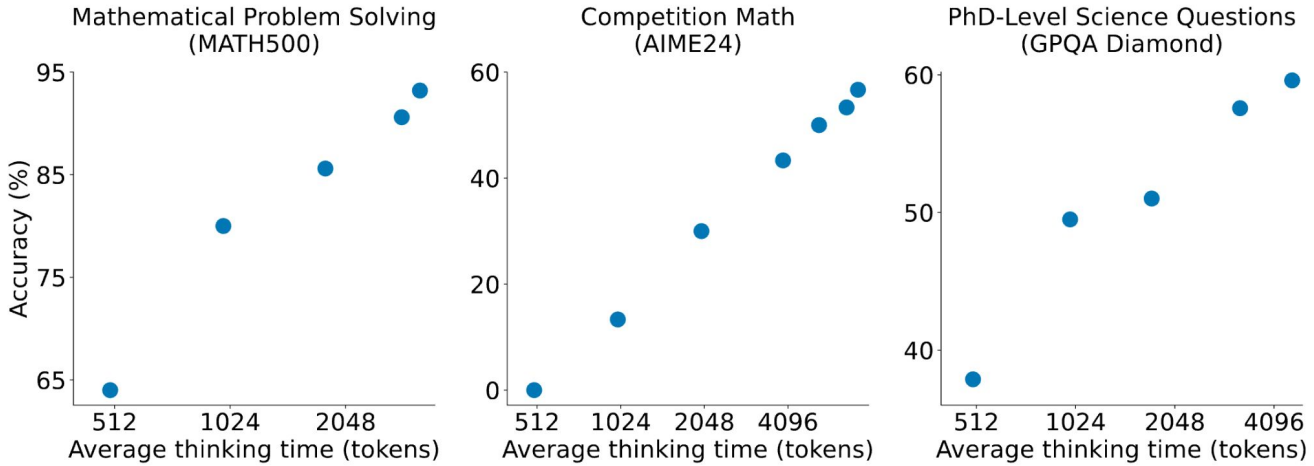
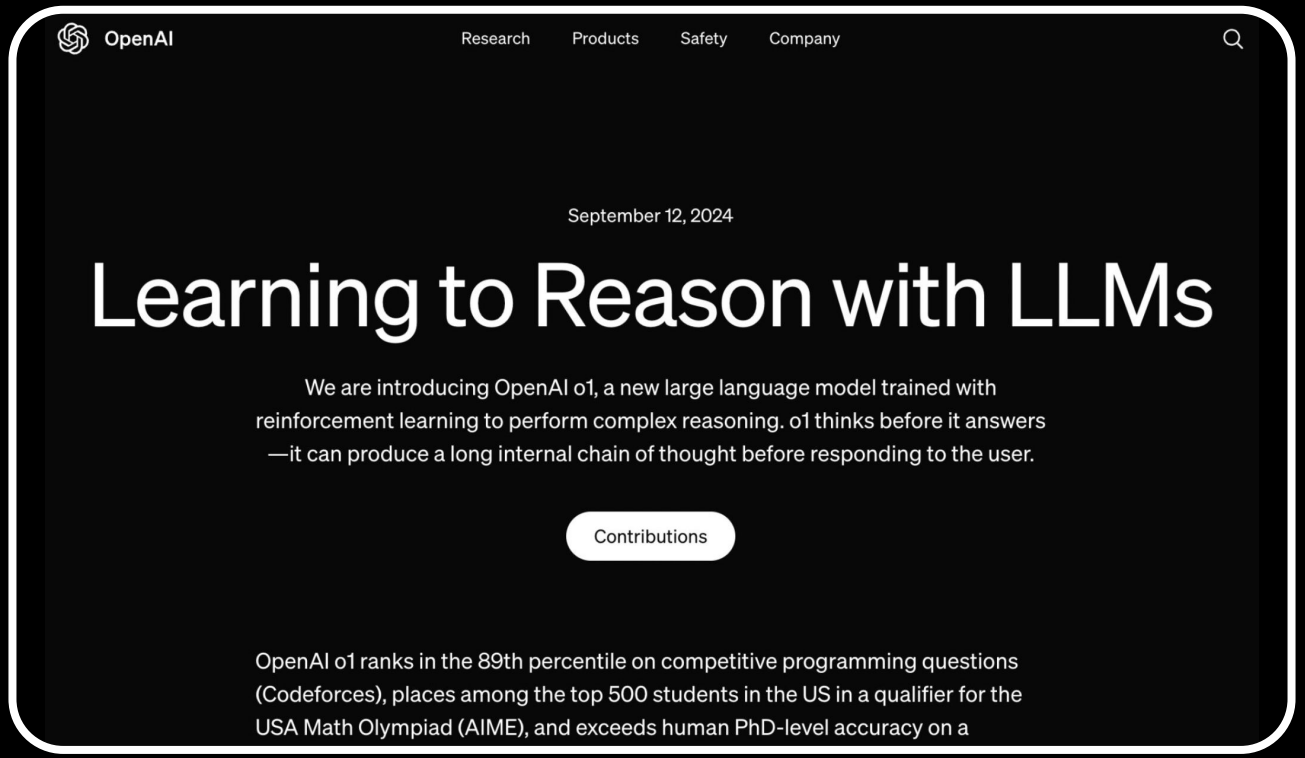
s1: Simple test-time scaling

Niklas Muennighoff * Zitong Yang * Weijia Shi * Xiang Lisa Li * Li Fei-Fei Hannaneh Hajishirzi
Luke Zettlemoyer Percy Liang Emmanuel Candès Tatsunori Hashimoto

