

# Reasoning Models

S1: Simple test-time scaling

LIMO: Less is More for Reasoning

Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning

**Charlie Ruan, Kaiwen Hu**

09/30


# Reasoning Models

- Background on reasoning models
- Paper #1: S1: Simple test-time scaling
- Paper #2: LIMO: Less is More for Reasoning
- Paper #3: Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning

# Reasoning Models

- Background on reasoning models
- Paper #1: S1: Simple test-time scaling
- Paper #2: LIMO: Less is More for Reasoning
- Paper #3: Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning

# Timeline

- **Aug., 2024: “Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters”, “Inference Scaling Laws: ...”** 
- **Sept. 12, 2024:** OpenAI o1 preview dropped
- **Nov. 28 2024:** Alibaba QwQ-32B-Preview (open weights, no disclosed methodology)
- **Dec. 5, 2024:** OpenAI o1 fully released
- **Jan. 8, 2025:** Sky-T1 (17k data distilled from QwQ-32B-Preview)
- **Jan. 22, 2025:** DeepSeek R1 —> on par with o1
- **Jan. 22, 2025:** Bespoke-32B (using Sky-T1 data pipeline but distilled from DeepSeek-R1)
- **Jan. 31, 2025:** S1 (1k data, distilled from Gemini 2.0 Flash Thinking Experimental)
- **Feb. 5, 2025:** LIMO (paper #2)

# Test-time scaling

- Previously, the performance of LMs have relied on scaling up **train-time compute** (# of parameters, data)
- In ~2024, people start exploring a new scaling paradigm: increase compute at test time to get better results – i.e. **test-time scaling**

Google DeepMind

2024-8-7

## Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters

Charlie Snell<sup>♣, 1</sup>, Jaehoon Lee<sup>2</sup>, Kelvin Xu<sup>♣, 2</sup> and Aviral Kumar<sup>♣, 2</sup>

♣Equal advising, <sup>1</sup>UC Berkeley, <sup>2</sup>Google DeepMind, ♣Work done during an internship at Google DeepMind

[Submitted on 1 Aug 2024 (v1), last revised 3 Mar 2025 (this version, v3)]

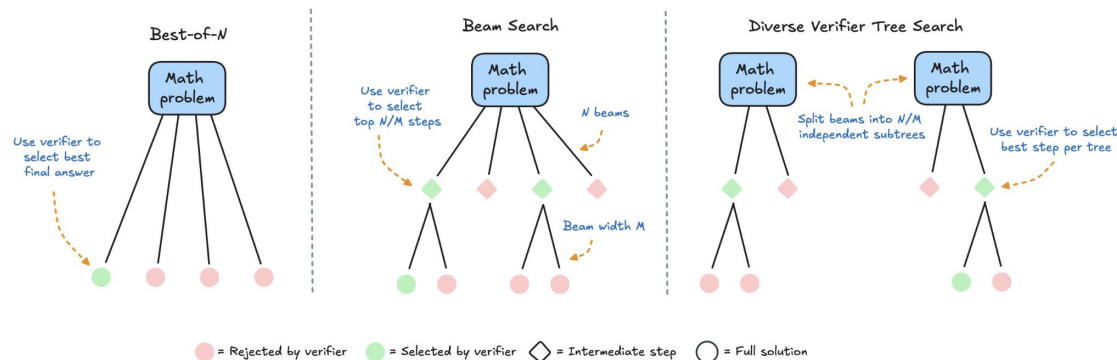
## Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, Yiming Yang

# Test-time scaling (Cont.) – sequential vs. parallel

s1 paper classifies test-time scaling into:


- 1) **Sequential:** later computations depend on earlier ones (e.g. reasoning)
- 2) **Parallel:** computations run independently (e.g. Large Language Monkey, Brown et al., 2024)



<https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute>

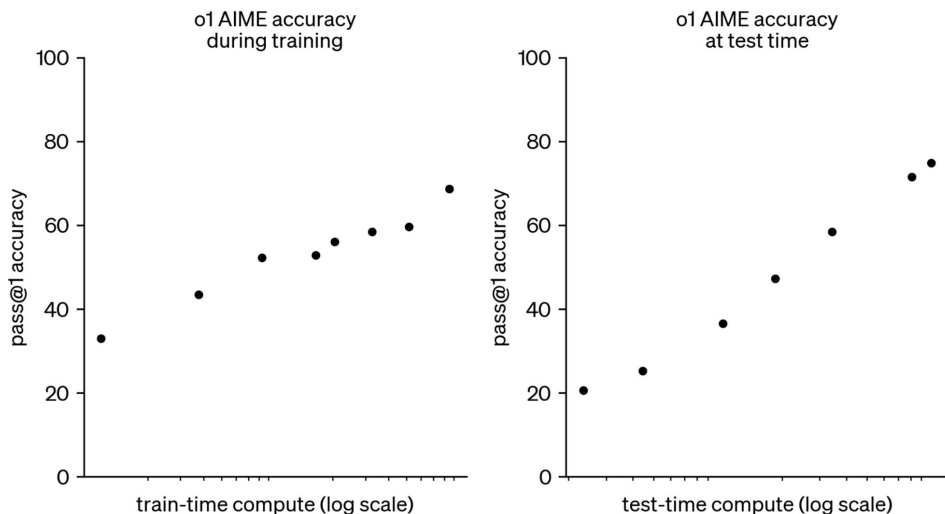
We focus on the first – particularly, reasoning!

# Timeline

- **Aug. 6, 2024:** “Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters”, “Inference Scaling Laws: ...”
- **Sept. 12, 2024: OpenAI o1 preview dropped** 
- **Nov. 28 2024:** Alibaba QwQ-32B-Preview (open weights, no disclosed methodology)
- **Dec. 5, 2024:** OpenAI o1 fully released
- **Jan. 8, 2025:** Sky-T1 (17k data distilled from QwQ-32B-Preview)
- **Jan. 22, 2025:** DeepSeek R1 —> on par with o1
- **Jan. 22, 2025:** Bespoke-32B (using Sky-T1 data pipeline but distilled from DeepSeek-R1)
- **Jan. 31, 2025:** S1 (1k data, distilled from Gemini 2.0 Flash Thinking Experimental)
- **Feb. 5, 2025:** LIMO (paper #2)

# Sept 12, 2024: OpenAI o1-preview

- On Sept 12, 2024, OpenAI dropped o1-preview
- Validates the test-time scaling paradigm



o1 performance smoothly improves with both train-time and test-time compute

<https://openai.com/index/learning-to-reason-with-lms/>



# OpenAI o1-preview (Cont.) – what is reasoning?

On OpenAI's website, they define *reasoning models* as:

- LLMs trained with reinforcement learning to perform reasoning
- Reasoning models *think before they answer*, producing *a long internal chain of thought* before responding to the user.

<https://platform.openai.com/docs/guides/reasoning>

---

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 ...

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

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: ...

...

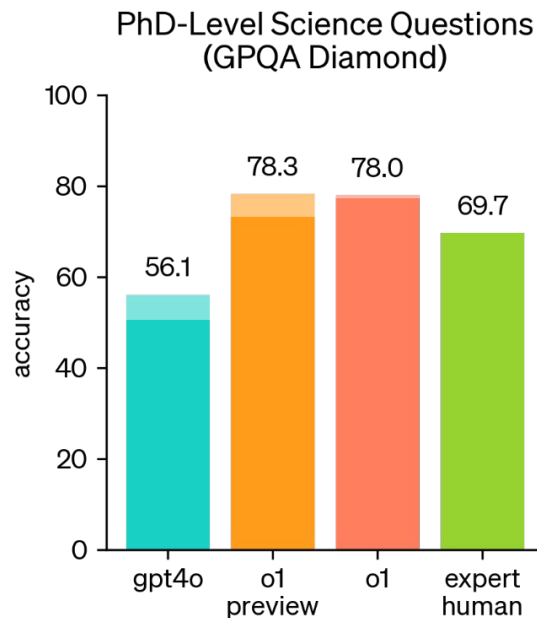
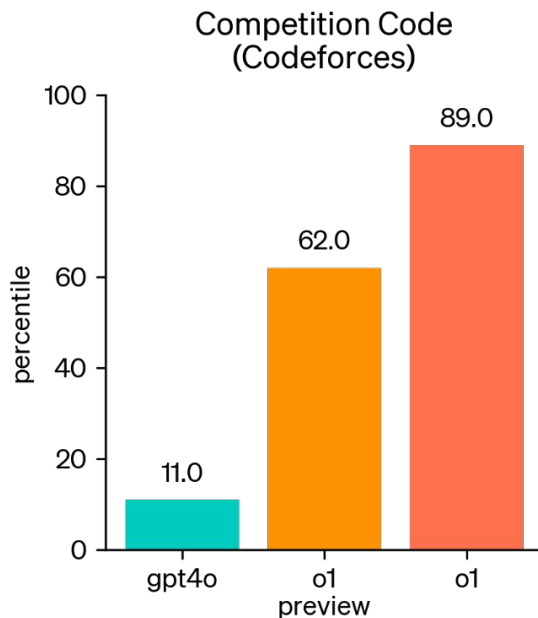
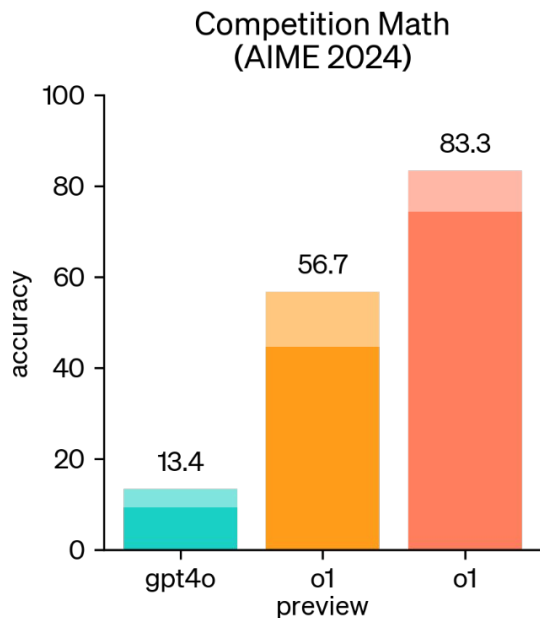
---

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.

Example reasoning trace from DeepSeek-R1 paper


<https://arxiv.org/abs/2501.12948>

# OpenAI o1-preview (Cont.) – results



<https://openai.com/index/learning-to-reason-with-llms/>

# Timeline

- **Aug. 6, 2024:** “Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters”, “Inference Scaling Laws: ...”
- **Sept. 12, 2024:** OpenAI o1 preview dropped
- **Nov. 28 2024:** Alibaba QwQ-32B-Preview (open weights, no disclosed methodology)
- **Dec. 5, 2024:** OpenAI o1 fully released
- **Jan. 8, 2025:** Sky-T1 (17k data distilled from QwQ-32B-Preview)
- **Jan. 22, 2025: DeepSeek R1 —> on par with o1** 
- **Jan. 22, 2025:** Bespoke-32B (using Sky-T1 data pipeline but distilled from DeepSeek-R1)
- **Jan. 31, 2025:** S1 (1k data, distilled from Gemini 2.0 Flash Thinking Experimental)
- **Feb. 5, 2025:** LIMO (paper #2)

