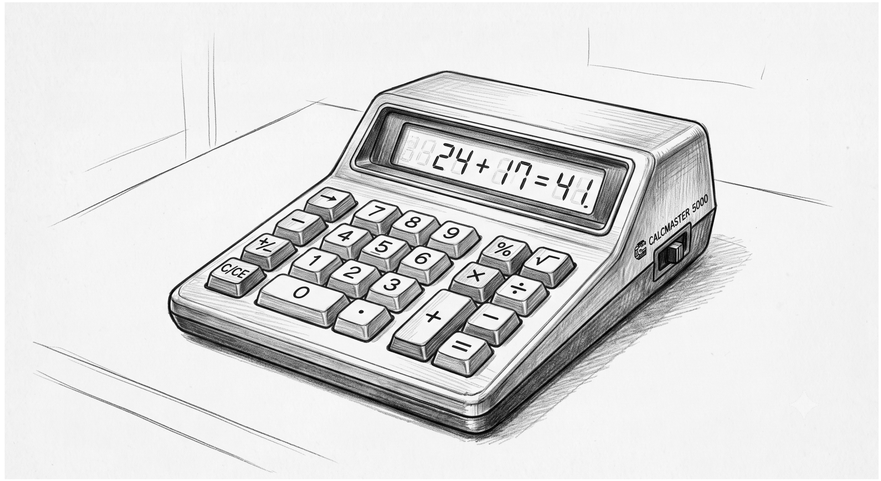
Continual Learning via Self-Distillation

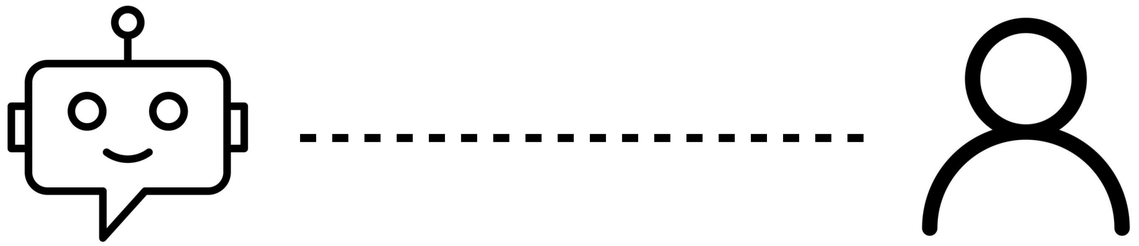
Idan Shenfeld



Instead of defining intelligence in terms of what an AI system can ***do***, focus on how well it can ***learn***.

Why should we care about the ability to learn?

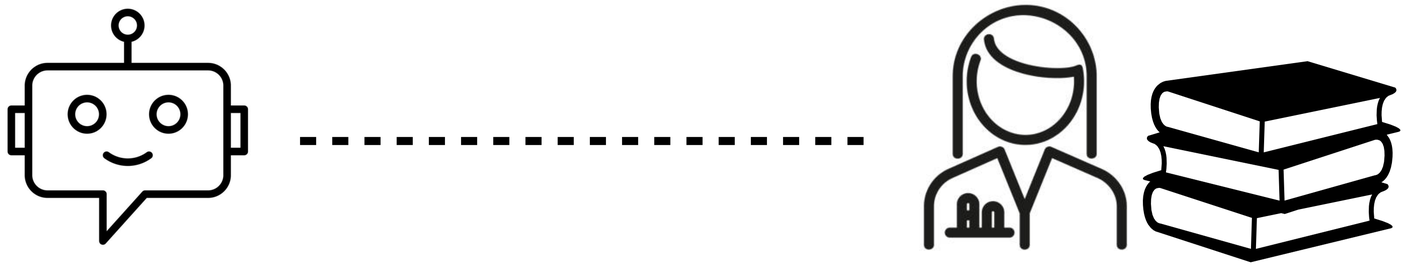
- Personalization - AI systems that learn and adapt to each user’s preferences.



Why should we care about the ability to learn?

- Personalization.

- Specialization - To solve hard problems, AI systems must absorb the relevant literature, context, and domain knowledge.

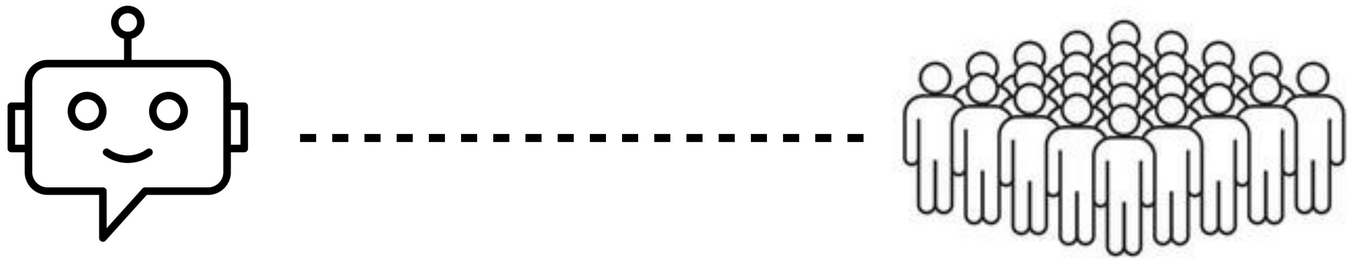


Why should we care about the ability to learn?

- Personalization.

- Specialization.

- Next Scaling Frontier - Deployed AI now processes more than 300 trillion tokens per day, roughly a web-scale corpus every day.



How Can Foundation Models Keep Learning?

Memory-based Approaches

Parameter-update Approaches

How Can Foundation Models Keep Learning?

Memory-based Approaches

- In-context Learning
- External Memory
- RAG
- Harness

Established but Limited

Parameter-update Approaches

How Can Foundation Models Keep Learning?

Memory-based Approaches

- In-context Learning
- External Memory
- RAG
- Recurrent NNs
- State-Space Models

Established but Limited

Parameter-update Approaches

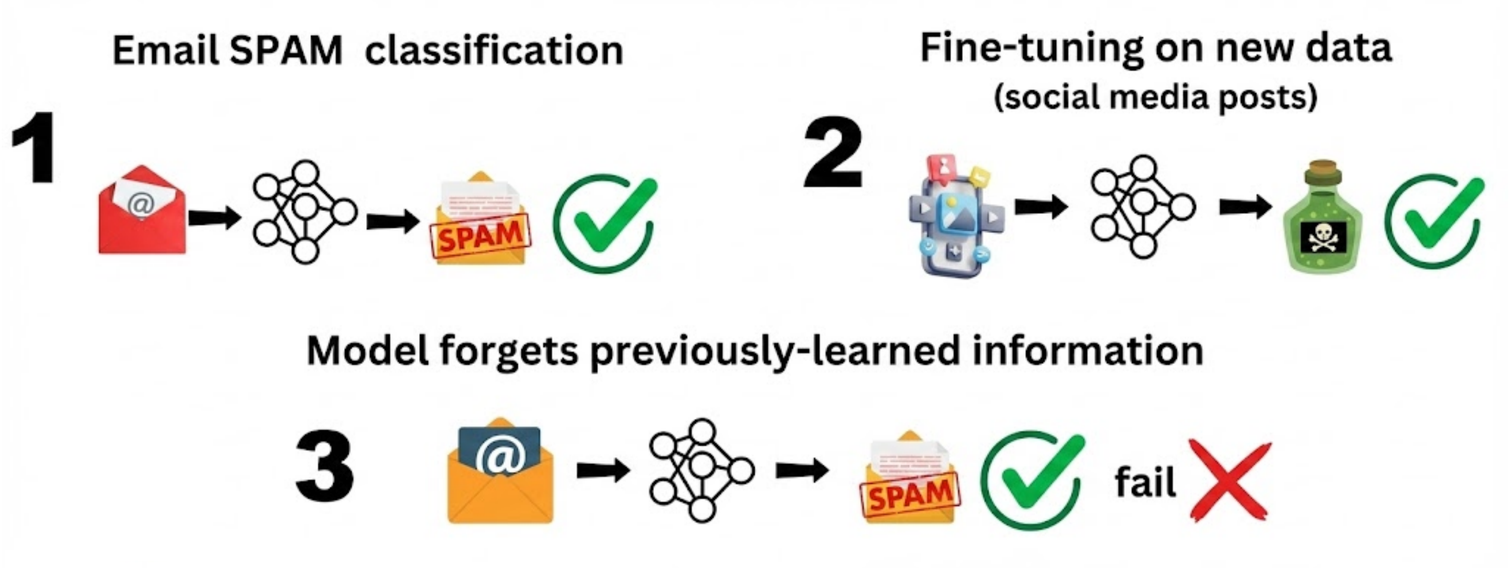
- Model Fine-Tuning
- Adapters
- Model Editing and Merging
- New algorithms?