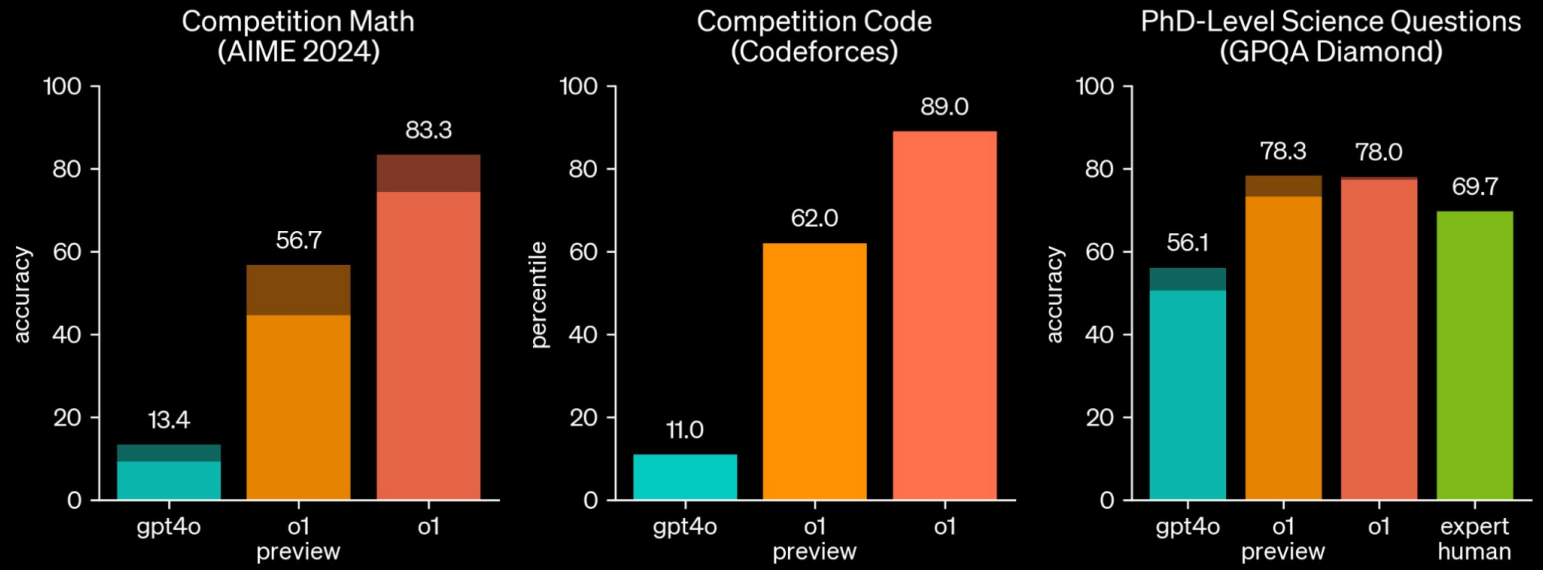


Niklas Muennighoff
X: @Muennighoff

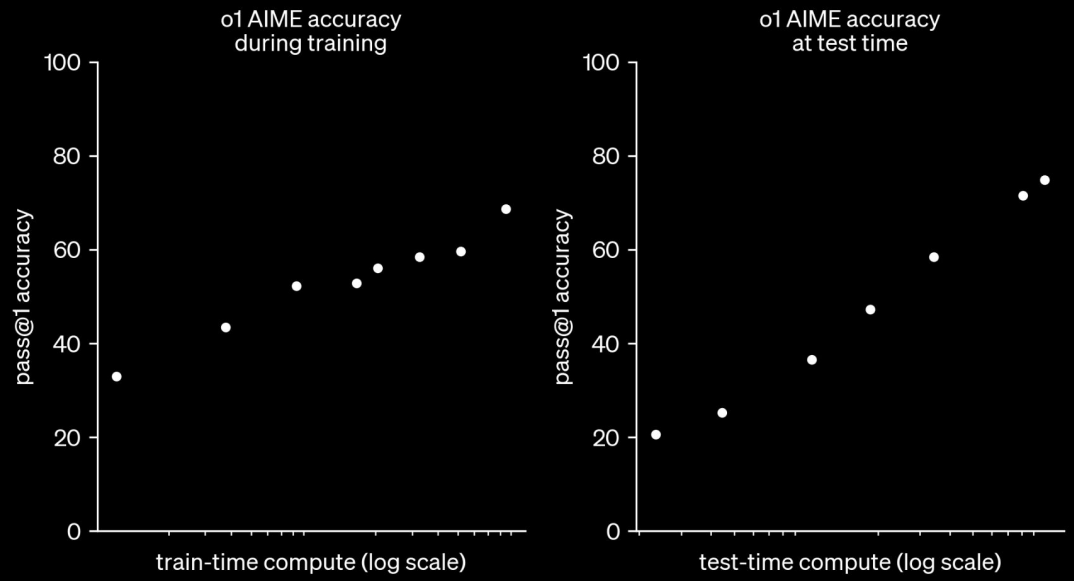
OpenAI o1


OpenAI
Research
Products
Safety
Company
September 12, 2024
Learning to Reason with LLMs
We are introducing OpenAI o1, a new large language model trained with reinforcement learning to perform complex reasoning. o1 thinks before it answers —it can produce a long internal chain of thought before responding to the user.
Contributions
OpenAI o1 ranks in the 89th percentile on competitive programming questions (Codeforces), places among the top 500 students in the US in a qualifier for the USA Math Olympiad (AIME), and exceeds human PhD-level accuracy on a

