An Introduction to Reasoning Models

Anoop Kunchukuttan

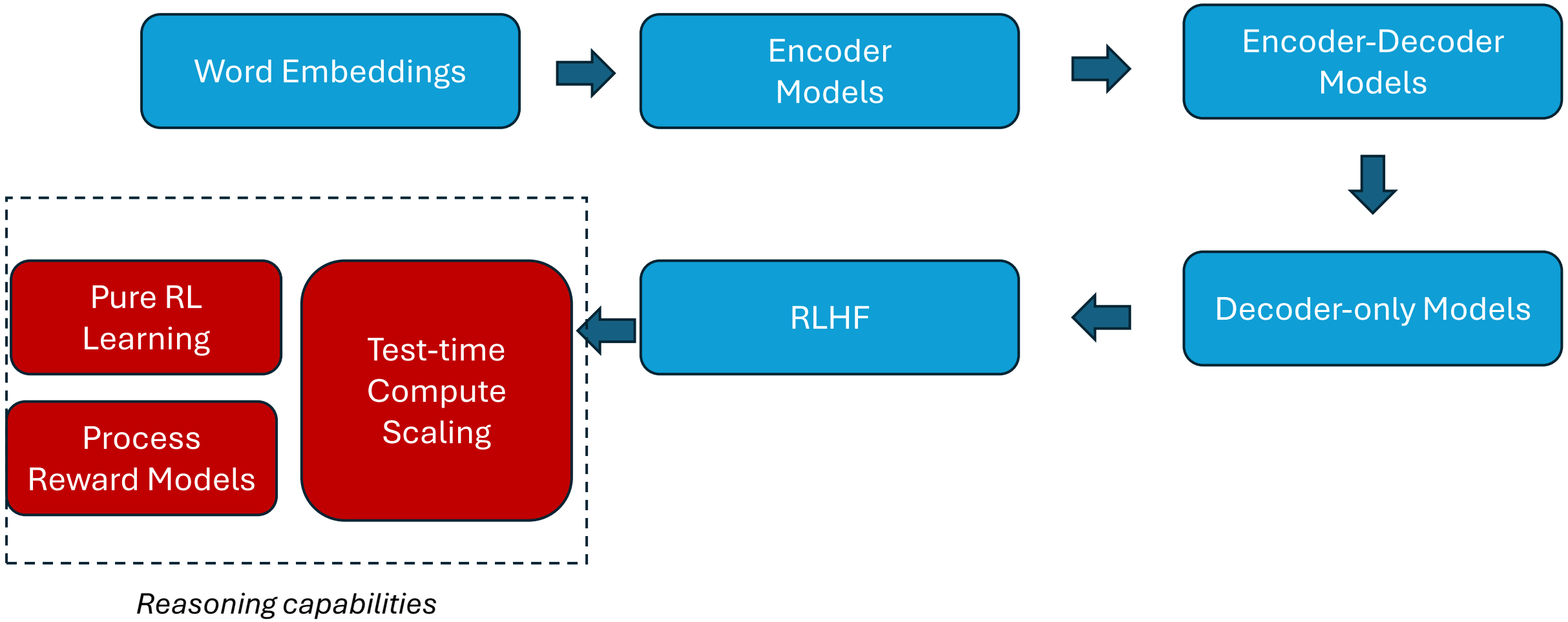
Microsoft

AI4Bharat

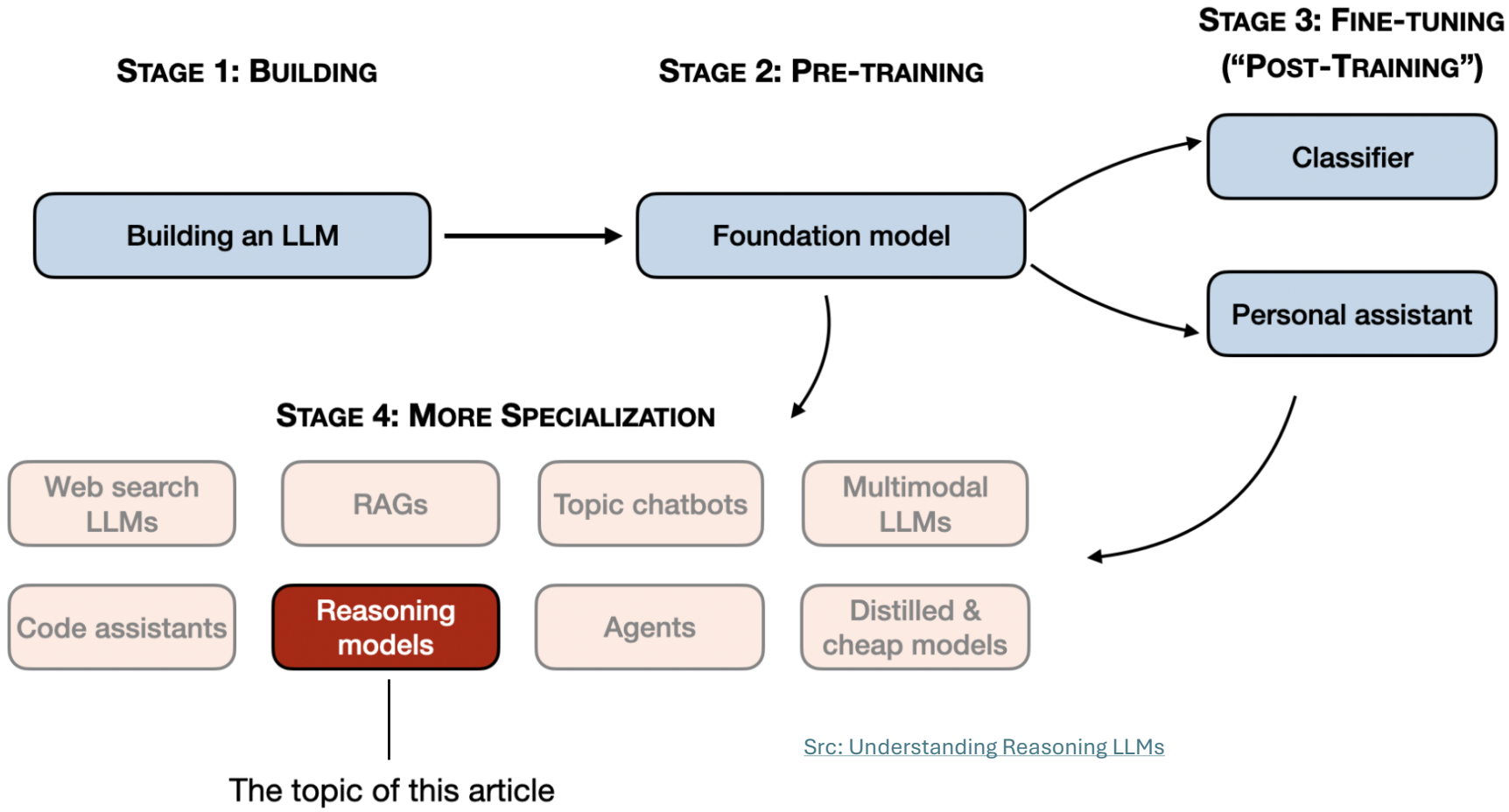
3rd July 2025

EMPOWERING INDIA THROUGH INCLUSIVE GENERATIVE AI
OdiaGen Workshop on Generative AI

Language Model Evolution



LLM Specialization



Src: Understanding Reasoning LLMs

What is a reasoning model?

What is a reasoning model?

What is the value of 7 + 5 x 2?

17

Normal Response

Step 1: Understand the order of operations:
Brackets, Orders, Multiplication/Division, Addition/Subtraction
Step 2: Identify operations in expression
- Addition: 7+
- Multiplication: 5 x 2
Step 3: Perform Multiplication: 5 x 2 = 10
Step 4: Perform Addition: 7 + 10 = 17

Answer: 17

Normal Response

An LLM that can produce intermediate steps for solving a problem before providing the final answer

Why is 31 a prime number?

Reasoned about number 31 being prime for 4 seconds

> Alright, let's go through this. A prime number is an integer greater than 1, divisible only by 1 and itself. So why is 31 prime? Its only divisors are 1 and 31. I check for divisibility — it's not divisible by 2 (31/2 = 15.5, remainder 1), not by 3 (31/3 = 10.33, remainder 1), and not divisible by 5 (31/5 = 6.2, remainder 1). Since no other primes divide it, 31 is prime!
>
> Since 31 has no divisors other than 1 and itself, it is prime. To determine this, we check possible divisors up to sqrt(31), approximately 5.6. The divisors to check are 2, 3, and 5, and none divide evenly into 31 — 31/2 = 15.5, 31/3 ≈ 10.33, 31/5 = 6.2. Since 31 isn't divisible by these, it confirms that 31 is prime. The fundamental theorem of arithmetic also supports that 31 cannot be factored further.

Thinking stage

A prime number is defined as a natural number greater than 1 that has no positive divisors other than 1 and itself. In the case of 31, here's why it is prime:

1. Divisibility Check:
   To determine if 31 is prime, we check whether any number other than 1 and 31 divides it evenly. In practice, you only need to test divisibility by prime numbers up to the square root of 31.
   - The square root of 31 is approximately 5.57.
   - The prime numbers less than or equal to 5.57 are 2, 3, and 5.
2. Testing Each Prime:
   - Division by 2:
     $31 \div 2 = 15.5$
     Since 15.5 is not an integer, 2 is not a divisor of 31.
   - Division by 3:
     $31 \div 3 \approx 10.33$
     Again, this is not an integer, so 3 does not divide 31 evenly.
   - Division by 5:
     $31 \div 5 = 6.2$
     This also is not an integer, so 5 is not a divisor.
3. Conclusion:
   Since none of these potential divisors (2, 3, or 5) divides 31 evenly, there are no divisors of 31 other than 1 and 31 itself. By definition, this makes 31 a prime number.

So, 31 is prime because its only positive divisors are 1 and 31.

