



STAT 453: Introduction to Deep Learning and Generative Models

Ben Lengerich

Lecture 23: Supervised Fine-tuning of LLMs

November 24, 2025

Reading: See course homepage



Today

- Optional HW5 out today
- [Project Presentation Sign-up](#)
 - **4 minute presentations!**
- Project Final Report
 - Due Friday December 12th
 - Submit PDF via Canvas
- Final Exam
 - December 17th, 5:05-7:05PM
 - **Science 180**
 - **Study Guide Released**



Today

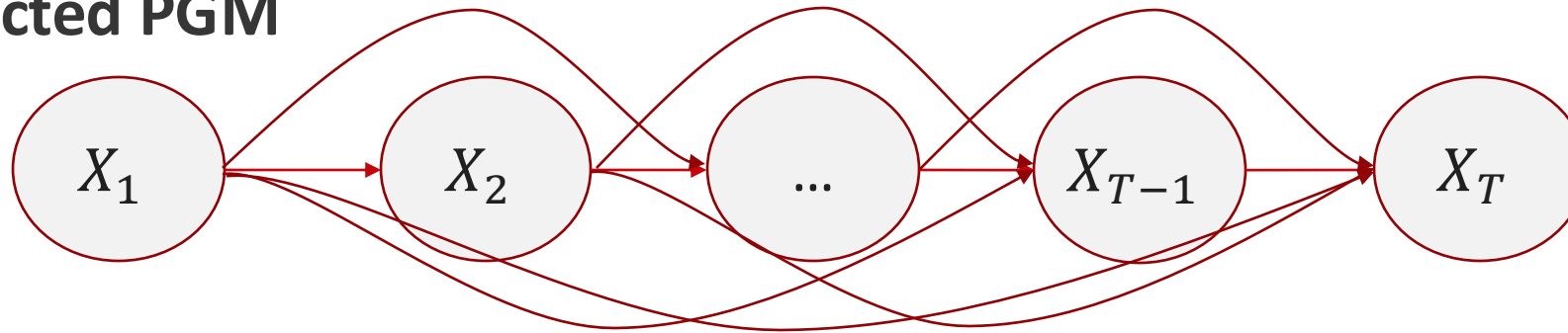
- Supervised Fine-tuning of LLMs
 - Alignment / Reinforcement Learning
- Efficient Parameter Fine-tuning / Personalization



