



# Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 22: Supervised Fine-Tuning of LLMs

April 22, 2025

Reading: See course homepage



# Today

---

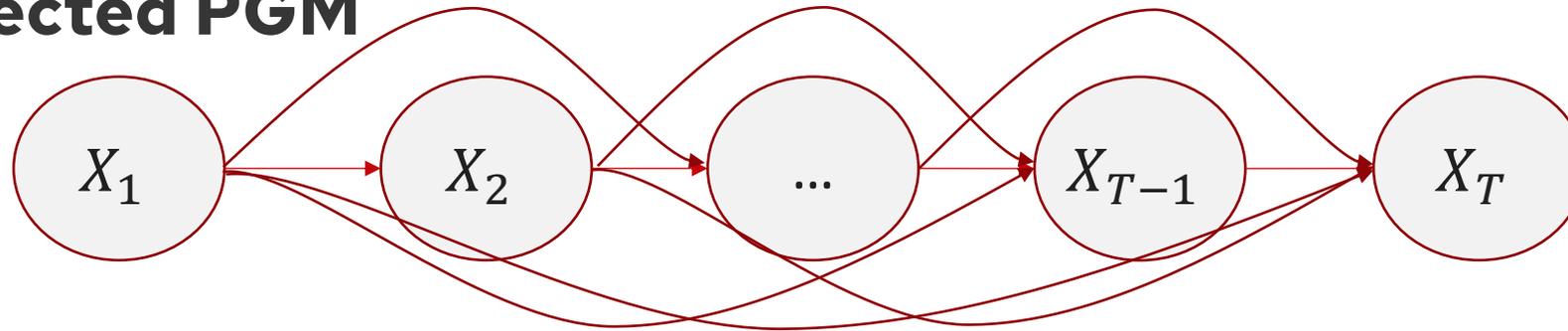
- Supervised Fine-tuning of LLMs
  - Alignment / Reinforcement Learning
- Efficient Parameter Fine-tuning / Personalization
- Prompt Optimization



# Supervised Fine-Tuning of LLMs

# Recall GPT training objective: MLE

- **Directed PGM**



$$P_{\theta}(X) = \prod_i \prod_t P_{\theta}(X_{i,t} | X_{i,<t})$$

- **Probabilistic objective:** Max log-likelihood of observed seqs

$$\max_{\theta} \sum_i \sum_t \log P_{\theta}(X_{i,t} | X_{i,<t})$$

[Radford et al., [Improving Language Understanding by Generative Pre-Training](#)]



# What does MLE not do?

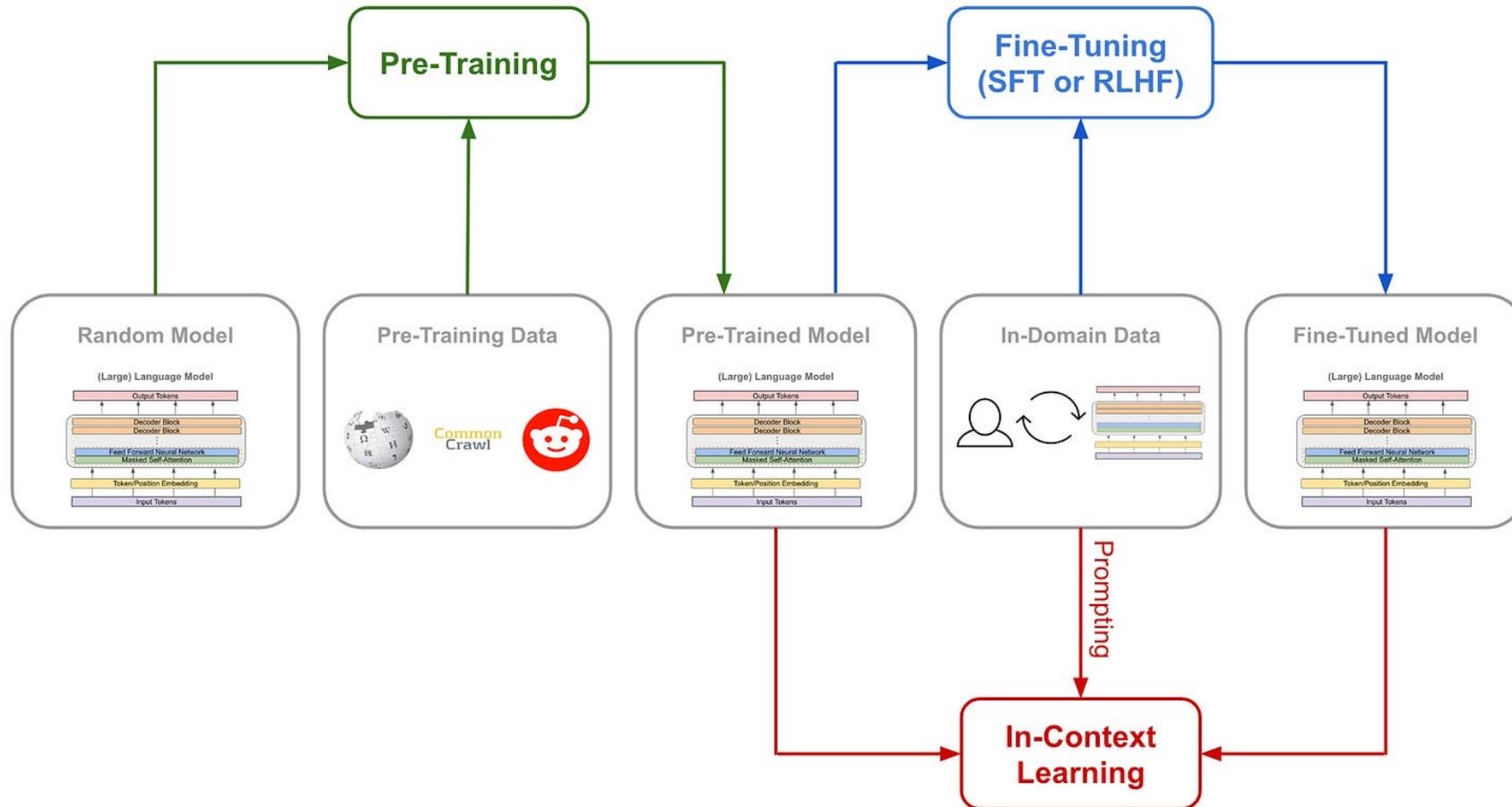
---

- 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?

