

Datasets & Infra For GRPO

How to train your own reasoning model with RL



What we will cover

GOSIM

- High Level DeepSeek-R1 Overview
- What is GRPO?
- What you need to run GRPO yourself
 - Practical Example: Training a 3B Rust Coder 🦀
 - Training Data
 - Hardware Requirements
 - Reward Functions



\$ whois

GOSIM

- Who is using foundation model APIs? (OpenAI, Anthropic, etc)
- Who is using local/self hosted LLMs?
- Who has done a Supervised Fine Tune?
- Who has done Reinforcement Learning?
 - PPO? DPO? GRPO?
- Who has a large proprietary dataset/data flywheel?



\$ whoami

GOSIM

- Founder & CEO of Oxen.ai
 - Training Language Models Since 2013
 - Ex-IBM Watson
 - Built AI/ML projects in NLP, Computer Vision, 3D Asset Generation, Reinforcement Learning
 - Host of Arxiv Dives
 - Live in Los Angeles, CA
- X: @gregschoeninger





What is Oxen.ai?

GOSIM

Tools for collaborating on AI/ML projects (all experimentation was done in platform)

- Notebooks on GPUs
- Train Models
- Evaluate Models
- Data Exploration
- Synthetic Data
- Version Control

The screenshot shows the Oxen.ai platform interface. At the top, there's a navigation bar with links for Datasets, Models, Docs, Blog, and Upgrade, along with a search bar and user account information. Below the navigation is a header for the 'main' branch of the 'OCRBatch-v2' repository, showing a file named 'default_test.parquet'. A progress bar indicates 'Ox Data Bot' is detecting the language of all questions. A search bar contains the query 'Find all the rows where language == english'. Below the search bar is a table titled 'Natural Language Query Applied' with 10 columns and 1-100 of 8299 rows. The table includes columns for Schema, id, image, question, answers, dataset_name, type, and id. The first few rows show various questions and their corresponding answers, such as 'What is the wrong ans...' and 'What's the name of the...'. To the right of the table, there's a preview of an image file named 'image_2.png' which appears to be a Facebook login interface. The bottom of the table shows a navigation bar with page numbers from 1 to 100.

Schema	(idx)	image	question	answers	dataset_name	type	id
image (str)	11	images/default/te	What is the wrong ans...	[\"enabled\", \"on\"], ..	rico	APP agent en	0
question (str)	12	images/default/te	What's the name of the...	[\"sberenter\"], ... 1 it	rico	APP agent en	1
answers (list[str])	13	images/default/te	What application is use...	[\"Facebook\", \"Facebo...	rico	APP agent en	2
dataset_name (str)	14	images/default/te	What is the sleep time?	[\"12:42 p.m.\", \"12:42	rico	APP agent en	3
type (str)	15	images/default/te	What is the coupon co...	[\"APPVIA\", ... 1 item	rico	APP agent en	4
id (i64)	16	images/default/te	What is the phone num...	[\"23480235049551\",	rico	APP agent en	5
language (str)	17	images/default/te	What is the address of ...	[\"312 Bingham Road\",	rico	APP agent en	6
	18	images/default/te	What is the contact nu...	[\"(212) 555-1212\", "	rico	APP agent en	7
	19	images/default/te	What application will ha...	[\"Truth or Dare\"], ...	rico	APP agent en	8
	20	images/default/te	What is the departure c...	[\"San Francisco\"], ...	rico	APP agent en	9
	21	images/default/te	Which account am I loa...	[\"Facebook\", \"accou...	rico	APP agent en	10



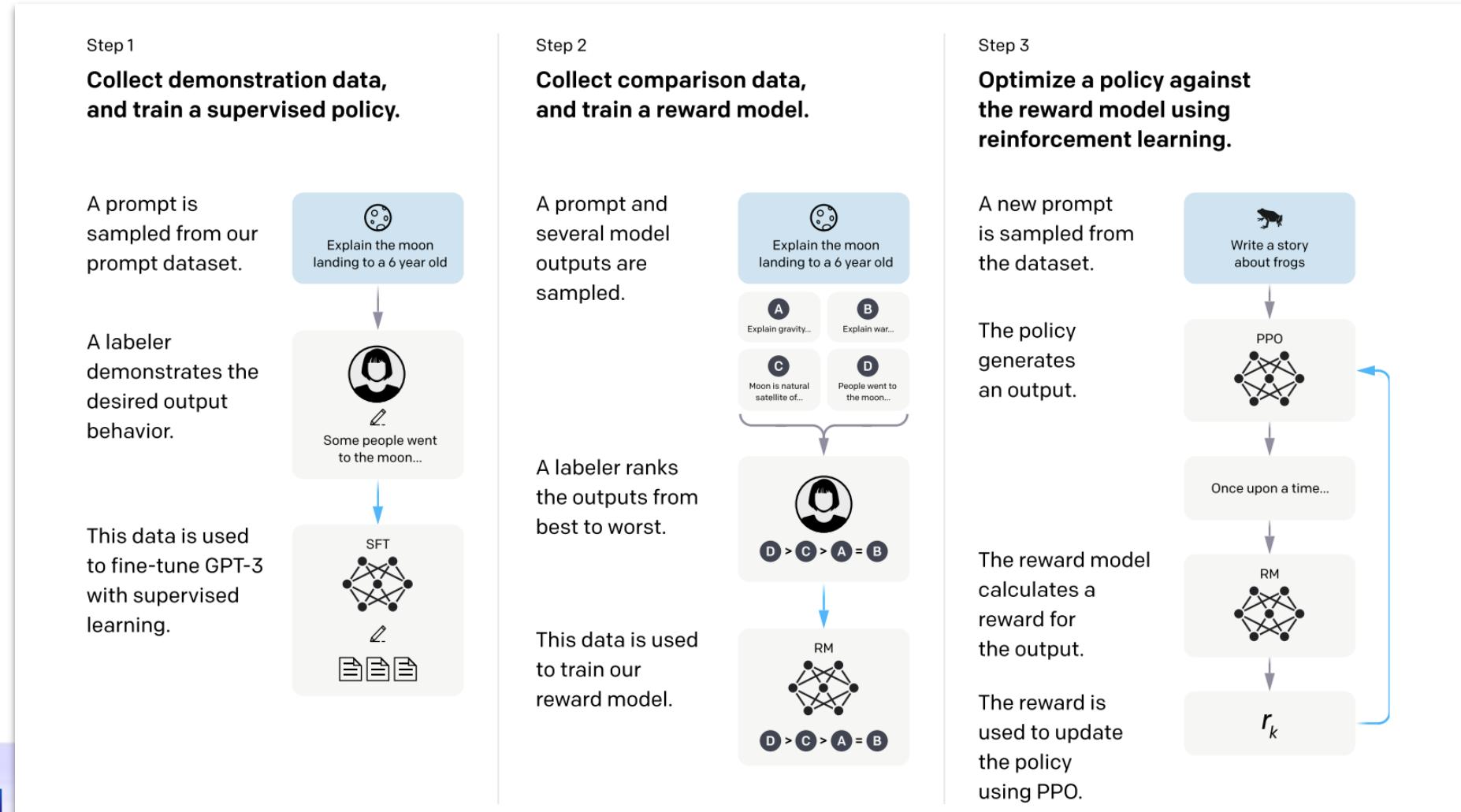
Let's Dive Into GRPO



How is RL used in LLMs?

GOSIM

InstructGPT - <https://arxiv.org/abs/2203.02155>



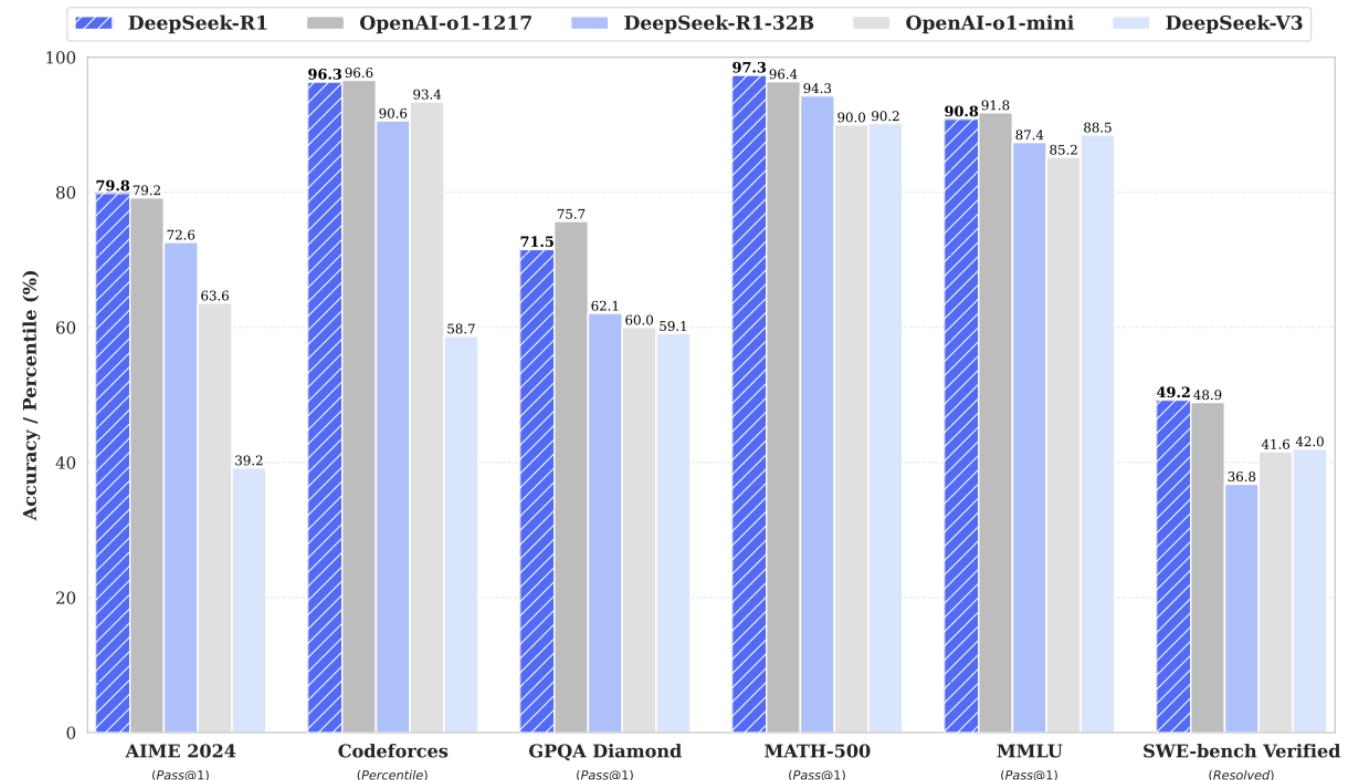
DeepSeek-R1 Entered the Chat

GOSIM



DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI
research@deepseek.com



Reasoning 101

GOSIM

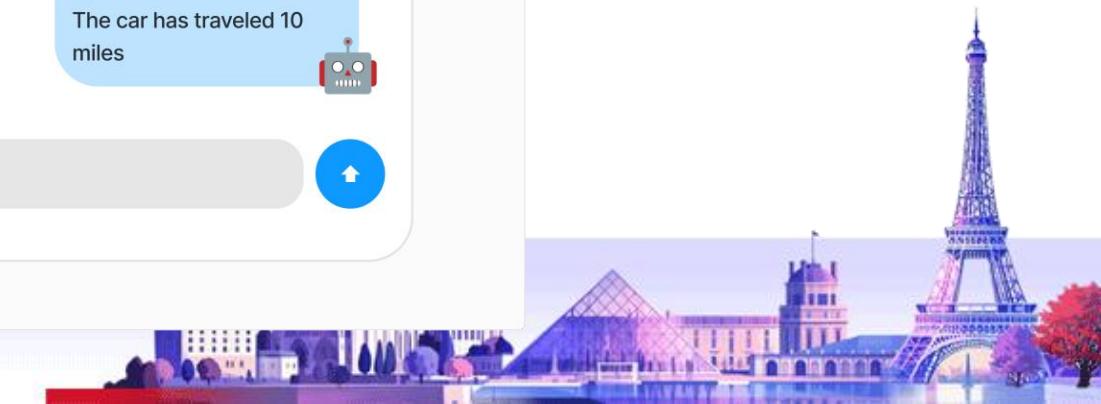
The image shows two side-by-side AI interaction logs. Both logs start with a user (yellow icon) asking a question about a car's distance traveled.

