unsloth

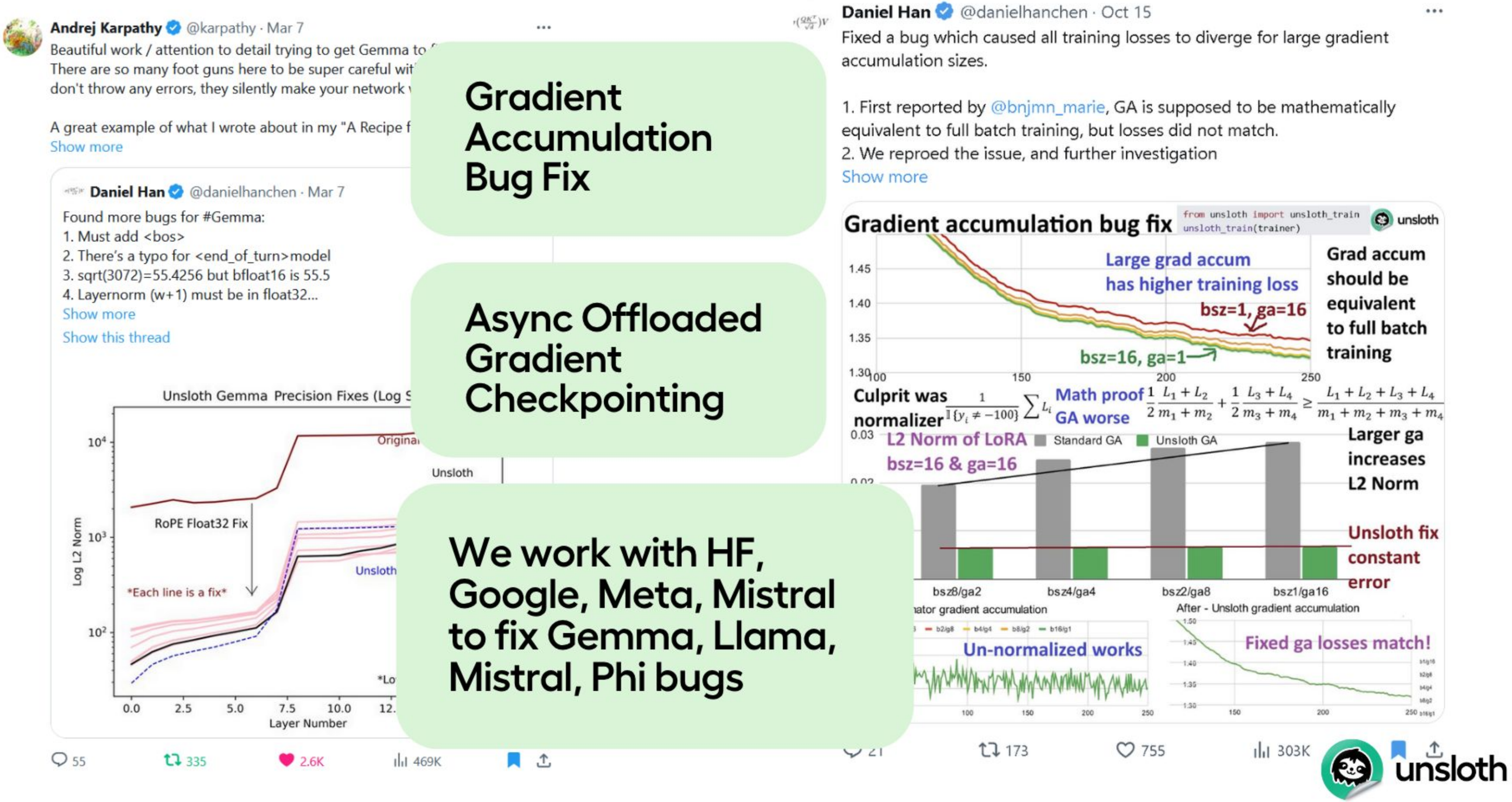
Deep-Dive: RL, Kernels, Agents & Quantization

www.unsloth.ai

Daniel Han

Co-founder

Access the slides


Andrej Karpathy @karpathy · Mar 7
Beautiful work / attention to detail trying to get Gemma to
There are so many foot guns here to be super careful wit
don't throw any errors, they silently make your network
A great example of what I wrote about in my "A Recipe f
Show more
Daniel Han @danielhanchen · Mar 7
Found more bugs for #Gemma:
1. Must add <bos>
2. There's a typo for <end_of_turn>model
3. sqrt(3072)=55.4256 but bfloat16 is 55.5
4. Layernorm (w+1) must be in float32...
Show more
Show this thread
Unsloth Gemma Precision Fixes (Log S
Original
Unsloth
RoPE Float32 Fix
Each line is a fix
Log L2 Norm
Layer Number
55
335
2.6K
469K
Gradient Accumulation Bug Fix
Async Offloaded Gradient Checkpointing
We work with HF, Google, Meta, Mistral to fix Gemma, Llama, Mistral, Phi bugs
Daniel Han @danielhanchen · Oct 15
Fixed a bug which caused all training losses to diverge for large gradient accumulation sizes.
1. First reported by @bnjmn_marie, GA is supposed to be mathematically equivalent to full batch training, but losses did not match.
2. We reproed the issue, and further investigation
Show more
Gradient accumulation bug fix
from unsloth import unsloth_train
unsloth_train(trainer)
unsloth
Large grad accum has higher training loss
bsz=1, ga=16
bsz=16, ga=1
Grad accum should be equivalent to full batch training
Culprit was normalizer
Math proof GA worse
L2 Norm of LoRA bsz=16 & ga=16
Standard GA
Unsloth GA
Larger ga increases L2 Norm
Unsloth fix constant error
bsz8/ga2
bsz4/ga4
bsz2/ga8
bsz1/ga16
After - Unsloth gradient accumulation
Un-normalized works
Fixed ga losses match!
173
755
303K
unsloth

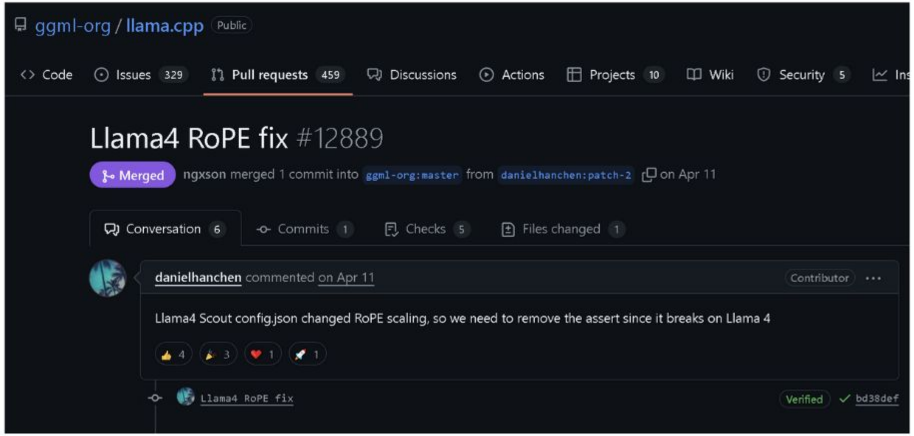
Open-source contributions

Simon Willison's Weblog

Phi-4 Bug Fixes by Unsloth (via) This explains why I was seeing weird `<|im_end|>` suffexes during my experiments with Phi-4 the other day: it turns out the Phi-4 tokenizer definition as released by Microsoft had a bug in it, and there was a small bug in the chat template as well.

Daniel and Michael Han figured this out and have now published GGUF files with their fixes on Hugging Face.

Posted 11th January 2025 at 1:20 am



Phi-3 + 4 Bug fixes Merged by Microsoft

Multiple Llama 4 fixes and contributions to llama.cpp

Recently worked with Qwen team on Qwen3 and Mistral on Devstral

unsloth

unsloth

10 million+

monthly downloads

10M+

monthly downloads on 🤗Hugging Face

40K

stars on GitHub

unslothai / unsloth

Type / to search

Code · Issues 959 · Pull requests 69 · Discussions · Actions · Wiki · Security 10 · Insights · Settings

unsloth Public

Sponsor · Edit Pins · Unwatch 234 · Fork 3.2k · Starred 39.9k

main · 11 Branches · 13 Tags · Go to file t · Add file · Code

danielhanchen Update rl.py ✓ 3340eaa · 7 hours ago · 1,324 Commits

.github	Update issue templates	3 days ago
images	Uploading HQ Unsloth Sticker	last month
tests	reroute merge logic language models + comprehensive tests...	11 hours ago
unsloth	Update rl.py	7 hours ago
.gitignore	MoE Kernel (#2465)	last month
CODE_OF_CONDUCT.md	Added missing code of conduct (#2416)	last month
CONTRIBUTING.md	Added missing code of conduct (#2416)	last month
LICENSE	Auto Healing Tokenizer (#283)	last year
README.md	Update README.md	last week
pyproject.toml	Bug fixes (#2651)	5 days ago
unsloth-cli.py	Bug fixes (#1516)	5 months ago

About

Finetune Qwen3, Llama 4, TTS, DeepSeek-R1 & Gemma 3 LLMs 2x faster with 70% less memory! 🦥

unsloth.ai

text-to-speech · tts · llama · lora · gemma · mistral · fine-tuning · finetuning · llm · llms · qlora · qwen · deepseek · unsloth · llama3 · deepseek-r1 · gemma3 · qwen3 · llama4 · llama-4

- Readme
- Apache-2.0 license
- Code of conduct
- Activity
- Custom properties
- 39.9k stars
- 234 watching
- 3.2k forks

Unsloth AI *Enterprise* Company Verified

https://unsloth.ai · unslothai · unslothai

Hugging Face

Activity Feed · Follow 5,809

AI & ML interests

Hey! We're focusing on making AI more accessible to everyone!

Dynamic 2.0 quants

DeepSeek R1-0528

Expect Bug fixes + updates constantly for chat templates, tokenizers etc.

GGUF, 4-bit, original

Collections 21

Collapse

DeepSeek R1 (All Versions)

DeepSeek-R1-0528 is here! The most powerful reasoning open L...

- unsloth/DeepSeek-R1-0528-GGUF
  Text Generation • Updated about 9 h... • 51.2k • 124
- unsloth/DeepSeek-R1-0528-Qwen3-8B-GGUF
  Text Generation • Updated 2 days ago • 116k • 116
- unsloth/DeepSeek-R1-0528-Qwen3-8B-unslot...
  Text Generation • Updated 5 days ago • 1.78k • 2
- unsloth/DeepSeek-R1-0528

Unsloth Dynamic 2.0 Quants

New 2.0 version of our Dynamic GGUF + Quants. Dynamic 2.0 ac...

- unsloth/DeepSeek-R1-0528-GGUF
  Text Generation • Updated about 9 h... • 51.2k • 124
- unsloth/DeepSeek-R1-0528-Qwen3-8B-GGUF
  Text Generation • Updated 2 days ago • 116k • 116
- unsloth/Devstral-Small-2505-GGUF
  Text2Text Generation • Updated 9 days ... • 113k • 80
- unsloth/Phi-4-reasoning-plus-GGUF

Qwen3

Qwen's new Qwen3 models. In Unsloth Dynamic 2.0, GGUF, 4-bi...

- unsloth/Qwen3-30B-A3B-GGUF
  Text Generation • Updated 4 days ago • 168k • 193
- unsloth/Qwen3-32B-GGUF
  Text Generation • Updated 10 days ago • 69k • 63
- unsloth/Qwen3-235B-A22B-GGUF
  Text Generation • Updated 10 days ago • 45k • 53

Llama 4

Meta's new Llama 4 multimodal models, Scout & Maverick. Incl...

- unsloth/Llama-4-Scout-17B-16E-Instruct-G...
  Image-Text-to-Text • Updated 5 days ago • 108k • 87
- unsloth/Llama-4-Maverick-17B-128E-Instru...
  Image-Text-to-Text • Updated 6 days ago • 43.1k • 23
- unsloth/Llama-4-Scout-17B-16E-Instruct
  Image-Text-to-Text • Updated 7 days ago • 2.99k • 55

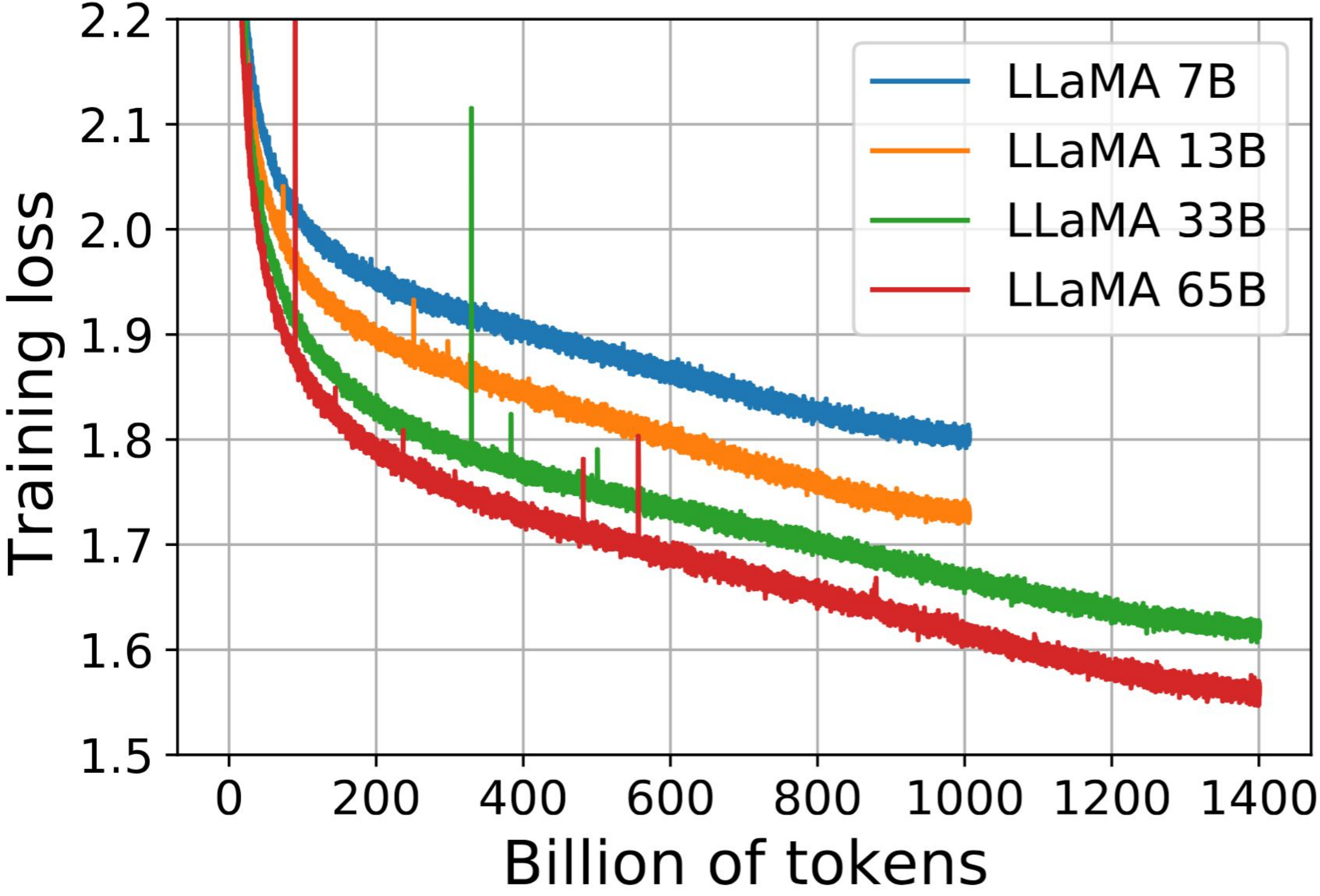
https://arxiv.org/pdf/2302.13971

LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron[*], Thibaut Lavril[*], Gautier Izacard[*], Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave[*], Guillaume Lample[*]

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: **Model sizes, architectures, and optimization hyper-parameters.**