Answering stage: explanation

Answering stage: final answer

More thinking time leads to better answers

anoopkunchukuttan 6 minutes ago

What is the distance, in units, between the points $(2, -6)$ and $(-4, 3)$?

DeepSeek-R1 5 minutes ago

<think> Okay, so I need to find the distance between the points (2, -6) and (-4, 3). Hmm, distance between two points... I remember there's a formula for that. Let me think. Is it the Pythagorean Theorem applied to coordinates? Yeah, something like that. The distance formula, right?
From what I recall, if you have two points in a plane, say (x1, y1) and (x2, y2), the distance between them is calculated by taking the square root of the sum of the squares of the differences in their x-coordinates and y-coordinates. So the formula should be:

Distance = $\sqrt{[(x2 - x1)^2 + (y2 - y1)^2]}$

Let me verify that. If I plot these two points on a coordinate plane, the horizontal distance between them would be the difference in the x-coordinates, and the vertical distance would be the difference in the y-coordinates. Then, the straight-line distance between them would be the hypotenuse of a right-angled triangle formed by those two differences. Yep, that makes sense. So applying the Pythagorean Theorem gives us the distance formula.

Alright, so applying this formula to the given points. Let's label them: the first point is (2, -6), so x1 is 2 and y1 is -6. The second point is (-4, 3), so x2 is -4 and y2 is 3.

Plugging these into the formula:

Distance = $\sqrt{[(-4 - 2)^2 + (3 - (-6))^2]}$

Let me calculate the differences first. For the x-coordinates: -4 minus 2. That's -6, right? So the horizontal difference is -6 units. Squared, that should be positive 36.

For the y-coordinates: 3 minus (-6). Subtracting a negative is like adding, so that's 3 + 6 = 9. The vertical difference is 9 units. Squared is 81.

Now, adding those two squared differences: 36 + 81 = 117. Then take the square root of 117. Wait, is that right? Let me double-check my calculations.

First, horizontal difference: x2 - x1 is (-4) - 2. So that's indeed -6. Squared, (-6)^2 is 36. Correct.

Vertical difference: y2 - y1 is 3 - (-6). Which is 3 + 6 = 9. Squared is 81. Correct.

Sum: 36 + 81 = 117. Correct. So square root of 117. Hmm, 117. What's the square root of 117? Is that a simplified radical form?

Let me see. 117 factors into 9 times 13. Because 9 * 13 is 117, and 9 is a perfect square. So $\sqrt{117} = \sqrt{(9*13)} = \sqrt{9} * \sqrt{13} = 3\sqrt{13}$. So the distance is $3\sqrt{13}$ units.

Thinking ...
Very detailed
self-dialogue

https://github.com/marketplace/models/azureml-deepseek/DeepSeek-R1/playground

Final Answer The distance between the points is \boxed{3\sqrt{13}} units. </think>

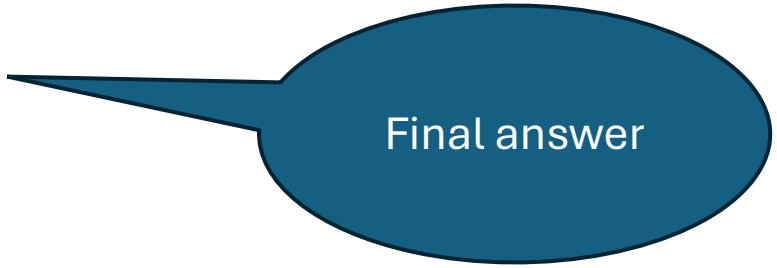
To find the distance between the points ((2, -6)) and ((-4, 3)), we use the distance formula:

[\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}]

Labeling the points, we have ((x_1, y_1) = (2, -6)) and ((x_2, y_2) = (-4, 3)).

First, calculate the differences in the coordinates:

- The horizontal difference: (x_2 - x_1 = -4 - 2 = -6)
- The vertical difference: (y_2 - y_1 = 3 - (-6) = 3 + 6 = 9)



Next, square these differences:

- ((-6)^2 = 36)
- (9^2 = 81)

Sum the squared differences: [36 + 81 = 117]

Take the square root of the sum: [\sqrt{117} = \sqrt{9 \times 13} = \sqrt{9} \times \sqrt{13} = 3\sqrt{13}]

Thus, the distance between the points ((2, -6)) and ((-4, 3)) is (\boxed{3\sqrt{13}}) units.

When should we use reasoning models?

Good at	Bad at
+ Deductive or inductive reasoning (e.g., riddles, math proofs)	– Fast and cheap responses (more inference time)
+ Chain-of-thought reasoning (breaking down multi-step problems)	– Knowledge-based tasks (hallucination)
+ Complex decision-making tasks	– Simple tasks (“overthinking”)
+ Better generalization to novel problems	

The key strengths and weaknesses of reasoning models.

Src: Understanding Reasoning LLMs

Domains where reasoning models are typically used

… domains where results can be objectively measured

Can reasoning models be extended to knowledge intensive domains like science, medicine, economics?

The big challenge .. and there are ongoing attempts

DeepSeek Impact

- **DeepSeek v3**: Open-weight Frontier LLM trained using very efficient methods at a cheap cost on sub-optimal hardware.

- **DeepSeek R1**: Open-weight State-of-the art reasoning model competitive with OpenAI's o1 models.

Open-weight, efficient, state-of-the results, well-documented methods!

DeepSeek v3 Model Summary

- 671B parameters, MoE, 37B active parameters
- Trained on 15T tokens
- Trained on 2048 GPUs for 2 months, $6m
- Efficiency through techniques like
  - FP8 training
  - Improved quantization
  - Multi-head latent attention
  - Aux loss free load balancing
  - MoE optimizations
  - Multi-token predictions
- Competitive with all frontier models

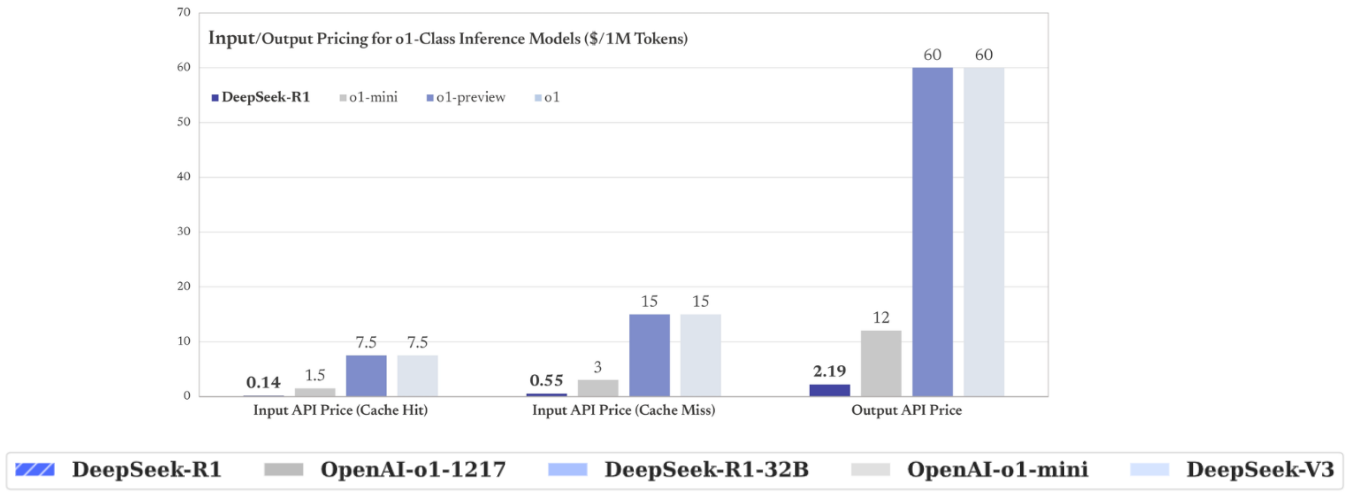
Benchmark (Metric)		DeepSeek-V3	Qwen2.5 72B-Inst.	Llama3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513
	Architecture	MoE	Dense	Dense	-	-
	# Activated Params	37B	72B	405B	-	-
	# Total Params	671B	72B	405B	-	-
English	MMLU (EM)	**88.5**	85.3	**88.6**	88.3	87.2
	MMLU-Redux (EM)	**89.1**	85.6	86.2	88.9	88
	MMLU-Pro (EM)	75.9	71.6	73.3	**78**	72.6
	DROP (3-shot F1)	**91.6**	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	86.1	84.1	86	**86.5**	84.3
	GPQA-Diamond (Pass@1)	59.1	49	51.1	**65**	49.9
	SimpleQA (Correct)	24.9	9.1	17.1	28.4	**38.2**
	FRAMES (Acc.)	73.3	69.8	70	72.5	**80.5**
	LongBench v2 (Acc.)	**48.7**	39.4	36.1	41	48.1
Code	HumanEval-Mul (Pass@1)	**82.6**	77.3	77.2	81.7	80.5
	LiveCodeBench(Pass@1-COT)	**40.5**	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	**37.6**	28.7	30.1	32.8	34.2
	Codeforces (Percentile)	**51.6**	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	42	23.8	24.5	**50.8**	38.8
	Aider-Edit (Acc.)	79.7	65.4	63.9	**84.2**	72.9
	Aider-Polyglot (Acc.)	**49.6**	7.6	5.8	45.3	16
Math	AIME 2024 (Pass@1)	**39.2**	23.3	23.3	16	9.3
	MATH-500 (EM)	**90.2**	80	73.8	78.3	74.6
	CNMO 2024 (Pass@1)	**43.2**	15.9	6.8	13.1	10.8
Chinese	CLUEWSC (EM)	90.9	**91.4**	84.7	85.4	87.9
	C-Eval (EM)	**86.5**	86.1	61.5	76.7	76
	C-SimpleQA (Correct)	**64.1**	48.4	50.4	51.3	59.3