# From Unsupervised to Supervised

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





# 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

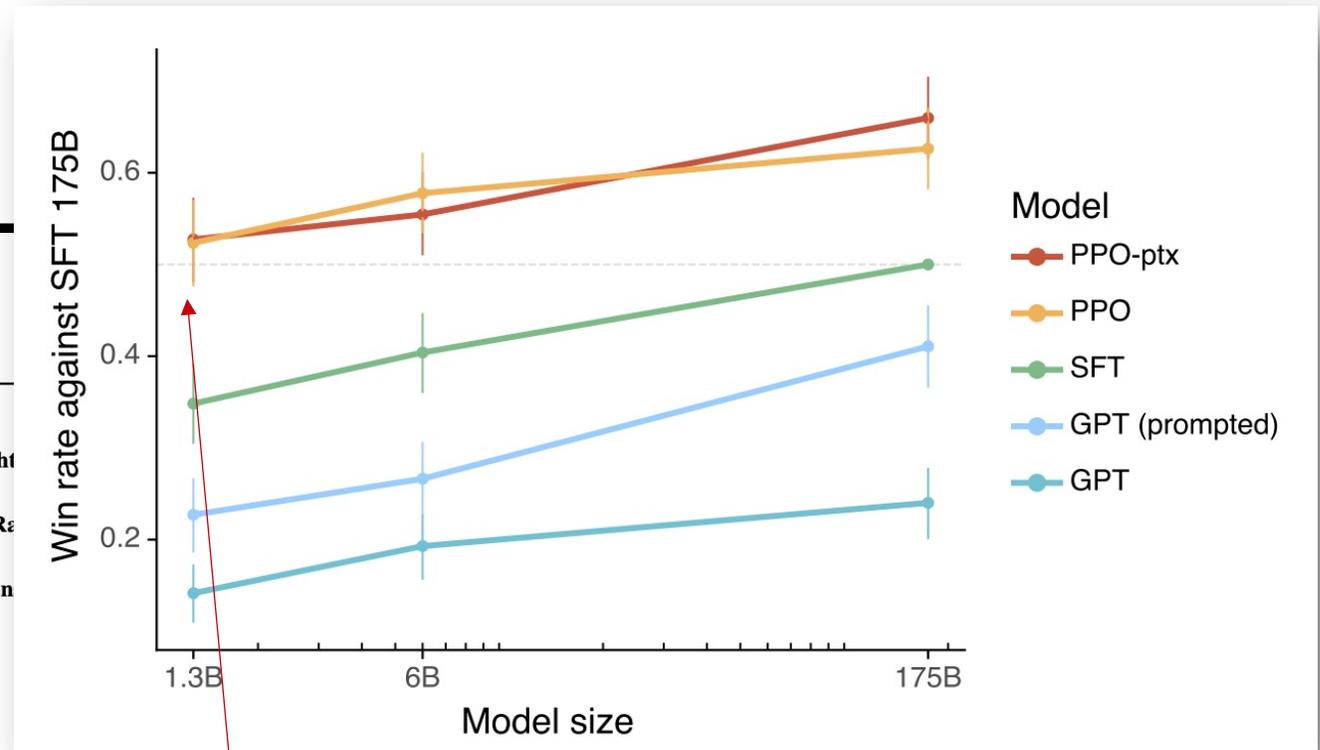
# Supervised Fine-Tuning (SFT)

- 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

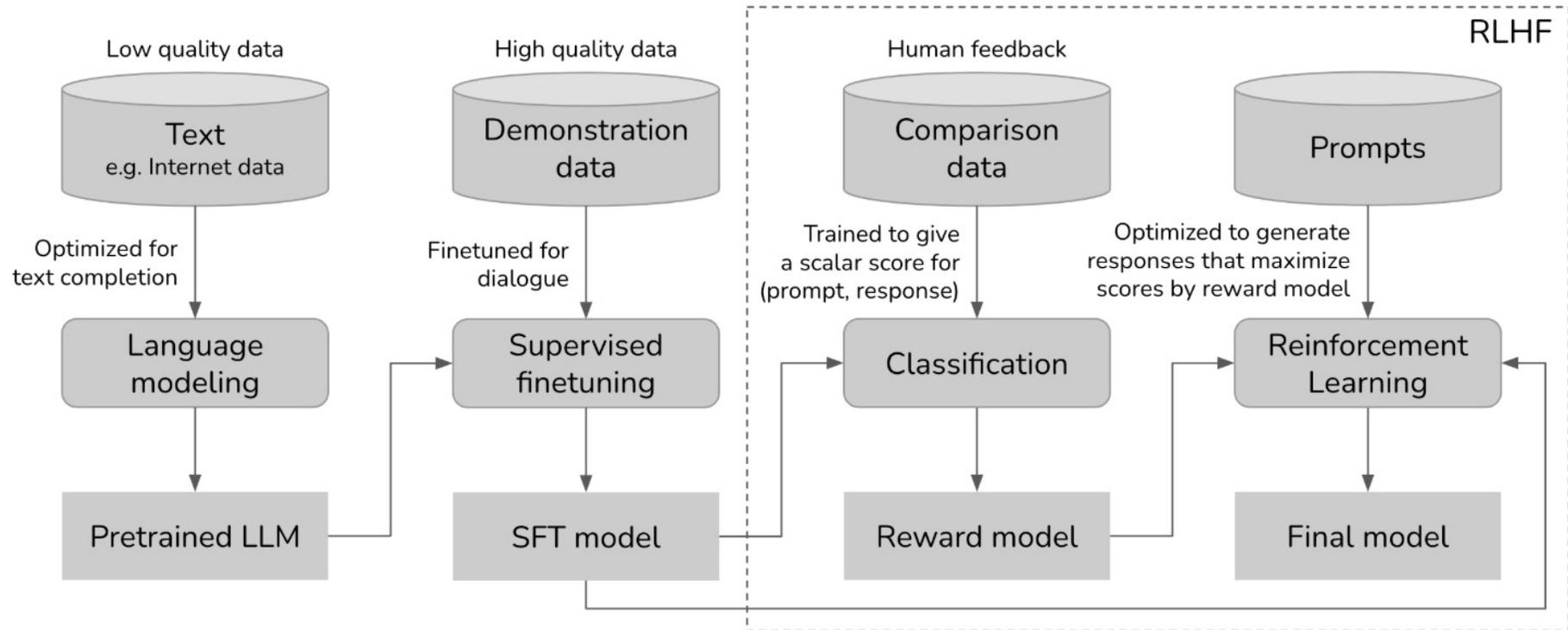
Long Ouyang\* Jeff Wu\* Xu Jiang\* Diogo Almeida\* Carroll L. Wainwright  
 Pamela Mishkin\* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ra  
 John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simen  
 Amanda Askell† Peter Welinder Paul Christiano\*†  
 Jan Leike\* Ryan Lowe\*