Supervised Fine-Tuning of LLMs

Recall GPT training objective: MLE

- **Directed PGM**



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

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

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

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

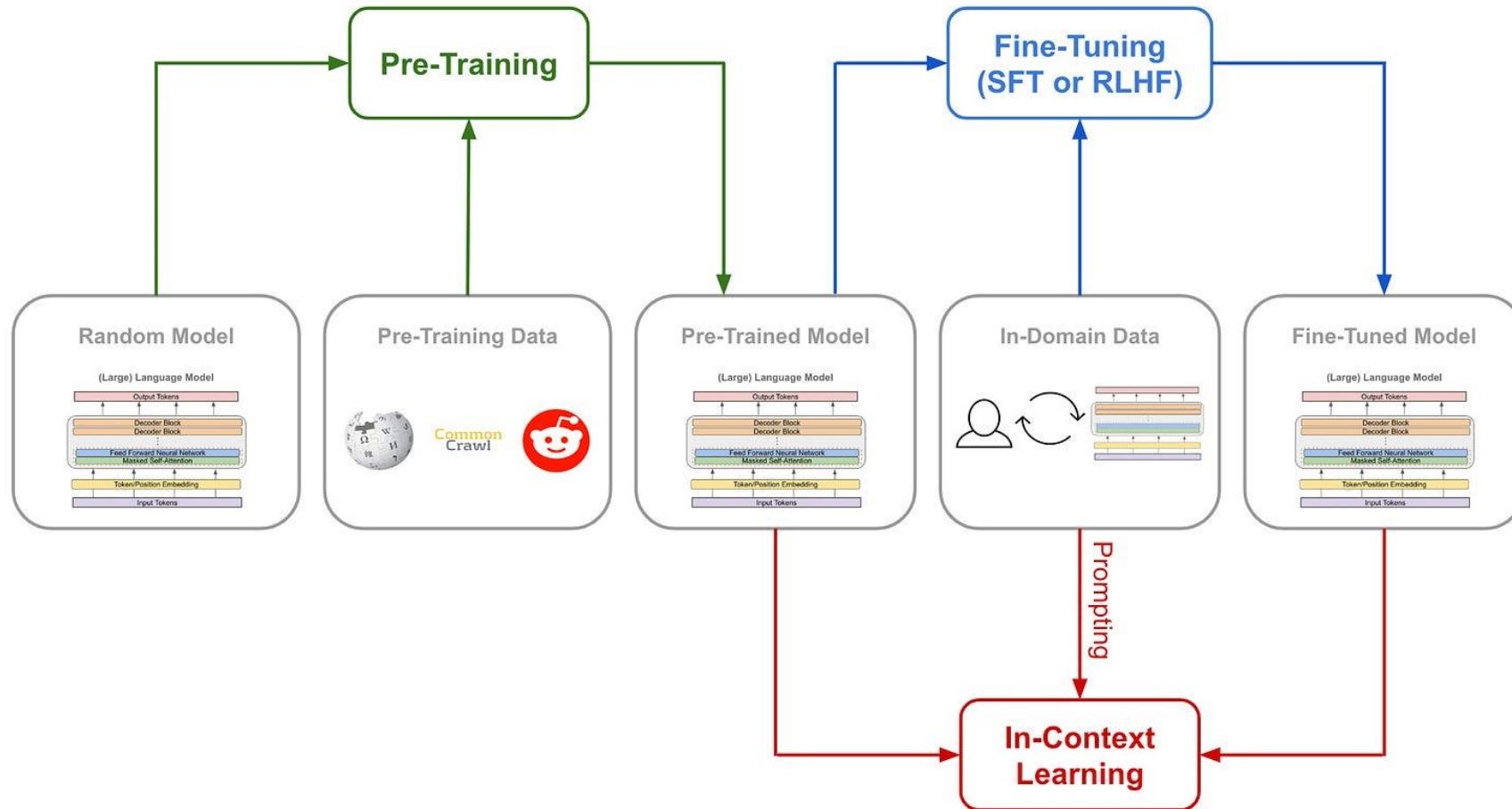
What does MLE not do?