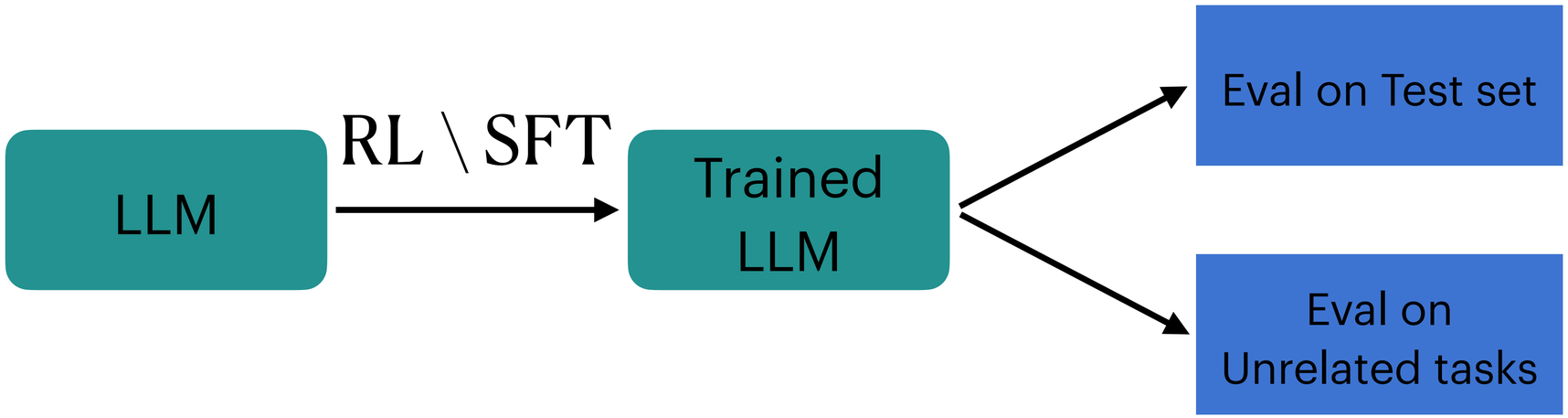
Open Research Questions

Part I

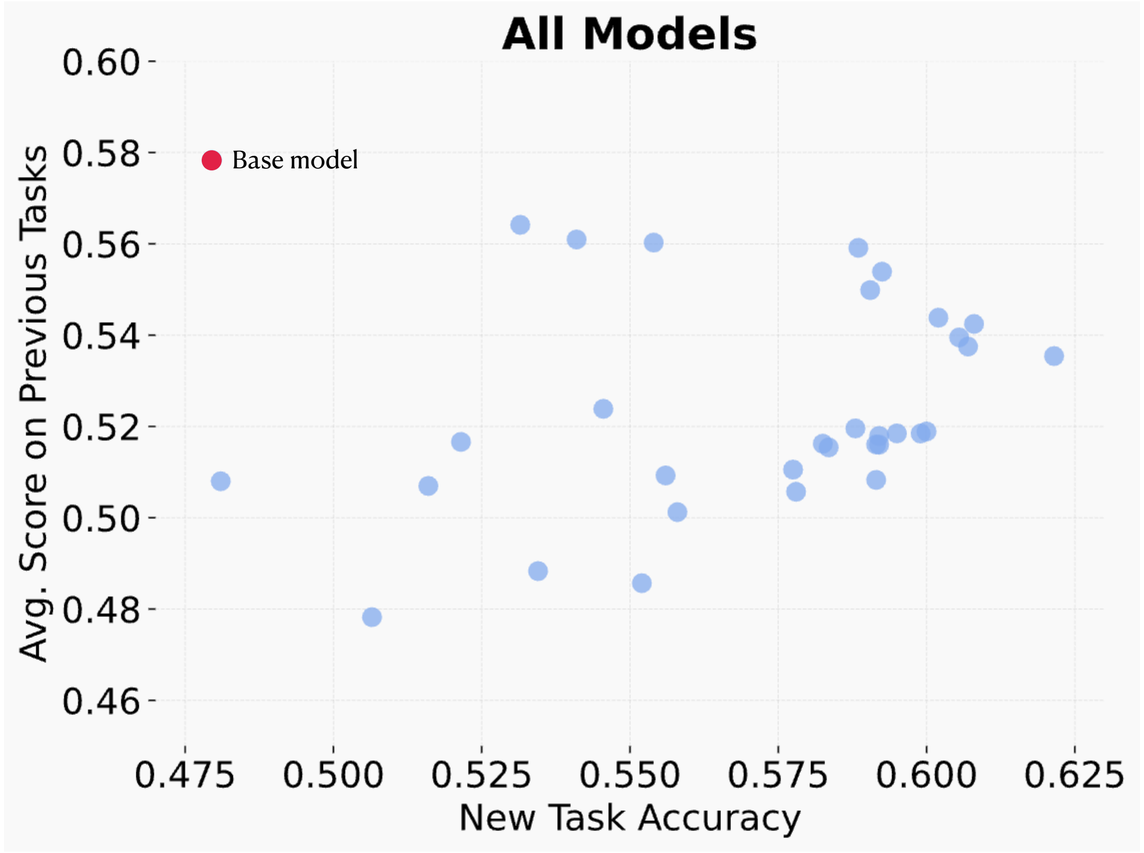
Catastrophic Forgetting

The phenomenon where a neural network loses previously learned information when it is trained on a new task

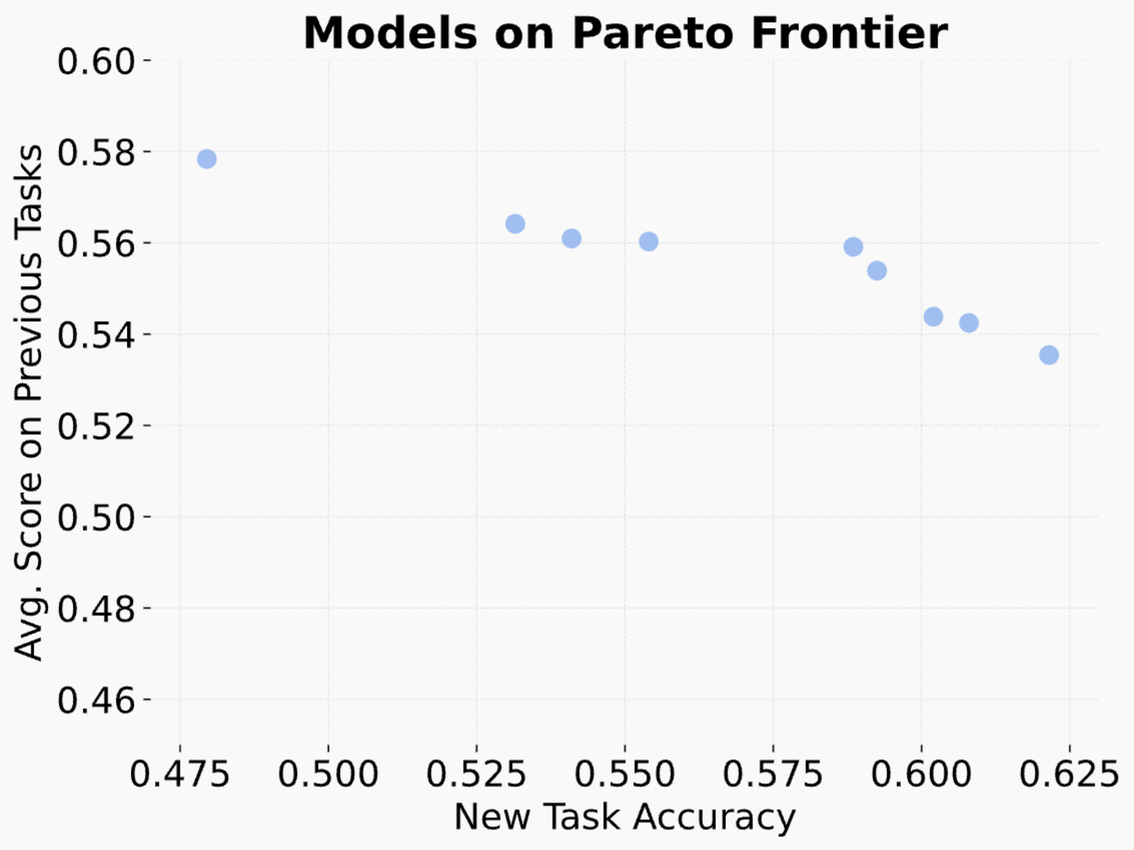




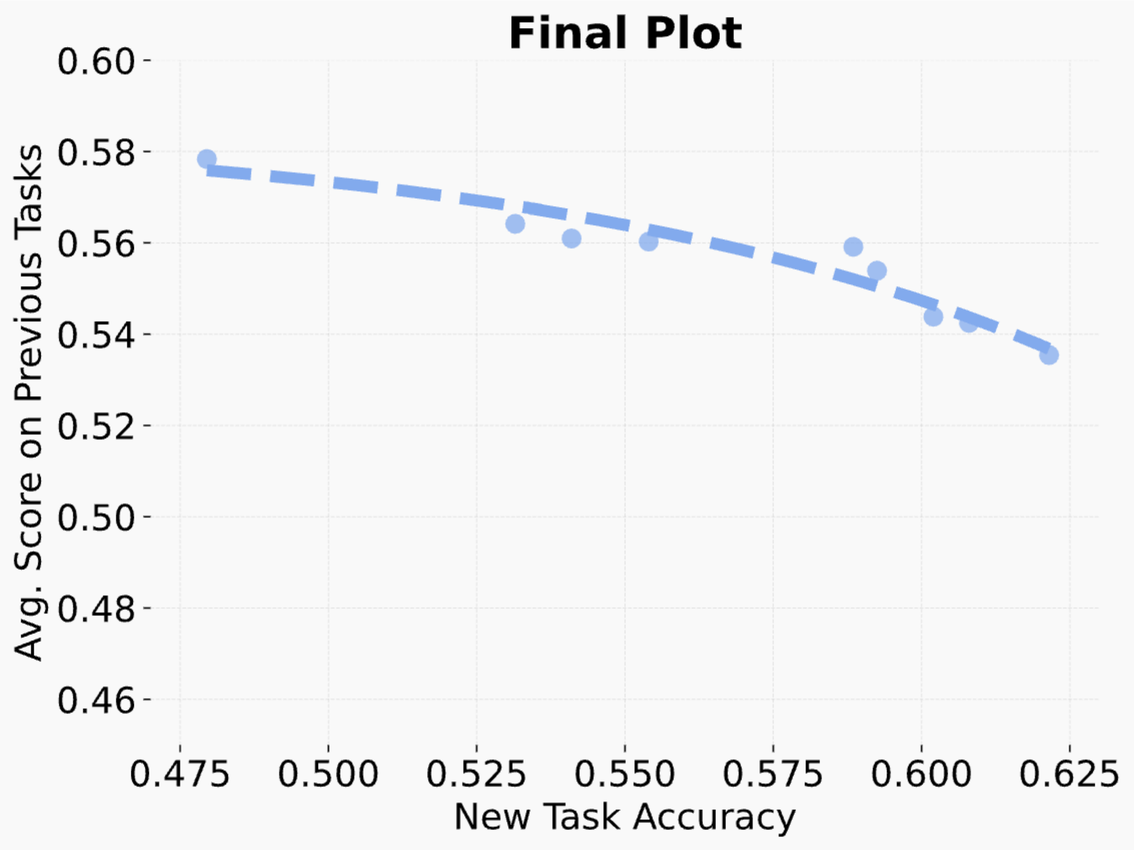
Shenfeld et al. 2025, *RL's Razor: Why Online Reinforcement Learning Forgets Less.*