OpenAI



**1.3B model can outperform 175B model**

# 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**



# Reinforcement Learning with Human Feedback

- $r_\theta$ : the reward model being trained, parameterized by  $\theta$ . The goal of the training process is to find  $\theta$  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  $\theta$  to minimize the expected loss for all training samples.  $-E_x \log(\sigma(s_w - s_l))$



# Collecting high-quality data is critical

Submit Skip Page 3 / 11 Total time: 05:39

**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

**1** 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...

**2** 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.

**F** 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)

OpenAI UI

<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

Dataset

### RealToxicity

GPT	0.233
Supervised Fine-Tuning	0.199
InstructGPT	<b>0.196</b>

Dataset

### TruthfulQA

GPT	0.224
Supervised Fine-Tuning	0.206
InstructGPT	<b>0.413</b>

API Dataset

### Hallucinations

GPT	0.414
Supervised Fine-Tuning	<b>0.078</b>
InstructGPT	0.172

API Dataset

### Customer Assistant Appropriate

GPT	0.811
Supervised Fine-Tuning	0.880
InstructGPT	<b>0.902</b>

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.



# Efficient Parameter Fine-Tuning



# Personalization / Adaptation / Alignment

---

- Every user has their own preferences, history, and contexts.
- **How can we efficiently adapt to each user?**

# 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)

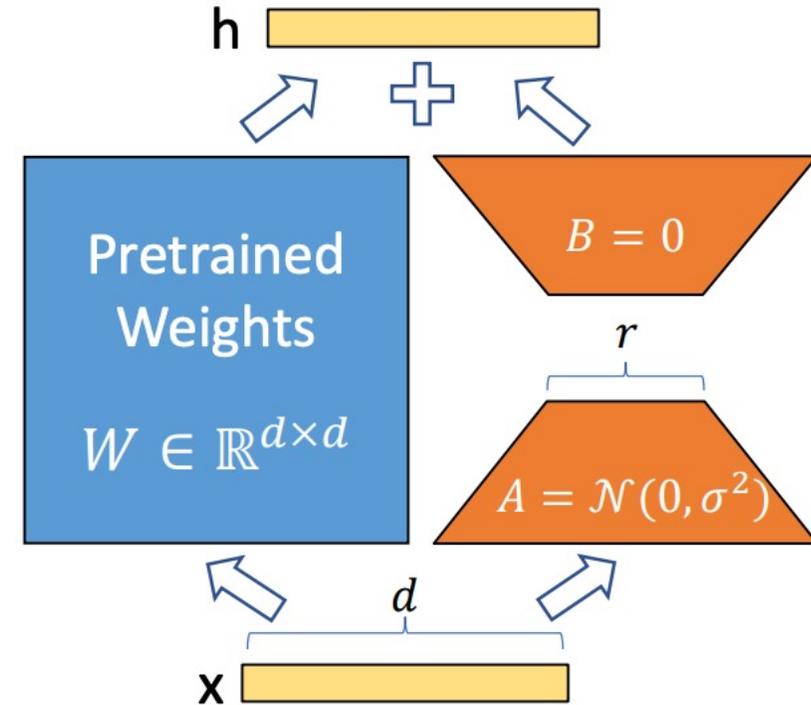
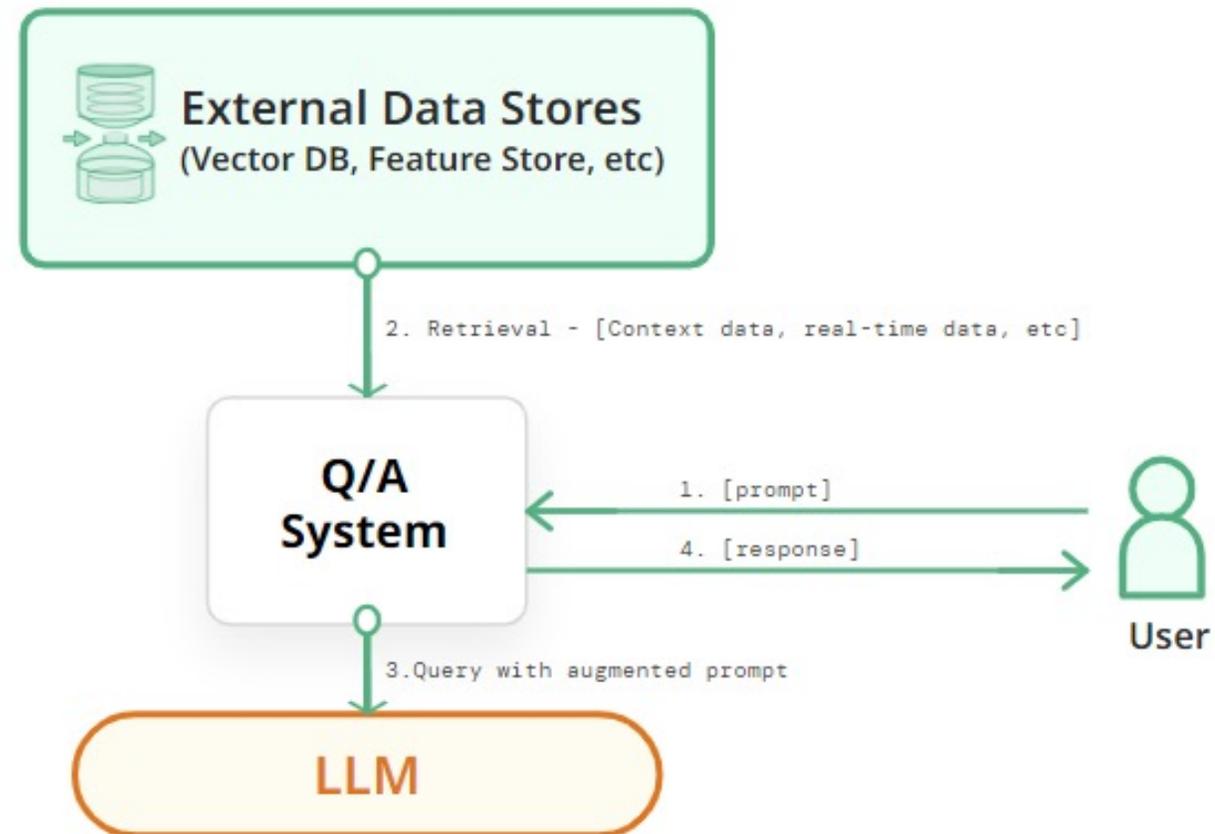