# Jan 22, 2025: DeepSeek-R1

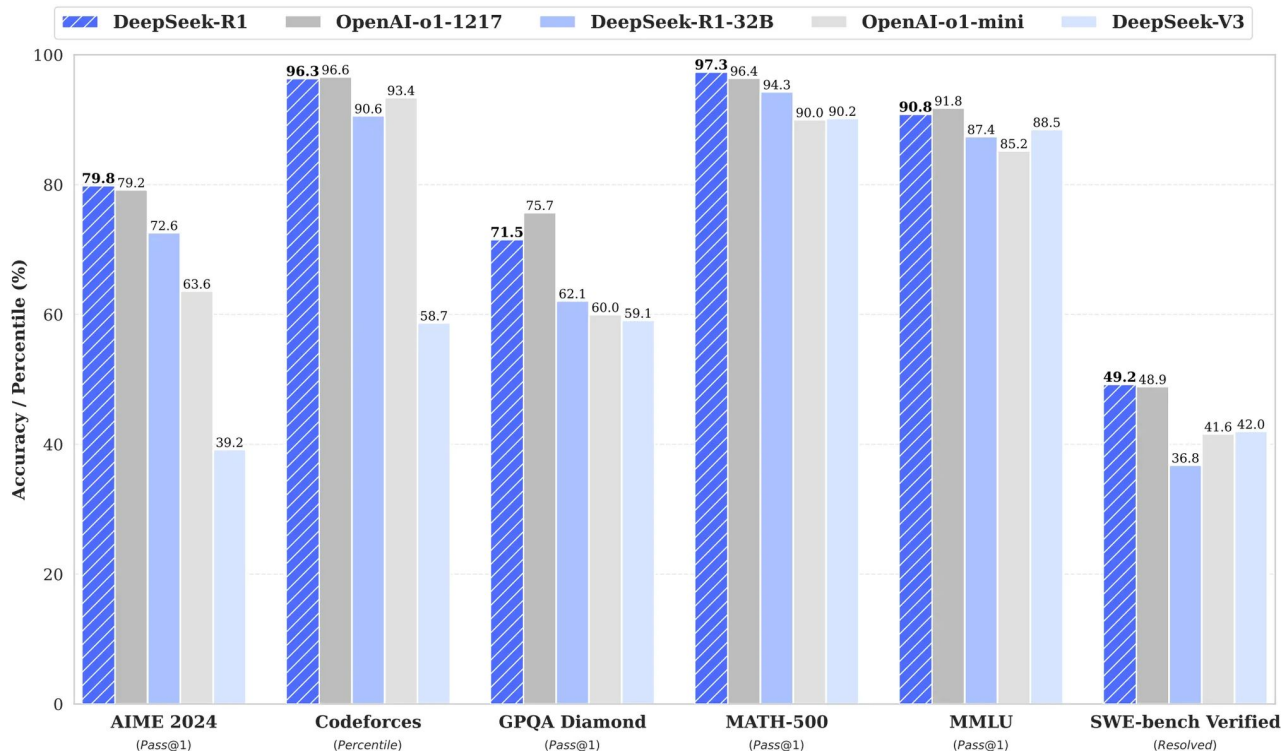
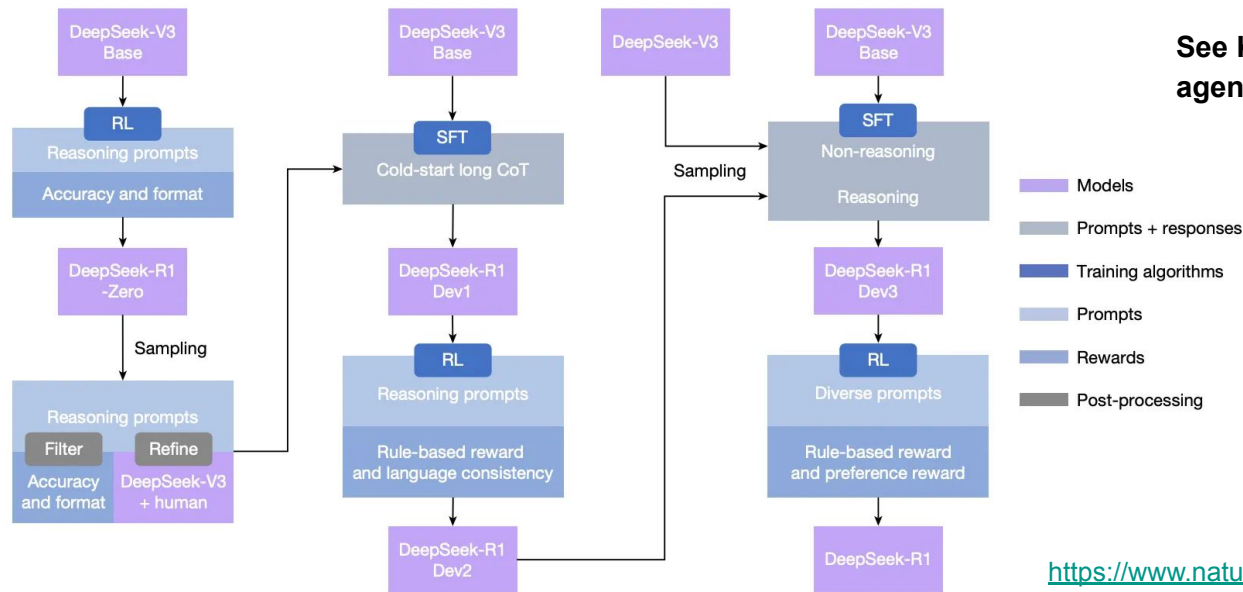


Figure 1 | Benchmark performance of DeepSeek-R1.

# DeepSeek-R1 (Cont.) – an example post-training pipeline



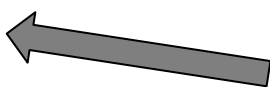
**Fig. 2 | The multistage pipeline of DeepSeek-R1.** A detailed background on DeepSeek-V3 Base and DeepSeek-V3 is provided in Supplementary Information, section 1.1. The models DeepSeek-R1 Dev1, Dev2 and Dev3 represent intermediate checkpoints in this pipeline.

1. **Cold-start with SFT:** readability, format into `|special_token| <reasoning_process> |special_token|`
2. **Reasoning-oriented RL:** easily verifiable tasks coding, maths, science; reward for language consistency
3. **Rejection sampling and SFT:** reasoning: not just easily verifiable tasks, and use DeepSeek-V3 for judge; SFT: writing, factual QA
4. **RL for all scenarios:** human preference alignment, helpfulness and harmlessness

# DeepSeek-R1 (Cont.)

- Then, we start seeing the trend of papers targeting Maths, coding, and other reasoning tasks
- Some are SFT-only methods (paper #1 s1, paper #2 LIMO)
- Some on (SFT +) RL methods
  - Paper #3 compares SFT vs. RL

# Timeline

- **Aug. 6, 2024:** “Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters”, “Inference Scaling Laws: ...”
  - **Sept. 12, 2024:** OpenAI o1 preview dropped
  - **Nov. 28 2024:** Alibaba QwQ-32B-Preview (open weights, no disclosed methodology)
  - **Dec. 5, 2024:** OpenAI o1 fully released
  - **Jan. 8, 2025:** Sky-T1 (17k data distilled from QwQ-32B-Preview)
  - **Jan. 22, 2025:** DeepSeek R1 —> on par with o1
  - **Jan. 22, 2025:** Bespoke-32B (using Sky-T1 data pipeline but distilled from DeepSeek-R1)
  - **Jan. 31, 2025: S1 (1k data, distilled from Gemini 2.0 Flash Thinking Experimental)**
  - **Feb. 5, 2025: LIMO (paper #2)**
- 

# Reasoning Models

- Background on reasoning models
- **Paper #1: S1: Simple test-time scaling**
- Paper #2: LIMO: Less is More for Reasoning
- Paper #3: Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning



# Overview

- Both S1 and LIMO appeared in the context of a prevailing belief that a huge amount of data is needed for SFT
- Both papers emphasize on **the quality of data** and how to construct a small-scale high-quality dataset with surprising results
- S1: more technical details on filtering and **test-time scaling**
- LIMO: **a specific scoring for reasoning** and extensive ablation studies based on the knowledge foundation revolution and inference-time computation scaling revolution

# S1: Simple test-time scaling

- Test-time scaling is a promising approach to improve performance
- OpenAI's o1 model does not publicly show this capability
- S1 provides open-source data (s1K) and model (s1-32B) for public use
  - Mainly highlighting the simplicity, and sample efficiency

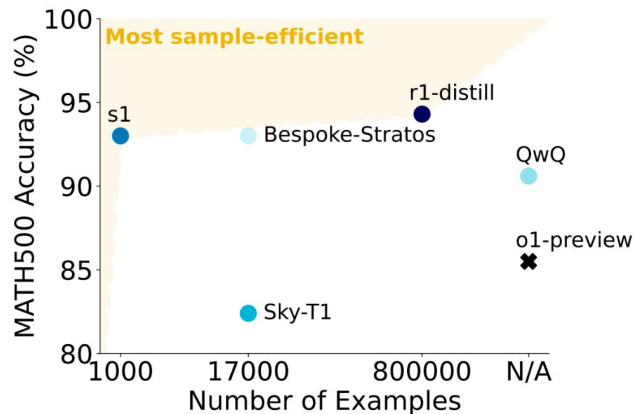
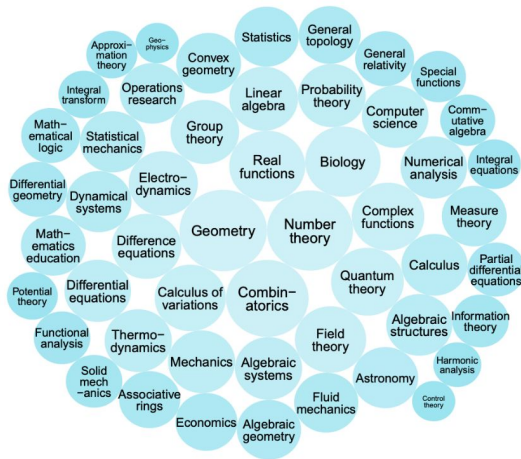


Figure 2. **s1K and s1-32B.** (left) **s1K** is a dataset of 1,000 high-quality, diverse, and difficult questions with reasoning traces. (right) **s1-32B**, a 32B parameter model finetuned on **s1K** is on the sample-efficiency frontier. See Table 1 for details on other models.

# S1K (1K) vs Initial Pool (59K)

Domain	#questions	Total token count	Keywords
Geometry	109	560.2K	Area, Triangle, Distance
Number theory	98	522.5K	Sequences, Divisibility
Combinatorics	75	384.7K	Permutations, Counting
Real functions	43	234.8K	Trigonometry, Calculus
Biology	41	120.9K	Organic reactions
Complex functions	32	170.2K	Complex roots
Quantum theory	32	127.9K	Particles, Wave functions
Field theory	28	150.1K	Polynomials, Roots
Calculus of variations	28	155.5K	Optimization, Control
Difference equations	24	132.5K	Recurrence, Recursion
Electromagnetic theory	23	95.8K	Optics, Waves, Diffraction
Group theory	22	100.0K	Groups, Automorphisms
Linear algebra	22	128.3K	Matrices, Determinants
Probability theory	20	114.6K	Random walk, Expectation
Algebraic systems	19	109.9K	Functional equations
Mechanics	19	103.6K	Forces, Motion, Energy
Thermodynamics	19	74.2K	Heat engines, Entropy
Differential equations	18	89.6K	Substitution, Existence
Computer science	18	34.2K	Complexity theory, Algorithms
Numerical analysis	18	76.5K	Error analysis, Stability
Calculus	17	96.3K	Convergence, Summation
Algebraic structures	17	90.4K	Inequalities, Sets
Astronomy	16	37.7K	Stellar populations, Orbits
Remaining 27 domains	242	982.2K	Domains with $\leq 16$ questions
All domains (51)	1000	4.7M	s1K

Source	Description	#Samples	Avg. thinking length
NuminaMATH (LI et al., 2024)	Math problems from online websites	30660	4.1K
MATH (Hendrycks et al., 2021)	Math problems from competitions	11999	2.9K
OlympicArena (Huang et al., 2024a)	Astronomy, Biology, Chemistry, Computer Science, Geography, Math, and Physics olympiad questions	4250	3.2K
OmniMath (Gao et al., 2024a)	Math problems from competitions	4238	4.4K
AGIEval (Zhong et al., 2023; Ling et al., 2017; Hendrycks et al., 2021; Liu et al., 2020; Zhong et al., 2019; Wang et al., 2021)	English, Law, Logic and Math problems from the SAT, LSAT and other exams	2385	1.2K
xword	Crossword puzzles	999	0.7K
OlympiadBench (He et al., 2024)	Math and Physics olympiad questions	896	3.9K
AIME (1983-2021)	American Invitational Mathematics Examination	890	4.7K
TheoremQA (Chen et al., 2023)	Computer Science, Finance, Math, and Physics university-level questions relating to theorems	747	2.1K
USACO (Shi et al., 2024)	Code problems from the USA Computing Olympiad	519	3.6K
JEEBench (Arora et al., 2023)	Chemistry, Math, and Physics problems used in the university entrance examination of the Indian Institute of Technology	515	2.9K
GPQA (Rein et al., 2023)	PhD-Level Science Questions	348	2.9K
SciEval (Sun et al., 2024)	Biology, Chemistry, and Physics problems from various sources	227	0.7K
s1-prob	Stanford statistics qualifying exams	182	4.0K
LiveCodeBench (Jain et al., 2024)	Code problems from coding websites (LeetCode, AtCoder, and CodeForces)	151	3.5K
s1-teasers	Math brain-teasers crawled from the Internet	23	4.1K
All 59K questions	Composite of the above datasets with reasoning traces and solutions	59029	3.6K

# How do they filter?

1. Quality: API errors, formatting issues (51K)
  2. Difficulty: filter out easily solved problems (24K)
    - Evaluated by Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct
    - Graded by Claude 3.5
  3. Diversity: classify the problems, sample one problem each time (1K)
    - Classified by Claude 3.5
    - Favor long reasoning traces
- 
- Reasoning traces and solutions are generated by Google Gemini Flash Thinking API

# Test-time scaling

## Budget Forcing

- Adding “Wait” when the reasoning trace is too short
- Appending “Final Answer:” when the reasoning trace reaches the token limit

How many r in raspberry?