o1 - Reasoning



o1 - Test-time scaling

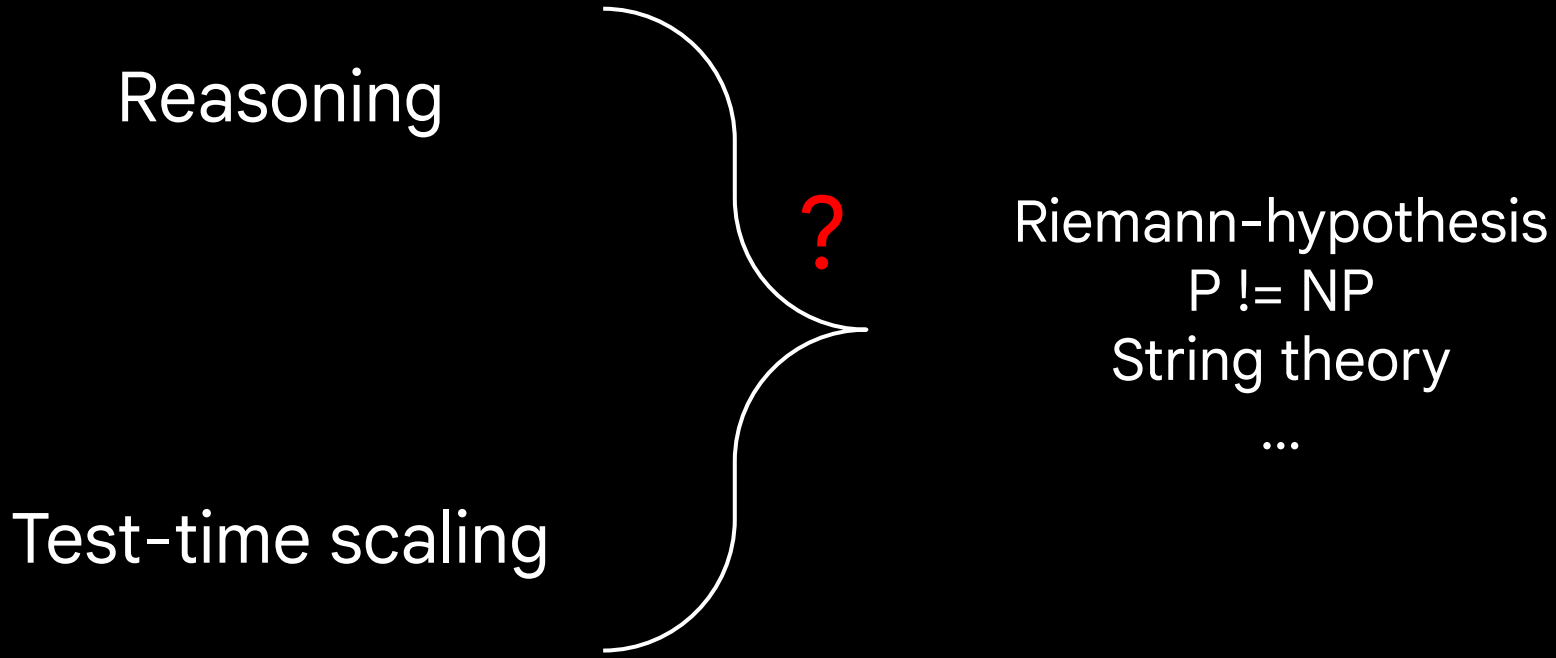


Reasoning

Test-time scaling

Reasoning

Test-time scaling


Reasoning
Test-time scaling
?
Riemann-hypothesis
P != NP
String theory
...

Simplest
Recipe?

s1
Simple test-time scaling

Data		Test-time scaling		
s1K	+	Budget forcing	=	s1

Data

Collect 59K questions

Collect 59K questions

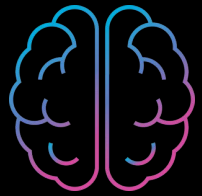
s1-prob

Show that the uniform distribution on \\([0,1]\\] is not the convolution of two independent, identically distributed variables.

Collect 59K questions

s1-prob

Show that the uniform distribution on \\([0,1]\\] is not the convolution of two independent, identically distributed variables.



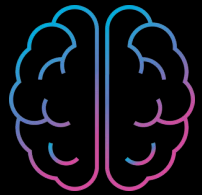
s1-teasers

Suppose you roll three fair 100-sided die. What is the expected value of the lowest roll?

Collect 59K questions

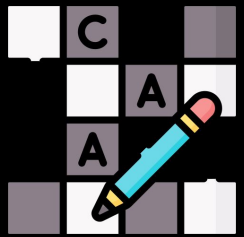
s1-prob

Show that the uniform distribution on \\([0,1]\\] is not the convolution of two independent, identically distributed variables.



s1-teasers

Suppose you roll three fair 100-sided die. What is the expected value of the lowest roll?



...

59K

59K

Quality

52K

59K

Quality

52K

Difficulty

24K

59K

Quality

52K

Difficulty

24K

Diversity

1K

Distill reasoning traces & answers

An often-repeated fun fact is that humans produce more power per unit volume than stars. If the sun were the same size, but it produced the same amount of power per unit volume as a human, what would its surface temperature be?...

Distill reasoning traces & answers

An often-repeated fun fact is that humans produce more power per unit volume than stars. If the sun were the same size, but it produced the same amount of power per unit volume as a human, what would its surface temperature be?...

s1: Google Gemini

The problem asks for the surface temperature of the sun if it had the same size but produced power per unit volume as a human.

First, let's find the power density of a human…

Distill reasoning traces & answers

An often-repeated fun fact is that humans produce more power per unit volume than stars. If the sun were the same size, but it produced the same amount of power per unit volume as a human, what would its surface temperature be?...