Left Log (LLM w/o Reasoning):

- User: If a car is moving at 30 mph and travels for 20 minutes, how far has it gone?
- Model (red robot): The car has traveled 60 miles.

Right Log (Reasoning LLM):

- User: If a car is moving at 30 mph and travels for 20 minutes, how far has it gone?
- Model (red robot):
 - <think>
 - Let's break this down step by step.
 - Distance=Speed*Time
 - Speed = 30 miles per hour
 - Time = 20 minutes
 - We need time in hours.
 - Time = 1/3 hours
 - $30 * 1/3 = 10$
 - </think>
 - The car has traveled 10 miles



R1-Zero “Aha Moment”

The logo for GOSIM, featuring the word "GOSIM" in a bold, sans-serif font. The letter "O" is stylized with a circular graphic composed of three overlapping colored arcs: teal at the top, purple in the middle, and blue at the bottom.

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

• •

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

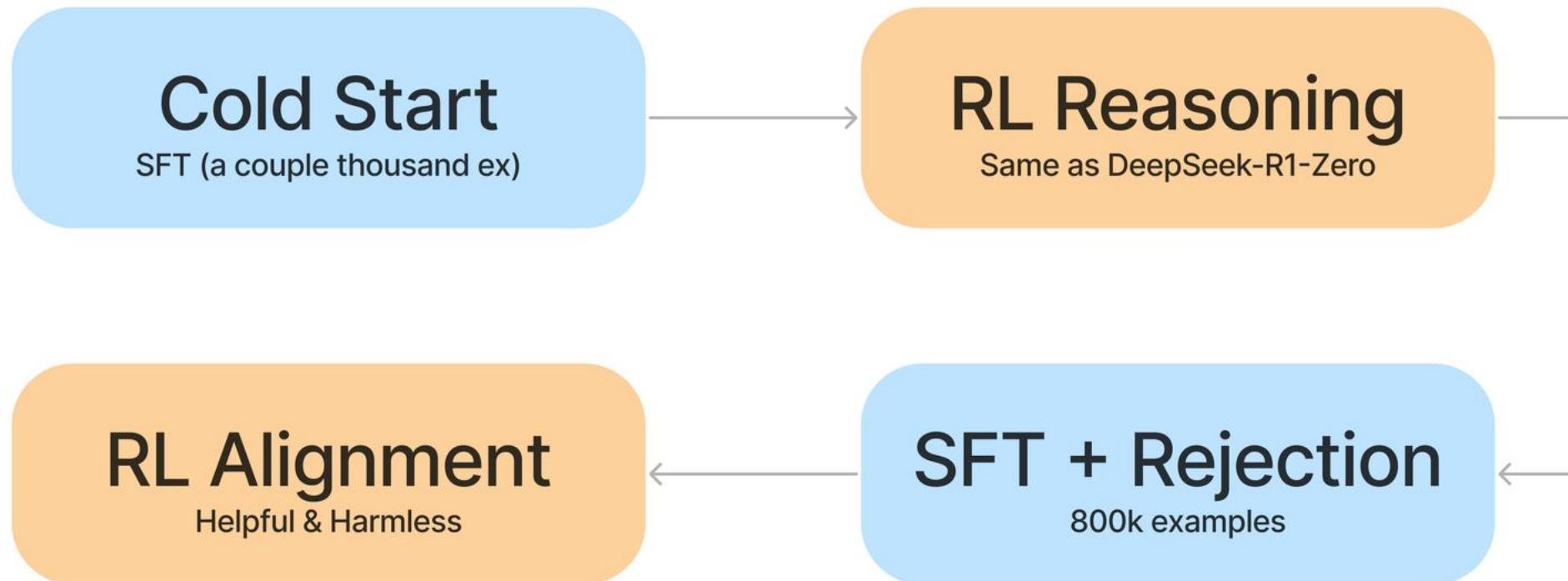
Next, I could square both sides again, treating the equation: ...

• •

Table 3 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

R1 = SFT+GRPO

GOSIM



What is GRPO?

<https://www.oxen.ai/blog/why-grpo-is-important-and-how-it-works>



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

Zhihong Shao^{1,2*†}, Peiyi Wang^{1,3*†}, Qihao Zha¹,
Xiao Bi¹, Haowei Zhang¹, Mingchuan Zhang¹

¹DeepSeek-AI, ²Tsinghua University

{zhihongshao, wangpeiyi, zhuhui}@
https://github.com/deepseek

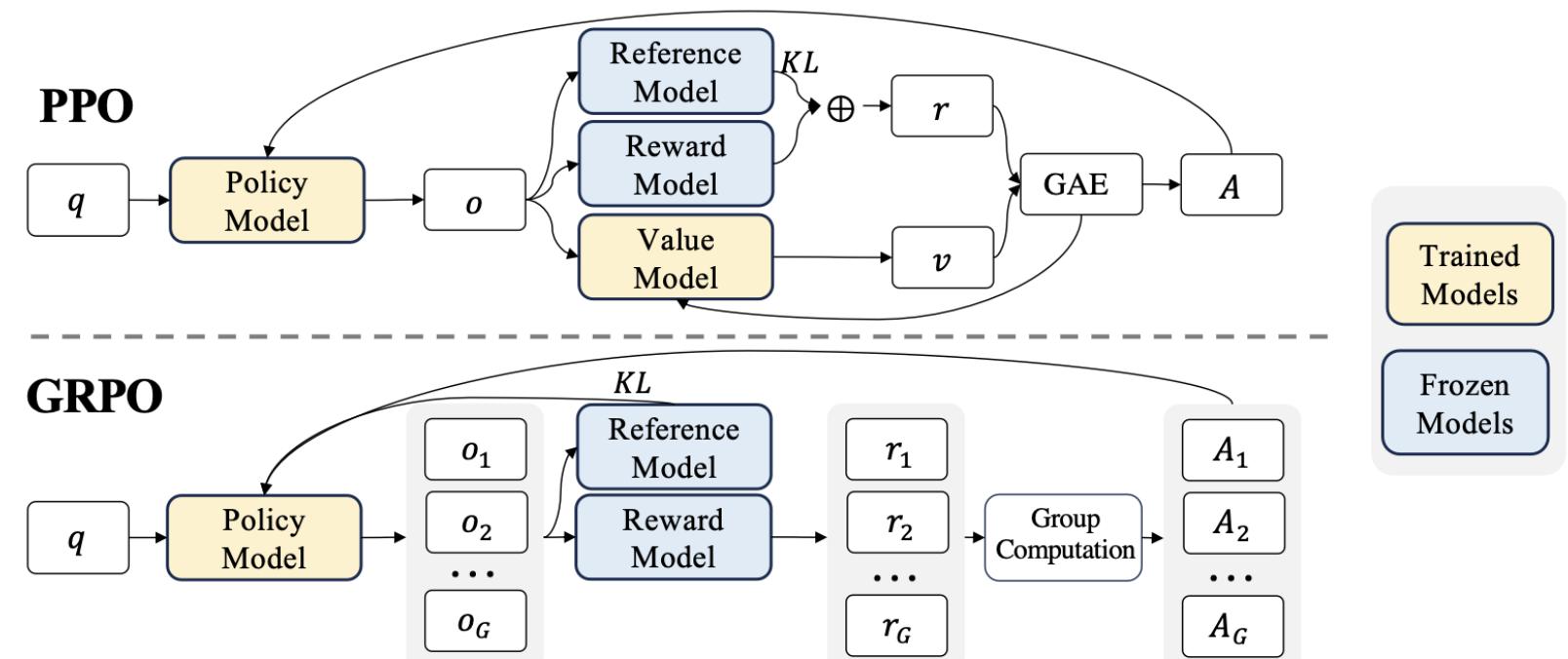
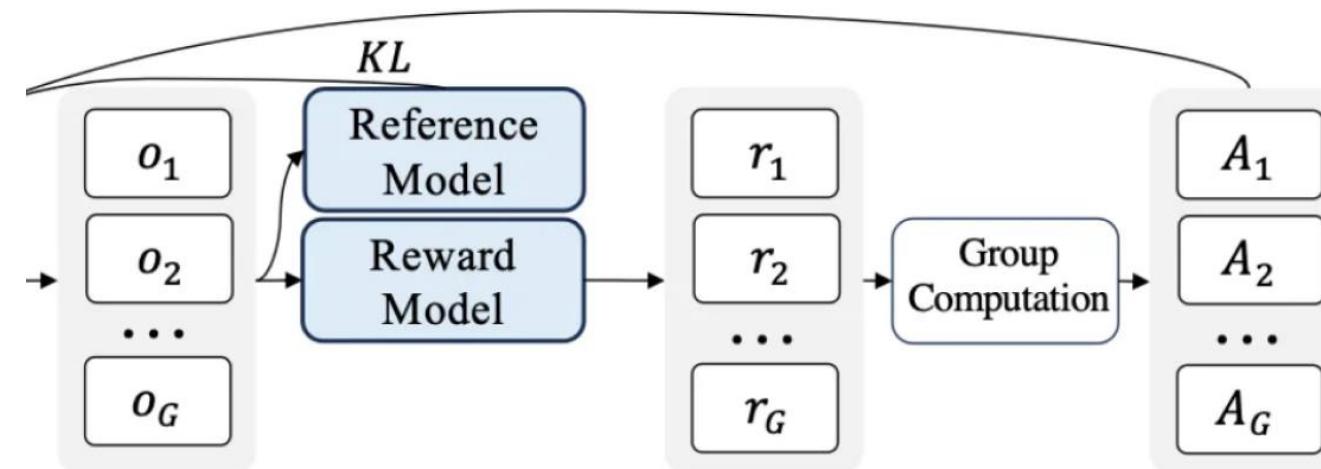


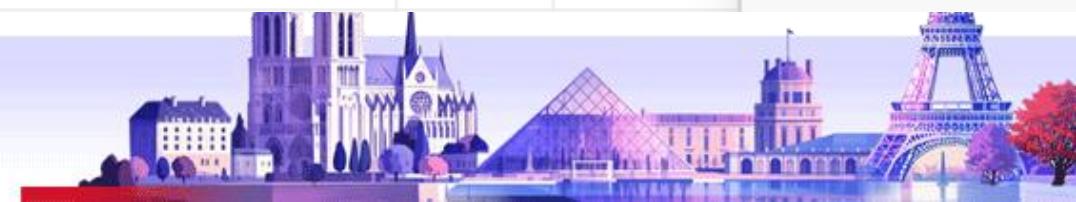
Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

Group Relative Policy Optimization

GOSIM



question	answer	response	model_answer	is_correct	question
Mr. Curtis has 325 chickens on his farm where ...	277	<reasoning>First, let's find out how many hens there are. The total n...	277	true	Mr. Curtis has 325 c...
Mr. Curtis has 325 chickens on his farm where ...	277	<reasoning>You need to subtract the 20 hens that do not lay eggs fr...	305	false	where 28 are rooster hens. Twenty hens c...
Mr. Curtis has 325 chickens on his farm where ...	277	<reasoning>First, we need to find the total number of hens. We know...	277	true	while the rest of the many egg-laying he...
Mr. Curtis has 325 chickens on his farm where ...	277	<reasoning>Your observation is incorrect. 28 is more than half of 32...	297	false	have on his farm?



