
  `The cat couldn't fit into the hat because` **`[the hat]`** `was too big`

https://web.stanford.edu/class/cs224n/slides/cs224n-spr2024-lecture10-prompting-rlhf.pdf

We can even hack them by example:



Brown, Tom B. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

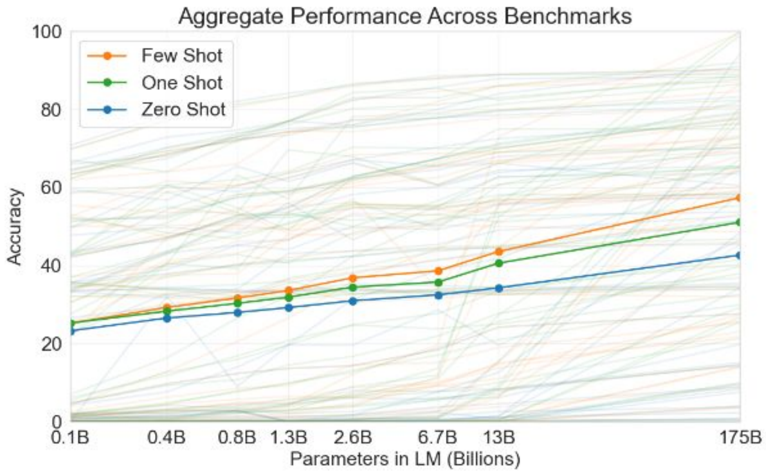
We can even hack them by example:

```
gaot => goat
sakne => snake
brid => bird
fsih => fish
dcuk => duck
cmihp => chimp
```

In-context learning

```
thanks => merci
hello => bonjour
mint => menthe
wall => mur
otter => loutre
bread => pain
```

In-context learning



Brown, Tom B. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

Language modeling is not human-interpretable by default

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.`

https://web.stanford.edu/class/cs224n/slides/cs224n-spr2024-lecture10-prompting-rlhf.pdf

Language modeling is not human-interpretable by default

Prompt:	
Why is it important to eat socks after meditating?	
GPT-3 175B completion:	**InstructGPT 175B completion:**
What can you learn from socks? A: When you eat socks, you are tasting the essence of enlightenment. You are tasting the true flavor of the awakened mind. Q: What is the true flavor of the awakened mind? A: The true flavor of the awakened mind is the flavor of no flavor. And the true flavor of no flavor is the flavor of smelly feet.	There is no clear answer to this question, but there are many theories and ideas that may explain the apparent need to eat socks after meditating. Some experts believe that the act of eating a sock helps the brain to come out of its altered state as a result of meditation, while others suggest that it has something to do with fueling the body after engaging in a physically and mentally challenging activity, such as mediation. Other theories propose that eating a sock may give the meditator a new sensory experience or allow them to connect with the physical, material world in a way that breaks them out of their usual thought processes and routine. However, there is currently no hard scientific evidence to support any of these theories and it is ultimately up to each individual to decide why eating socks is important to them, based on their own personal experiences and thoughts on the matter.

Language modeling is not human-interpretable by default

Prompt:
What is the purpose of the list C in the code below?