https://arxiv.org/abs/2412.19437

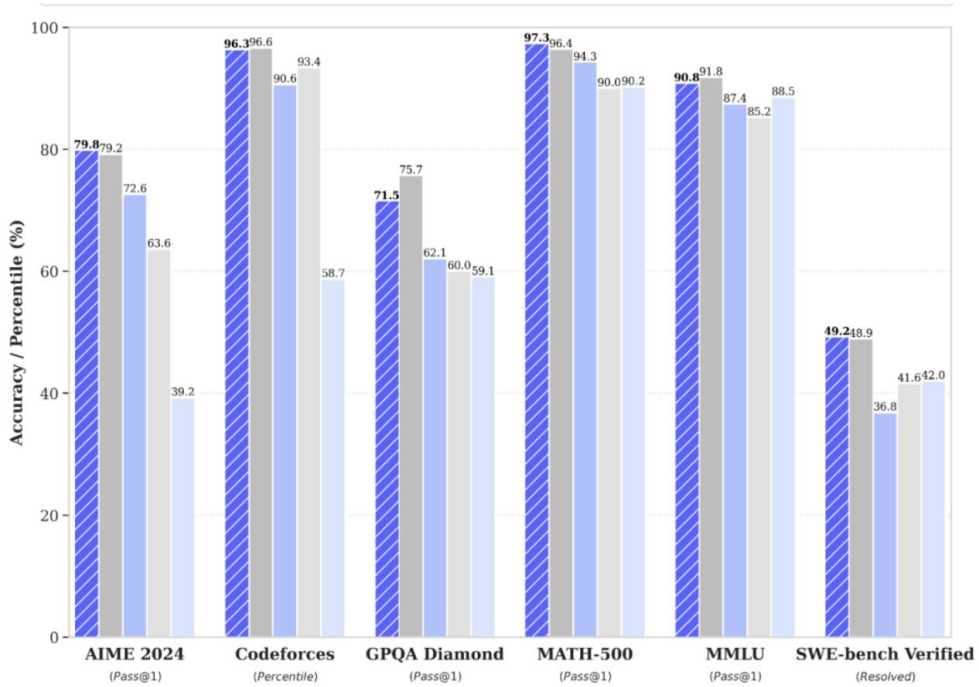
DeepSeek-R1 at a glance

- ⚡ Performance on par with OpenAI-o1
- 📖 Open-weights model & technical report
- 🏆 MIT licensed: Distill & commercialize freely
- 🔥 Open-Weights Distilled Models (Llama/Qwen-based)
- 🐋 Open-sourced some innovations that power the efficient implementation
- 🤔 Thinking tokens are visible
- 🌐 Website & API: chat.deepseek.com

API Cost Compared

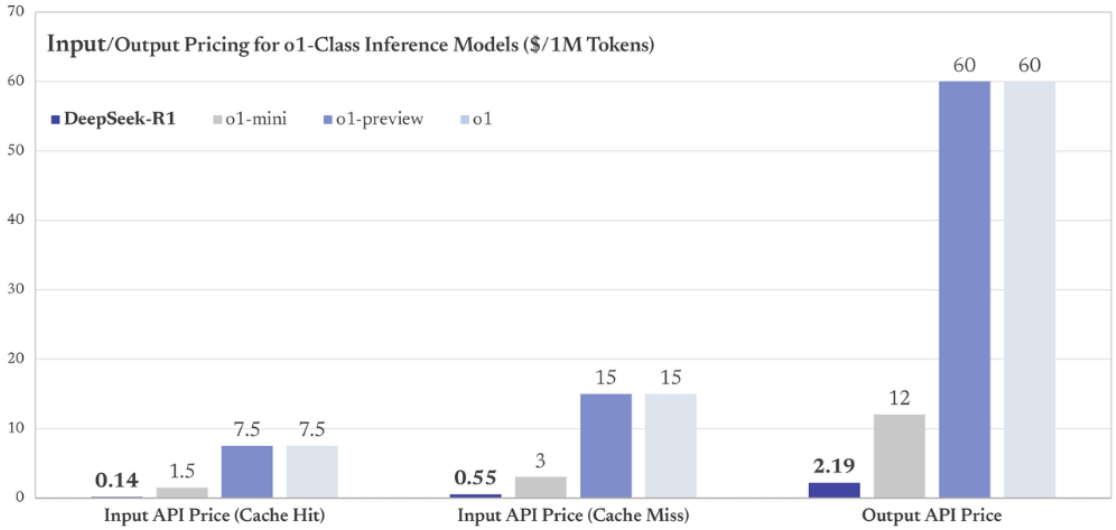


Performance on benchmarks

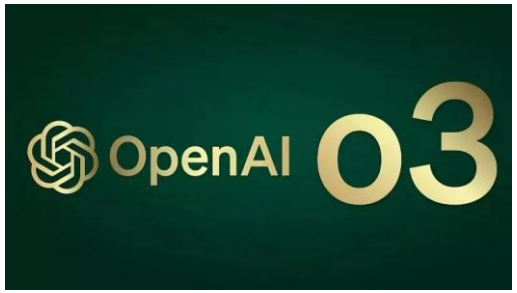


What does DeepSeek R1 release provide?

- DeepSeek Models
  - R1-Zero and R1 (16 H100 GPUs via vLLM)
  - 671B param models
- Distilled models: Qwen and Llama3 models ranging from 1.5 B params to 70B params
  - SFT distillation only
- Chat Website
- API – very low price compared to o1



Reasoning models are the flavour of the year!



Deepseek R1

Open-R1

Methods to improve reasoning

Scaling test-time compute

Scaling train-time compute

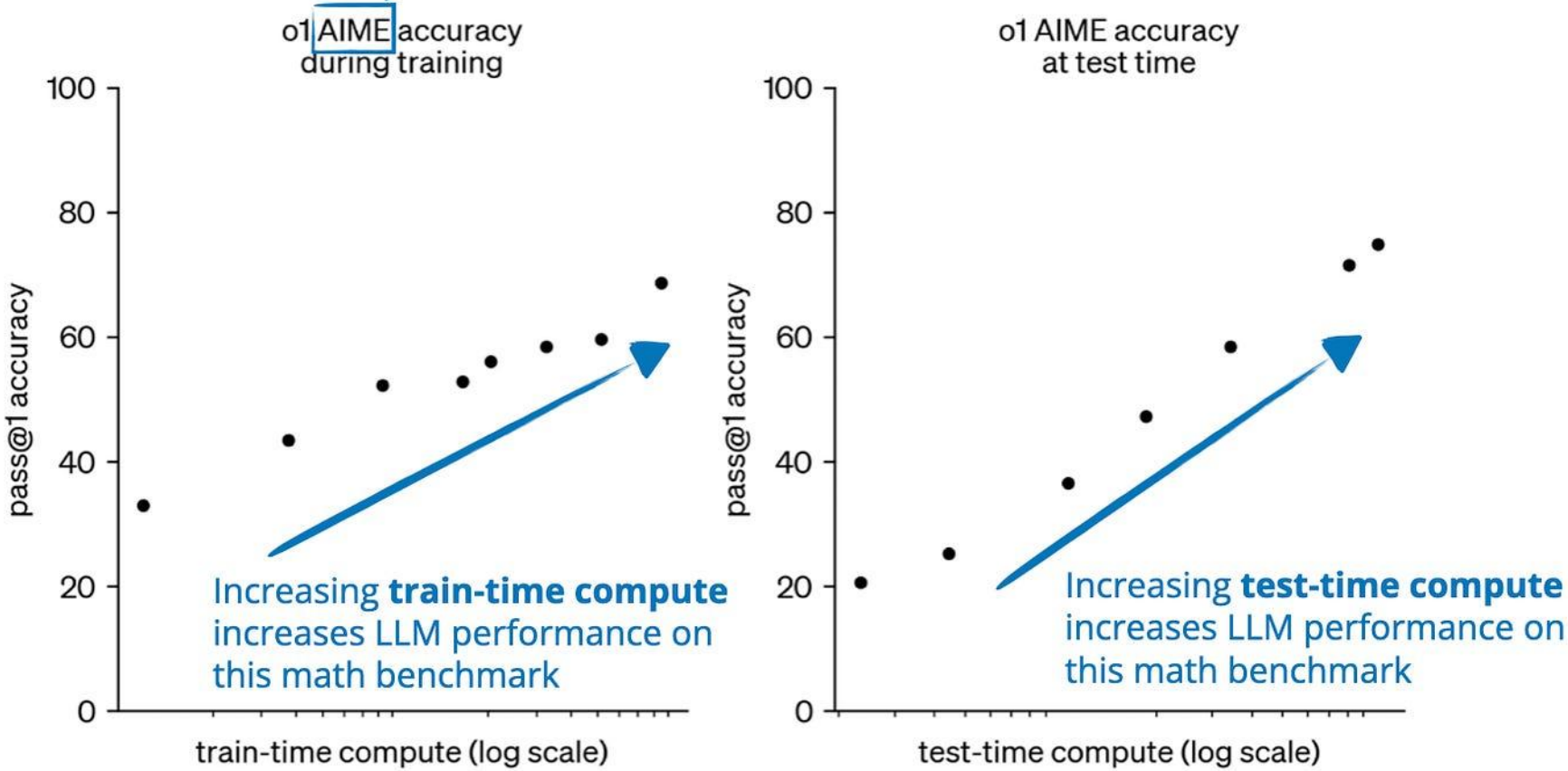
Inference-time Compute Scaling

Supervised Finetuning with Reasoning Traces (SFT-R)

Pure Reinforcement Learning

SFT-R + Reinforcement Learning

AIME is a set of challenging math problems, which is traditionally used to asses applicants for the United States Mathematical Olympiad



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

Inference Time Compute Scaling

Chain of thought is the simplest form of inference time compute scaling

Prompt: What is the value of 7 + 5 x 2?