LLaMA 7B
LLaMA 13B
LLaMA 33B
LLaMA 65B
Training loss
Billion of tokens
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
0
200
400
600
800
1000
1200
1400

google/gemma-3-27b-it

Image-Text-to-Text • Updated Mar 21 • 382k • • 1.41k

google/gemma-3-27b-pt

Image-Text-to-Text • Updated Mar 21 • 18.8k • 93

14 trillion tokens

meta-llama/Llama-4-Scout-17B-16E-Instruct

Image-Text-to-Text • Updated 11 days ago • 287k • • 933

meta-llama/Llama-4-Scout-17B-16E

Image-Text-to-Text • Updated Apr 9 • 32.7k • 173

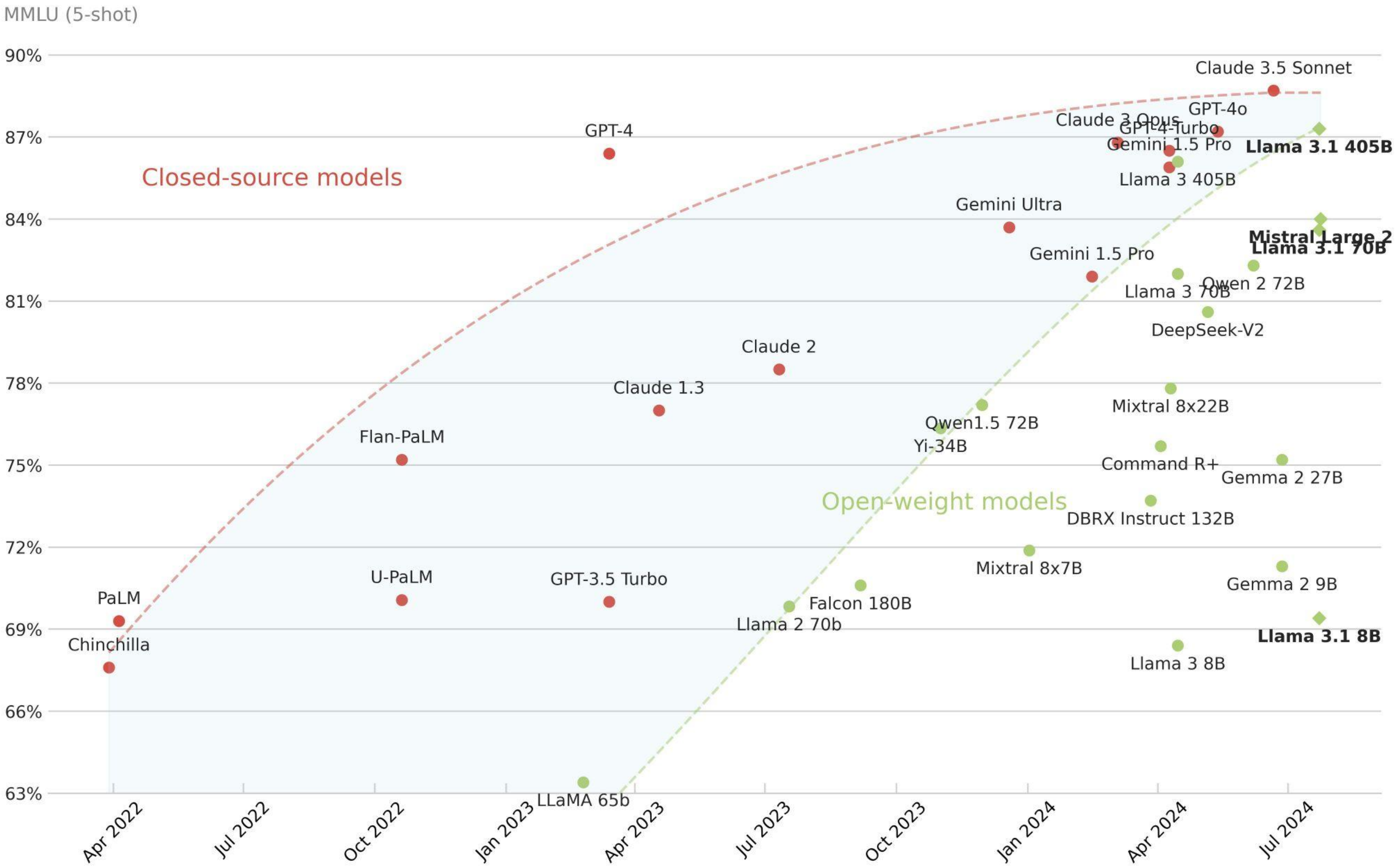
30 trillion tokens

Access the slides

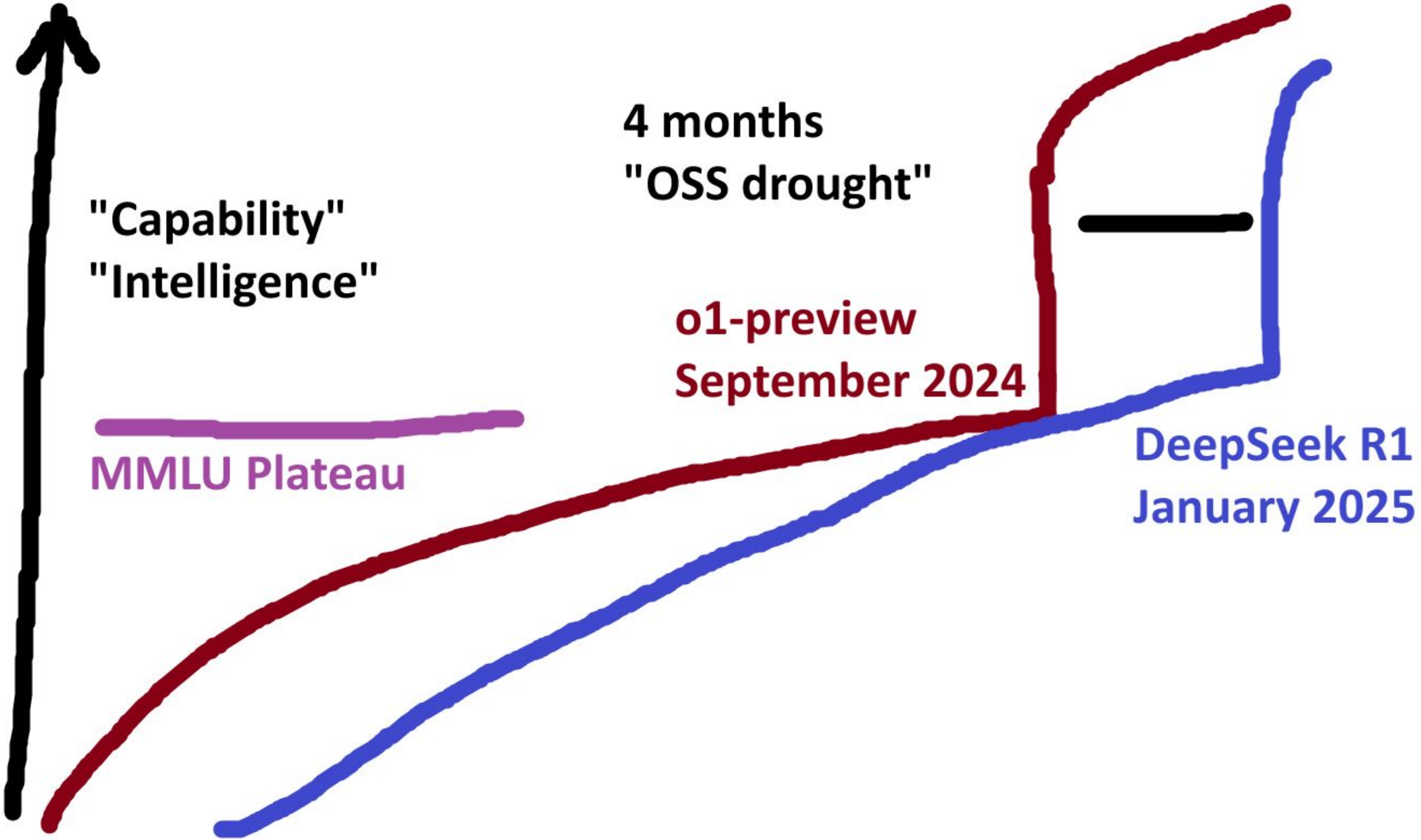
https://docs.unsloth.ai/ai-engineers-2025

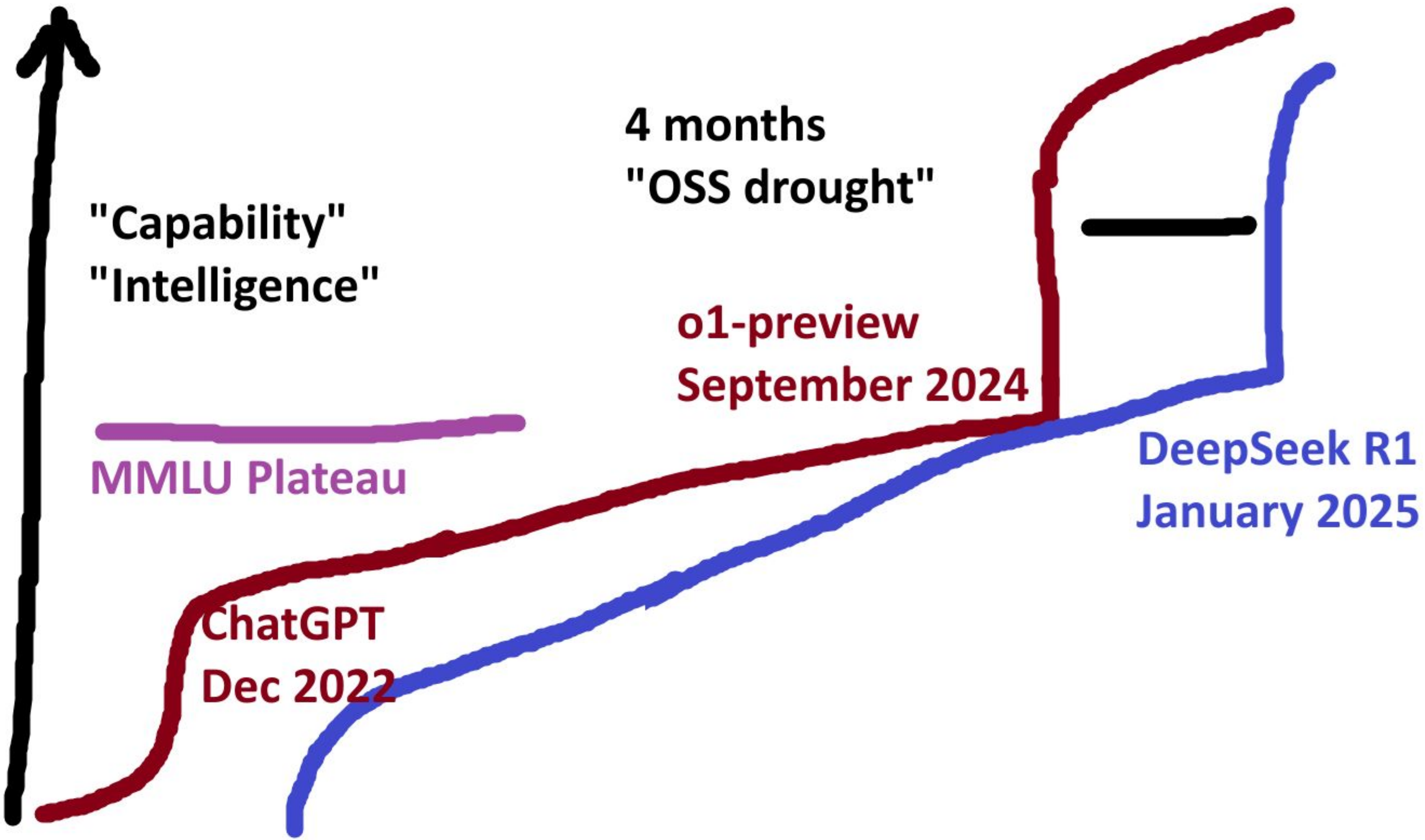
Closed-source vs. open-weight models

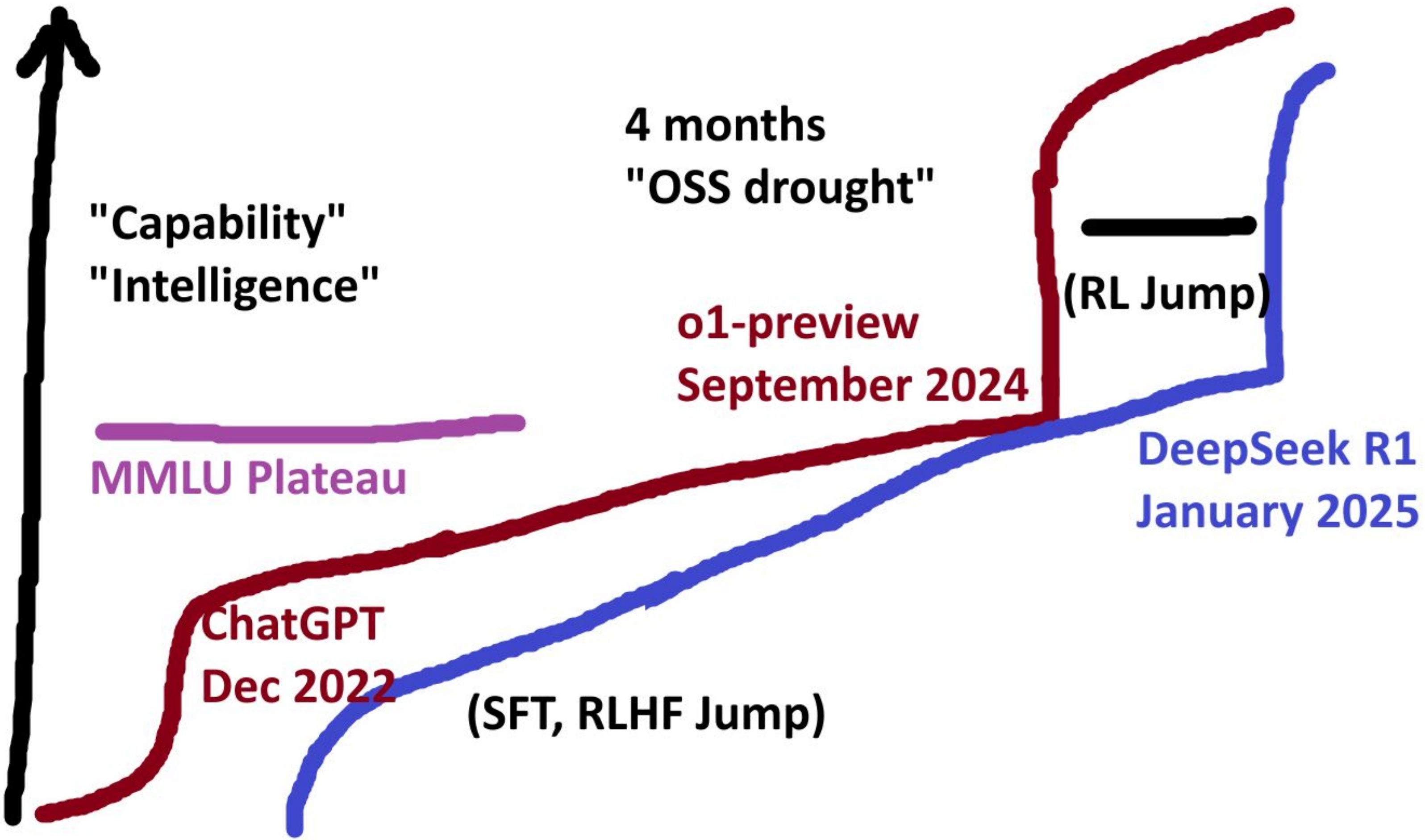
Llama 3.1 405B closes the gap with closed-source models for the first time in history.



https://x.com/maximelabonne/status/1816416043511808259


4 months
"OSS drought"
"Capability"
"Intelligence"
o1-preview
September 2024
MMLU Plateau
DeepSeek R1
January 2025


"Capability"
"Intelligence"
MMLU Plateau
ChatGPT
Dec 2022
4 months
"OSS drought"
o1-preview
September 2024
DeepSeek R1
January 2025



4 months
"OSS drought"
"Capability"
"Intelligence"
(RL Jump)
o1-preview
September 2024
MMLU Plateau
DeepSeek R1
January 2025
ChatGPT
Dec 2022
(SFT, RLHF Jump)

How Much Information Does the Machine Need to Predict?

Y LeCun

Reinforcement Learning (cherry)

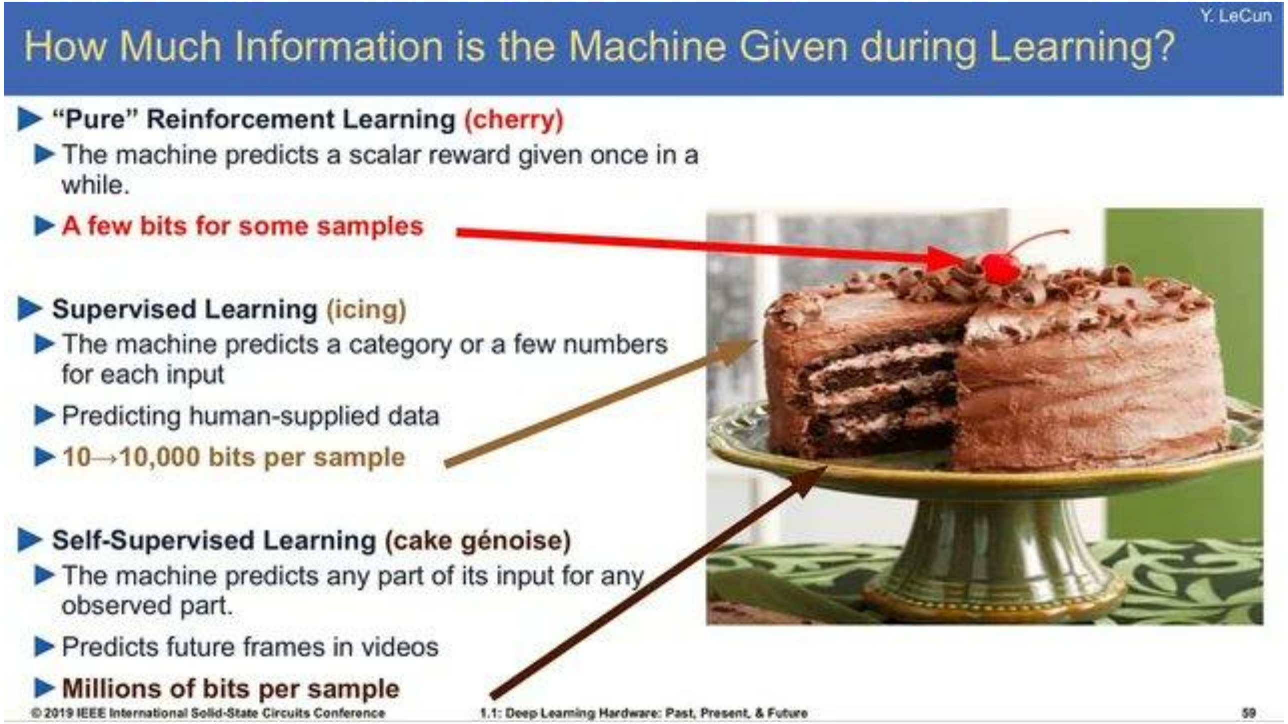
- The machine predicts a scalar reward given once in a while.
- **A few bits for some samples**