s1: Google Gemini

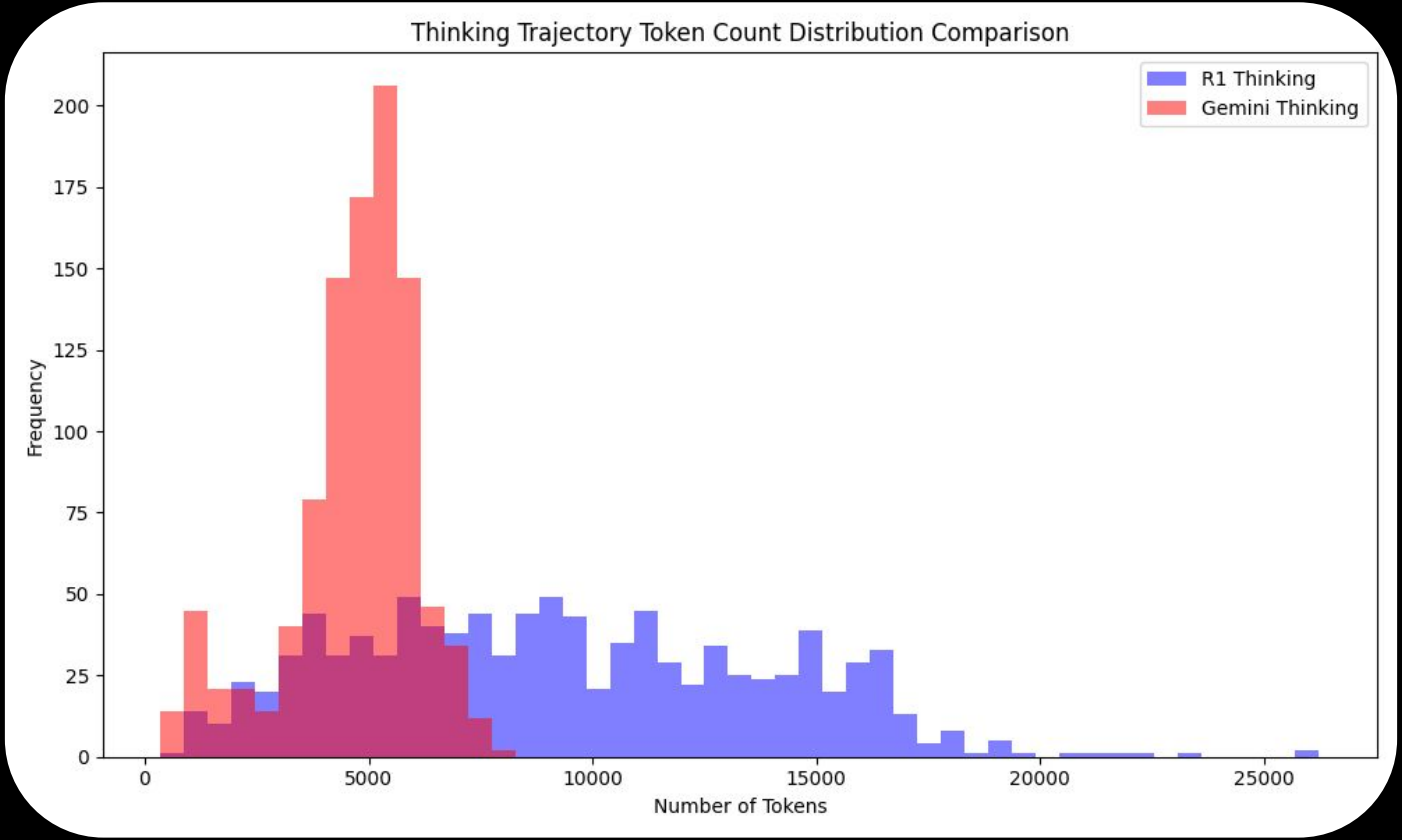
The problem asks for the surface temperature of the sun if it had the same size but produced power per unit volume as a human.

First, let's find the power density of a human…

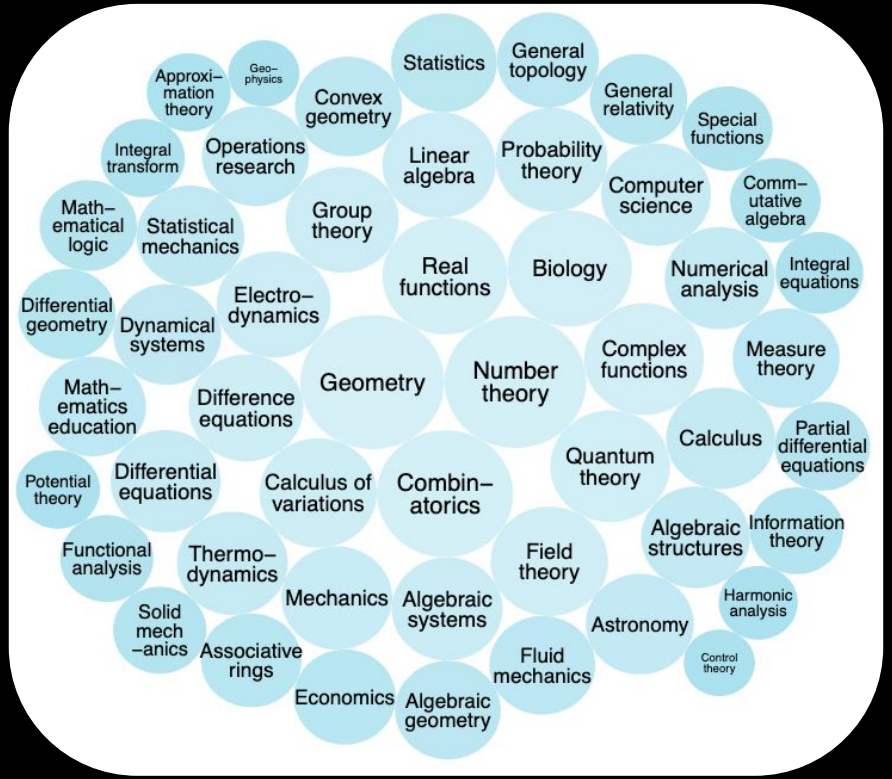
s1.1: DeepSeek r1

Okay, so there's this fun fact saying that humans produce more power per unit volume than stars. The question is asking if the Sun were the same size as a human (but with the same power per unit volume as a human), what would its surface temperature be? The answer choices are given, and I need to figure out which one is correct…