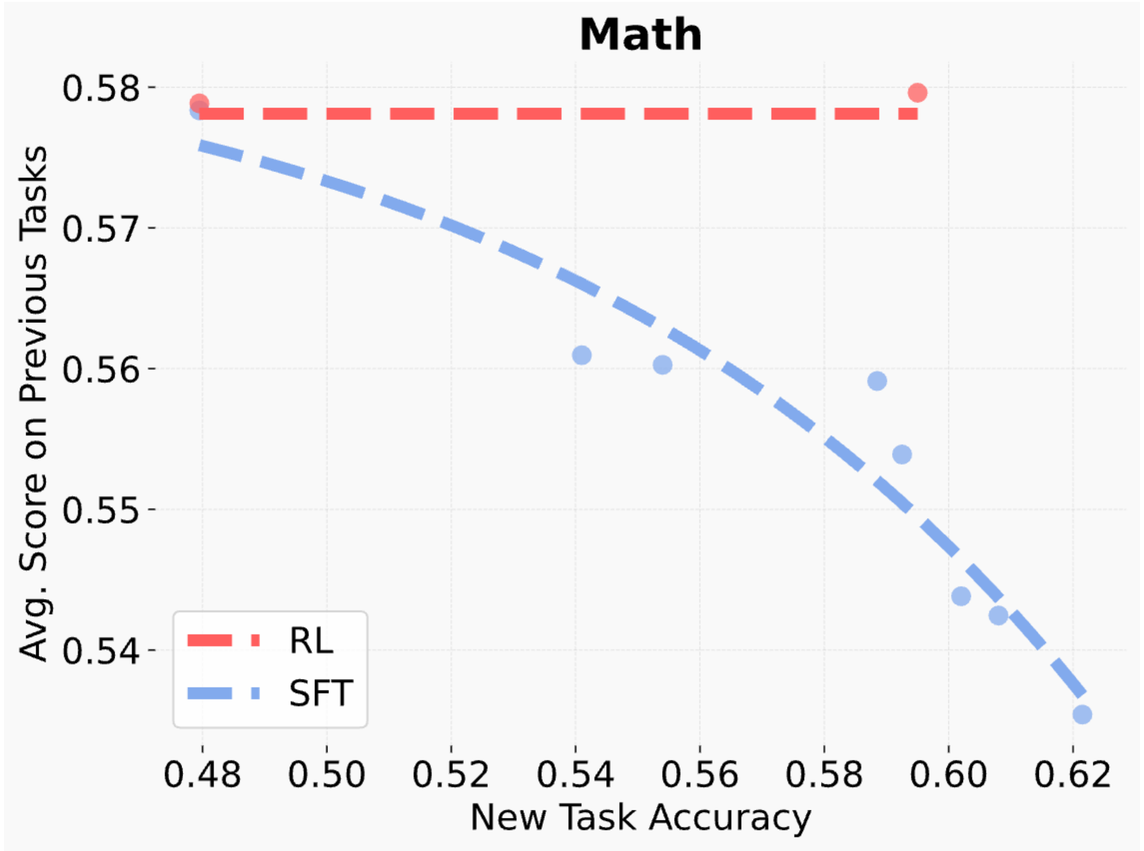
Shenfeld et al. 2025, *RL's Razor: Why Online Reinforcement Learning Forgets Less.*



Shenfeld et al. 2025, *RL's Razor: Why Online Reinforcement Learning Forgets Less.*

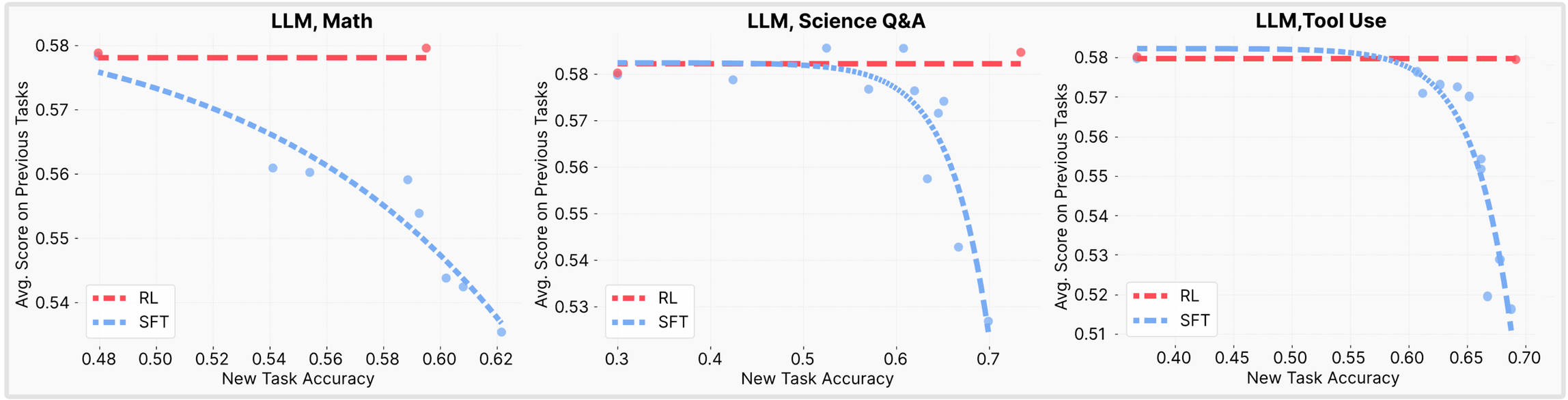


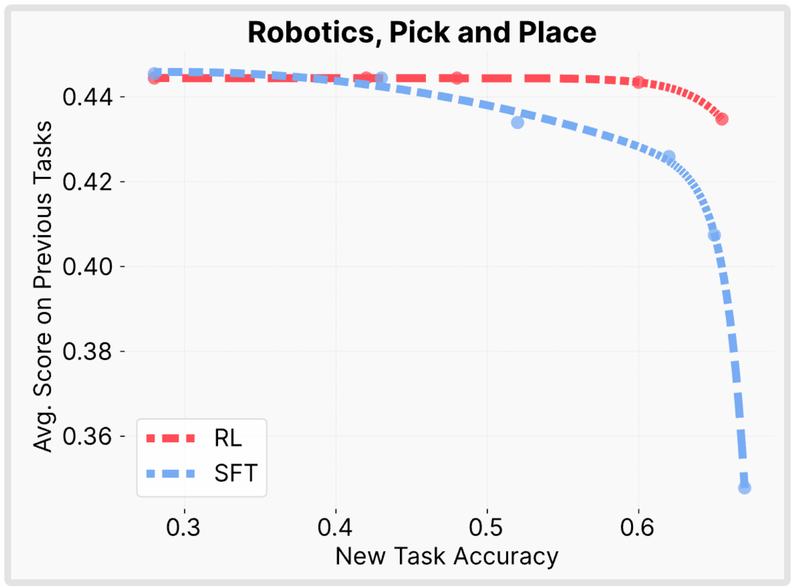
Shenfeld et al. 2025, *RL's Razor: Why Online Reinforcement Learning Forgets Less.*



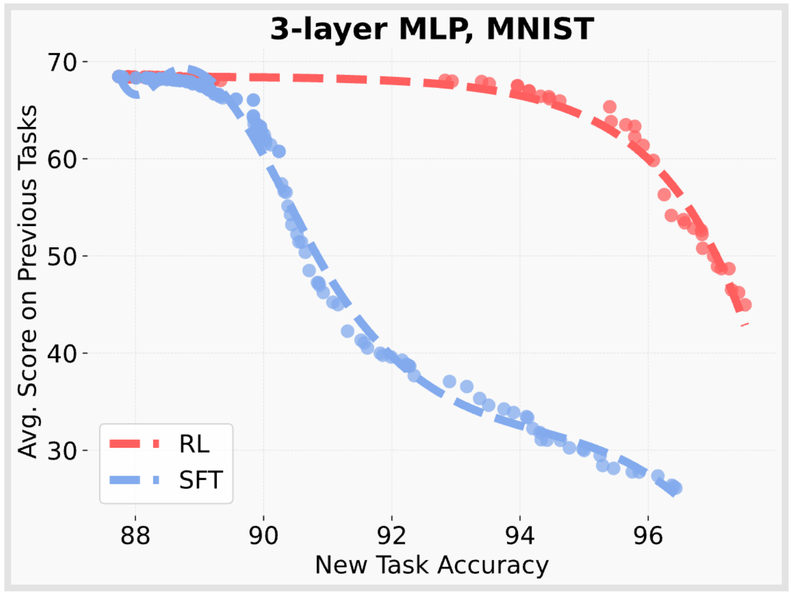
Qwen 3B-Instruct on Open-Reasoner-Zero dataset; SFT annotations from R1.

Shenfeld et al. 2025, *RL's Razor: Why Online Reinforcement Learning Forgets Less.*


LLM, Math
Avg. Score on Previous Tasks
New Task Accuracy
RL
SFT
0.58
0.57
0.56
0.55
0.54
0.48
0.50
0.52
0.54
0.56
0.58
0.60
0.62
LLM, Science Q&A
Avg. Score on Previous Tasks
New Task Accuracy
RL
SFT
0.58
0.57
0.56
0.55
0.54
0.53
0.3
0.4
0.5
0.6
0.7
LLM,Tool Use
Avg. Score on Previous Tasks
New Task Accuracy
RL
SFT
0.58
0.57
0.56
0.55
0.54
0.53
0.52
0.51
0.40
0.45
0.50
0.55
0.60
0.65
0.70



OpenVLA in SimplerEnv simulator; SFT annotations from RT-1.



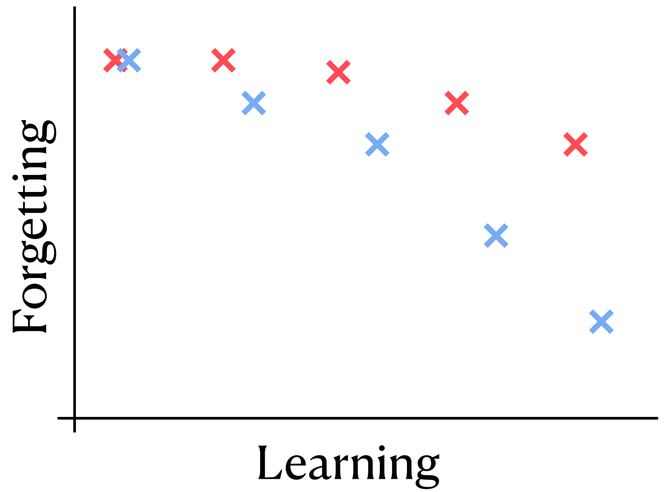
FashionMNIST and ParityMNIST in pertaining; Fine-tuning on ParityMNIST.

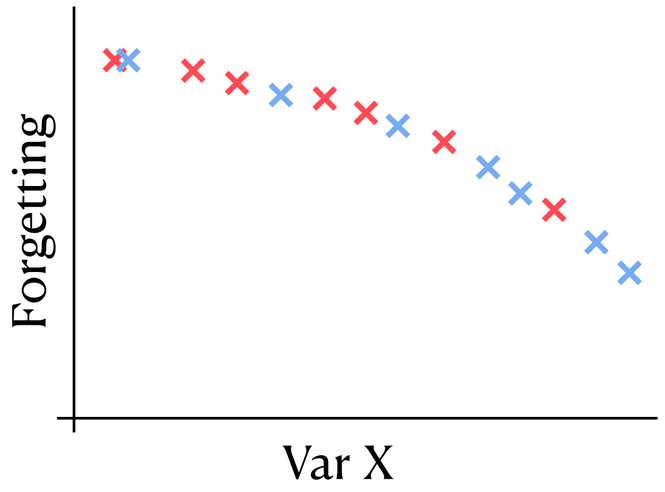
Shenfeld et al. 2025, *RL's Razor: Why Online Reinforcement Learning Forgets Less.*

So RL forgets less, but why?

