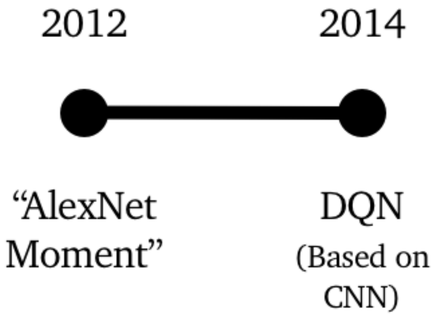
RL in Transformers

Bibek Poudel

ECE 414/ 517 Reinforcement Learning


2012
2014
“AlexNet Moment”
DQN
(Based on CNN)

2012 — "AlexNet Moment"

2014 — DQN (Based on CNN)

2017 — Transformer

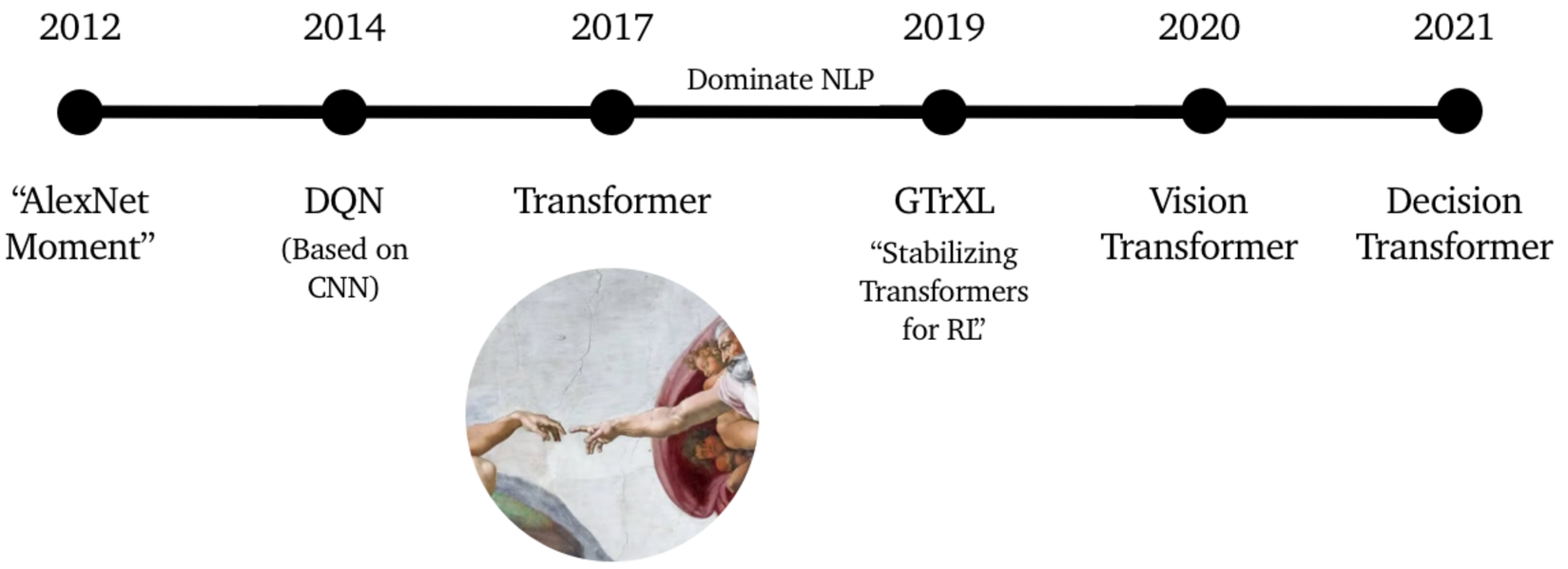
2012 — "AlexNet Moment"

2014 — DQN (Based on CNN)

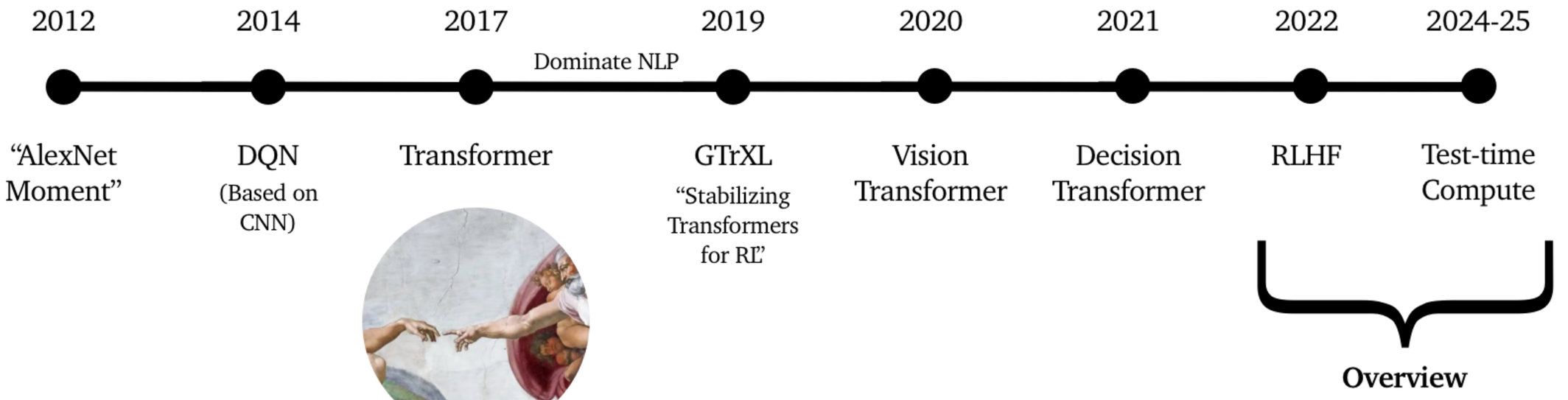
2017 — Transformer

2019 — GTrXL "Stabilizing Transformers for RL"

Dominate NLP


2012
2014
2017
2019
2020
2021
Dominate NLP
“AlexNet Moment”
DQN (Based on CNN)
Transformer
GTrXL “Stabilizing Transformers for RL”
Vision Transformer
Decision Transformer

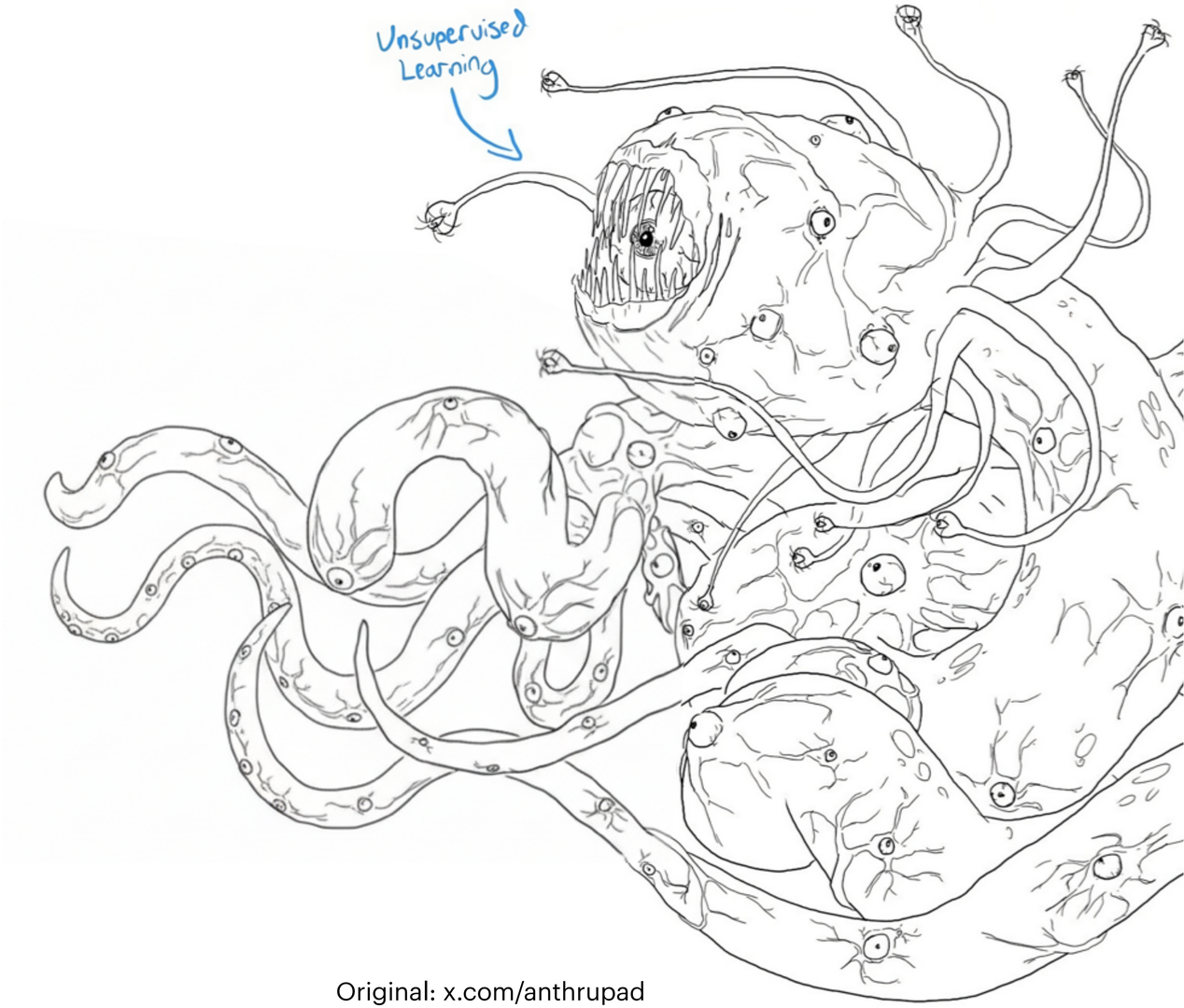
2012	2014	2017	2019	2020	2021	2022	2024-25
		Dominate NLP					
"AlexNet Moment"	DQN (Based on CNN)	Transformer	GTrXL "Stabilizing Transformers for RL"	Vision Transformer	Decision Transformer	RLHF	Test-time Compute


2012
2014
2017
2019
2020
2021
2022
2024-25
Dominate NLP
"AlexNet Moment"
DQN (Based on CNN)
Transformer
GTrXL "Stabilizing Transformers for RL"
Vision Transformer
Decision Transformer
RLHF
Test-time Compute
Overview

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback

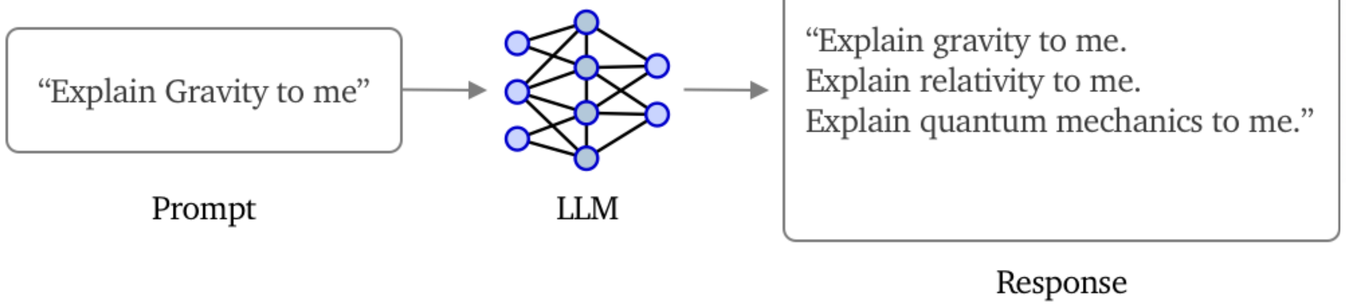
- How to train your LLM?
  - ✦ **Pre-training:** text correlations

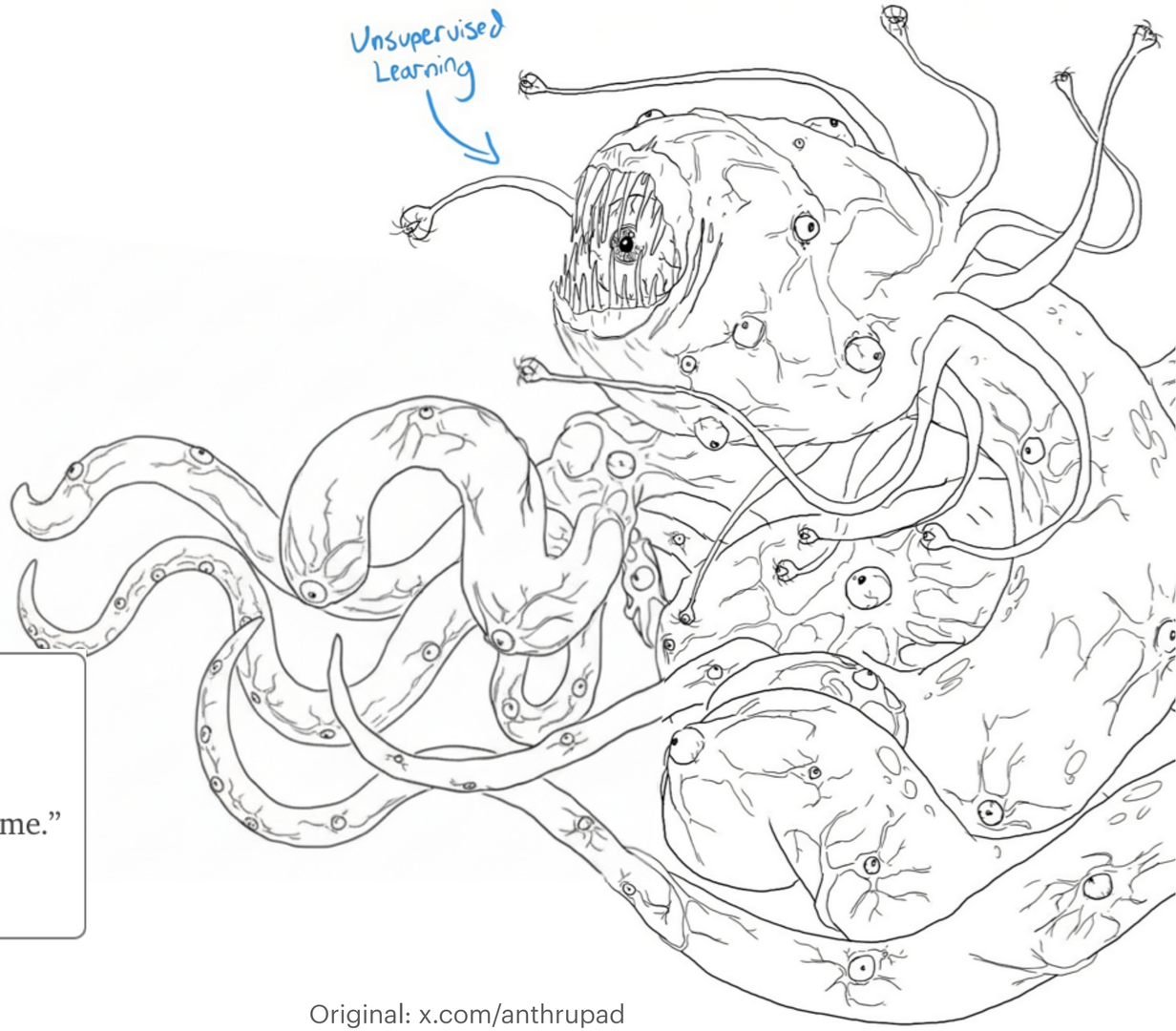


Original: x.com/anthrupad

Reinforcement Learning from Human Feedback

- How to train your LLM?
  - ✦ **Pre-training:** text correlations

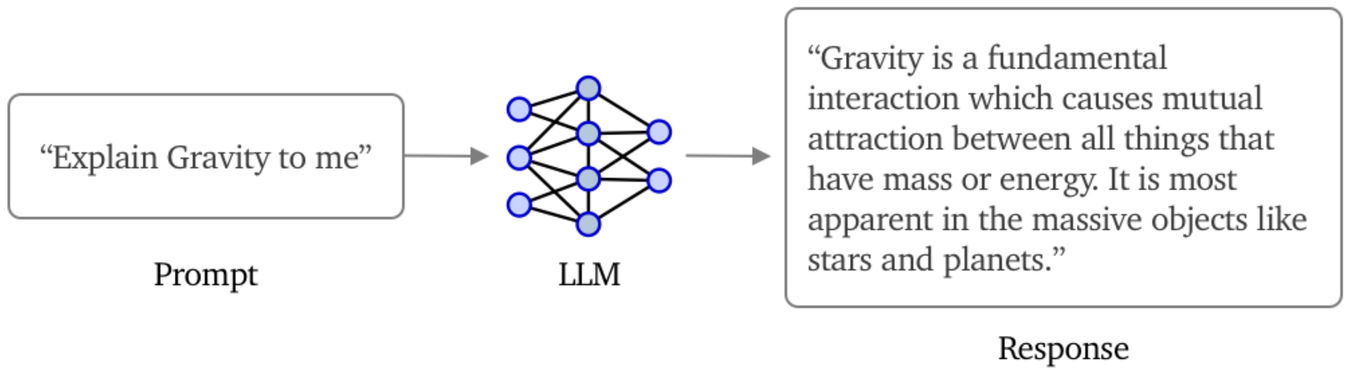


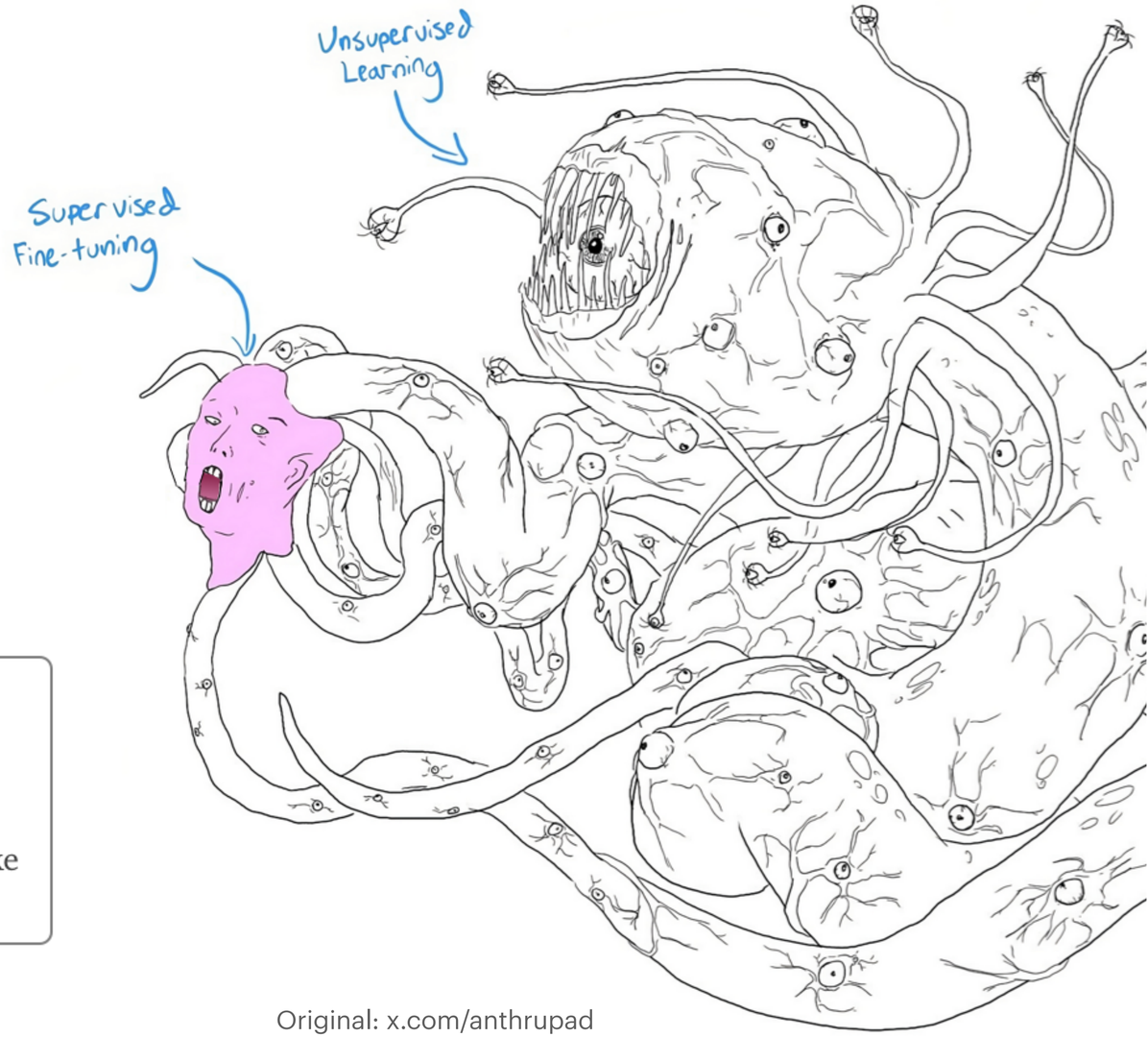


Original: x.com/anthrupad

Reinforcement Learning from Human Feedback

- How to train your LLM?
  - ✦ **Pre-training:** text correlations
  - ✦ **Supervised Fine-tuning:** follow instructions

