STAT 992: Foundation Models for Biomedical Data

Ben Lengerich

Lecture 11: Prompts, In-Context Learning, Test-Time Adaptation

March 2, 2026

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?`

  ≡ **Is** `P(...because` **`the cat`** `was too big) >=`
  `P(...because` **`the hat`** `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 _____.
Target word: Gabriel

LAMBADA (language modeling w/ long discourse dependencies) [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	**35.13**	45.99	**87.65**	**83.4**	**29.41**
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**

[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]:

```
SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook
the San Francisco
...
overturn unstable
objects. TL;DR:
```

2018 SoTA

Supervised (287K)

Select from article

"Too Long, Didn't Read"

"Prompting"?

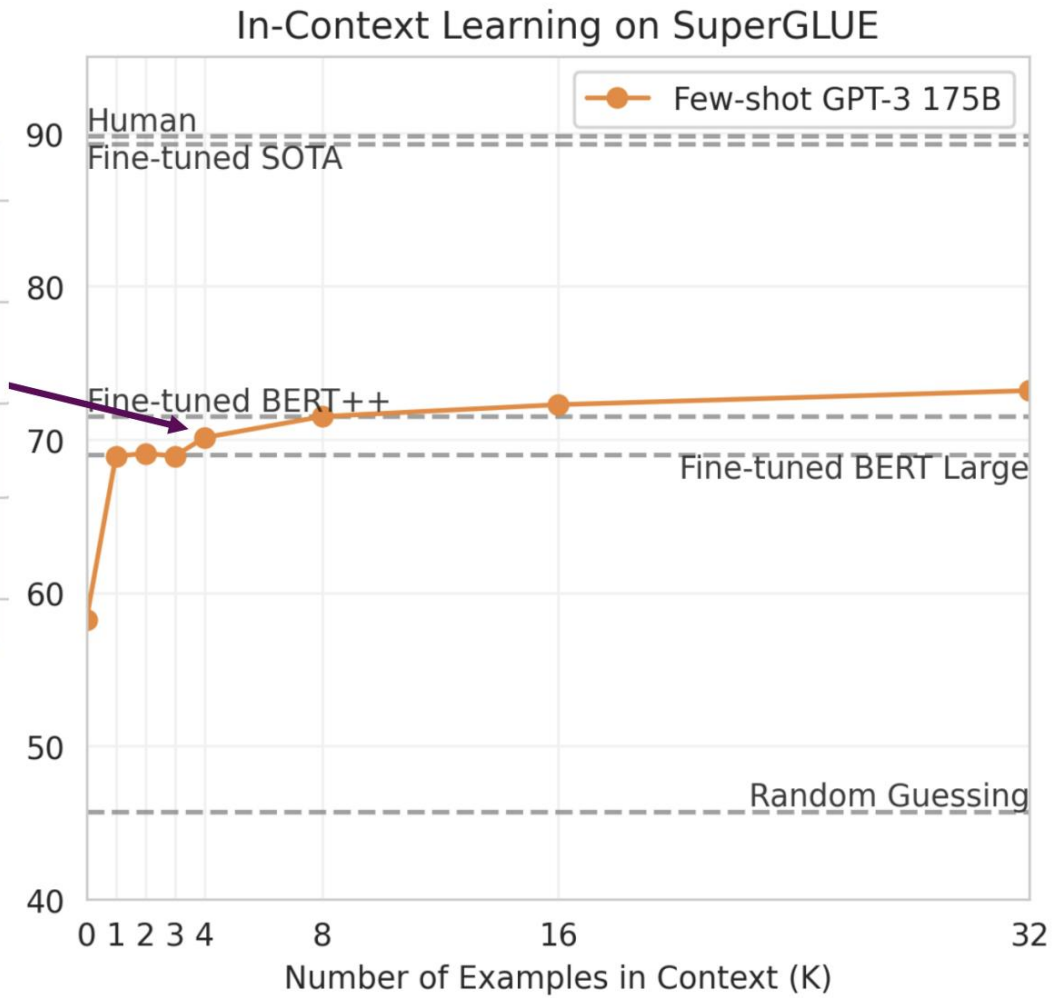
	ROUGE		
	R-1	R-2	R-L
Bottom-Up Sum	**41.22**	**18.68**	**38.34**
Lede-3	40.38	17.66	36.62
Seq2Seq + Attn	31.33	11.81	28.83
GPT-2 TL;DR:	29.34	8.27	26.58
Random-3	28.78	8.63	25.52