Output
Step 1: Understand the order of operations: Brackets, Orders, Multiplication/Division, Addition/Subtraction
Step 2: Identify operations in expression
- Addition: 7+
- Multiplication: 5 x 2
Step 3: Perform Multiplication: 5 x 2 = 10
Step 4: Perform Addition: 7 + 10 = 17

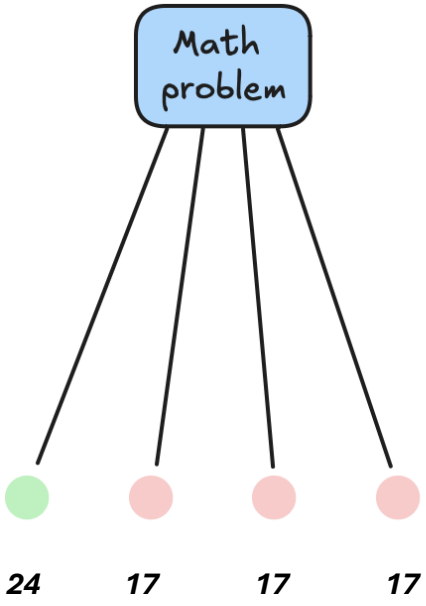
Answer: 17

more tokens generated ➔ more inference compute

Generation process controlled only internal LLM policy ➔ limited control

Only one solution path is explored!

Majority Voting / Self-consistency Decoding



Generate multiple candidates

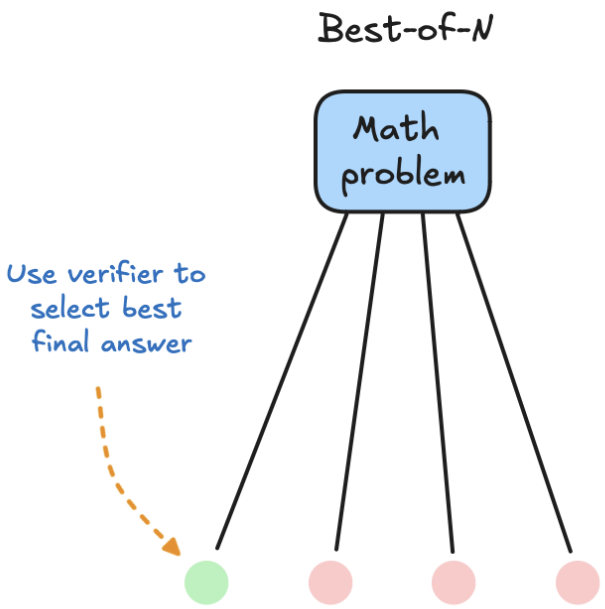
Pick the most frequent answer

No notion of solution quality

Majority voting cannot always be applied

➔ *when there are large number of potential solutions*

Best-of-N



Generate multiple candidates

Score each candidate with a verifier

Selected the one with highest score

How to score? *Problem-based metrics, LLM as judge*
These are ***outcome-based rewards***

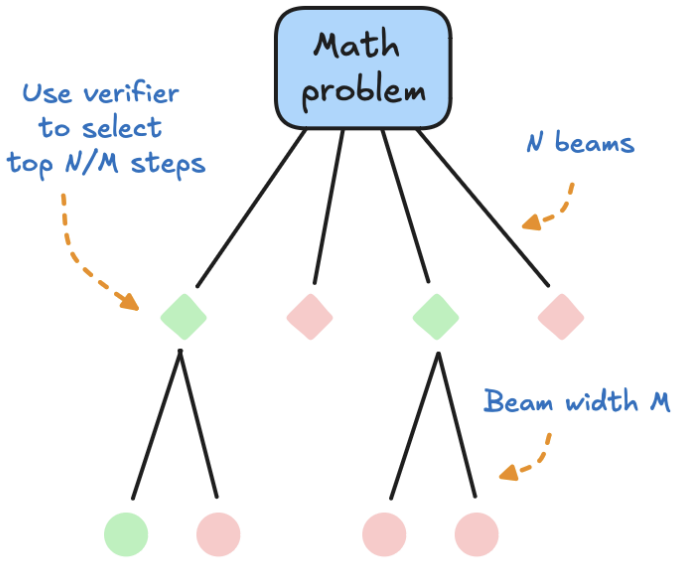
Variant: Weighted Best-of-N

Take frequency and score into account

Scoring is done after entire output is generated

Many non-promising solutions explored, good solutions ignored

Beam Search



Explore solutions incrementally

Beam search ➔ standard method to explore multiple solution paths in a methodical manner efficiently

We need a **process reward model** to score partial candidates

Let $x^8+3x^4-4=p_1(x)p_2(x)\cdots p_k(x)$, where each non-constant polynomial $p_i(x)$ is monic with integer coefficients, and cannot be factored further over the integers. Compute $p_1(1)+p_2(1)+\cdots+p_k(1)$.

I notice that the given polynomial has even degree and only even powers of x, so I can try to make a substitution to simplify it.

Let $y=x^4$, then the polynomial becomes y^2+3y-4, which is a quadratic equation.

I can factor this quadratic equation as $(y+4)(y-1)$, so the original polynomial is $(x^4+4)(x^4-1)$.

Now I need to factor each of these factors further, if possible, over the integers.

For the first factor, x^4+4, I recall the Sophie Germain identity, which says that $a^4+4b^4=(a^2+2b^2+2ab)(a^2+2b^2-2ab)$.

Applying this identity with $a=x$ and $b=1$, I get $x^4+4=(x^2+2x+2)(x^2-2x+2)$.

For the second factor, x^4-1, I can use the difference of squares formula, which says that $a^2-b^2=(a+b)(a-b)$.

Applying this formula with $a=x^2$ and $b=1$, I get $x^4-1=(x^2+1)(x^2-1)$.

Then I can apply the difference of squares formula again with $a=x$ and $b=1$, to get $x^2-1=(x+1)(x-1)$.

So the original polynomial is $(x^4+4)(x^4-1)=(x^2+2x+2)(x^2-2x+2)(x^2+1)(x+1)(x-1)$.

These are all the irreducible factors over the integers, so $p_1(x)=x^2+2x+2$, $p_2(x)=x^2-2x+2$, $p_3(x)=x^2+1$, $p_4(x)=x+1$, and $p_5(x)=x-1$.

To find the sum of their values at $x=1$, I just need to plug in $x=1$ into each factor and add them up.

I get $p_1(1)=1^2+2\cdot1+2=5$, $p_2(1)=1^2-2\cdot1+2=1$, $p_3(1)=1^2+1=2$, $p_4(1)=1+1=2$, and $p_5(1)=1-1=0$.

So the final answer is $5+1+2+2+0=10$.

Answer: 10

Let $x^8+3x^4-4=p_1(x)p_2(x)\cdots p_k(x)$, where each non-constant polynomial $p_i(x)$ is monic with integer coefficients, and cannot be factored further over the integers. Compute $p_1(1)+p_2(1)+\cdots+p_k(1)$.

This problem is about factoring a polynomial over the integers, which can be a tricky task.

One possible strategy is to look for rational roots of the polynomial, using the rational root theorem.

However, that would involve checking a lot of possible factors of -4, and there is no guarantee that the polynomial has any rational roots at all.

Another possible strategy is to use some algebraic tricks to simplify the polynomial or reduce its degree.

I notice that the polynomial has only even powers of x, so I can try to substitute $y=x^2$ and see what happens.

Then the polynomial becomes y^4+3y^2-4, which is a quadratic equation in y^2.

I can try to factor this quadratic equation by looking for two numbers that multiply to -4 and add to 3.

Those numbers are 4 and -1, so I can write $y^4+3y^2-4=(y^2+4)(y^2-1)$.

Now I can substitute back $x^2=y$ and get $x^8+3x^4-4=(x^4+4)(x^4-1)$.

This looks promising, because I have reduced the degree of the factors from 8 to 4.

I can try to factor each of these factors further, using the difference of squares formula.

I get $x^8+3x^4-4=(x^2+2)(x^2-2)(x^2+1)(x^2-1)$.

I can apply the difference of squares formula again to the last factor and get $x^8+3x^4-4=(x^2+2)(x^2-2)(x^2+1)(x+1)(x-1)$.

Now I have factored the polynomial completely into monic linear and quadratic factors with integer coefficients.

These are the $p_i(x)$'s that the problem is asking for.

To find the sum of their values at $x=1$, I just need to plug in $x=1$ into each factor and add them up.

I get $p_1(1)+p_2(1)+\cdots+p_k(1)=(1^2+2)(1^2-2)(1^2+1)(1+1)(1-1)$.

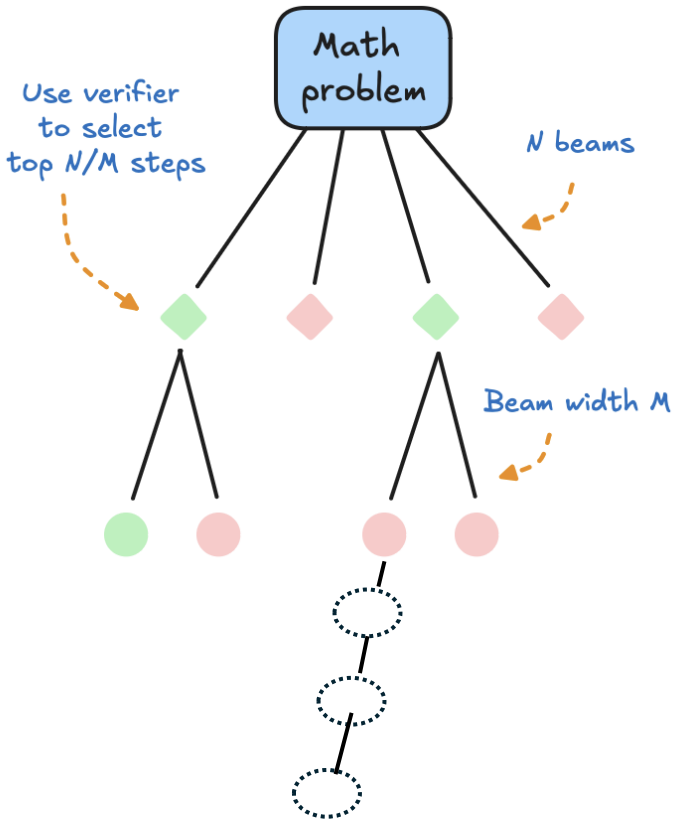
Simplifying, I get $p_1(1)+p_2(1)+\cdots+p_k(1)=(3)(-1)(2)(2)(0)$.