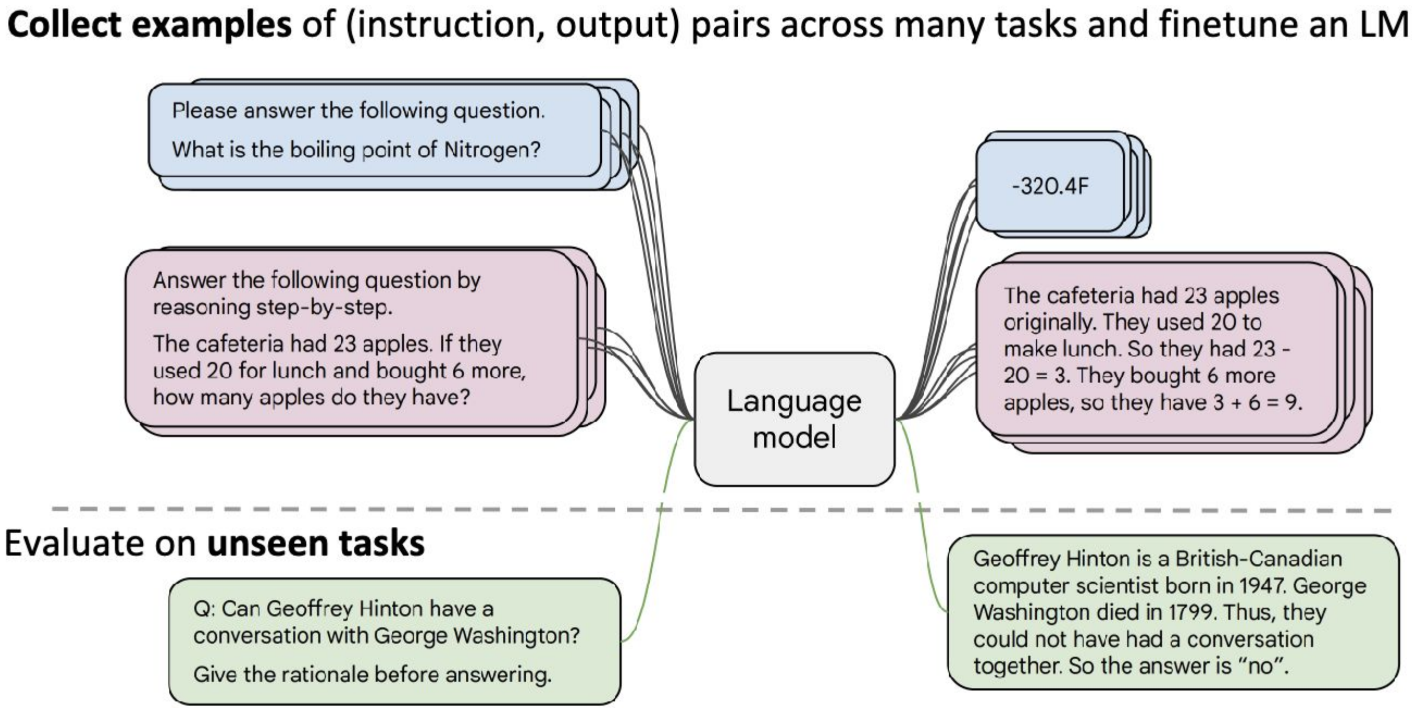
```
def binomial_coefficient(n, r):
     C = [0 for i in range(r + 1)];
     C[0] = 1;
     for i in range(1, n + 1):
          j = min(i, r);
          while j > 0:
               C[j] += C[j - 1];
               j -= 1;
     return C[r]
```

GPT-3 175B completion:
A. to store the value of C[0]
B. to store the value of C[1]
C. to store the value of C[i]
D. to store the value of C[i - 1]

InstructGPT 175B completion:
The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function.