Predictive Variable

An **objective-agnostic** measurement that explains forgetting.

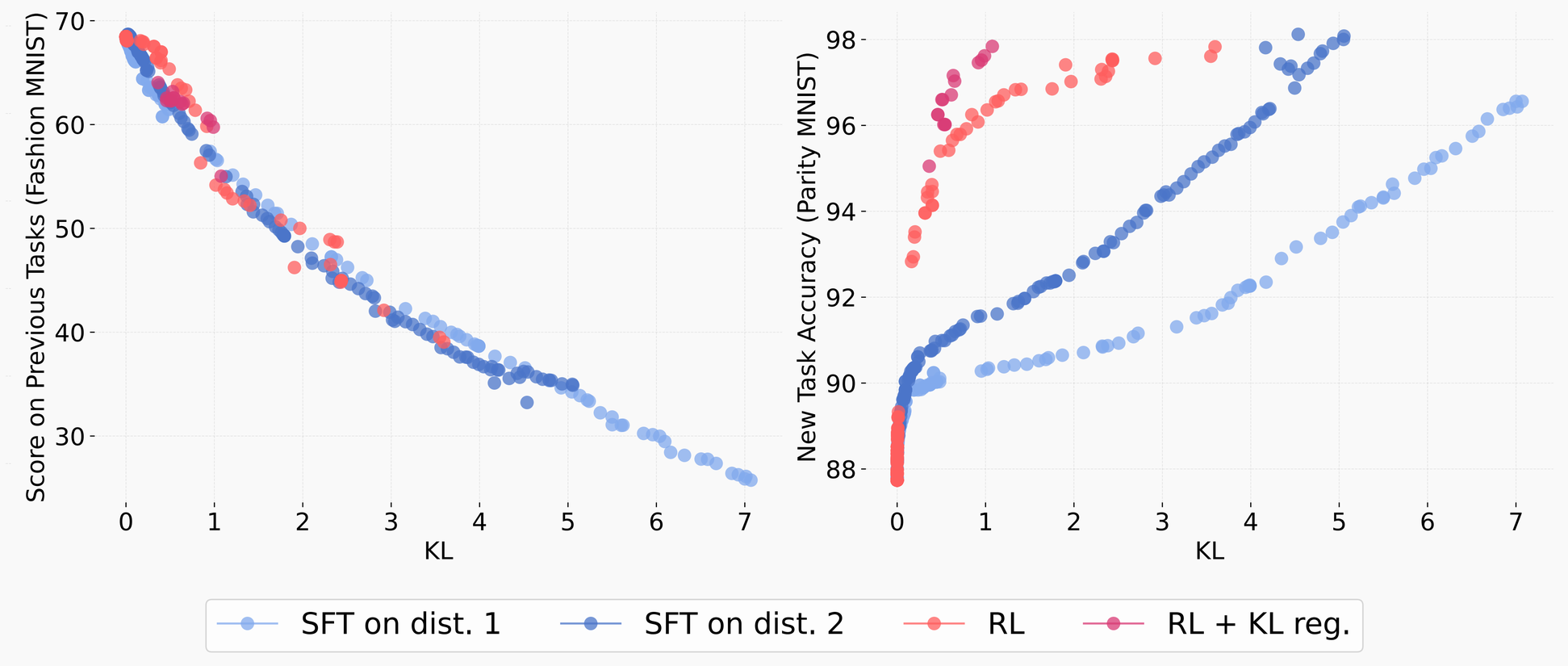




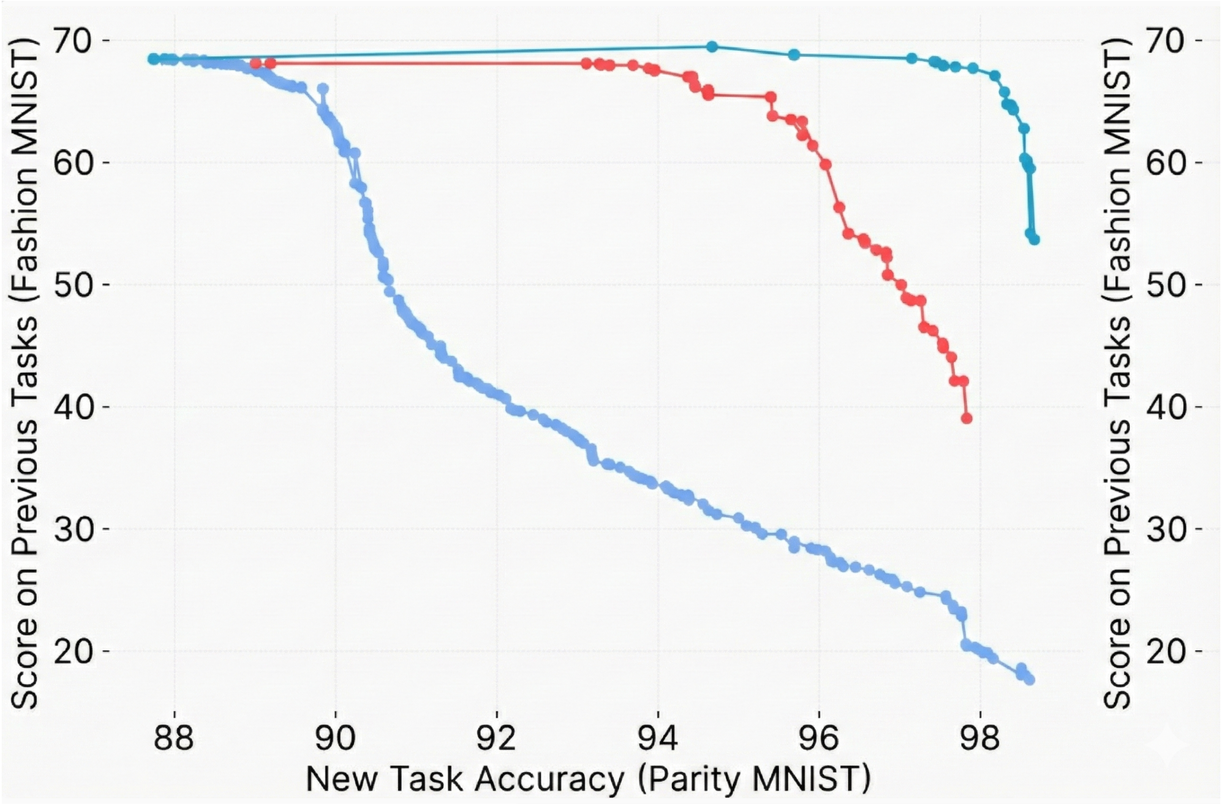
Predictive Variable

Forgetting is almost perfectly predicted by the *KL divergence* between the fine-tuned and base policy *evaluated on the new task*.

$$\mathbb{E}_{x \sim \tau_{new}} \left[\mathrm{KL}(\pi_0(\cdot \mid x) \,||\, \pi(\cdot \mid x)) \right]$$


Score on Previous Tasks (Fashion MNIST)
70
60
50
40
30
0
1
2
3
4
5
6
7
KL
New Task Accuracy (Parity MNIST)
98
96
94
92
90
88
0
1
2
3
4
5
6
7
KL
SFT on dist. 1
SFT on dist. 2
RL
RL + KL reg.

SFT on Optimal Annotations



Policy Gradient Objective

$$-\sum_{x\sim D, y\sim\pi}[R(x,y)\log\pi(x,y)]$$

Cross-Entropy Loss

$$-\sum_{x\sim D, y\sim\pi^*}[\log\pi(x,y)]$$

Policy Gradient Objective

$$-\sum_{x\sim D, y\sim\pi} [R(x,y)\log \pi(x,y)]$$

Cross-Entropy Loss

$$-\sum_{x\sim D, y\sim\pi^*} [\log \pi(x,y)]$$

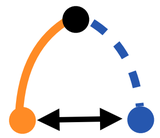
On-policy vs. Off-policy

On-Policy Learning

Given an teacher model π_T, we want to fit a student model π to its distribution

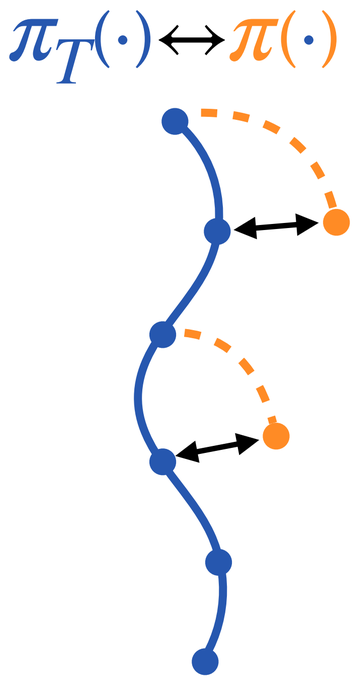
Given an teacher model π_T, we want to fit a student model π to its distribution

$\pi_T(\cdot) \leftrightarrow \pi(\cdot)$

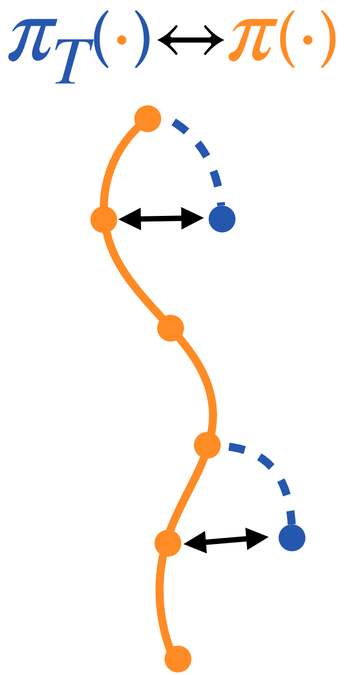


Adapted from slides by Gokul Swamy.

Given an teacher model π_T, we want to fit a student model π to its distribution



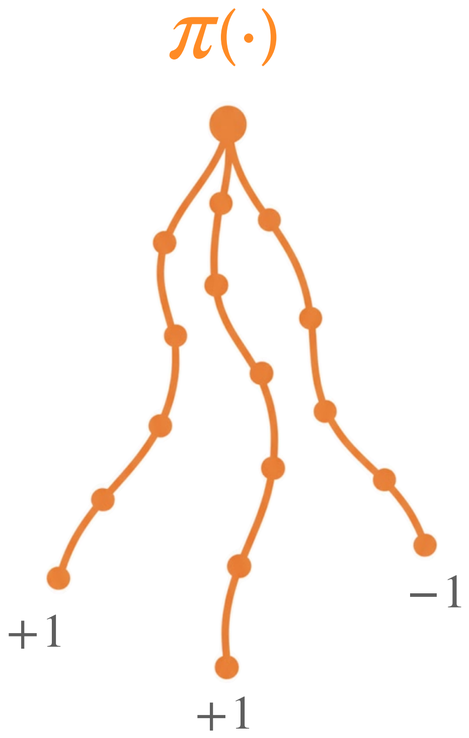
Offline\Off-Policy



On-Policy

Adapted from slides by Gokul Swamy.