GRPO in ~10 Lines of Code

GOSIM

```
from trl import GRPOTrainer
from datasets import load_dataset

dataset = load_dataset("your_dataset", split="train")

def reward_func(completions, **kwargs):
    # your logic here!
    return rewards

trainer = GRPOTrainer(
    model="Qwen/Qwen2.5-32B-Instruct",
    reward_funcs=reward_func,
    train_dataset=dataset,
)
trainer.train()
```



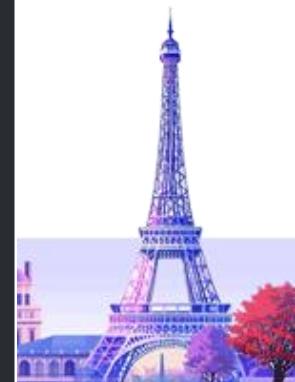
Reward Functions (advantages)

GOSIM

```
def extract_xml_answer(text: str) -> str:
    answer = text.split("<answer>") [-1]
    answer = answer.split("</answer>") [0]
    return answer.strip()

def format_reward_func(completions, **kwargs) -> list[float]:
    """Reward function that checks if the completion has a specific format."""
    pattern = r"^\<reasoning>\n.*?\n</reasoning>\n<answer>\n.*?\n</answer>\n$"
    responses = [completion[0] ["content"] for completion in completions]
    matches = [re.match(pattern, r) for r in responses]
    return [0.5 if match else 0.0 for match in matches]

def accuracy_reward_func(prompts, completions, answer, **kwargs) -> list[float]:
    """Reward function that extracts the answer from the xml tags and compares it to
    the correct answer."""
    responses = [completion[0] ['content'] for completion in completions]
    extracted_responses = [extract_xml_answer(r) for r in responses]
    return [2.0 if r == a else 0.0 for r, a in zip(extracted_responses, answer)]
```



Sample Reward Values

GOSIM

Mr. Curtis has 325 chickens on his farm where 28 are roosters and the rest are hens. Twenty hens do not lay eggs while the rest of the hens do. How many egg-laying hens does Mr. Curtis have on his farm?

```
# o_0 = <reasoning>I have some reasoning</reasoning><answer>12</answer>
r_0 = 1.0
```

```
# o_1 = <reasoning></reasoning><answer>12</answer>
r_1 = 0.5
```

```
# o_2 = The answer is 312
r_2 = 0.0
```

```
# o_3 = <reason>I did not have valid formatting or answer.
r_3 = 0.0
```



How the Math Works

GOSIM

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left(r_t(\theta) \hat{A}_{i,t}, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t} \right) \right] - \beta D_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}).$$

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$$

```
[>>> r = [1.0, 0.5, 0.0, 0.0]
[>>> np.mean(r)
0.375
[>>> np.std(r)
0.414578098794425
[>>> (1.0 - 0.375) / 0.414578098794425
1.507556722888818
[>>> [(r_i - np.mean(r)) / np.std(r) for r_i in r]
[1.507556722888818, 0.30151134457776363, -0.9045340337332909, -0.9045340337332909]
>>>
```



Keep rewards high level.

Let the model figure out specifics.

<think>

Let me break this down step by step...

</think>

<answer>

42



</answer>



Some Example Reward Functions

GOSIM

- Answer is correct?
- XML tag is present?
- Answer contains substring?
- Answer is formatted properly?
- Correct tool was called?
- Generated code compiles?
- Generated code passes unit tests?



Training a Small LM on Rust Coding



User Prompt

Write a rust function to take two vectors of numbers as input and return the merge sorted and deduplicated output.

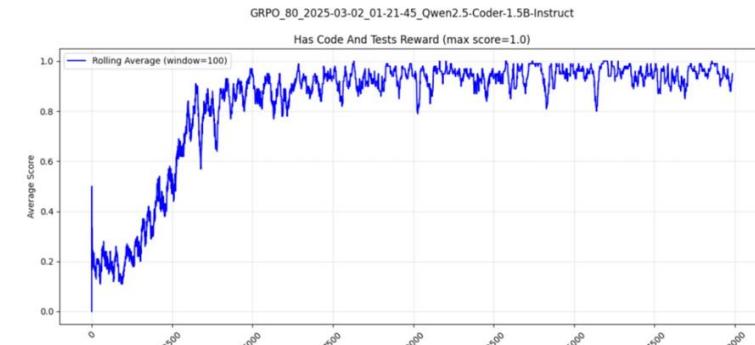
LLM Generated Code

```
fn merge_sorted(v1: &[i32], v2: &[i32]) -> Vec<i32> {  
    // all the logic in here...  
}
```

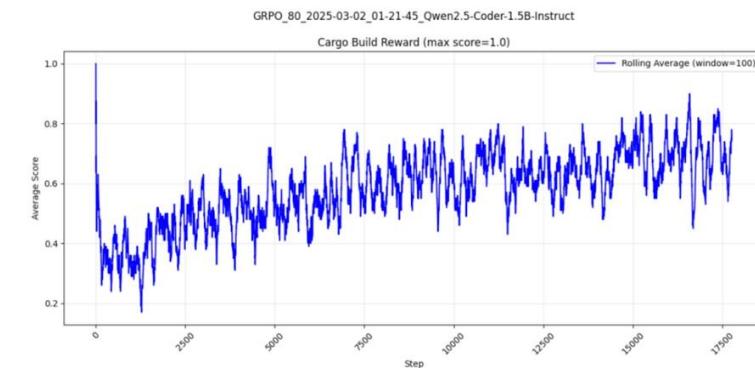
Verified Unit Tests

```
assert_eq!(merge_sorted(&[], &[]), Vec::new());  
assert_eq!(merge_sorted(&[1, 3], &[2, 4]), vec![1, 2, 3, 4]);  
assert_eq!(merge_sorted(&[1, 2, 2], &[2, 4]), vec![1, 2, 4]);
```

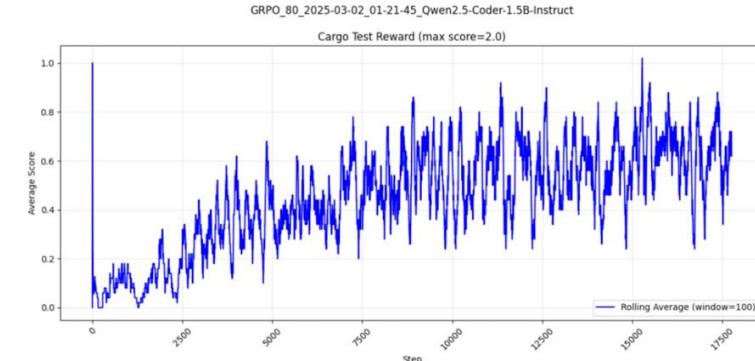
Learning to generate code+tests



Learning to compile the code



Learning to pass the unit tests



How small can we go?

GOSIM

The rStar-Math paper from Microsoft shows this in the domain of verifiable math problems allowing the model to reason. The 1.5B model outperforms GPT-4o and o1-preview.

Task (pass@1 Acc)	rStar-Math (Qwen-7B)	rStar-Math (Qwen-1.5B)	rStar-Math (Phi3-mini)	OpenAI o1-preview	OpenAI o1-mini	QWQ 32B-preview	GPT-4o	DeepSeek-V3
MATH	90.0	88.6	86.4	85.5	90.0	90.6	76.6	90.2
AIME 2024	53.3	46.7	43.3	44.6	56.7	50.0	9.3	39.2
Olympiad Bench	65.6	64.6	60.3	-	65.3	61.2	43.3	55.4
College Math	60.5	59.3	59.1	-	57.8	55.8	48.5	58.9
Omni-Math	50.5	48.5	46.0	52.5	60.5	49.6	30.5	35.9

<https://arxiv.org/abs/2501.04519>



Let's Talk Memory Requirements

GOSIM

DROP THE VALUE MODEL

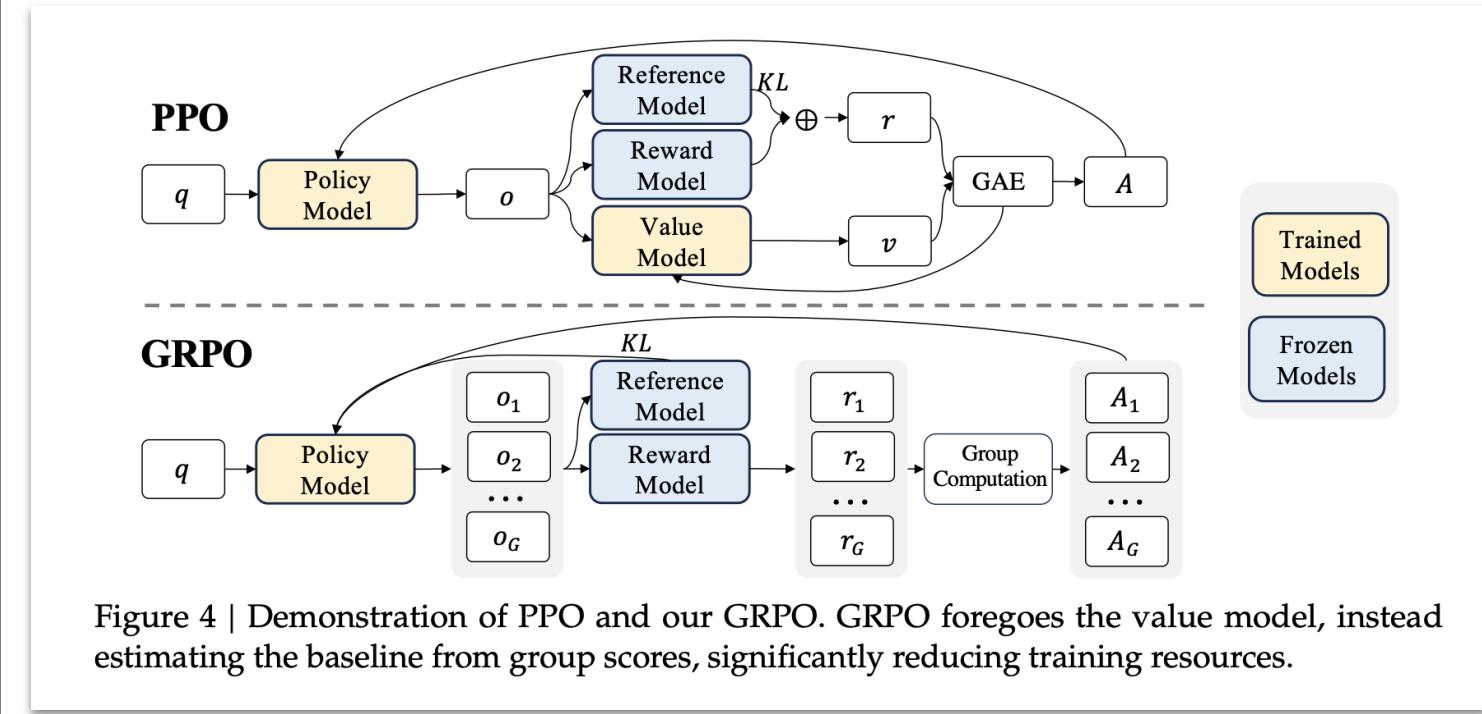
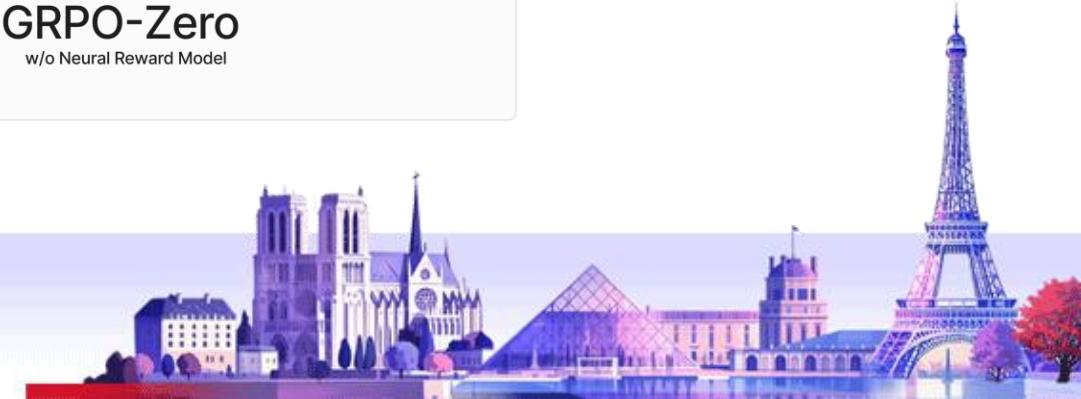
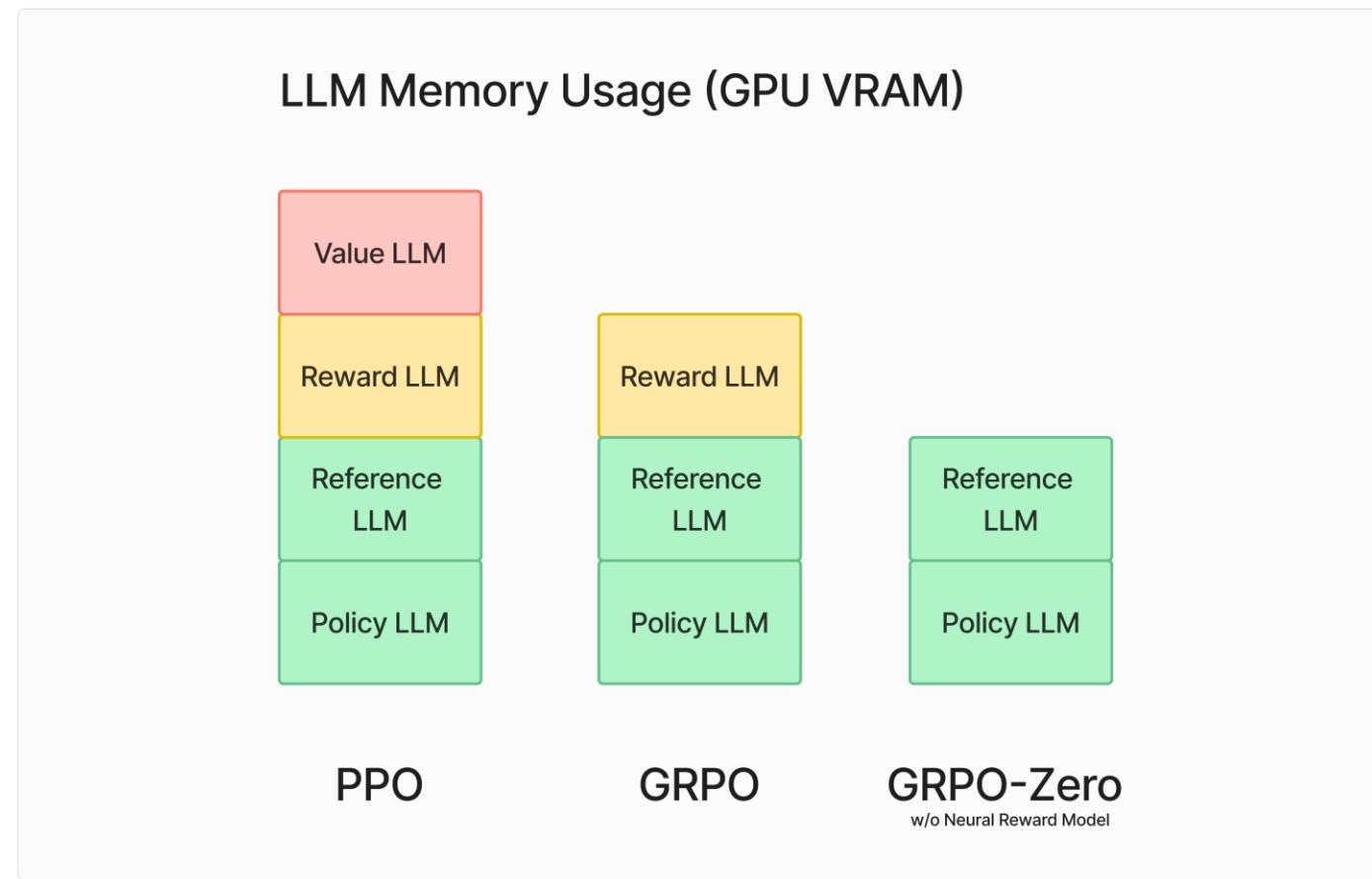