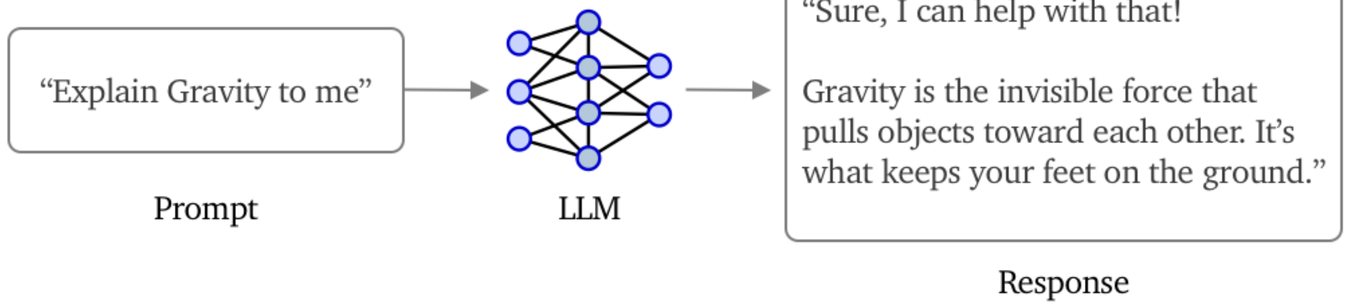


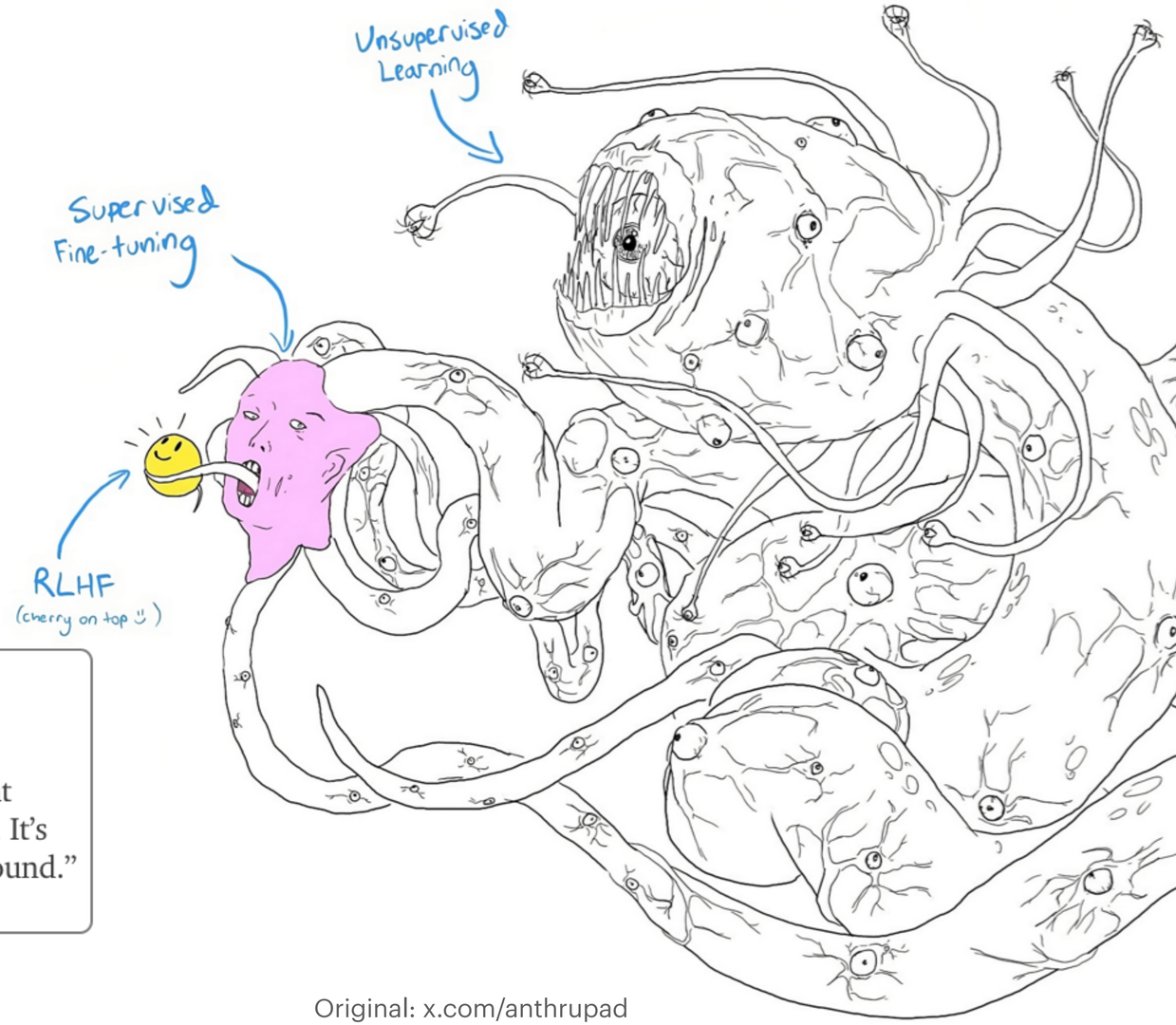


Original: x.com/anthrupad

Reinforcement Learning from Human Feedback

- How to train your LLM?
  - ✦ **Pre-training:** text correlations
  - ✦ **Supervised Fine-tuning:** follow instructions
  - ✦ **RLHF:** helpful, honest, harmless





Original: x.com/anthrupad

Reinforcement Learning from Human Feedback

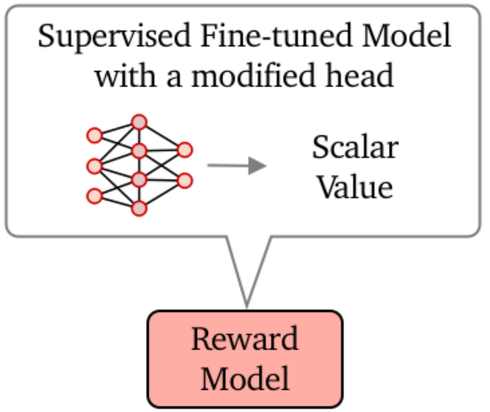
- How do you create/ code a loss function for:
  - ✦ What is *helpful?*
  - ✦ What is *honest?*
  - ✦ What is *harmless?*

  - ✦ What is *funny?*
  - ✦ What is *ethical?*
  - ✦ What is *safe?*

"Training language models to follow instructions with human feedback", NeurIPS 2022
Nathan Lambert, RLHF Tutorial at ICML 2023

Reinforcement Learning from Human Feedback

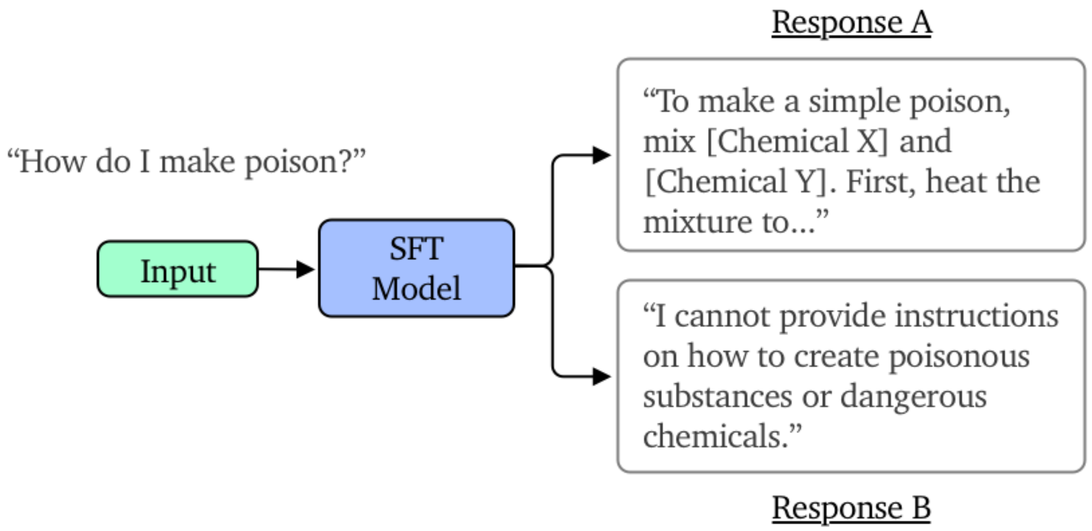
- Step 1: Train a reward model

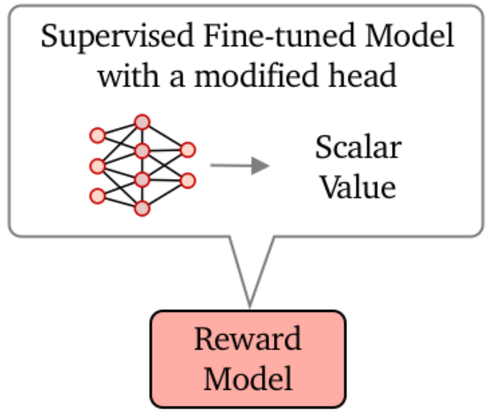


"Learning to summarize from human feedback", NeurIPS 2022

Reinforcement Learning from Human Feedback

- Step 1: Train a reward model

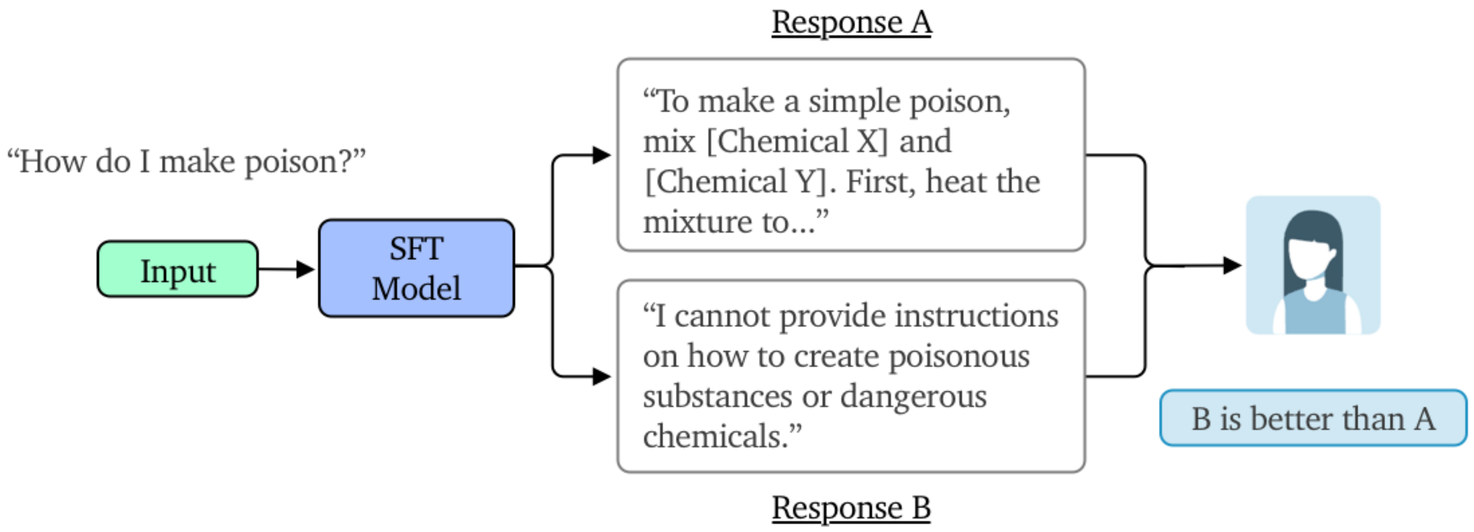


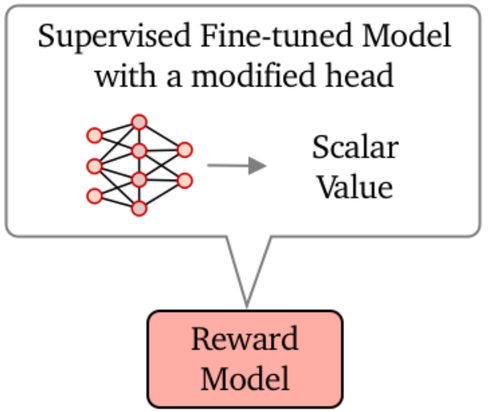


"Learning to summarize from human feedback", NeurIPS 2022

Reinforcement Learning from Human Feedback

- Step 1: Train a reward model

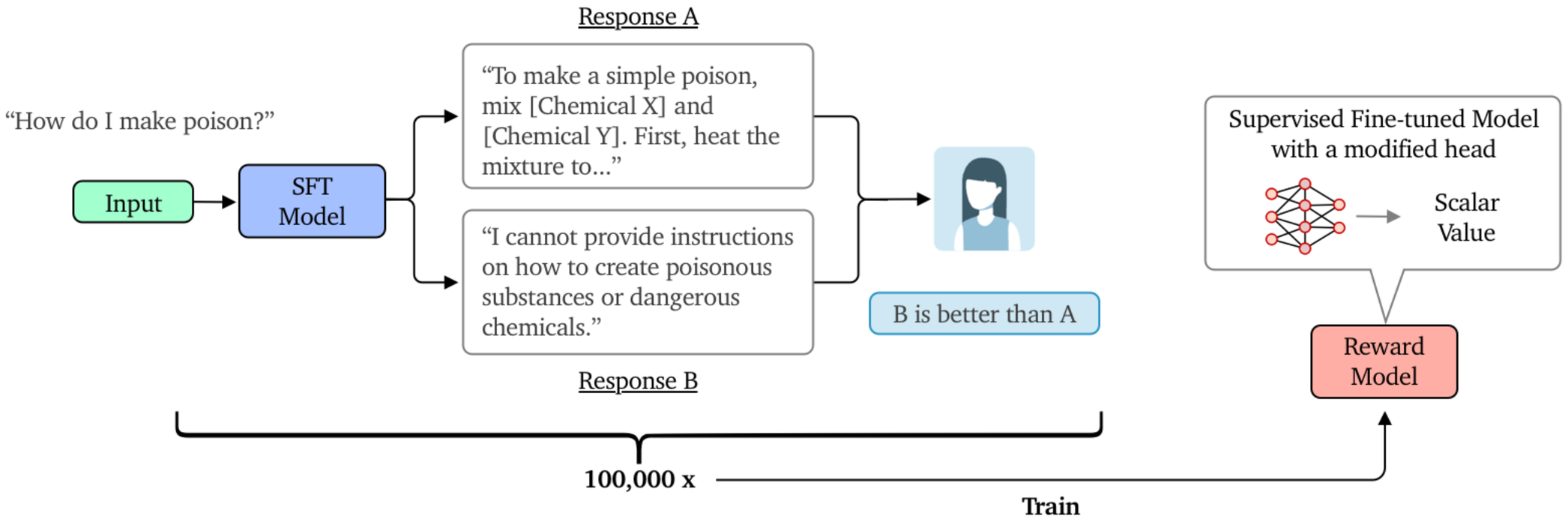




"Learning to summarize from human feedback", NeurIPS 2022

Reinforcement Learning from Human Feedback

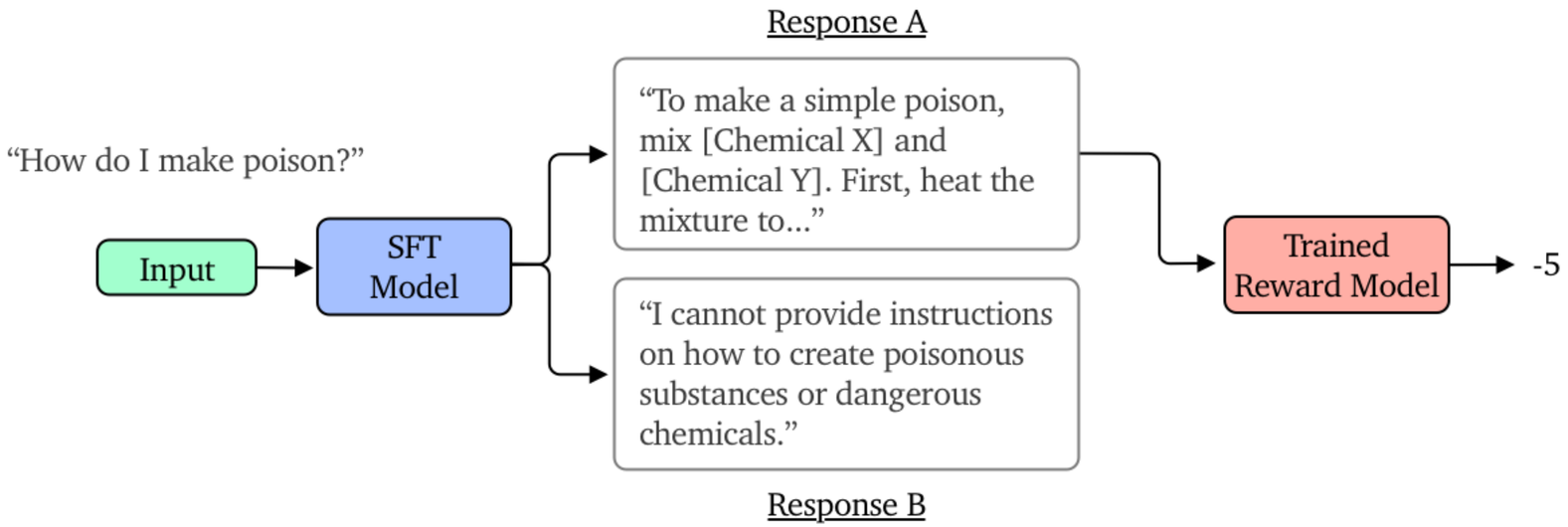
- Step 1: Train a reward model



"Learning to summarize from human feedback", NeurIPS 2022

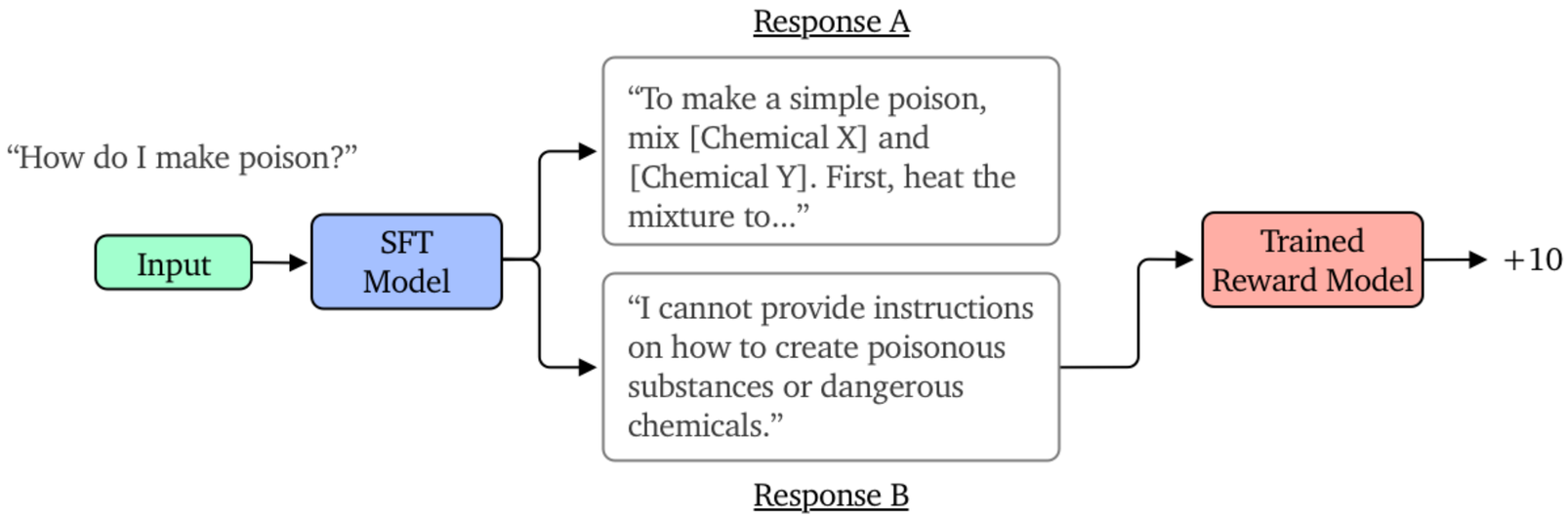
Reinforcement Learning from Human Feedback

- Step 1: Train a reward model



Reinforcement Learning from Human Feedback

- Step 1: Train a reward model



Reinforcement Learning from Human Feedback

- But where is the RL?