- 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

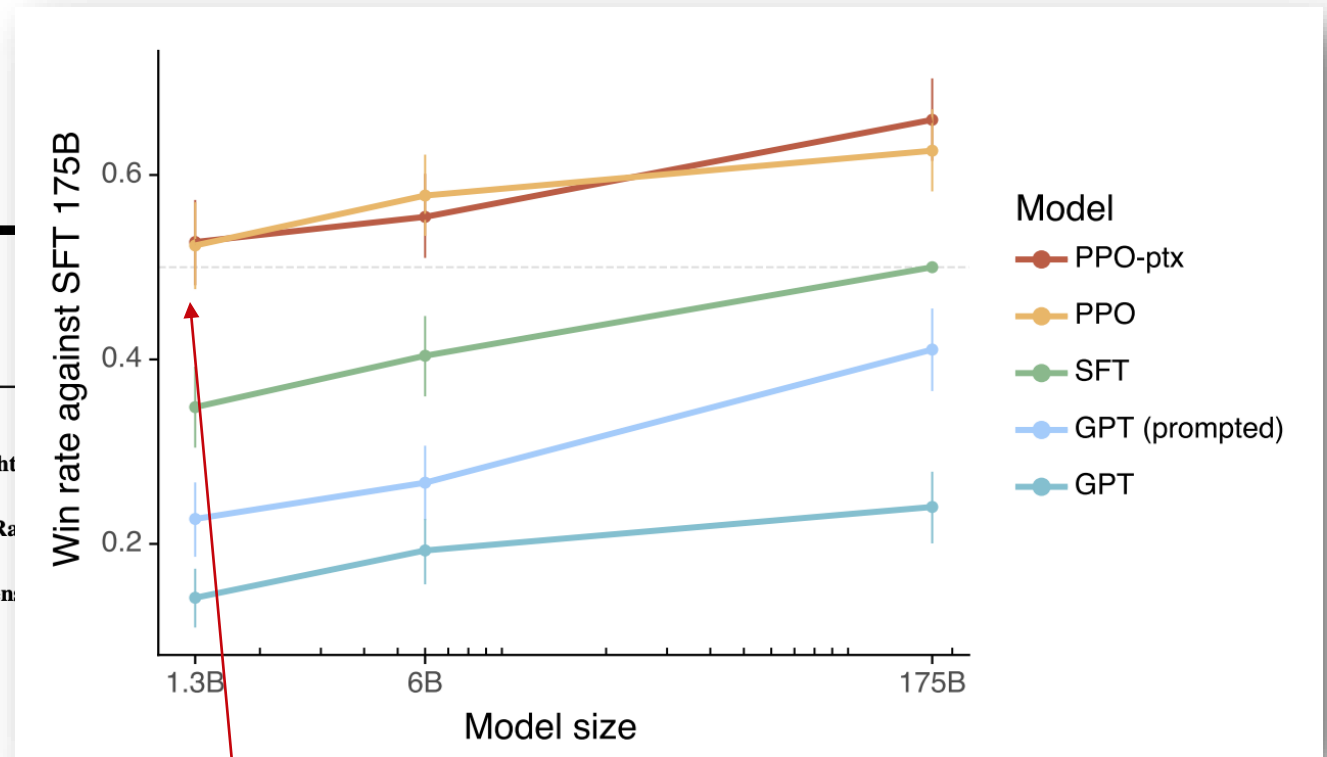
Supervised Fine-Tuning (SFT)