When we don't have a teacher, we use a Reward Function instead



On-Policy RL

Adapted from slides by Gokul Swamy.

Policy Gradient Objective

$$-\sum_{x\sim D, y\sim\pi} [R(x,y)\log \pi(x,y)]$$

Cross-Entropy Loss

$$-\sum_{x\sim D, y\sim\pi^*} [\log \pi(x,y)]$$

On-policy vs. Off-policy

Policy Gradient Objective

$$-\sum_{x\sim D, y\sim\pi} [R(x,y) \log \pi(x,y)]$$

Cross-Entropy Loss

$$-\sum_{x\sim D, y\sim\pi^*} [\log \pi(x,y)]$$

Negative Examples

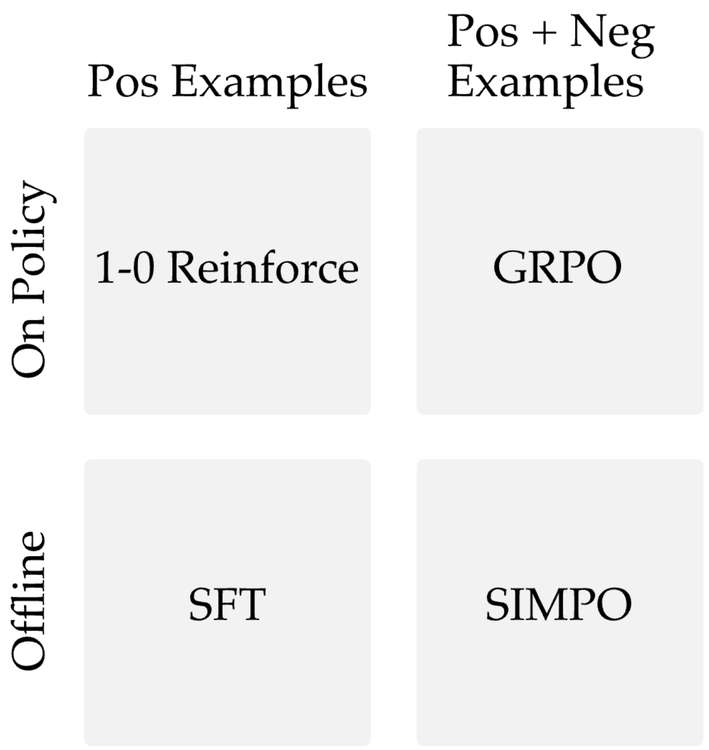
Policy Gradient Objective

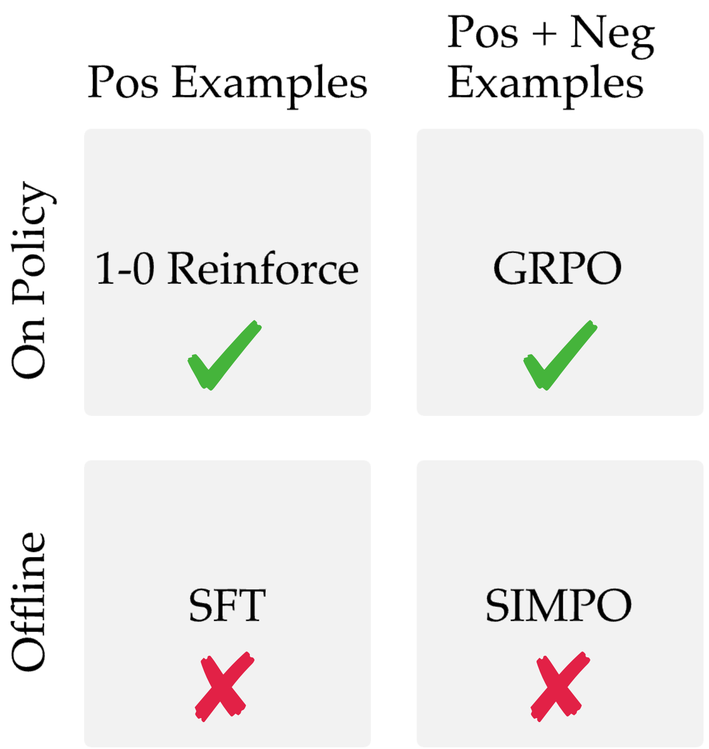
$$-\sum_{x\sim D, y\sim\pi} [R(x,y)\log \pi(x,y)]$$