[Radford et al., 2019]

“In-Context Learning”

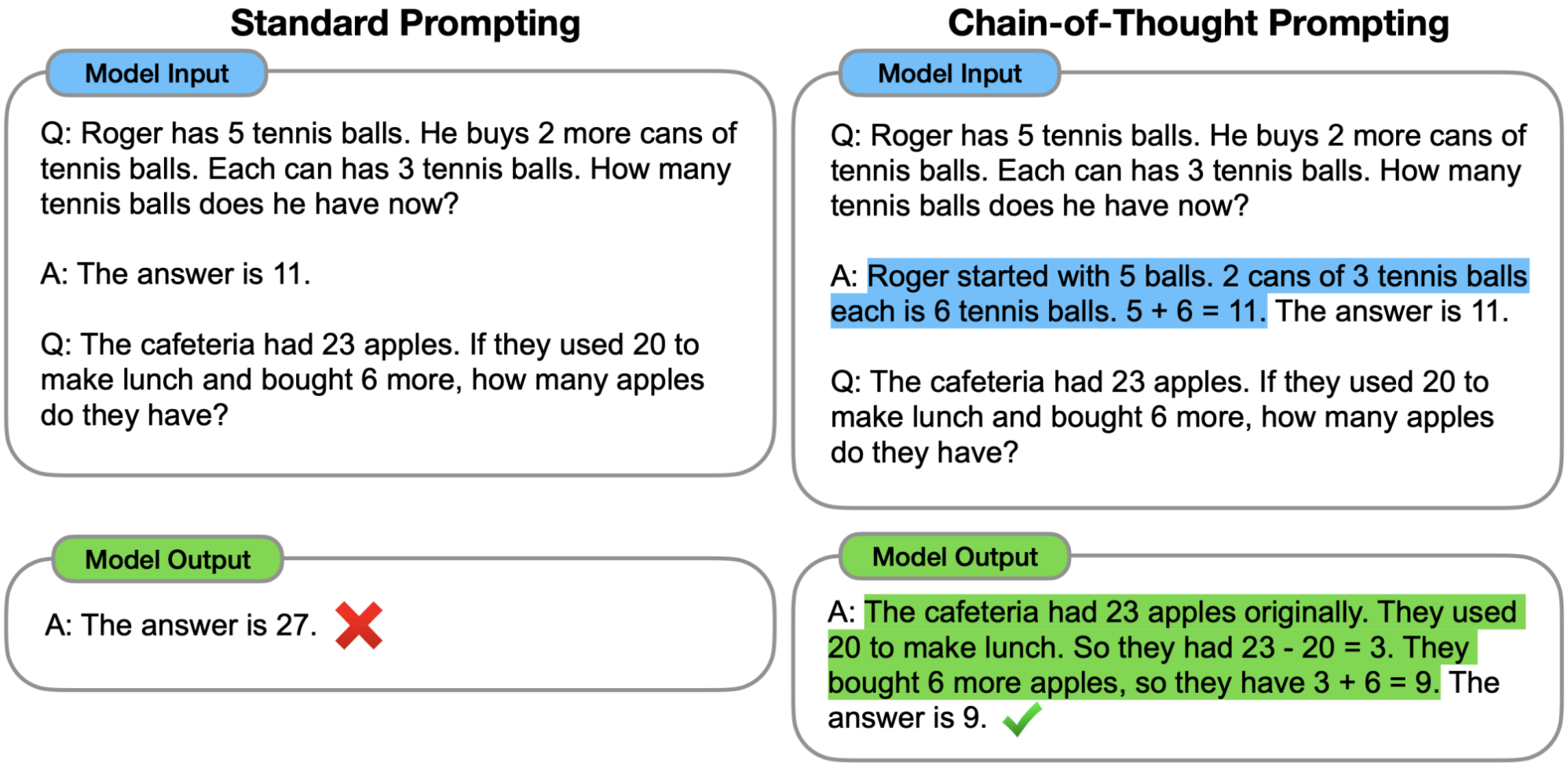
Few-shot