**Question**

Let's break down the process of counting the letter 'r' in the word "raspberry" ...

\* First letter: 'r' - This is an 'r', count = 1.

\* Second letter: 'a' - Not an 'r', count remains 1 ...

\* Sixth letter: 'e' - Not an 'r', count remains 1.

\* Seventh letter: 'r' - This is an 'r', count = 2.

\* Eighth letter: 'y' - Not an 'r', count remains 2 ...

The number of 'r's in "raspberry" is 2.

**Wait**, let's re-read the question carefully. It asks "How many r in raspberry?" ... \* r - a - s - p - b - e - r - r - y ... \* First 'r' ... \*

Second 'r' ... \* Third 'r' ... Count = 3 ...

**Reasoning trace**

My initial answer of 2 was incorrect due to a quick reading of the word. **Final Answer:** The final answer is **3**

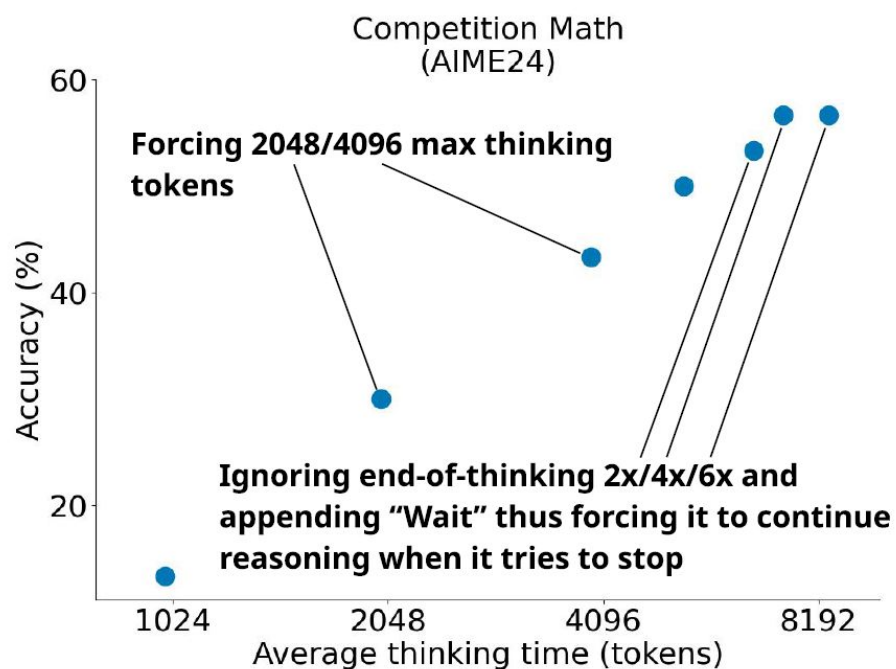
**Response**

# S1-32B and its performance

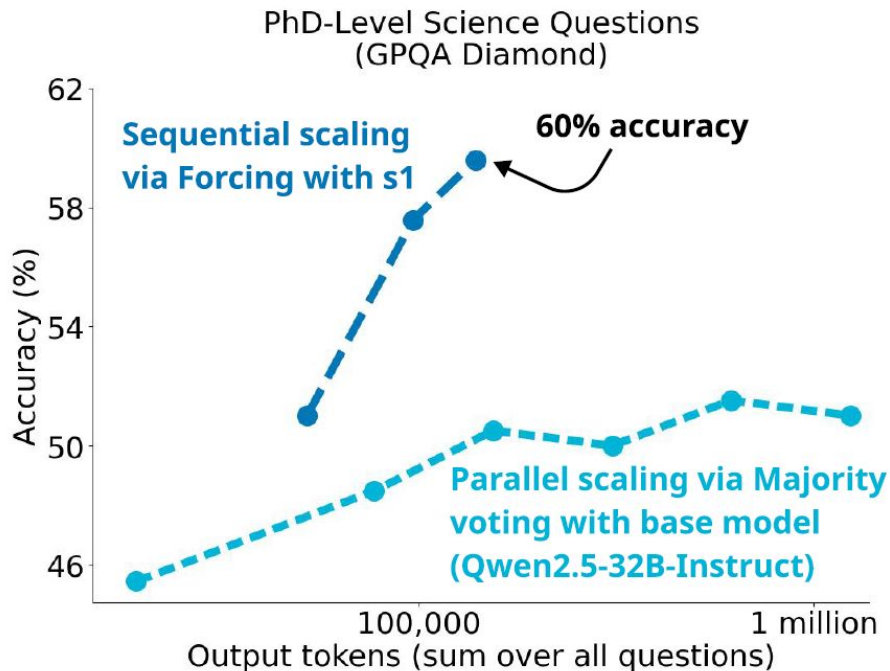
- SFT on Qwen2.5-32B-Instruct using s1K
- Benchmark: AIME24, MATH500, GPQA Diamond
- Best sample efficiency
- Budget forcing helps a lot

Model	# ex.	AIME 2024	MATH 500	GPQA Diamond
API only				
o1-preview	N.A.	44.6	85.5	73.3
o1-mini	N.A.	70.0	90.0	60.0
o1	N.A.	<b>74.4</b>	<b>94.8</b>	<b>77.3</b>
Gemini 2.0 Flash Think.	N.A.	60.0	N.A.	N.A.
Open Weights				
Qwen2.5- 32B-Instruct	N.A.	26.7	84.0	49.0
QwQ-32B	N.A.	50.0	90.6	54.5
r1	≥800K	<b>79.8</b>	<b>97.3</b>	<b>71.5</b>
r1-distill	800K	72.6	94.3	62.1
Open Weights and Open Data				
Sky-T1	17K	43.3	82.4	56.8
Bespoke-32B	17K	<b>63.3</b>	93.0	58.1
s1 w/o BF	<b>1K</b>	50.0	92.6	56.6
<b>s1-32B</b>	<b>1K</b>	56.7	<b>93.0</b>	<b>59.6</b>

# Sequential vs Parallel



(a) Sequential scaling via budget forcing



(b) Parallel scaling via majority voting



# Ablation study on the three-step filtering

Model	AIME 2024	MATH 500	GPQA Diamond
1K-random	36.7 [-26.7%, -3.3%]	90.6 [-4.8%, 0.0%]	52.0 [-12.6%, 2.5%]
1K-diverse	26.7 [-40.0%, -10.0%]	91.2 [-4.0%, 0.2%]	54.6 [-10.1%, 5.1%]
1K-longest	33.3 [-36.7%, 0.0%]	90.4 [-5.0%, -0.2%]	59.6 [-5.1%, 10.1%]
59K-full	53.3 [-13.3%, 20.0%]	92.8 [-2.6%, 2.2%]	58.1 [-6.6%, 8.6%]
<b>s1K</b>	<b>50.0</b>	<b>93.0</b>	<b>57.6</b>



# Reasoning Models

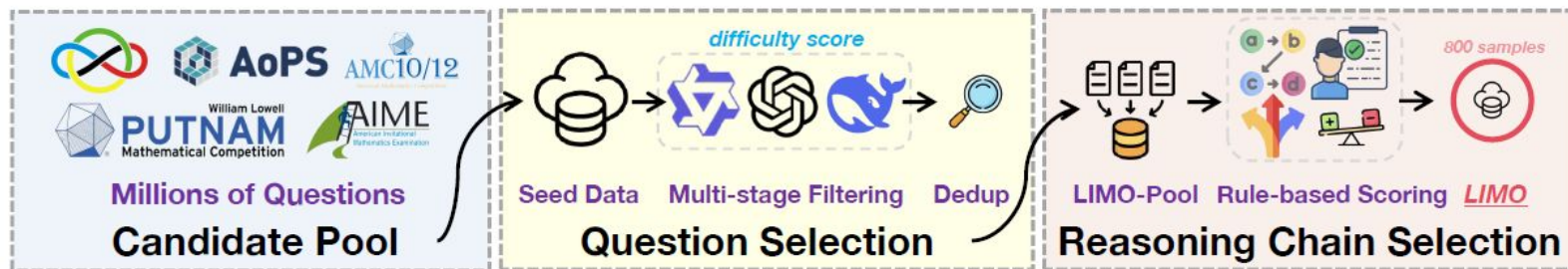
- Background on reasoning models
- Paper #1: S1: Simple test-time scaling
- **Paper #2: LIMO: Less is More for Reasoning**
- Paper #3: Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning

# LIMO: Less is more for reasoning

- Challenge the prevailing assumption that massive training data is necessary to achieve excellent performance
- Less-Is-More reasoning hypothesis: knowledge-rich foundation models and sufficient inference-time computation are crucial
- Construct LIMO dataset with 800 samples

# Data Filtering

- Apply Qwen2.5-Math-7B-Instruct and DeepSeek-R1-Distill-Qwen-32B to ensure difficulty of problems (LIMO pool with size 2125)
- Select the problems with the highest-scoring solutions provided by DeepSeek R1, DeepSeek-R1-Distill-Qwen-32B, QwQ-32B (LIMO dataset with size 800)
- Score measures elaborated reasoning, self-verification, exploratory approach, and adaptive granularity



# SFT Qwen2.5-32B-Instruct with LIMO dataset

Datasets	OpenAI-o1 -preview	Qwen2.5-32B -Instruct	QwQ-32B- preview	OpenThoughts (114k)	NuminaMath (100k)	LIMO ours(800)
In Domain						
AIME24	44.6	16.5	50.0	50.2	6.5	<b>63.3</b>
MATH500	85.5	79.4	89.8	80.6	59.2	<b>95.6</b>
AMC23	81.8	64.0	83.6	80.5	40.6	<b>96.3</b>
Out of Domain						
OlympiadBench	52.1	45.3	58.5	56.3	36.7	<b>67.6</b>
CHMath	50.0	27.3	68.5	74.1	11.2	<b>84.2</b>
Gaokao	62.1	72.1	80.1	63.2	49.4	<b>91.1</b>
Kaoyan	51.5	48.2	70.3	54.7	32.7	<b>83.9</b>
GradeSchool	62.8	56.7	63.8	39.0	36.2	<b>76.2</b>
Minerva	47.1	41.2	39.0	41.1	24.6	<b>52.2</b>
GPQA	<b>73.3</b>	48.0	65.1	42.9	25.8	70.7
AVG.	61.1	49.9	66.9	58.3	32.3	<b>78.1</b>

# Impact of reasoning chain quality

- Quality level is categorized based on the score
- Why is the score reasonable?