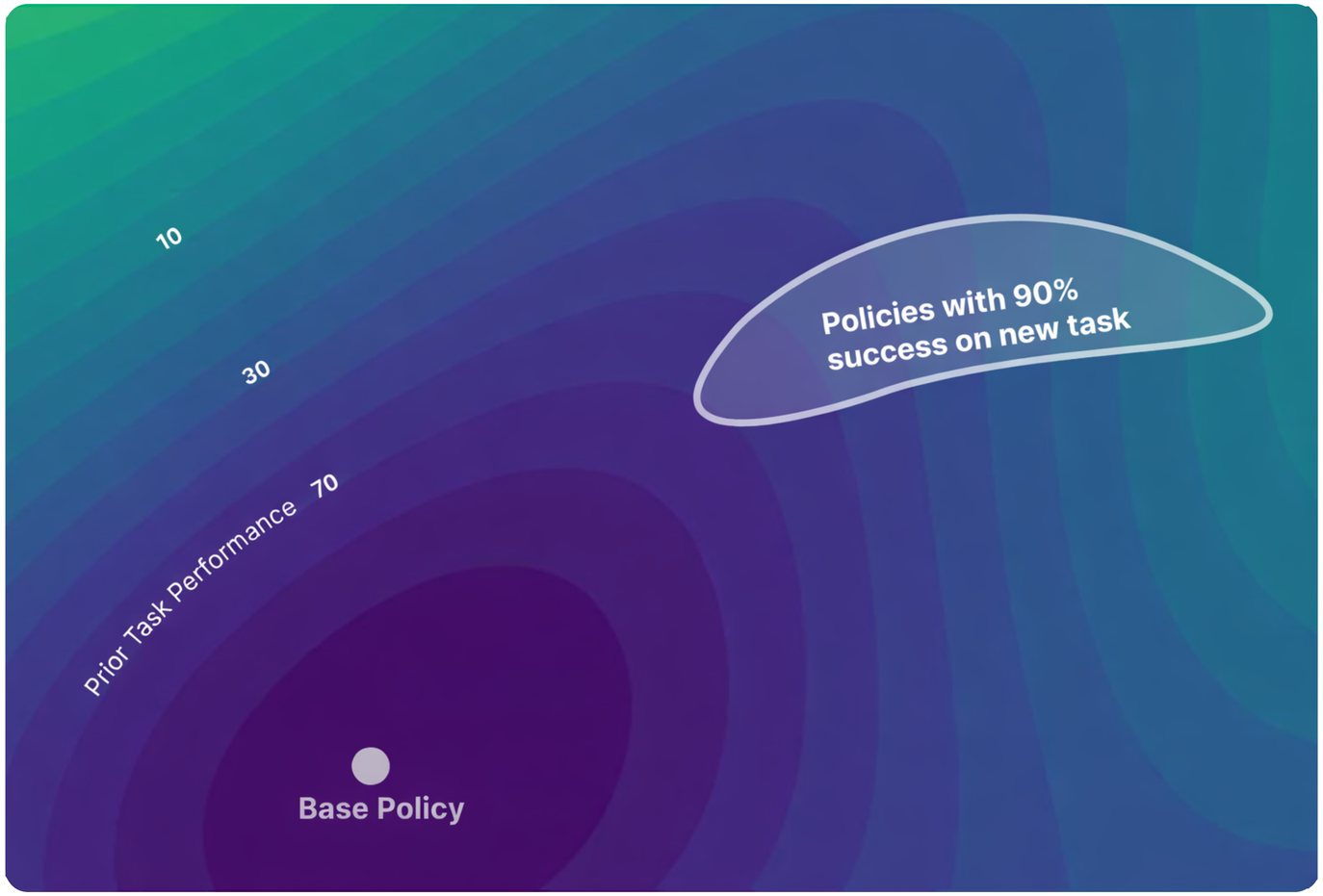
Cross-Entropy Loss

$$-\sum_{x\sim D, y\sim\pi^*} [\log \pi(x,y)]$$

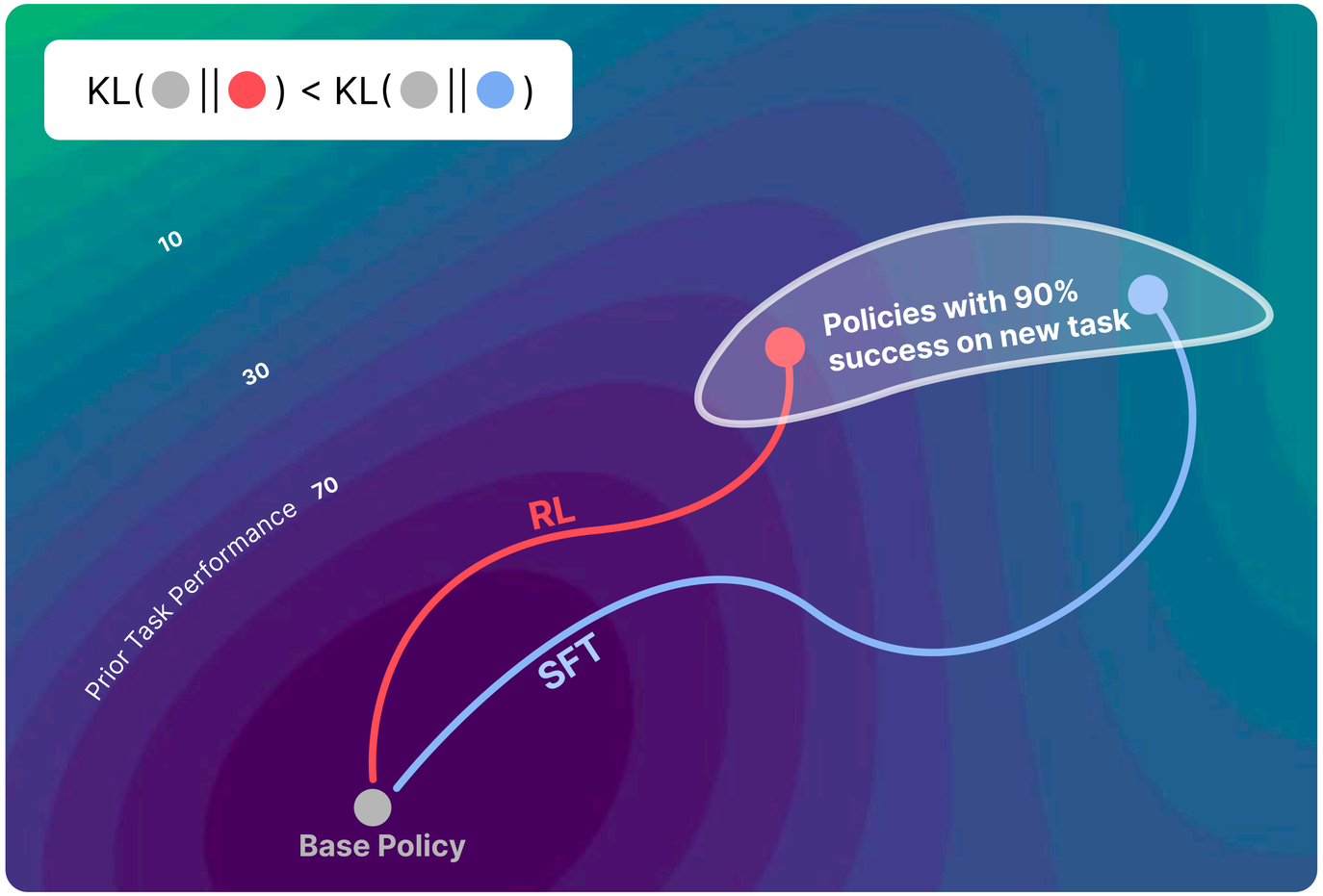
Negative Examples or On-Policy


Pos Examples
Pos + Neg Examples
On Policy
Offline
1-0 Reinforce
GRPO
SFT
SIMPO


Pos Examples
Pos + Neg Examples
On Policy
Offline
1-0 Reinforce
GRPO
SFT
SIMPO



Shenfeld et al. 2025, *RL's Razor: Why Online Reinforcement Learning Forgets Less.*



Shenfeld et al. 2025, *RL's Razor: Why Online Reinforcement Learning Forgets Less.*

Part II

- Main takeaway - to learn without forgetting, we should do on-policy learning.

- On-policy RL is limited to scalar rewards.

- But there are many types of learning signals:
  - Textual feedback
  - Hints
  - Detailed Instructions
  - And more

In-Context Learning

- The ability of Large Models to change their behavior based their context.
- Doesn’t have to be examples - works well with feedback, scalar rewards, hints and more.
- Transient effect.

Self-Distillation

Prompt x: "what are the roots of $x^2 - 9$?"

Context c: "Use the equation $a^2 - b^2 = (a - b)(a + b)$"