Supervised Learning (icing)

- The machine predicts a category or a few numbers for each input
- **10→10,000 bits per sample**

Unsupervised Learning (cake)

- The machine predicts any part of its input for any observed part.
- Predicts future frames in videos
- **Millions of bits per sample**



Deep Learning and the Future of AI September 2016!!!

https://youtu.be/_1Cyyt-4-n8?si=UdDAB5bkhqJs68pz&t=3717

Training Stages

GPT 4 Base
Claude 4 Base
Gemini Base

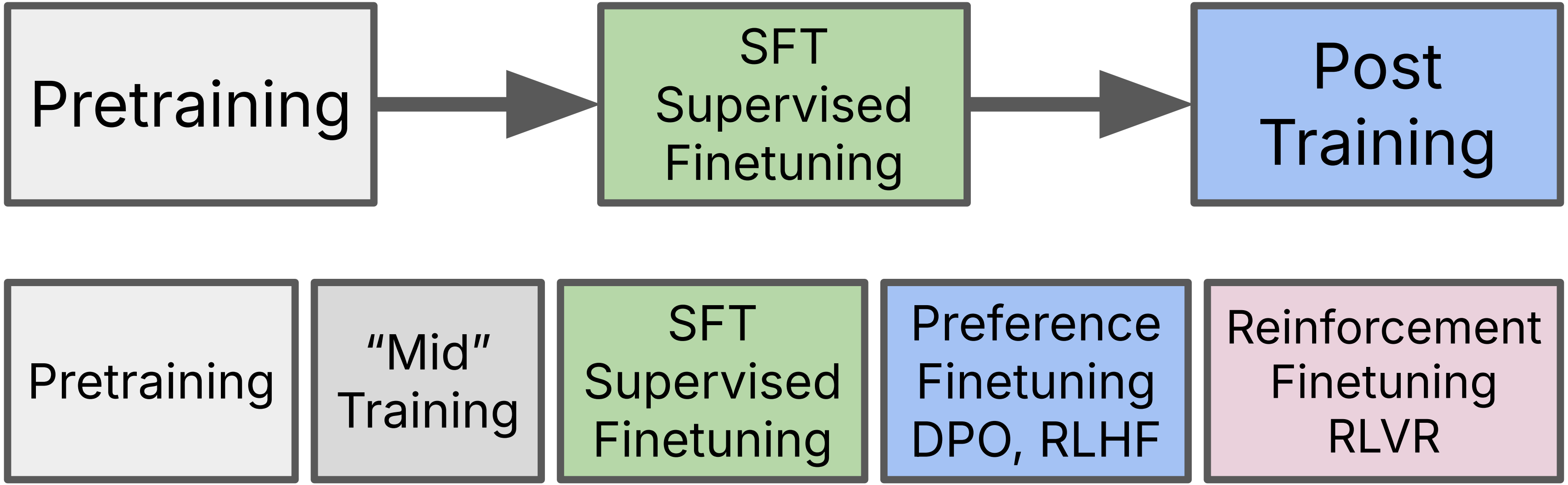
GPT 4 / ChatGPT 4
Claude 4 Opus
Gemini 2.5 Pro

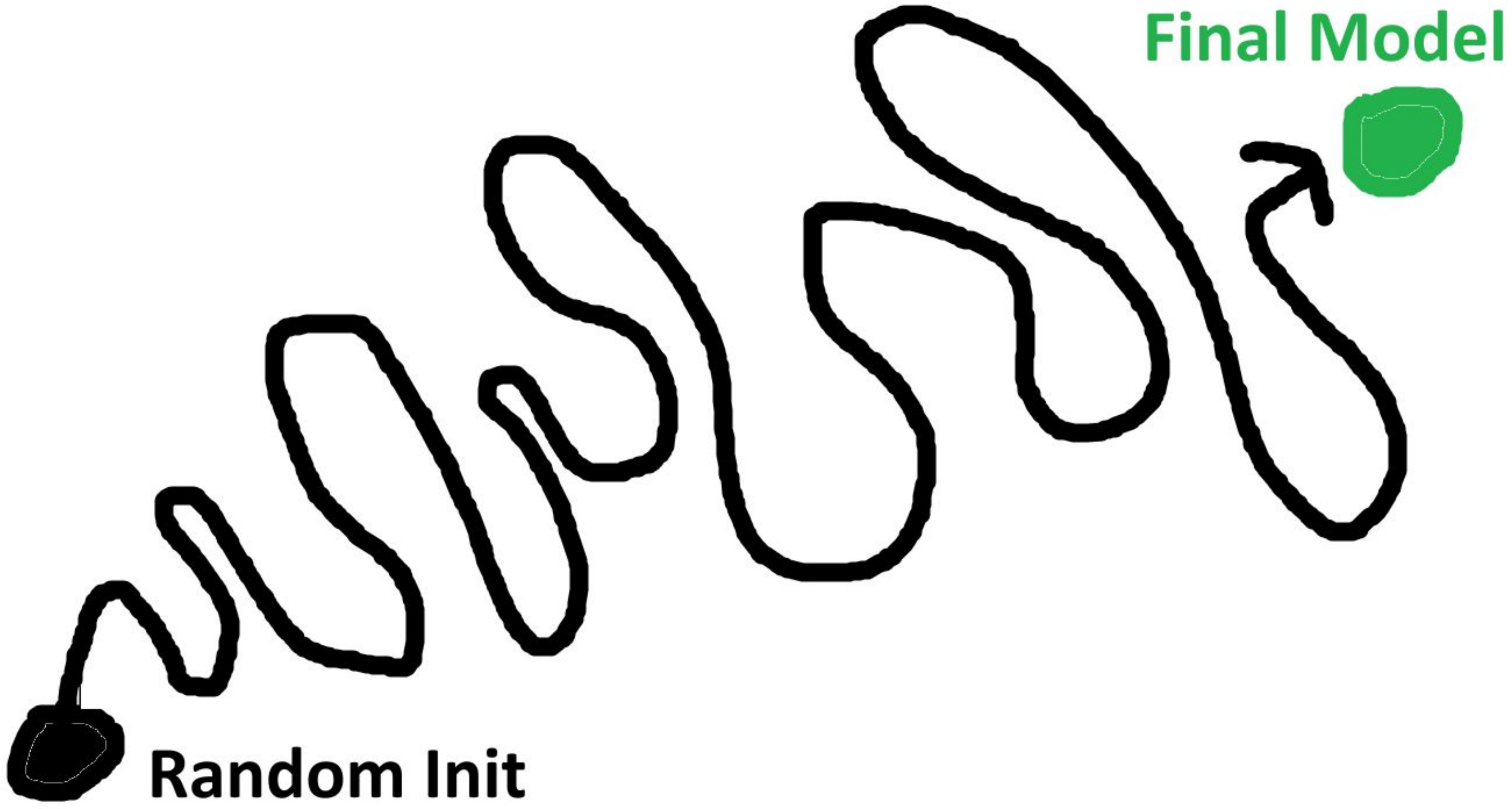
Training Stages

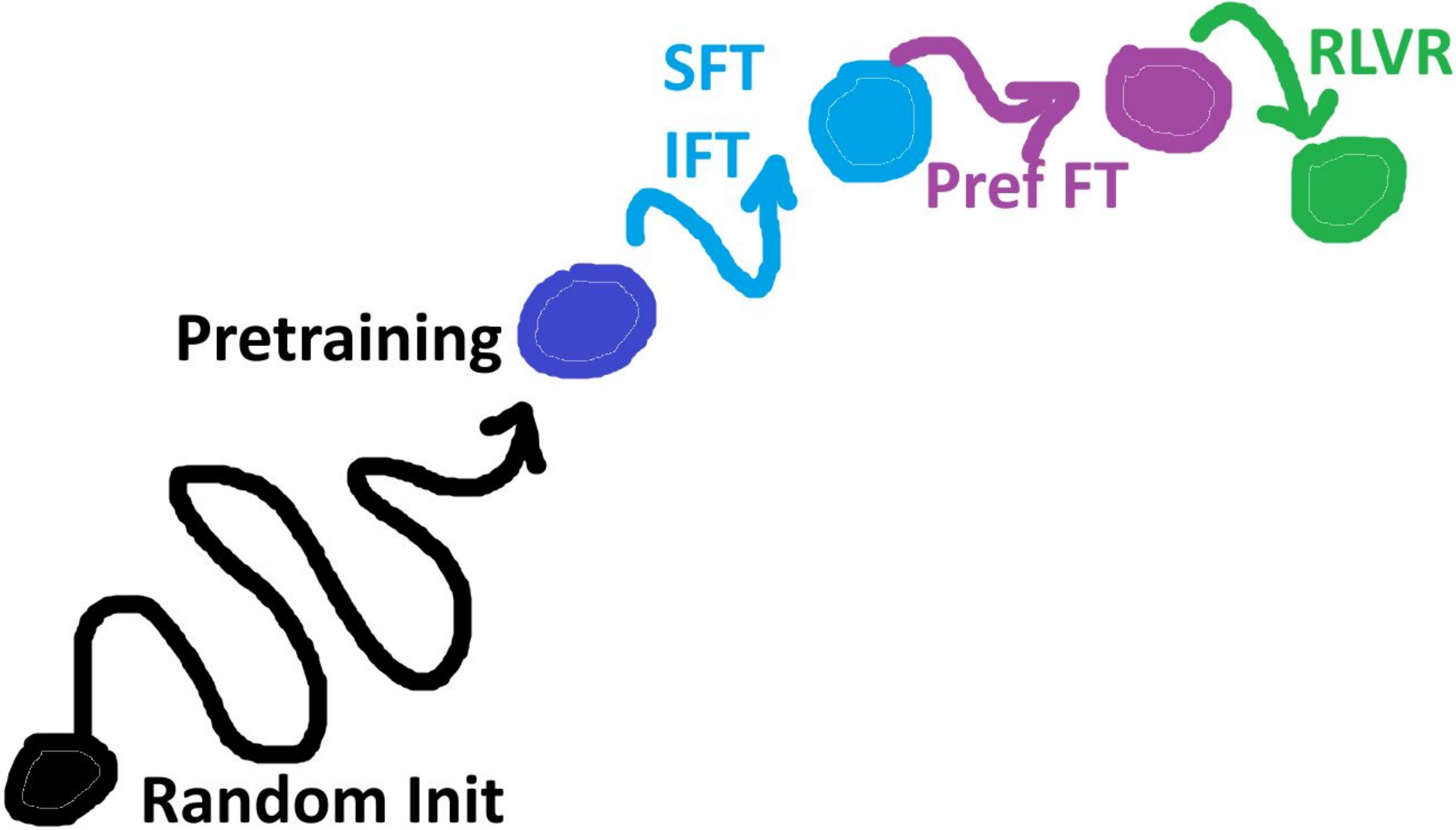
Gemma 3 PT
Llama 4
Qwen 3 Base
Mistral Small Base
Llama 2

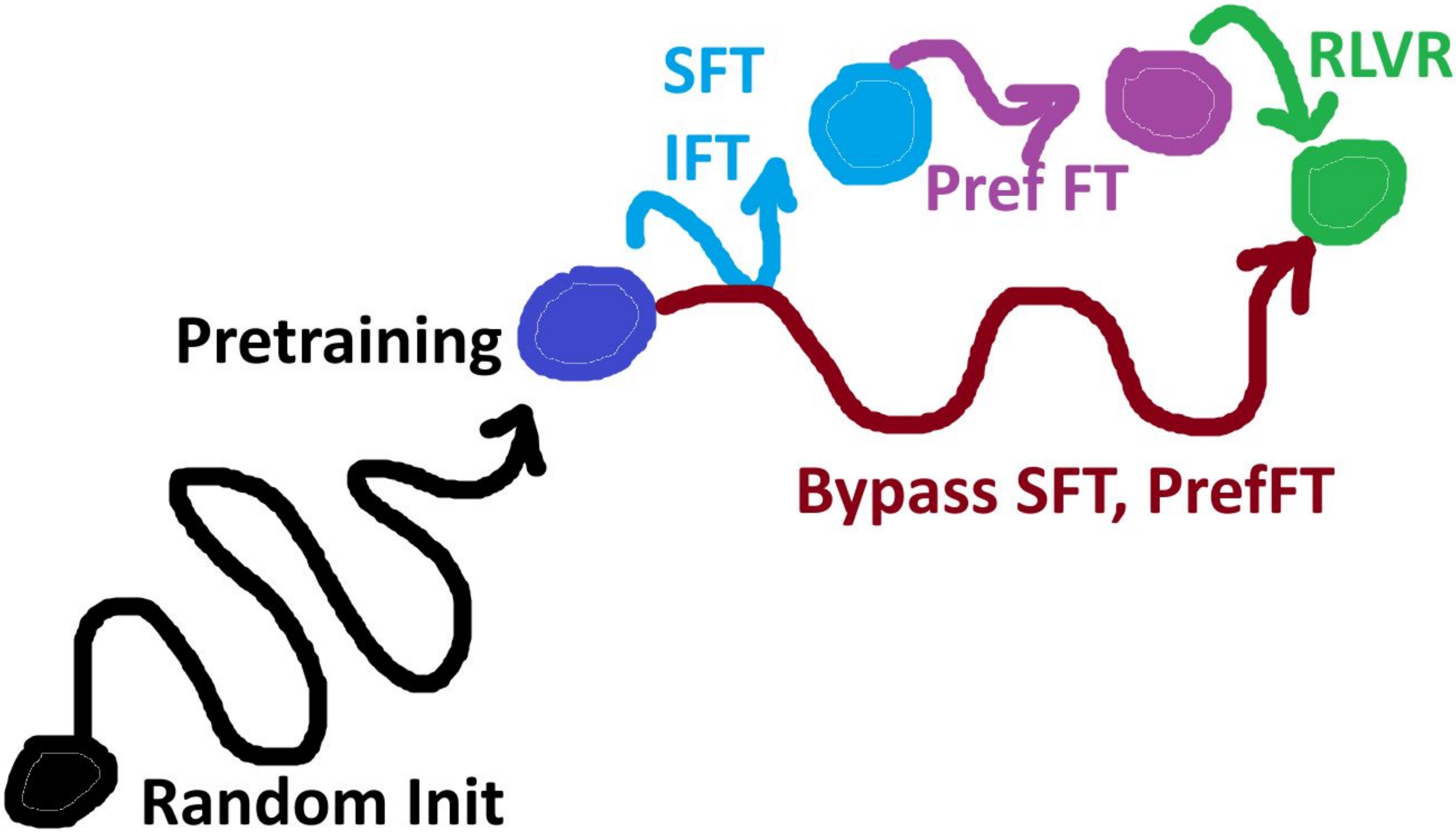
Gemma 3 IT
Llama 4 Instruct
Qwen 3
Mistral Small Instruct
Llama 2 Chat