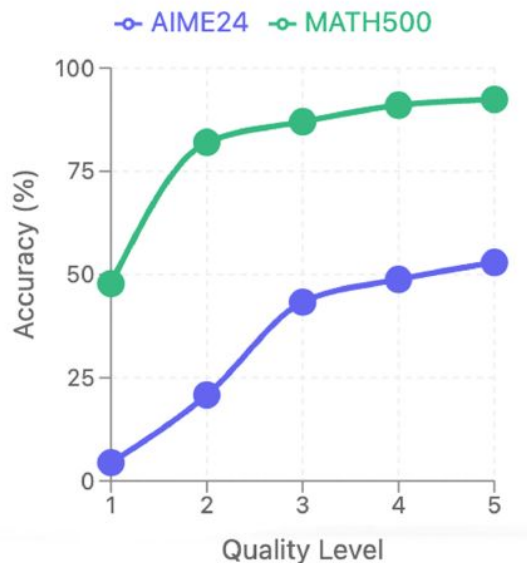
**Elaborated Reasoning:** Comprehensive exploration of logical steps without premature conclusions

**Self-Verification:** Regular validation of intermediate results and logical consistency

**Exploratory Approach:** Consideration of multiple possibilities before reaching conclusions

**Adaptive Granularity:** Appropriate detail level across simple and complex deductions

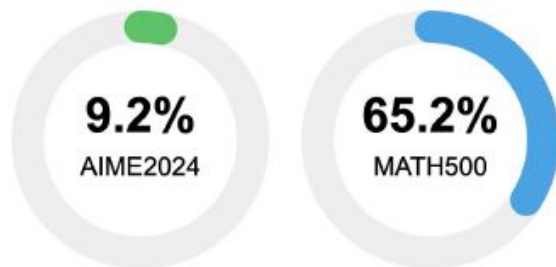
To quantify these qualities, we implemented a rule-based scoring system that calculated weighted metrics for each dimension. Elaborated Reasoning was measured by solution length (30% weight); Self-Verification through frequency of validation-related words like "check" and "verify" (20% weight); Exploratory Approach by counting tentative expressions such as "perhaps" and "might" (25% weight); and Adaptive Granularity via connective phrases like "therefore" and "since" (25% weight). All keyword frequencies were normalized by text length to ensure fair comparison across solutions of different sizes.



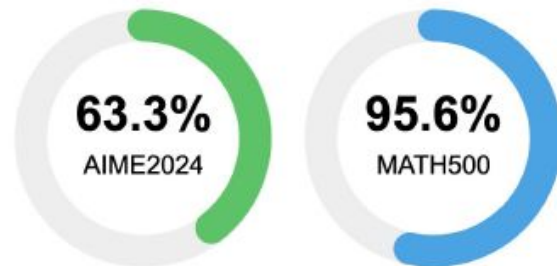
# Impact of the backbone quality

- Identical architecture
- Qwen2.5's pretraining data has higher quality, particularly in mathematical and code-related contents

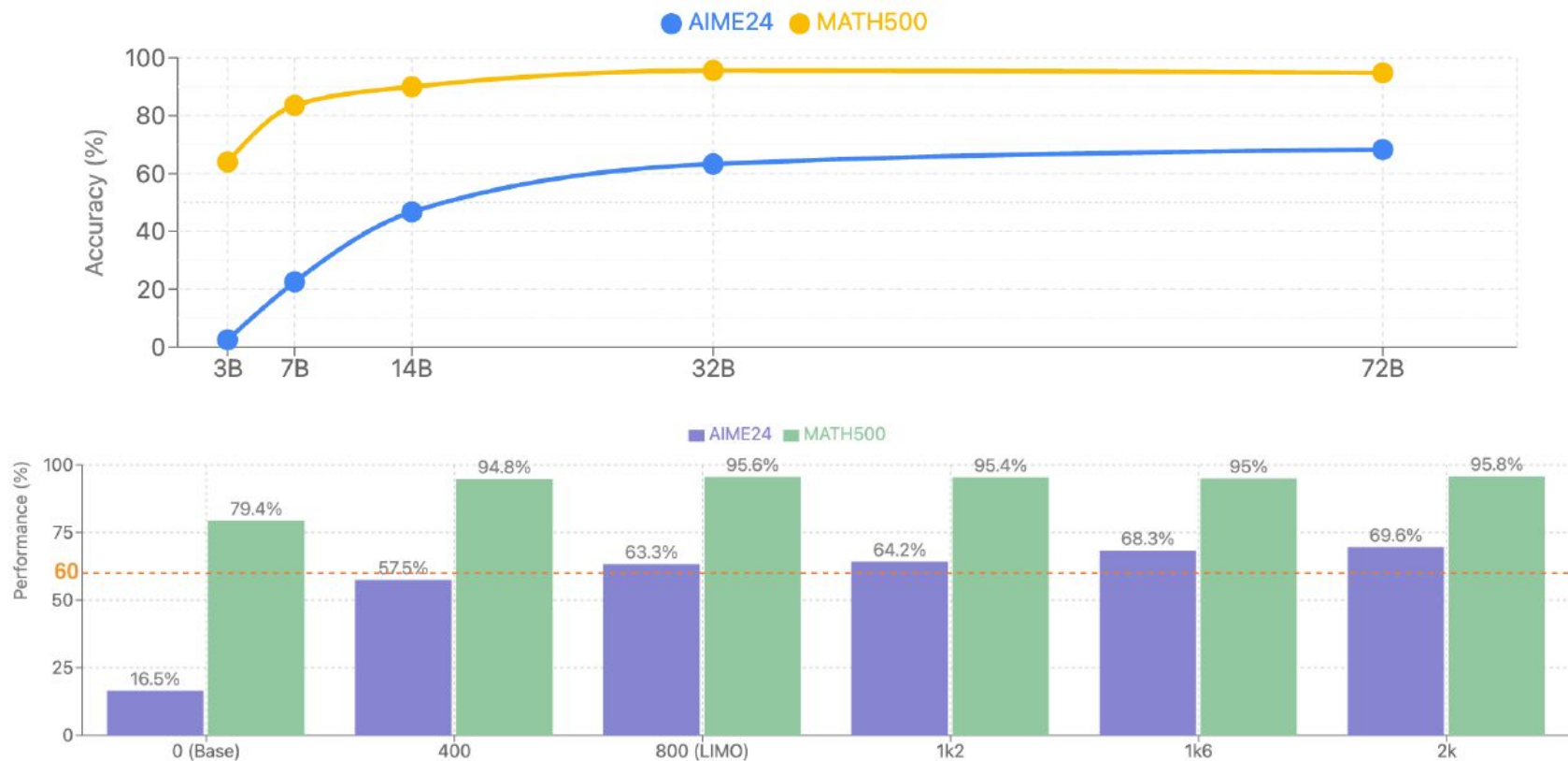
**Qwen1.5-32B-Chat**



**Qwen2.5-32B-Instruct**



# Scaling model size and dataset size



# Takeaway & Questions

- Quality matters more than quantity
- SFT is an approach to awaken the inherent reasoning ability of a strong pre-trained model
- How to define data quality and reasoning quality?
- Is it reasonable to assume longer CoT translates to better reasoning or higher difficulty of the problem?
- Is SFT truly able to generalize to out-of-distribution tasks?



# Reasoning Models

- Background on reasoning models
- Paper #1: S1: Simple test-time scaling
- Paper #2: LIMO: Less is More for Reasoning
- **Paper #3: Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning**

# TLDR of the paper

- **Background:** models' performance on MATH and AIME improve week by week (avalanche of papers on improving Maths benchmarks)
- **Research question:** “do these gains reflect broader problem-solving ability or just narrow overfitting?”
- **Approach:** evaluate >20 open-weight reasoning-tuned models across a broad suite of tasks (maths itself, non-math reasoning, non-reasoning)
- **Observation:** RL'd models generalize better!
- **Further investigation (focus of this presentation):** controlled study of RL vs. SFT; looked at metrics on hidden state shift, and token shift

# I. Evaluation on Transferability – Benchmarks Used

- **Math reasoning:** MATH500, AIME24, AIME25, OlympiadBench
- **Other reasoning:** Scientific QA, coding, agent planning, logical deduction
  - LiveCodeBench, GPQA-Diamond, ACPBench, HeadQA
- **Non-reasoning:** instruction following, conversational QA
  - CoQA, IFEval, HaluEval, MC-TACO

# I. Evaluation on Transferability (Cont.) – Results

## Transferability Index

How reasoning models can transfer their capabilities  
from one domain to another

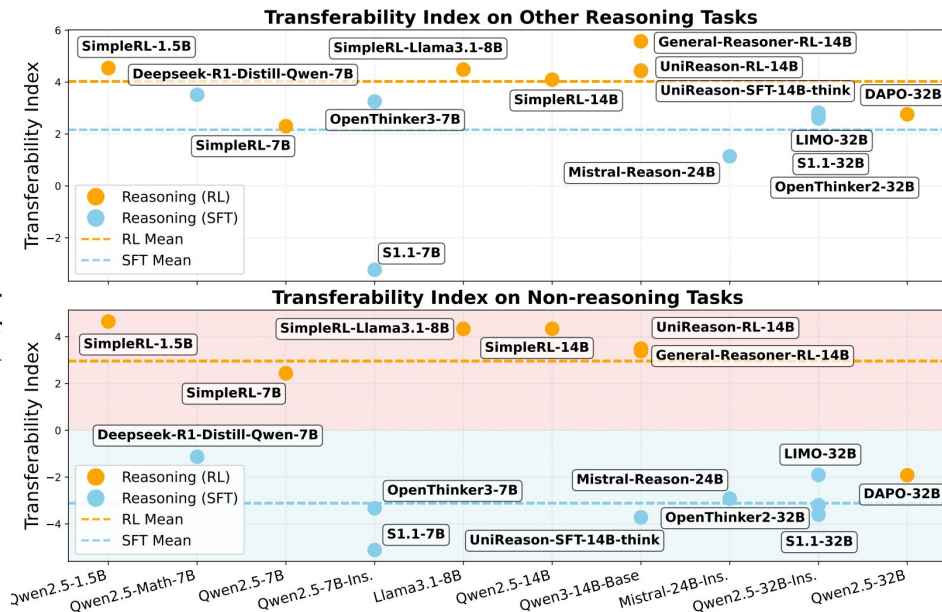
Let  $\mathcal{B}_g$  be the set of benchmarks in group  $g \in \{\text{math}, \text{other}, \text{non}\}$ , corresponding to each of our task groups: math reasoning, other reasoning and non-reasoning. Let  $|\mathcal{B}_g|$  be its cardinality. For each benchmark  $b \in \mathcal{B}_g$  we have scores  $R_b^{\text{model}}$  and  $R_b^{\text{base}}$ . We define the group-level relative gain as the average of per-benchmark gains:

$$\Delta R_g = \frac{1}{|\mathcal{B}_g|} \sum_{b \in \mathcal{B}_g} \frac{R_b^{\text{model}} - R_b^{\text{base}}}{R_b^{\text{base}}}, \quad g \in \{\text{math}, \text{other}, \text{non}\}.$$

Next, the two Transferability Indices are

$$\text{TI}_{\text{other}}(\%) = \frac{\Delta R_{\text{other}}}{\Delta R_{\text{math}}} \times 100, \quad \text{TI}_{\text{non}}(\%) = \frac{\Delta R_{\text{non}}}{\Delta R_{\text{math}}} \times 100.$$

The TI value is compared against 0, any TI above zero indicates positive transfer observed.



**Figure 2** | Transferability of mathematical reasoning to other reasoning and non-reasoning tasks. The Transferability Index measures a model's ability to transfer performance from mathematics to other domains, with positive values indicating successful transfer and negative values indicating performance degradation. Details of this metric can be found in Section 2.1. RL models consistently outperform SFT models, regardless of model size, architecture, or training data, demonstrating superior transferability. Values are displayed using a signed logarithmic transformation for better visualization.