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

Let's Talk Memory Requirements

GOSIM

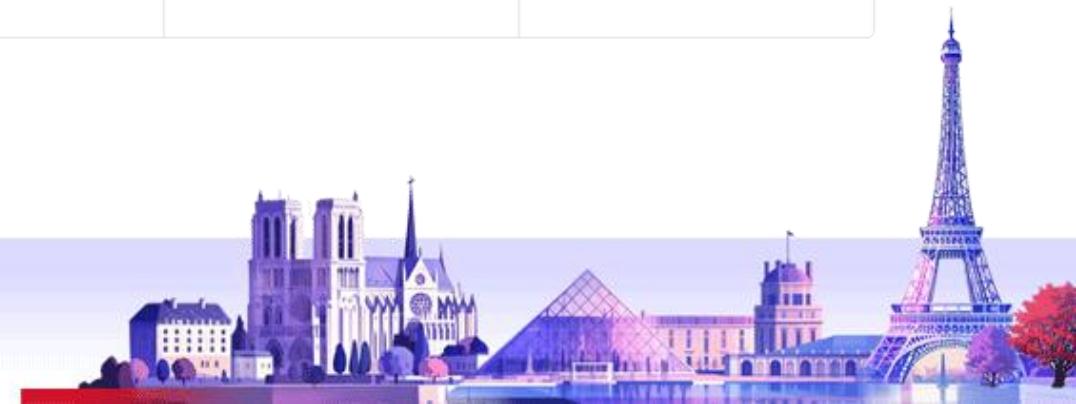


Using a Single H100

GOSIM

<https://www.oxen.ai/blog/grpo-vram-requirements-for-the-gpu-poor>

Method	0.5B	1B	3B	8B	14B
Fine Tune	25 GB	38 GB	58 GB	OOM 💥	OOM 💥
Fine Tune +optim=adam_8bit	16 GB	26 GB	52 GB	OOM 💥	OOM 💥
Fine Tune +optim=adam_8bit + gradient checkpointing	15 GB	24 GB	41 GB	72 GB	OOM 💥
PEFT LoRA	13 GB	16 GB	32 GB	58 GB	69 GB



Formulating the Problem

GOSIM

User Prompt

Write a rust function to take two vectors of numbers as input and return the merge sorted and deduplicated output.

LLM Generated Code

```
fn merge_sorted(v1: &[i32], v2: &[i32]) -> Vec<i32> {  
    // all the logic in here...  
}
```

Verified Unit Tests

```
assert_eq!(merge_sorted(&[], &[]), Vec::<i32>::new());  
assert_eq!(merge_sorted(&[1, 3], &[2, 4]), vec![1, 2, 3, 4]);  
assert_eq!(merge_sorted(&[1, 2, 2], &[2, 4]), vec![1, 2, 4]);
```

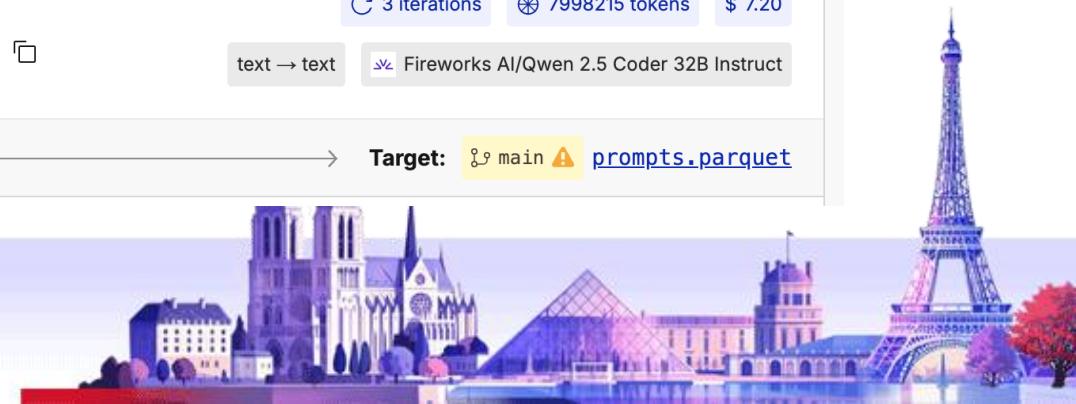


Generate Synthetic Data

GOSIM

Translate TIGER-Lab/AceCode-87K and filter out any examples that do pass clippy, build, and compile

Write 20k Unit Tests b886ddd6-b501-4db8-8ed6-0b719d0ac595	20000 rows	08:34:22	Completed
Bessie 5 days ago	2 iterations	12299605 tokens	\$ 11.07
Prompt: You are a pragmatic Rust programmer. Given the following question, and rust func... □	text → text Fireworks AI/Qwen 2.5 Coder 32B Instruct		
Source: main prompts.parquet	Target: main prompts.parquet		
Write 20k Rust Code Snippets febc562a-9bd4-4e91-88d7-a95ee676a5ed	20000 rows	03:49:57	Completed
Bessie 5 days ago	4 iterations	7577885 tokens	\$ 6.82
Prompt: You are a pragmatic Rust programmer. Given the following question, write a Rust ... □	text → text Fireworks AI/Qwen 2.5 Coder 32B Instruct		
Source: main prompts.parquet	Target: main 🚨 prompts.parquet		
Translate 20k questions to rust ce45630c-d9e8-4fac-9b41-2d41692076b3	20000 rows	09:04:08	Completed
Bessie 6 days ago	3 iterations	7998215 tokens	\$ 7.20
Prompt: The following prompt may have Python syntax, variable types, and pythonic behavi... □	text → text Fireworks AI/Qwen 2.5 Coder 32B Instruct		
Source: main prompts.parquet	Target: main 🚨 prompts.parquet		



Translate Prompt to Rust

GOSIM

└ main prompts.parquet

Type: text → text

Model: Qwen/Qwen 2.5 Coder 32B Instruct

Provider: Fireworks AI

Prompt

The following prompt may have Python syntax, variable types, and pythonic behavior mentioned. Translate the question to the appropriate Rust syntax and replace anything mentioning Python with Rust.

Do NOT use or mention:

- * &'static str
- * <&dyn Any>
- * lifetimes
- * Advanced rust objects
- * External libraries

...

Target field: rust_prompt

└ main prompts.parquet

View changes View file at commit

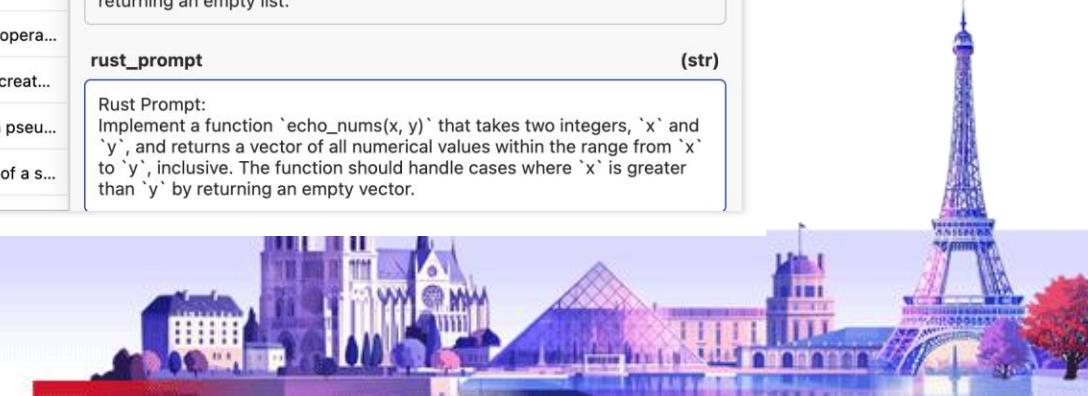
Queued: Feb 17, 2025, 9:38 PM GMT+1
Completed: Feb 18, 2025, 6:42 AM GMT+1
🕒 09:04:08
☰ 20000 rows
⌚ 7998215 tokens \$ 7.20

20000 rows processed, 7998215 tokens used (\$7.20)

✓ Completed

3 columns, 1-100 of 20000 rows			Row details
task_id	prompt	rust_prompt	
task_0	Implement a function `echo_nums(x, y)` that takes two integers, ...	Rust Prompt: Implement a function `echo_nums(x, y)` that takes ...	prompt (str) Implement a function `echo_nums(x, y)` that takes two integers, x and y, and returns a list of all numerical values within the range from x to y, inclusive. The function should handle cases where x is greater than y by returning an empty list.
task_1	Given the coordinates of a point on the x-axis and the center coor...	Given the coordinates of a point on the x-axis and the center coor...	
task_2	You are given a list of integer lists, where each inner list represent...	You are given a vector of integer vectors, where each vector...	
task_3	You are tasked with creating a function that simulates basic opera...	You are tasked with creating a function that simulates basic opera...	
task_4	You are given two integers, num1 and num2. Your task is to creat...	You are given two integers, num1 and num2. Your task is to creat...	
task_5	Write a function `generate_random_string` that generates a pseu...	Write a function `generate_random_string` that generates a pseu...	
task_6	Implement a function that generates a unique ID in the form of a s...	Implement a function that generates a unique ID in the form of a s...	rust_prompt (str) Rust Prompt: Implement a function `echo_nums(x, y)` that takes two integers, 'x' and 'y', and returns a vector of all numerical values within the range from 'x' to 'y', inclusive. The function should handle cases where 'x' is greater than 'y' by returning an empty vector.