Finetuning everywhere

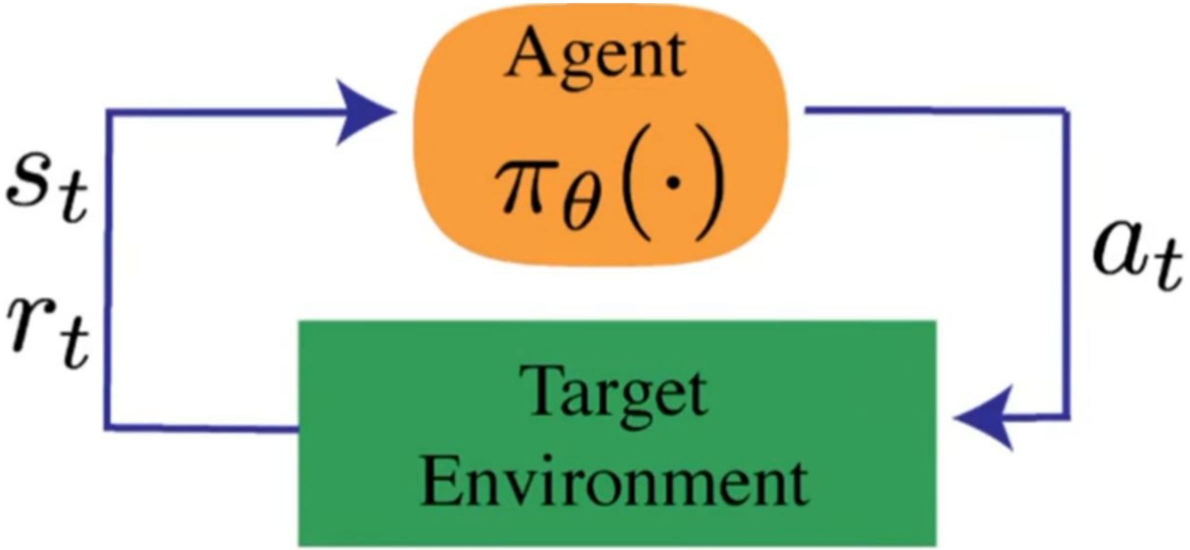



Final Model
Random Init


SFT
IFT
RLVR
Pref FT
Pretraining
Random Init


SFT
IFT
RLVR
Pref FT
Pretraining
Bypass SFT, PrefFT
Random Init

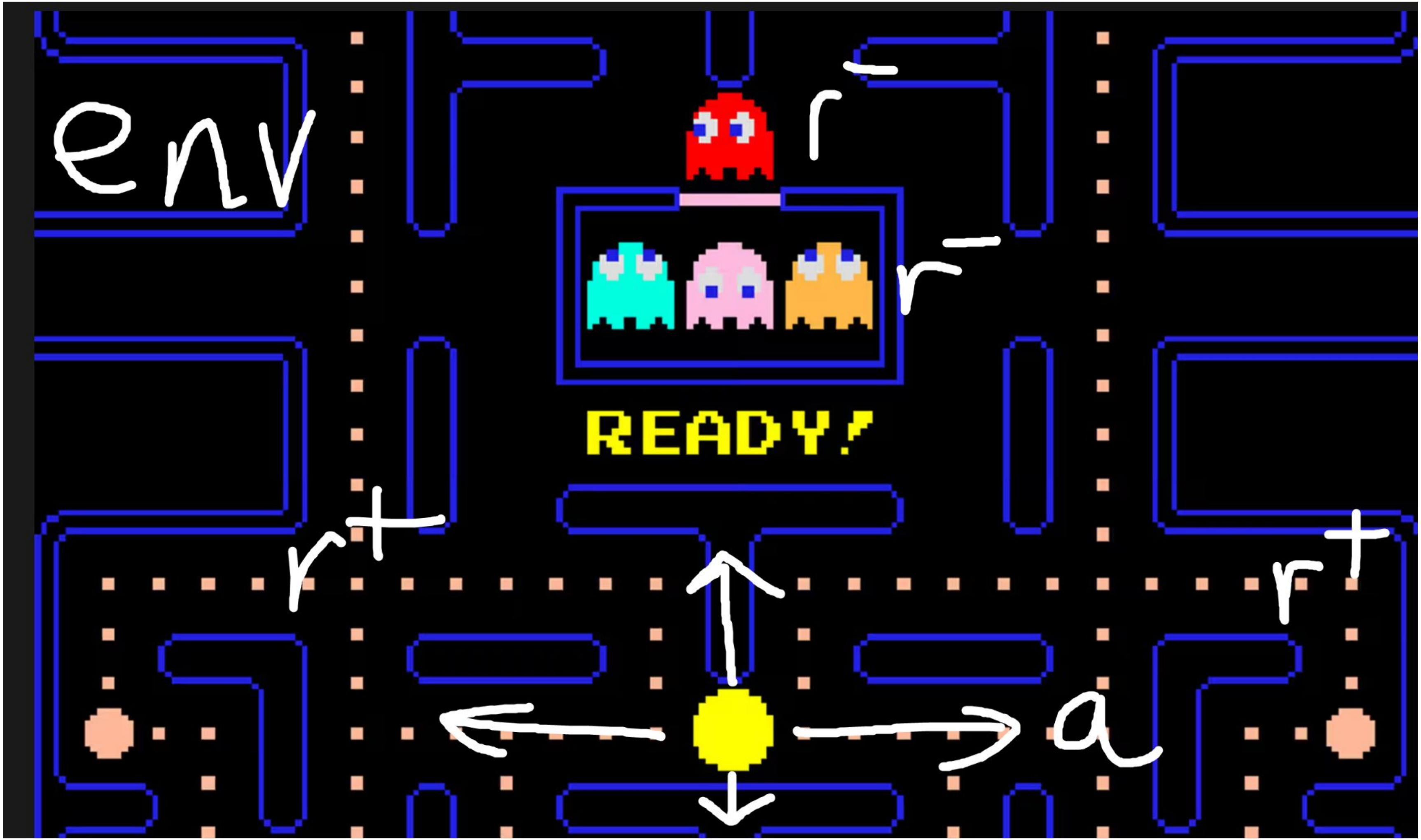
Agents in the old sense



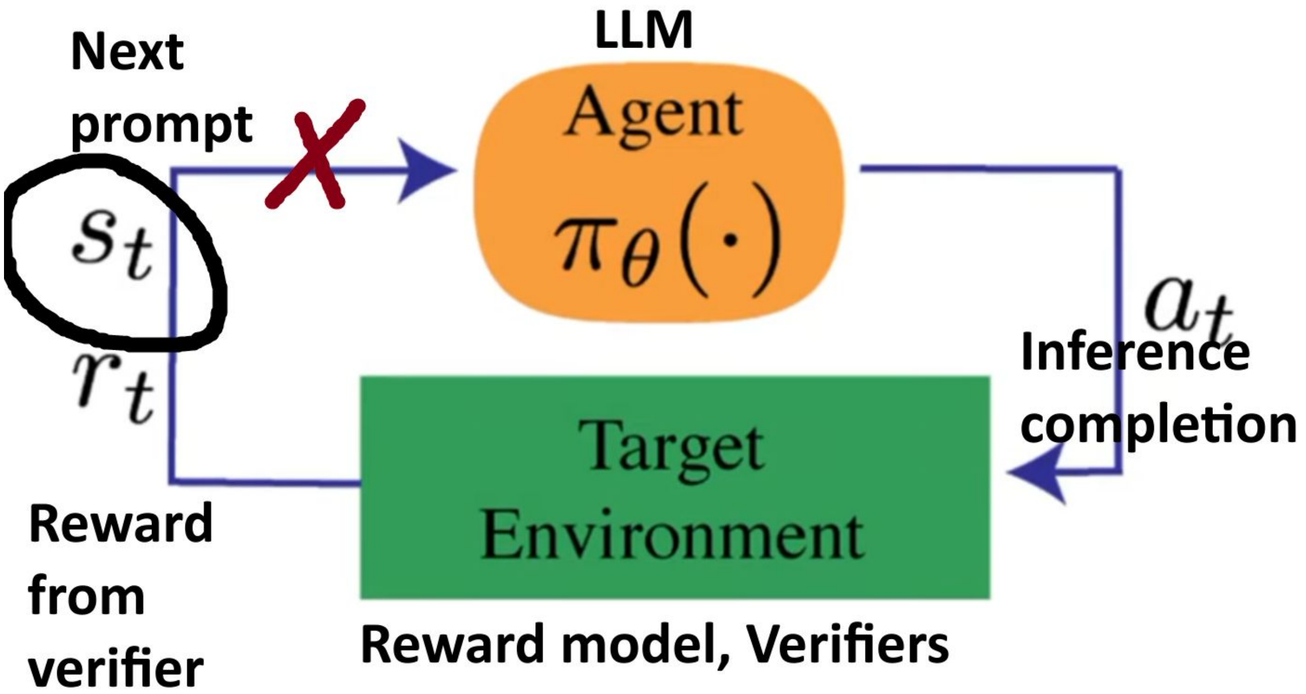
Some notation:

s_t : state
r_t : reward
a_t : action
$a_t \sim \pi_\theta(s_t)$: policy

https://youtu.be/zYeIqzULzr0?si=RawclaFMYwL7xPrU **Experimenting with Reinforcement Learning with Verifiable Rewards (RLVR) Nathan Lambert**


env
r^-
r^-
READY!
r^+
r^+
a

Reinforcement Learning for LLMs



0 7 8 A D

1

s 2 -10 E

What is 2+2?

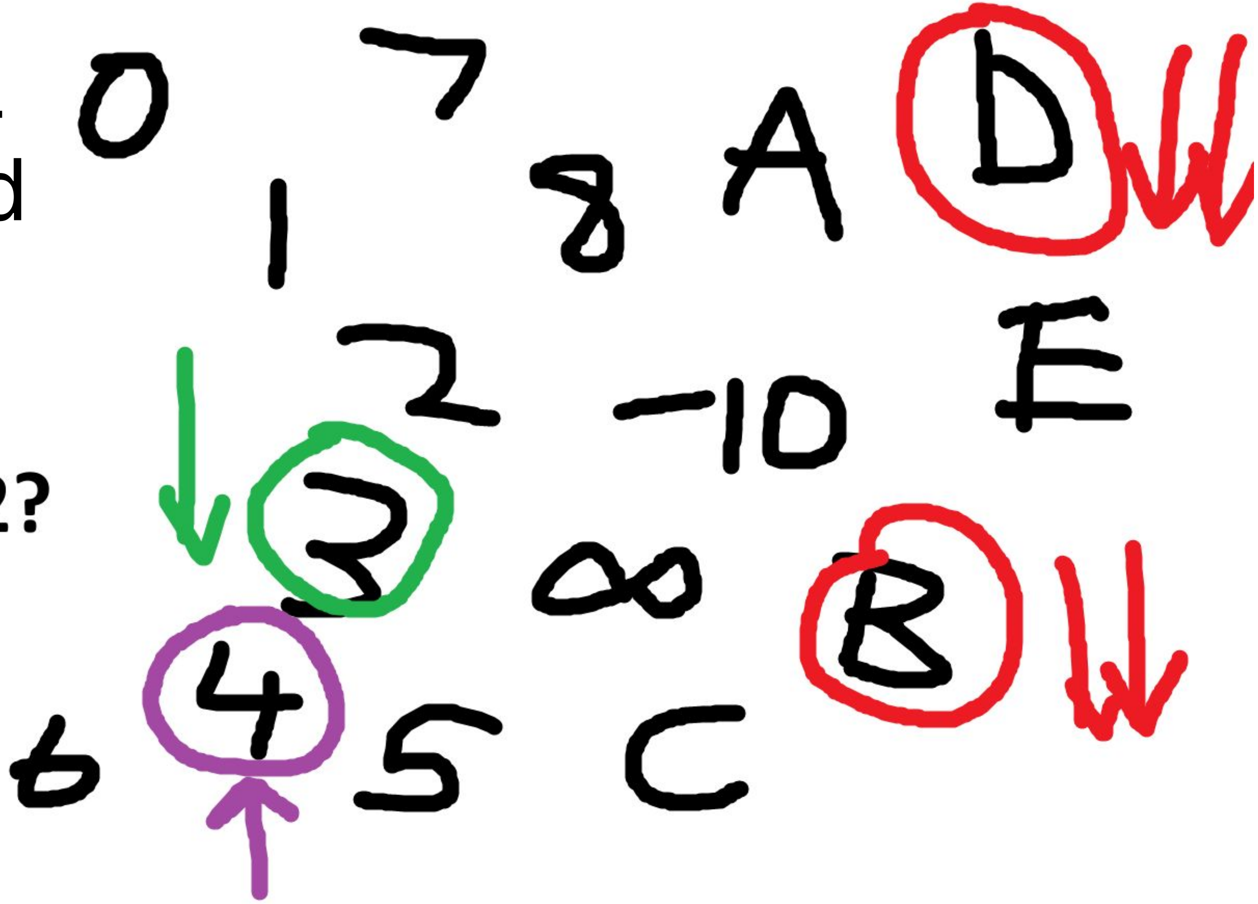
3 ∞ B

6 4 5 C a

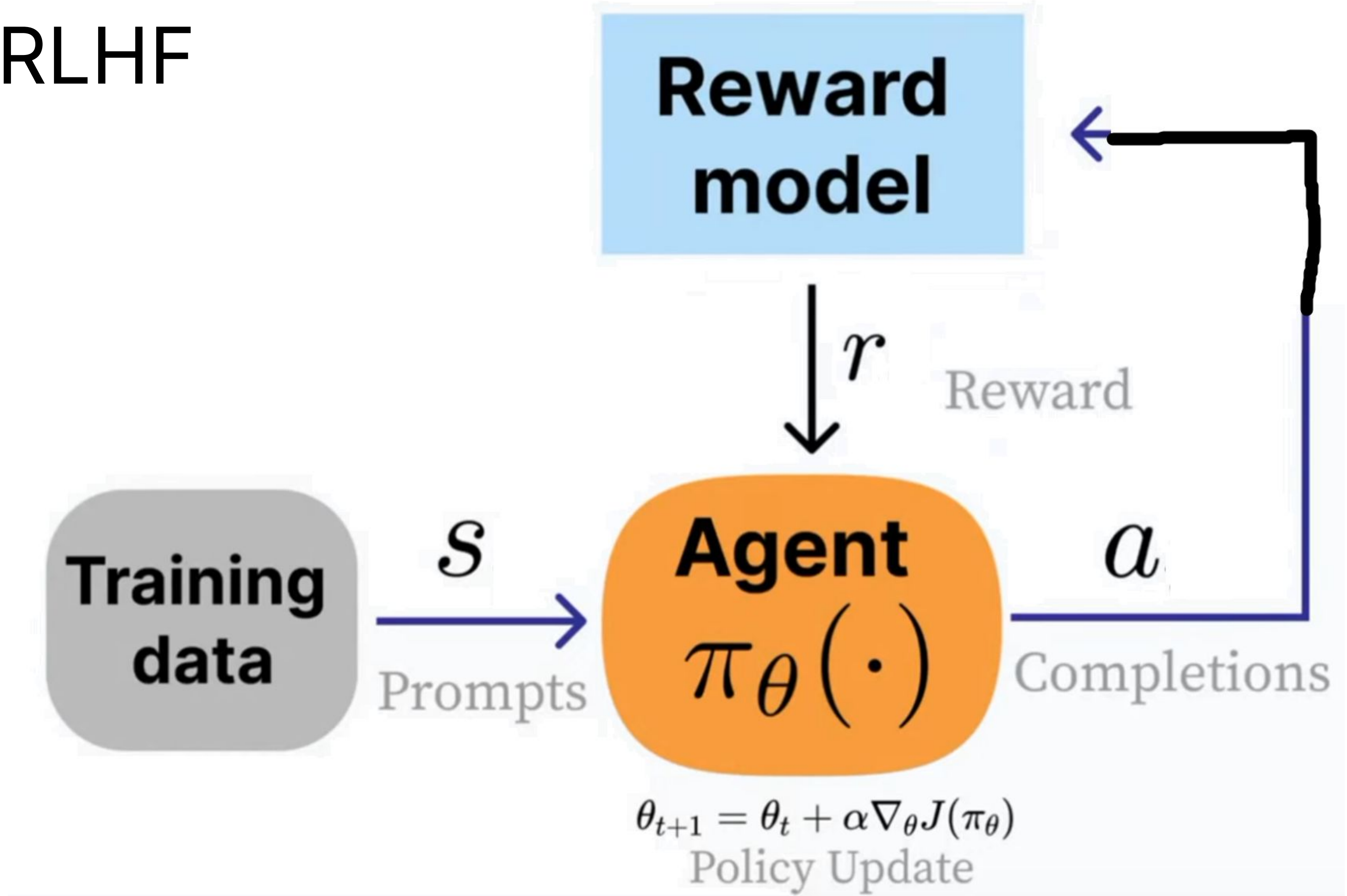
+1

Goal of RL
More good
Less bad

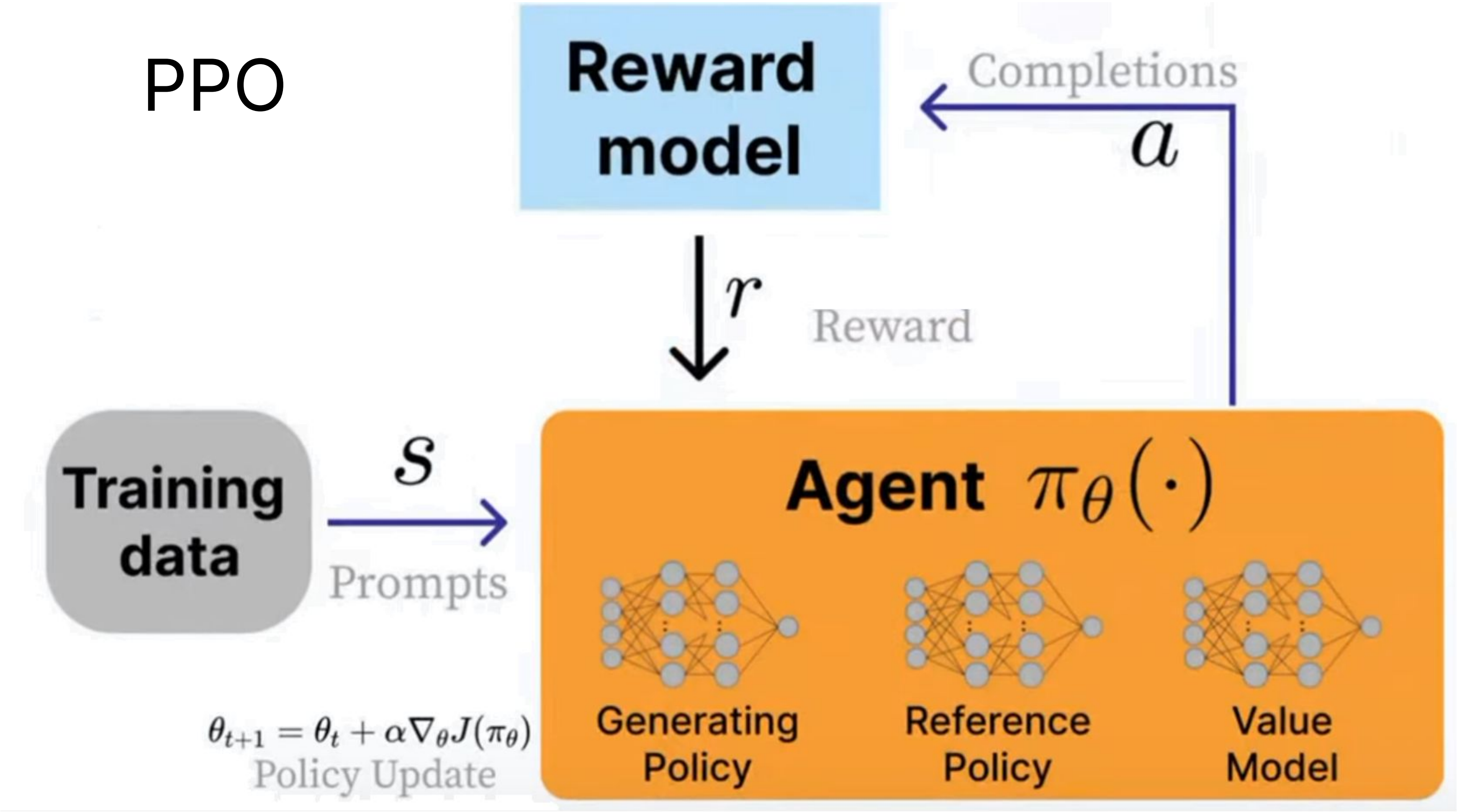
What is 2+2?