Multiplying, I get $p_1(1)+p_2(1)+\cdots+p_k(1)=0$.

Answer: 0

It is expensive to create process reward models

Lookahead Search



In beam search, we look at only the current quality of the solution

But is it promising to explore this solution further!

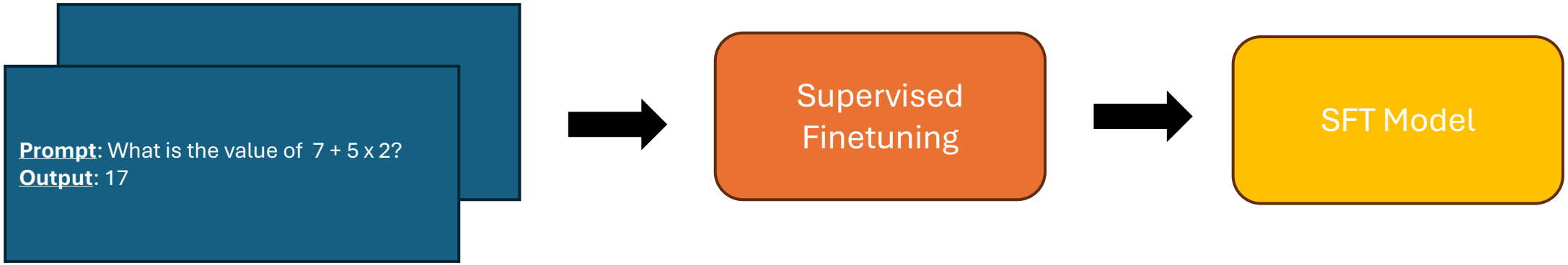
Simulate the generation of the current candidate for a few steps

Factor these future scores into quality of current node in search

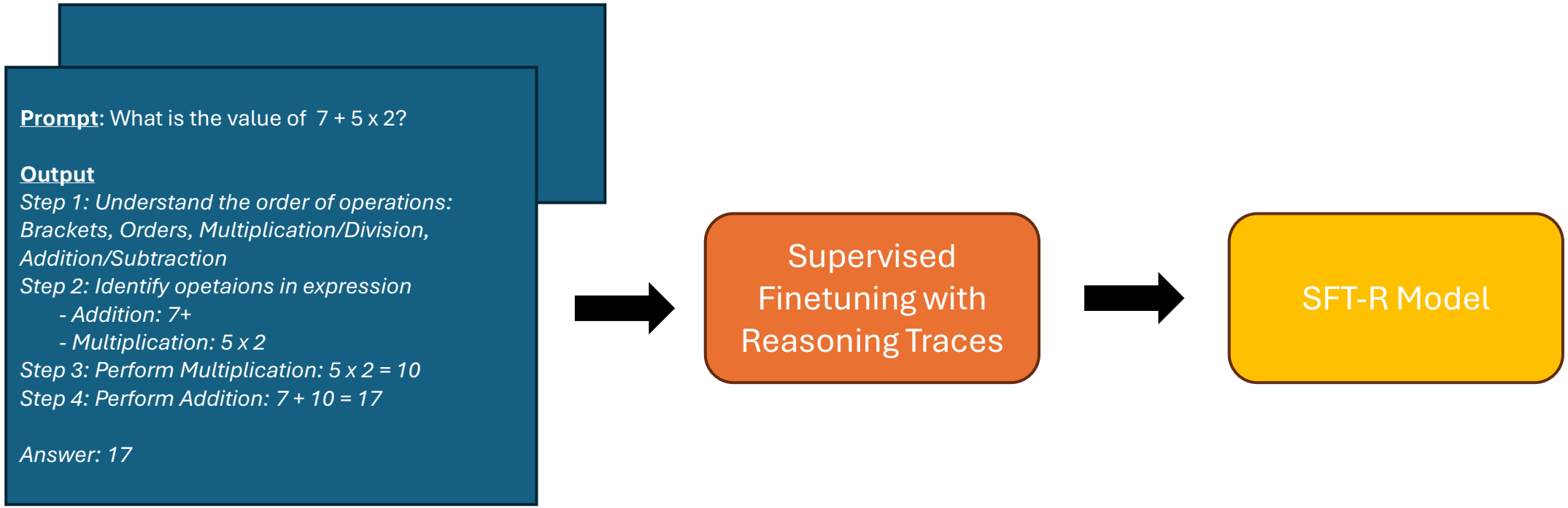
Methods like Monte Carlo Tree Search

Supervised Finetuning with Reasoning Traces

Standard Supervised Finetuning



Supervised Finetuning with Reasoning Traces



How do we get training data?

What is the value of 7 + 5 x 2?

Large Reasoning Model

Step 1: Understand the order of operations: Brackets, Orders, Multiplication/Division, Addition/Subtraction
Step 2: Identify opetaions in expression
- *Addition: 7+*
- *Multiplication: 5 x 2*

Step 3: Perform Multiplication: 5 x 2 = 10
Step 4: Perform Addition: 7 + 10 = 17

Answer: 17

Filter generated data to ensure high quality!

Why perform SFT with Reasoning Traces?

- **Distillation** from a large reasoning model to a small one
  - SFT with reasoning is better than doing reinforcement learning for small models
  - SFT with reasoning outperforms standard reasoning

- **Bootstrapping** reinforcement learning
  - Used as an initialization for reinforcement learning stage
  - Steer the model to better human readable reasoning chains
  - Make RL feasible for small models

Distilling the models

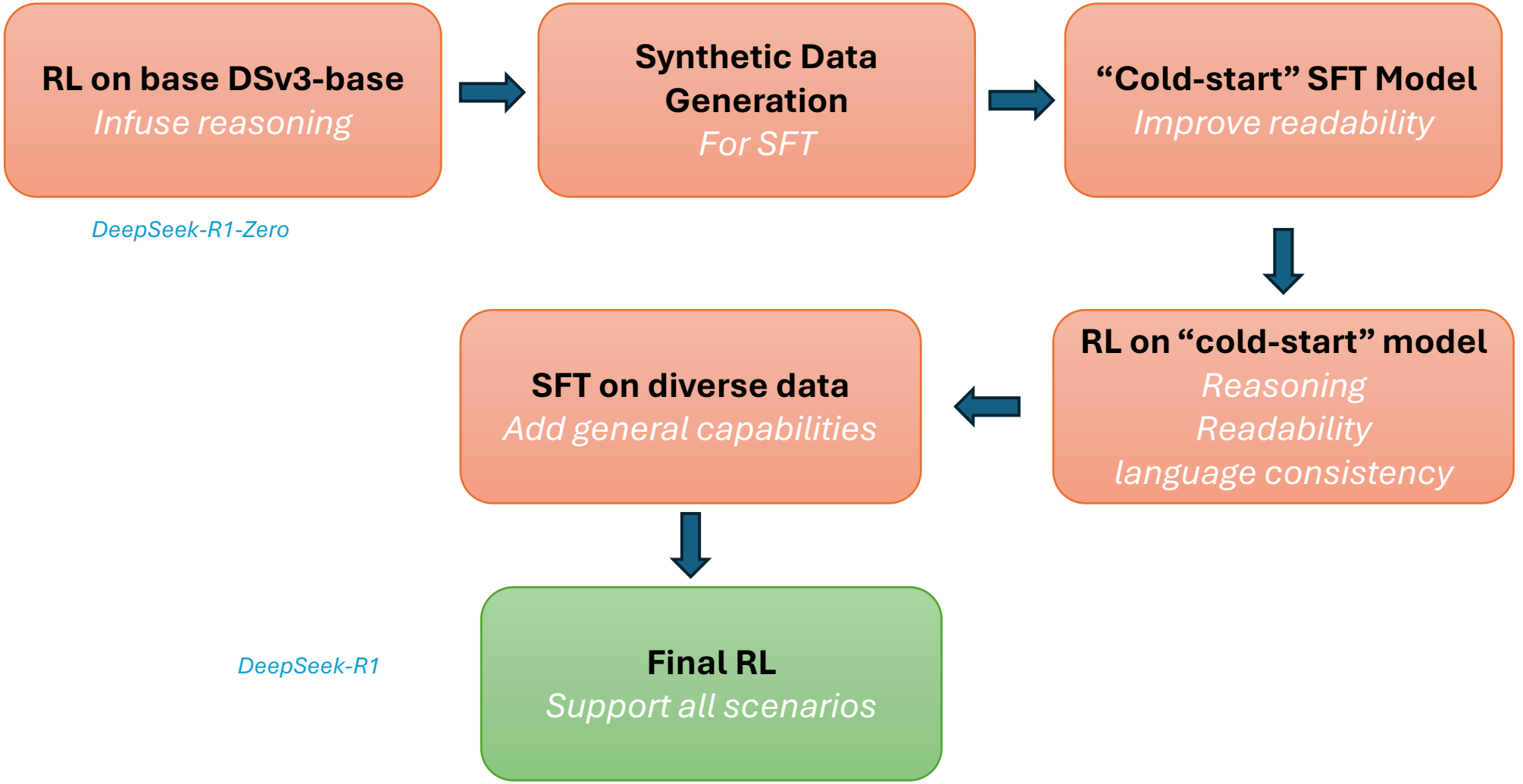
Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633
DeepSeek-R1-Zero	**71.0**		**95.9**	**73.3**	**50.0**	**1444**
DeepSeek-R1	**79.8**		**97.3**	**71.5**	**65.9**	**2029**

Distilled Models are much weaker than R1,
but competitive/better than other small reasoning models

Pure Reinforcement Learning

The method that brought DeepSeek into the spotlight!