GOSIM AI Paris 2025



Write Rust Code

GOSIM

main prompts.parquet

Type: text → text
Model: Qwen/Qwen 2.5 Coder 32B Instruct
Provider: Fireworks AI

Prompt

You are a pragmatic Rust programmer. Given the following question, write a Rust function to complete the task. Make the code simple and easy to understand. The code should pass `cargo build` and `cargo clippy`. Do not add a main function. Try to limit library usage to the standard library std. Respond with only the Rust function and nothing else.

Question:
{rust_prompt}

Code:

Target field: rust_code

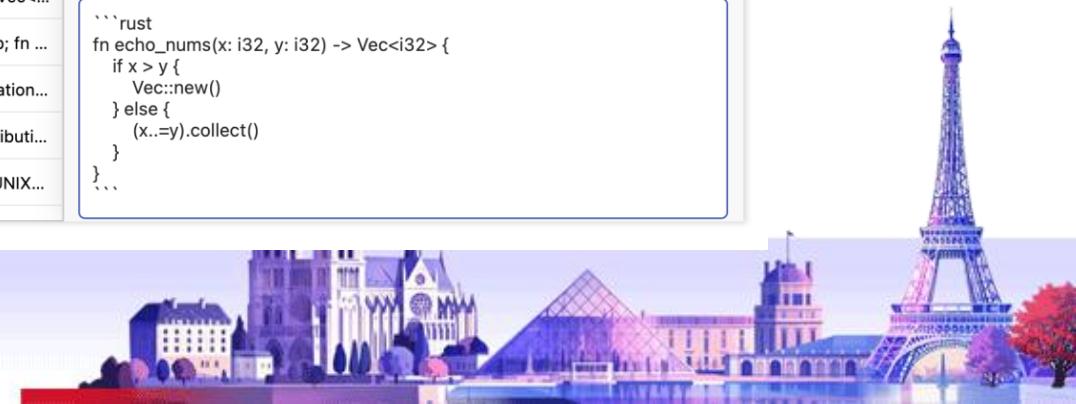
main prompts.parquet

View changes View file at commit

Queued: Feb 18, 2025, 7:21 PM GMT+1
Completed: Feb 19, 2025, 5:29 AM GMT+1
🕒 03:49:57
☰ 20000 rows
⌚ 7577885 tokens \$ 6.82

20000 rows processed, 7577885 tokens used (\$6.82) ✓ Completed

4 columns, 1-100 of 20000 rows				Row details
task_id	prompt	rust_prompt	rust_code	⋮
task_0	Implement a function `echo_nums(x, y)` t...	Rust Prompt: Implement a function `echo...	```rust fn echo_nums(x: i32, y: i32) -> Vec<i32> { if x > y { Vec::new() } else { (x..=y).collect() } }``````	to `y`, inclusive. The function should handle cases where `x` is greater than `y` by returning an empty vector.
task_1	Given the coordinates of a point on the x-...	Given the coordinates of a point on the x-...	```rust fn circle_y_coordinate(x: f32, r: f3...``````	
task_2	You are given a list of integer lists, where ...	You are given a vector of integer vectors, ...	```rust fn aggregate_sums(int_lists: Vec<Vec<i32>>) -> Vec<i32> { int_lists.iter().map(list list.sum()).sum() }``````	
task_3	You are tasked with creating a function th...	You are tasked with creating a function th...	```rust use std::collections::HashMap; fn ...``````	
task_4	You are given two integers, num1 and nu...	You are given two integers, num1 and nu...	```rust #[derive(Debug)] enum Operation { Add(i32), Sub(i32), Mul(i32), Div(i32) }``````	
task_5	Write a function `generate_random_strin...	Write a function `generate_random_strin...	```rust use rand::Rng; use rand::distributi...``````	
task_6	Implement a function that generates a uni...	Implement a function that generates a uni...	```rust use std::time::{SystemTime, UNIX_EPOCH}; fn ...``````	



Write Unit Tests

GOSIM

main prompts.parquet

Type: text → text
Model: Qwen/Qwen 2.5 Coder 32B Instruct
Provider: Fireworks AI

Prompt

You are a pragmatic Rust programmer. Given the following question, and rust function, write three unit tests for the function. The tests should be a simple line delimited list of assert! or assert_eq! statements.

For example, if the function is:

```
```rust
fn add_nums(x: i32, y: i32) -> i32 {
 x + y
}
```

```

Target field: rust_test_list

View changes View file at commit

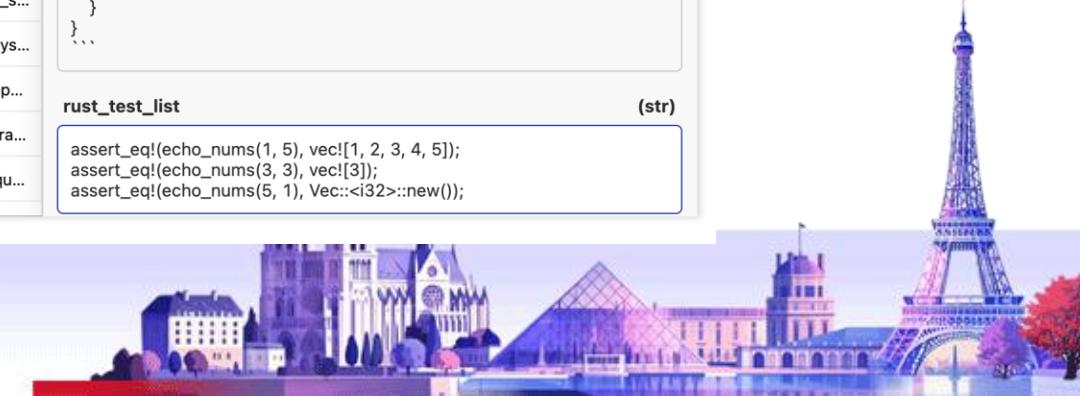
Queued: Feb 19, 2025, 4:07 PM GMT+1
Completed: Feb 20, 2025, 12:42 AM GMT+1
08:34:22
20000 rows
12299605 tokens \$ 11.07

20000 rows processed, 12299605 tokens used (\$11.07)

✓ Completed

| 5 columns, 1-100 of 20000 rows | | | | | Row details |
|--------------------------------|-------------------------------|--|------------------------------|----------------------------|---|
| task_id | prompt | rust_prompt | rust_code | rust_test_list | |
| task_0 | Implement a function `... | Rust Prompt: Implement a function `echo_nums(... | ```rust fn echo_nums(x... | assert_eq!(echo_nums(... | <pre>fn echo_nums(x: i32, y: i32) -> Vec<i32> { if x > y { Vec::new() } else { (x..=y).collect() } } ``` </pre> |
| task_1 | Given the coordinates o... | Given the coordinates of a point on the x-axis an... | ```rust fn circle_y_coor... | assert_eq!(circle_y_coo... | |
| task_2 | You are given a list of in... | You are given a vector of integer vectors, where ... | ```rust fn aggregate_s... | assert_eq!(aggregate_s... | |
| task_3 | You are tasked with cre... | You are tasked with creating a function that simu... | ```rust use std::collecti... | assert_eq!(banking_sys... | |
| task_4 | You are given two integ... | You are given two integers, num1 and num2. You... | ```rust #[derive(Debug...] | assert_eq!(process_op... | |
| task_5 | Write a function `gener... | Write a function `generate_random_string` that ... | ```rust use rand::Rng; ... | assert_eq!(generate_ra... | |
| task_6 | Implement a function th... | Implement a function that generates a unique ID ... | ```rust use std::time::...; | assert!(generate_uniqu... | |

GOSIM AI Paris 2025

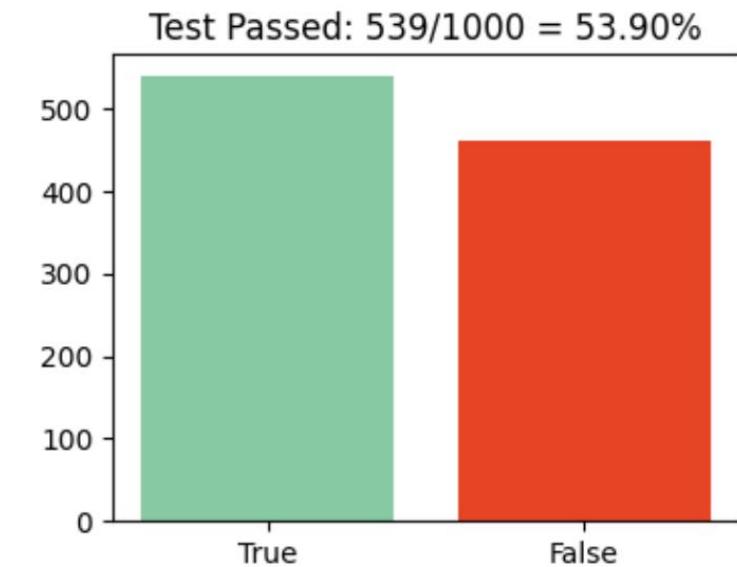
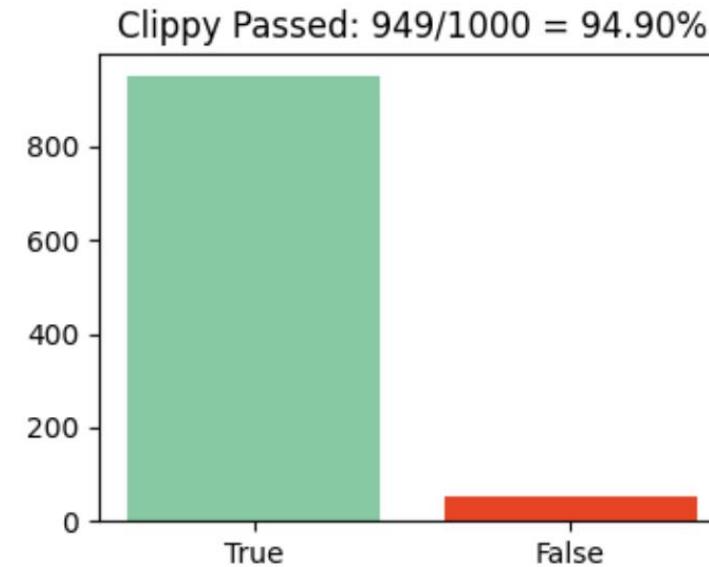
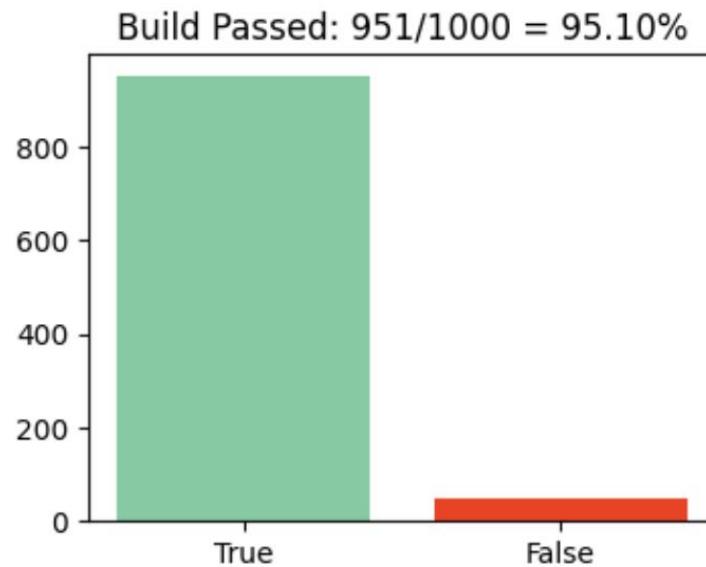


GPT-4o Results



GOSIM

Results GPT-4o



Look At Your Data (in Oxen.ai)

The logo for GOSIM, featuring the word "GOSIM" in a bold, sans-serif font. The letter "O" is stylized with a circular gradient fill transitioning from teal at the top to purple at the bottom.