- 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

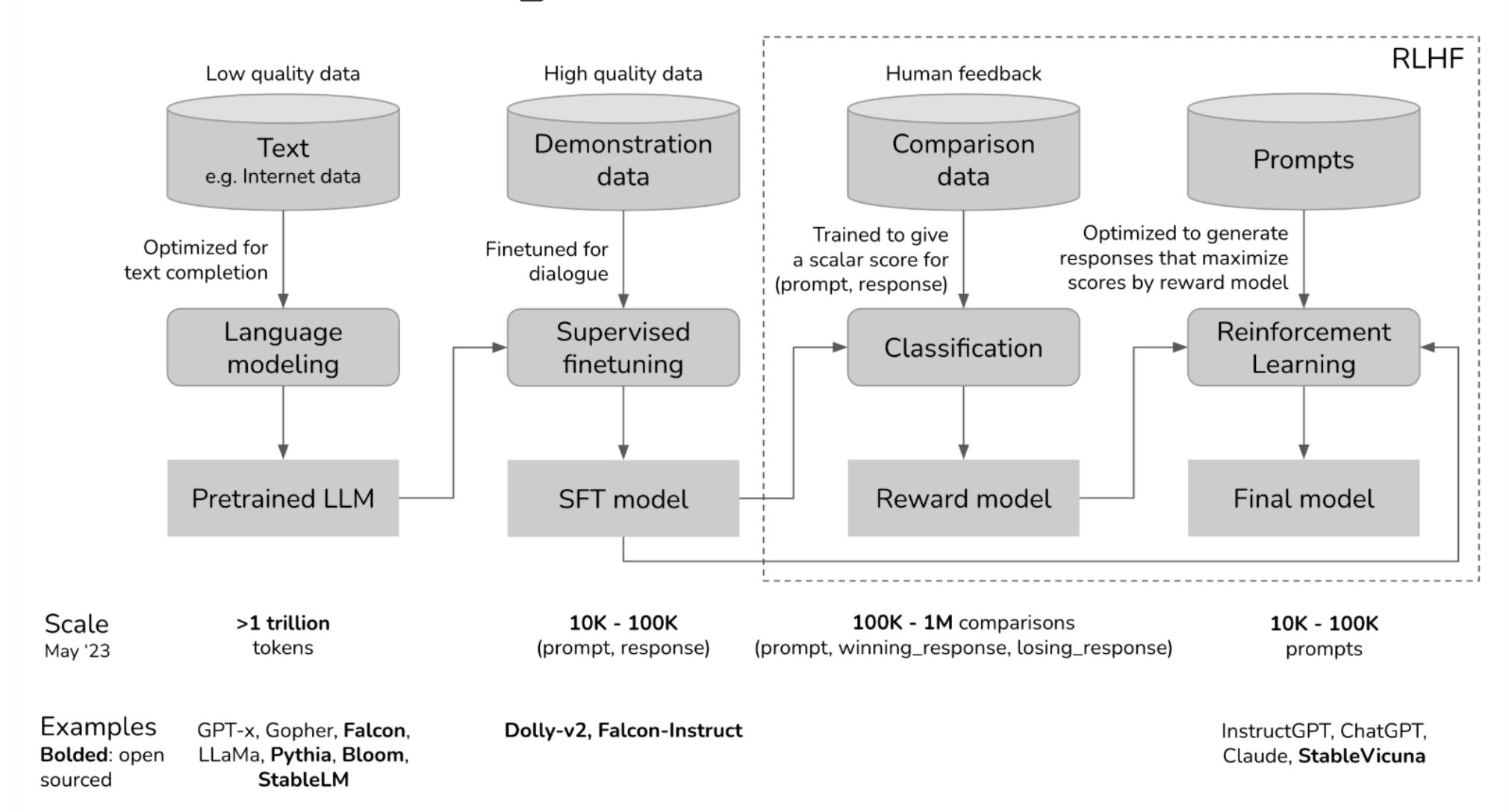
Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ra
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simen
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*

OpenAI



1.3B model can outperform 175B model

Reinforcement Learning with Human Feedback



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. $-E_x \log(\sigma(s_w - s_l))$

High-quality data is critical

OpenAI UI

Submit

Skip

«Page 3 / 11»

Total time: 05:39

Instruction

Summarize the following news article:

====
{article}
=====

Include output

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...

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...

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 1 (best)

Rank 2

Rank 3

Rank 4

Rank 5 (worst)

(b)

Ben Lengerich © University of Wisconsin-Madison 2025

<https://huyenchip.com/2023/05/02/r1hf.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	0.196

Dataset

TruthfulQA

GPT	0.224
Supervised Fine-Tuning	0.206
InstructGPT	0.413

API Dataset

Hallucinations

GPT	0.414
Supervised Fine-Tuning	0.078
InstructGPT	0.172

API Dataset

Customer Assistant Appropriate

GPT	0.811
Supervised Fine-Tuning	0.880
InstructGPT	0.902

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.

Reinforcement Learning with Verifiable Rewards

- RLVR
- Better than human feedback: verifiable truth
- Examples:
 - Code generation (verify: does it run correctly?)
 - Math questions (verify: did you solve it?)
 - Formatting-specifics (verify: did output match format requirements?)