Sampling Response $y \sim \pi(\cdot \mid x)$ leads to incorrect solution

Sampling Response $y \sim \pi(\cdot \mid x, c)$ leads to correct solution

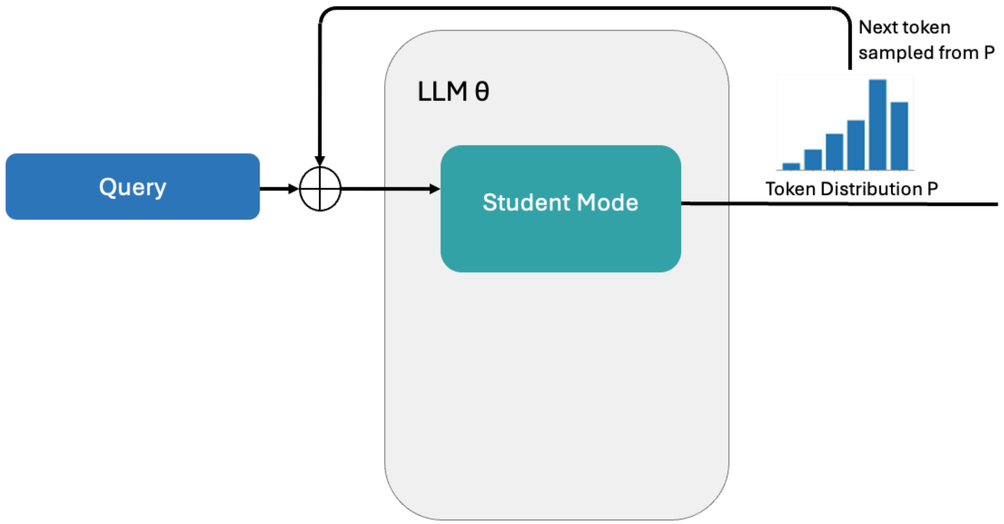
Self-Distillation

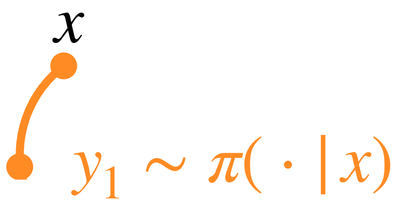
Prompt x, Response y, Context c



Self-Distillation

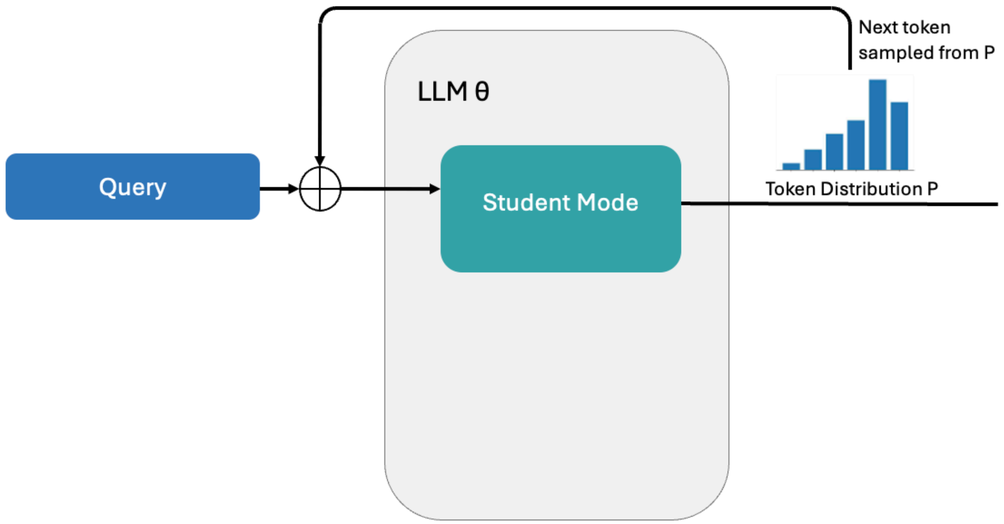
Prompt x, Response y, Context c

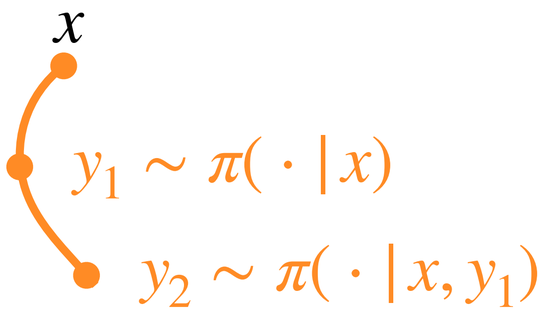




Self-Distillation

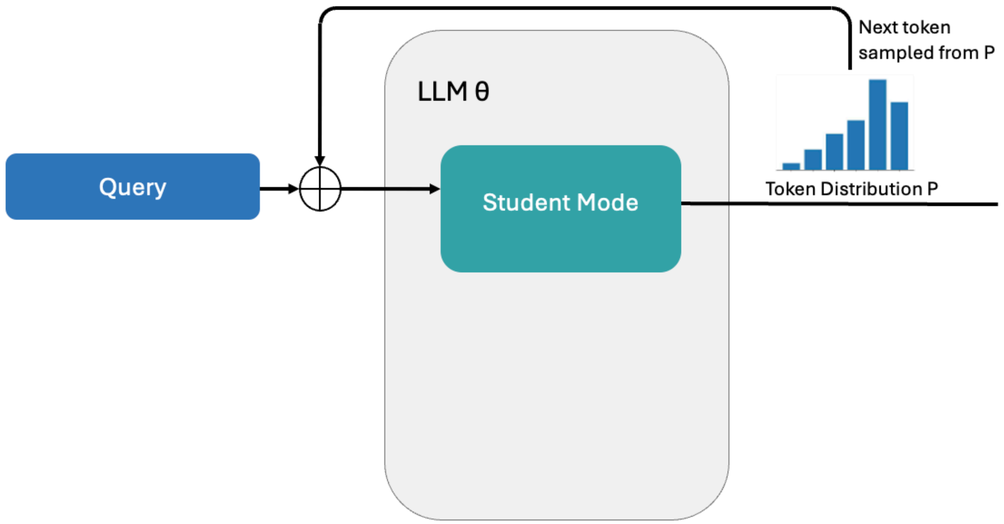
Prompt x, Response y, Context c

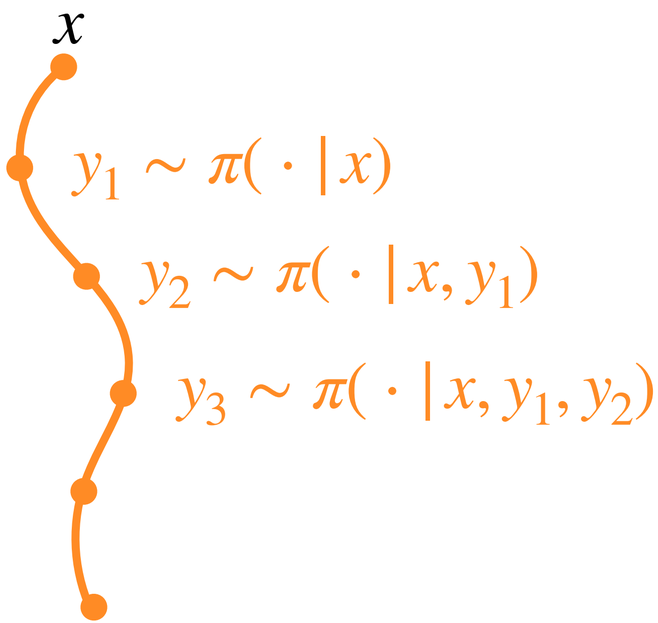




Self-Distillation

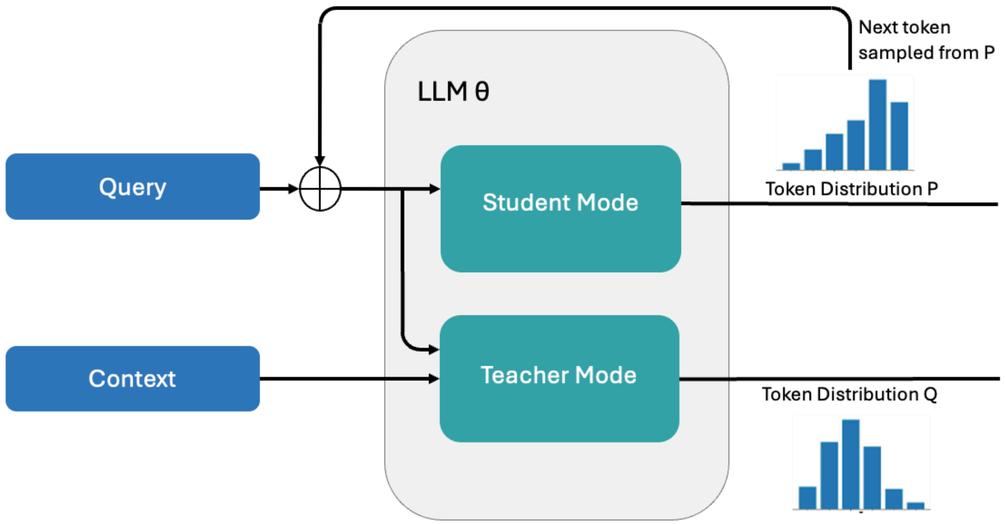
Prompt x, Response y, Context c

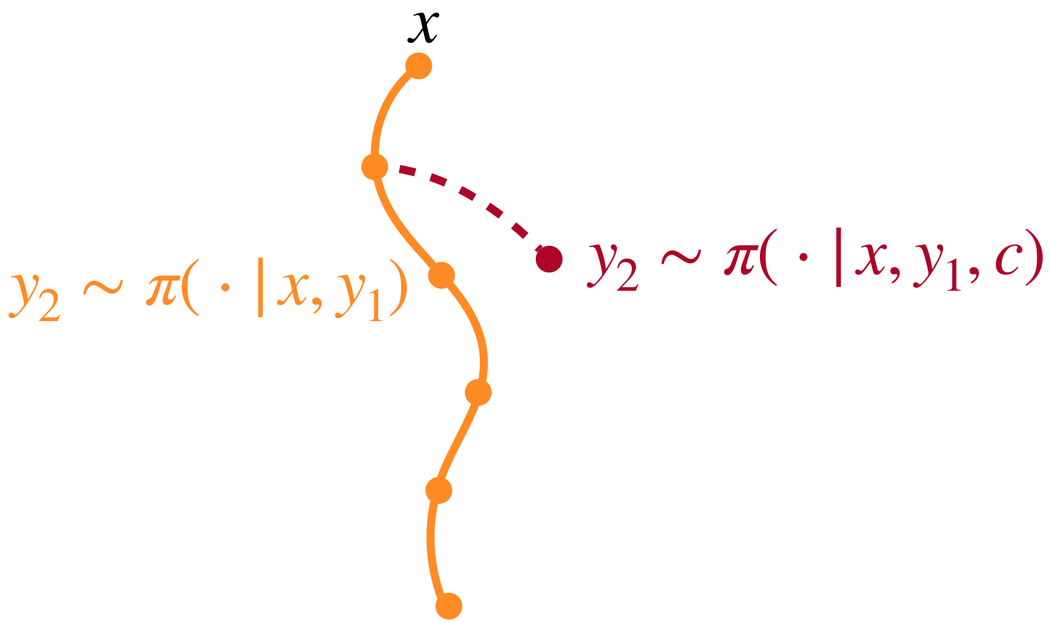




Self-Distillation

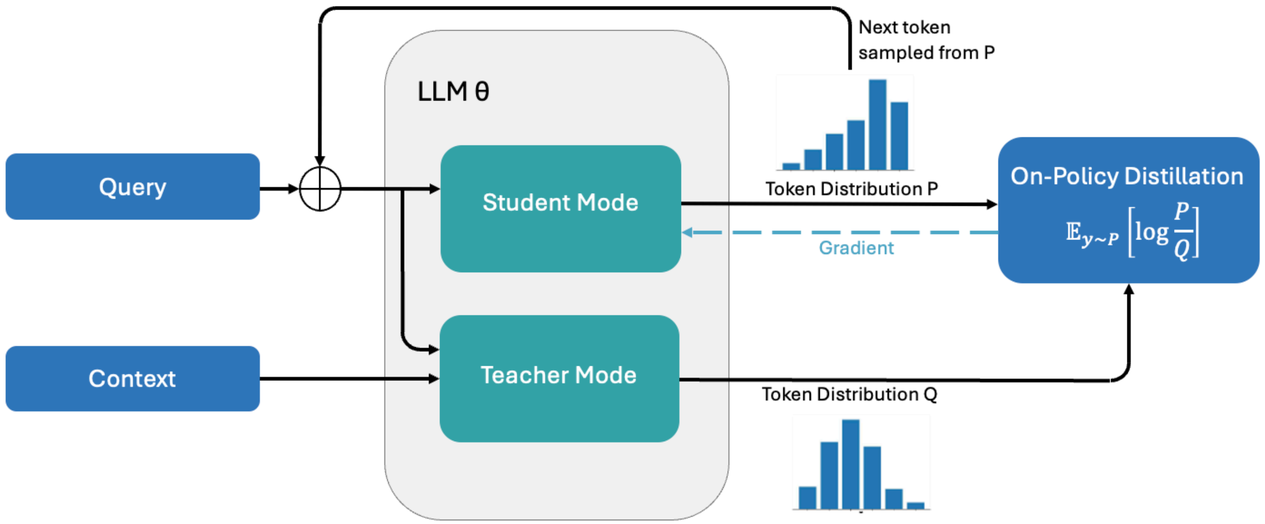
Prompt x, Response y, Context c

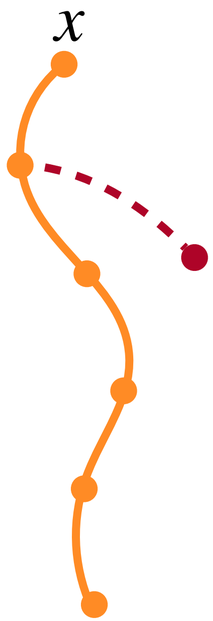




Self-Distillation

Prompt x, Response y, Context c

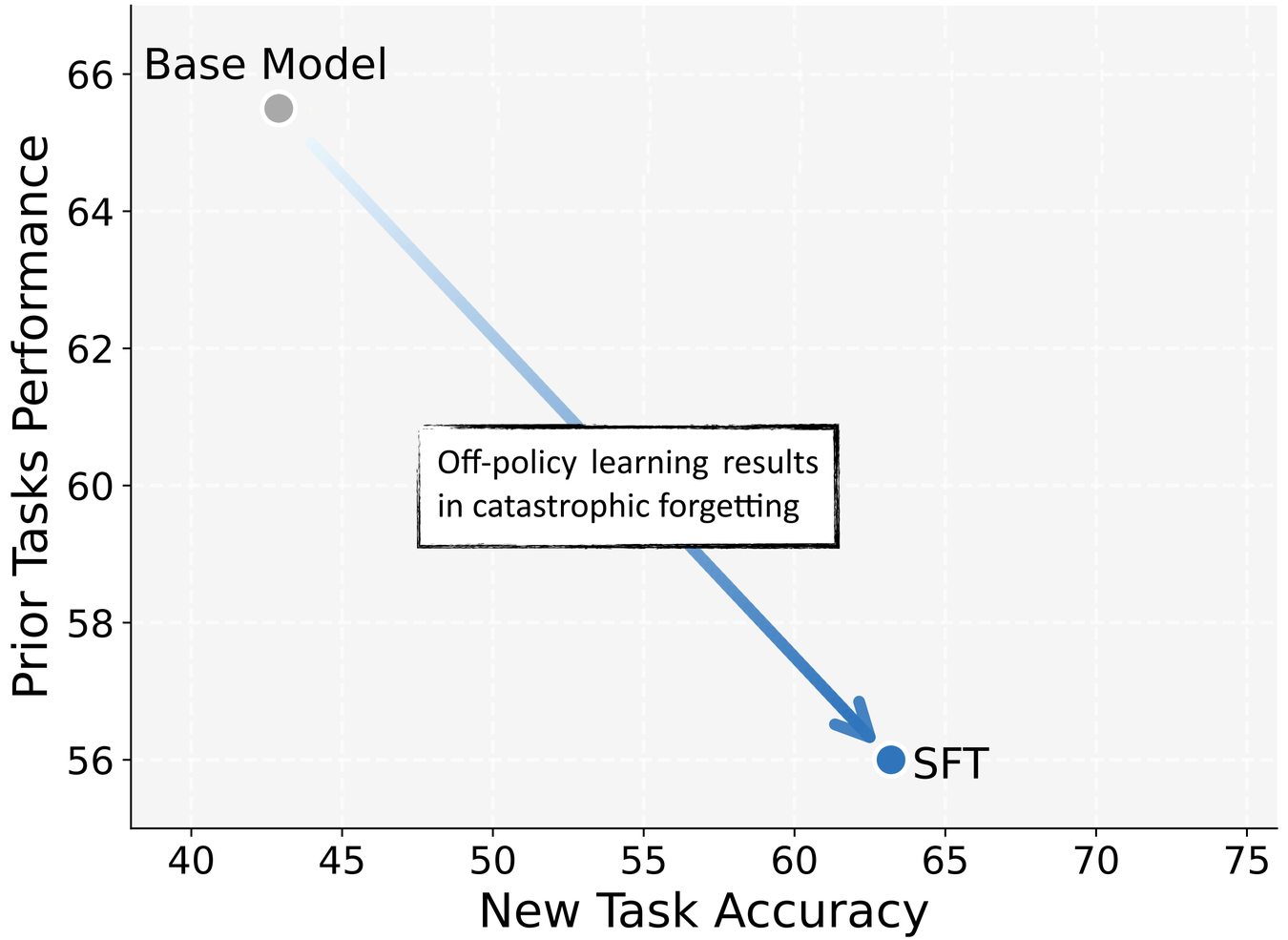




$$y_2 \sim \pi(\cdot \mid x, y_1) \longleftrightarrow y_2 \sim \pi(\cdot \mid x, y_1, c)$$