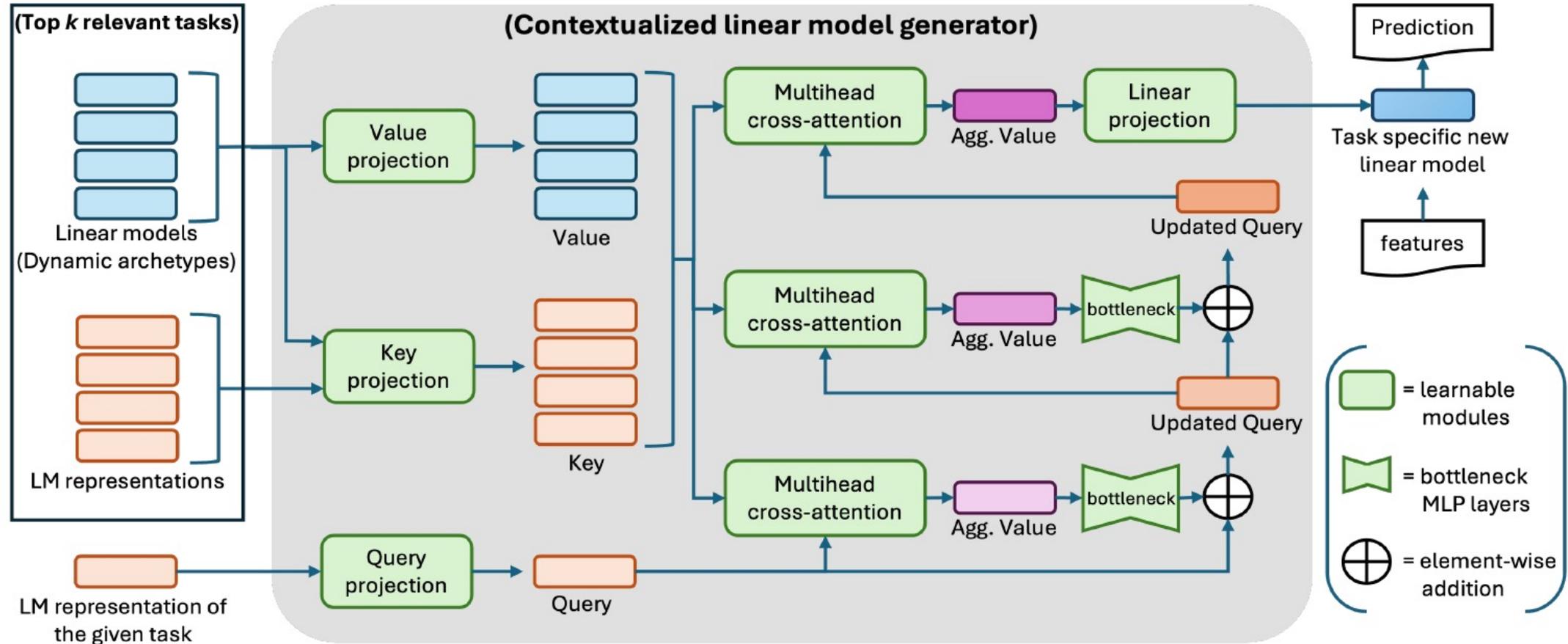
Figure 1: Our reparametrization. We only train  $A$  and  $B$ .