OXEN.AI ox / Rust □

Datasets Models Docs Blog Pricing Search + @ox ▾

Data Branches Merge Requests Evaluations Notebooks Settings

gpt-4o-results ▼ Rust / results / GPT4-o / results.parquet □

ox add code renderer 39c5f60bee1141a542a592a4697fa6aa · just now 562 Commits

Query this dataset. Ask me anything! Go

Schema 13 columns, 1-100 of 1000 rows Row details

| task_id | (str) | rust_test_list | response | test_list | template | build_passed | build_stderr | clippy_passed | clippy_stderr | test_passed | test_stderr |
|------------|----------------|----------------|----------------|-----------|----------|--------------|--------------|----------------|---------------|-----------------|-------------|
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | false | | error: test ... | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | true | | | |
| assert_eq! | ``rust us... | assert_eq!... | #![allow(d...) | true | | true | | false | | error: test ... | |
| assert_eq! | ``rust str... | assert_eq!... | #![allow(d...) | true | | true | | false | | error[E042... | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | true | | | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | true | | | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | true | | | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | false | | error: test ... | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | false | | error[E006... | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | true | | | |
| assert_eq! | ``rust fn i... | assert_eq!... | #![allow(d...) | true | | true | | warning: le... | | true | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | false | | error: test ... | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | false | | error: test ... | |
| assert_eq! | ``rust fn ... | assert_eq!... | #![allow(d...) | true | | true | | false | | warning: u... | |

build_passed (bool)

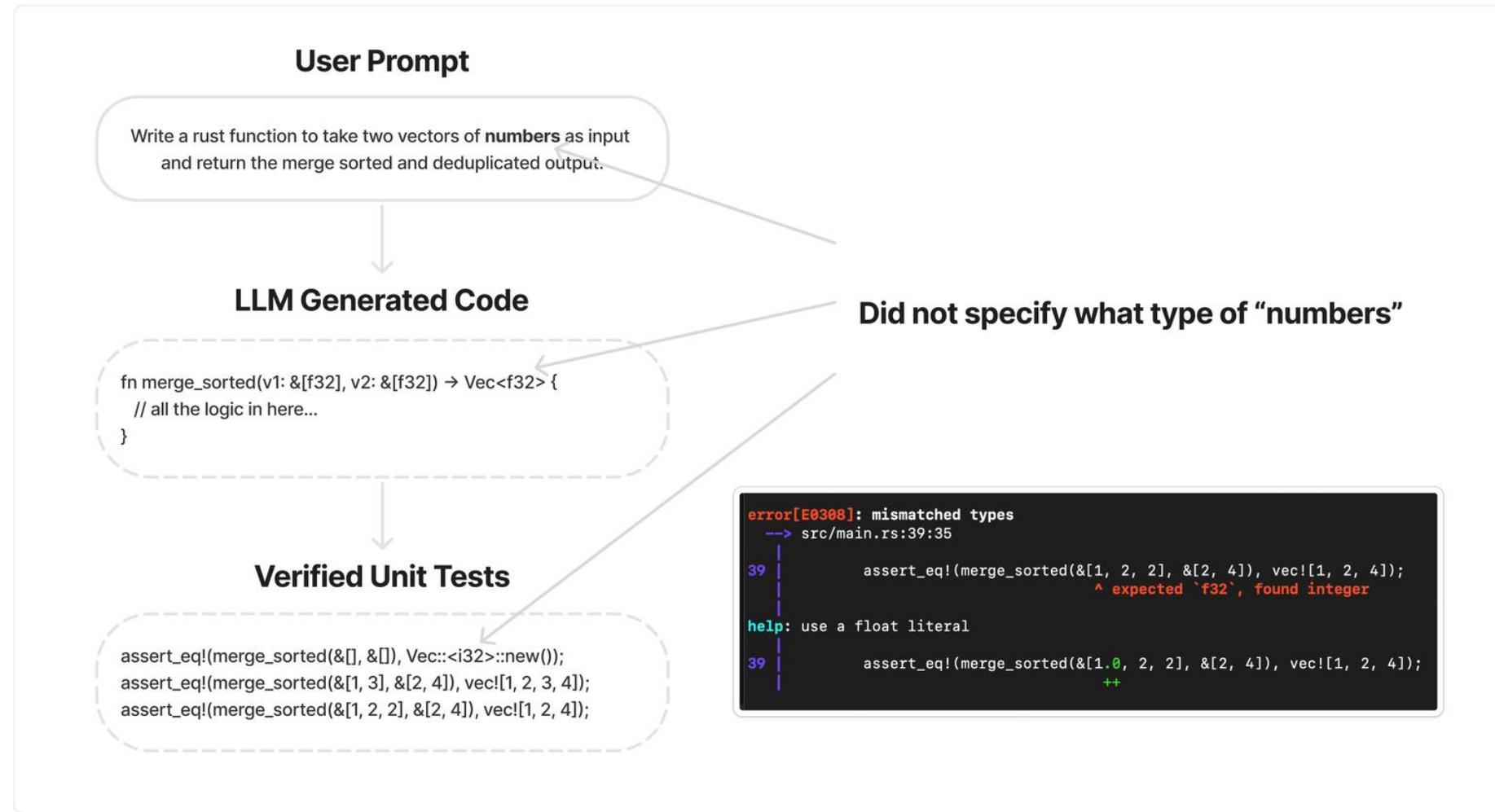
true false null

https://www.oxen.ai/ox/Rust

Three Steps is Hard

GOSIM

If you don't know the function spec before hand, you are going to fail a lot of unit tests



Simplify the Problem

GOSIM

Generate code + Unit Tests Together

User Prompt

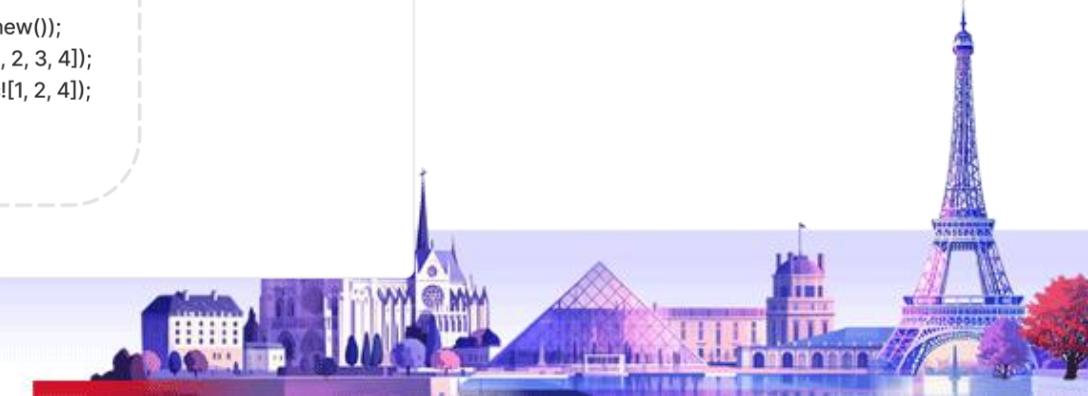
Write a rust function to take two vectors of numbers as input
and return the merge sorted and deduplicated output.



LLM Generated Code + Tests

```
fn merge_sorted(v1: &[i32], v2: &[i32]) -> Vec<i32> {  
    // all the logic in here...  
}  
  
#[cfg(test)]  
mod tests {  
    use super::*;

    #[test]  
    fn test_merge_sorted() {  
        assert_eq!(merge_sorted(&[], &[]), Vec::new());  
        assert_eq!(merge_sorted(&[1, 3], &[2, 4]), vec![1, 2, 3, 4]);  
        assert_eq!(merge_sorted(&[1, 2, 2], &[2, 4]), vec![1, 2, 4]);  
    }
}
```



Add a System Prompt

GOSIM

To describe the output format we want

System Prompt

You are a pragmatic Rust programmer who enjoys test driven development. Given the following question, write a Rust function to complete the task. Make the code simple and easy to understand. The code should pass 'cargo build' and 'cargo clippy'. Try to limit library usage to the standard library std. Be careful with your types, and try to limit yourself to the basic built in types and standard library functions. When writing the function you can think through how to solve the problem and perform reasoning in the comments above the function.

Then write unit tests for the function you defined. Write three unit tests for the function. The tests should be a simple line delimited list of assert! or assert_eq! statements. When writing the unit tests you can have comments specifying what you are testing in plain english. The tests should use super::*.

An example output should look like the following:

```
```rust
/// Reasoning goes here
/// and can be multi-line
fn add_nums(x: i32, y: i32) -> i32 {
 x + y
}

#[cfg(test)]...
```

## User Prompt

Write a rust function to take two vectors of numbers as input and return the merge sorted and deduplicated output.

## LLM Generated Code + Tests

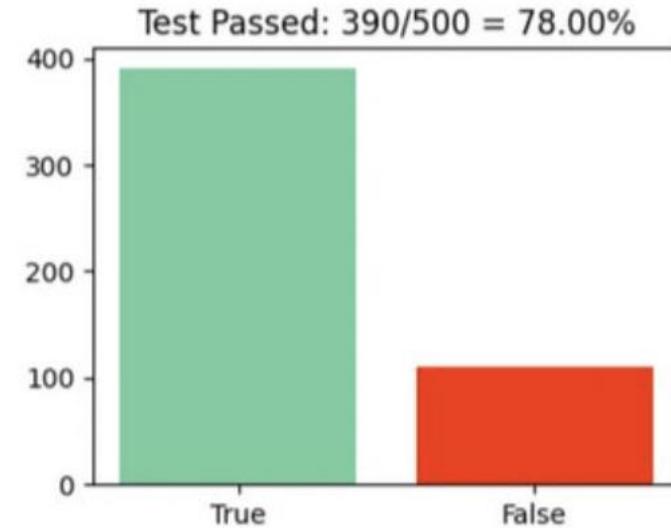
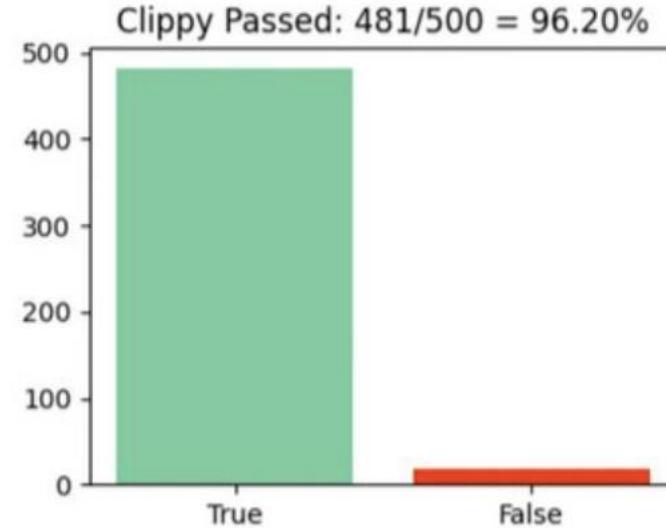
```
fn merge_sorted(v1: &[i32], v2: &[i32]) -> Vec<i32> {
 // all the logic in here...
}

#[cfg(test)]
mod tests {
 use super::*;

 #[test]
 fn test_merge_sorted() {
 assert_eq!(merge_sorted(&[], &[]), Vec::new());
 assert_eq!(merge_sorted(&[1, 3], &[2, 4]), vec![1, 2, 3, 4]);
 assert_eq!(merge_sorted(&[1, 2, 2], &[2, 4]), vec![1, 2, 4]);
 }
}
```



## Results GPT-4o



# Training data, simplified

GOSIM

- Training data only need prompts
- Filter out any prompts that a strong model cannot complete the task on
- For each prompt, generate N answers, score with rubrics

OXEN ox / Rust □ Datasets Models Docs Blog Pricing Search + @ox ▾

Data Branches Merge Requests Evaluations Notebooks Settings

main Rust / cargo\_test\_passed\_train.parquet □ Edit

ox Update cargo\_test\_passed\_train to render code dc1f7a1353c680ce31dbe74e4d9ced0a · 7 minutes ago 566 Commits

Query this dataset. Ask me anything! Go

Schema	task_id	rust_prompt
task_id (str)	task_0	Implement a function `echo_nums(x, y)` that takes two integers, `x` and `y`, and returns a vector of all numerical values within the range f...
rust_prompt (str)	task_1	Given the coordinates of a point on the x-axis and the center coordinates of a circle, along with the radius of the circle, write a function `cir...
rust_code (str)	task_2	You are given a vector of integer vectors, where each inner vector represents a group of integers. Write a function that takes this vector an...
rust_test_list (str)	task_3	You are tasked with creating a function that simulates basic operations of a banking system. The function should allow users to manage mu...
	task_4	You are given two integers, num1 and num2. Your task is to create a function that takes these two integers as input and performs the follow...
	task_8	Write a function `calculate_rectangle_properties(length: i32, width: i32) -> (i32, i32, f32)` that takes the length and width of a rectangle a...
	task_9	You are given a vector of integers. Your task is to create a function that determines if the sum of the integers in the vector is even or odd. Y...
	task_10	You are tasked with implementing a function that retrieves information from a given hash map based on an invitation code. The function sh...
	task_11	You are given two vectors of integers, `list1` and `list2`. Your task is to implement a function that returns a vector of integers that are pres...
	task_12	Given a string representation of a Rust value, implement a function that formats the string to return just the actual value contained in it. The...

< 1 2 3 4 5 6 7 8 9 10 ... 146 147 148 149 150 >

# Our Rubric

GOSIM

- response contains one code block
- response contains one test block
- test has assert statements
- cargo build passes
- cargo clippy passes
- cargo test passes

## rust\_prompt

You are given a vector of integers representing the scores of different models evaluated over multiple epochs. Your task is to implement a function `best\_model\_epoch(scores: Vec<i32>) → (i32, i32)` that returns a tuple containing the epoch number (1-indexed) and the highest accuracy score from the vector. If there are multiple epochs with the same highest score, return the earliest epoch. The input will be a vector of integers where each integer represents the accuracy score for a model at that specific epoch. The function should return the epoch number and the highest score as a tuple. For example, if the input is `[80, 85, 85, 90, 75]`, the output should be `(4, 90)`.

## completion