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



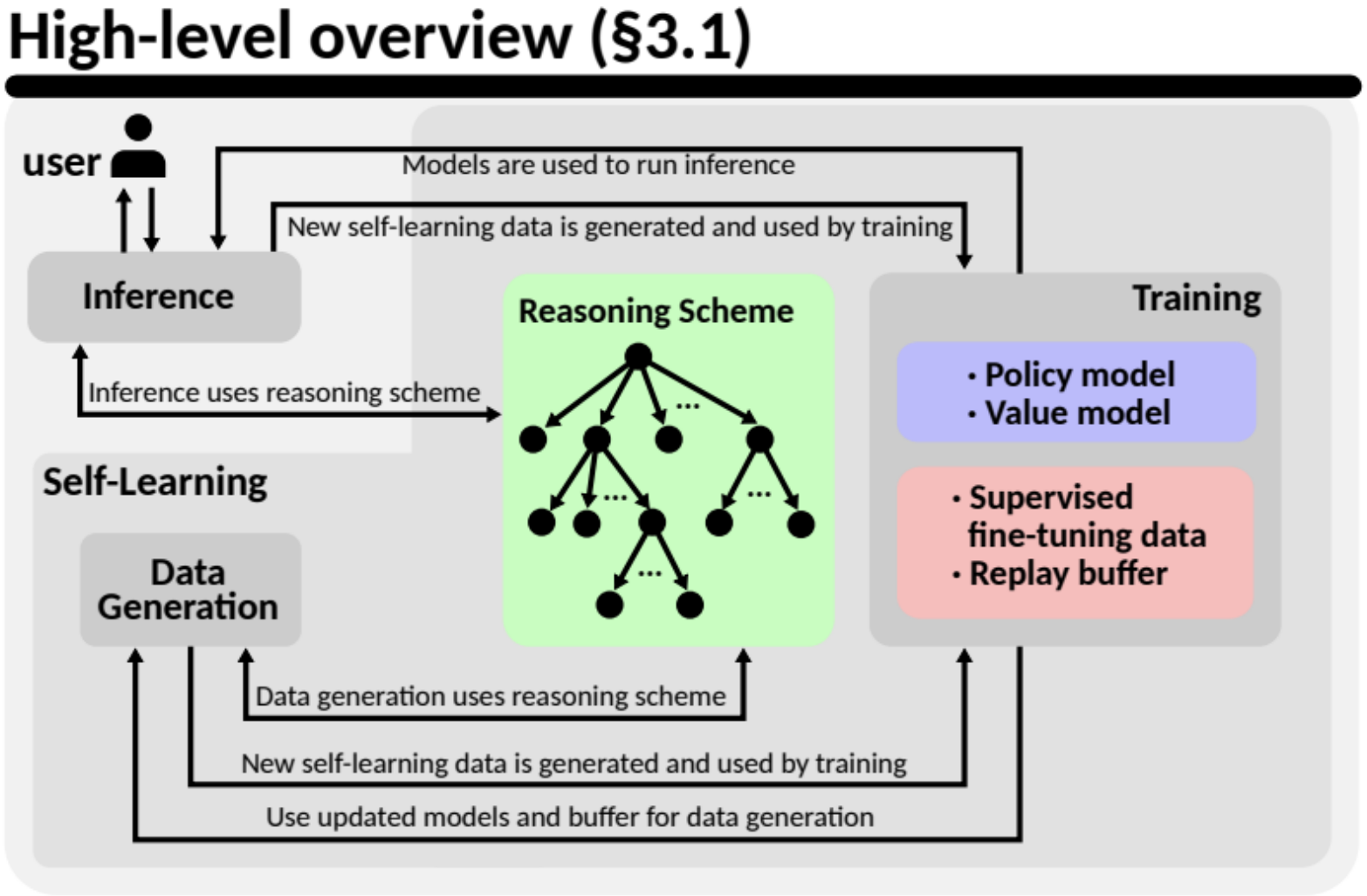
[Brown et al., 2020]