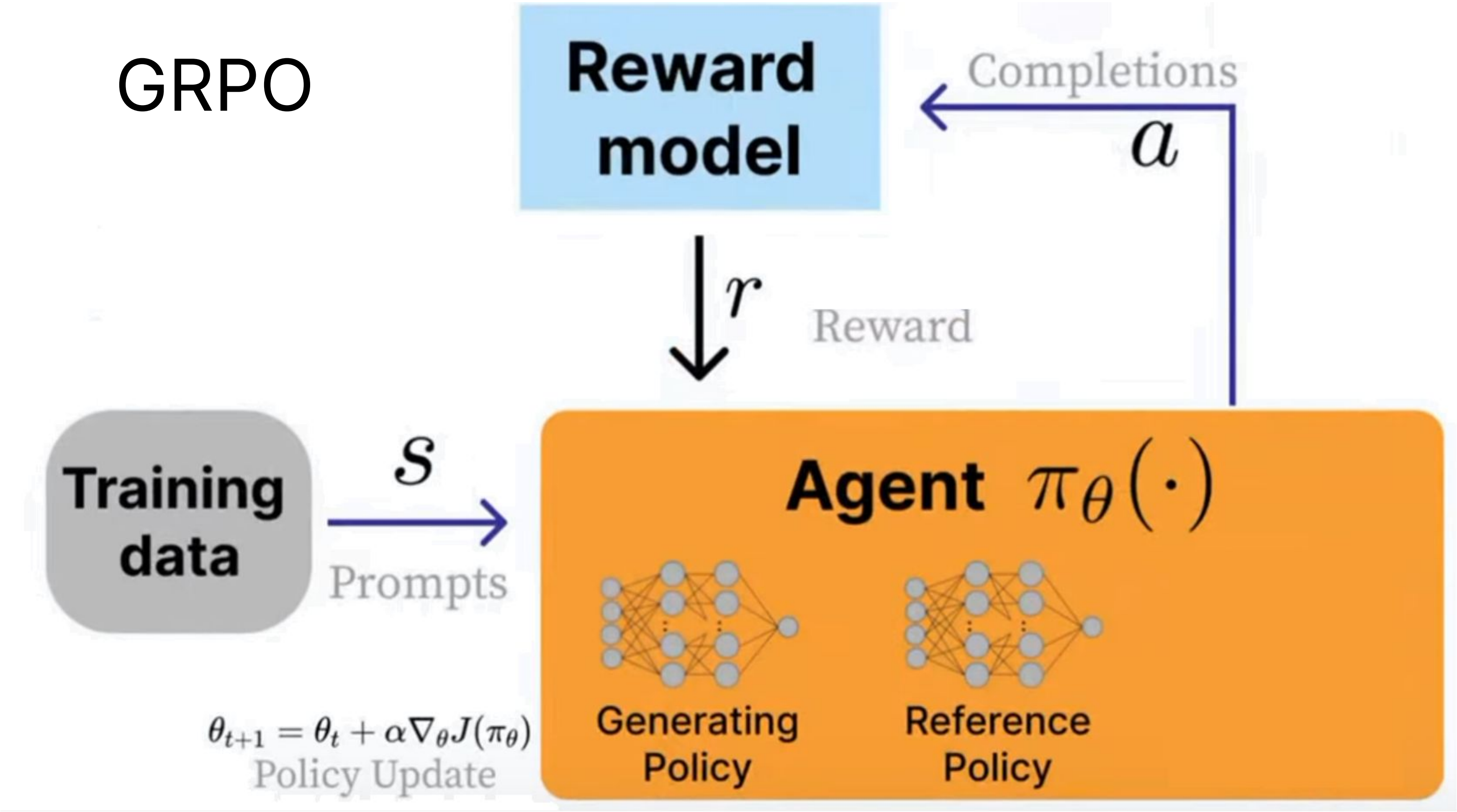
RLHF


Reward model
r
Reward
Training data
s
Prompts
Agent
$\pi_\theta(\cdot)$
a
Completions
$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\pi_\theta)$
Policy Update

PPO

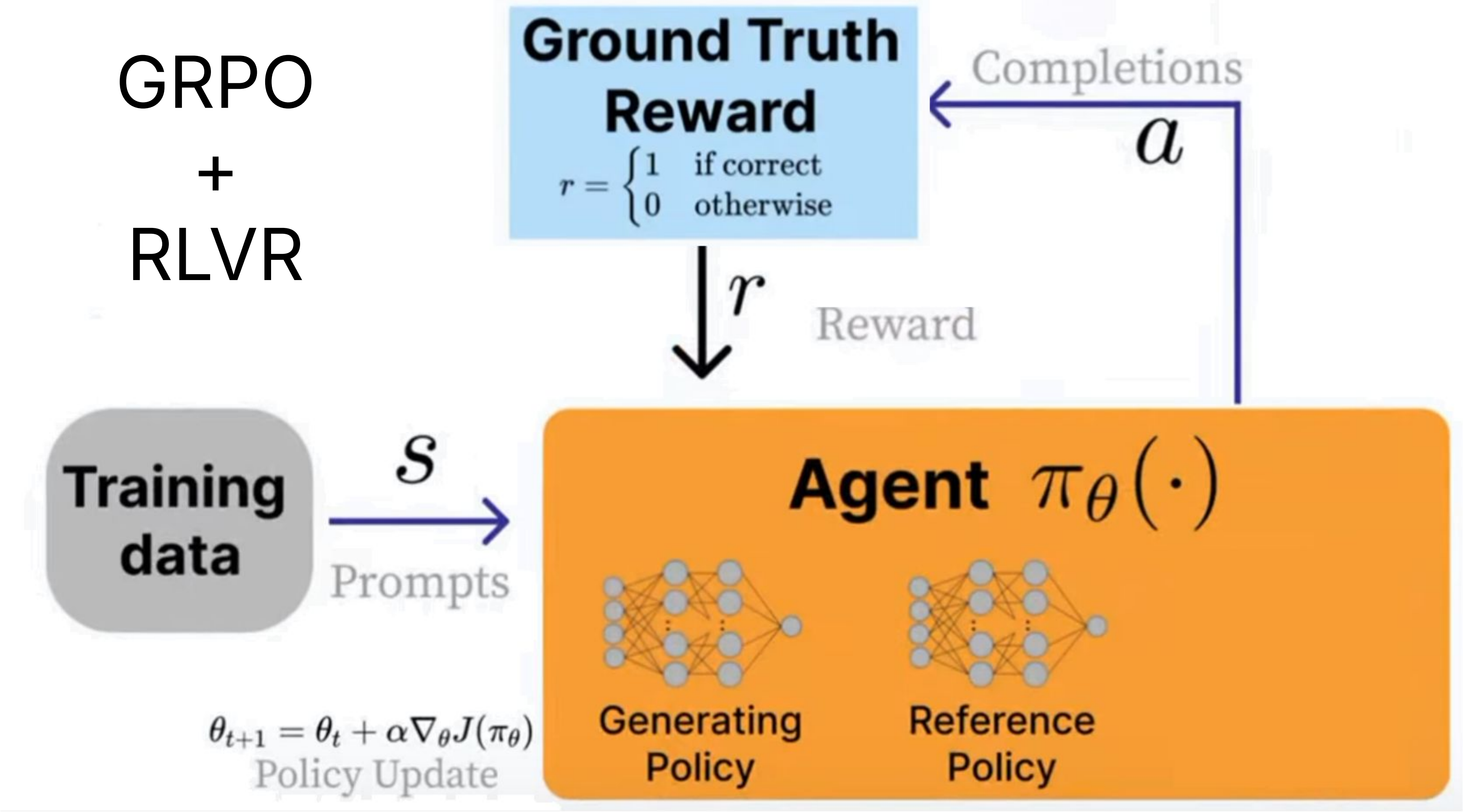

Reward model
Completions
a
r
Reward
Training data
s
Prompts
Agent $\pi_\theta(\cdot)$
Generating Policy
Reference Policy
Value Model
$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\pi_\theta)$
Policy Update

GRPO

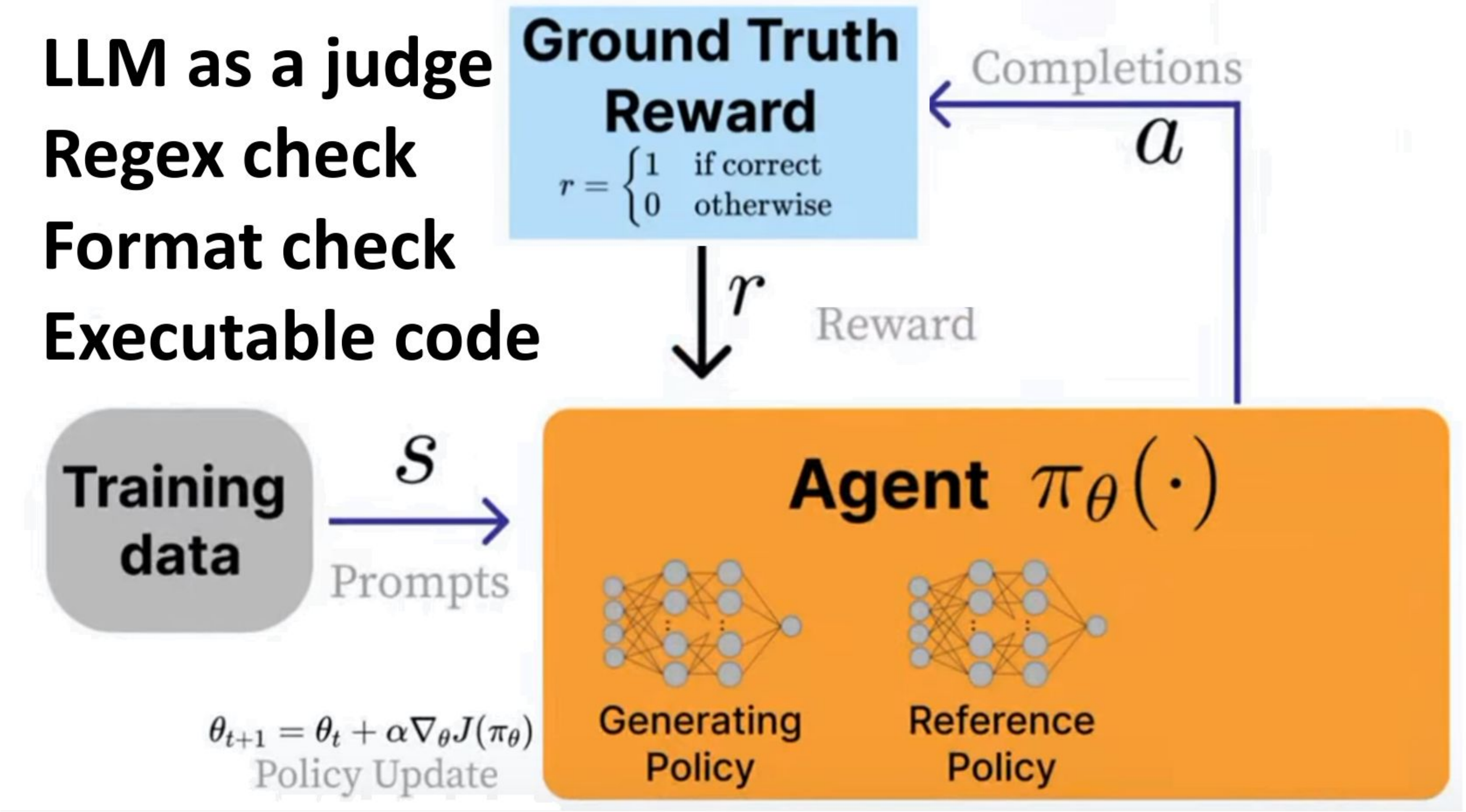


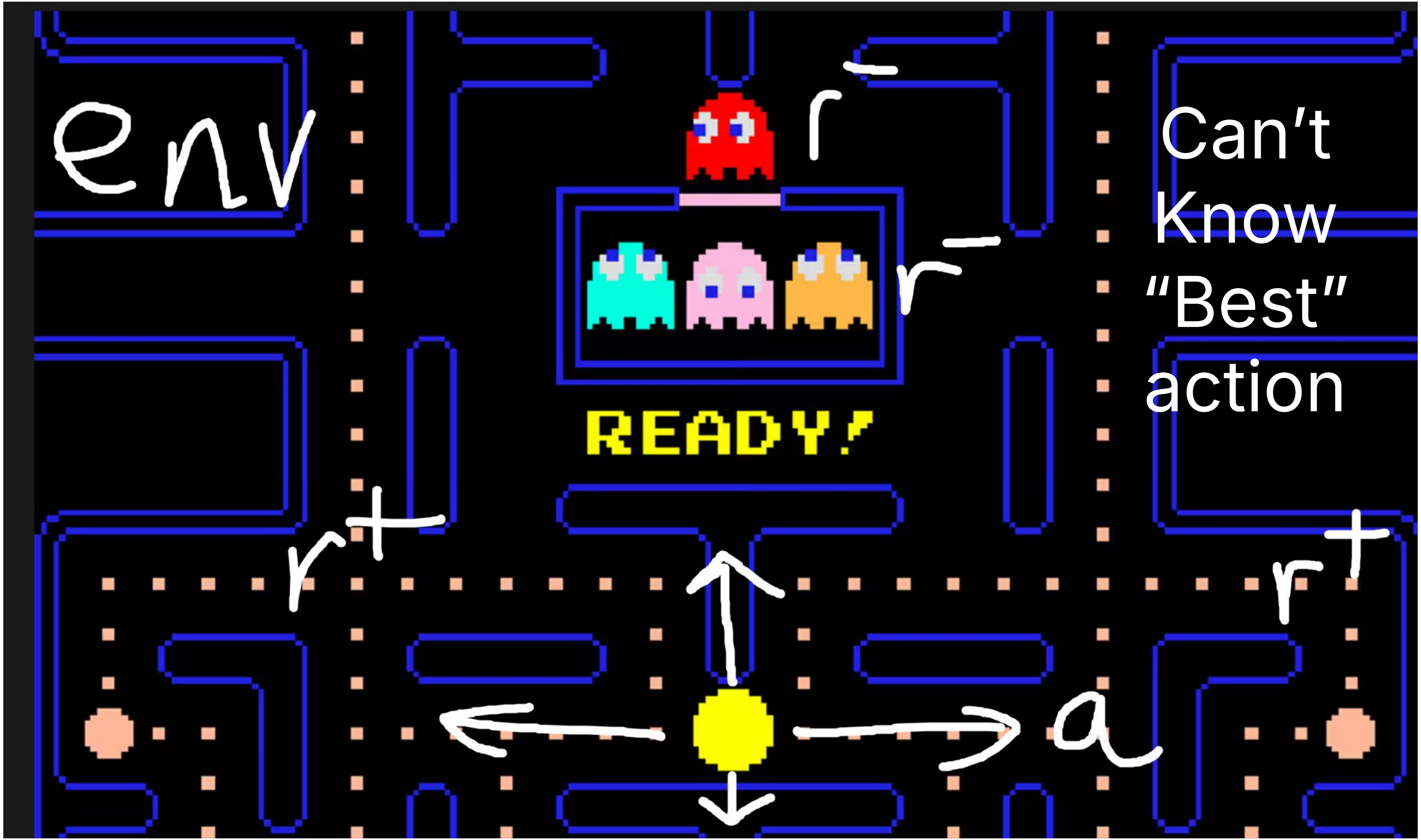
Reward model
Completions
a
r
Reward
Training data
s
Prompts
Agent $\pi_\theta(\cdot)$
$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\pi_\theta)$
Policy Update
Generating Policy
Reference Policy

GRPO
+
RLVR



Ground Truth Reward
$r = \begin{cases} 1 & \text{if correct} \\ 0 & \text{otherwise} \end{cases}$
Completions
a
r
Reward
Training data
s
Prompts
Agent $\pi_\theta(\cdot)$
Generating Policy
Reference Policy
$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\pi_\theta)$
Policy Update


LLM as a judge
Regex check
Format check
Executable code
Ground Truth Reward
$r = \begin{cases} 1 & \text{if correct} \\ 0 & \text{otherwise} \end{cases}$
Completions
a
r
Reward
Training data
s
Prompts
Agent $\pi_\theta(\cdot)$
Generating Policy
Reference Policy
$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\pi_\theta)$
Policy Update


env
r^-
r^-
Can't
Know
"Best"
action
READY!
r^+
r^+
a

Maximize

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_\tau \left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) R_t \right]$$

Action given state

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_\tau \left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta (a_t | s_t) R_t \right]$$