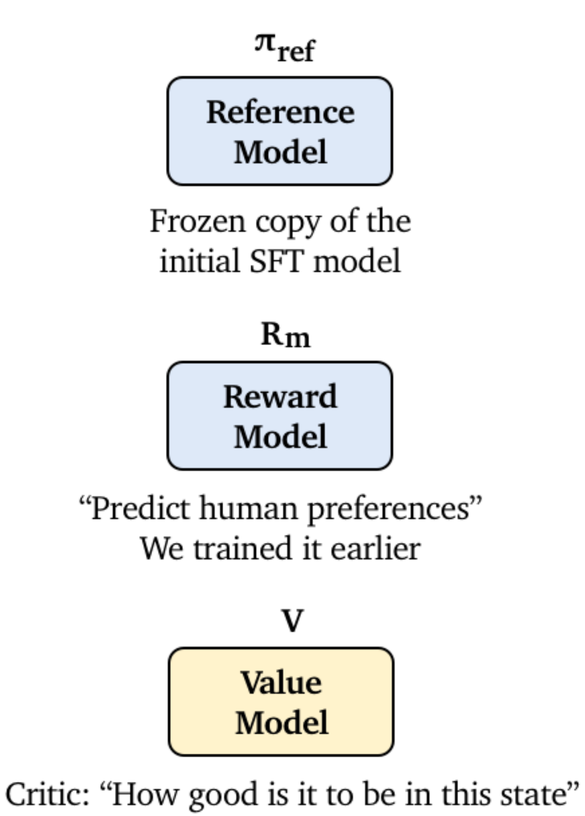
Reinforcement Learning from Human Feedback

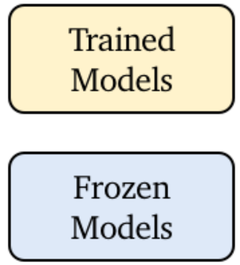
- Step 2: Align the LLM using PPO

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

- Step 2: Align the LLM using PPO

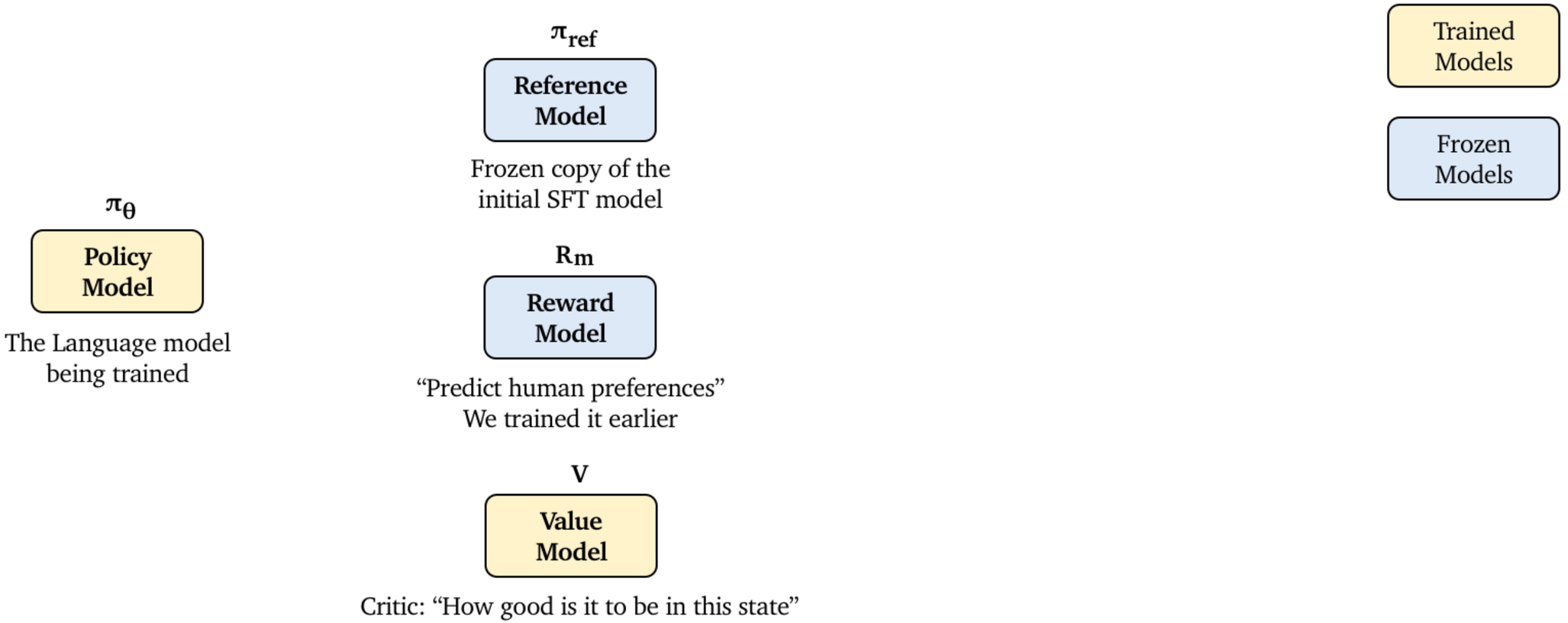




DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

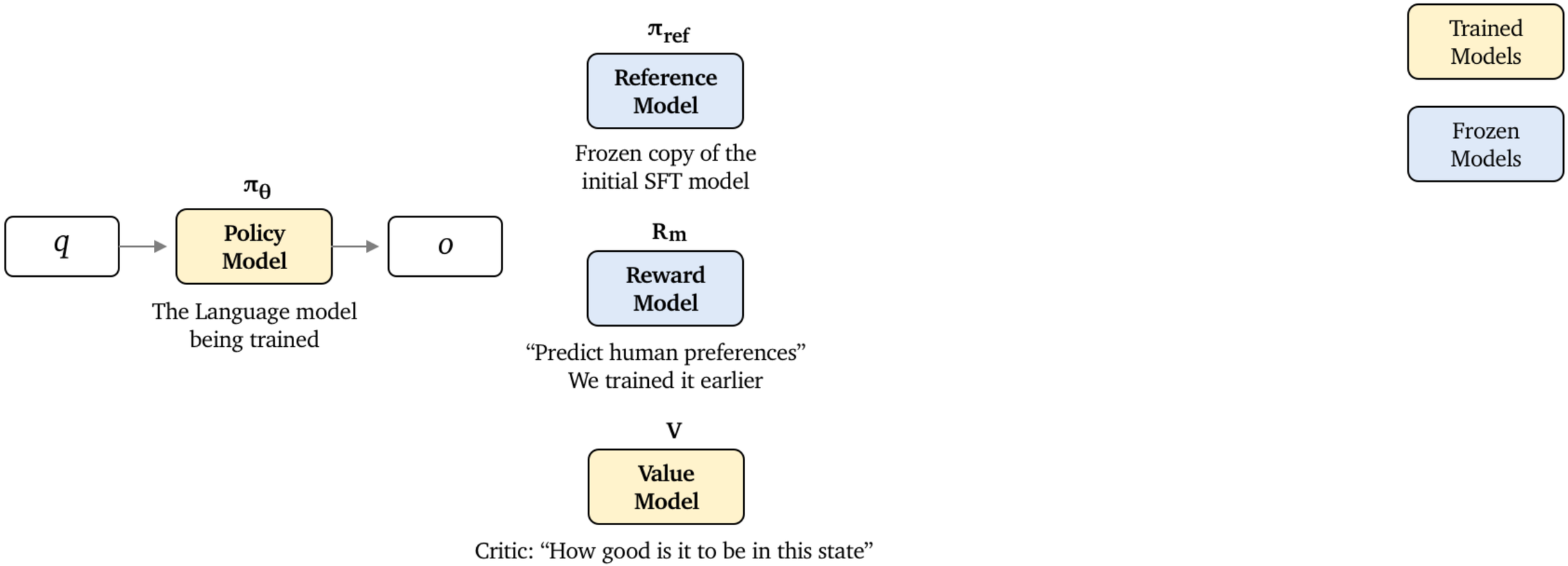
- Step 2: Align the LLM using PPO



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

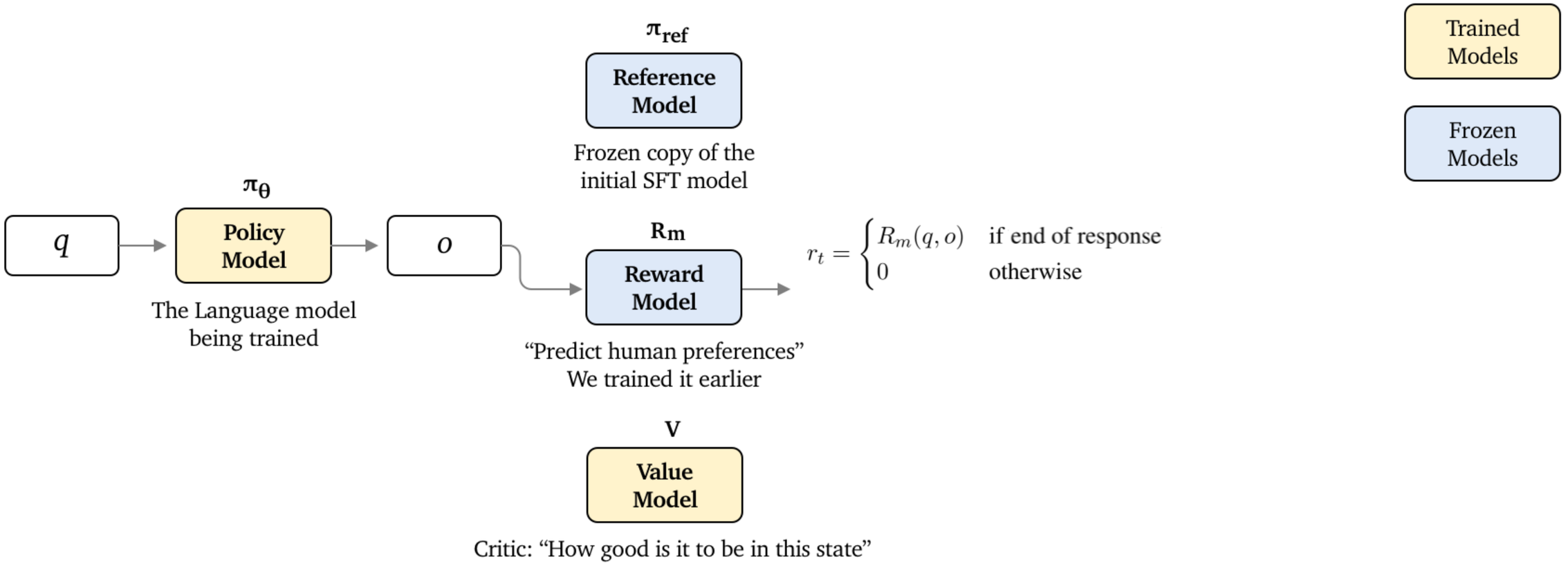
Reinforcement Learning from Human Feedback

- Step 2: Align the LLM using PPO



Reinforcement Learning from Human Feedback

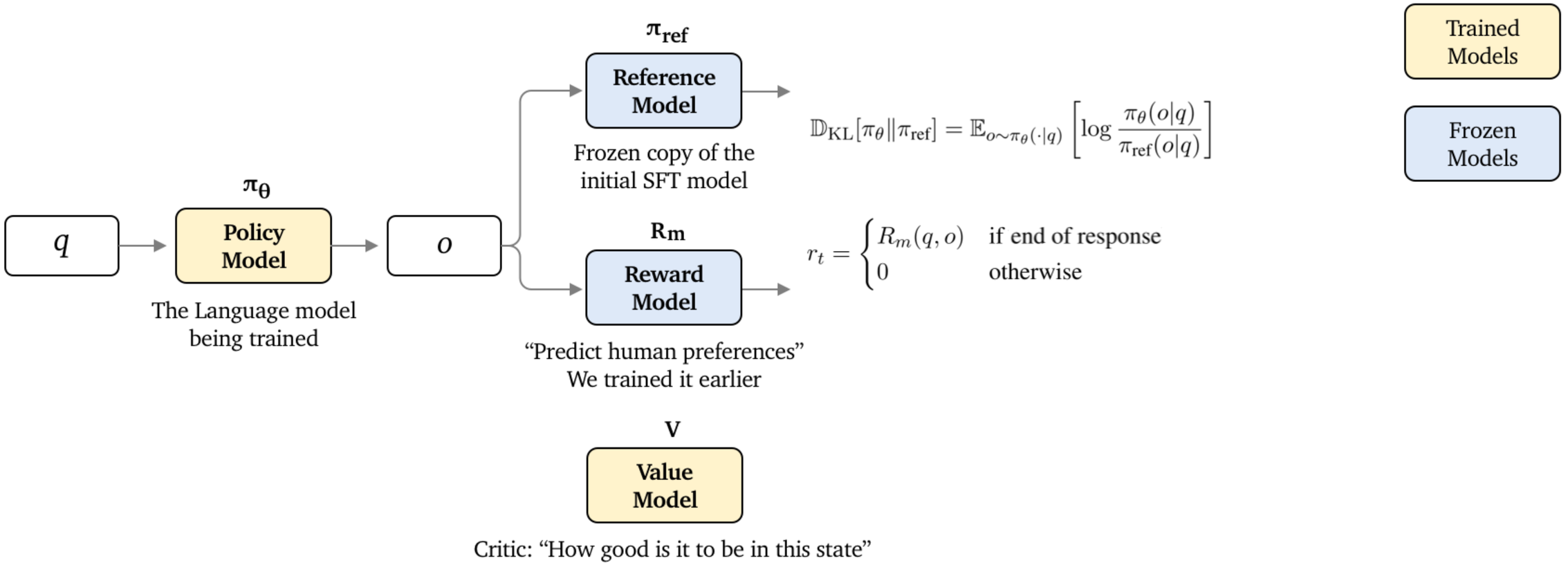
- Step 2: Align the LLM using PPO



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

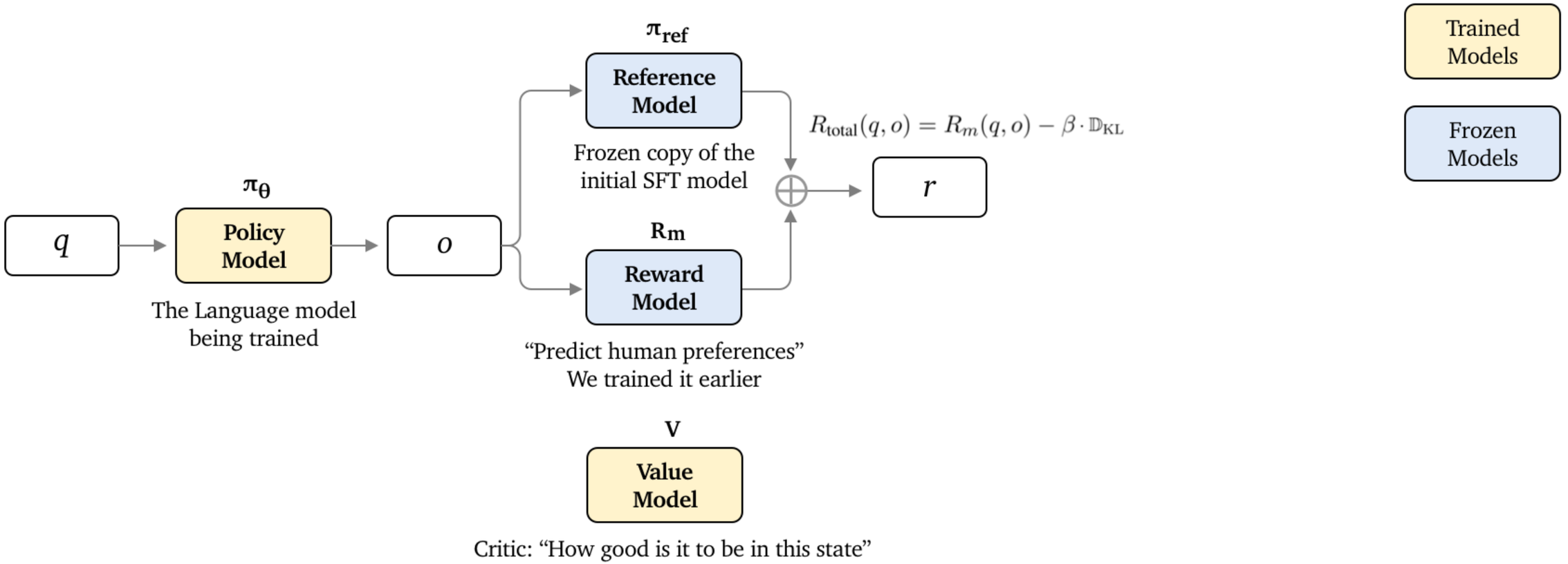
- Step 2: Align the LLM using PPO



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

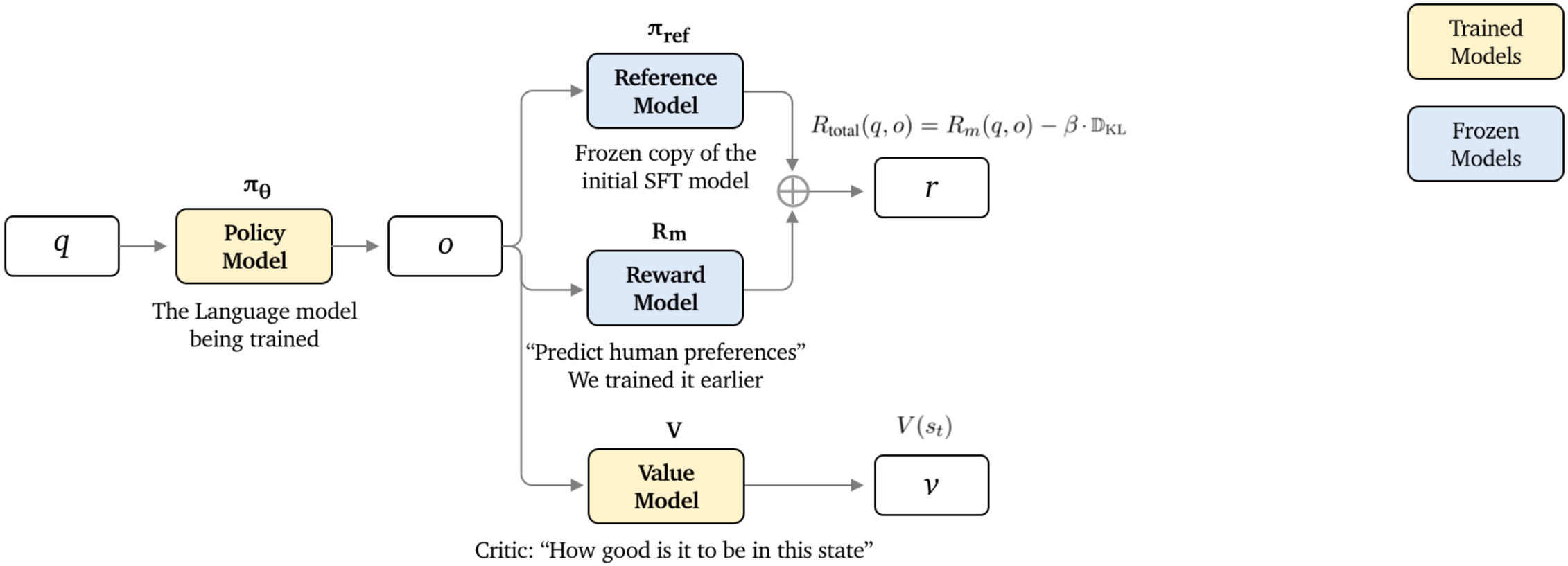
- Step 2: Align the LLM using PPO



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

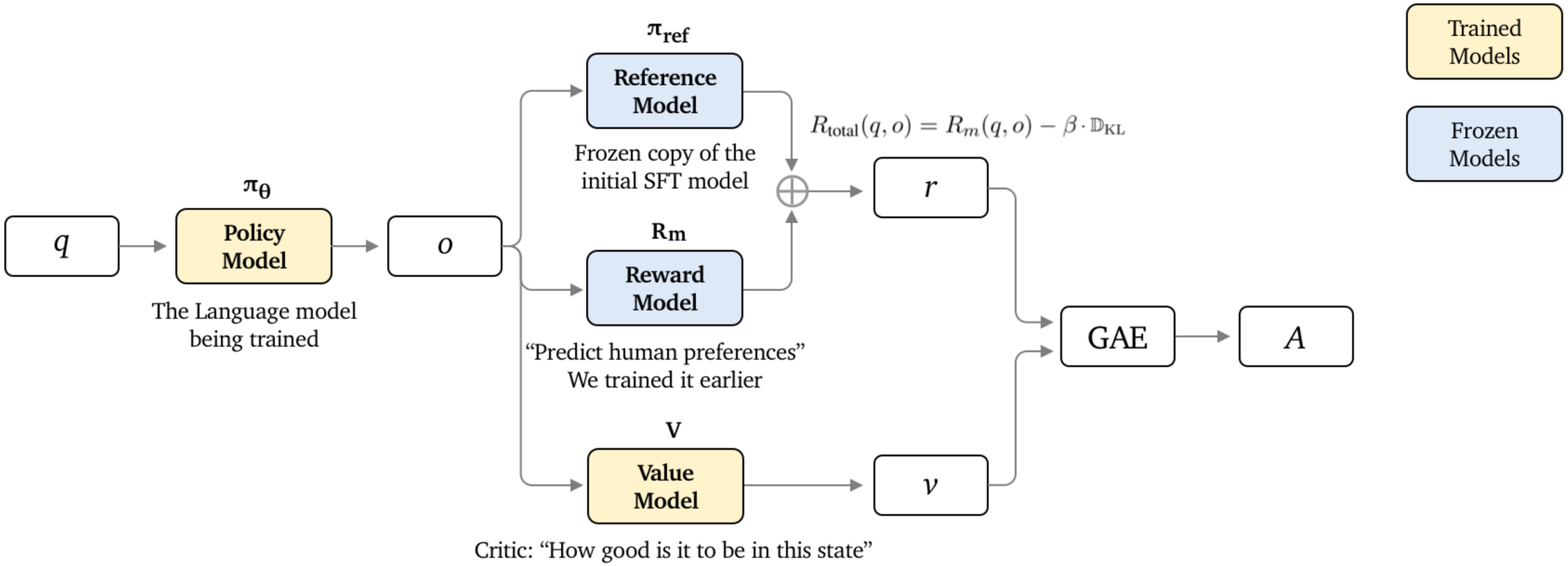
- Step 2: Align the LLM using PPO



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

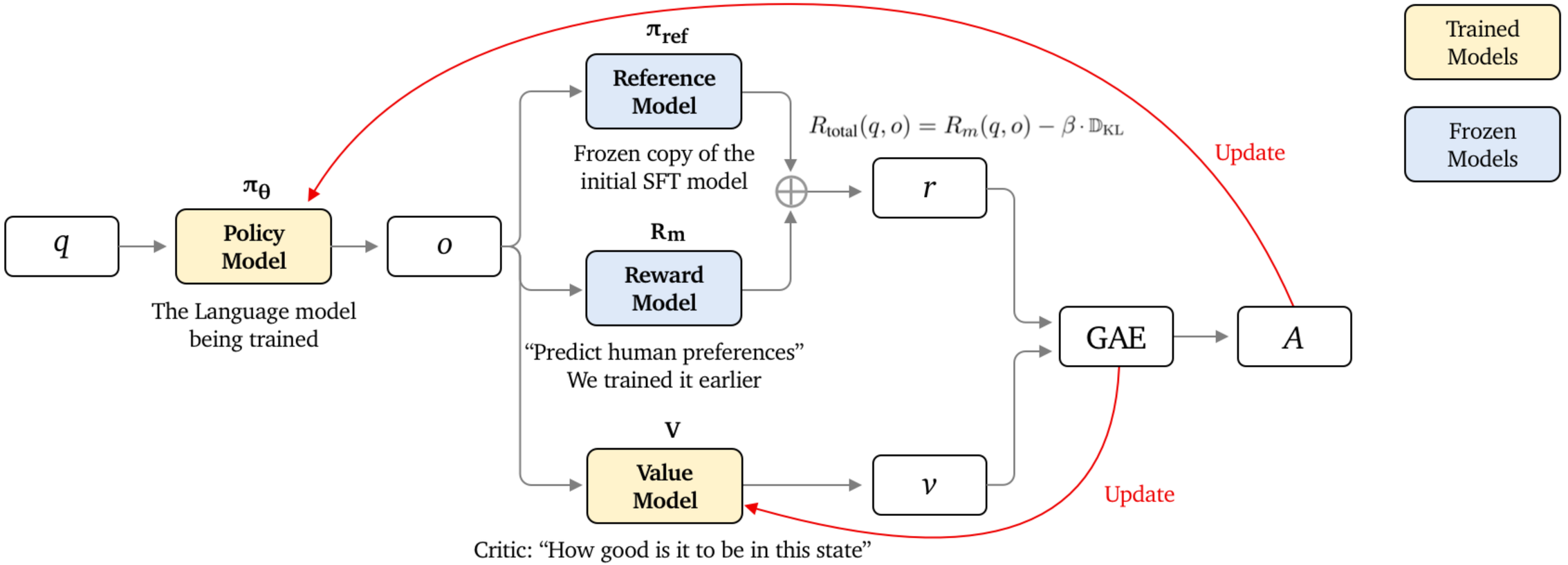
- Step 2: Align the LLM using PPO



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

- Step 2: Align the LLM using PPO



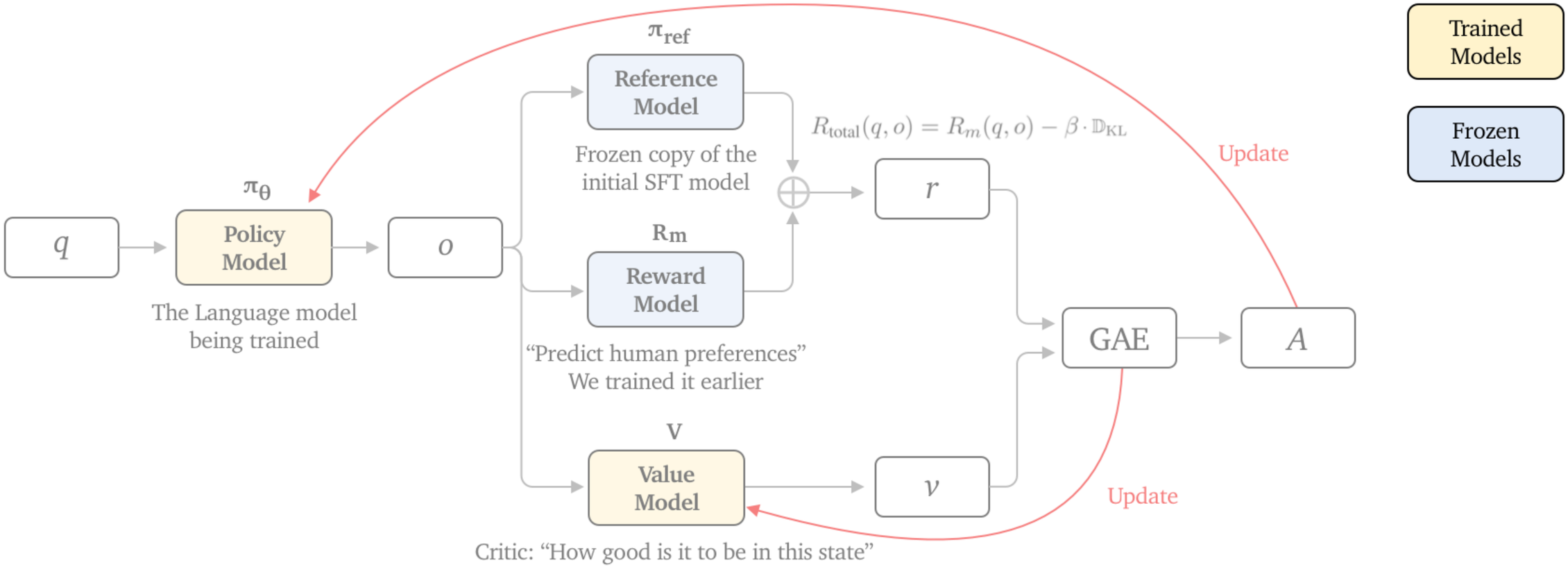
DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