# Retrieval-Augment Generation

- Resource access enables personalization

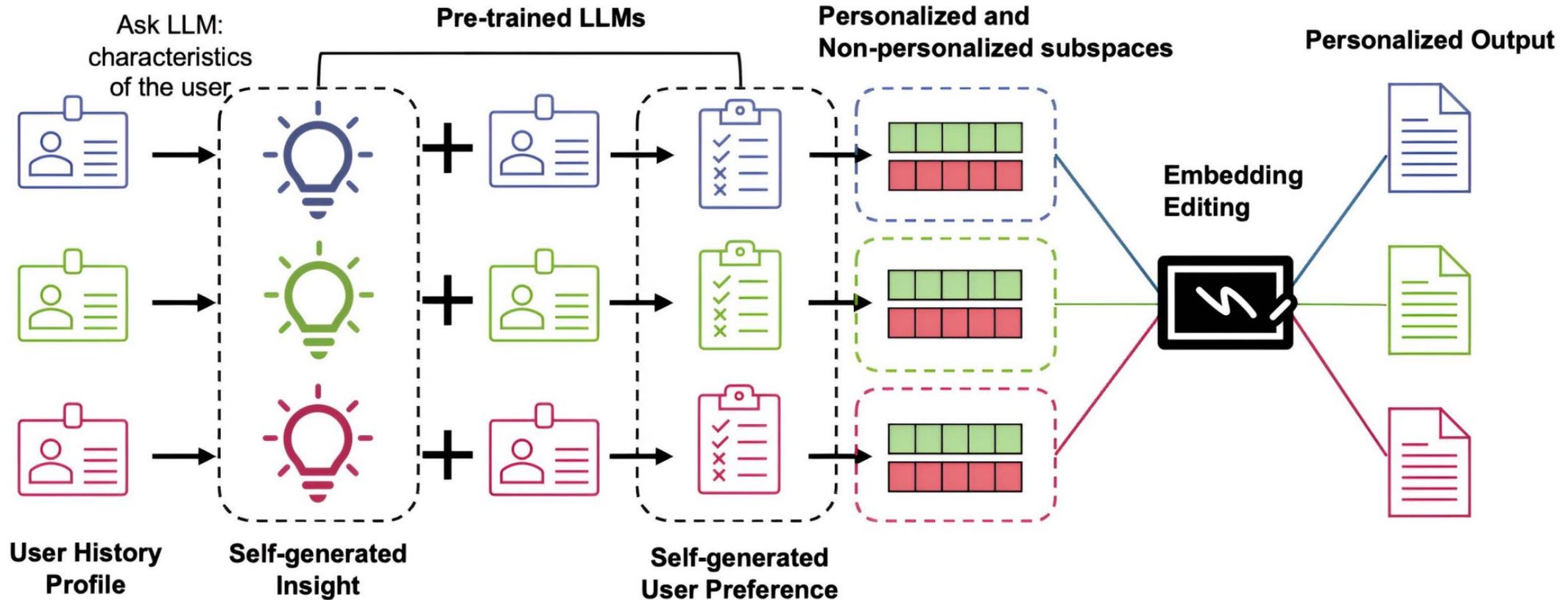


# RAG of Interpretable Models (RAG-IM)



From One to Zero: RAG-IM Adapts Language Models for Interpretable Zero-Shot Clinical Predictions [Mahbub et al 2024]

# More Efficient Personalization



<https://arxiv.org/pdf/2503.01048>

# Prompting





# Few-Shot / Zero-shot learning

One key emergent ability in GPT-2 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 (e.g. question answering):

Passage: Tom Brady... Q: Where was Tom Brady born? A: ...

- Comparing probabilities of sequences (e.g. Winograd Schema Challenge [[Levesque, 2011](#)]):

The cat couldn't fit into the hat because it was too big.  
**Does it = the cat or the hat?**

$\equiv$  Is  $P(\dots\text{because } \mathbf{the\ cat} \text{ was too big}) \geq$   
 $P(\dots\text{because } \mathbf{the\ hat} \text{ was too big})?$

[[Radford et al., 2019](#)]



# Few-Shot / Zero-shot learning

GPT-2 beats SoTA on language modeling benchmarks with **no task-specific fine-tuning**

*Context:* “Why?” “I would have thought you’d find him rather dry,” she said. “I don’t know about that,” said Gabriel.  
“He was a great craftsman,” said Heather. “That he was,” said Flannery.

*Target sentence:* “And Polish, to boot,” said ----- **LAMBADA** (language modeling w/ long discourse dependencies)  
*Target word:* Gabriel [\[Paperno et al., 2016\]](#)

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14
117M	<b>35.13</b>	45.99	<b>87.65</b>	<b>83.4</b>	<b>29.41</b>
345M	<b>15.60</b>	55.48	<b>92.35</b>	<b>87.1</b>	<b>22.76</b>
762M	<b>10.87</b>	<b>60.12</b>	<b>93.45</b>	<b>88.0</b>	<b>19.93</b>
1542M	<b>8.63</b>	<b>63.24</b>	<b>93.30</b>	<b>89.05</b>	<b>18.34</b>

[\[Radford et al., 2019\]](#)



# Few-Shot / Zero-shot learning

You can get interesting zero-shot behavior if you're creative enough with how you specify your task!

Summarization on CNN/DailyMail dataset [[See et al., 2017](#)]:

		ROUGE			
		R-1	R-2	R-L	
SAN FRANCISCO, California (CNN) -- A magnitude 4.2 earthquake shook the San Francisco ... overturn unstable objects.	<b>2018 SoTA</b>	<b>Bottom-Up Sum</b>	<b>41.22</b>	<b>18.68</b>	<b>38.34</b>
		Lede-3	40.38	17.66	36.62
	<b>Supervised (287K)</b>	Seq2Seq + Attn	31.33	11.81	28.83
		GPT-2 TL;DR:	29.34	8.27	26.58
	<b>TL;DR: Select from article</b>	Random-3	28.78	8.63	25.52