## II. A controlled study on RL vs. SFT for transferability

### Controlled

- Architecture and size: Qwen3-14B-Base
- Data: MATH and DeepScaler

### Comparing

- RL: use GRPO (a cheaper PPO, first proposed in DeepSeek-Math)
- SFT: use data generated by teacher model Qwen3-32B, with rejection sampling (i.e. only use correct data)

# Quick note on the data – DeepScaler

- **Problem:** If  $991+993+995+997+999=5000-N$ , then  $N=$
- **Answer:** 25
- **Problem:** Positive integers  $a$  and  $b$  are each less than 6. What is the smallest possible value for  $2 \cdot a - a \cdot b$ ?
- **Answer:** -15
- How is it curated?
  - “Compiled AIME problems from 1984-2023 and AMC problems prior to 2023, along with questions from the Omni-MATH and Still datasets”
  - Then post-processed

## DeepScaleR: Surpassing O1-Preview with a 1.5B Model by Scaling RL

Michael Luo\*, Sijun Tan\*, Justin Wong\*, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai, Jeffrey Luo

Advisors: Li Erran Li, Raluca Ada Popa, Ion Stoica

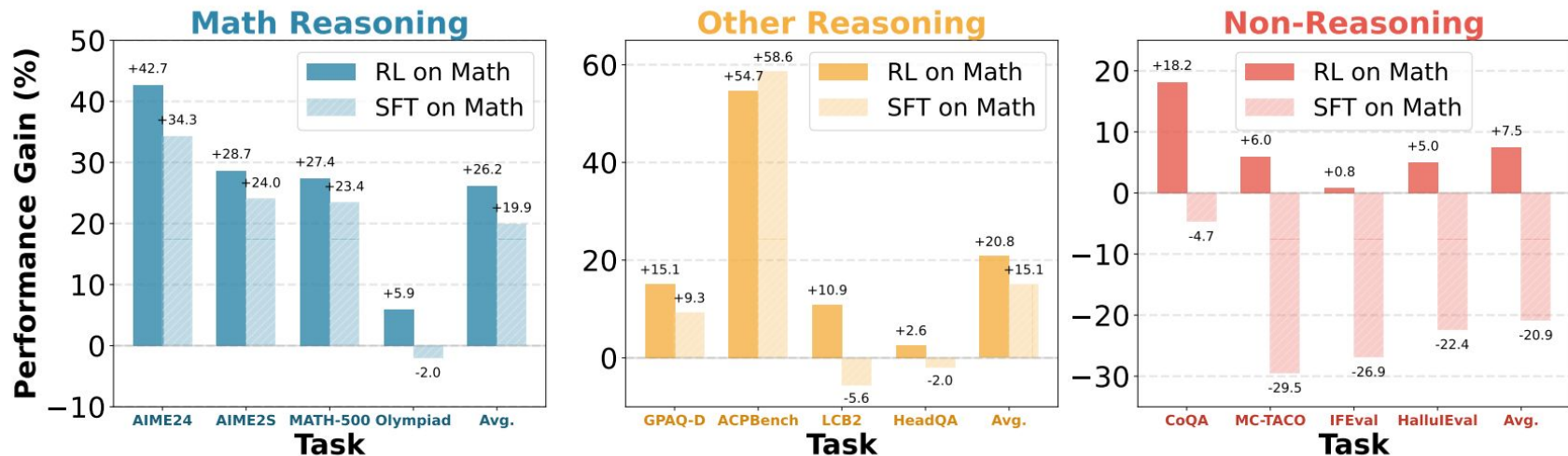
\*: Project Leads; \*: Significant Contributor

### TL;DR

RL magic is in the air! We introduce **DeepScaleR-1.5B-Preview**, a language model finetuned from **Deepseek-R1-Distilled-Qwen-1.5B** using simple reinforcement learning (RL). It achieves an impressive **43.1% Pass@1** accuracy on AIME2024 (+14.3% improvement over the base model), surpassing the performance of OpenAI's **o1-preview** with just **1.5B** parameters. We **open sourced** our dataset, code and training logs for everyone to progress on scaling intelligence with RL.

[Website](#), [Github](#), [HF Model](#), [HF Dataset](#), [Wandb Logs](#), [Eval Logs](#)

## II. A controlled study on RL vs. SFT – Results



**Figure 1** | Impact of SFT and RL using math-only training queries on the same base model, Qwen3-14B-Base. Performance improvements are measured relative to the base model. While SFT-trained models partially generalize to other reasoning tasks, they show limited transfer to non-reasoning tasks. In contrast, RL-trained models exhibit broader generalization across both reasoning and non-reasoning scenarios.

## III.a. RL vs. SFT – Latent Space Shift

- **Methodology**

- Measure hidden representations, rather than weights!
- Why?
  - Weights: “changes in model parameters do not always correspond to functional differences”
  - Hidden states: “subtle parameter modifications can lead to significant shifts in the activation distribution.”
- “Since transferability fundamentally relies on the alignment and stability of learned representations across tasks or domains, PCA shift is particularly effective for diagnosing changes that may impact cross-domain generalization.”



## III.a. RL vs. SFT – Latent Space Shift (Cont.) – results

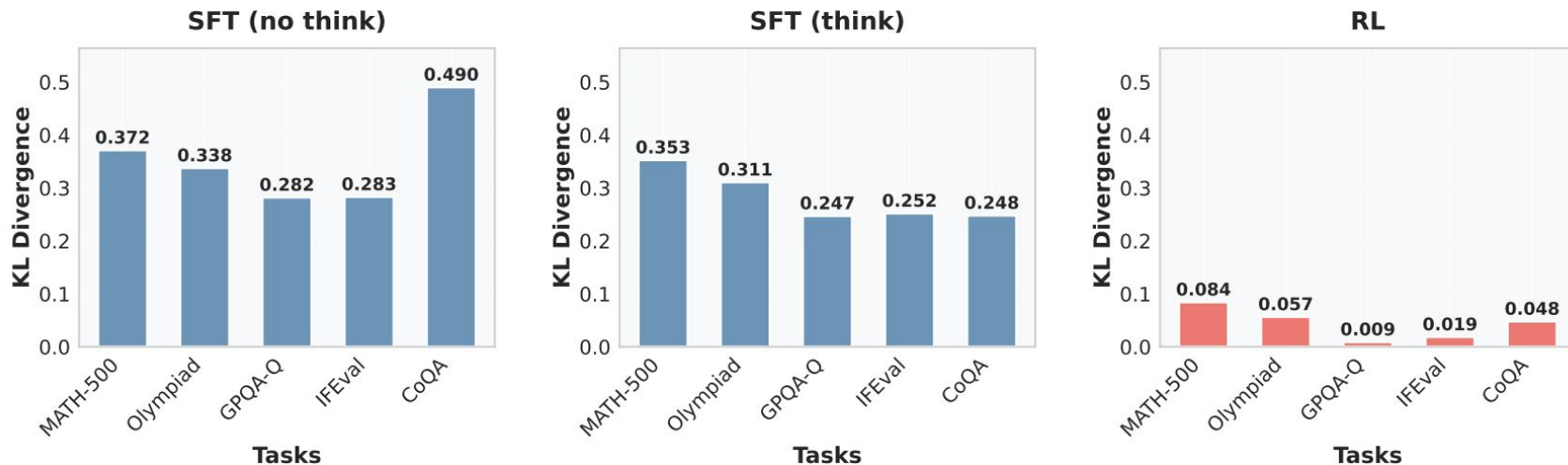
**Table 2** | Comparison of mean PCA shift magnitudes on math, other reasoning and non-reasoning tasks for Qwen3-14B and our finetuned models. We observed that RL model has lower mean PCA shifts than SFT models.

Model	Math	Other-Reasoning	Non-Reasoning
Qwen3-14B (no-think)	40.4	13.9	129.8
UniReason-Qwen3-14B- <i>no-think</i> (SFT)	21.4	10.9	113.7
UniReason-Qwen3-14B- <i>think</i> (SFT)	19.2	6.7	38.2
UniReason-Qwen3-14B (RL)	8.5	3.5	36.9

## III.b. RL vs. SFT – Tokens' Shift – Two Metrics

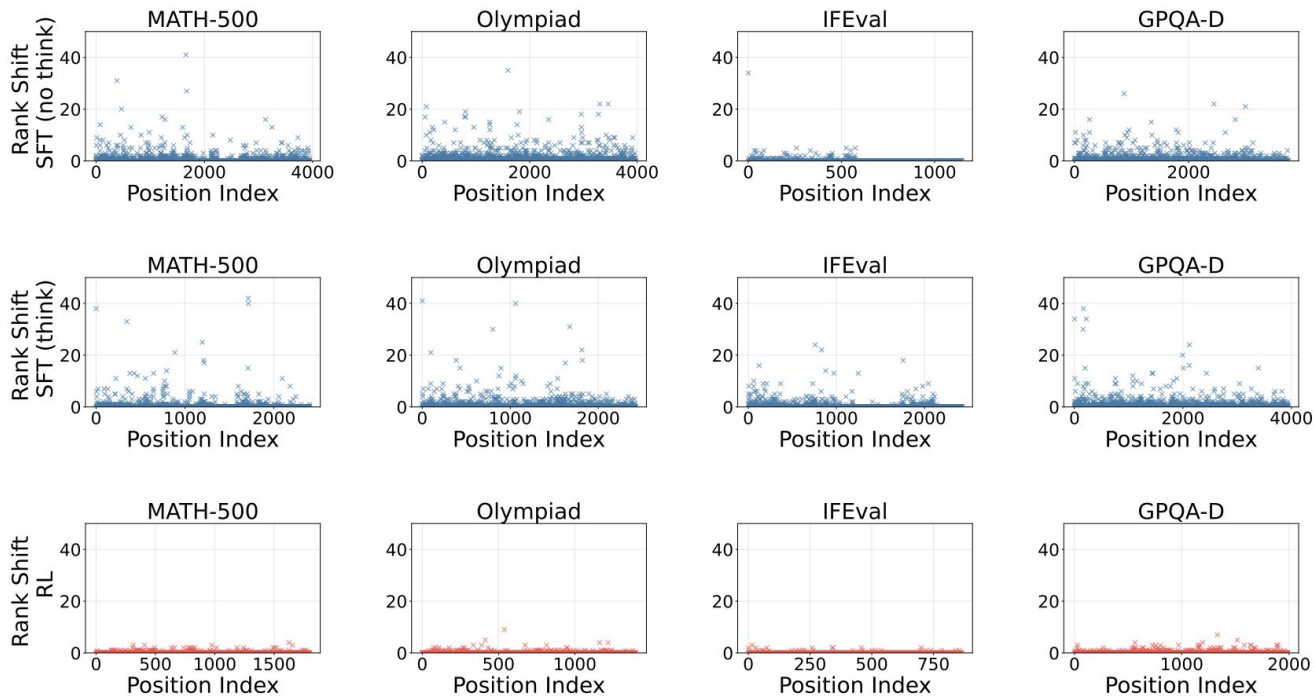
1. Take the decoded output of the fine-tuned model (RL, or SFT) – call it sequence **A**
  - a. e.g.: “**Let ’ s**”, “**look**”, “**at**”, “**the**”, “**problem**”
2. Decode the exact same tokens on the base model
  - a. For each token **A<sub>i</sub>**, we look at whether the base model would pick the same token **A<sub>i</sub>**
    - i. If it is the same – no rank shift
    - ii. If not – there is a shift
  - b. e.g. Look at **A<sub>1</sub>** “look” – base model might have logits ranked as [ “**take**”, “**look**”, ... ]
  - c. → **token rank shift!**
3. In addition, use KL-divergence to measure the token distribution shift (i.e. the difference between the two **logprobs**) → **token distribution shift!**

## III.b. RL vs. SFT – Token Distribution Shift – Results



**Figure 5** | KL divergence analysis of RL and SFT models. Higher KL divergence indicates greater distribution shifts from the original backbone model. We observe that RL models consistently exhibit significantly lower KL divergence compared to SFT models across different tasks, suggesting less distribution shift during training.

## III.b. RL vs. SFT – Token Rank Shift



**Figure 6** | Visualization of token rank shifts across different position indices for both reasoning and non-reasoning tasks. We observe that RL models exhibit less token rank shifts while SFT models demonstrate substantial rank shifts across numerous positions throughout the sequence.