- Step 2: Align the LLM using GRPO **“DeepSeek Moment”**

Reinforcement Learning from Human Feedback

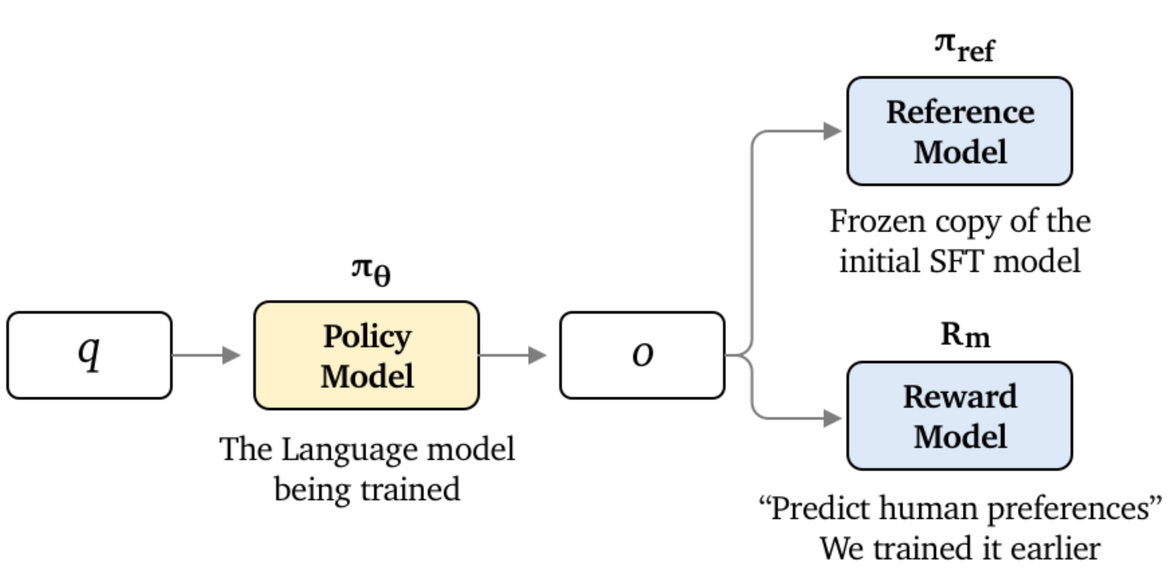
- Step 2: Align the LLM using GRPO **“DeepSeek Moment”**

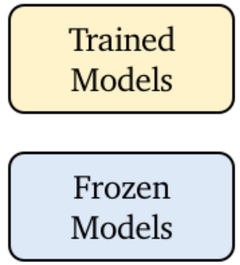


DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

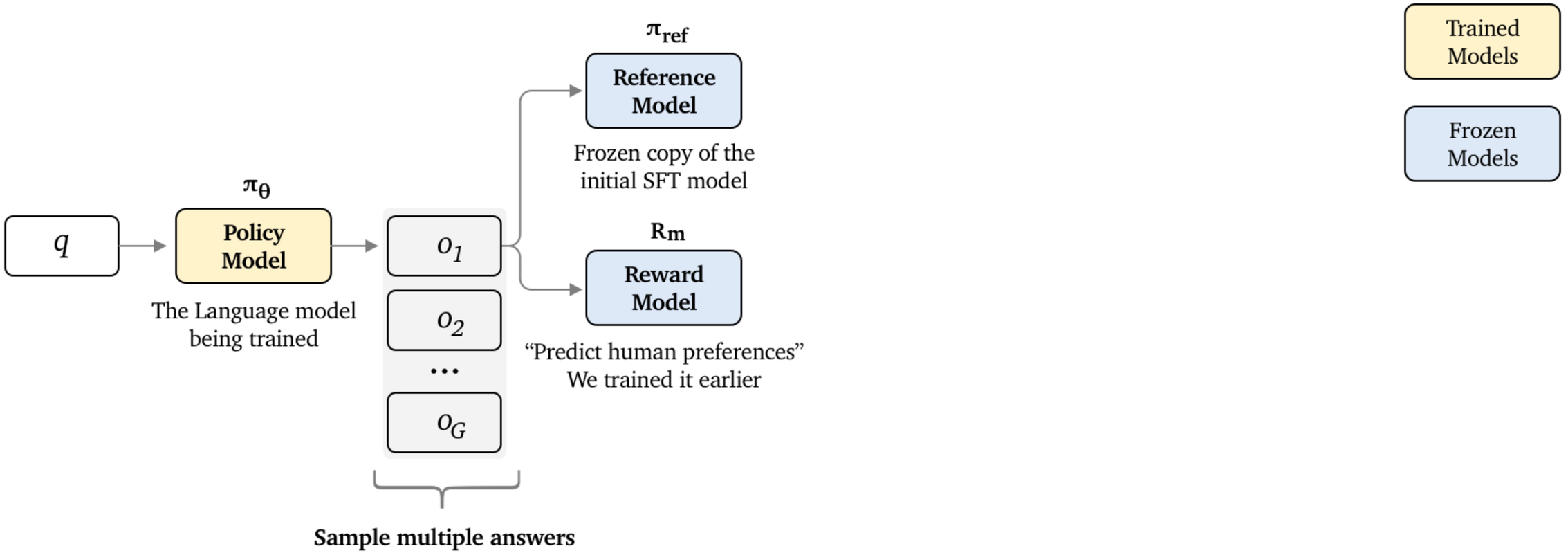
- Step 2: Align the LLM using GRPO **“DeepSeek Moment”**





Reinforcement Learning from Human Feedback

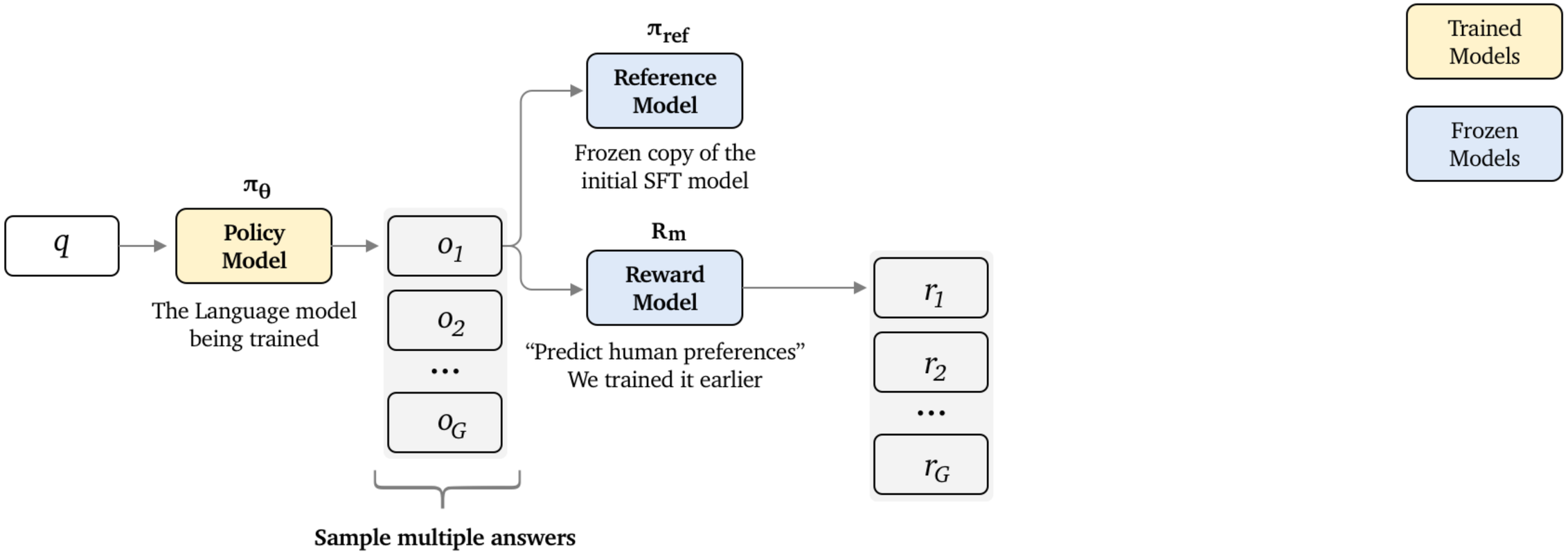
- Step 2: Align the LLM using GRPO **“DeepSeek Moment”**



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

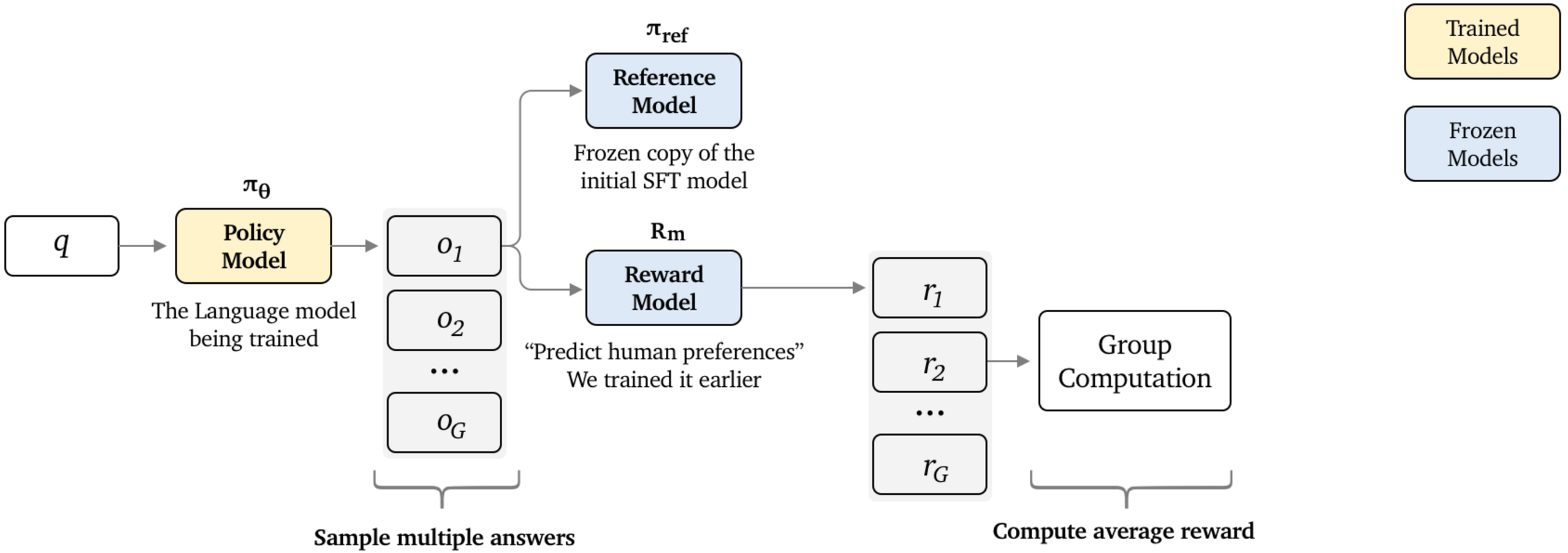
- Step 2: Align the LLM using GRPO **“DeepSeek Moment”**



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

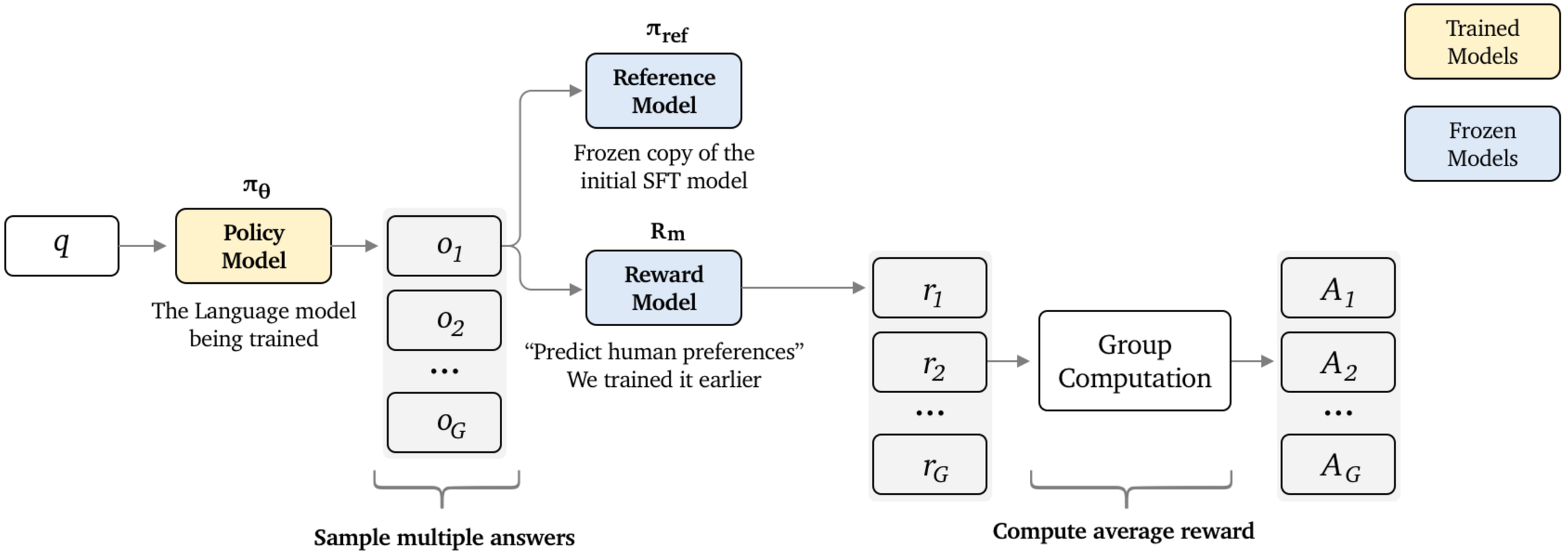
- Step 2: Align the LLM using GRPO **“DeepSeek Moment”**



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

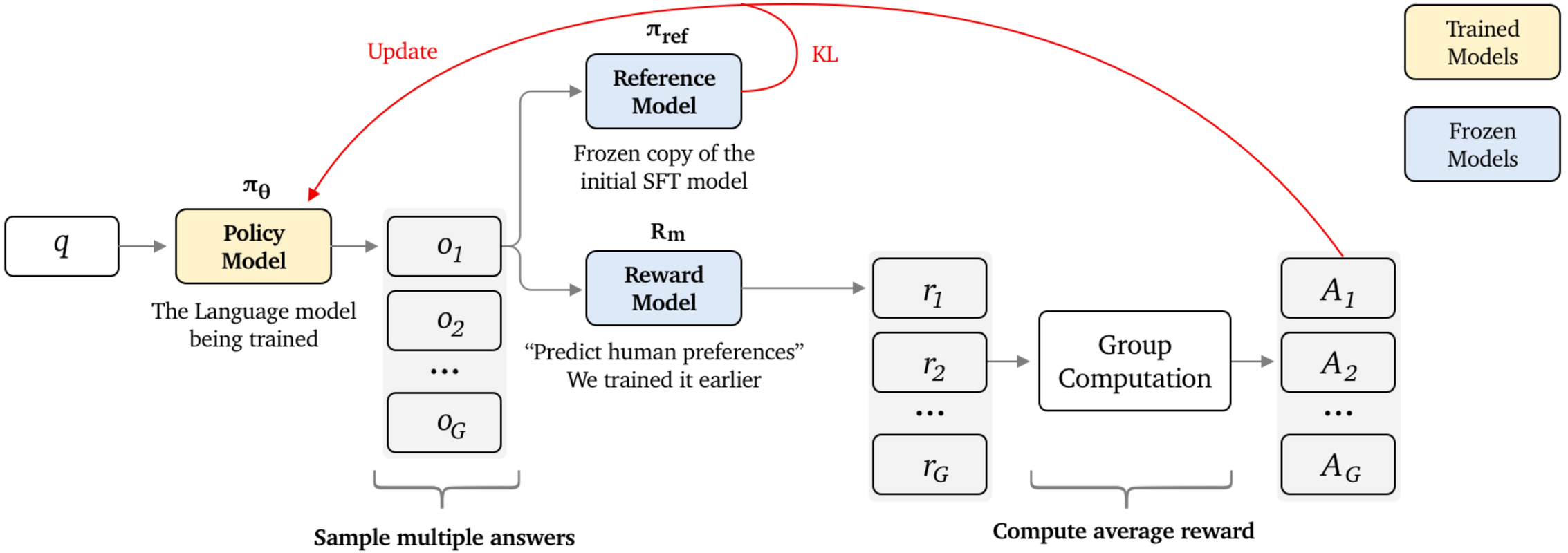
- Step 2: Align the LLM using GRPO **“DeepSeek Moment”**



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Reinforcement Learning from Human Feedback

- Step 2: Align the LLM using GRPO **“DeepSeek Moment”**



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Model

Alignment with RLHF Summary

- Reward model is central to the process
- PPO: Need a value model to calculate advantage
- GRPO: No value model (less GPU memory), use group average for advantage

Test-time Compute

Test-time Compute

- Neural Scaling Laws: how does performance (test error) change with
  - ✦ Compute
  - ✦ Dataset size
  - ✦ Number of parameters

“Scaling Laws for Neural Language Models”, 2020
https://www.youtube.com/watch?v=5eqRuVp65eY

Test-time Compute

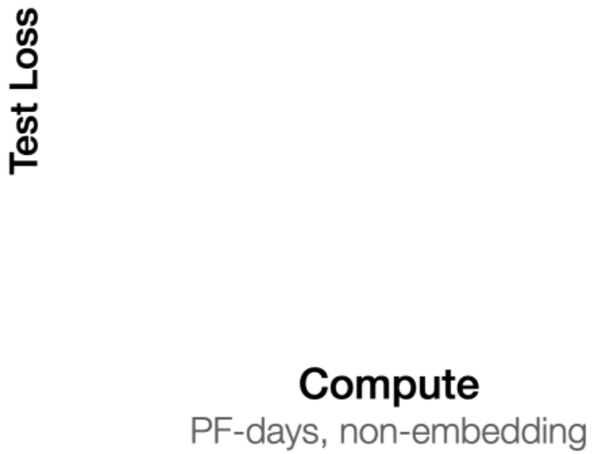
- Neural Scaling Laws: how does performance (test error) change with
  - ✦ Compute
  - ✦ Dataset size
  - ✦ Number of parameters

- Finding: “Performance has a power-law relationship with each of the three scale factors”

“Scaling Laws for Neural Language Models”, 2020
https://www.youtube.com/watch?v=5eqRuVp65eY