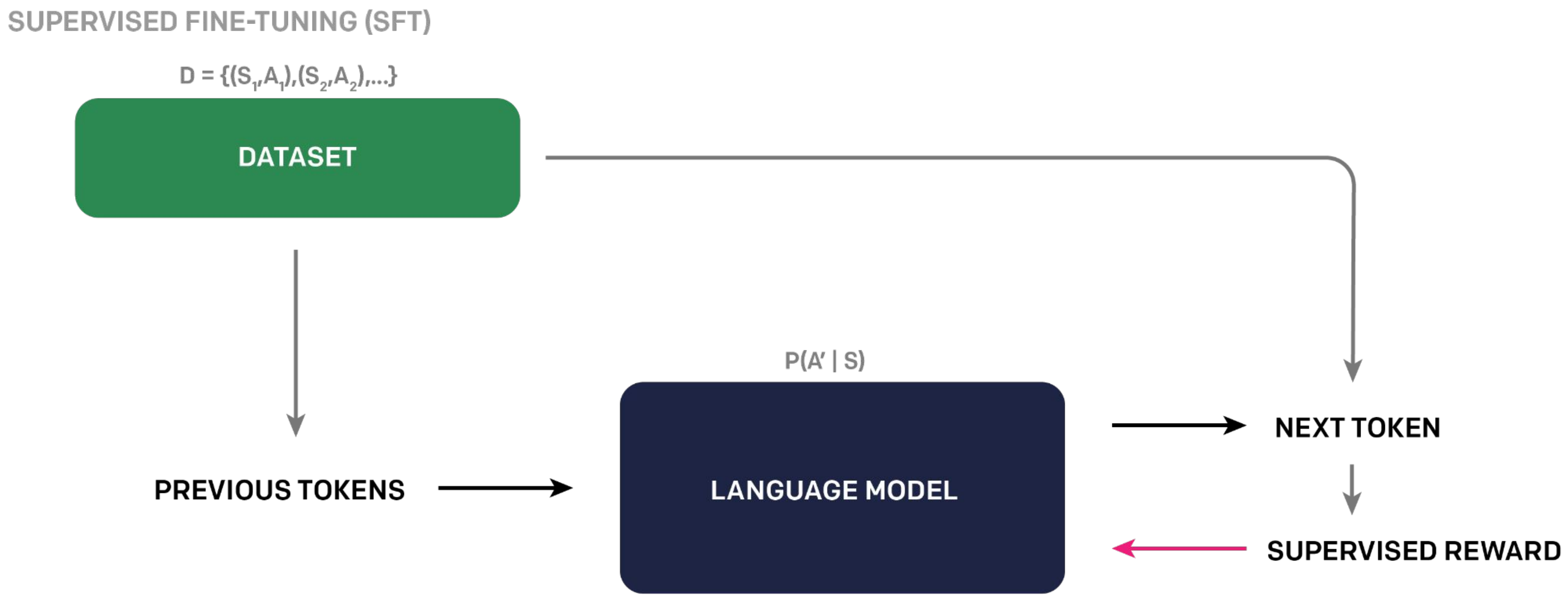
So what can we do?
We fine-tune!

From SFT to IRL to RLHF to DPO



Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." Journal of Machine Learning Research 25.70 (2024): 1-53.

From SFT to IRL to RLHF to DPO



From SFT to IRL to RLHF to DPO

Limitations of instruction fine-tuning:

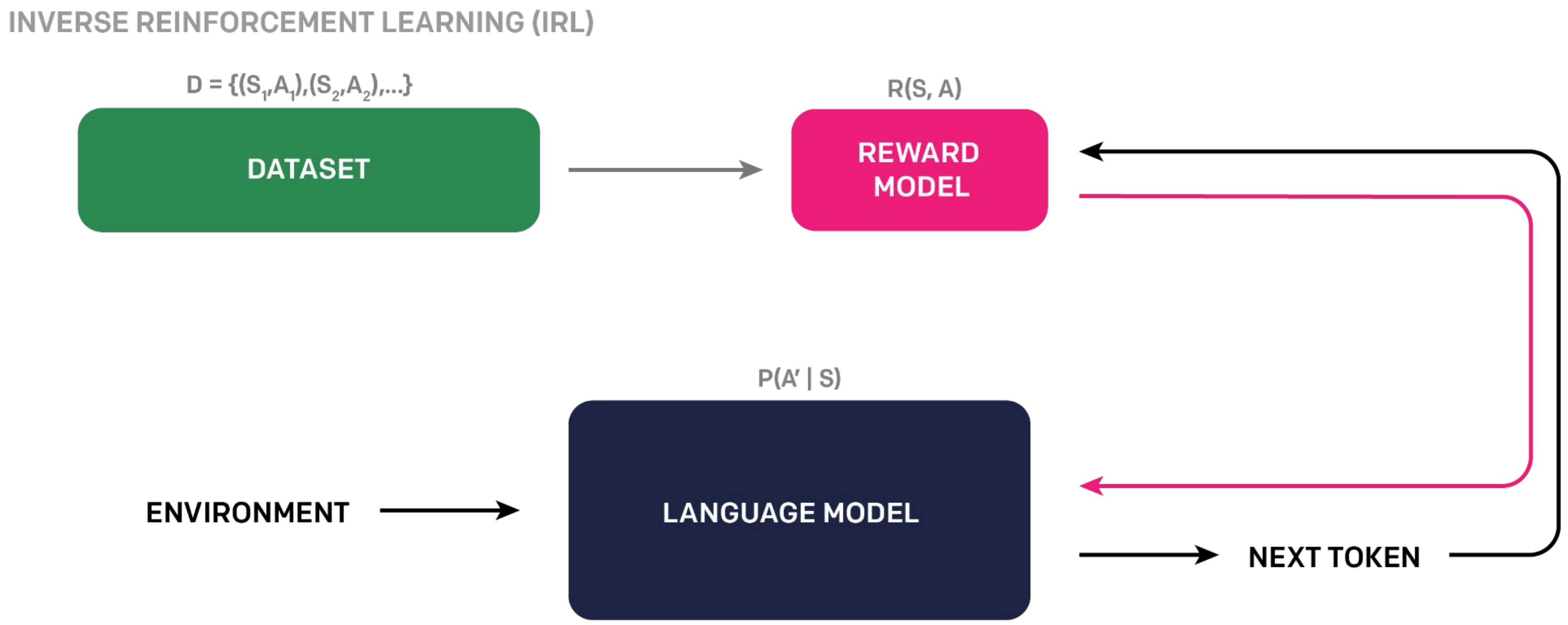
From SFT to IRL to RLHF to DPO

Limitations of instruction fine-tuning:

- It's **expensive** to collect ground truth
- Tasks such as open-ended creative generation have no correct answer
- Language modeling penalizes all token-level mistakes equally (but some are worse than others)
- Humans often generate sub-optimal answers

Can we explicitly model human preferences?

From SFT to IRL to RLHF to DPO



From SFT to IRL to RLHF to DPO

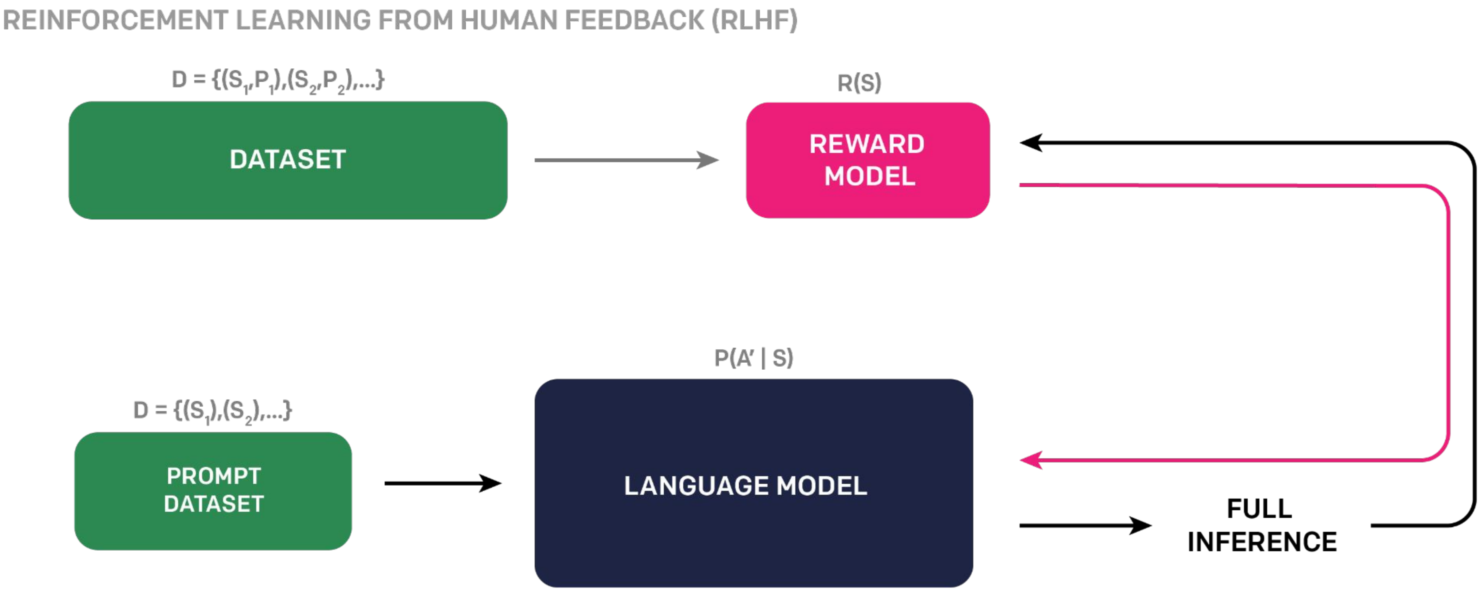
What's wrong with learning human policies directly?