Parameter Efficient 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)

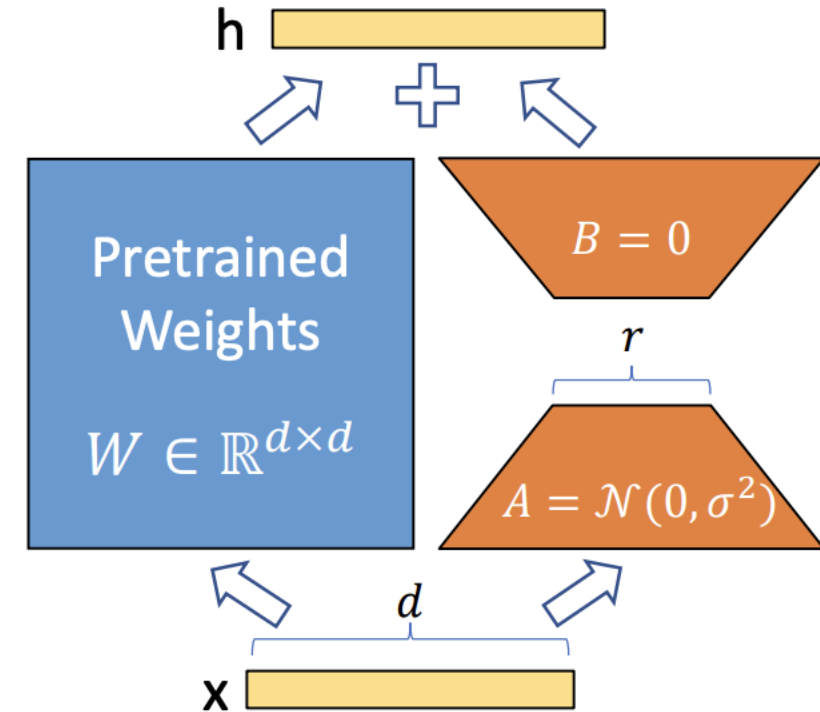
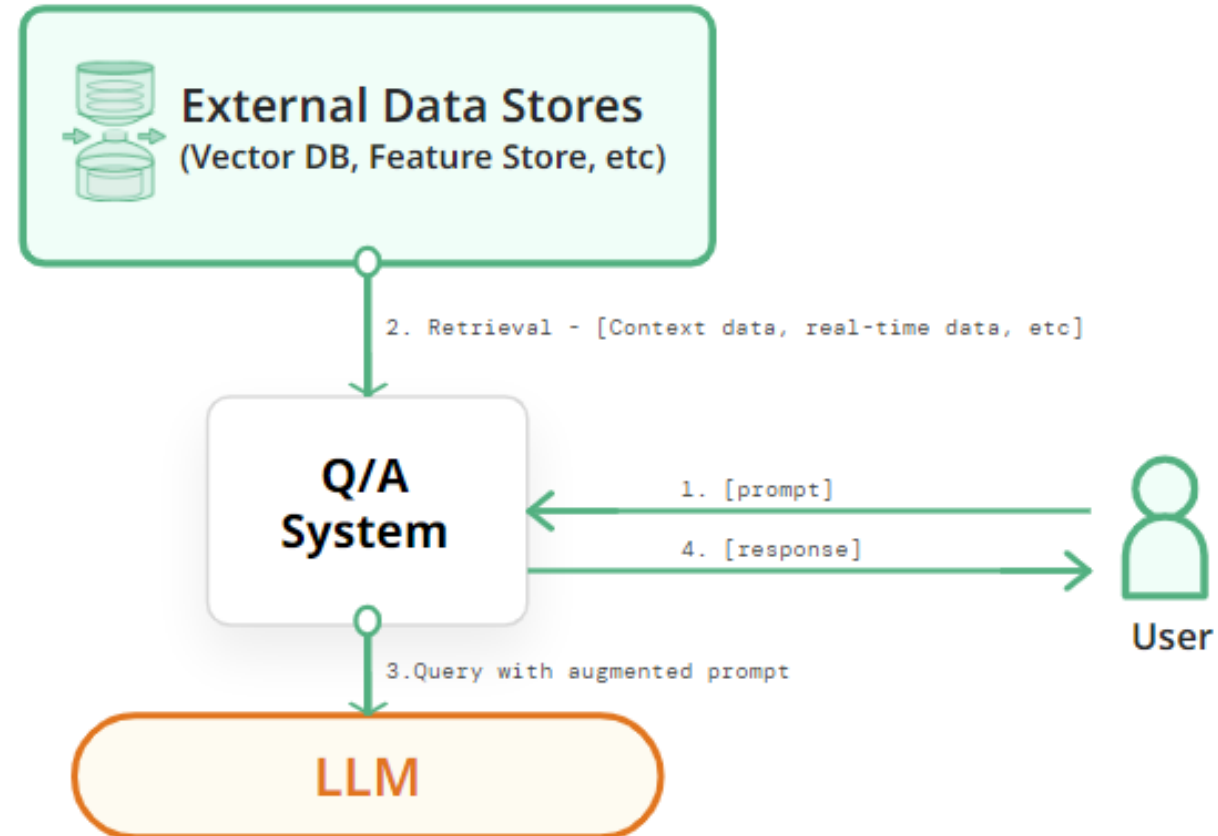