Test-time Compute

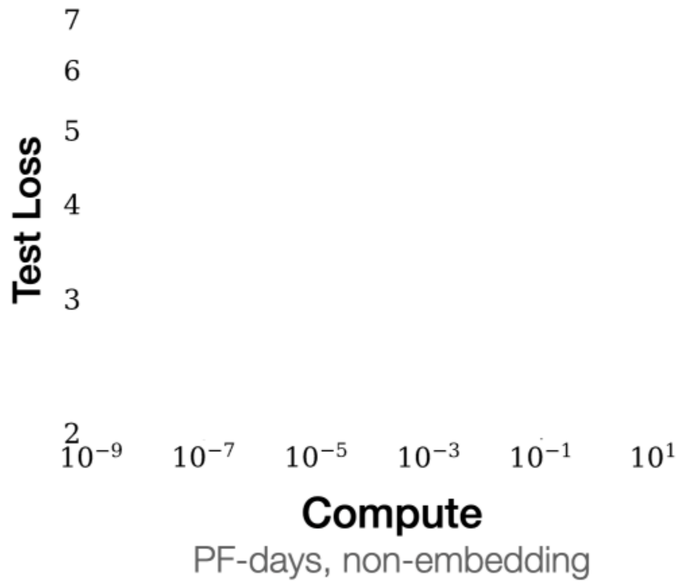
- “Performance has a power-law relationship with each of the three scale factors”

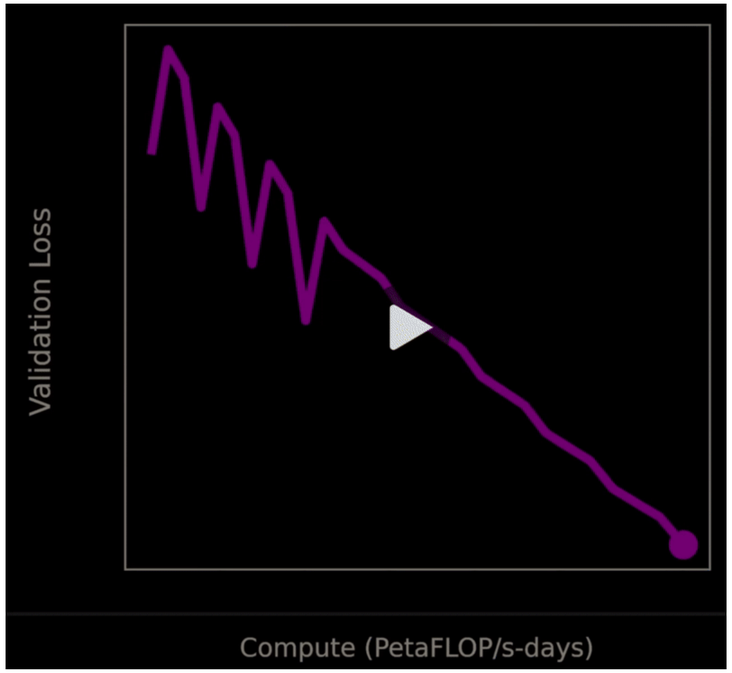


“Scaling Laws for Neural Language Models”, 2020
https://www.youtube.com/watch?v=5eqRuVp65eY

Test-time Compute

- “Performance has a power-law relationship with each of the three scale factors”

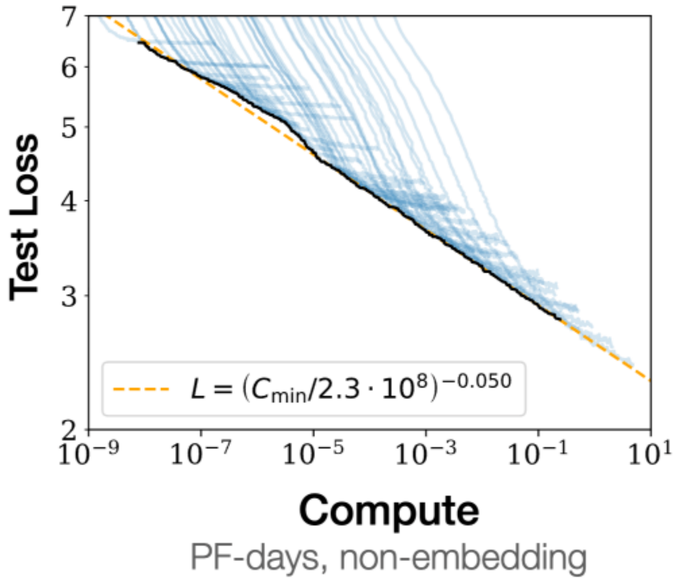




“Scaling Laws for Neural Language Models”, 2020
https://www.youtube.com/watch?v=5eqRuVp65eY

Test-time Compute

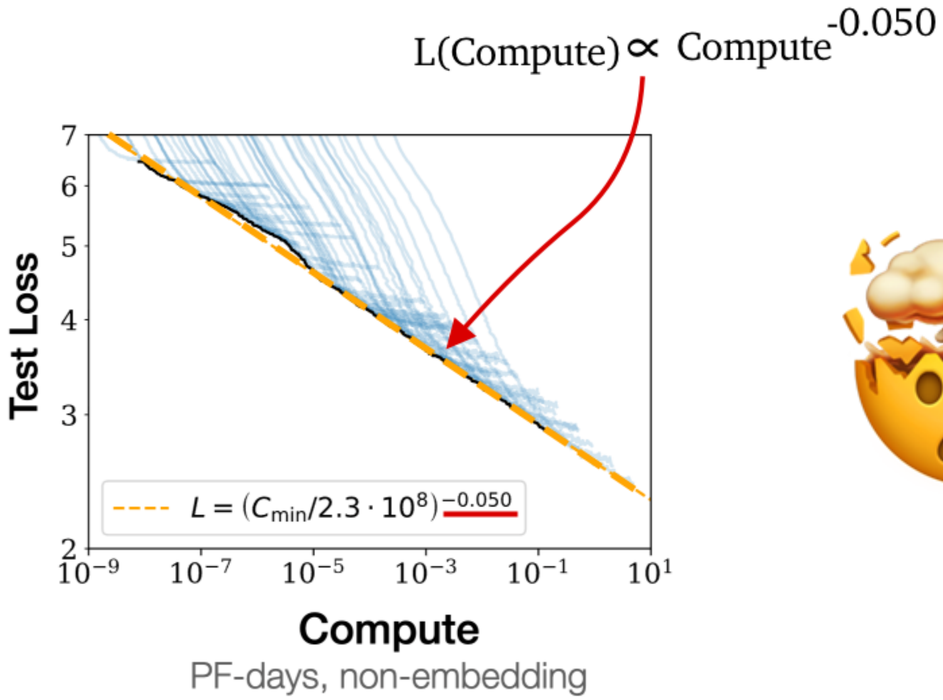
- “Performance has a power-law relationship with each of the three scale factors”



“Scaling Laws for Neural Language Models”, 2020
https://www.youtube.com/watch?v=5eqRuVp65eY

Test-time Compute

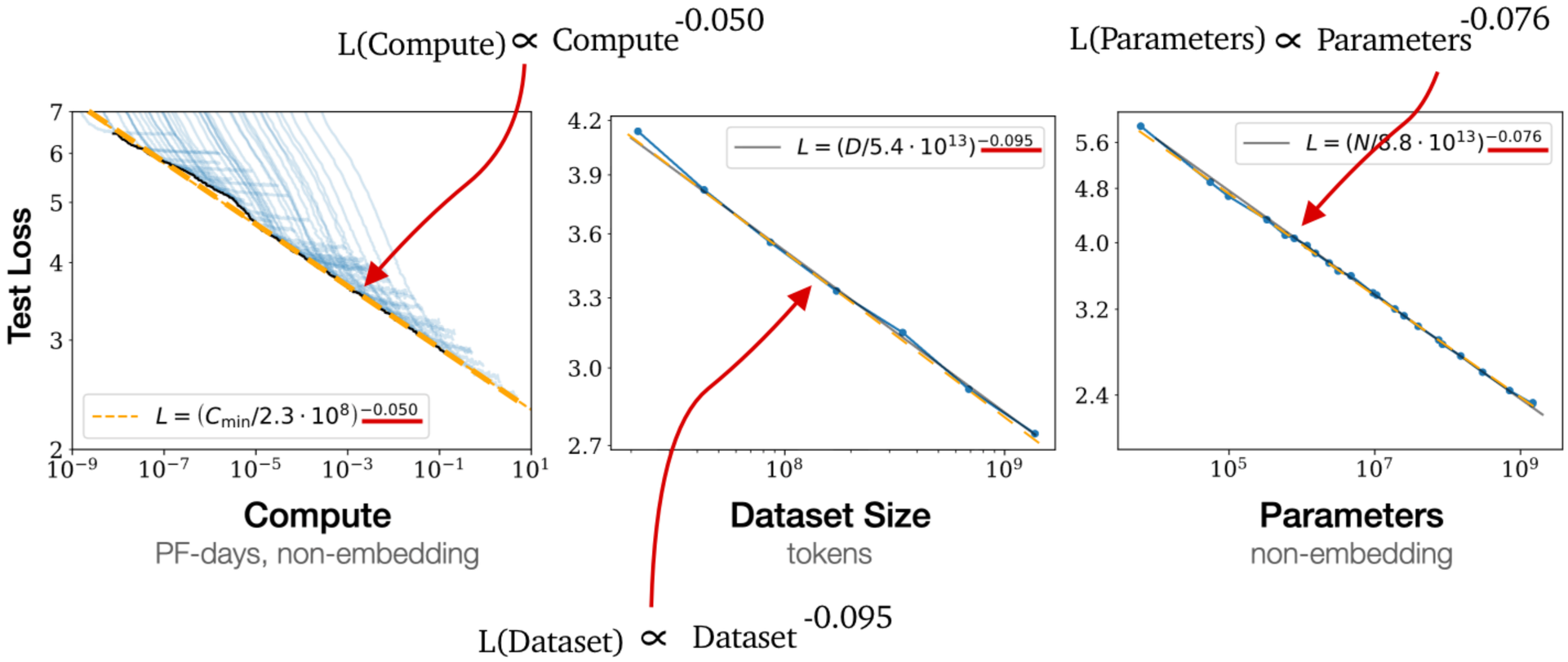
- “Performance has a power-law relationship with each of the three scale factors”



“Scaling Laws for Neural Language Models”, 2020
https://www.youtube.com/watch?v=5eqRuVp65eY

Test-time Compute

- “Performance has a power-law relationship with each of the three scale factors”



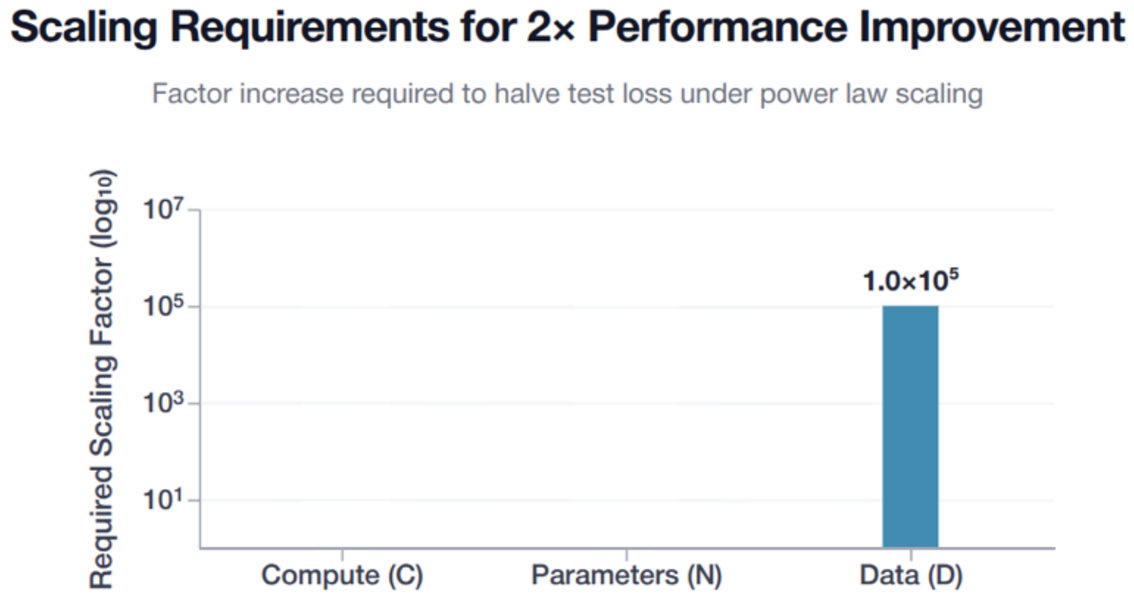
“Scaling Laws for Neural Language Models”, 2020
https://www.youtube.com/watch?v=5eqRuVp65eY

Test-time Compute

Computation at Scale > Algorithmic Breakthrough *

Test-time Compute (~3 years later)

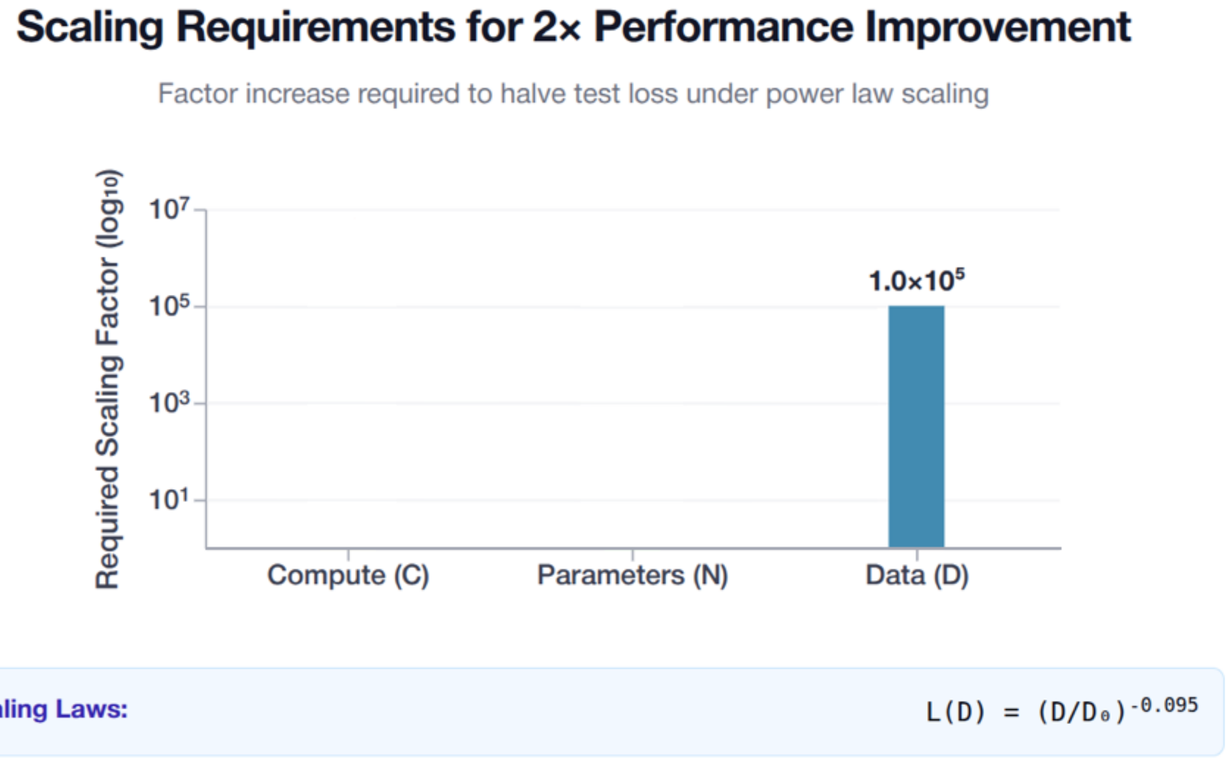
- But… napkin math



Scaling Laws: $L(D) = (D/D_0)^{-0.095}$

Test-time Compute (~3 years later)

- But… napkin math
- Useful tokens from internet ~10T tokens
- GPT-4 and Llama 3.1 ~ 15T tokens
- Increase that by 10^5 = 1.5 Quintillion



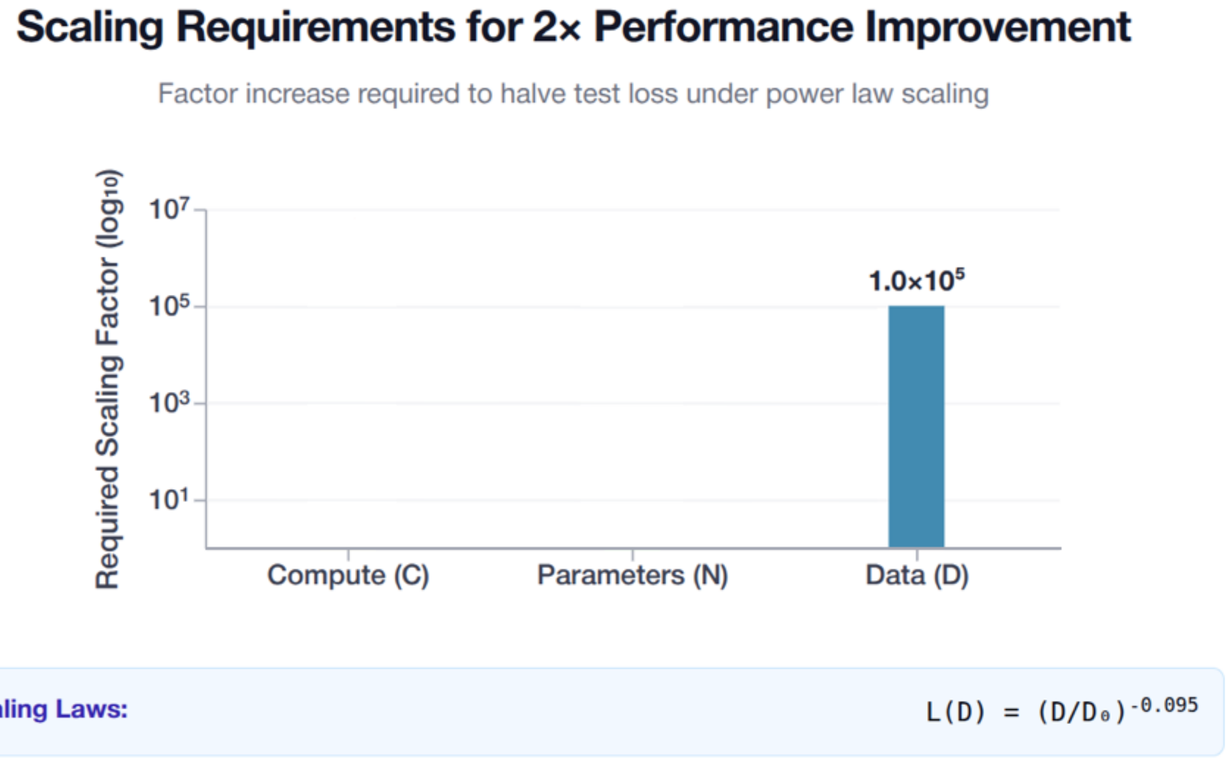
Ilya Sutskever: "Sequence to sequence learning with neural networks: what a decade", NeurIPS 2024
https://www.youtube.com/watch?v=1yvBqasHLZs

Test-time Compute (~3 years later)

- But… napkin math
- Useful tokens from internet ~10T tokens
- GPT-4 and Llama 3.1 ~ 15T tokens
- Increase that by 10^5 = 1.5 Quintillion

“the data is not growing,
we have but one internet”

“we have achieved peak data,
there will be no more”

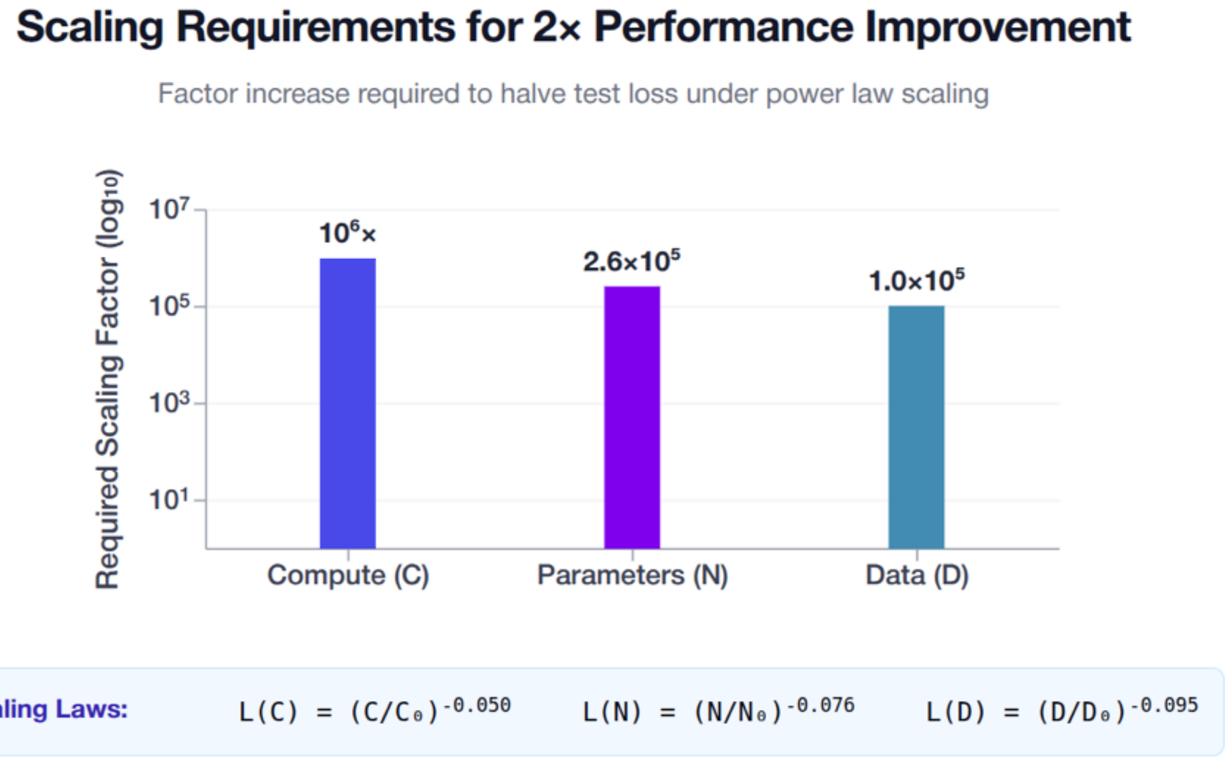


Ilya Sutskever: "Sequence to sequence learning with neural networks: what a decade”, NeurIPS 2024
https://www.youtube.com/watch?v=1yvBqasHLZs

Test-time Compute (~3 years later)

- But… napkin math
- Scaling other dimensions, diminishing returns

“Intelligence scales with the log of compute”