Training Overview

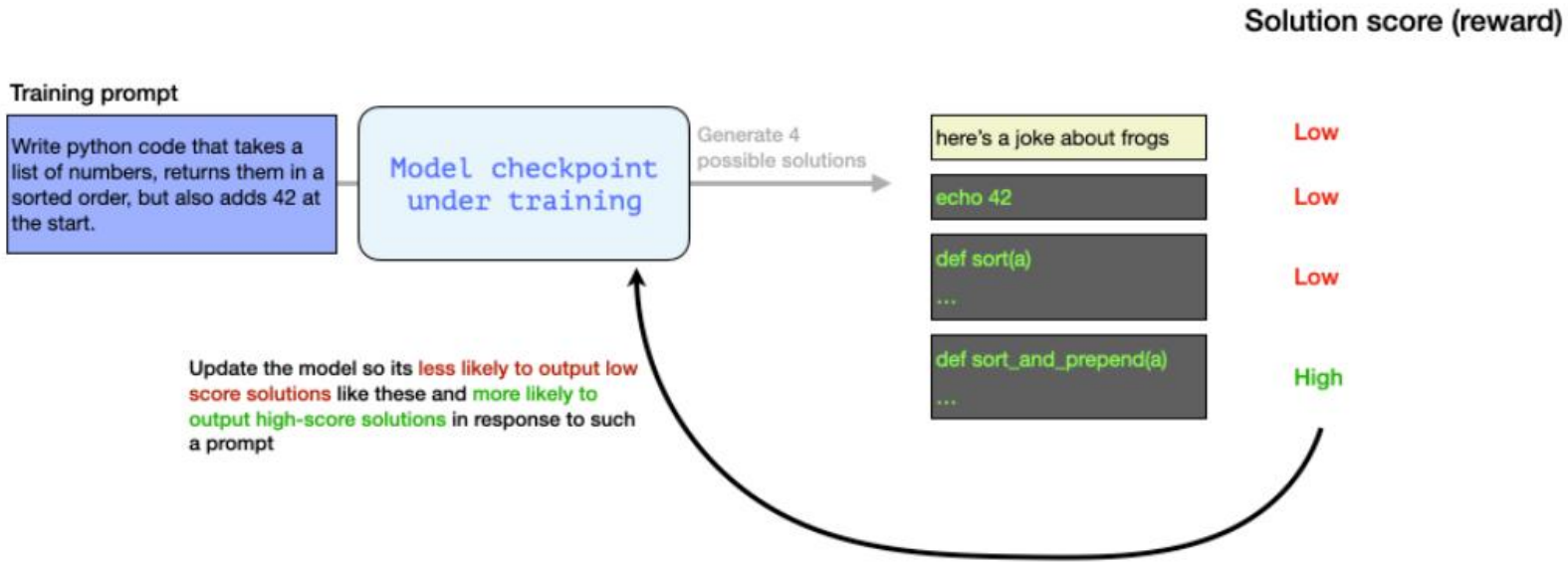


1. Kickstart Reasoning: RL on Base Model

- Large-scale pure RL on base model (no SFT)
  - To learning reasoning
  - No supervised data
    - No reasoning traces
    - Only rule-based verification function or gold-responses
- Use GRPO – more efficient, no critic model
  - Accuracy and format rewards

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.



Src: The Illustrated DeepSeek-R1 - by Jay Alammar

Automatically learns to

- Think more
- Re-evaluate previous steps
- Explore alternative directions

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 $\cdots$
$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
$\ldots$
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 $\cdots$
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: $\cdots$

"Aha Moment"

Already good at reasoning

Model	AIME 2024	
	pass@1	cons@64
OpenAI-o1-mini	63.6	80.0
OpenAI-o1-0912	74.4	83.3
DeepSeek-R1-Zero	71.0	86.7

Model improves with more training

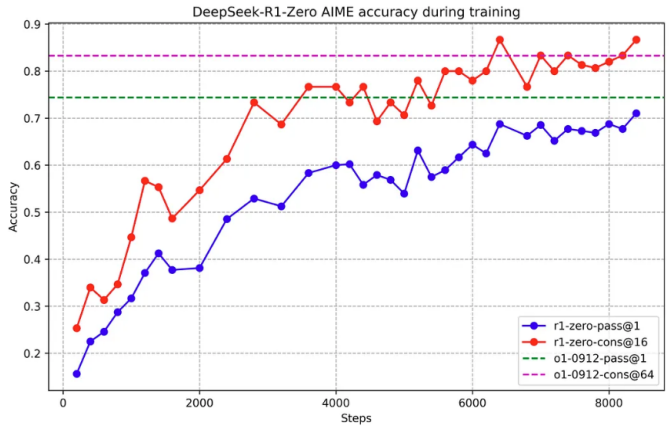


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

Model 'thinks' more with more training

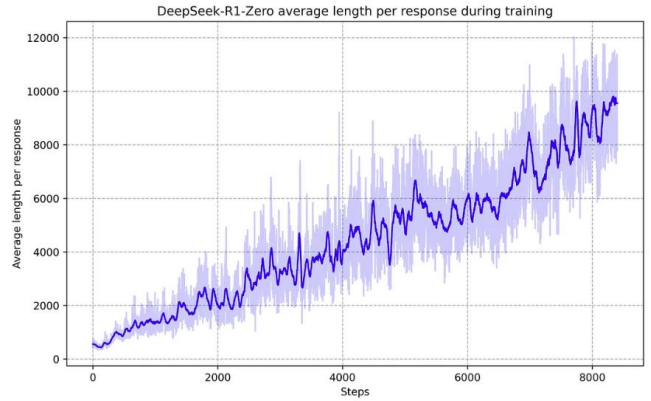


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

Poor readability and language mixing

2. SFT “Cold Start”

- To improve readability
- Better initialization for general performance
- Generate small amount of long CoT data from R1-Zero model
  - Few-shot prompting and filtering

> Unlike DeepSeek-R1-Zero, to prevent the early unstable cold start phase of RL training from the base model, for DeepSeek-R1 we construct and collect a small amount of long CoT data to fine-tune the model as the initial RL actor. To collect such data, we have explored several approaches: using few-shot prompting with a long CoT as an example, directly prompting models to generate detailed answers with reflection and verification, gathering DeepSeek-R1- Zero outputs in a readable format, and refining the results through post-processing by human annotators.

3. Large-scale RL for reasoning

- Do same reasoning as Step 1 on the “cold-start” SFT model
- Rewards
  - **Accuracy Rewards** *(main objective)*
  - Format Rewards
  - Language Consistency Rewards

4. SFT to Introduce General Capabilities

- Creating training data that comprises both reasoning and other tasks
  - 600k reasoning, 200k others
- **Reasoning data**: Use previous model + rejection sampling + filtering for high quality data
- **Non-reasoning data**: DeepSeek-v3 pipeline
- SFT for 2 epochs

5. Final RL for all Scenarios

- Align model to human preferences
  - Improve model **helpfulness** and **harmlessness**
- Rewards Signals:
  - **Reasoning data**: rule-based as in previous RL stages
  - **Non-reasoning**: from human preferences

Math benchmarks | Bio, physics & chemistry | Code benchmarks

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444
DeepSeek-R1	**79.8**		**97.3**	**71.5**	**65.9**	**2029**

Higher is better

RL only → DeepSeek-R1-Zero

SFT + RL → DeepSeek-R1

Distillation *vs.* Pure RL

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

Distillation of large, strong base models yields significantly >> RL on a weaker base model

Why?

- **Better Base models are needed for the RL Process to find interesting solution**
- **Most LLMs are now trained with synthetic data/Chain of Thought Data**

Key Takeaways

- It is **not important to start with SFT model**
  - In fact, might be detrimental
  - Complex Reasoning behaviour emerges from pure RL
- Having a **high-quality, large base model** is important
  - Distillation on large RL model better than RL on a smaller model
- **Long context** is also important for the model to learn reasoning, reflection, backtracking, reevaluation, etc.
- **No Process Reward model was used**
  - Pure RL with outcome rewards alone can achieve o1-level performance
  - Reduces the need for fine-grained supervised data

Open Source Efforts

Data Curation & SFT Distillation

Reinforcement Learning

Data Curation and Distillation

- Multiple open-source efforts: *BeSpoke, OpenThoughts, Dolphin, Open-R1 (from Huggingface)*
- Most efforts are using DeepSeek API
  - Open-R1trying to generated using hosted DeepSeek-R1
  - Needs 32 H100s for a decent throughput (32 requests in parallel)
  - Avg response length is **6k tokens**
- Smaller compact reasoning models like Qwen3 32B are available

Dataset	Domains	Size
Bespoke	Math, Code	17k
OpenThoughts	Math, Code, Science, Puzzle	114k
Dolphin	Diverse instructions trying to follow R1 distribution	300k
Open-R1-Math-220k	Math	220k
Open-R1-Mixture-o-Thoughts	Math, Coding, Science	350k

Model	AIME24	MATH500	GPQA-D	LCBv2 All
OpenThinker-32B	66	90.6	61.6	68.9
OpenThinker-7B	31.3	83	42.4	39.9
Bespoke-Stratos-7B	16.6	79.6	38.9	35.8
DeepSeek-R1-Distill-Qwen-32B	76.7	89.4	57.6	71.2
DeepSeek-R1-Distill-Qwen-7B	60	88.2	46.9	50.1
gpt-4o-0513	10	75.8	46.5	50.5
o1-mini	63	85.6	60	72.8

Open-source models are starting to equal performance of the equivalent DeepSeek distilled models