Total Gradient

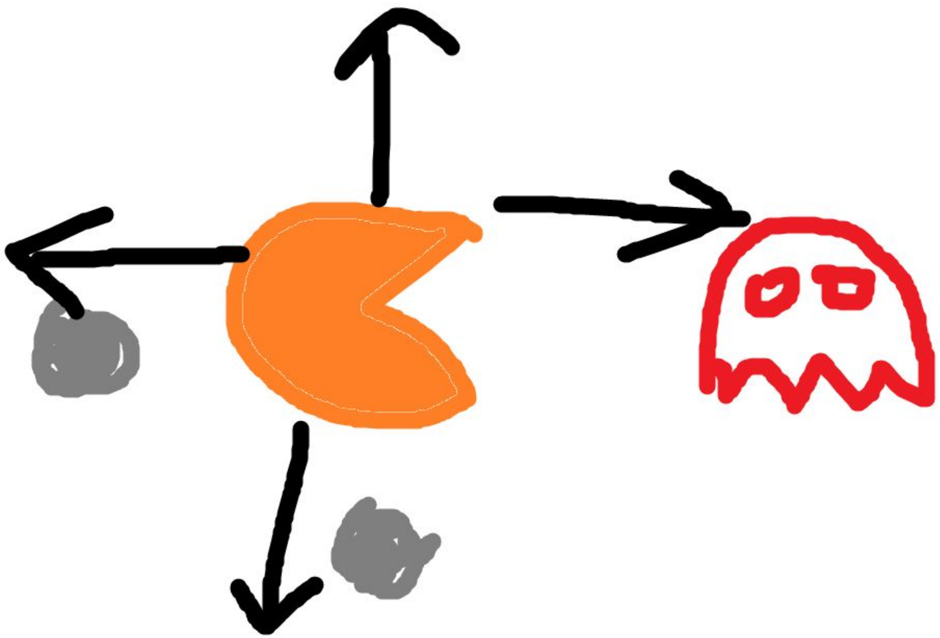
"Calculate gradient"

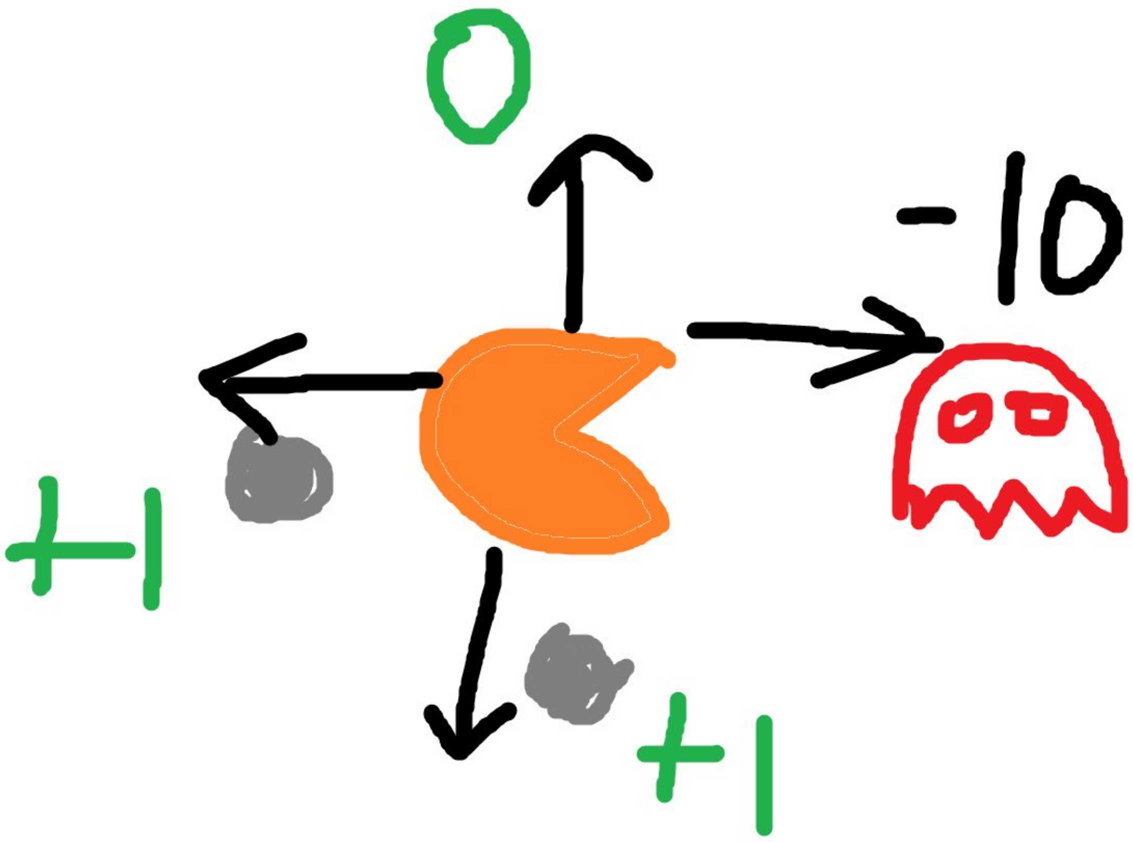
Policy LLM

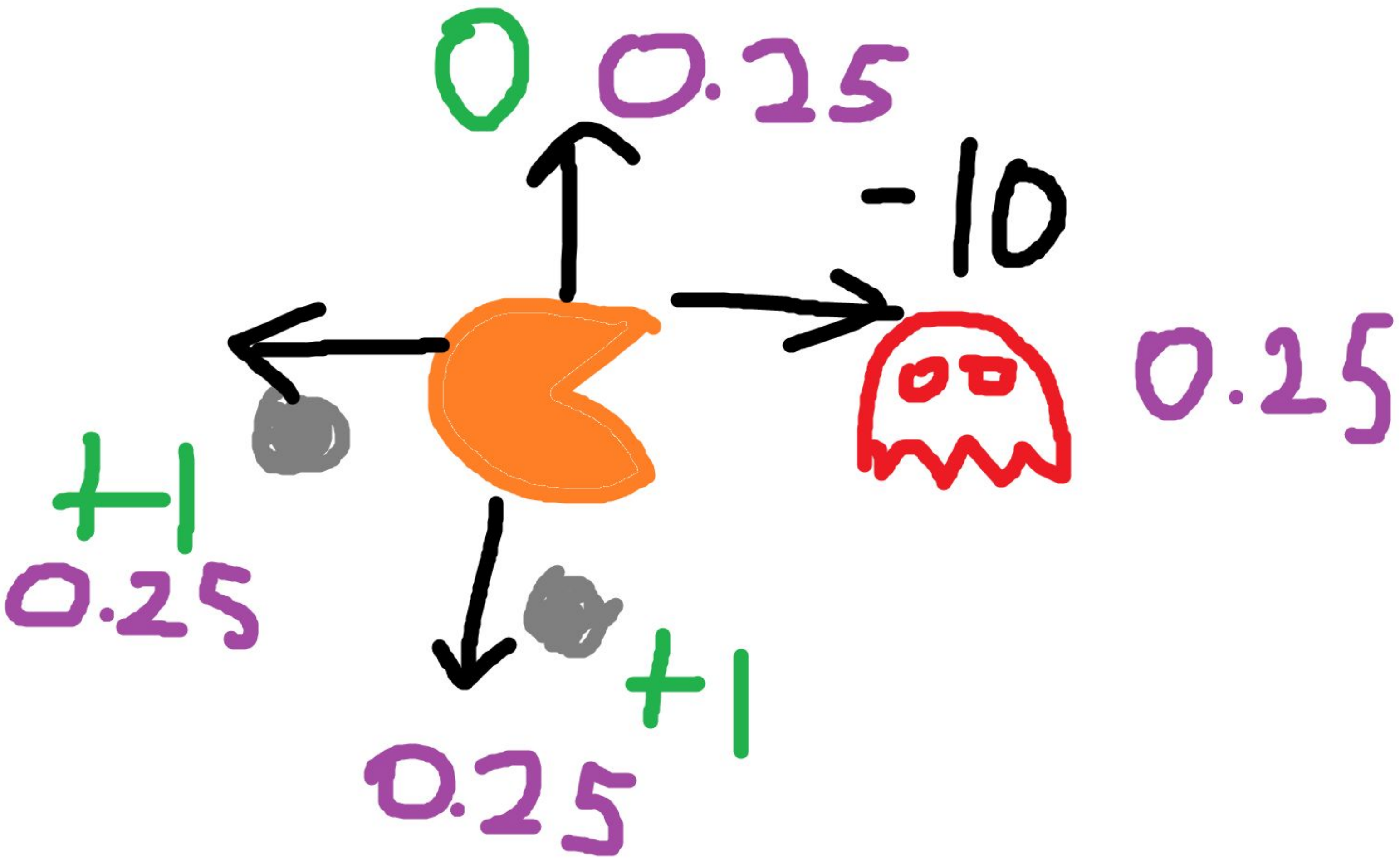
Reward

Maximize

$$\frac{d}{dW} \log P(\text{action}|\text{state}) \cdot \text{reward}$$

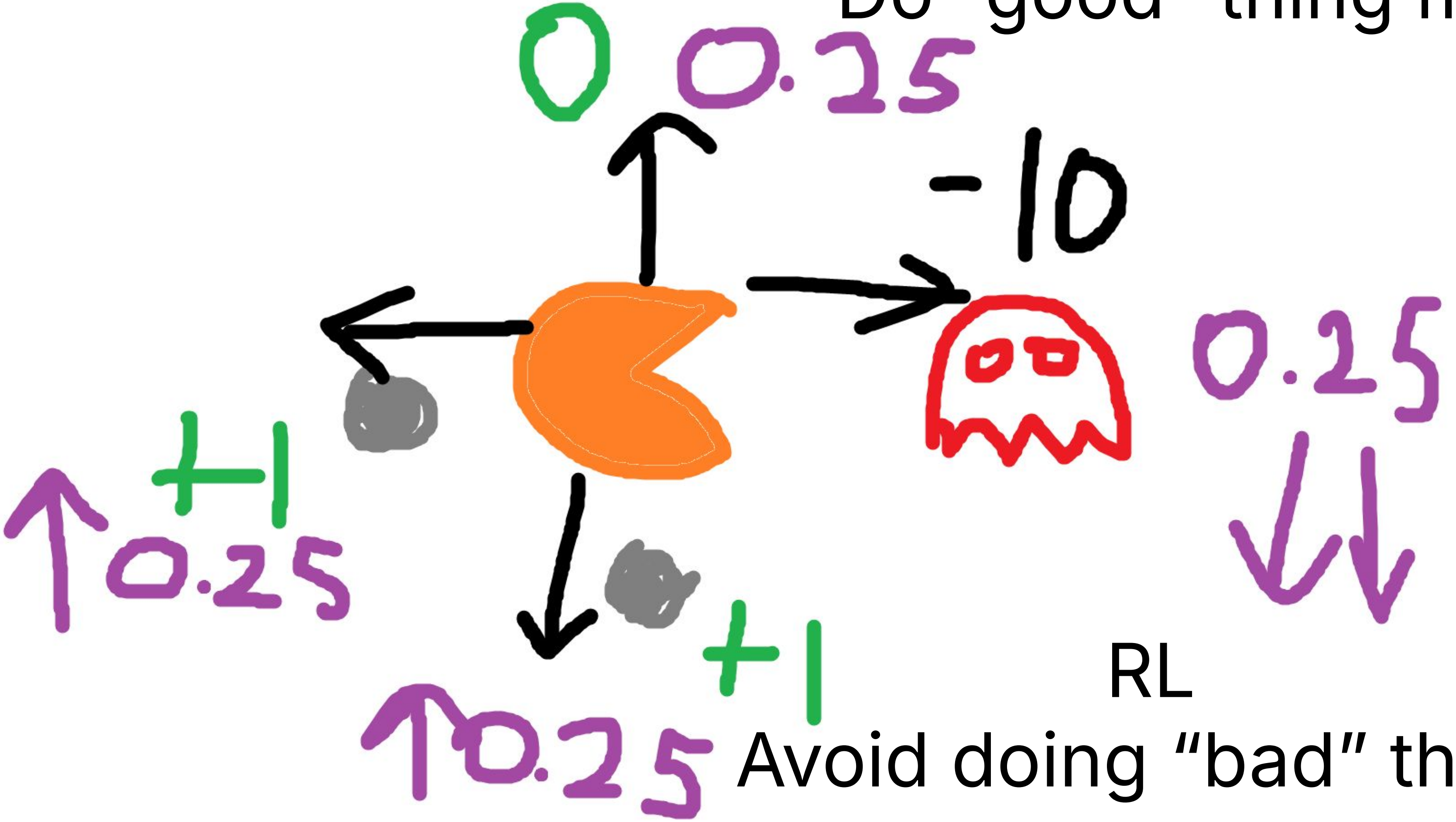



0
-10
+1
+1



0
0.25
-10
0.25
+1
0.25
+1
0.25

Do “good” thing more



RL

Avoid doing “bad” thing

	P(action|state)	Reward	P(A|S)*Reward	(logP)*R
Up	0.25	0	0	0
Down	0.25	1	0.25	-0.6020599
Left	0.25	1	0.25	-0.6020599
Right	0.25	-10	-2.5	6.0205999
				4.8164799

	P(action|state)	Reward	P(A|S)*Reward	(logP)*R
Up	0.2	0	0	0
Down	0.2	1	0.2	-0.6989700
Left	0.2	1	0.2	-0.6989700
Right	0.4	-10	-4	3.9794000
				2.5814600

	P(action|state)	Reward	P(A|S)*Reward	(logP)*R
Up	0.3	0	0	0
Down	0.3	1	0.3	-0.5228787
Left	0.3	1	0.3	-0.5228787
Right	0.1	-10	-1	10
				8.9542425

REINFORCE
Advantage = Normalized Reward from "baseline"

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, A_t\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, (G_t - b)\right]$$

Value Function / Model

$$\nabla_\theta J(\theta) \;=\; \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\,(G_t - b)\Big]$$

Estimates
What is the **average reward** if we see the current state

Value Function / Model

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\,(G_t - b)\right]$$

A < 0 if “worse” than average
A > 0 if “better” than average
Do action more better than average

Proximal Policy Optimization

Maximize J()

$$J(\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}A, \mathrm{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}, 1-\varepsilon, 1+\varepsilon\right)A\right)$$

Proximal Policy Optimization

$$J(\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}A, \mathrm{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}, 1-\varepsilon, 1+\varepsilon\right)A\right)$$

$$J(\theta) = \qquad \left(\frac{\pi_\theta(a|s)}{\qquad}A, \qquad\qquad\qquad\qquad\qquad\qquad\right)$$

Proximal Policy Optimization

What we want to maximize

$$J(\theta) = \min\left(\frac{\pi_\theta(a|s)}{\underbrace{\pi_{\theta_{old}}(a|s)}_{\text{Model that created action}}}A, \operatorname{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}, 1-\varepsilon, 1+\varepsilon\right)A\right)$$

Maximize Likelihood ratio

Proximal Policy Optimization

Top	Bot	Top/Bot	
0.01	0.01	1	
0.01	0.99	0.01	Likely, but we do not want
0.99	0.01	99	Not likely, but we want
0.99	0.99	1	

Proximal Policy Optimization

If we just maximize will “reward hack”

What is 2+2?
To solve this question, we need to ...
...
Hello Hello Hello Hello ← **NOT LIKELY**
4 ← Correct answer though

Proximal Policy Optimization

$$J(\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A, \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}, \underbrace{1-\varepsilon, 1+\varepsilon}_{} \right) A \right)$$

"Trust region"

Do not trust large gradient updates

Eps = 0.1, 0.2, 0.3

1-E = 0.8, 1+E=1.2

PPO - KL term

Deviation from SFT model

KL Divergence

$$r_t = r_\varphi(q, o_{\leq t}) - \beta \log \frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{ref}(o_t|q, o_{<t})}$$

Beta = 0.05

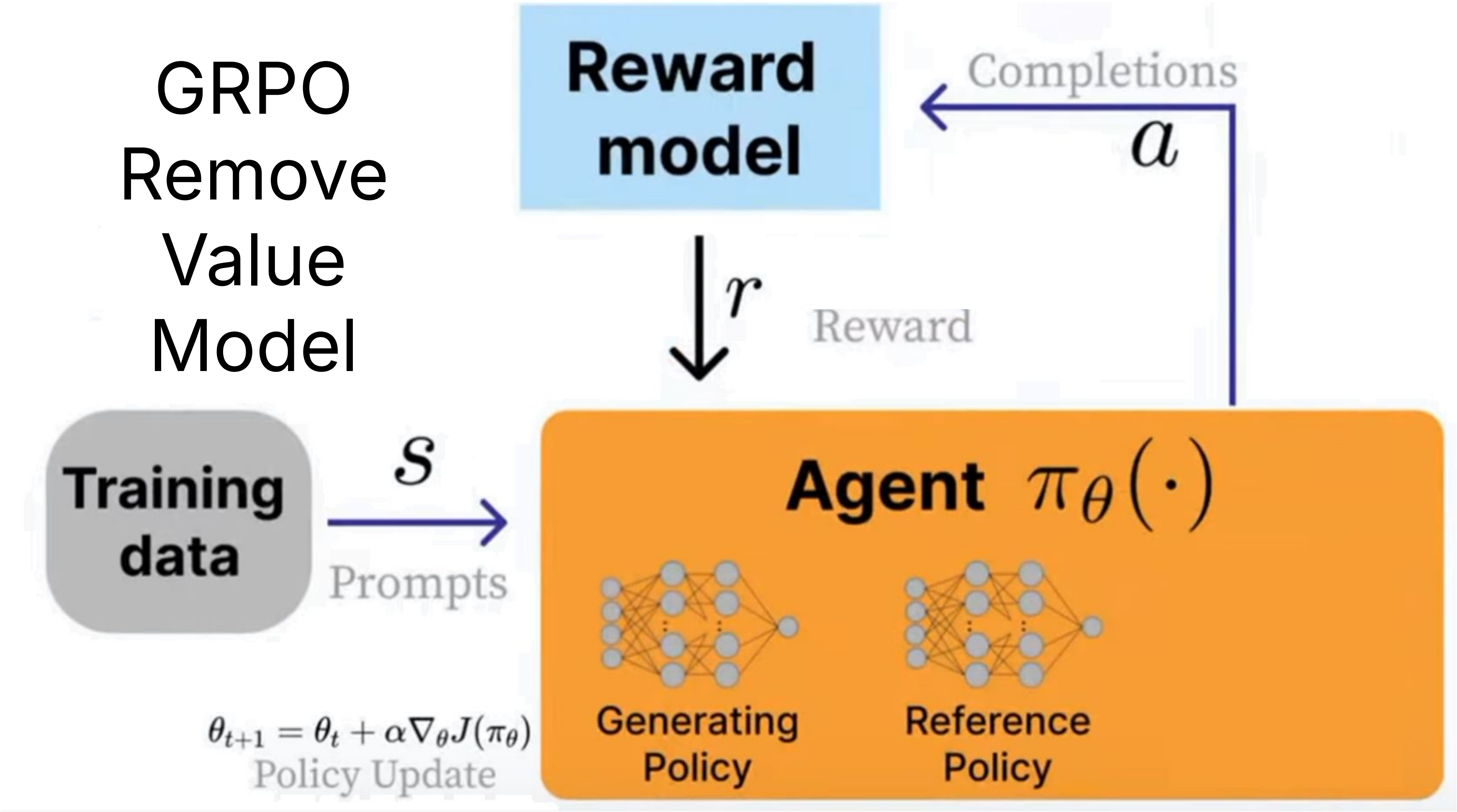
Deviates too much = bad, so like a "tax"