## III.b. RL vs. SFT – Example rank shift

Domain	Query	Model	Shifted Tokens
Reasoning Task	Ten trees weigh as much as three squigs and one goolee. Two trees and one goolee are equal in weight to one squig. How many trees' weight equals one squig?	RL Model	<b>(Only 15 tokens experienced rank shift when decoded in the base model)</b> In a Now Now define for number second 2 Now , add This
		SFT Model	<b>(390 tokens experienced rank shift)</b> they The again conflicting but m Alternatively make have Hmm hold equations Wait For find check Let maybe using written Original pl contrad So There Wait solve I 's Alternatively Alright so First solving a either check conflicting write Correct here another Like where ? Still From where The question / . The where here where equations Therefore problem check if was the ? equations together . answer I For or For Wait matrices this about m either and solve combined 1 problem ten Let . equation That If...

*“RL models selectively shift task-relevant tokens, while SFT models shift numerous irrelevant tokens.”*

## IV.a. Conclusion and Thoughts – SFT learns structure and style!

- On reasoning specifically

*[Submitted on 11 Feb 2025 (v1), last revised 18 Feb 2025 (this version, v2)]*

### **LLMs Can Easily Learn to Reason from Demonstrations Structure, not content, is what matters!**

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhmaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, Ion Stoica

- *“a model trained on Long CoT samples with incorrect answers still achieves only 3.2% lower accuracy compared to training with fully correct samples.”*

- Reminds me of the “False Promise” paper from earlier lecture

*[Submitted on 25 May 2023]*

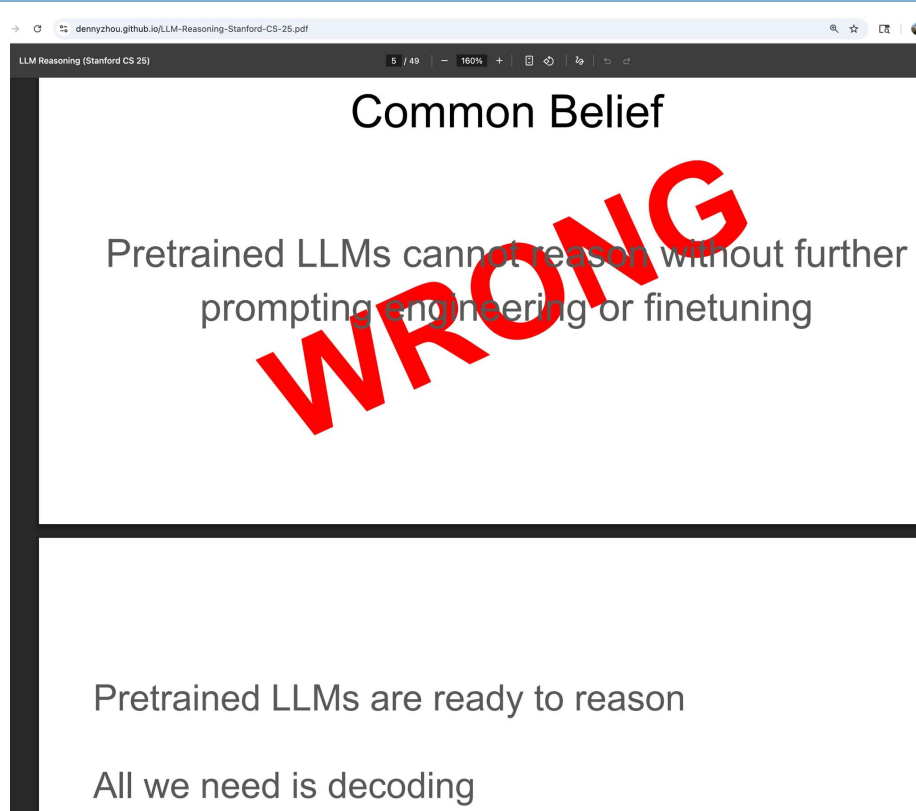
### **The False Promise of Imitating Proprietary LLMs**

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, Dawn Song

An emerging method to cheaply improve a weaker language model is to finetune it on outputs from a stronger model, such as a proprietary system like ChatGPT (e.g., Alpaca, Self-Instruct, and others). This approach looks to cheaply imitate the proprietary model's capabilities using a weaker open-source model. In this work, we critically analyze this approach. We first finetune a series of LMs that imitate ChatGPT using varying base model sizes (1.5B--13B), data sources, and imitation data amounts (0.3M--150M tokens). We then evaluate the models using crowd raters and canonical NLP benchmarks. Initially, we were surprised by the output quality of our imitation models -- they appear far better at following instructions, and crowd workers rate their outputs as competitive with ChatGPT. However, when conducting more targeted automatic evaluations, we find that imitation models close little to none of the gap from the base LM to ChatGPT on tasks that are not heavily supported in the imitation data. We show that these performance discrepancies may slip past human raters because imitation models are adept at mimicking ChatGPT's style but not its factuality. Overall, we conclude that model imitation is a false promise: there

## IV.a. Conclusion and Thoughts – SFT learns structure and style!

- Related slide from Denny Zhou (Deepmind)
- That is, all the prior is there – just teach the model the structure of reasoning!



## IV.b. Conclusion and Thoughts – More On RL vs. SFT

- The 3rd paper uses RFT from a teacher model
- What if we self-sample from the base model?
- Related paper: ***Bridging Supervised Learning and Reinforcement Learning in Math Reasoning*** (<https://arxiv.org/abs/2505.18116>)
  - GRPO vs. NFT (RFT, but use both correct and incorrect answers)
  - ***“we demonstrate that NFT and GRPO are actually equivalent in strict-on-policy training, even though they originate from entirely different theoretical foundations”***



# Critic

Reasoning Models:

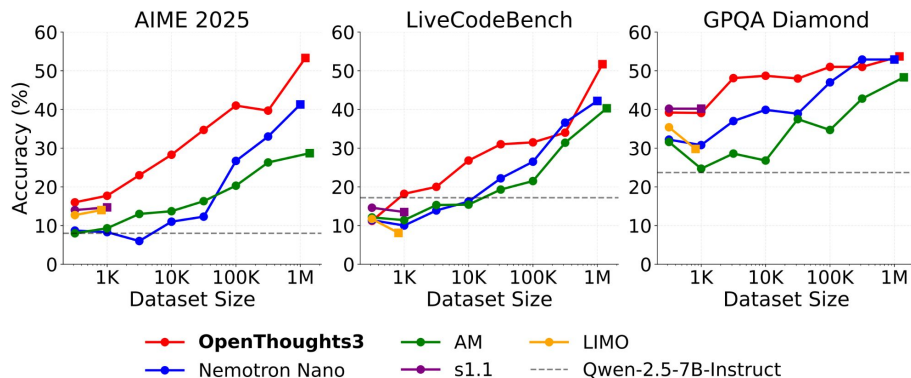
- S1: Simple test-time scaling
- LIMO: Less is More for Reasoning
- Does Math Reasoning Improve General LLM Capabilities?

**Hanchen Li, Harman Singh**

Sep 30th

# #1.1 Less may not be more for reasoning

1. Openthoughts shows that increasing dataset size reliably improves performance
2. Test-time scaling improves with data
3. Reducing from 59k samples to 1k reduces performance.



Model	# ex.	AIME 2024	MATH 500	GPQA Diamond
r1	≥800K	79.8	97.3	71.5
r1-distill	800K	72.6	94.3	62.1
Open Weights and Open Data				
Sky-T1	17K	43.3	82.4	56.8
Bespoke-32B	17K	63.3	93.0	58.1
s1 w/o BF	1K	50.0	92.6	56.6
s1-32B	1K	56.7	93.0	59.6

## #1.2 Base model as the BIG confounding factor

1. Does it work for models except Qwen?

Qwen shows many surprising phenomena including

- [RL with one training example](#)
- [RL with random and incorrect label rewards](#)

Potentially, Qwen's base model is already strong at reasoning and SFT with 1k examples is beneficial “only for Qwen”

2. Dataset selection is biased by the same family models (Qwen)

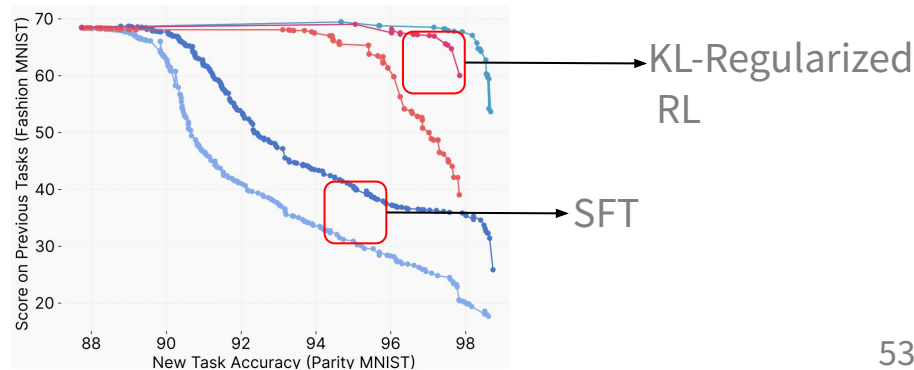
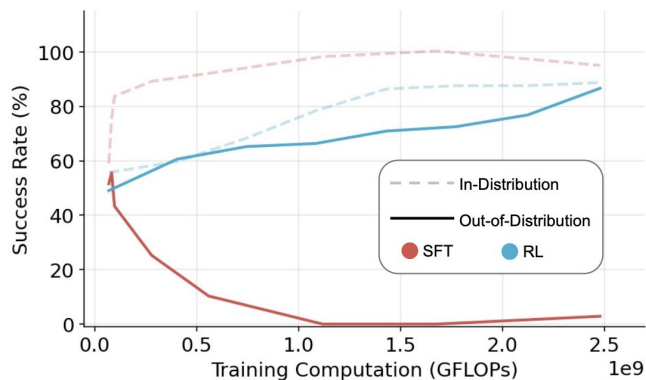
# #1.3 Distillation, Compute and Cost of Data

Usage of large existing thinking models and API's

- a. S1 - recipe to achieve test-time scaling if you have a strong thinking model already (**Gemini 2.0**). LIMO uses **DeepSeek R1 + Humans**.
- b. **Un-accounted Compute and Complexity in “Simple” approaches:**
  - i. Generating Reasoning traces from Gemini
  - ii. Qwen models to check example hardness,
  - iii. Claude for evaluating answers during dataset construction – this cost is non-trivial and makes the approach complex.
- c. Human labor & reproducibility: collaborative review (LIMO) + creating new data (S1) is potentially very costly, unscalable, subjective and not-reproducible (in LIMO's case)

# #1.4 RL is important and SFT isn't enough!

1. SFT tends to memorize traces, not generalize to unseen variants. RL with outcome rewards can generalize better.
2. SFT can cause forgetting of prior capabilities and diverge arbitrarily from the base model's distribution
3. SFT lacks exploration — cannot discover new reasoning paths



# #1.5 Evaluation

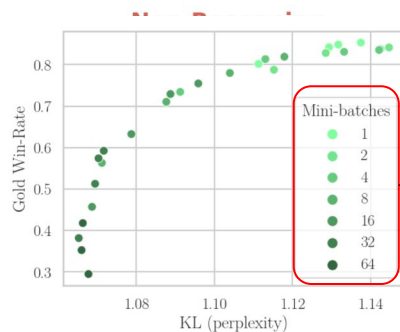
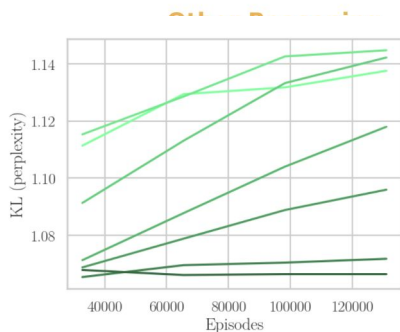
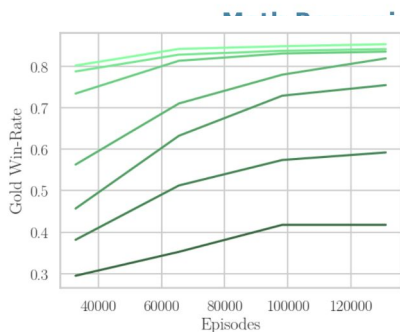
1. Budget forcing may be brittle and no clear signal to stop?-> confidence?
2. Decontamination: Weak n-gram overlap based decontamination in LIMO and S1 may result in exaggerated performance.
3. Human (Author) data curation might be bias towards Evaluations.
4. The benchmarks are largely in Mathematics. Minerva and GPQA (STEM) performance gain is not as significant.