Learning from Expert Demonstrations

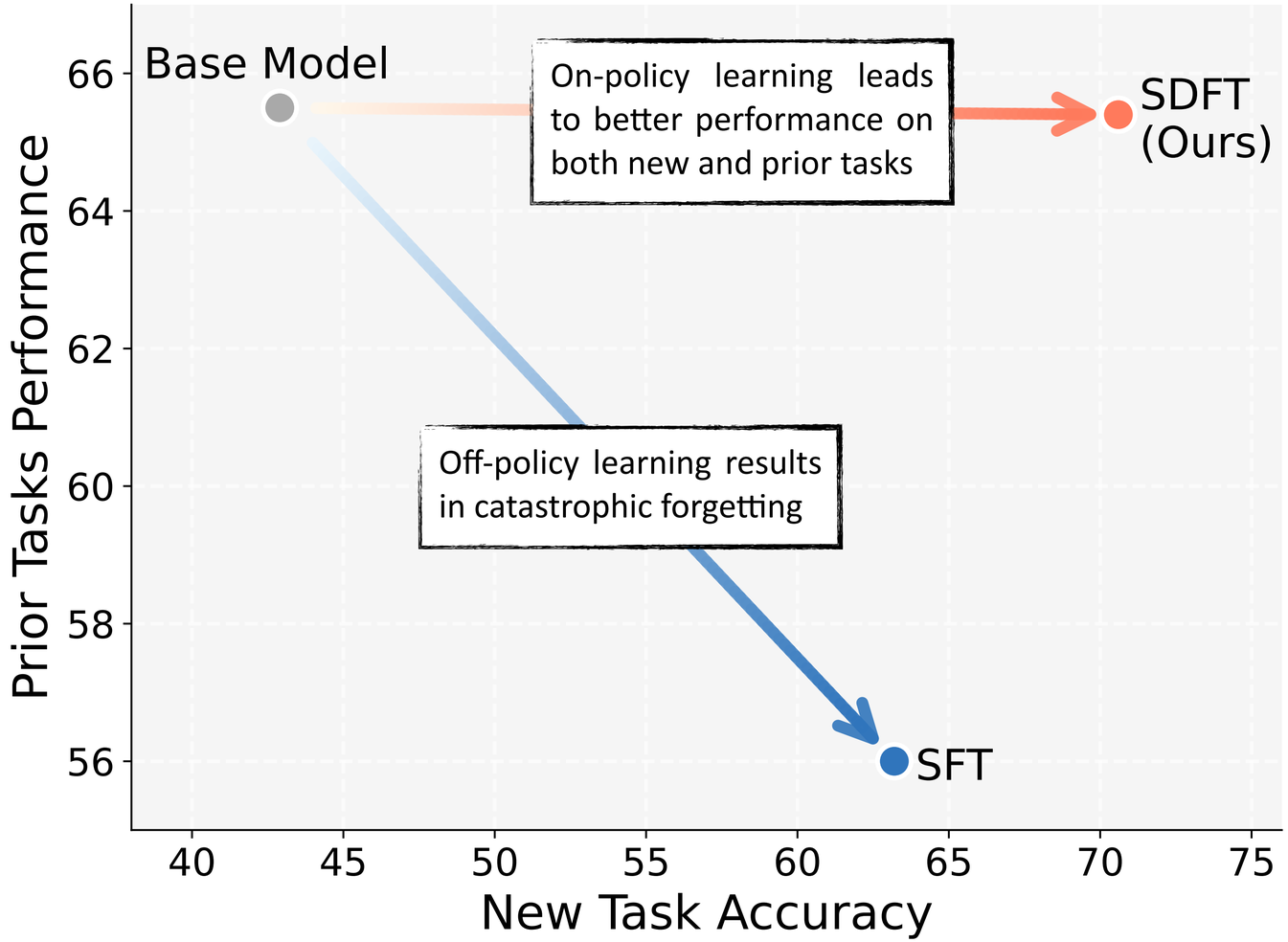
Qwen2.5 7B-Instruct on ToolAlpaca dataset.



Shenfeld et al. 2026, *Self-Distillation Enables Continual Learning.*

Learning from Expert Demonstrations

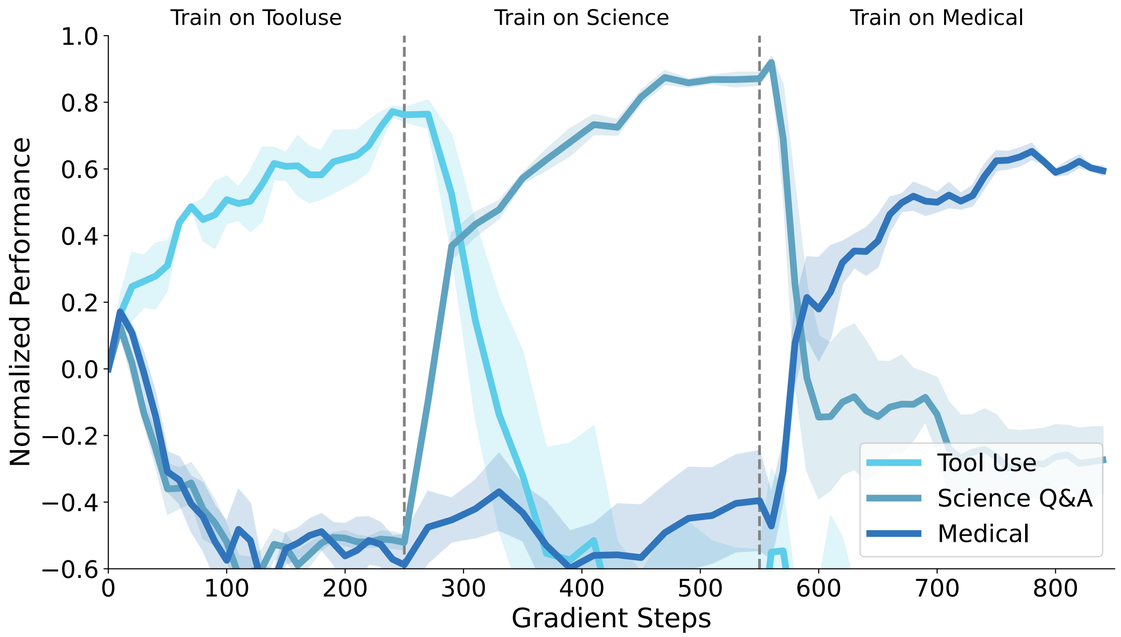
Qwen2.5 7B-Instruct on ToolAlpaca dataset.



Shenfeld et al. 2026, *Self-Distillation Enables Continual Learning.*

Learning from Expert Demonstrations

SFT

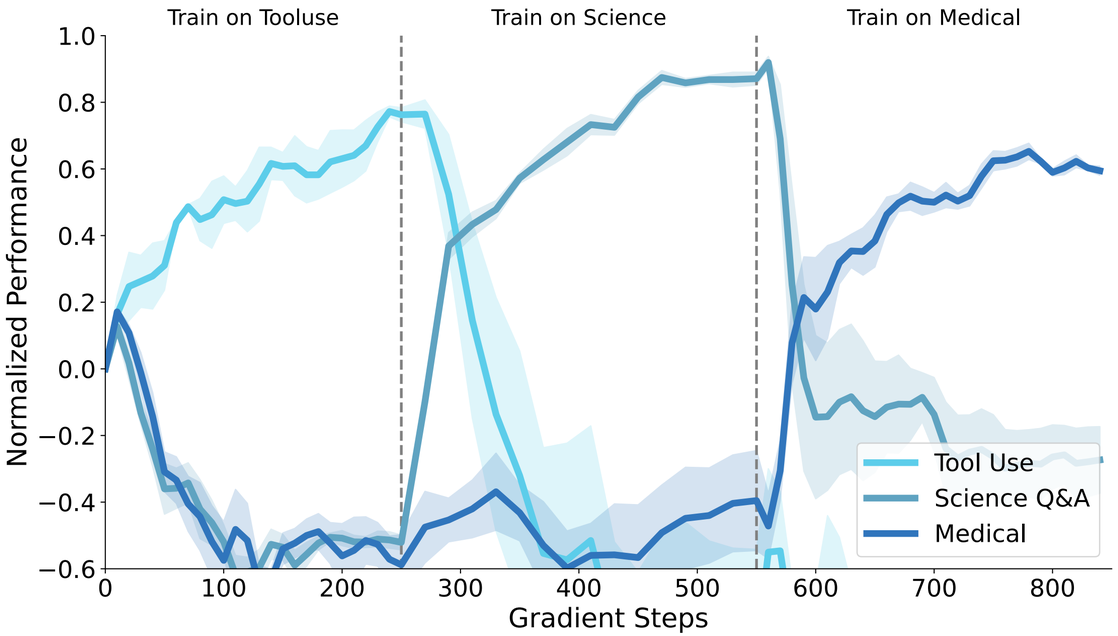


Qwen2.5 7B-Instruct trained sequentially on ToolAlpaca, ScieKnowEval Chemistry and Clinical reasoning questions from HuatuoGPT-o1 dataset.

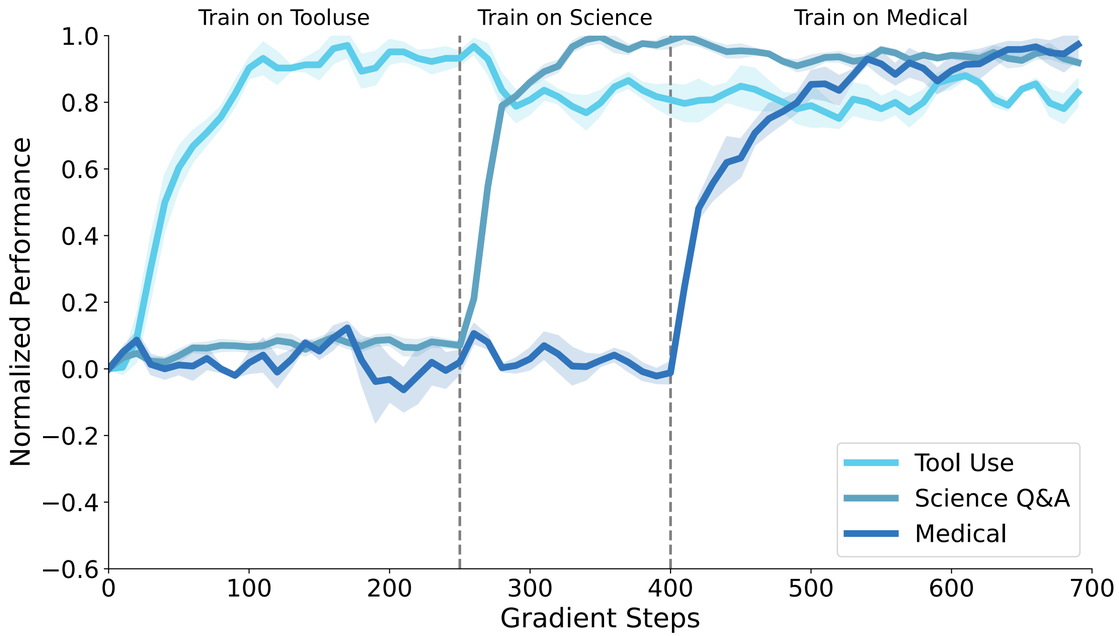
Shenfeld et al. 2026, *Self-Distillation Enables Continual Learning.*

Learning from Expert Demonstrations

SFT



Self-Distillation



Qwen2.5 7B-Instruct trained sequentially on ToolAlpaca, ScieKnowEval Chemistry and Clinical reasoning questions from HuatuoGPT-o1 dataset.

Shenfeld et al. 2026, *Self-Distillation Enables Continual Learning.*

Questions?