Proximal Policy Optimization

Maximize J()

$$J(\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}A, \mathrm{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}, 1-\varepsilon, 1+\varepsilon\right)A\right)$$

GRPO Remove Value Model



Group Relative Policy Optimization

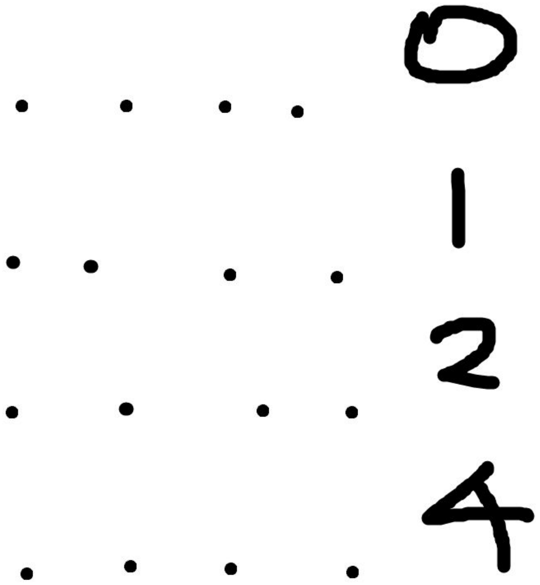
Value model to estimate base / "average" rewards removed

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\,(G_t - b)\right]$$

Do action more better than average

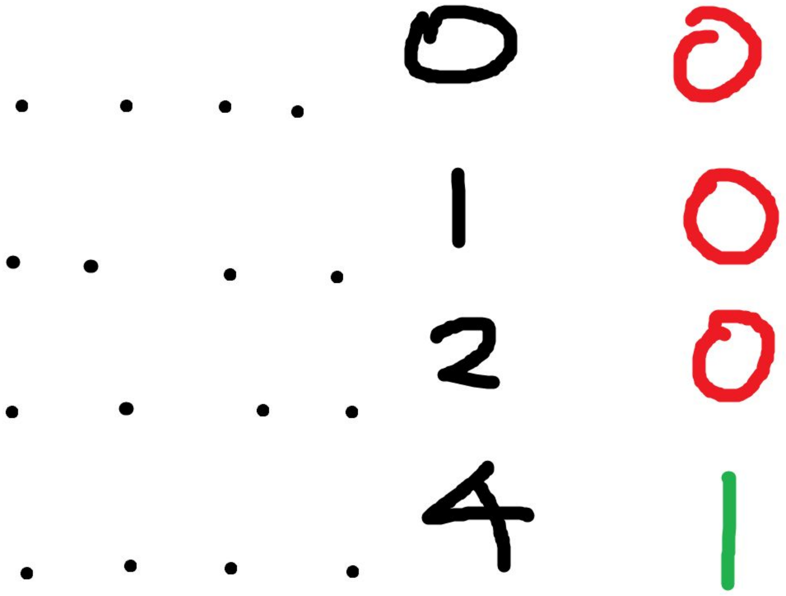
Do “rollout” ie inference sampling

What is 2+2?



Take statistics

What is 2+2?



Take statistics

Advantage = Z Score

No more value model!

$$A_i = \frac{r_i - \mathrm{mean}(r_1, r_2, \cdots, r_G)}{\mathrm{std}(r_1, r_2, \cdots, r_G)}$$

Taking Z Scores

What is 2+2?	Prediction	Reward	Mean	Std	(R-M)/S
	0	0			-0.866
	1	0			-0.866
	2	0			-0.866
	4	1			1.443
			0.375	0.43301	

Why “Group Relative”?

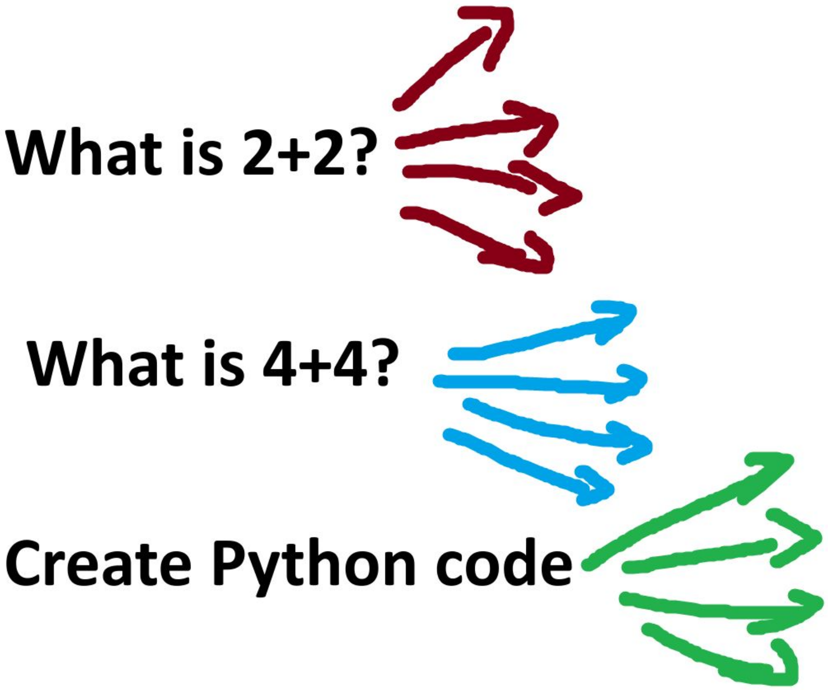
What is 2+2?	Prediction	Reward	Mean	Std	(R-M)/S
	0	0			-0.866
	1	0			-0.866
	2	0			-0.866
	4	1			1.443
			0.375	0.43301	

What is 2+2?	Prediction	Reward	Mean	Std	(R-M)/S
	0	0			-0.866
	1	0			-0.866
	2	0			-0.866
	4	1			1.443
			0.375	0.43301	

What is 2+2?	Prediction	Reward	Mean	Std	(R-M)/S
	0	0			-0.866
	1	0			-0.866
	2	0			-0.866
	4	1			1.443
			0.375	0.43301	

What is 2+2?	Prediction	Reward	Mean	Std	(R-M)/S
	0	0			-0.866
	1	0			-0.866
	2	0			-0.866
	4	1			1.443
			0.375	0.43301	

Why “Group Relative”?



GRPO full form

$$J(\theta) = \frac{1}{G} \sum_{i=1}^{G} \left(\min \left(\frac{\pi_\theta(a_i|s)}{\pi_{\theta_{old}}(a_i|s)} A_i, \mathrm{clip} \left(\frac{\pi_\theta(a_i|s)}{\pi_{\theta_{old}}(a_i|s)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta D_{KL}(\pi_\theta || \pi_{ref}) \right)$$

Resources

https://rlhfbook.com/c/11-policy-gradients.html

https://www.youtube.com/watch?v=bAWV_yrqx4w

Colab GRPO Notebook

https://colab.research.google.com/github/unslothai/notebooks/blob/main/nb/Qwen3_(4B)-GRPO.ipynb

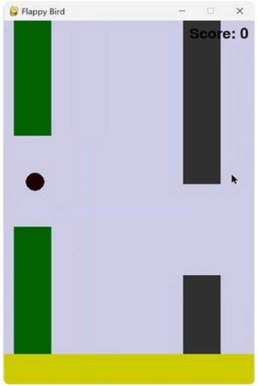
✨ Finetune for Free

Notebooks are beginner friendly. Read our guide. Add your dataset, click "Run All", and export your finetuned model to GGUF, Ollama, vLLM or Hugging Face.

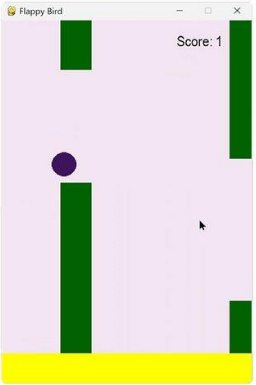
Unsloth supports	Free Notebooks	Performance	Memory use
Qwen3 (14B)	▶️ Start for free	2x faster	70% less
Qwen3 (4B): GRPO	▶️ Start for free	2x faster	80% less

deepseek R1

Dynamic 1.58-bit GGUF



R1 original 8-bit



R1 dynamic 1-bit

To test the quantized models, we asked DeepSeek R1 to create a Flappy Bird game with 3 tries (pass@3), scoring it on 10 criteria (e.g., random colors, shapes, and Python compatibility). We used seeds 3407, 3408, and 3409, with a temperature of 0.6.

The left is from chat.deepseek.com, the right is 1.58bit. As you can see, the 1.58bit's code exceutes a fully functional Flappy Bird game. 1.58bit runs at ~140 tokens/sec with 160GB VRAM.

unslot

deepseek

R1-0528

Dynamic 1-bit GGUFs

Qwen3

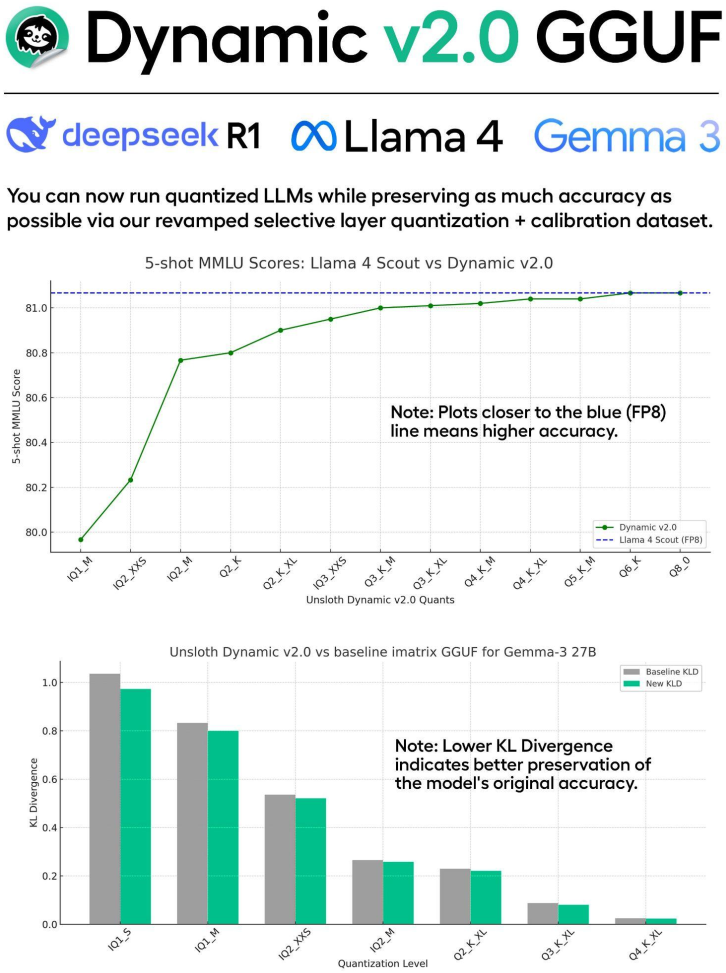
Support includes R1-0528 Qwen3-8B Distill GGUFs

Naive quantization causes the model to break with loops, gibberish and produce poor code. Our dynamic quants solve this.

By studying R1-0528's architecture, we selectively quantize layers to higher bits (like 4-bit), and layers like MOE to lower bits.

The 1.78-bit quant can fit in 1x 24GB VRAM GPU with offloading for fast generation inference at ~10 tokens/sec.