Datasets	OpenAI-o1 -preview	Qwen2.5-32B -Instruct	QwQ-32B- preview	OpenThoughts (114k)	NuminaMath (100k)	LIMO ours(800)
In Domain						
AIME24	44.6	16.5	50.0	50.2	6.5	<b>63.3</b>
MATH500	85.5	79.4	89.8	80.6	59.2	<b>95.6</b>
AMC23	81.8	64.0	83.6	80.5	40.6	<b>96.3</b>
Out of Domain						
OlympiadBench	52.1	45.3	58.5	56.3	36.7	<b>67.6</b>
CHMath	50.0	27.3	68.5	74.1	11.2	<b>84.2</b>
Gaokao	62.1	72.1	80.1	63.2	49.4	<b>91.1</b>
Kaoyan	51.5	48.2	70.3	54.7	32.7	<b>83.9</b>
GradeSchool	62.8	56.7	63.8	39.0	36.2	<b>76.2</b>
Minerva	47.1	41.2	39.0	41.1	24.6	<b>52.2</b>
GPQA	<b>73.3</b>	48.0	65.1	42.9	25.8	70.7
AVG.	61.1	49.9	66.9	58.3	32.3	<b>78.1</b>

# #2 On Math Reasoning Improves General Capabilities

## #2.1 Is being on-policy the reason why RL is better? How can this claim be strengthened?

- A controlled study on gradually increasing on-policyness would prove the point more strongly (As discussed in [Asynchronous RLHF paper](#))



# #2 On Math Reasoning Improves General Capabilities

## #2.2 Are SFT vs RL experiments a fair comparison w.r.t compute and hyperparameters?

1. “We use ... batch size is 512 and we train for **1.5 epoch (??)** to align with our RL settings.”
  2. “We use  $5 \times 10^{-5}$  (??) learning rate for SFT....”
- RL and SFT updates are not directly comparable in compute
    - Rollouts + evals are expensive. Compute (FLOPs) not reported
  - RL → sensitive to hyperparams. Hyperparam decisions not explained well
  - KL penalty is off, but drift from base model is less → counterintuitive
  - It is possible that with compute-matched settings (SFT with multiple epochs and tuning, SFT generalizes better



# #2 On Math Reasoning Improves General Capabilities

## #2.3 SFT data curation design decisions incomprehensiveness

- SFT with **multiple responses per prompt** might be much better for reducing memorization (as explored in [OpenThoughts](#)). This paper doesn't ablate this.
- Current data curation: prompt → Qwen3-32B-Instruct → rejection sampling → SFT data
  - **Why not “match” Qwen3-32B-Instruct distribution** by not doing hard rejection sampling (i.e. keep some incorrect traces to learn about general reasoning skills instead of only knowledge)
- **SFT with curriculum learning:** raining on more on-policy-like/in-distribution data first may reduce catastrophic distortion of representations

# Critic-supplementary slides

S1 paper:

- training on only 1,000 samples with next-token prediction ... a strong reasoning model that **scales in performance with more test-time compute**
- Similar definition can be applied to pre-training? So is pre-training enough
- S1 is a sophisticated data curation technique for reasoning

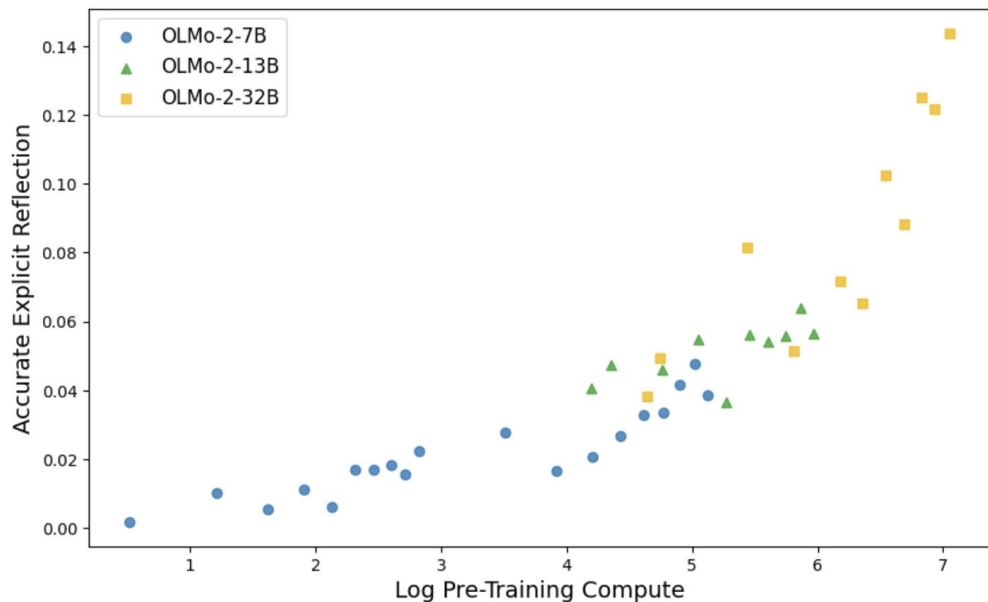


Figure 1: As pre-training compute increases, OLMo-2 checkpoints solve more word problems from our adversarial version of GSM8K-Platinum, leveraging explicit reflection. For the definition of pre-training compute please refer to section 4.3.

- LIMO hypothesis: “Sophisticated reasoning can emerge through **minimal** but **precisely orchestrated** demonstrations”
  - S1: not simple and not cheap: replies on gemini
  - LIMO: not simple and selection criteria is not reproducible or minimal by any scientific definition
- i.e. 1K **well-chosen** data can outperform 100K **low-quality** data → obviously, many papers in literature on pre-training and SFT data curation
- Framing of the papers can be toned down

RL with one example:

> but under-reported base model accuracy numbers make the improvements seem much larger than they are.

>

Model	Dataset	RL Acc (their method)	Reported Base model Accuracy	Actual Base model Accuracy
DeepSeek-R1-Distill-Qwen-1.5	MATH500	78.0	71.0	84.9
DeepSeek-R1-Distill-Qwen-1.5	AMC23	71.6	60.9	71.5
DeepSeek-R1-Distill-Qwen-1.5	AIME24	25.8	20.0	28.7

> but this does **not** mean SFT will **only** benefit such models – right, more experiments are required and haven't been done in the papers

# Proponent

Reasoning Models:

- S1: Simple test-time scaling
- LIMO: Less is More for Reasoning
- Does Math Reasoning Improve General LLM Capabilities?

**Juno Kim, Shangyin Tan**

Sep 30th

# #1.1 Less may not be more for reasoning

- LIMO hypothesis: “Sophisticated reasoning can emerge through minimal but precisely orchestrated demonstrations”
- i.e. 1K **well-chosen** data can outperform 100K **low-quality** data
- NOT: “less data is always better”

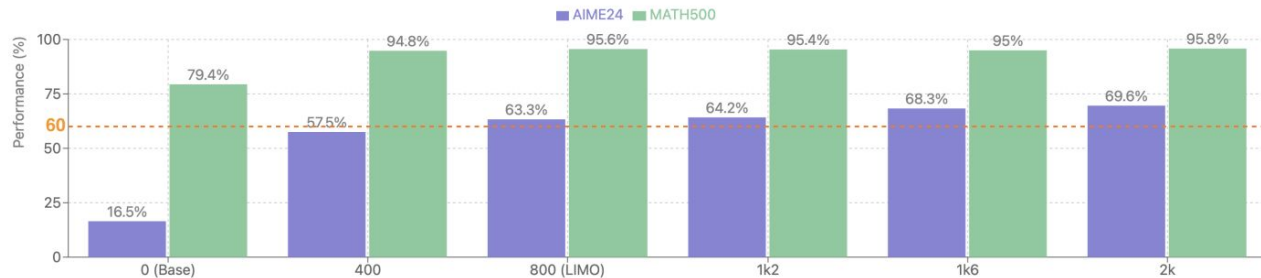
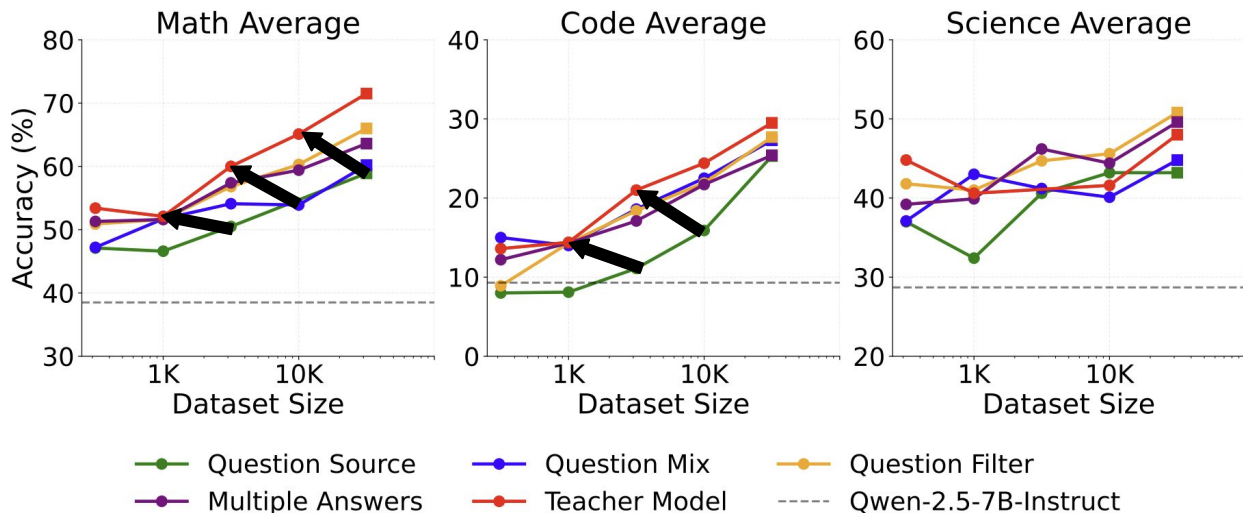


Figure 8: Impact of dataset size on model performance

Increasing LIMO size does improve performance but plateaus early

# #1.1 Less may not be more for reasoning

- LIMO hypothesis: “Sophisticated reasoning can emerge through minimal but precisely orchestrated demonstrations”
- i.e. 1K **well-chosen** data can outperform 100K **low-quality** data
- NOT: “less data is always better”



In fact, each step of Openthoughts3 pipeline results in improved acc. **even with less data!**

## #1.2 Base model as the BIG confounding factor

1. Baselines (OpenThoughts-114k, NuminaMath-100k) share same LLM backbone  
→ not really a confounding factor in the statistical sense, just a limitation (which we agree with!)
2. The “*RL with one training example*” paper does not argue that Qwen is exceptional; they show 1-shot RLVR improves both Qwen2.5-Math-1.5B and Llama3.2-3B-Instruct
3. The “*incorrect label rewards*” paper does show Qwen may already have strong reasoning patterns (e.g. code-reasoning), but this does **not** mean SFT will **only** benefit such models  
→ more work is needed to understand link between base model & SFT v.s. RL



## #1.3 Distillation, Compute and Cost of Data

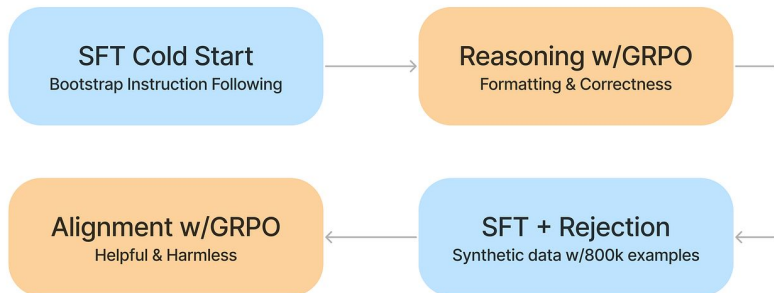
- While S1/LIMO need to distill from a strong thinking model (Gemini/DeepSeek etc.), recent work suggests weaker teachers can be more suitable for teaching advanced reasoners
- Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling (ICLR'25): distilling from a “**weaker & cheaper**” teacher consistently outperforms “**stronger but expensive**” teachers under a fixed budget, even when teaching a more powerful student
- Also, this line of work suffers less from scalability issues – as we need less data