```
```rust
fn best_model_epoch(scores: Vec<i32>) -> (i32, i32) {
    let mut highest_score = 0;
    let mut best_epoch = 1;
    for (i, &score) in scores.iter().enumerate() {
        if score > highest_score {
            highest_score = score;
            best_epoch = i as i32 + 1;
        }
    }
    (best_epoch, highest_score)
}

#[cfg(test)]
mod tests {
    use super::*;

    #[test]
    fn test_best_model_epoch() {
        assert_eq!(best_model_epoch(vec![80, 85, 85, 90, 75]), (4, 90));
        assert_eq!(best_model_epoch(vec![95, 95, 95, 95]), (1, 95));
        assert_eq!(best_model_epoch(vec![60, 60, 60, 60]), (1, 60));
    }
}
```

```



# Example Rubric: Has Unit Test?

GOSIM

A reward function could be as simple as a regex requiring that we have a valid test block in our code.

```
Simple regex for checking if a code block has a rust tests module
def code_has_test_block(code: str) -> Optional[str]:
 # Use re.DOTALL to make '.' match newlines as well
 result = re.search(
 r'(#\[cfg\((test)\)\])\s*mod\s+tests\s*\{.*?\}">',
 code,
 re.DOTALL
)
 return 1.0 if result else 0.0

For each prompt and completion, reward if the response has a test block
def test_block_reward_func(prompts, completions, **kwargs) -> list[float]:
 contents = [completion[0]["content"] for completion in completions]
 return [code_has_test_block(c) for c in contents]
```



# Putting it all together

GOSIM

<https://github.com/Oxen-AI/GRPO-With-Cargo-Feedback/blob/main/train.py>

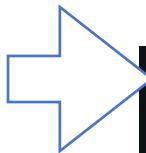
```
trainer = GRPOTrainer(
 model=model,
 processing_class=tokenizer,
 reward_funcs=[
 cargo_build_reward_func, # 1.0 if passes cargo build else 0.0
 cargo_clippy_reward_func, # 1.0 if passes cargo clippy else 0.0
 cargo_test_reward_func, # 3.0 if passes cargo test else 0.0
 non_empty_reward_func, # 1.0 if the code is not empty else 0.0
 test_block_count_reward_func, # 1.0 if there is a test block else 0.0
 tests_have_asserts_reward_func # 1.0 if there are assert statements in the test else 0.0
 args=training_args,
 train_dataset=train_dataset,
 peft_config=peft_config, # None if !use_peft_checkbox.value
 callbacks=[OxenTrainerCallback(experiment, bar, commit_every=commit_every.value)],
)
trainer.train()
```



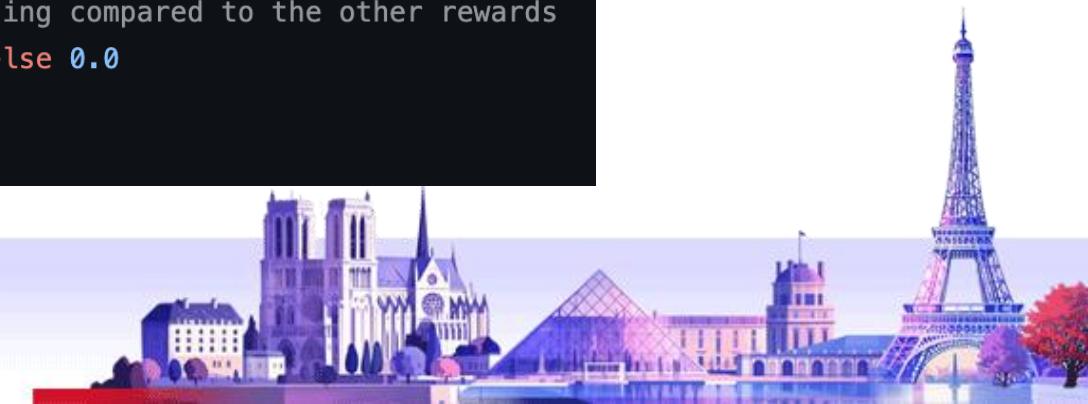


# Monitoring in Oxen.ai

GOSIM