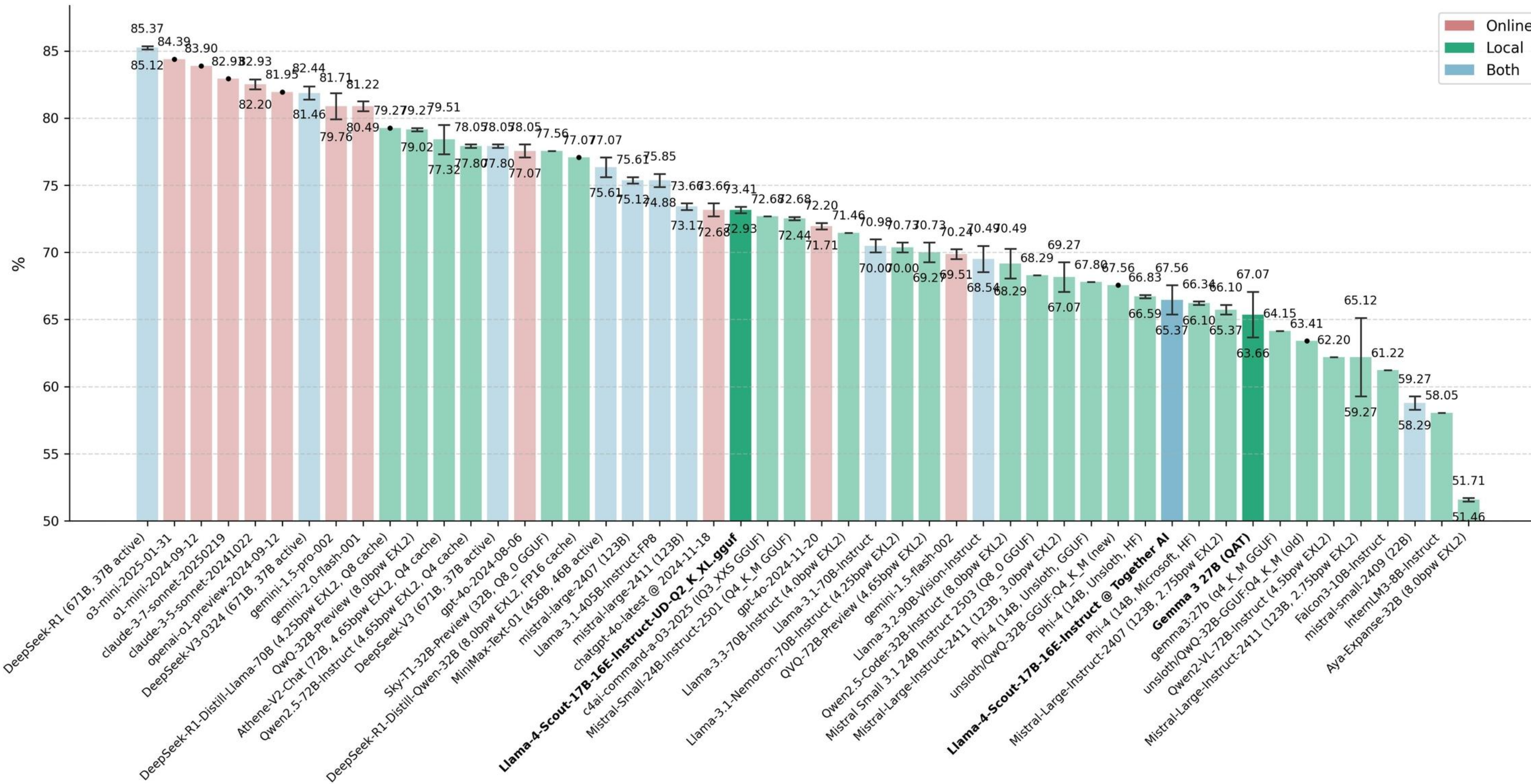
unsloth



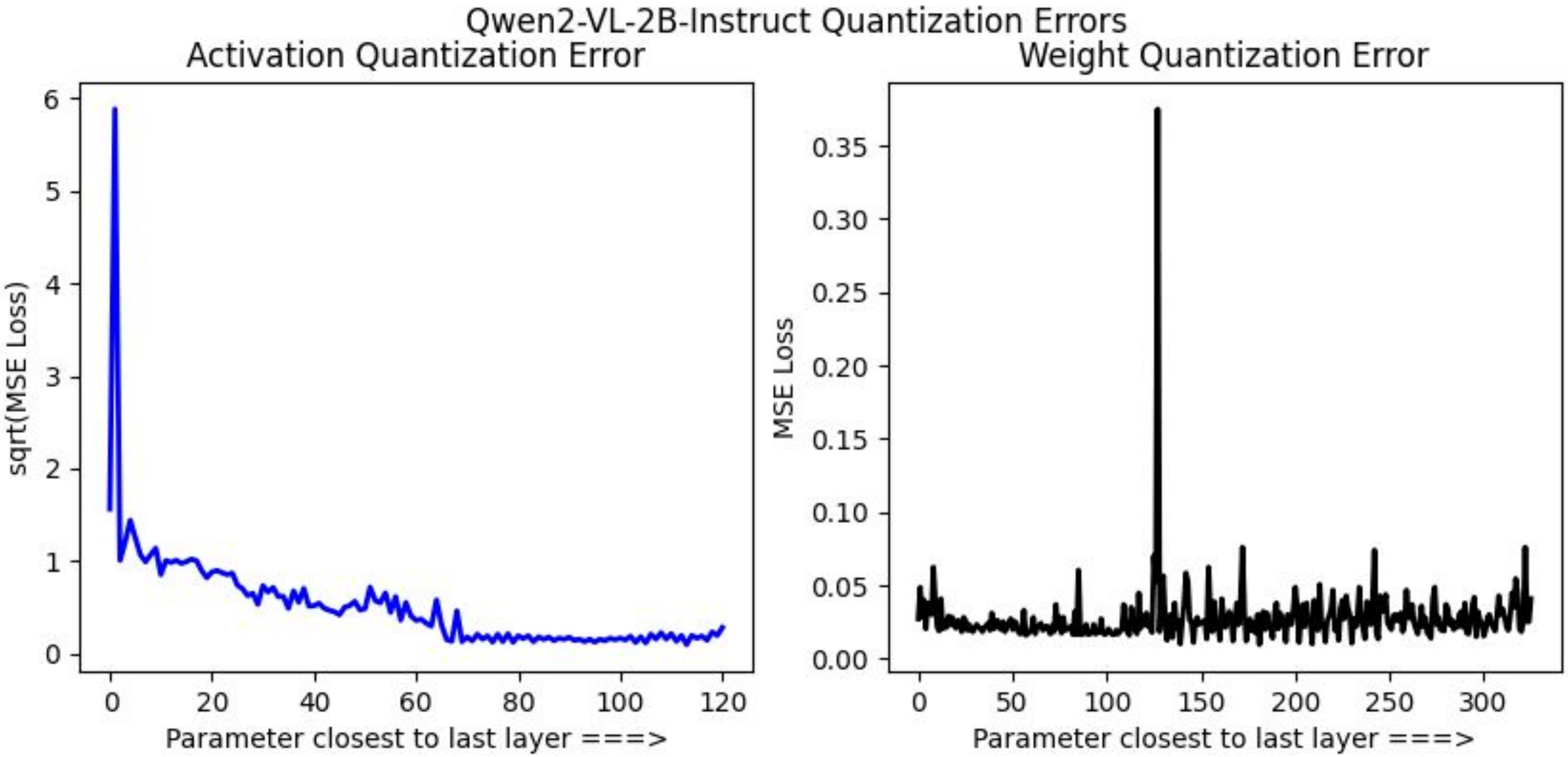
Quantize MoE layers heavily

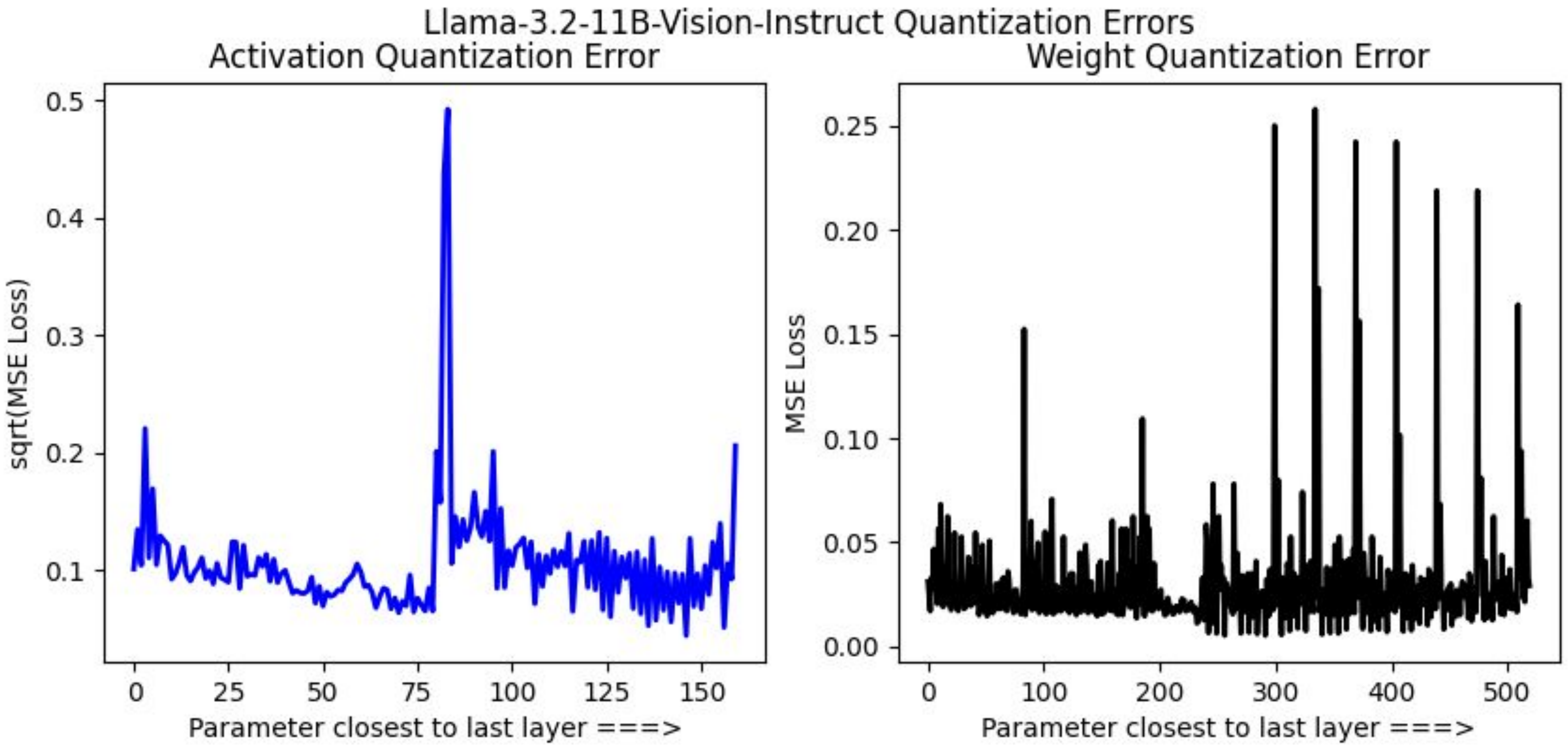
Leave attention
Shared experts
In higher precision

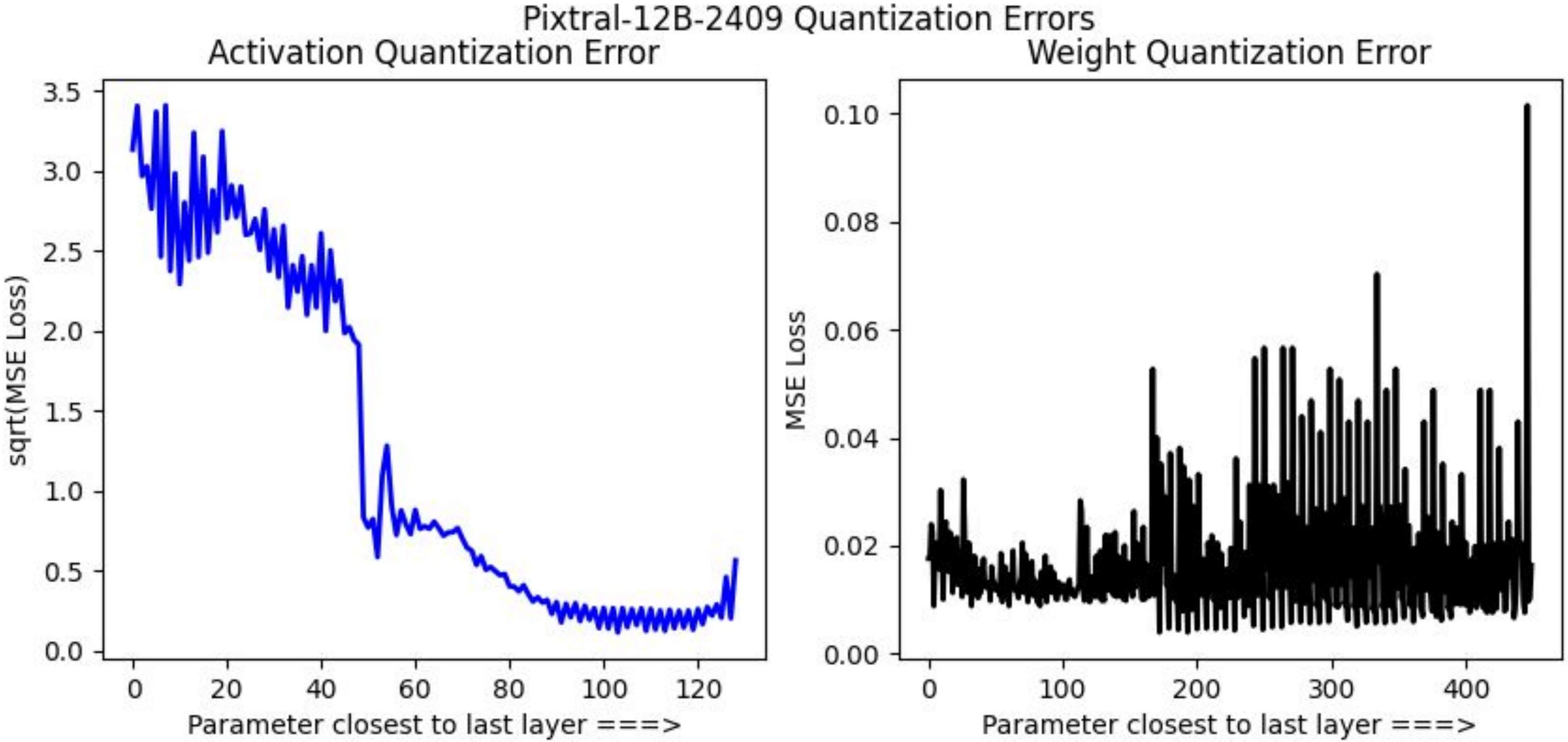
Wolfram Ravenwolf's MMLU-Pro Computer Science LLM Benchmark Results (2025-04-08)



Qwen2-VL-2B-Instruct	**Description**	**Size**	**Result**
16bit	The image shows a train traveling on tracks.	4.11GB	✅
Default 4bit all layers	The image depicts a vibrant and colorful scene of a coastal area.	1.36GB	❌
Unsloth quant	The image shows a train traveling on tracks.	1.81GB	✅


Qwen2-VL-2B-Instruct Quantization Errors
Activation Quantization Error
sqrt(MSE Loss)
0
1
2
3
4
5
6
0
20
40
60
80
100
120
Parameter closest to last layer ===>
Weight Quantization Error
MSE Loss
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
50
100
150
200
250
300
Parameter closest to last layer ===>

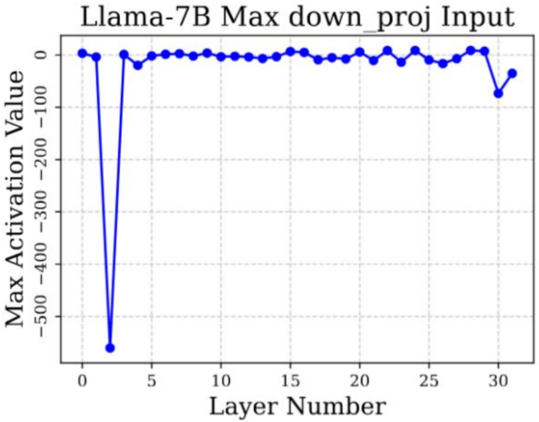

Llama-3.2-11B-Vision-Instruct Quantization Errors
Activation Quantization Error
Weight Quantization Error
sqrt(MSE Loss)
MSE Loss
Parameter closest to last layer ===>
Parameter closest to last layer ===>


Pixtral-12B-2409 Quantization Errors
Activation Quantization Error
sqrt(MSE Loss)
Parameter closest to last layer ===>
Weight Quantization Error
MSE Loss
Parameter closest to last layer ===>

Model	No.	Type	Weight	Coordinates
Llama 7B	2	mlp	down_proj	[3968, 7003]
Llama 13B	2	mlp	down_proj	[2231, 2278]
	2	mlp	down_proj	[2231, 6939]
Llama 30B	3	mlp	down_proj	[5633, 12817]
	3	mlp	down_proj	[5633, 17439]
	10	mlp	down_proj	[5633, 14386]
Llama2 7B	1	mlp	down_proj	[2533, 7890]
Llama2 13B	3	mlp	down_proj	[4743, 7678]
Mistral-7B v0.1	1	mlp	down_proj	[2070, 7310]

Model	No.	Type	Weight	Coordinates
OLMo-1B 0724-hf	1	mlp	down_proj	[1764, 1710]
	1	mlp	down_proj	[1764, 8041]
OLMo-7B 0724-hf	1	mlp	down_proj	[269, 7467]
	2	mlp	down_proj	[269, 8275]
	7	mlp	down_proj	[269, 453]
	24	mlp	down_proj	[269, 2300]
Phi-3 mini-4k-instruct	2	mlp	down_proj	[525, 808]
	2	mlp	down_proj	[1693, 808]
	2	mlp	down_proj	[1113, 808]
	4	mlp	down_proj	[525, 2723]
	4	mlp	down_proj	[1113, 2723]
	4	mlp	down_proj	[1693, 2723]

Table 2: Super Weight Directory. The above layer numbers, layer types, and weight types can be directly applied to Huggingface models. For example, for Llama-7B on Huggingface, access the super weight using `layers[2].mlp.down_proj.weight[3968, 7003]`.



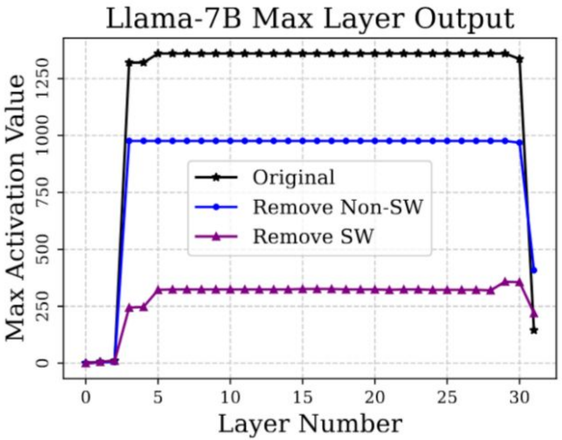


Figure 3: How to identify the Super Weight for Llama-7B. `down_proj` input features a large maximum-magnitude activation only in Layer 2, where the super activation first appeared. The value's channel index, e.g., `7003`, tells the row of SW. `down_proj` output likewise features a large maximum-magnitude activation at Layer 2. This value's channel index, e.g., `3968`, gives us the column of the SW.

Figure 4: The super activation persists throughout the entire model, at exactly the same magnitude, starting after Layer 2. Pruning the super weight decreases the super activation's magnitude by 75%.

Format Name	Element Data Type	Element Bits (d)	Scaling Block Size (k)	Scale Data Type	Scale Bits (w)
MXFP8	FP8 (E5M2) FP8 (E4M3)	8	32	E8M0	8
MXFP6	FP6 (E3M2) FP6 (E2M3)	6	32	E8M0	8
MXFP4	FP4 (E2M1)	4	32	E8M0	8
MXINT8	INT8	8	32	E8M0	8

Table 1. Format names and parameters of concrete MX-compliant formats.

	GB200 Grace Blackwell Superchip	H100 NVL[1]
Configuration	1 Grace CPU : 2 Blackwell GPU	2x H100 NVLink
FP4 Tensor Core[2]	40 PFLOPS	
FP8/FP6 Tensor Core[2]	20 PFLOPS	7,916 teraFLOPs[2]
FP16/BF16 Tensor Core[2]	10 PFLOPS	3,958 teraFLOPs[2]
GPU Memory | Bandwidth	Up to 384 GB HBM3e | 16 TB/s	188GB

Blackwell GPU		
FP8	20 PFLOPS	2.5X Hopper
NEW FP6	20 PFLOPS	2.5X
NEW FP4	**40 PFLOPS**	**5X**
HBM Model Size	740B param	6X
HBM Bandwidth	34T param/s	5X

	Exponent	Mantissa	~ Space E+M^2	x fp32
Float32	8	23	537	1
Float16	5	10	105	5
Bfloat16	8	7	57	9
Float8 E4M3	4	3	13	41
Float8 E5M2	5	2	9	60
1.58bit (E5M2)	5	2	7	77
Float4	2	1	3	179
1.58bit (Float4)	2	1	3	179

Float4 sign=1 exponent=1 mantissa=3

Wikipedia table:

	0 ... 0	**0 ... 1**	**1 ... 0**	**1 ... 1**
... 00 ...	0	0.5	−0	−0.5
... 01 ...	1	1.5	−1	−1.5
... 10 ...	2	3	−2	−3
... 11 ...	Inf	NaN	−Inf	NaN

Use torch.compile!

```
{'TYPE_CHECKING': False,
 'inplace_padding': True,
 'can_inplace_pad_graph_input': False,
 'enable_auto_functionalized_v2': True,
 'debug': False,
 'disable_progress': True,
 'verbose_progress': False,
 'fx_graph_cache': True,
 'fx_graph_remote_cache': None,
 'bundle_triton_into_fx_graph_cache': True,
 'autotune_local_cache': True,
 'autotune_remote_cache': None,
 'bundled_autotune_remote_cache': None,
 'force_disable_caches': False,
 'sleep_sec_TESTING_ONLY': None,
 'custom_op_default_layout_constraint': 'needs_fixed_stride_order',
 'triton_kernel_default_layout_constraint': 'needs_fixed_stride_order',
 'cpp_wrapper': False,
 'cpp_cache_precompile_headers': True,
 'online_softmax': True,
```

Thank you!

Star us on GitHub

Join our Discord

www.unsloth.ai

Don’t forget to grab stickers