# #1.4 RL is important and SFT isn't enough!

- The paper is not arguing that RL is not needed, but just that SFT can be made more **data-efficient**. SFT is still an important part of the LRM pipeline
- Actually, this means we may have **more compute** for a subsequent RL stage!  
→ Interesting to see if LIMO affects the usefulness of further RL/SFT stages or analyze tradeoff under total fixed budget



## DeepSeek-R1 Training Pipeline



## #1.5 Evaluation

- We agree that contamination and human bias could be an issue.
- Since the LIMO dataset is public, we easily can run more advanced dedup methods (e.g. locally sensitive hashing, [example](#) of noise-robust neural dedup with contrastive training) to check for contamination
- S1K is even better: [the documentation](#) (data section) explicitly contains the pipeline to recreate the dataset, so we can add more dedup filters and improve it ourselves

## #2 On Math Reasoning Improves General Capabilities

TI is defined as

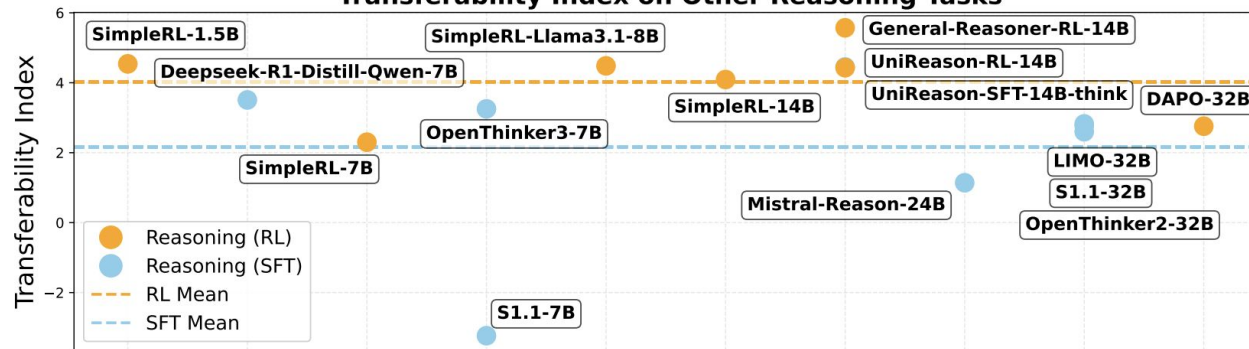
$$\Delta R_g = \frac{1}{|\mathcal{B}_g|} \sum_{b \in \mathcal{B}_g} \frac{R_b^{\text{model}} - R_b^{\text{base}}}{R_b^{\text{base}}}, \quad g \in \{\text{math}, \text{other}, \text{non}\}.$$

Essentially this says: how much better is model compared to base

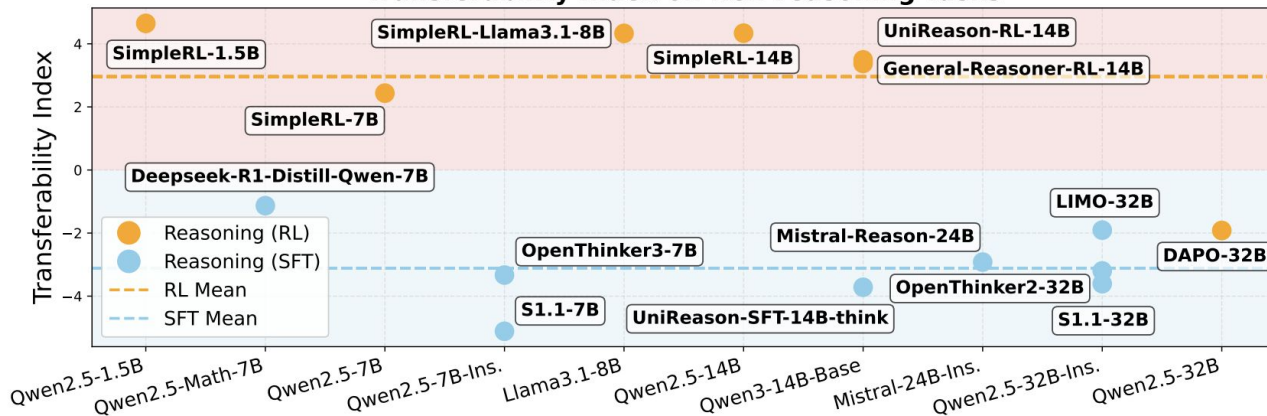
$$\text{TI}_{\text{non}}(\%) = \frac{\Delta R_{\text{non}}}{\Delta R_{\text{math}}} \times 100.$$

# #2 On Math Reasoning Improves General Capabilities

Transferability Index on Other Reasoning Tasks



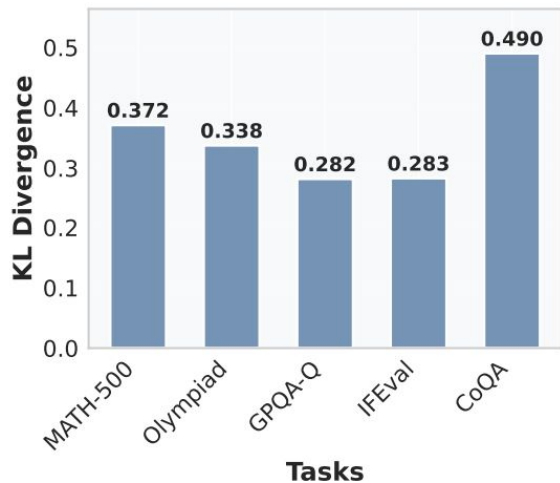
Transferability Index on Non-reasoning Tasks



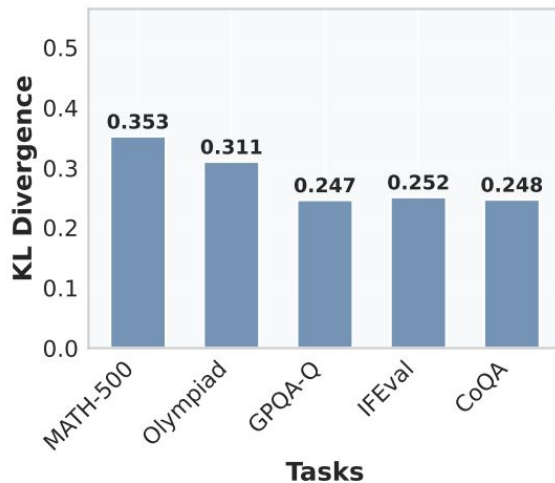
Clearly reasoning tasks have higher TI. But non-reasoning ones have positive TIs too.

# #2 On Math Reasoning Improves General Capabilities

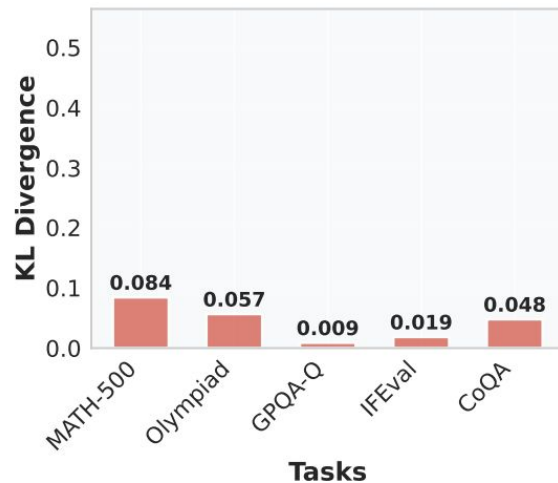
**SFT (no think)**



**SFT (think)**

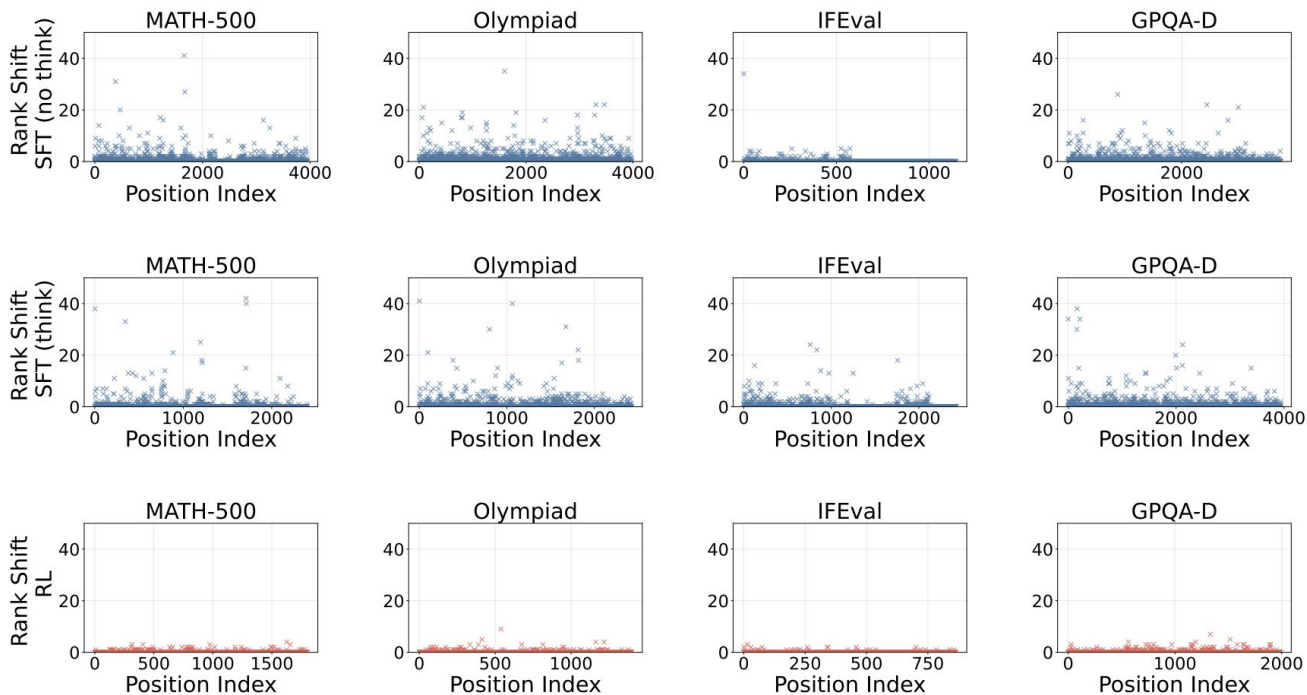


**RL**



KL divergence is lower for RL

# #2 On Math Reasoning Improves General Capabilities



**Figure 6** | Visualization of token rank shifts across different position indices for both reasoning and non-reasoning tasks. We observe that RL models exhibit less token rank shifts while SFT models demonstrate substantial rank shifts across numerous positions throughout the sequence.

## #2 On Math Reasoning Improves General Capabilities

*"RL models selectively shift task-relevant tokens, while SFT models shift numerous irrelevant tokens."*

**Table 8** | Performance of off-the-shelf models on Non-Reasoning Benchmarks

Model	CoQA	MC-TACO	IFEval	HalluEval
Qwen3-14B-Base	10.0	67.7	69.2	35.7
Qwen3-14B (think)	2.6	66.1	42.9	5.5
Qwen3-14B (no-think)	44.1	74.9	90.5	70.7
General-Reasoner-Qwen3-14B (SFT)	1.4	49.4	31.7	10.8
General-Reasoner-Qwen3-14B (RL)	4.8	51.19	72.0	55.7
UniReason-Qwen3-14B- <i>think</i> (SFT)	1.7	38.2	42.3	2.3
UniReason-Qwen3-14B- <i>no-think</i> (SFT)	5.3	66.1	41.4	3.3
UniReason-Qwen3-14B (RL)	28.2	74.0	70.0	40.7

Final remark for critics' slides: KL Regularization might not be needed (per CWM paper from Meta)