From SFT to IRL to RLHF to DPO

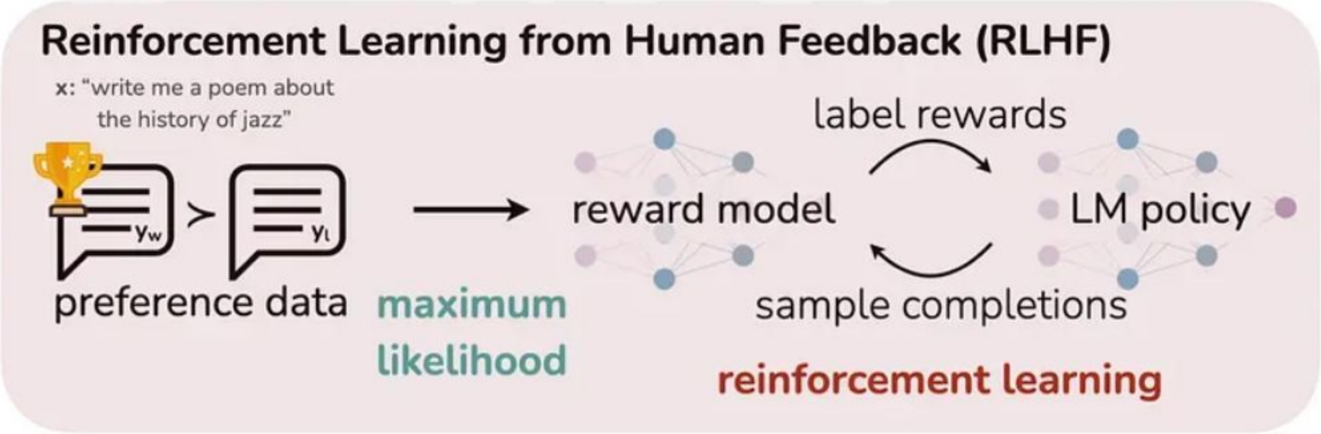
What’s wrong with learning human policies directly?

- **Expensive** to collect expert demonstrations
- Experts are assumed to be optimal (They’re still usually not)
- Experts must EXIST (and it’s hard to transfer to new tasks where they don’t)