Longer traces



s1K: 1000 **high-quality**, **difficult** & **diverse** samples



Test-time scaling

Budget Forcing

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 $\boxed{3}$ **Response**

Other

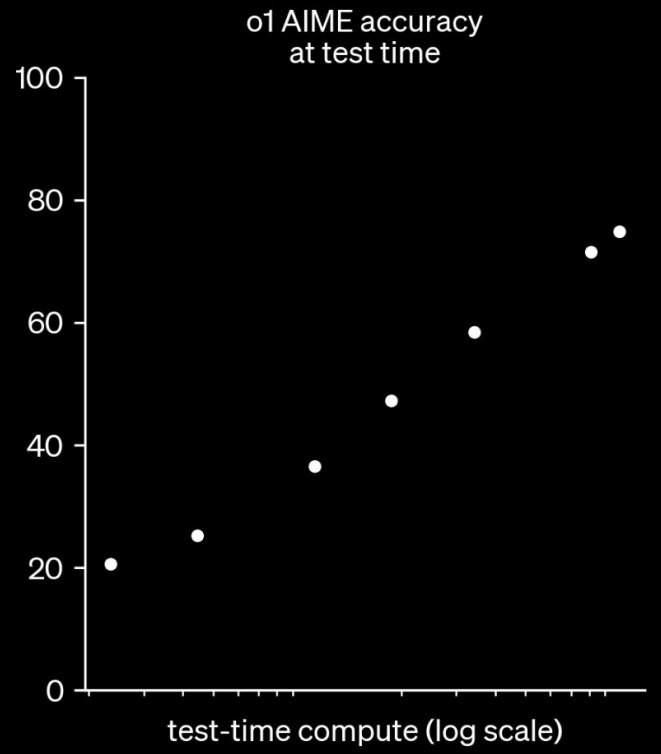
Conditional length-control

Rejection sampling

Control

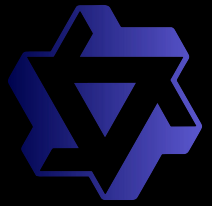
Scaling

Performance



Training & Results

Training



Qwen2.5-32B

16 H100s

26min

Models

SFT / Distill

s1/s1.1

r1-distill

Sky-T1

Bespoke-Stratos

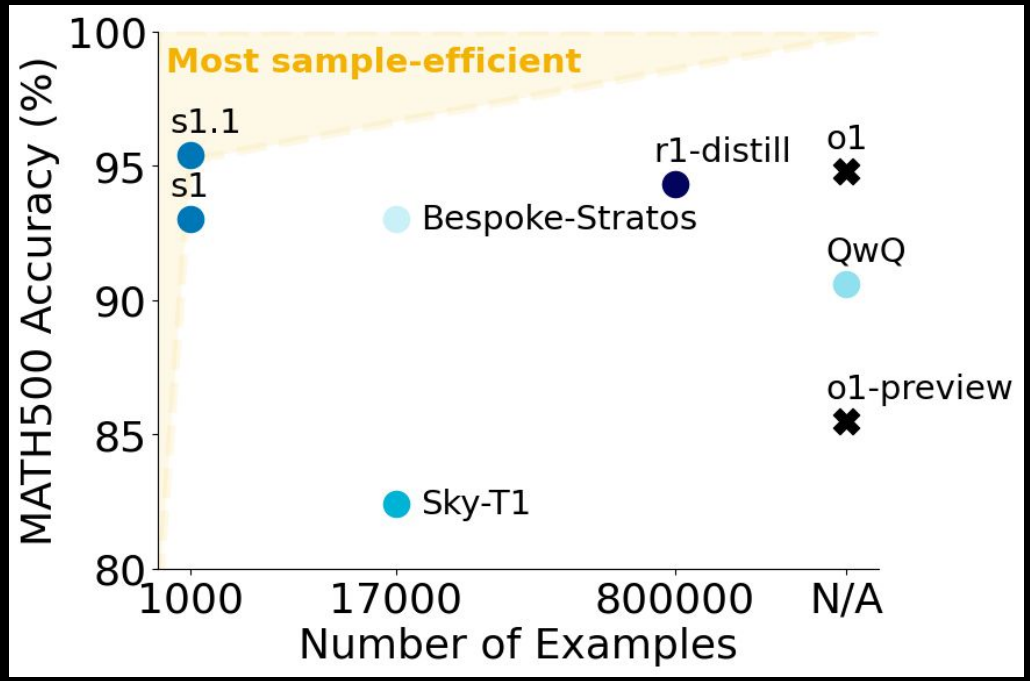
Models

SFT / Distill	(SFT+) RL