Chain-of-Thought



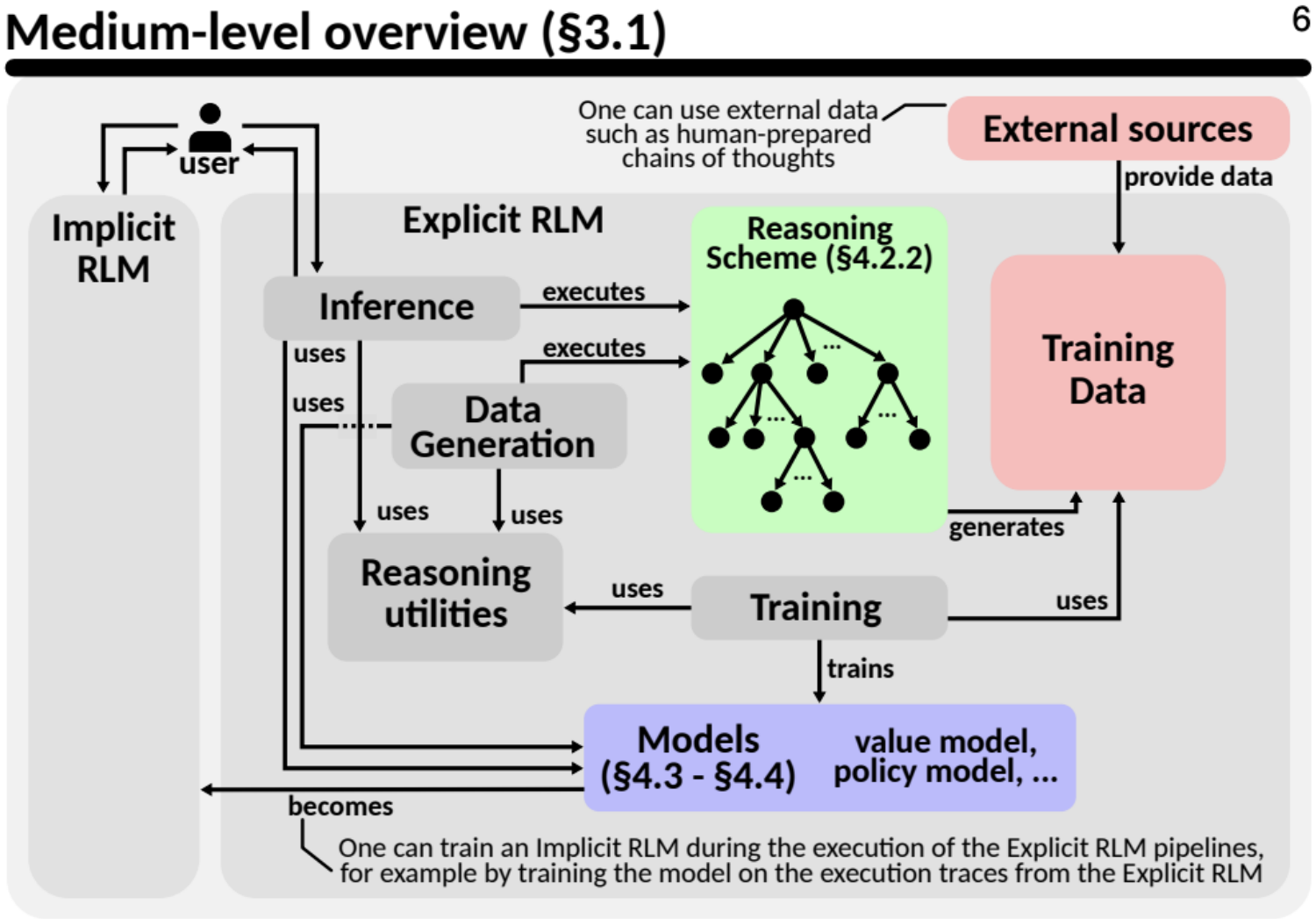
Wei, et al. (2023) Chain-of-Though Prompting Elicits Reasoning in LLMs