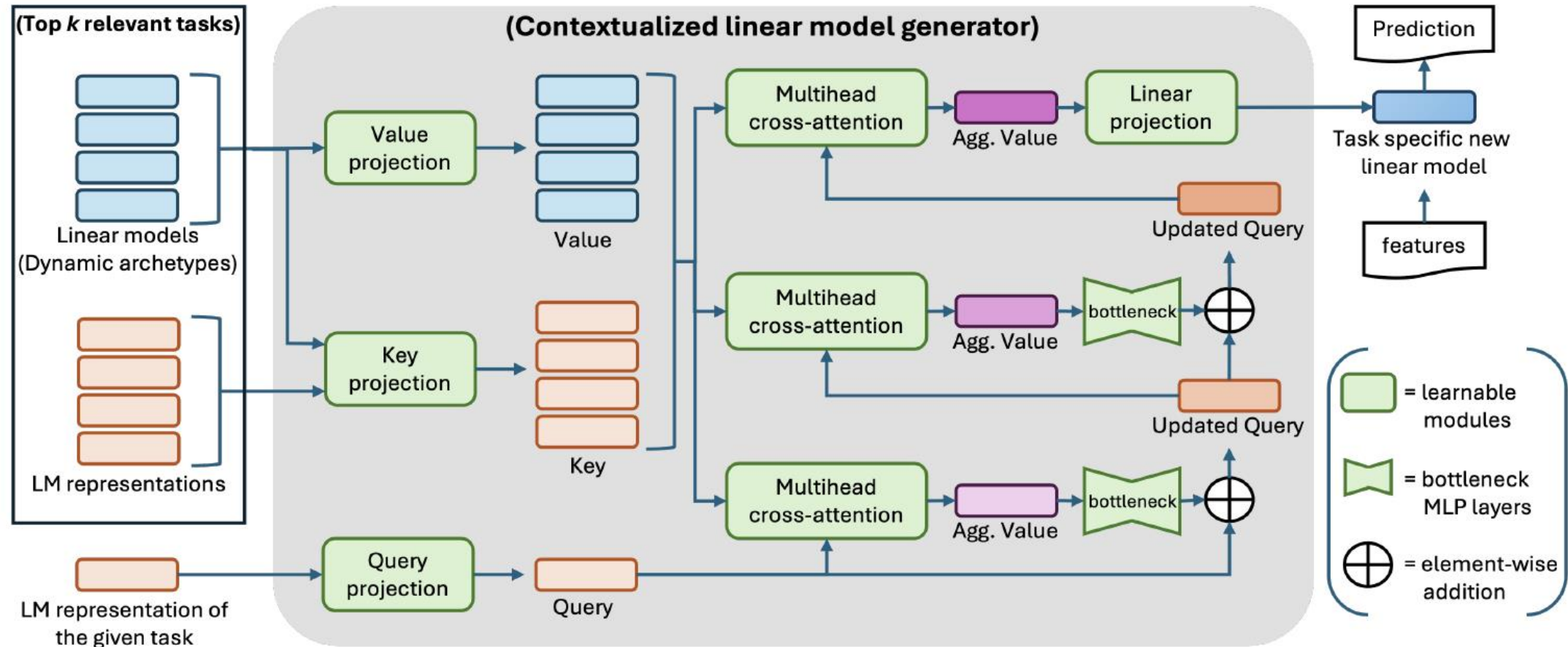
Figure 1: Our reparametrization. We only train A and B .