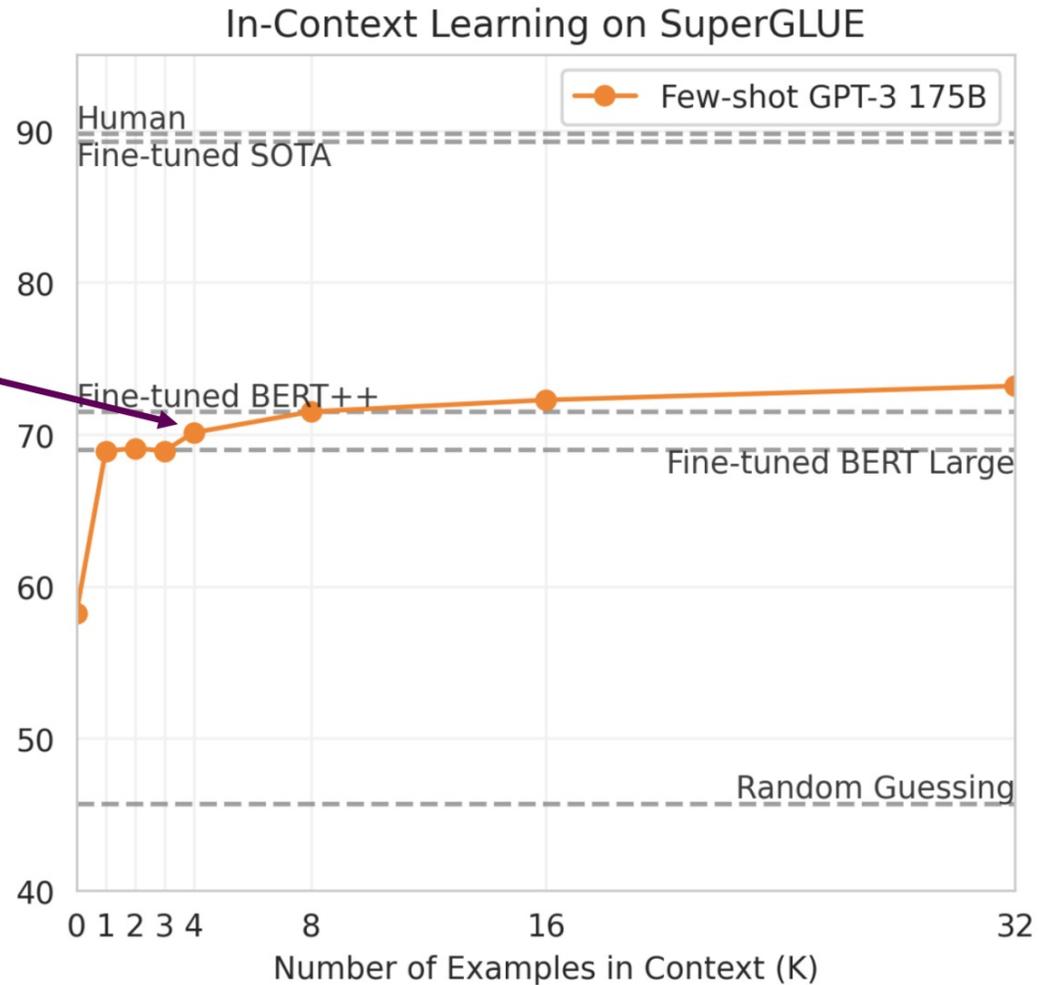
“Too Long, Didn’t Read”  
“Prompting”?

[[Radford et al., 2019](#)]

# "In-Context Learning"

## Few-shot

1 Translate English to French:  
2 sea otter => loutre de mer  
3 peppermint => menthe poivrée  
4 plush girafe => girafe peluche  
5 cheese => .....



[Brown et al., 2020]

# Chain-of-Thought

## Standard Prompting

### Model 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: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. ❌

## Chain-of-Thought Prompting

### Model 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: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

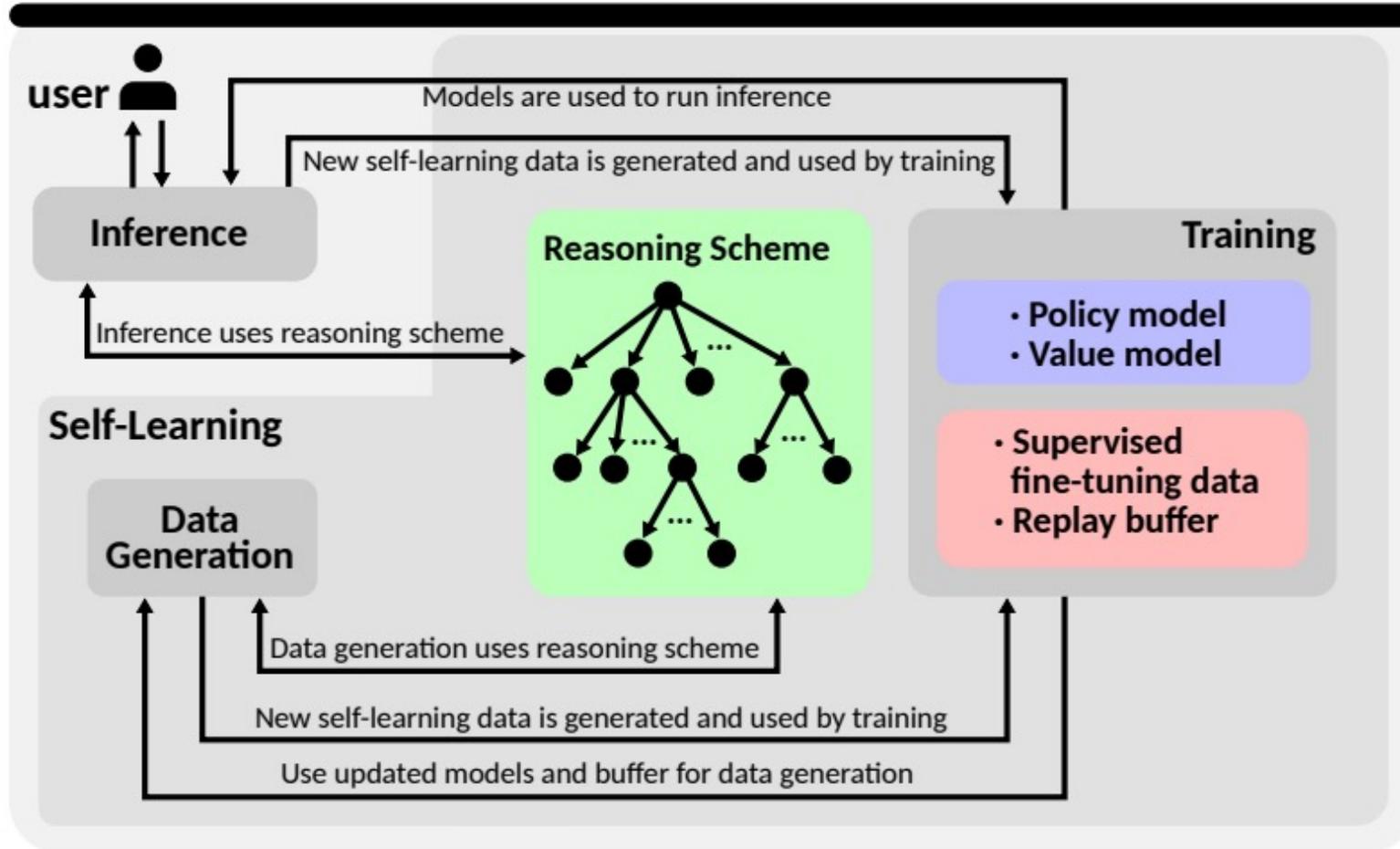
### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅

Wei, et al. (2023) Chain-of-Thought Prompting Elicits Reasoning in LLMs

# Reasoning Models

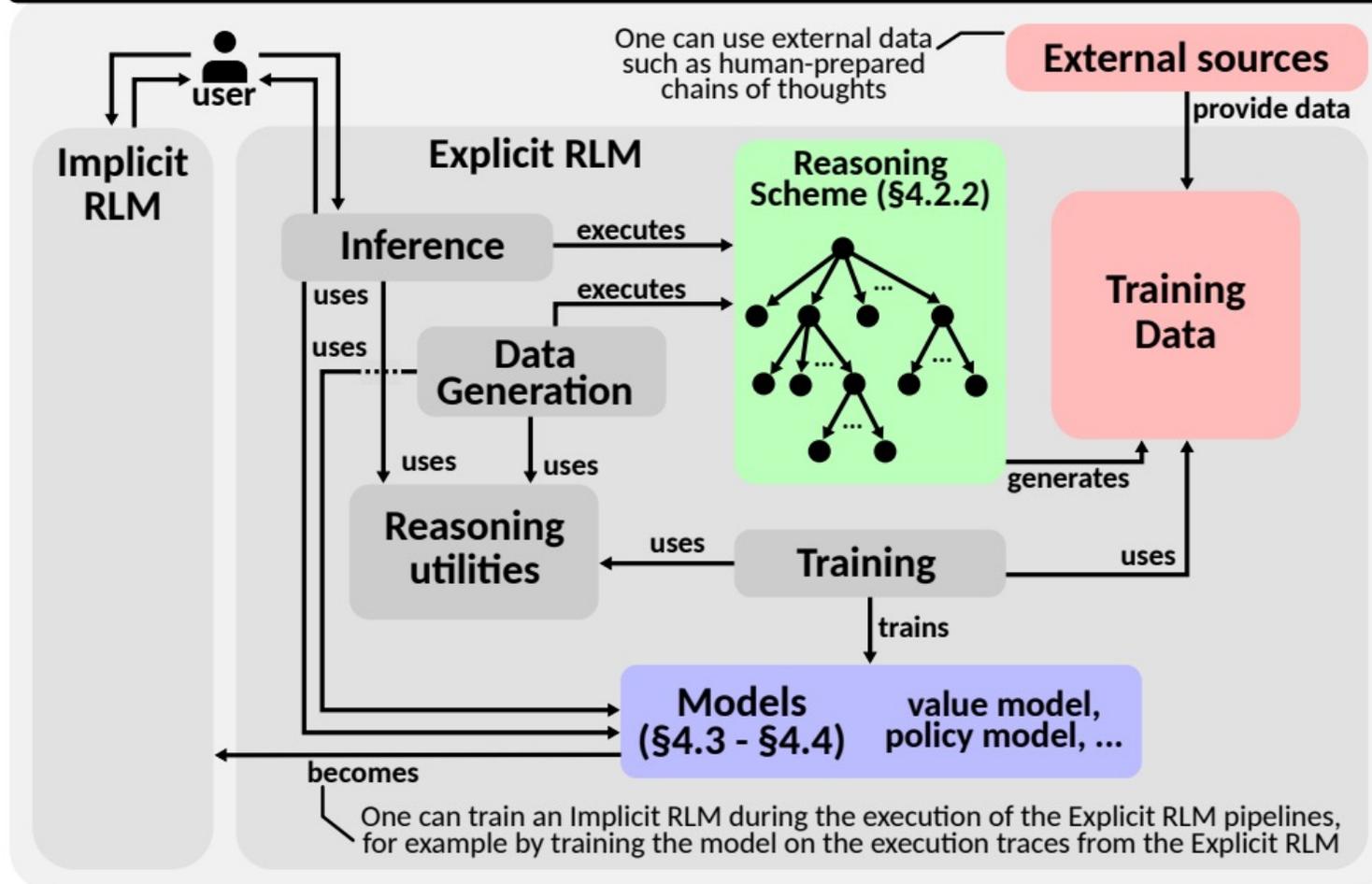
## High-level overview (§3.1)