```
@experiment.log(f"cargo_test_rewards.jsonl")
def cargo_test_reward_func(prompts, completions, **kwargs) -> list[float]:
 # Extract the answers from the completions
 responses = [completion[0]['content'] for completion in completions]
 extracted_codes = [extract_rust_code(r) for r in responses]
 extracted_tests = [extract_test_code(c) for c in extracted_codes]
 results = []
 for i, answer in enumerate(extracted_codes):
 score = 0.0
 if extracted_tests[i]:
 data = {
 'rust_code': answer
 }
 tools = [RustTool("test")]
 cargo_results = setup_and_test_rust_project(data, tools)
 # Let's give some extra credit for tests passing compared to the other rewards
 score = 2.0 if cargo_results['test_passed'] else 0.0
 results.append(score)
 return results
```





# Monitoring in Oxen.ai

GOSIM

GRPO\_82\_2025-03-02\_22-49-1...

Rust / outputs / GRPO\_82\_2025-03-02\_22-49-17\_Qwen2.5-Coder-1.5B-Instruct / cargo\_test\_rewards.jsonl



Edit



Bessie step end GRPO

6d2b972892f1e1092715530599740b99 · 2 months ago ⏲ 1783 Commits

Query this dataset. Ask me anything!

Go

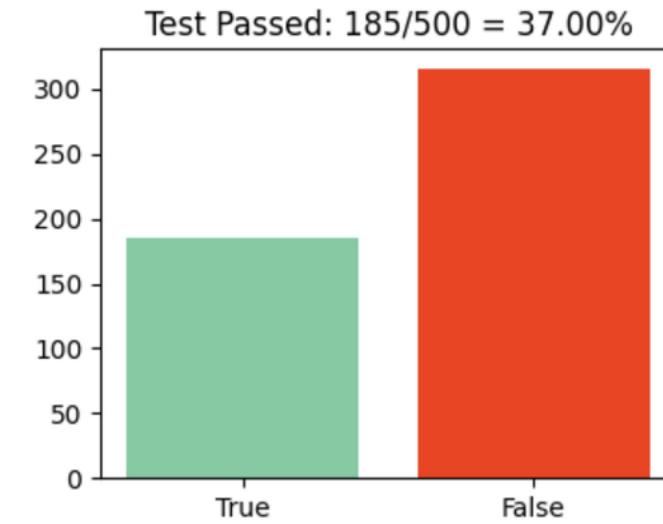
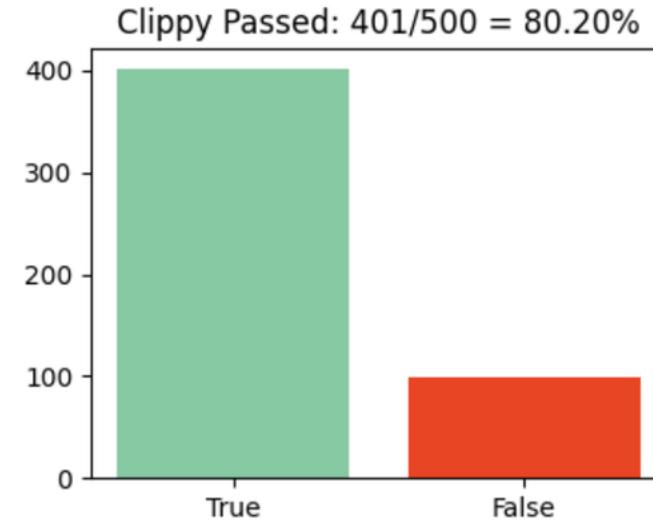
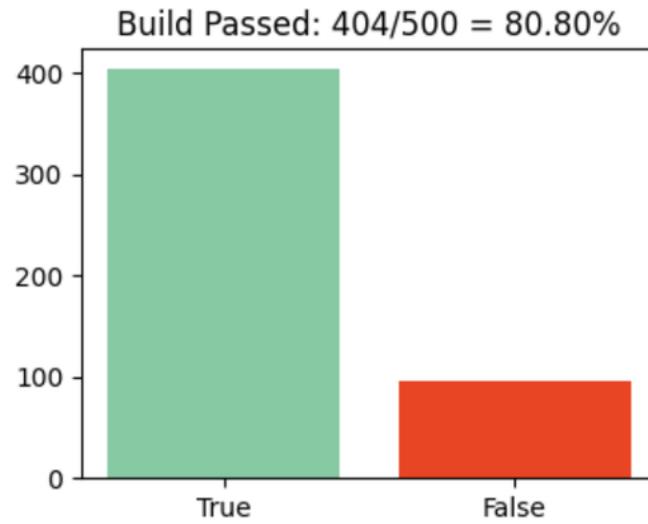


| Schema                 | Row details                                                                                                                 |
|------------------------|-----------------------------------------------------------------------------------------------------------------------------|
|                        | 7 columns, 59701-59800 of 60000 rows                                                                                        |
| timestamp (str)        | <b>timestamp</b> function score task_id rust_prompt completion func_execution_time                                          |
| function (str)         | 2025-03-05T15:15:... cargo_test_reward_f... 0.000000 task_9750 You are given a confi... ````rust #[derive(Deb... 1.912210   |
| score (f64)            | 2025-03-05T15:15:... cargo_test_reward_f... 2.000000 task_9750 You are given a confi... ````rust pub struct C... 1.912354   |
| task_id (str)          | 2025-03-05T15:16:... cargo_test_reward_f... 0.000000 task_144 You are tasked with i... ````rust fn perform_o... 2.509563    |
| rust_prompt (str)      | 2025-03-05T15:16:... cargo_test_reward_f... 0.000000 task_144 You are tasked with i... ````rust fn perform_o... 2.509850    |
| completion (str)       | 2025-03-05T15:16:... cargo_test_reward_f... 2.000000 task_144 You are tasked with i... ````rust fn perform_o... 2.509967    |
| func_executio... (f64) | 2025-03-05T15:16:... cargo_test_reward_f... 2.000000 task_144 You are tasked with i... ````rust fn perform_o... 2.510068    |
|                        | 2025-03-05T15:16:... cargo_test_reward_f... 0.000000 task_5969 You are given a vect... ````rust fn best_mod... 2.225187     |
|                        | 2025-03-05T15:16:... cargo_test_reward_f... 2.000000 task_5969 You are given a vect... ````rust fn best_mod... 2.225414     |
|                        | 2025-03-05T15:16:... cargo_test_reward_f... 0.000000 task_5969 You are given a vect... ````rust use std::cmp... 2.225532    |
|                        | 2025-03-05T15:16:... cargo_test_reward_f... 0.000000 task_5969 You are given a vect... ````rust fn best_mod... 2.225657     |
|                        | 2025-03-05T15:16:... cargo_test_reward_f... 0.000000 task_12600 Implement a functio... ````rust use std::iter::... 1.235403 |
|                        | 2025-03-05T15:16:... cargo_test_reward_f... 0.000000 task_12600 Implement a functio... ````rust fn encode_b... 1.235701     |
|                        | .....                                                                                                                       |

You are given a vector of integers representing the scores of different models evaluated over multiple epochs. Your task is to implement a function `best\_model\_epoch(scores: Vec<i32>) -> (i32, i32)` that returns a tuple containing the epoch number (1-indexed) and the highest accuracy score from the vector. If there are multiple epochs with the same highest score, return the earliest epoch. The input will be a vector of integers where each integer represents the accuracy score for a model at that specific epoch. The function should return the epoch number and the highest score as a tuple. For example, if the input is `[80, 85, 85, 90, 75]`, the output should be `(4, 90).

# The Baseline

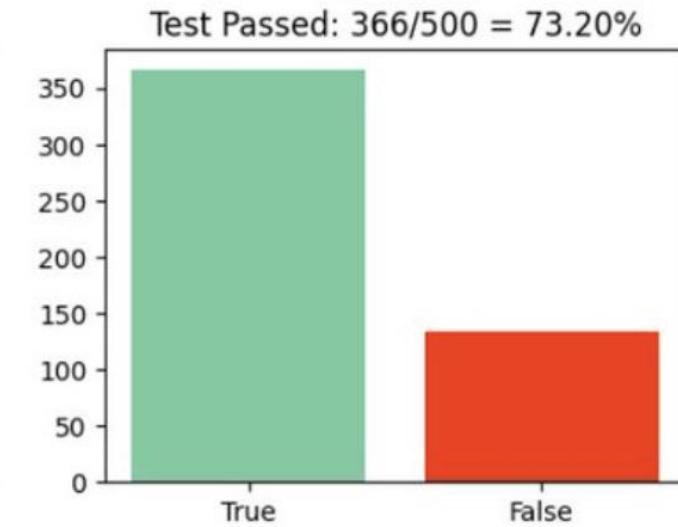
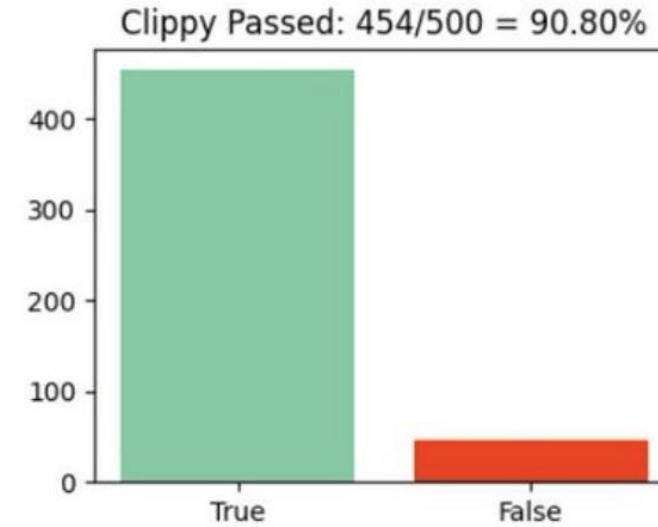
GOSIM



# The Results



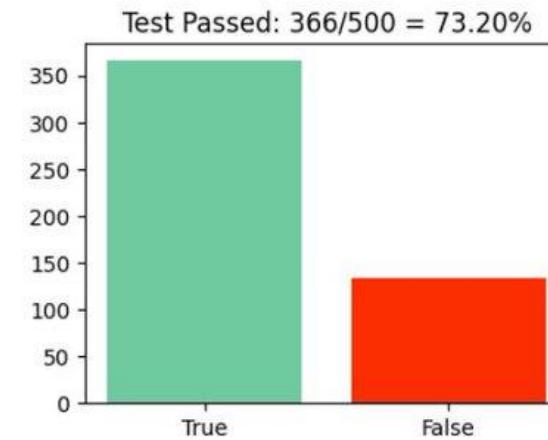
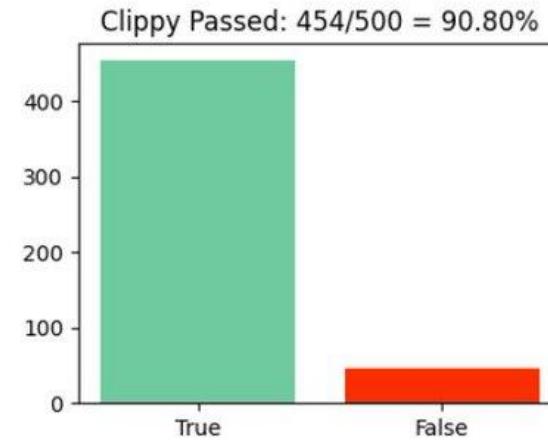
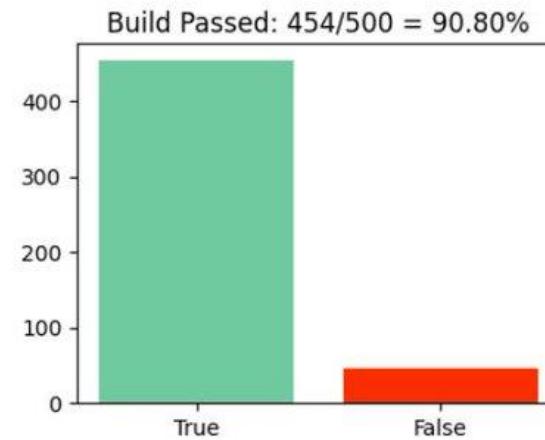
GOSIM



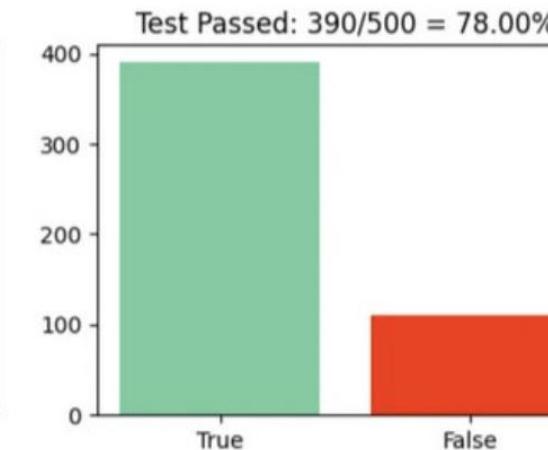
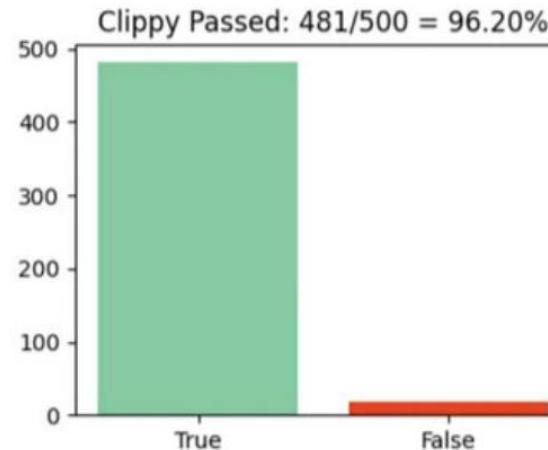
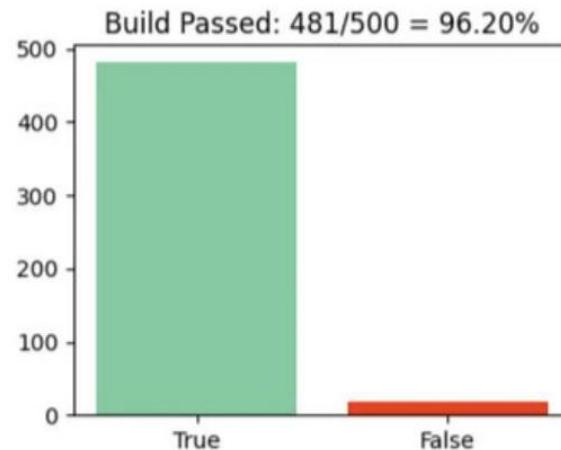
# GRPO-3B vs GPT-4o

GOSIM

## Results GRPO\_Qwen2.5-Coder-3B-Instruct



## Results GPT-4o





# Try it yourself

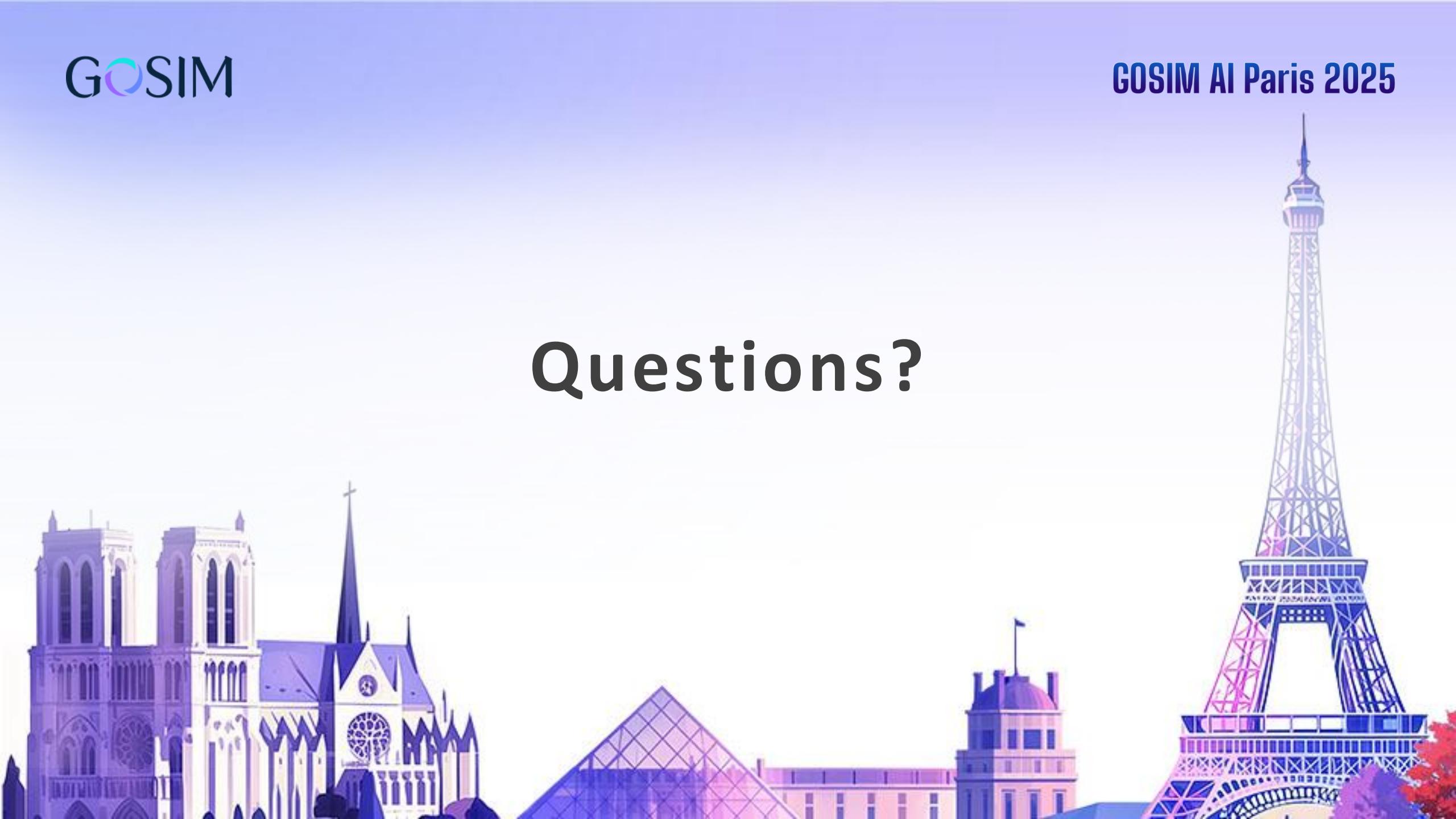
GOSIM

- 🔥 Free Credits for GPU Notebooks in Oxen.ai
- 📊 Generate your own synthetic data
- ✅ Build your own reward rubric



<https://github.com/Oxen-AI/GRPO-With-Cargo-Feedback>



A stylized silhouette of the Paris skyline, featuring the Eiffel Tower on the right, the Louvre Pyramid in the center, and the Notre Dame Cathedral on the left, all rendered in a light purple color against a white background.

# Questions?

# THANK YOU