Retrieval-Augment Generation

- Resource access enables personalization

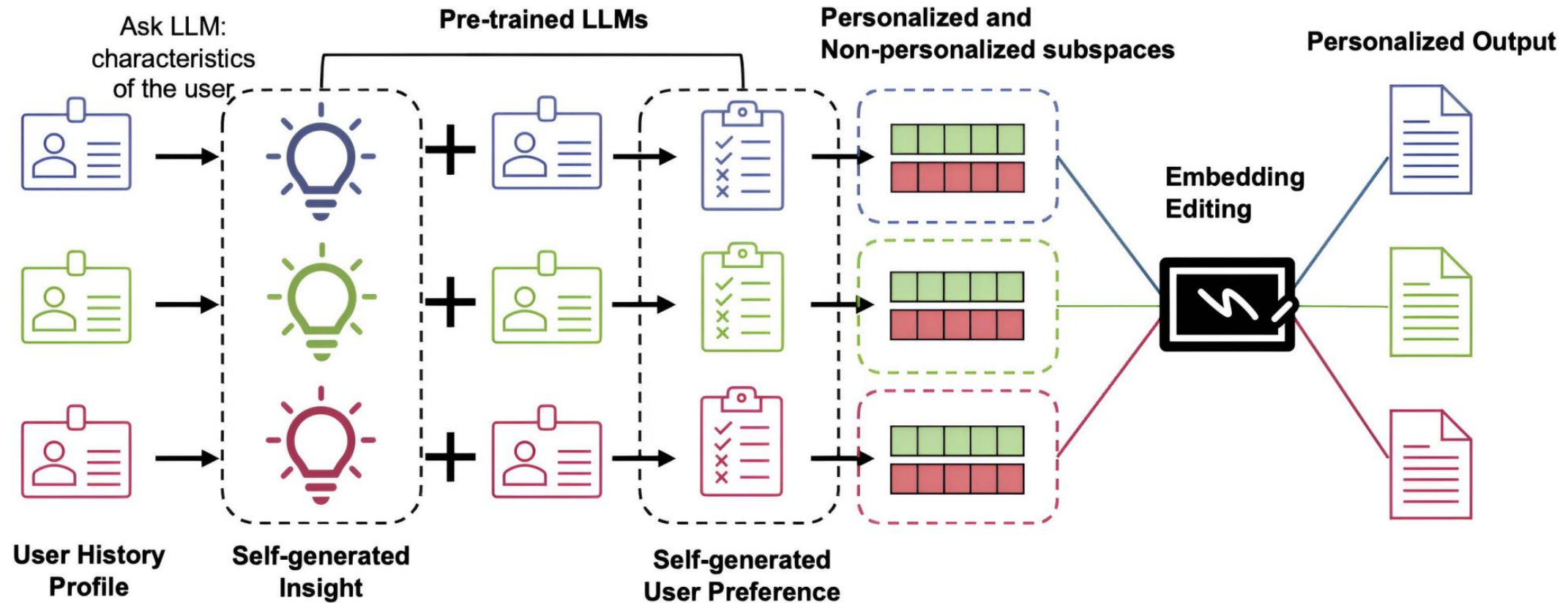


RAG of Interpretable Models (RAG-IM)



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>

Questions?