Ilya Sutskever: "Sequence to sequence learning with neural networks: what a decade”, NeurIPS 2024
https://www.youtube.com/watch?v=1yvBqasHLZs

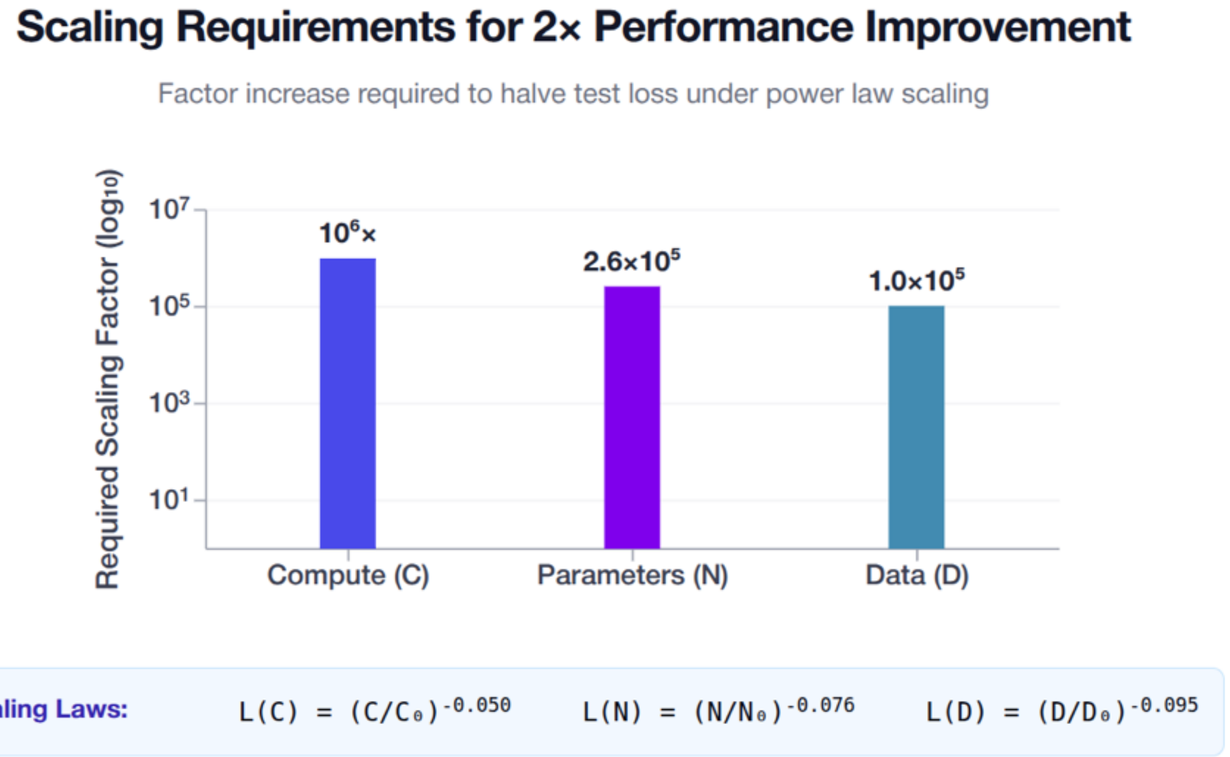
Test-time Compute (~3 years later)

- But… napkin math
- Scaling other dimensions, diminishing returns

“Intelligence scales with the log of compute”

“The age of pre-training scaling has ended”

“Deep learning has hit a wall”



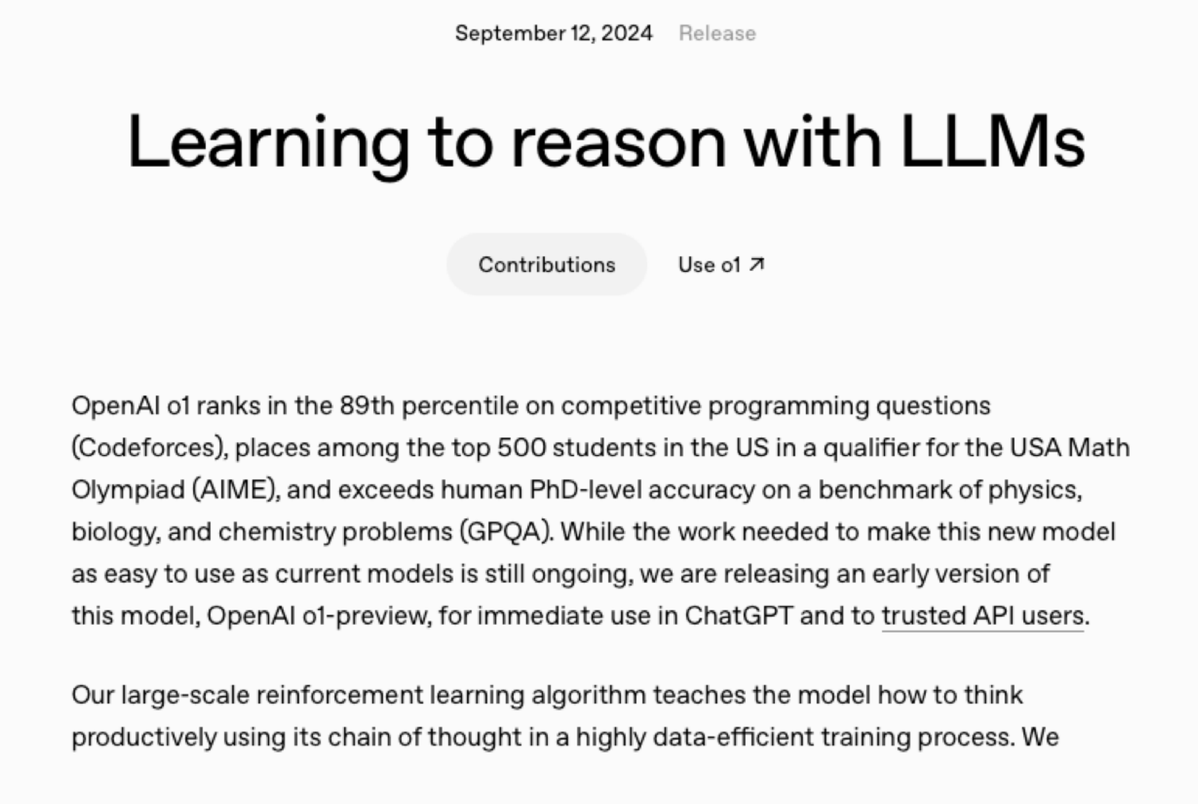
Ilya Sutskever: "Sequence to sequence learning with neural networks: what a decade”, NeurIPS 2024
https://www.youtube.com/watch?v=1yvBqasHLZs

Test-time Compute

- Until… RL to the rescue

Test-time Compute

- OpenAI o1, a new kind of model: “*Spend more time thinking before they respond*”
- *“Can reason through complex tasks and solve harder problems than previous models”*



https://openai.com/index/introducing-openai-o1-preview/

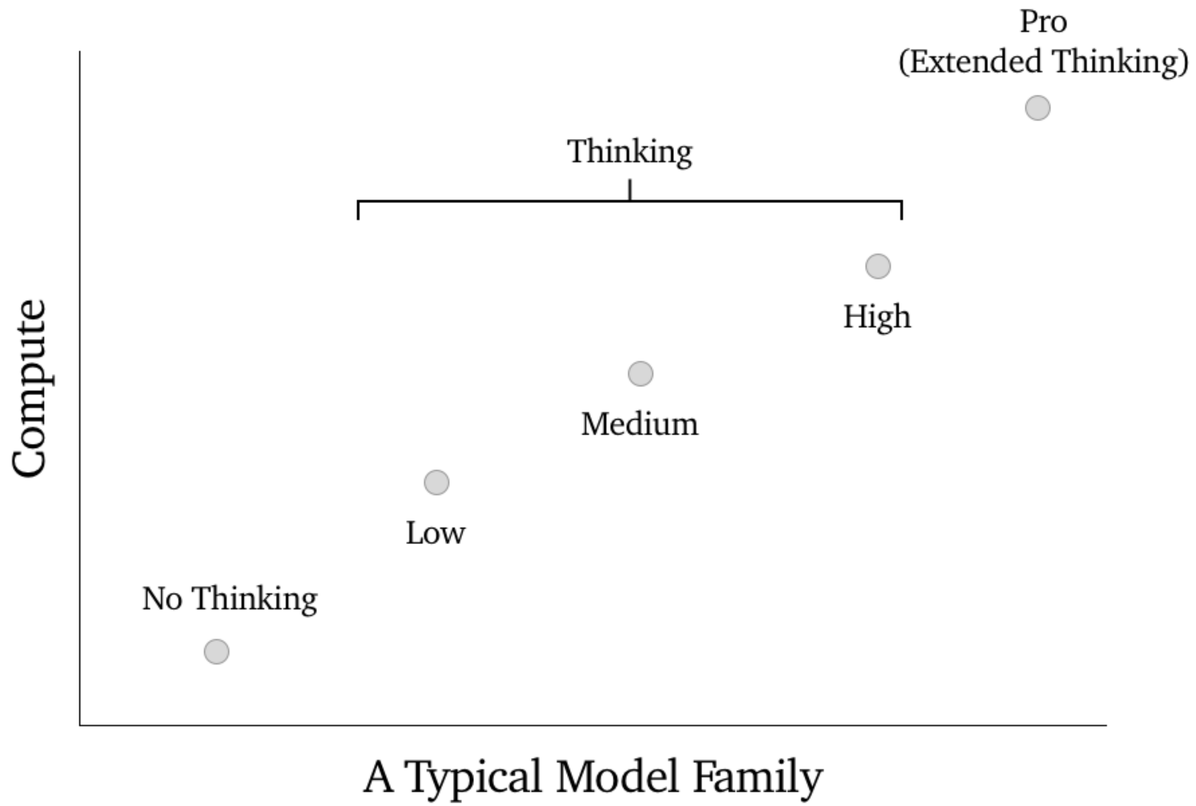
Test-time Compute

- OpenAI o1, a new kind of model: “*Spend more time thinking before they respond*”
- *“Can reason through complex tasks and solve harder problems than previous models”*
- A new scaling law:

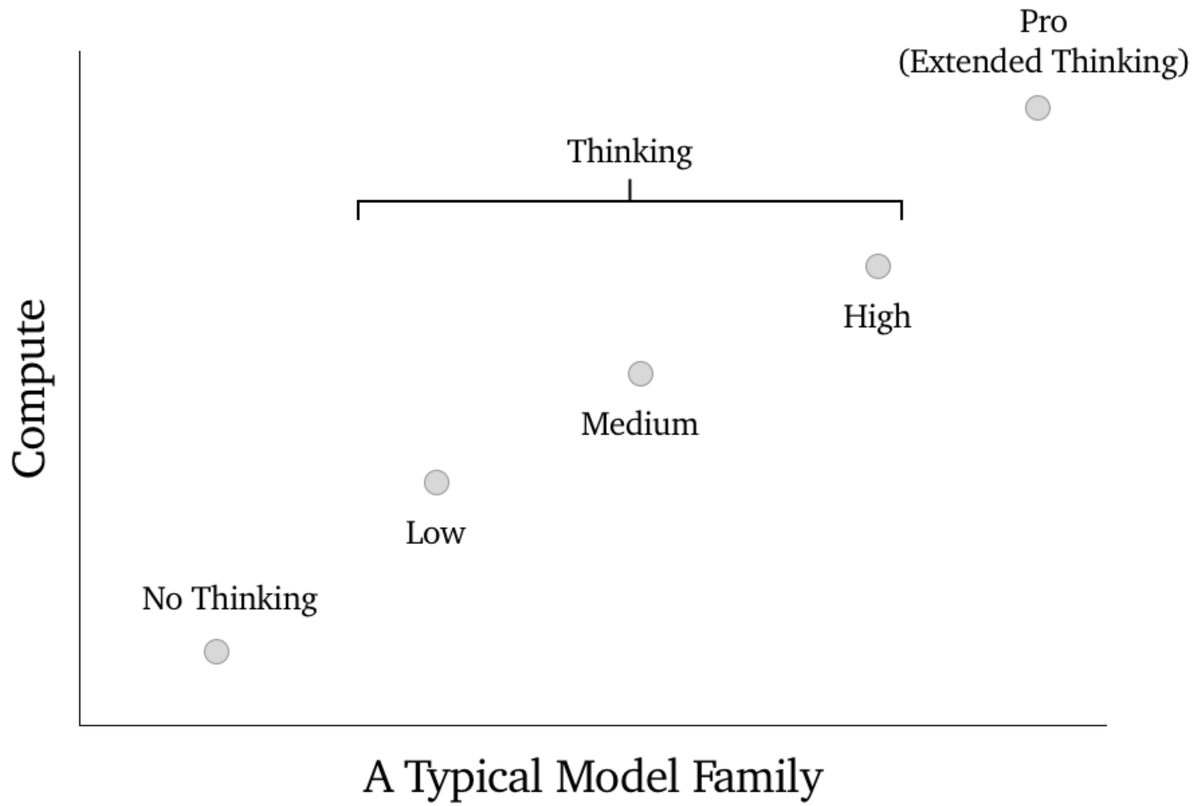
“**Performance scales with the amount of compute applied during generation of an answer**”

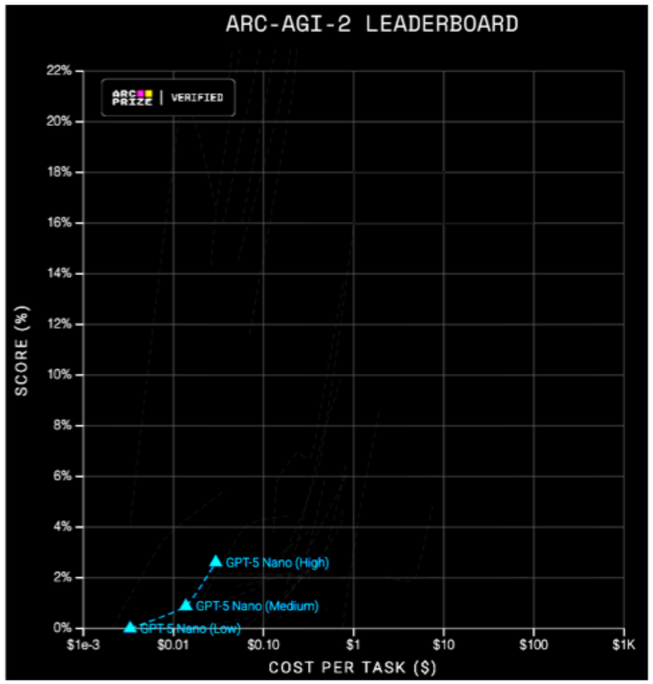
https://openai.com/index/introducing-openai-o1-preview/

Test-time Compute



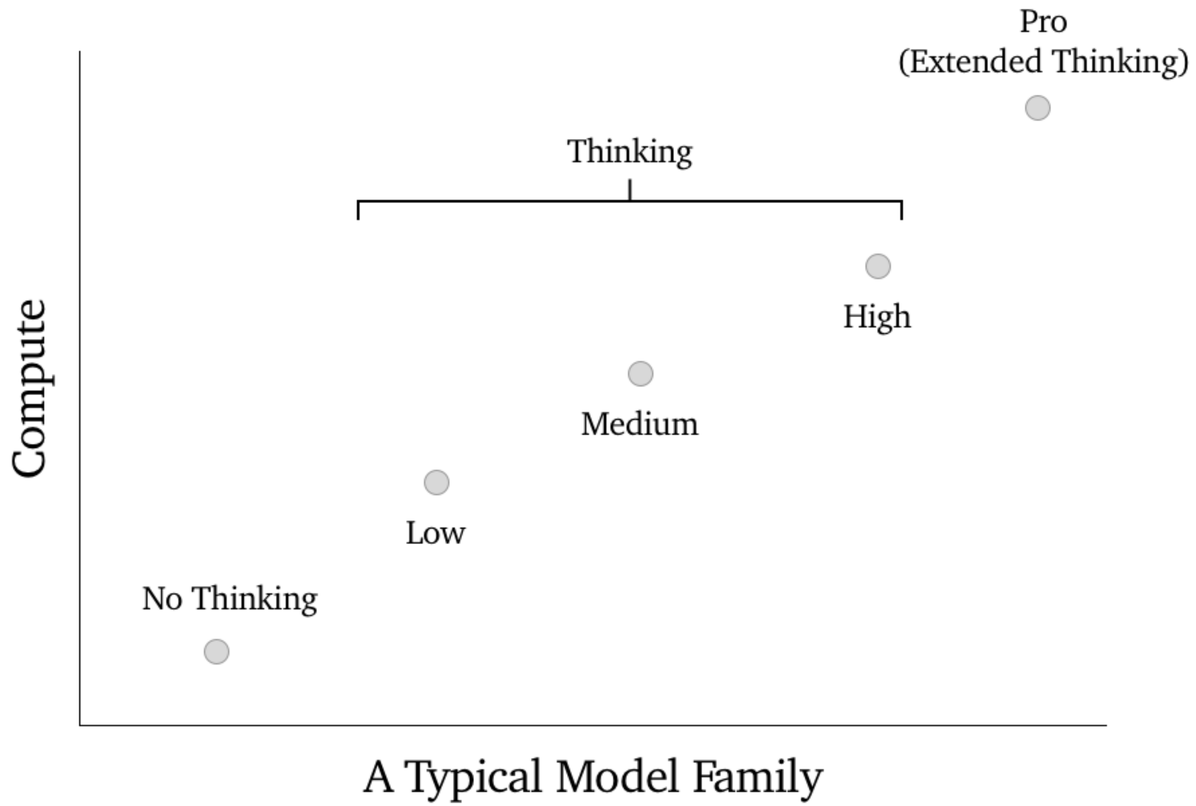
Test-time Compute

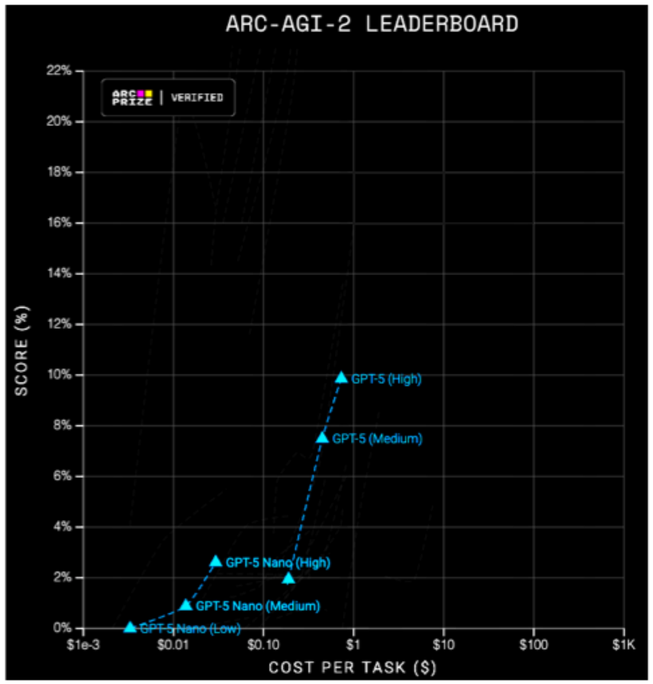




https://arcprize.org/leaderboard

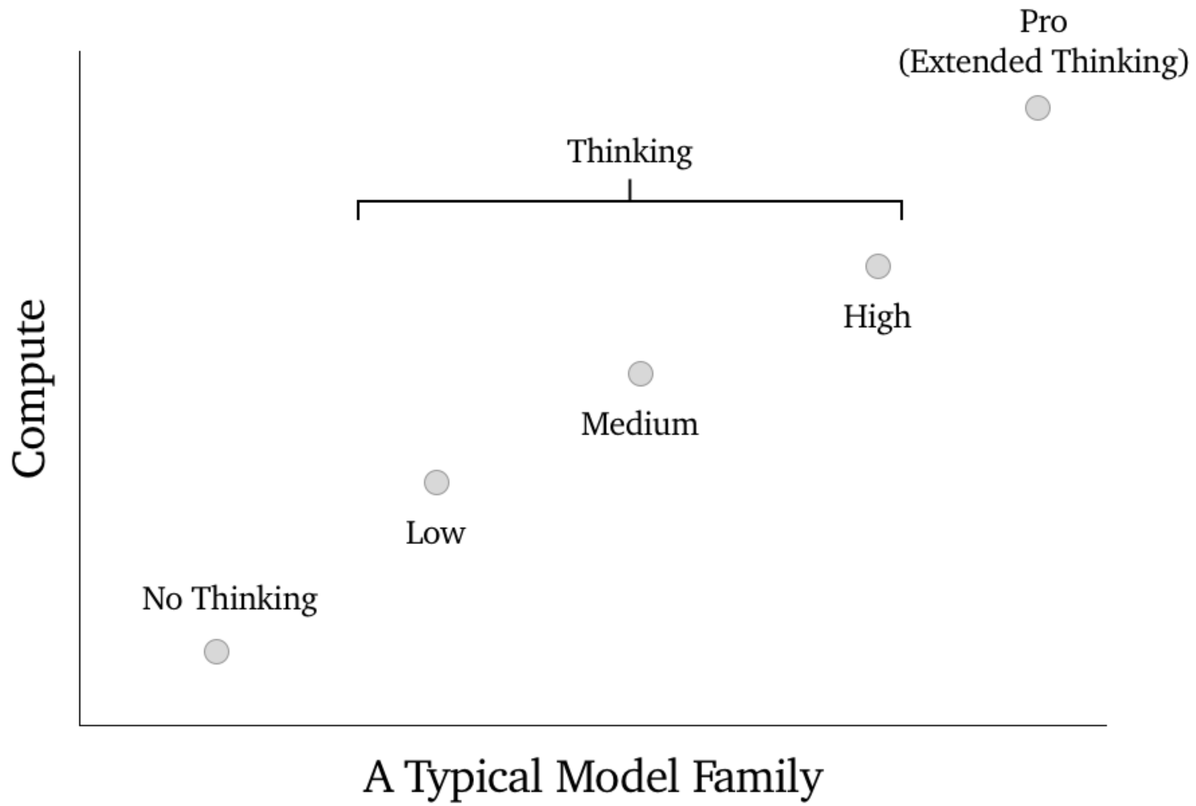
Test-time Compute

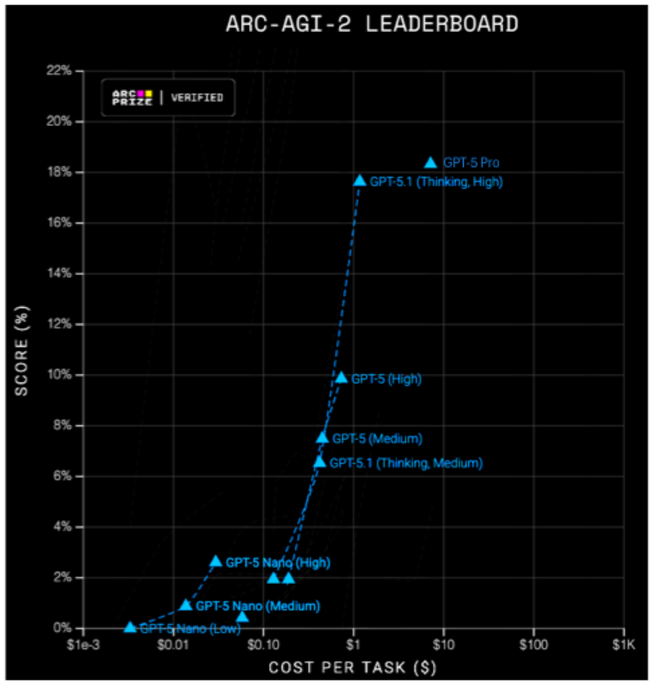




https://arcprize.org/leaderboard

Test-time Compute





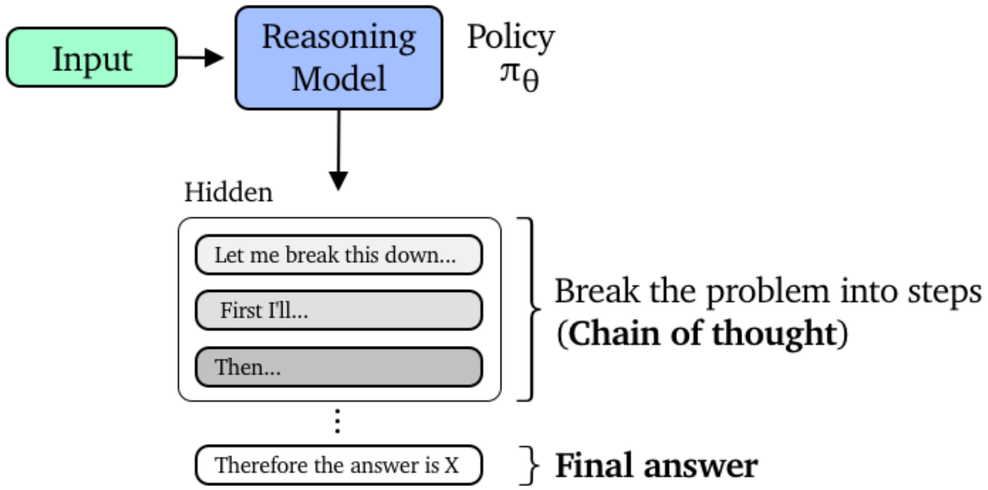
https://arcprize.org/leaderboard

Test-time Compute

- But where is the RL?