Reasoning Models



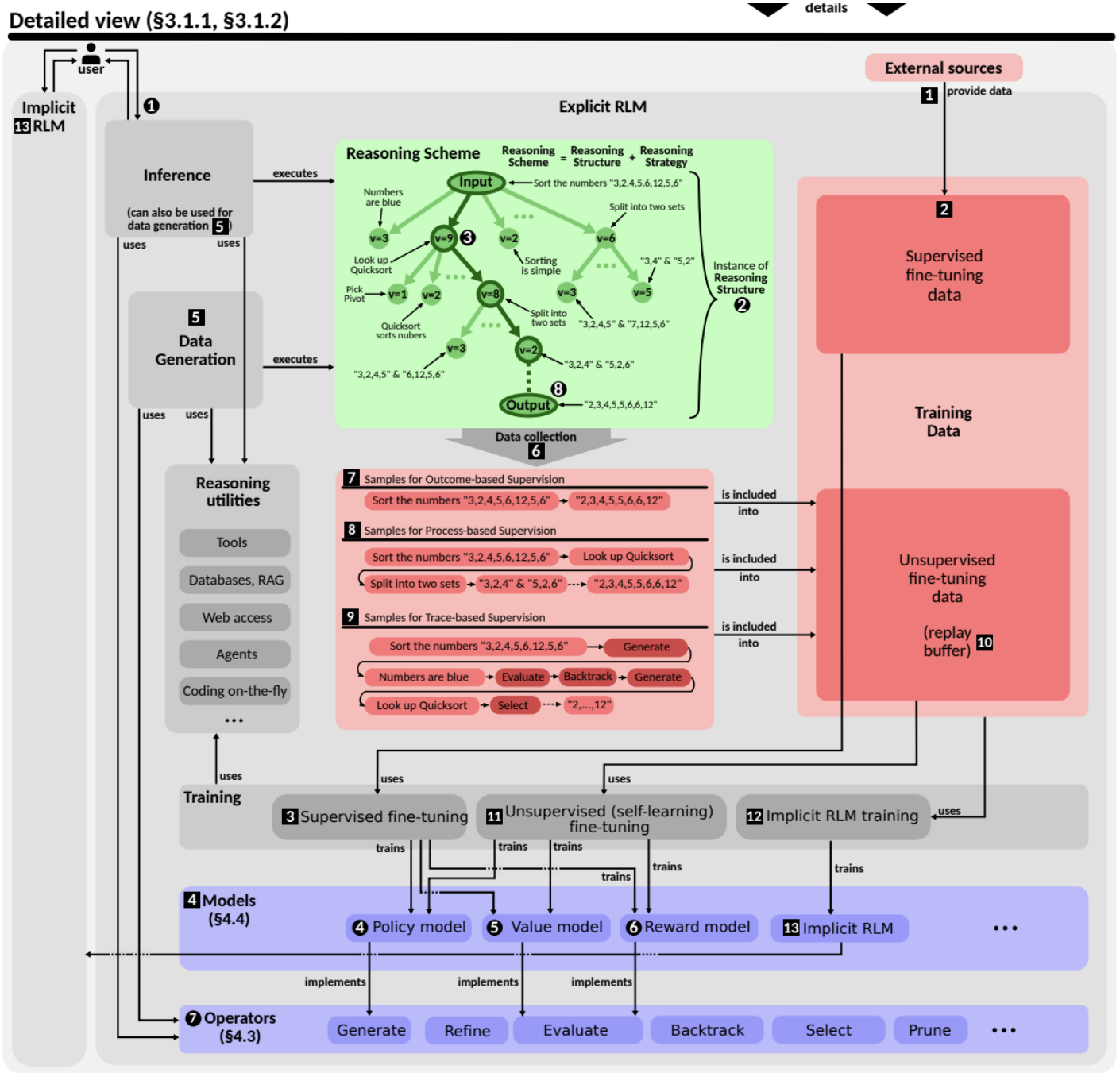
https://arxiv.org/pdf/2501.11223

Reasoning Models



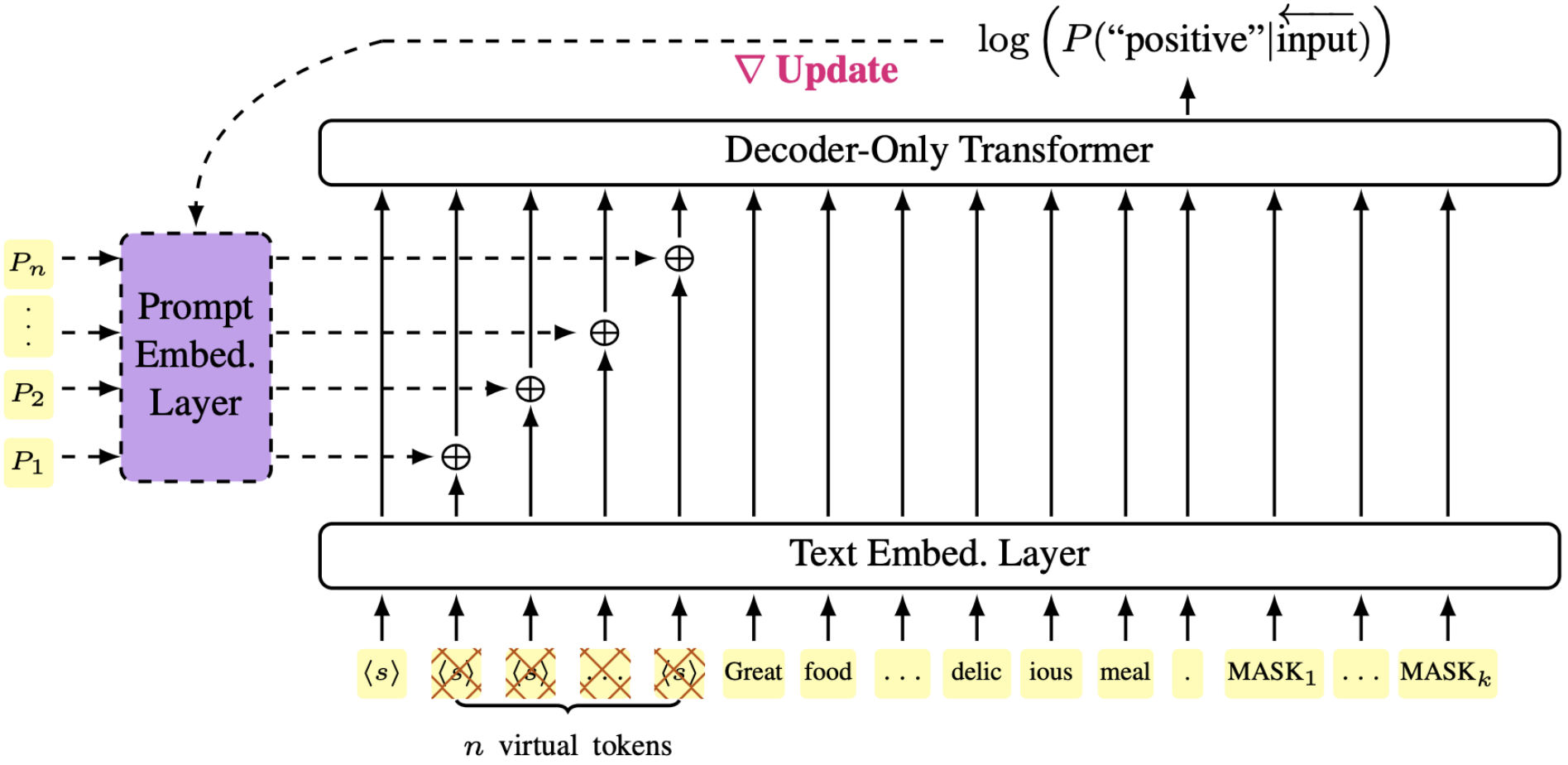
https://arxiv.org/pdf/2501.11223

Reasoning Models