s1/s1.1	r1
r1-distill	
Sky-T1	
Bespoke-Stratos	

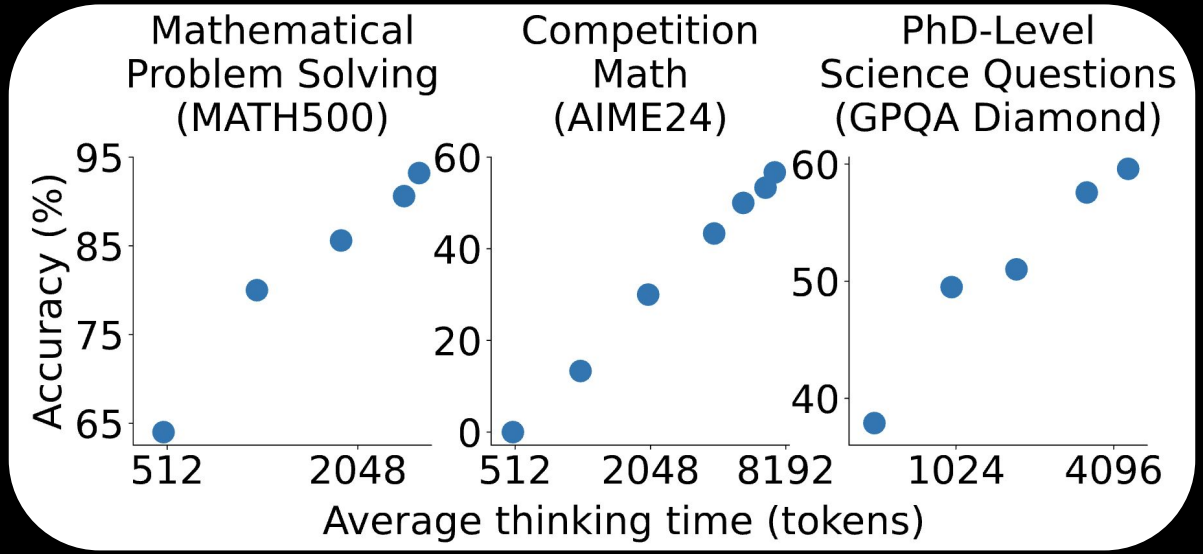
Models

SFT / Distill	(SFT+) RL	?
s1/s1.1	r1	QwQ
r1-distill		o1
Sky-T1		
Bespoke-Stratos		

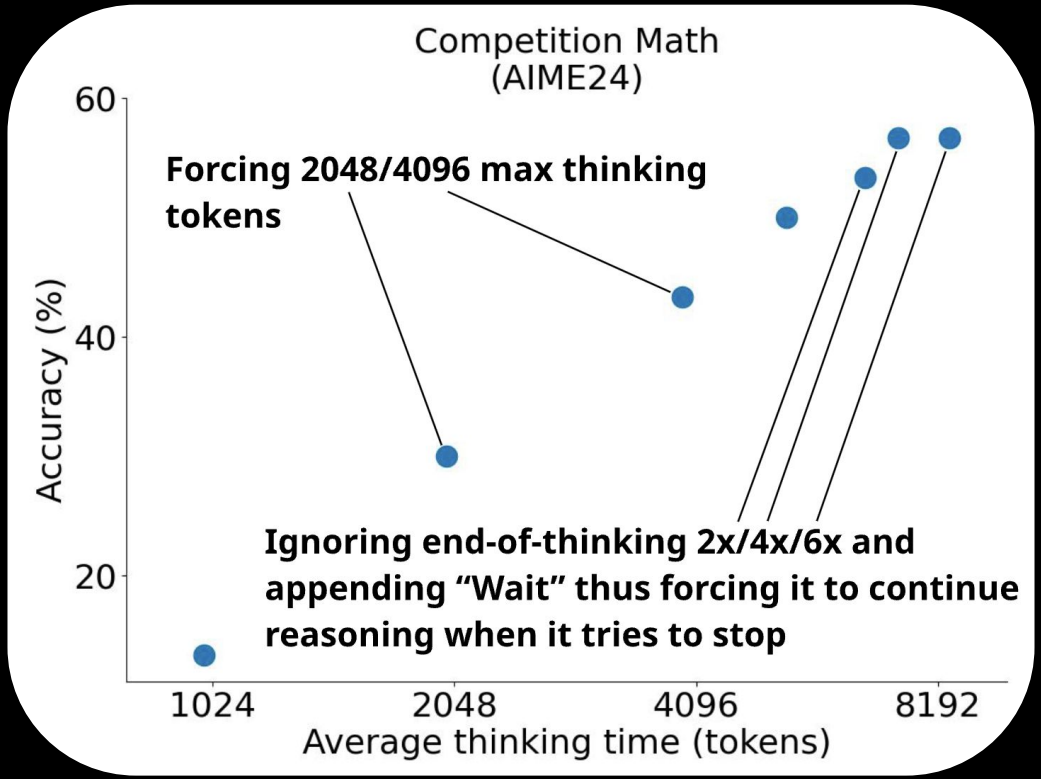
s1 - Strong reasoning



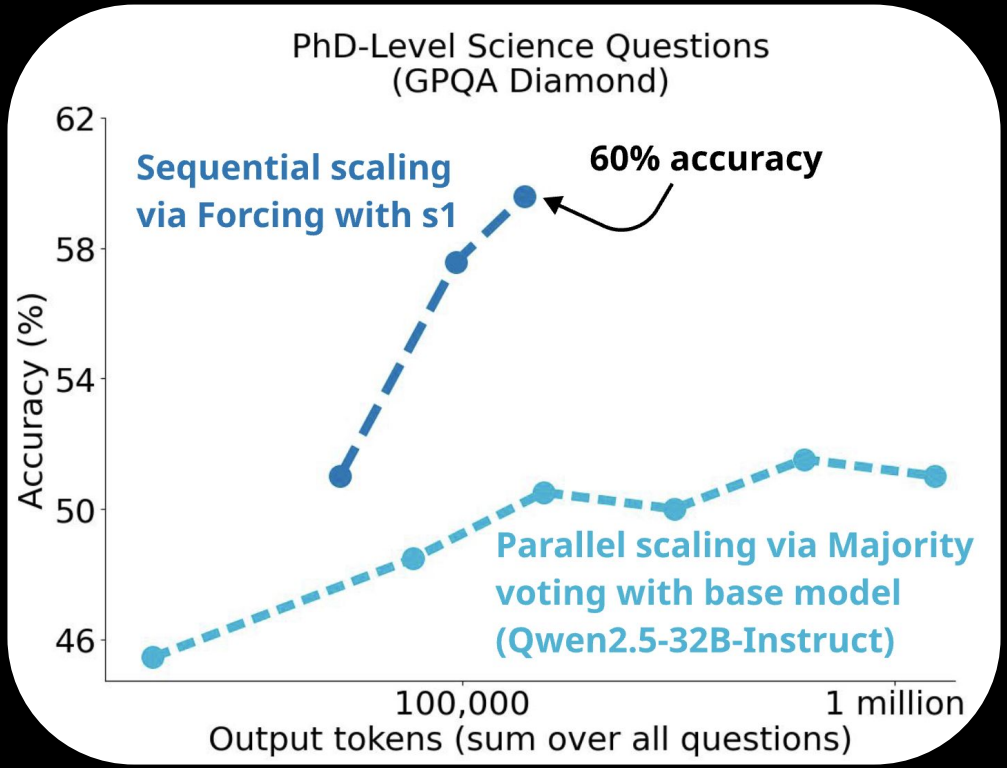
s1 - Test-time scaling



Zooming in


Competition Math
(AIME24)
Forcing 2048/4096 max thinking tokens
Ignoring end-of-thinking 2x/4x/6x and appending “Wait” thus forcing it to continue reasoning when it tries to stop
Accuracy (%)
60
40
20
1024
2048
4096
8192
Average thinking time (tokens)