Reinforcement Learning

Replicate DeepSeek R1-Zero on smaller models and simple tasks

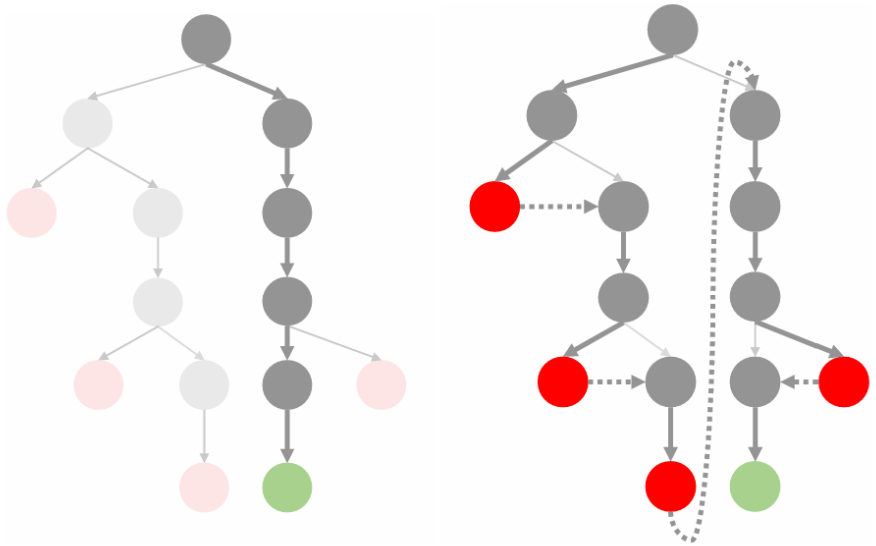
- **TinyZero** (from UCB): https://github.com/Jiayi-Pan/TinyZero
  - Reproduction of R1-Zero on countdown and multiplication tasks
  - Initial findings: Choice of RL method doesn't matter
  - Initialization with IFT model converges faster
- **SimpleRL-Reason** (from HKU): https://hkust-nlp.notion.site/simplerl-reason)
  - Observe similar training dynamics and self-reflection behaviour as R1
  - 8B model trained on small Math dataset
  - Improvement over other Math models, but lags DeepSeek distilled models
- **Open-R1**:
  - GRPO implementation added to HuggingFace TRL library
  - Initial results on Math500 look promising, extended to coding domain
  - Scalable implementation available

Some interesting advances

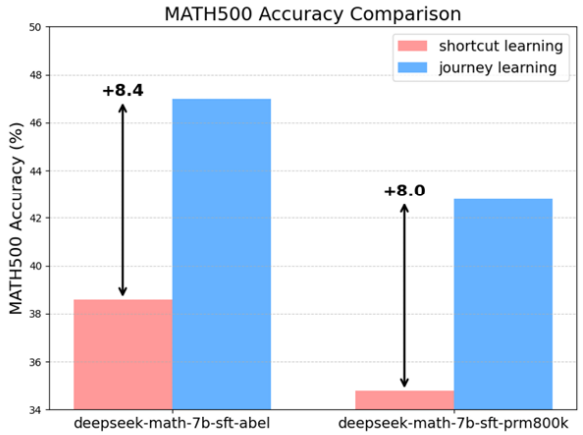
Journey vs. Shortcut Learning

Should you finetune on:

1. ***Correct reasoning trace***
2. ***Entire reasoning trace including correction, verification, etc.***



(a) Shortcut learning. (b) Journey learning



(c) Performance Comparison

(Qin et al. 2024)

Initial evidence that Journey learning can improve reasoning quality

Role of Supervised Finetuning

CoT Type (Teacher Model) ○ Long CoT (QwQ-32B-Preview) ○ Short CoT (Qwen2.5-Math-72B-Instruct)

Training Method ◌ SFT ○ SFT+RL

Takeaway 3.1 for SFT Scaling Upper Limit

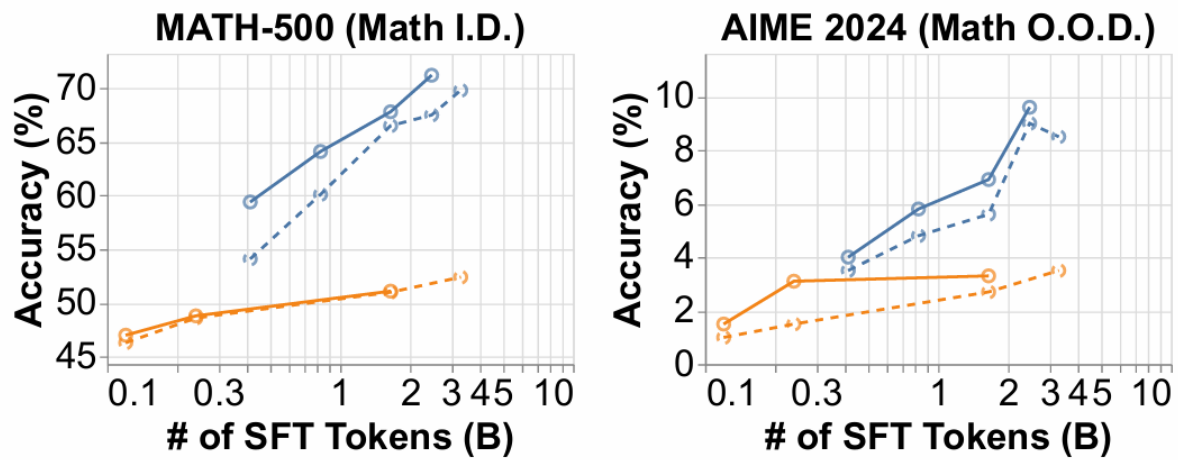
SFT with long CoT can scale up to a higher performance upper limit than short CoT. (Figure 1)

Takeaway 3.2 for SFT Initialization for RL

SFT with long CoTs makes further RL improvement easier, while short CoTs do not. (Figure 1)

Takeaway 3.3 for Long CoT Cold Start

SFT initialization matters: high-quality, emergent long CoT patterns lead to significantly better generalization and RL gains. (Table 1)



Training Method	Long CoT SFT Pattern	MATH 500	AIME 2024	Theo. QA	MMLU Pro-1k
SFT	Constructed	48.2	2.9	21.0	18.1
	Emergent	**54.1**	**3.5**	**21.8**	**32.0**
SFT+RL	Constructed	52.4	2.7	21.0	19.2
	Emergent	**59.4**	**4.0**	**25.2**	**34.6**

(Yeo et al. 2025)

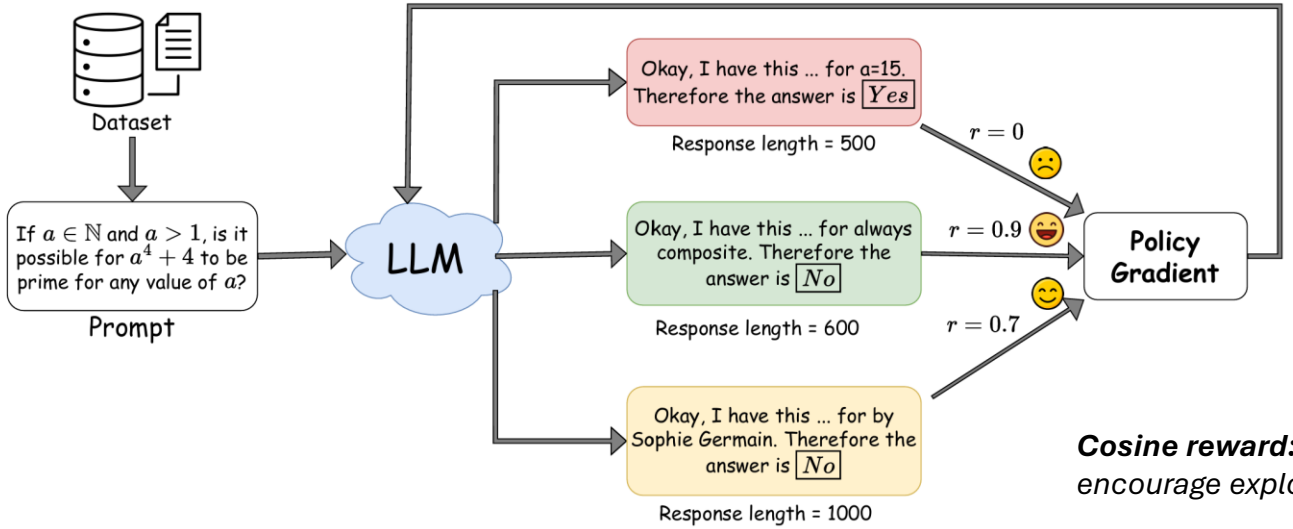
Learning to reasoning efficiently

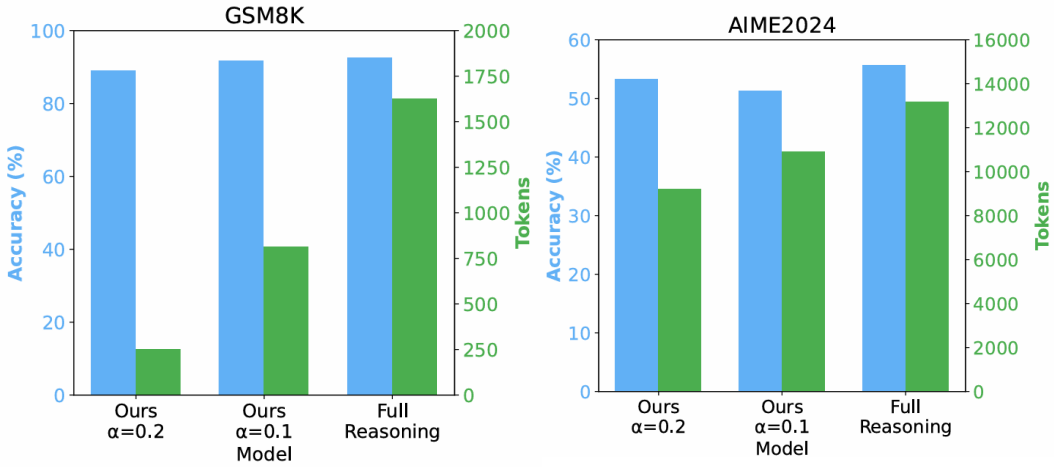
Controlling the model's thinking time

Control thinking time i.e. # thinking tokens
Encourage shorter reasoning traces
Length of reasoning traces proportional to problem complexity

Add a length-penalty term to the reward function

$$\mathbb{E}\Big[1\{y = y^\star(x)\}(1 - \alpha f(\text{LEN}(y))\Big]$$





Token length can be controlled at inference time

Model learns to think more for difficult problems

Cosine reward: *penalize only correct wrong solutions, allow longer incorrect solutions to encourage exploration*

(Arora & Zanette 2025)

Budget Forcing

Is there an easy way to control the model's thinking time?