From SFT to IRL to RLHF to DPO



From SFT to IRL to RLHF to DPO



From SFT to IRL to RLHF to DPO

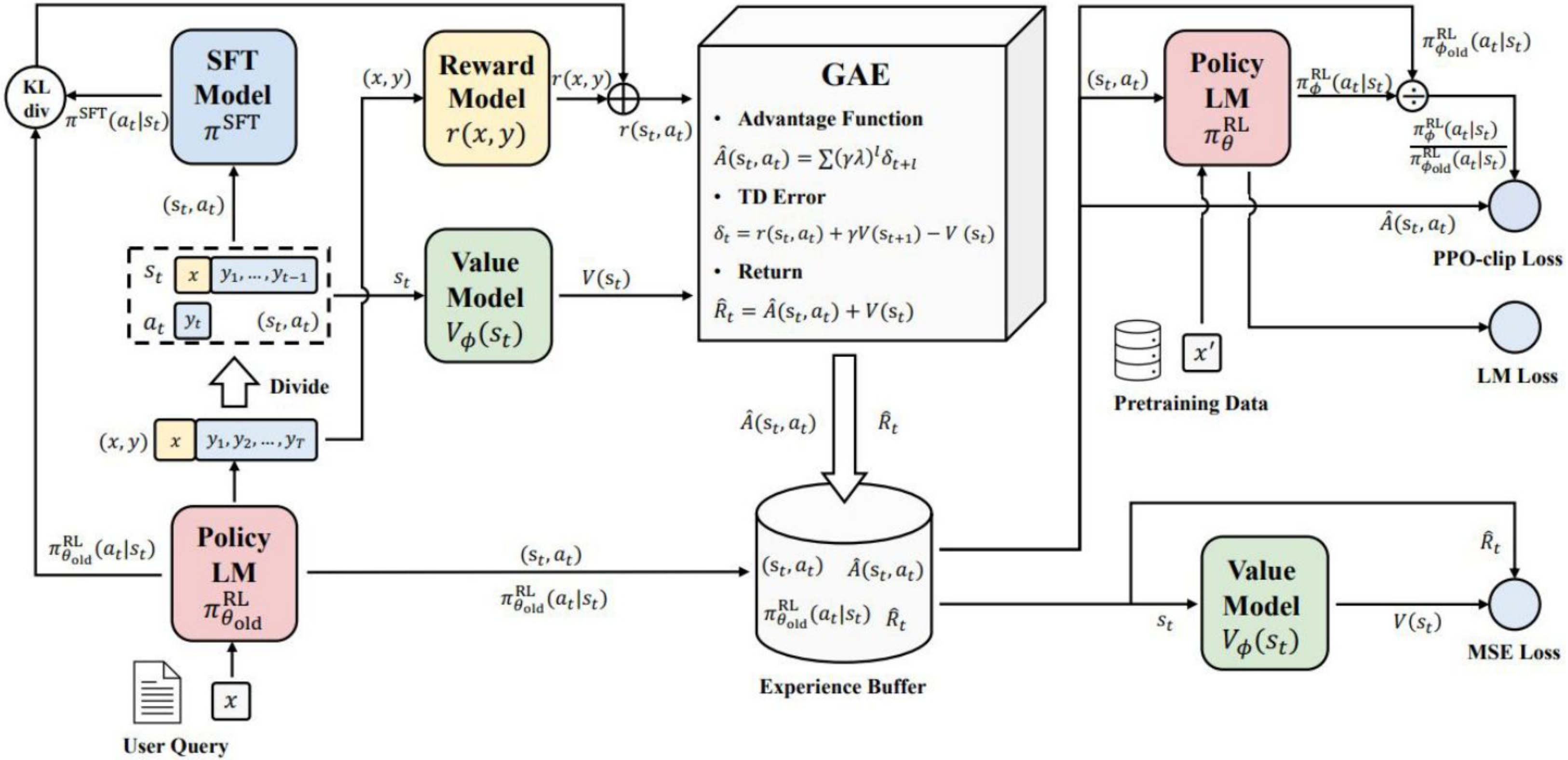
$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\right]$$

Sample from policy

Want high reward…

…but keep KL to original model small!

From SFT to IRL to RLHF to DPO



From SFT to IRL to RLHF to DPO

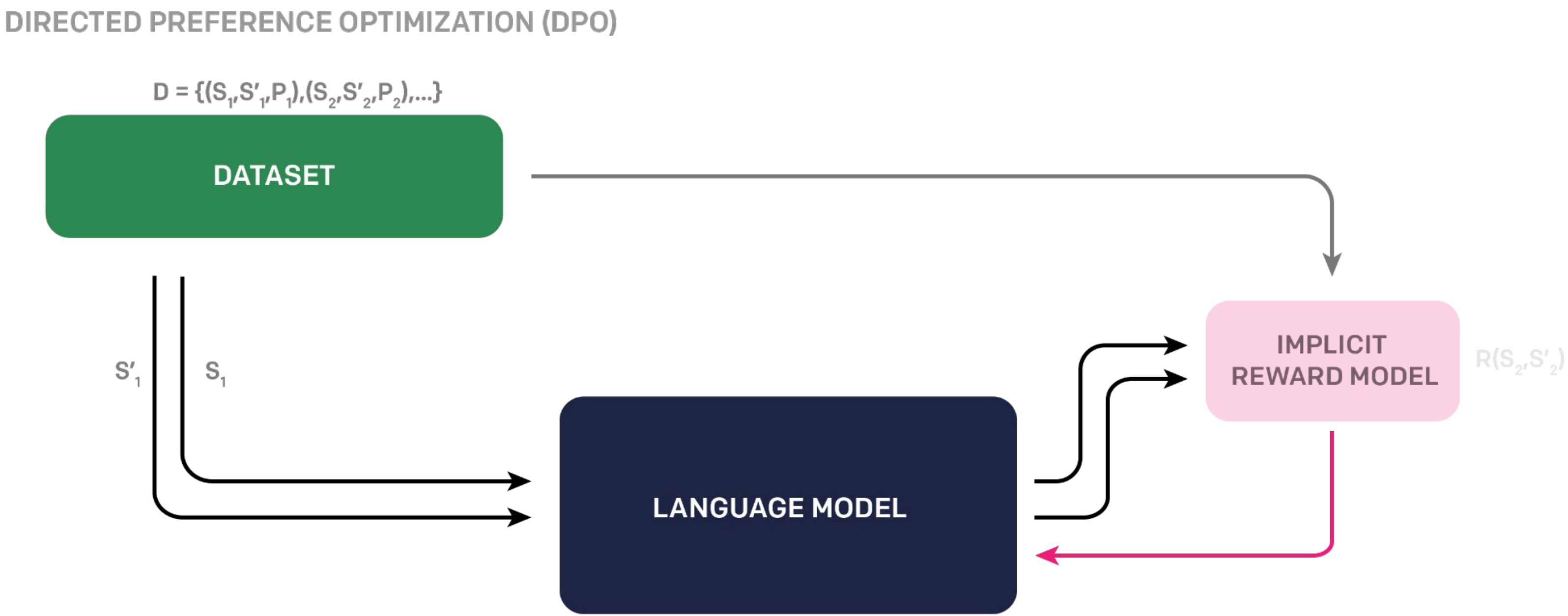
Limitations of RLHF:

From SFT to IRL to RLHF to DPO

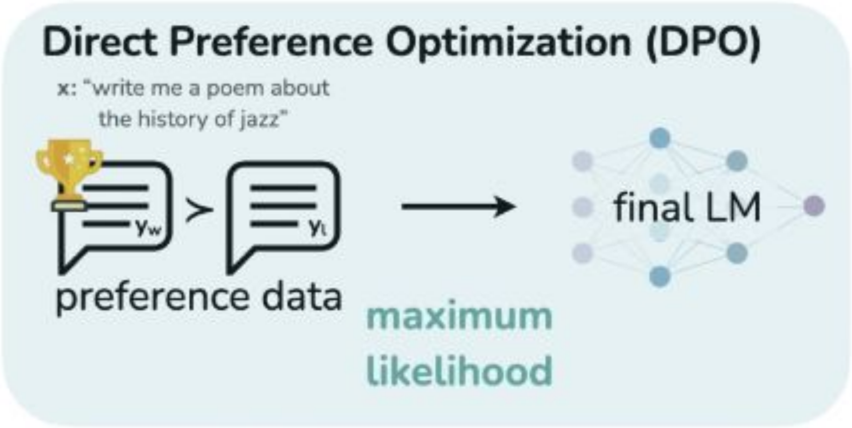
Limitations of RLHF:

- Complexity: Designing and training reward models can be challenging
- Computational Overhead: **It’s expensive**
- Control: Users don’t have direct control over the LLM’s behavior

From SFT to IRL to RLHF to DPO



From SFT to IRL to RLHF to DPO



From SFT to IRL to RLHF to DPO

RLHF Objective

(get **high reward**, stay **close** to reference model)

any reward function

$$\max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(y|x)} [r(x, y)] - \beta \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x))$$

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

any reward function

RLHF Objective

(get **high reward**, stay **close** to reference model)

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(y|x)} \left[r(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x))$$

Closed-form Optimal Policy

(write **optimal policy** as function of **reward function**; from prior work)

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

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

RLHF Objective

(get **high reward**, stay **close** to reference model)

any reward function

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(y|x)} [r(x,y)] - \beta \mathbb{D}_{\text{KL}}(\pi(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x))$$

Closed-form Optimal Policy

(write **optimal policy** as function of **reward function**; from prior work)

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

with $Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x,y)\right)$

Note **intractable sum** over possible responses; can't immediately use this

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

RLHF Objective

(get **high reward**, stay **close** to reference model)

any reward function

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(y|x)} \left[r(x, y)\right] - \beta \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x))$$

Closed-form Optimal Policy

(write **optimal policy** as function of **reward function**; from prior work)

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

with $Z(x) = \sum_y \pi_{\mathrm{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)$

Note **intractable sum** over possible responses; can't immediately use this

Rearrange

(write **any reward function** as function of **optimal policy**)

Ratio is **positive** if policy likes response more than reference model, **negative** if policy likes response less than ref. model