Sequential vs Parallel



Data ablations

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%]
s1K	50.0	93.0	57.6

Scaling ablations

BF = Budet Forcing

T/S/C-CC = Token/Step/Class-Conditonal Control

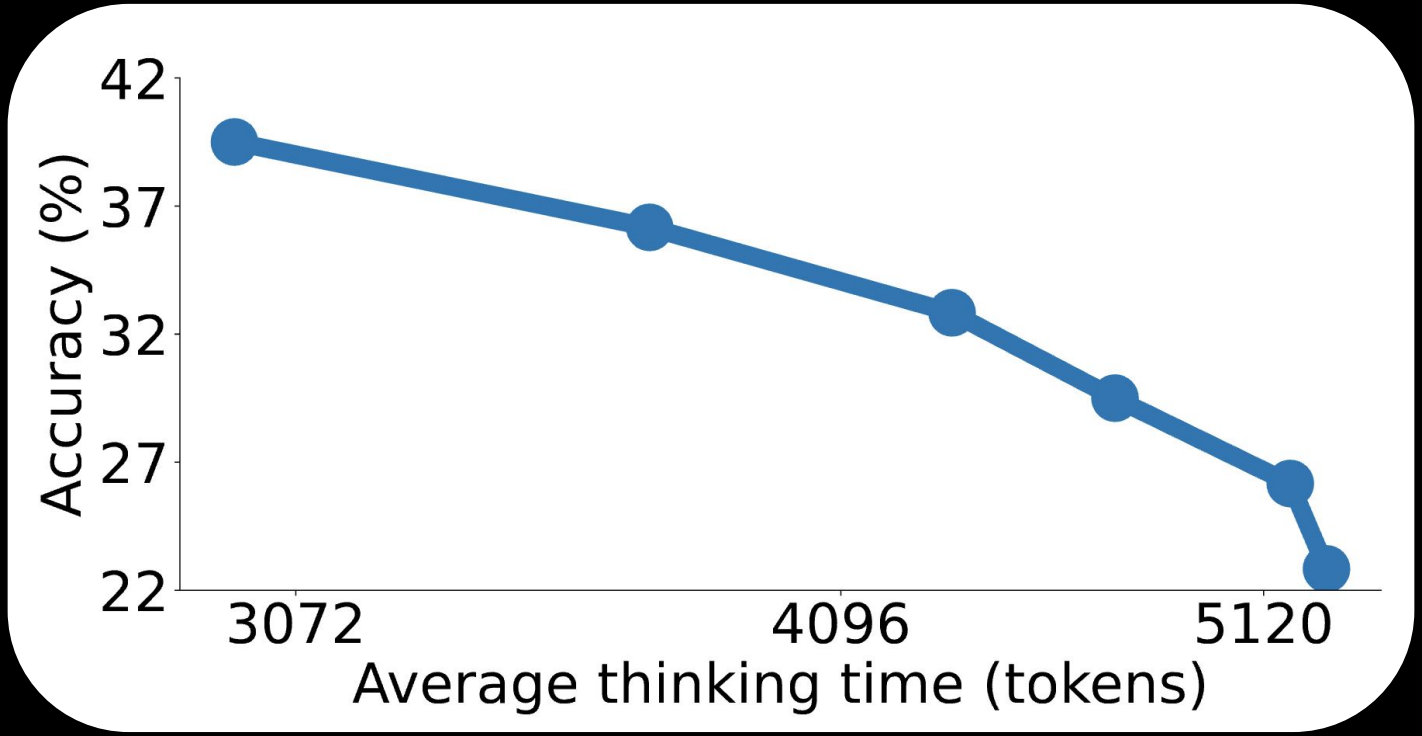
RS = Rejection Sampling

Method	Control	Scaling	Performance
BF	**100%**	15	**56.7**
TCC	40%	-24	40.0
TCC + BF	**100%**	13	40.0
SCC	60%	3	36.7
SCC + BF	**100%**	6	36.7
CCC	50%	**25**	36.7
RS	**100%**	-35	40.0

Scaling ablations

Model	AIME 2024	MATH 500	GPQA Diamond
No extrapolation	50.0	**93.0**	57.6
2x without string	50.0	90.2	55.1
2x “Alternatively”	50.0	92.2	**59.6**
2x “Hmm”	50.0	**93.0**	**59.6**
2x “Wait”	**53.3**	**93.0**	**59.6**