https://arxiv.org/pdf/2501.11223

Soft Prompting



https://arxiv.org/pdf/2306.04735

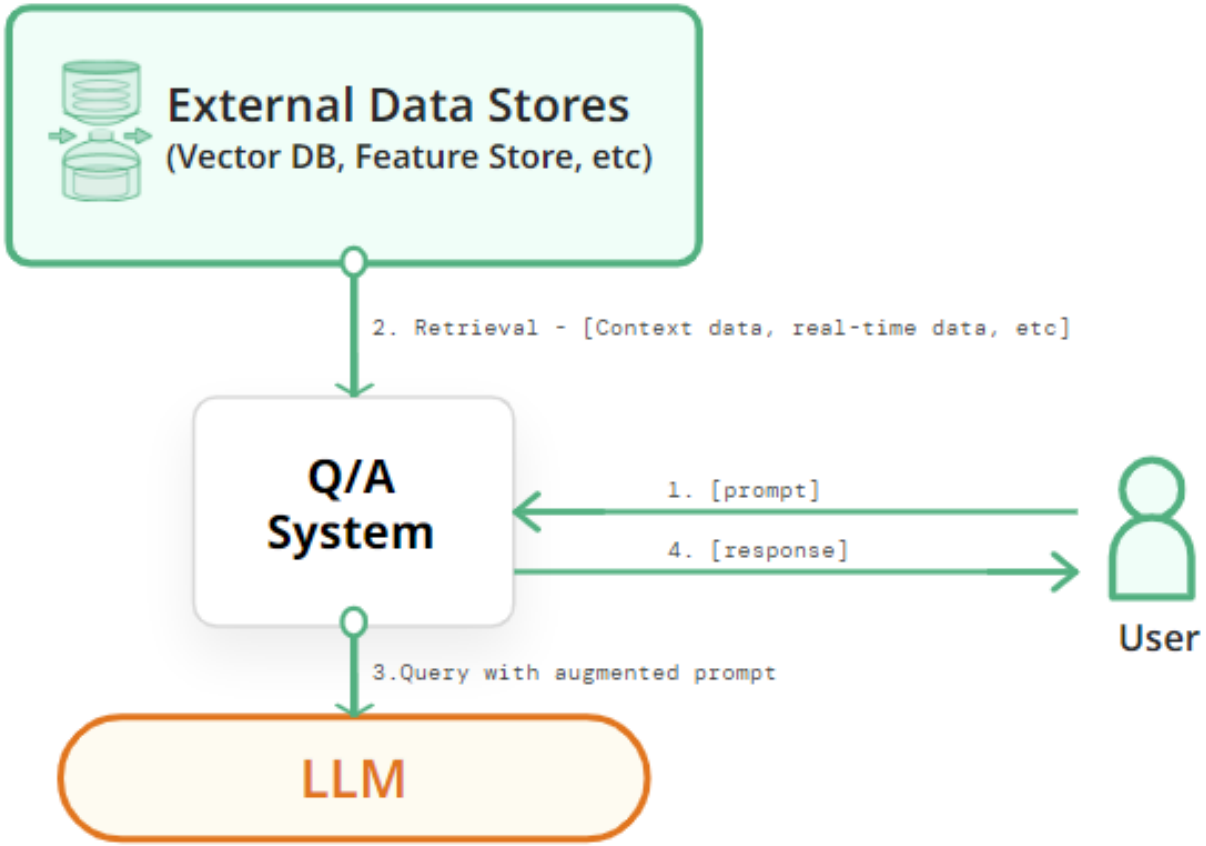
Parameter Efficient Fine-Tuning

Personalization / Adaptation / Alignment

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

Retrieval-Augment Generation

- Resource access enables personalization



https://www.hopsworks.ai/dictionary/retrieval-augmented-generation-llm

Questions?