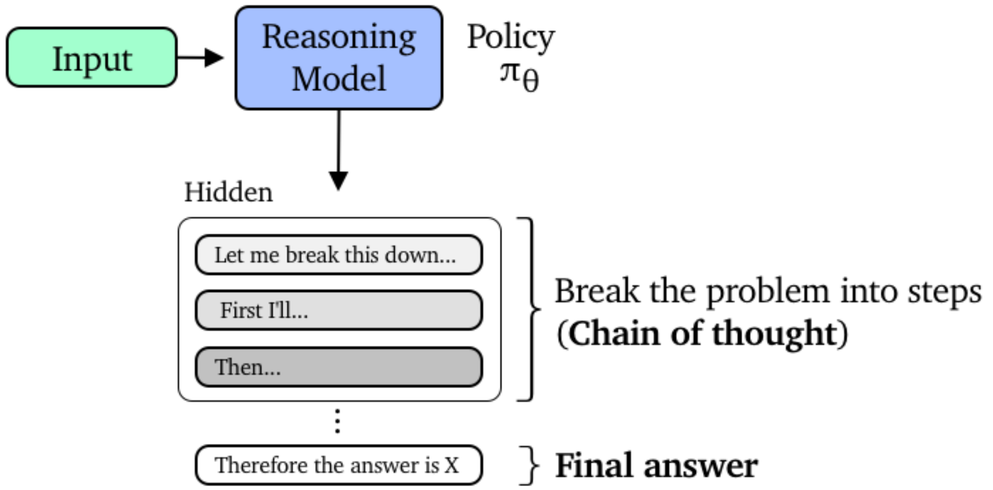
Test-time Compute

- But where is the RL?

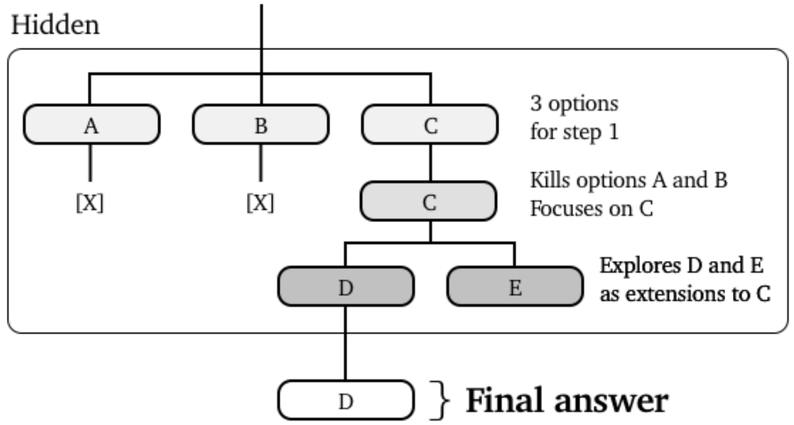


Test-time Compute

- But where is the RL?



Branches into multiple possibilities (**Tree of thought**)

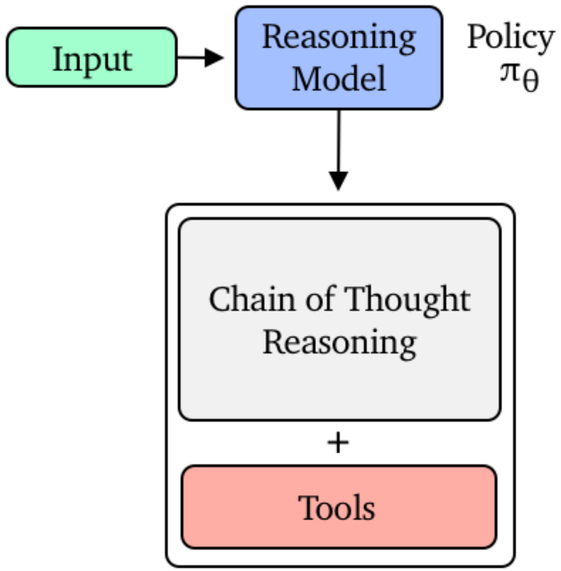


More complex traces (**Graph of thought**)

- Self-correction
- Merge
- Backtrack

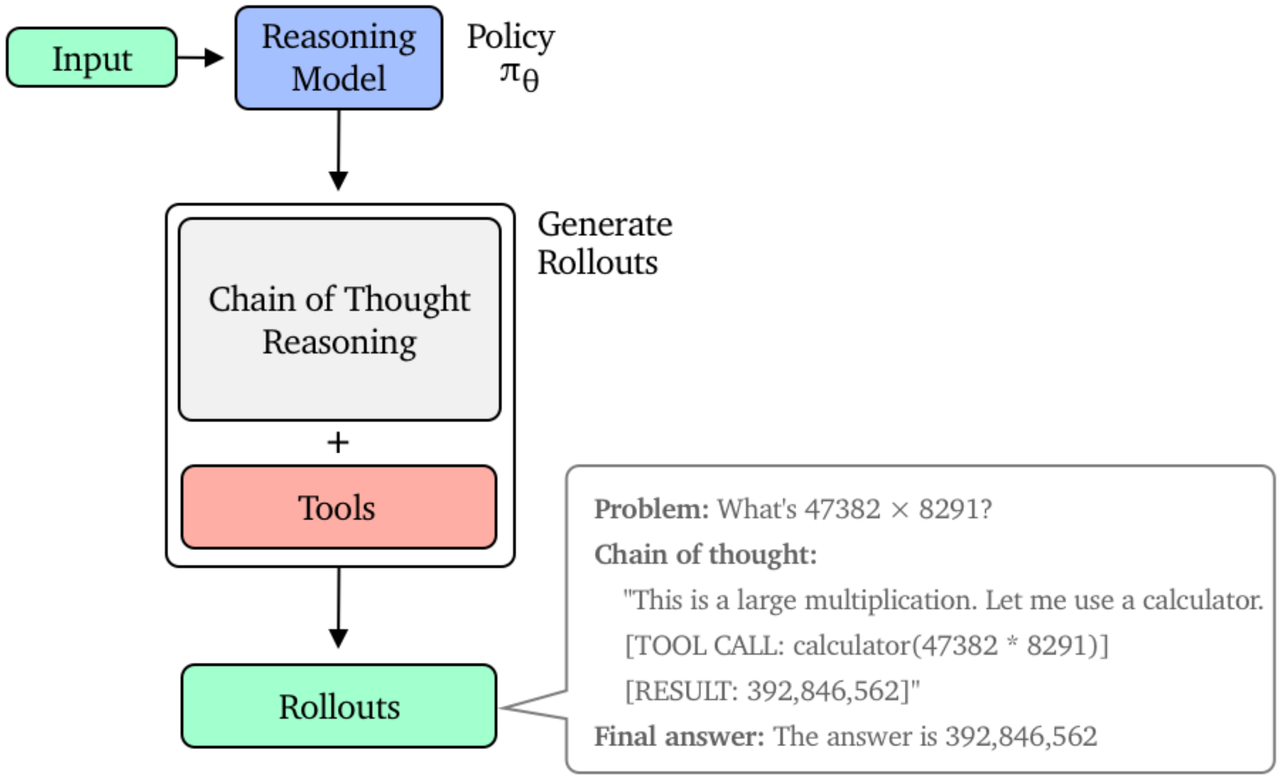
Test-time Compute

- But where is the RL?



Test-time Compute

- But where is the RL?

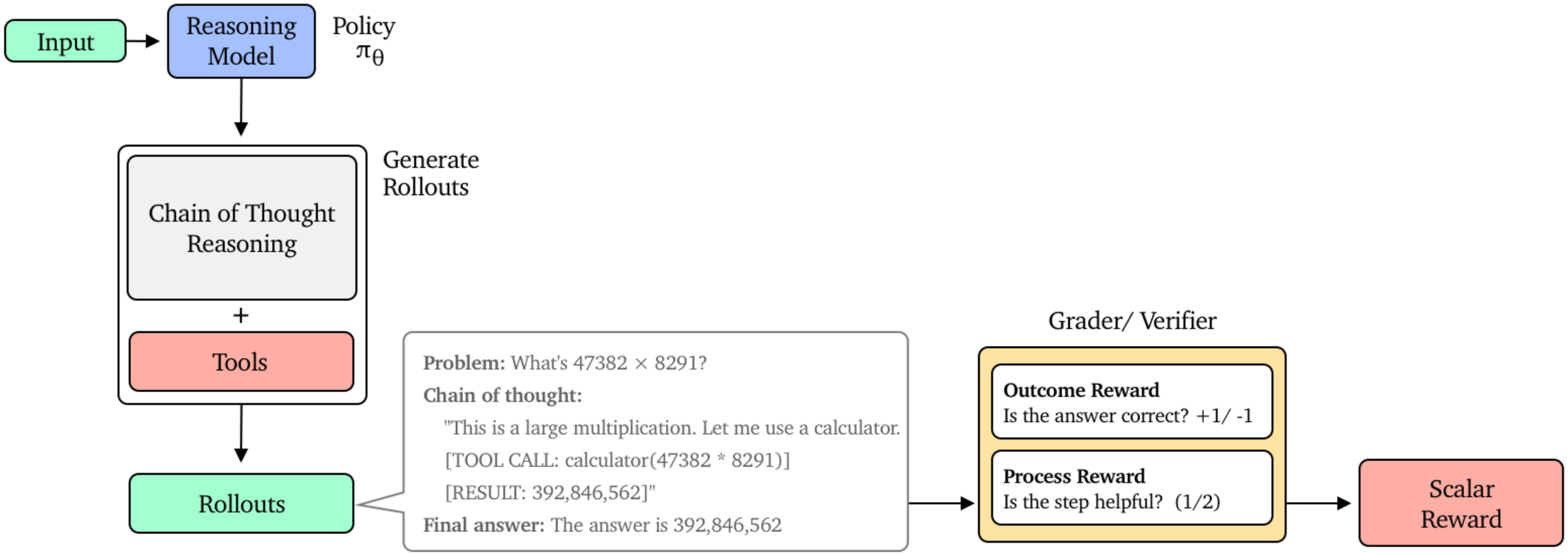


Test-time Compute

- But where is the RL?

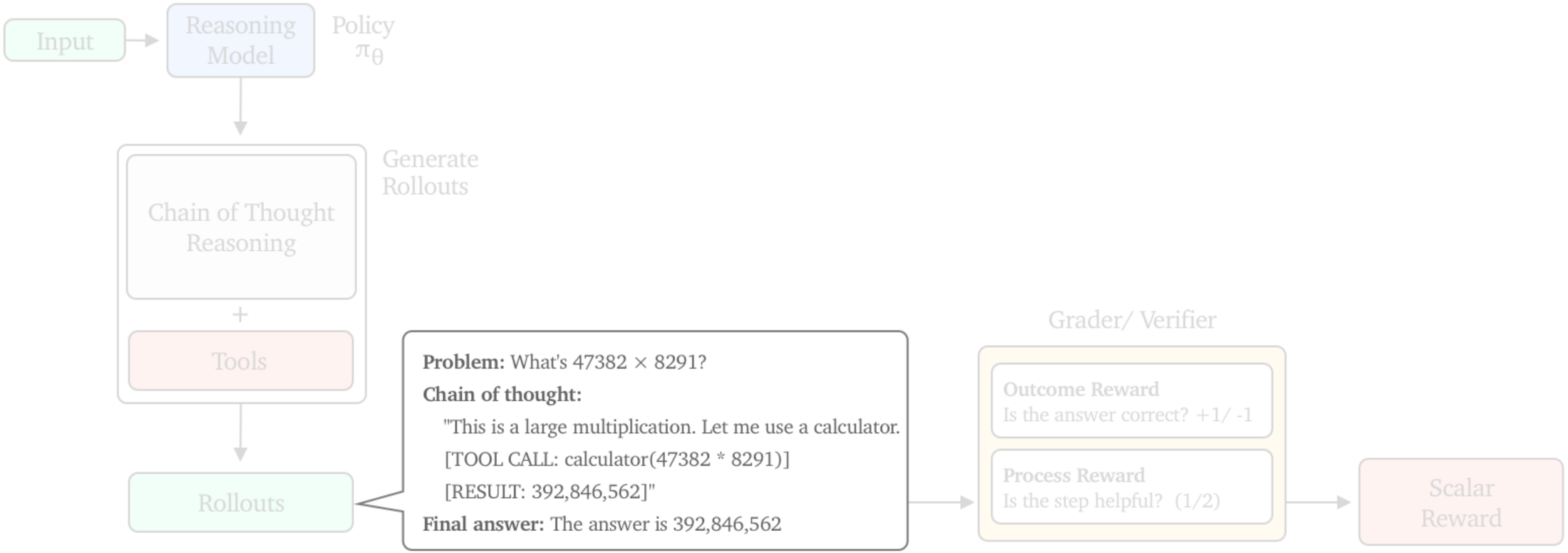
Test-time Compute

- But where is the RL?



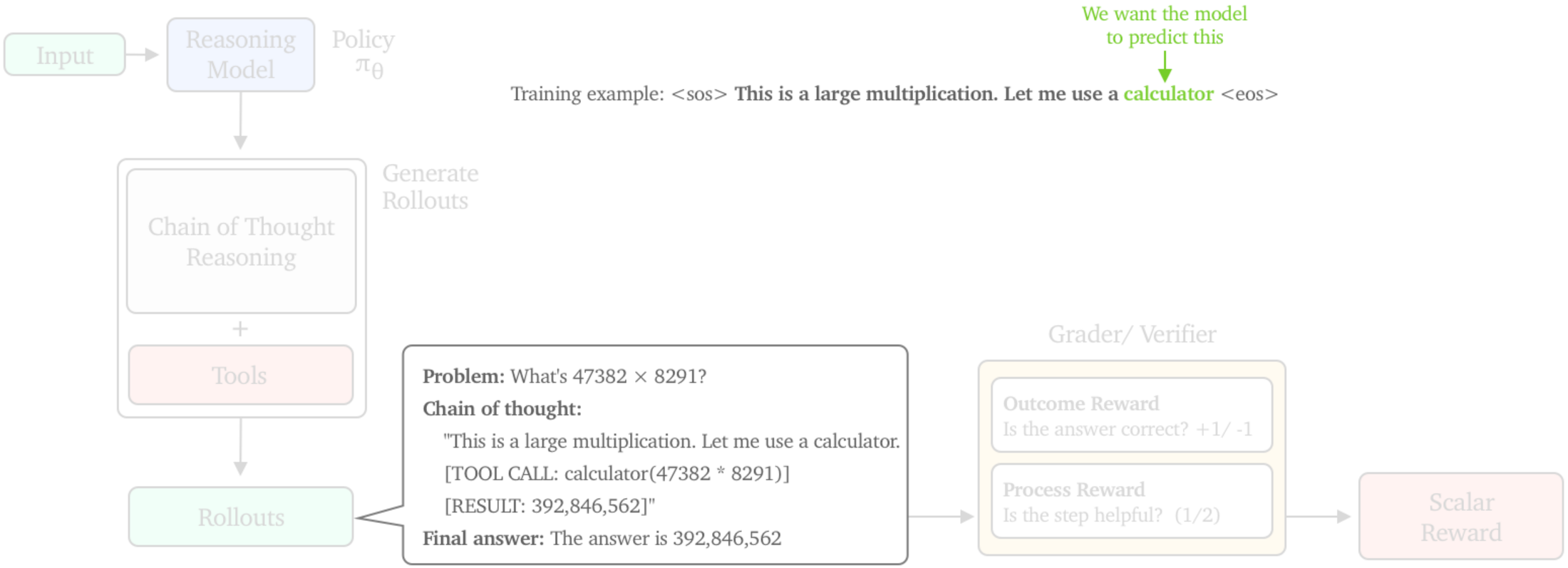
Test-time Compute

- But where is the RL?



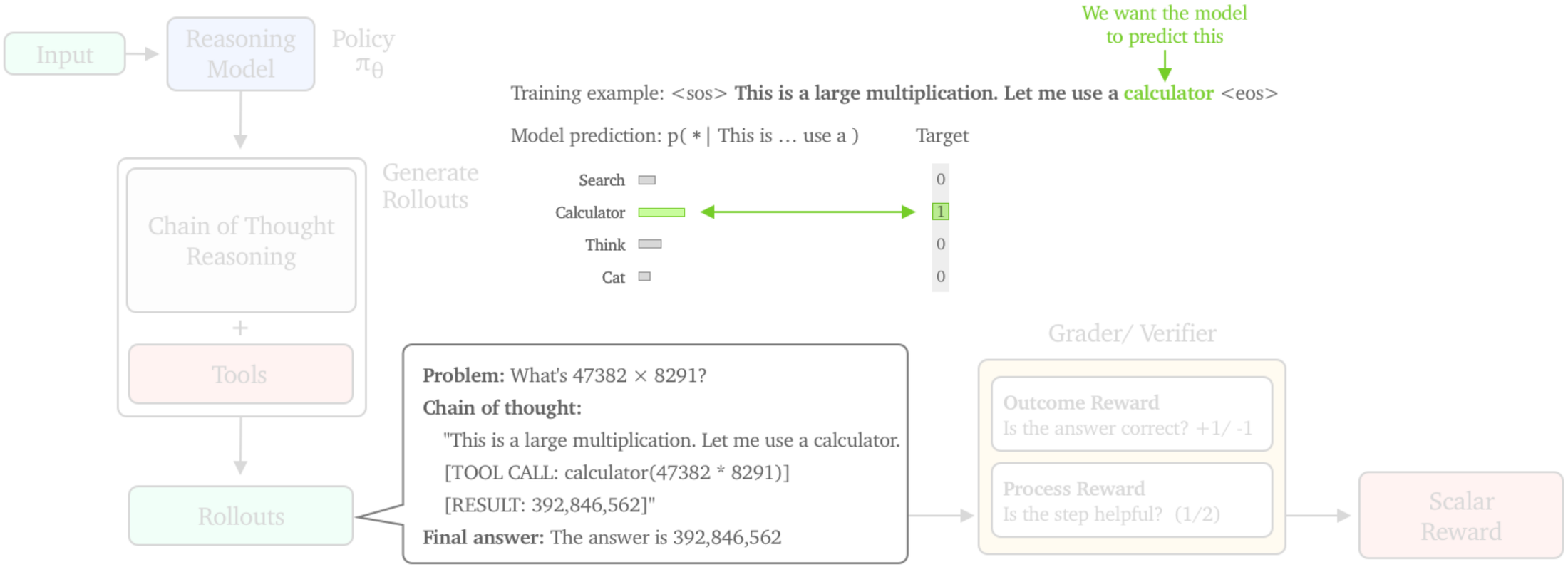
Test-time Compute

- But where is the RL?