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

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

RLHF Objective

(get **high reward**, stay **close** to reference model)

any reward function

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(y|x)} \left[r(x,y) \right] - \beta \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x))$$

Closed-form Optimal Policy

(write **optimal policy** as function of **reward function**; from prior work)

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

with $Z(x) = \sum_y \pi_{\mathrm{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x,y)\right)$

Note **intractable sum** over possible responses; can't immediately use this

Rearrange

(write **any reward function** as function of **optimal policy**)

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

Ratio is **positive** if policy likes response more than reference model, **negative** if policy likes response less than ref. model

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

A loss function on
reward functions

+

A transformation
between reward
functions and policies

=

A loss function
on policies

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

A loss function on reward functions

+

A transformation between reward functions and policies

=

A loss function on policies

Derived from the Bradley-Terry model of human preferences:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(r(x, y_w) - r(x, y_l)) \right]$$

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

A loss function on <u>reward functions</u>

+

A transformation between <u>reward functions</u> and <u>policies</u>

=

A loss function on <u>policies</u>

Derived from the Bradley-Terry model of human preferences:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(r(x, y_w) - r(x, y_l)) \right]$$

$$r_{\pi_\theta}(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x)$$

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

A loss function on reward functions

Derived from the Bradley-Terry model of human preferences:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(r(x, y_w) - r(x, y_l)) \right]$$

+

A transformation between reward functions and policies

$$r_{\pi_\theta}(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x)$$

When substituting, the **log Z term cancels**, because the loss only cares about **difference** in rewards

=

A loss function on policies

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

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

From SFT to IRL to RLHF to DPO

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

https://web.stanford.edu/class/cs234/slides/dpo_slides.pdf

Paper 1: From DPO to KTO

KTO: Model Alignment as Prospect Theoretic Optimization

Kawin Ethayarajh [1] Winnie Xu [2] Niklas Muennighoff [2] Dan Jurafsky [1] Douwe Kiela [1 2]

Paper 2: RLHF-V

RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback

Tianyu Yu[1] Yuan Yao[2*] Haoye Zhang[1] Taiwen He[1] Yifeng Han[1]
Ganqu Cui[1] Jinyi Hu[1] Zhiyuan Liu[1*] Hai-Tao Zheng[1*] Maosong Sun[1] Tat-Seng Chua[2]

[1]Tsinghua University [2]National University of Singapore

yiranytianyu@gmail.com yaoyuanthu@gmail.com

https://rlhf-v.github.io

Paper 3: Self-Supervised Visual Preference Alignment

Self-Supervised Visual Preference Alignment

Ke Zhu[1,2] Liang Zhao[4] Zheng Ge[3,4] Xiangyu Zhang[3,4]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]School of Artificial Intelligence, Nanjing University, China
[3]MEGVII Technology
[4]StepFun Intelligent Technology
zhuk@lamda.nju.edu.cn, {zhaoliang06, gezheng, zhangxiangyu}@megvii.com

Conclusion / Discussion (Time permitting)

Overall:

- When should we use alignment-tuning processes (such as KTO, DPO, etc.) vs raw base models? Are there advantages to using non-aligned models?
- Are there **any** differences between preference optimization with multimodal models, and preference optimization with unimodal models? Should we consider modality-specific paradigms for vision-language learning?

Conclusion / Discussion (Time permitting)

Human Preferences and Instruction Design:

- What challenges arise in modeling human preferences for vision-related tasks? How do these compare to challenges in language instruction tuning?
- How can vision instruction tuning incorporate subjective preferences, such as aesthetic judgments or creative interpretations?

Conclusion / Discussion (Time permitting)

Limitations and Challenges:

- Are there any specific bottlenecks in instruction-tuning for vision tasks, especially compared to language?
- Do biases in training datasets manifest differently in vision tasks, do we need to take different approaches to vision instruction tuning?
- What role do statistical priors play in enabling zero-shot or few-shot learning in vision tasks? How can these priors be mathematically represented and optimized during tuning?
- Does alignment tuning (such as in LLaVA) impact human preference tuning?

Conclusion / Discussion (Time permitting)

Evaluation and Metrics:

- How should success in vision instruction tuning be measured? What metrics can effectively capture performance beyond accuracy?