# Reasoning Models

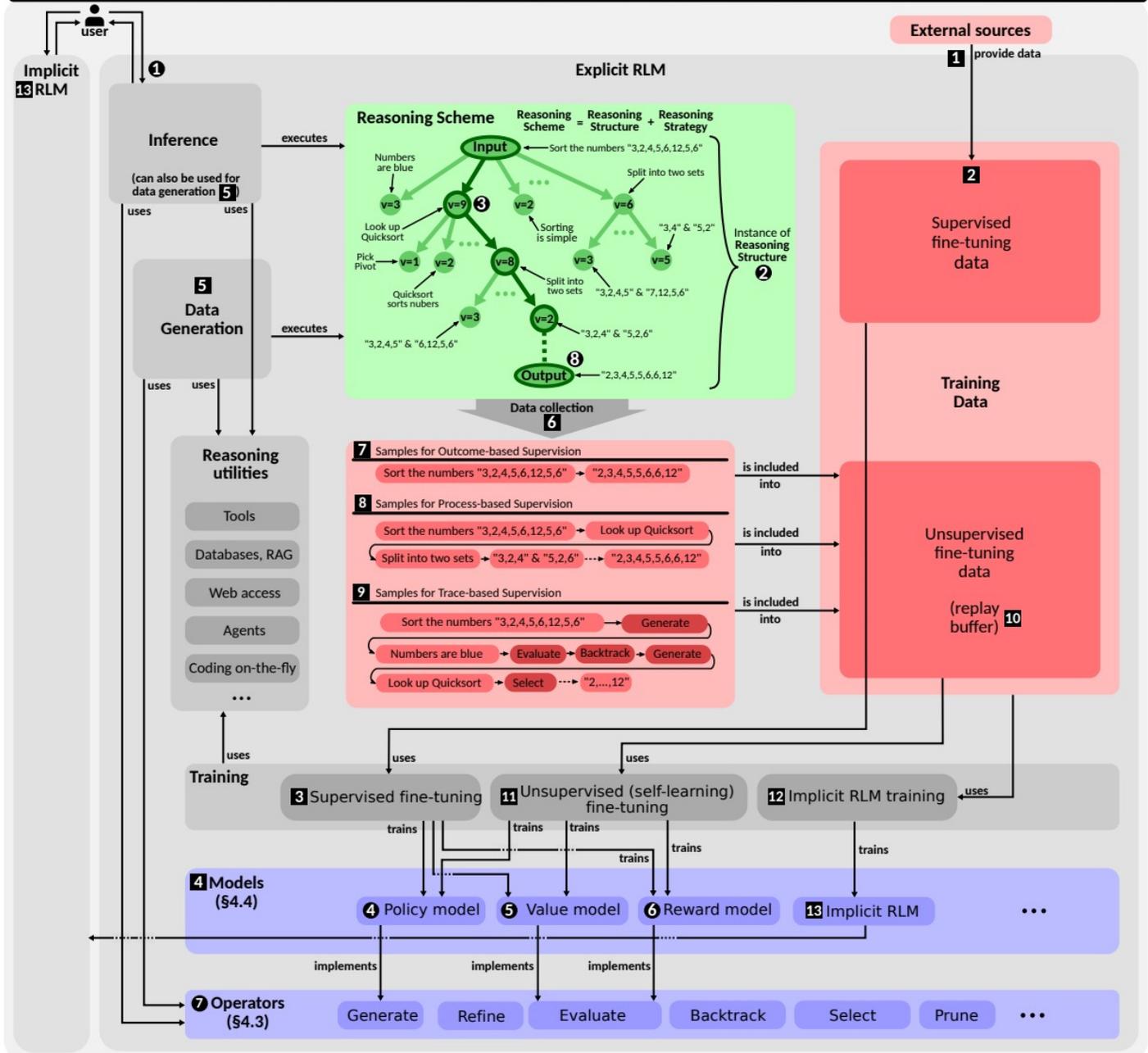
## Medium-level overview (§3.1)

6

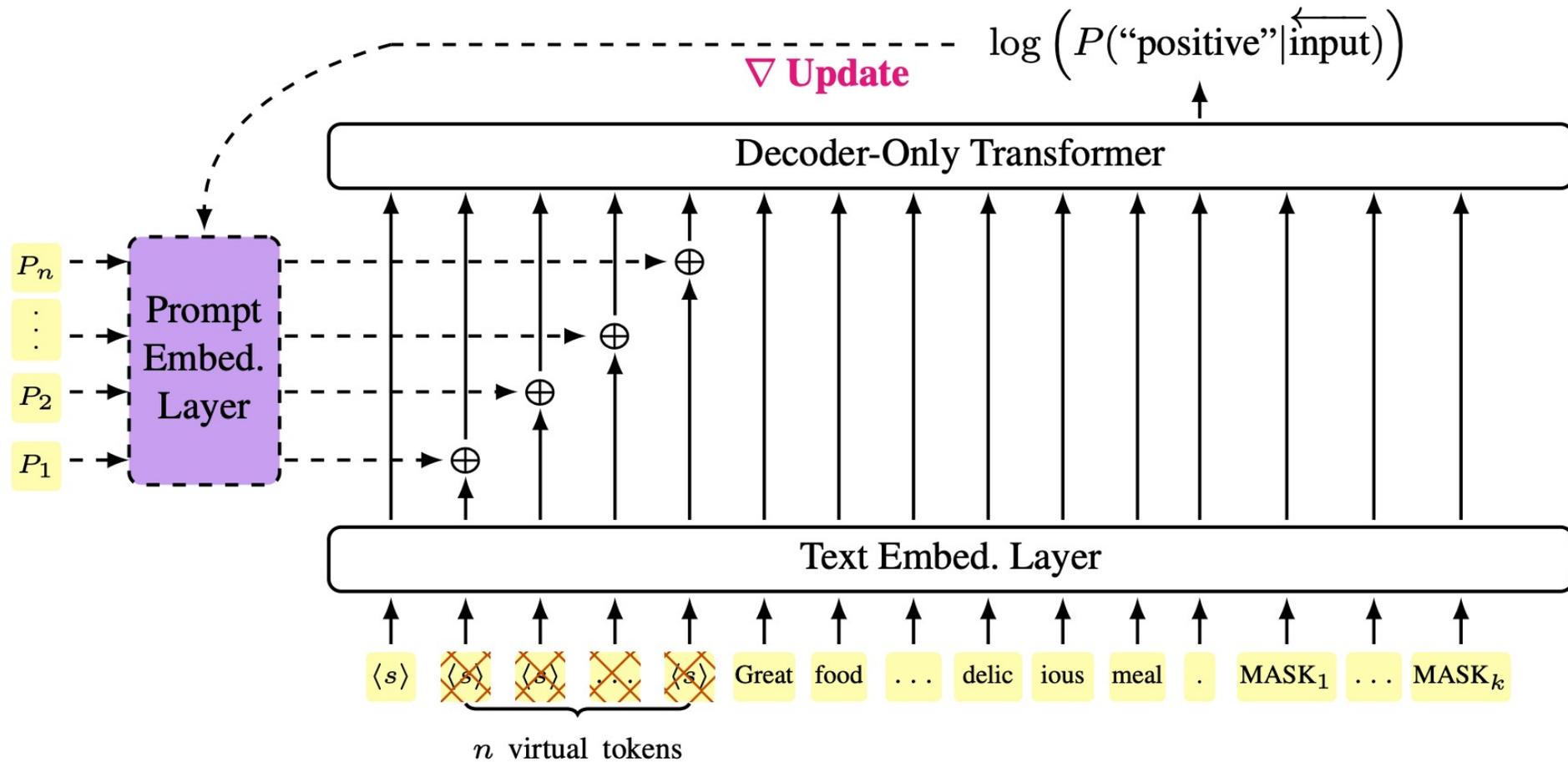




# Reasoning Models



# Soft Prompting



Questions?