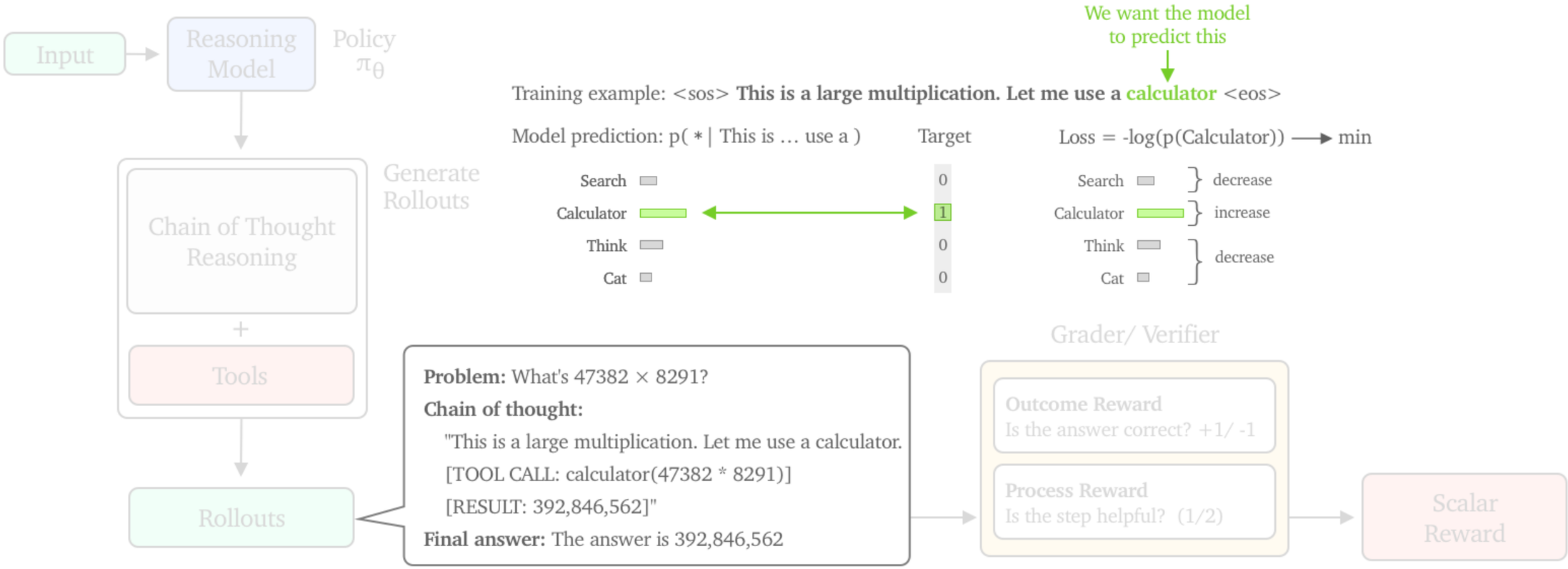
Test-time Compute

- But where is the RL?



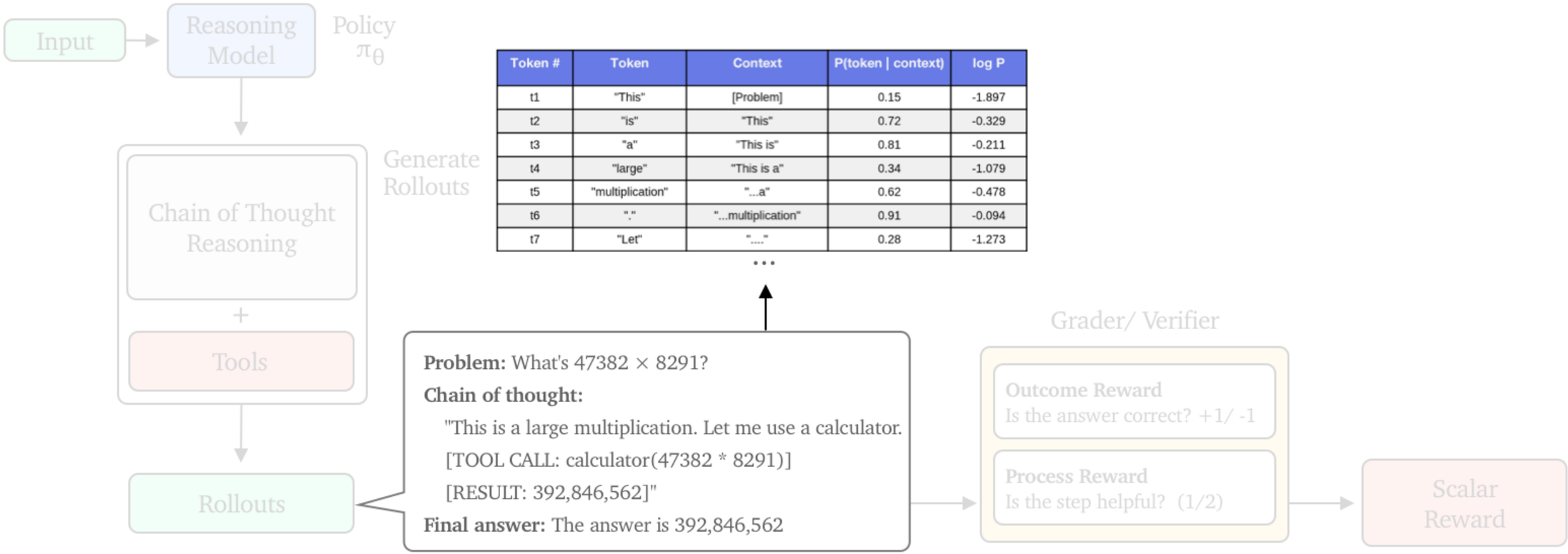
Test-time Compute

- But where is the RL?



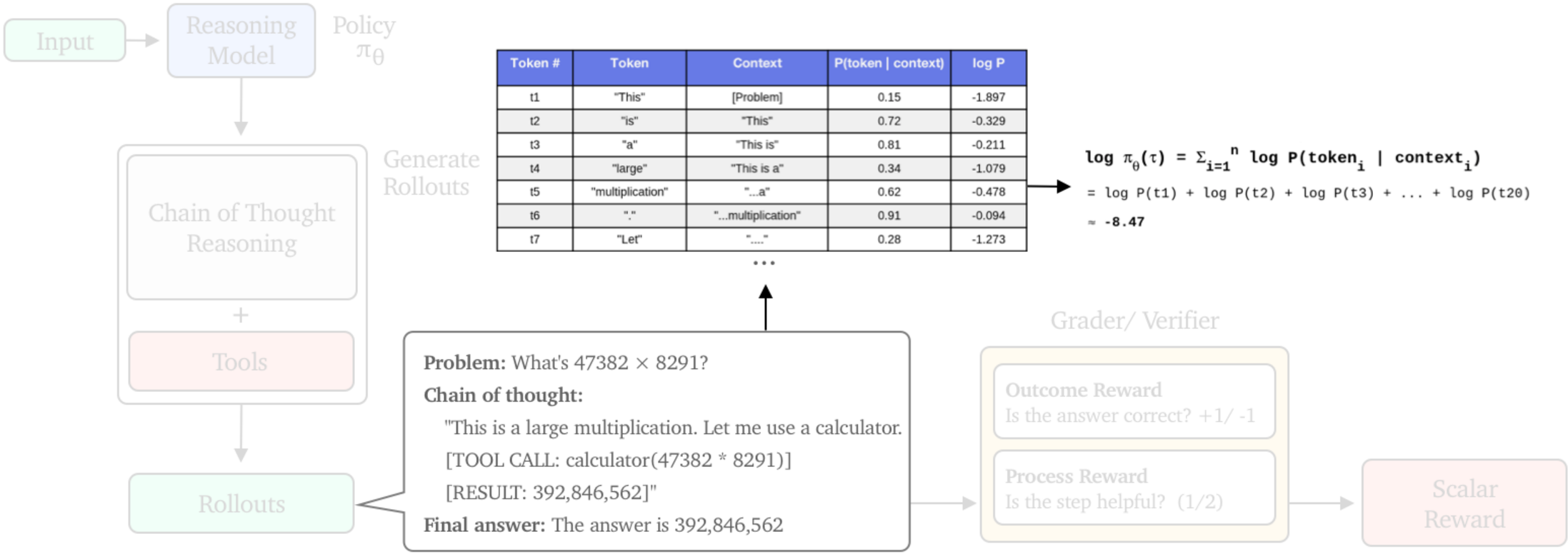
Test-time Compute

- But where is the RL?



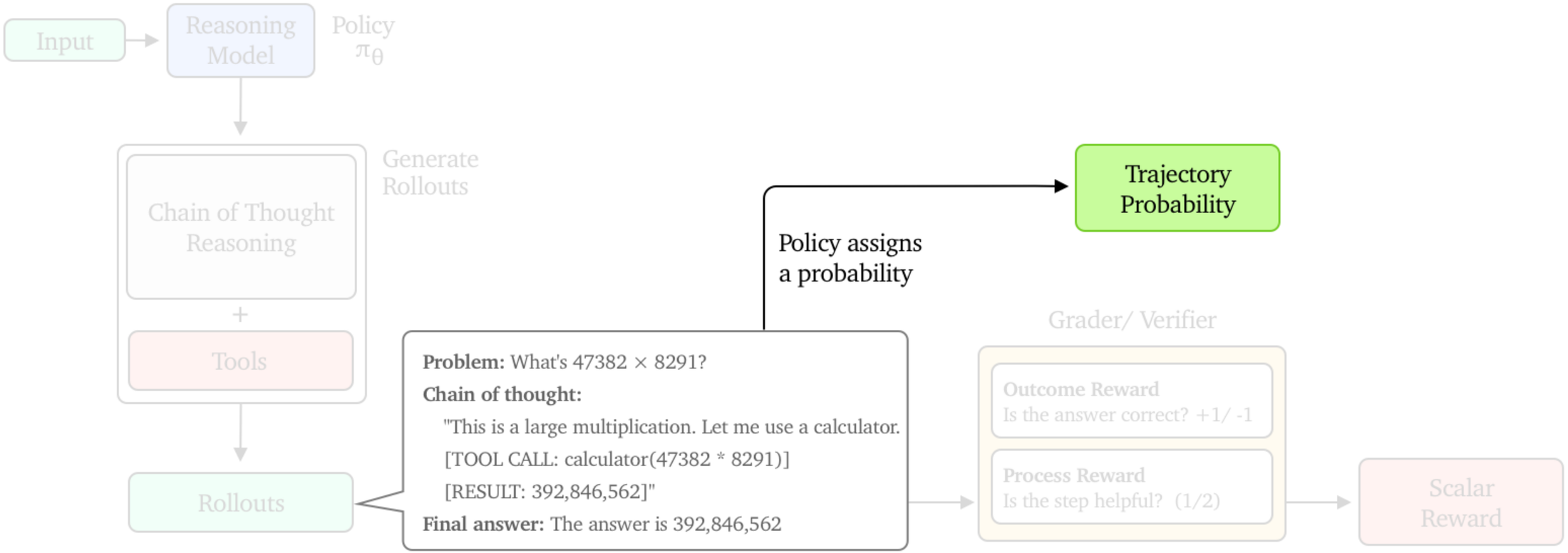
Test-time Compute

- But where is the RL?



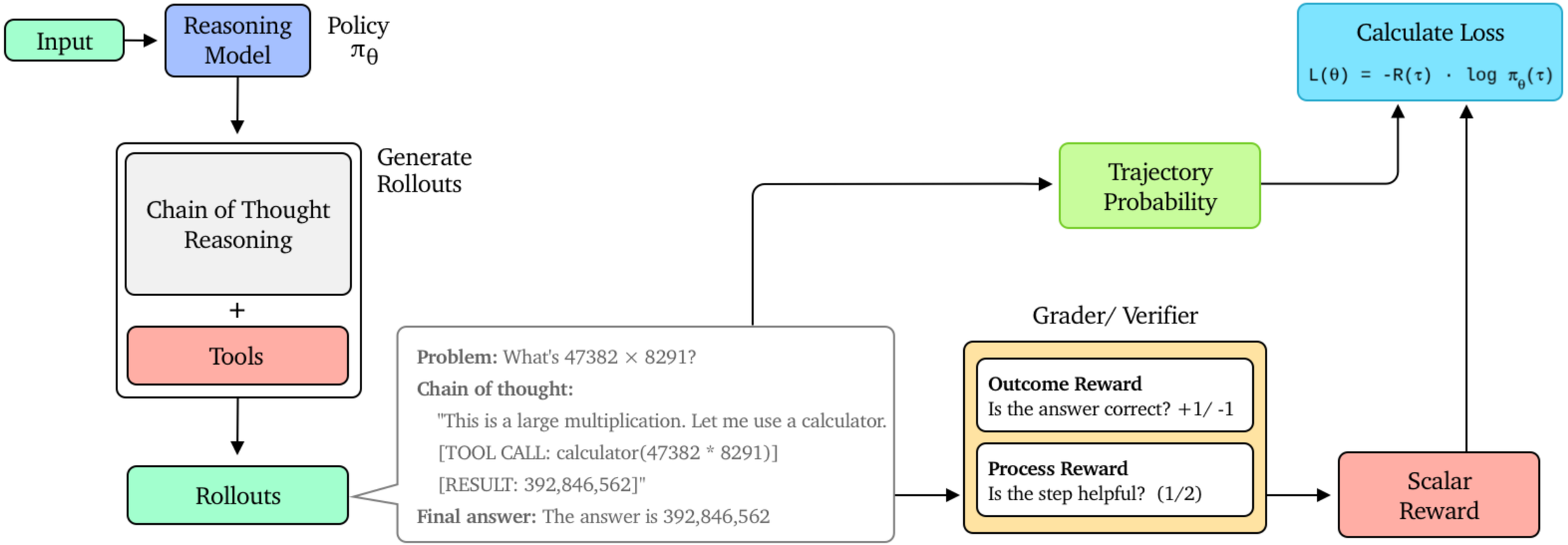
Test-time Compute

- But where is the RL?



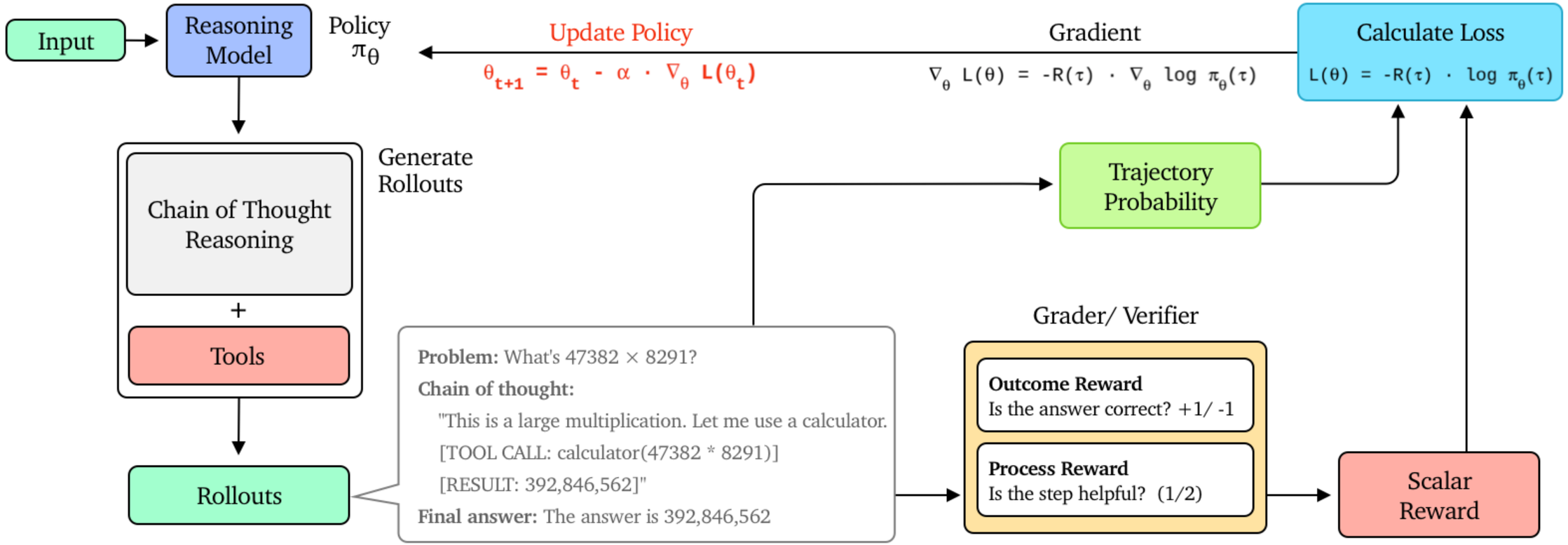
Test-time Compute

- But where is the RL?



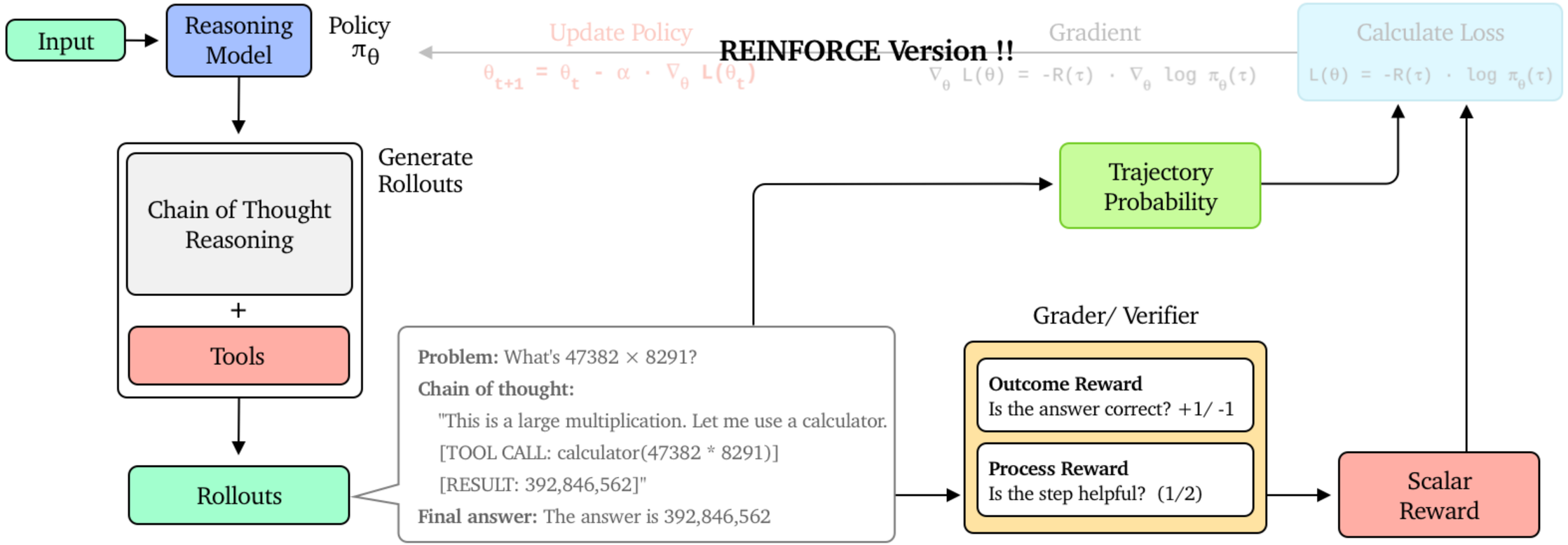
Test-time Compute

- But where is the RL?



Test-time Compute

- But where is the RL?

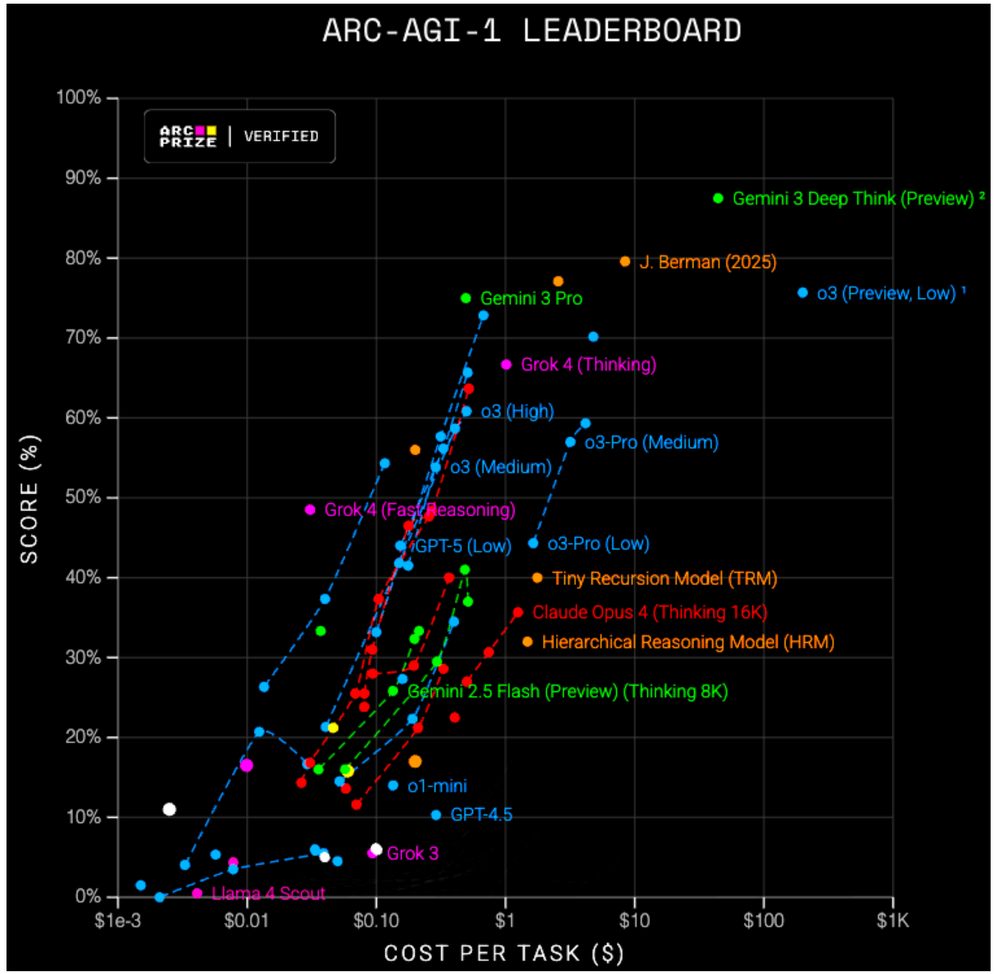


Test-time Compute Summary

- RL can be used to reinforce the thinking process of models
- Using REINFORCE, PPO, GPRO algorithms

Test-time Compute Summary

- RL can be used to reinforce the thinking process of models
- Using REINFORCE, PPO, GPRO algorithms
- Does it work?



Thank You!