Rejection sampling

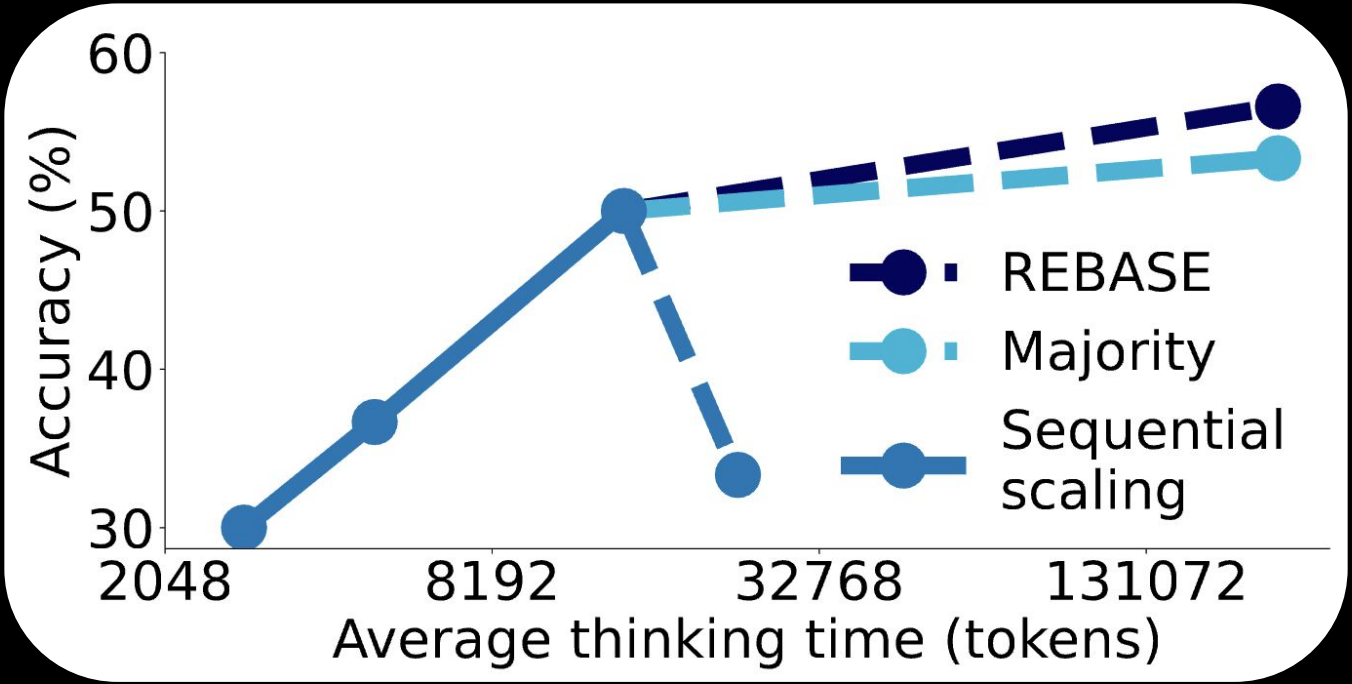


Scaling limits

Flattening

Context Window

Scaling further



Minimal recipe for **Reasoning** & **Test-time scaling**

s1K + Budget forcing = s1

Thanks!

Open-source on GitHub: simplescaling/s1
arxiv.org/abs/2501.19393

Niklas Muennighoff
X: @Muennighoff

Acknowledgements: Thanks to my amazing co-authors: Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, Tatsunori Hashimoto. Thanks to Ryan Marten for help with generating traces from DeepSeek r1. Thanks to Stanford for the Stanford Marlowe GPU cluster. Thanks to Alexander M. Rush, Andrew Ilyas, Banghua Zhu, Chenglei Si, Chunting Zhou, John Yang, Ludwig Schmidt, Samy Jelassi, Suhas Kotha, Tengyu Ma, Xuechen Li, Yu Sun and Yue Zhang for very constructive discussions. Thanks to Jason Wei for slide inspiration.

Backup

Token-Conditional Control

Tokens instructed (→)	1024	2048	4096	8192	16384
No intervention at test-time					
Thinking tokens	7939	7158	8263	7108	7500
Answer tokens	689	669	659	722	724
AIME24	26.7	30.0	33.3	33.3	40.0
Forcing end of thinking when token budget is reached					
Thinking tokens	1024	2048	4031	5664	6330
Answer tokens	15	15	142	722	691
AIME24	3.3	30.0	33.3	33.3	40.0

Step-Conditional Control

Steps instructed (→)	16	32	64	128	256
No intervention at test-time					
Steps used	123	90	80	82	136
Tokens per step	60	70	69	66	56
Thinking tokens	7252	6277	5396	5552	7551
Answer tokens	665	653	735	777	754
AIME24	33.3	23.3	33.3	36.7	33.3
Forcing end of thinking when 0 steps are reached					
Steps used	16	32	59	78	136
Tokens per step	96	94	80	70	56
Thinking tokens	1517	2963	4636	5409	7551
Answer tokens	1111	788	799	794	754
AIME24	23.3	23.3	33.3	36.7	33.3

Class-Conditional Control

Prompt appended to the question after two newlines	AIME24	MATH500	GPQA
Answer after a short amount of thinking. Do not spend excessive time double-checking your work.	30.0% / 8033	90.4% / 2537	56.6% / 4177
Answer after a long amount of thinking. If you feel like you are finished early, spend the extra time trying to double-check your work until you are absolutely sure that you have the correct answer.	36.7% / 9651	91.4% / 3875	51.0% / 4827
Without generic prompt appending	50.0% / 6109	93.0% / 3298	57.6% / 3510

s1.1

Model	# Examples	AIME 2024	AIME 2025 I
API only			
o1-preview	N/A	44.6	37.5
o3-mini-low	N/A	56.3	44.2
o3-mini-medium	N/A	75.8	66.7
o3-mini-high	N/A	**83.8**	**76.7**
Open Weights			
QwQ-32B	N.A.	46.7	37.2
r1	≫800K	**79.8**	**65.0**
r1-distill-Llama-70B	800K	57.1	51.4
r1-distill-Qwen-14B	800K	61.7	46.7
r1-distill-Qwen-32B	800K	58.3	46.1
Open Weights and Open Data			
LIMO	817	56.3	44.5
s1 vanilla	1K	50.0	26.7
s1 with Budget Forcing "Wait" 1x	1K	53.3	26.7
s1.1 vanilla	1K	**56.7**	53.3
s1.1 with Budget Forcing "Wait" 1x	1K	**56.7**	**60.0**