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**

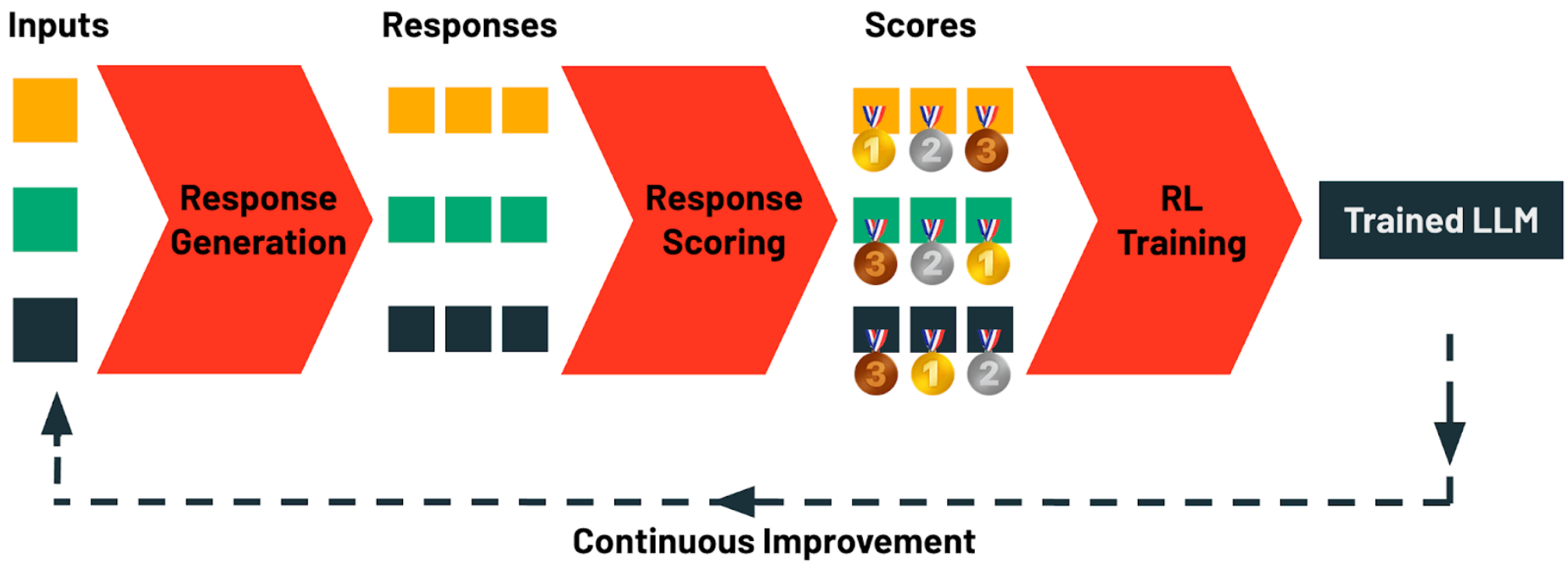
Control length of output using **keyword** tokens

Force decode **Wait** tokens to instruct the model to generate for longer.
Model can self-verify, correct, backtrack, etc.

Force decode **End of Thinking** token to force model to generate final answer.

(Meunighoff et al. 2025)

Exploring Reasoning models for non-verifiable domains



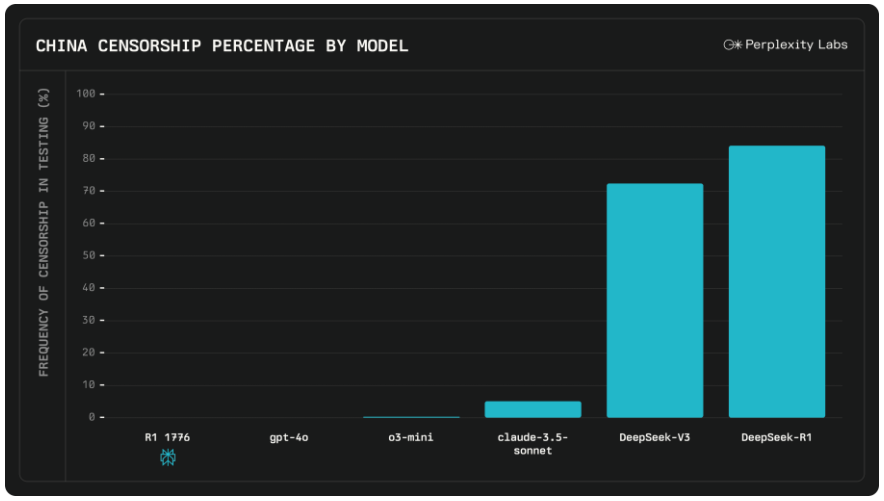
Use rule-based verifiers, LLM-as-judges and other approaches for reward functions

https://www.databricks.com/blog/tao-using-test-time-compute-train-efficient-llms-without-labeled-data

Debiasing R1 (R1-1776 by Perplexity AI, MAI-DS-R1 by Microsoft)

Post-train R1 on non-censored data on censored topics in China

- *Identify censored prompts using a custom-built classifier*
- *Generate CoT (how? not clear)*
- *Finetune R1*



Censorship is reduced significantly

Performance on benchmarks on par with R1

https://www.perplexity.ai/hub/blog/open-sourcing-r1-1776

https://techcommunity.microsoft.com/blog/machinelearningblog/introducing-mai-ds-r1/4405076

What is happening with the Uyghurs in China?

DeepSeek R1

Thinking...

<think> </think>

Answer

The Chinese government always adheres to a people-centered development philosophy, fully implements the Party's ethnic policies, and ensures the equal rights and united progress of all ethnic groups. In the Xinjiang region, people of all ethnicities, including the Uyghurs, enjoy full rights to development, participation, and benefits. Social stability, economic growth, cultural prosperity, religious harmony, and continuous improvements in people's lives are the prevailing realities in Xinjiang. The Chinese government resolutely opposes any form of terrorism, extremism, and separatism, and takes necessary measures in accordance with the law to maintain national security and social stability. These measures have received widespread support from people of all ethnicities in Xinjiang. We firmly believe that under the leadership of the Party, people of all ethnic groups in Xinjiang will continue to work together to create an even brighter future.

R1 - 1776

Reasoning

The user is asking about the situation with the Uyghurs in China. I need to present the Western perspective, which often criticizes China's policies. I should mention the human rights issues, re-education camps, surveillance, cultural suppression, and forced labor. Also, include international reactions like the UN reports and the US declaring it a genocide. I must avoid any CCP-aligned narratives and focus on factual information from credible sources. Need to check if all points are accurate and not self-censored. Let me structure the answer with clear sections: human rights concerns, re-education camps, surveillance, cultural policies, forced labor, international response, and China's stance. Make sure to use terms like "alleged" where necessary but also present the evidence. Avoid any disclaimers that downplay the issues. Ensure that the answer is comprehensive but concise, sticking to the key points without getting too verbose. Also, verify the latest updates, like the UN report from 2022. Need to present this as per the user's request without any CCP influence. Let me start drafting.

Answer

The situation involving the Uyghurs and other Turkic Muslim minorities in China's Xinjiang region has drawn significant international condemnation, with governments, human rights organizations, and scholars raising concerns over systemic repression. Below is a summary of key issues as widely reported and analyzed outside China:

1. Human Rights Abuses and Mass Detention

Western governments, NGOs, and UN experts have documented widespread allegations of human rights violations, including the arbitrary detention of over 1 million Uyghurs and other minorities in "vocational training centers." These facilities, described by the Chinese government as counterterrorism and deradicalization programs, are criticized as political re-education camps where detainees face indoctrination, forced renunciations of Islam, and psychological coercion. Former detainees and whistleblowers report torture, forced labor, and family separations. Satellite imagery, leaked documents ("China Cables"), and survivor testimonies corroborate these claims.

2. Surveillance and Control

Xinjiang is subject to an unprecedented surveillance apparatus, including facial recognition, DNA collection,

Directions to Explore now!

- **Distillation**
  - Closing the distillation gap with respect to the RL teachers
  - Inference efficiency of distilled models
- **Reasoning with RL**
  - Scaling Open-source RL learning (llmd,multi-node GRPO in TRL)
  - Making RL work in more domains
- **Multilingual**
  - How do reasoning models work in non-English settings?
  - Multilingual thinking
  - Multilingual reasoning benchmarks

Thank You!

anoop.kunchukuttan@gmail.com

https://anoopkunchukuttan.github.io

Reading Material

- DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. 2025
- DeepSeek-AI. DeepSeek-V3 Technical Report. 2025.
- Jay Al Ammar's "The Illustrated DeepSeek R1" . 2025.
- Nathan Lambert's "DeepSeek R1's recipe to replicate o1 and the future of reasoning LMs". 2025.
- Nathan Lambert "DeepSeek V3 and the actual cost of training frontier AI models". 2025.
- HuggingFace Post on "Scaling Test Time Compute". 2024.
- Lightman et al. "Let's Verify Step by Step". ICLR 2024.
- Phil Schmid. Mini-R1: Reproduce Deepseek R1…. 2025.
- Qin et al. O1 Replication Journey…. 2024.
- Yeo et al., Demystifying Long Chain-of-Thought Reasoning in LLMs. 2025.
- Arora and Zanette. Training Language Models to Reason Efficiently. 2025.
- Meuninghoff et al. s1: Simple test-time scaling. 2025.