An Introduction to Reasoning Models with DeepSeek R1

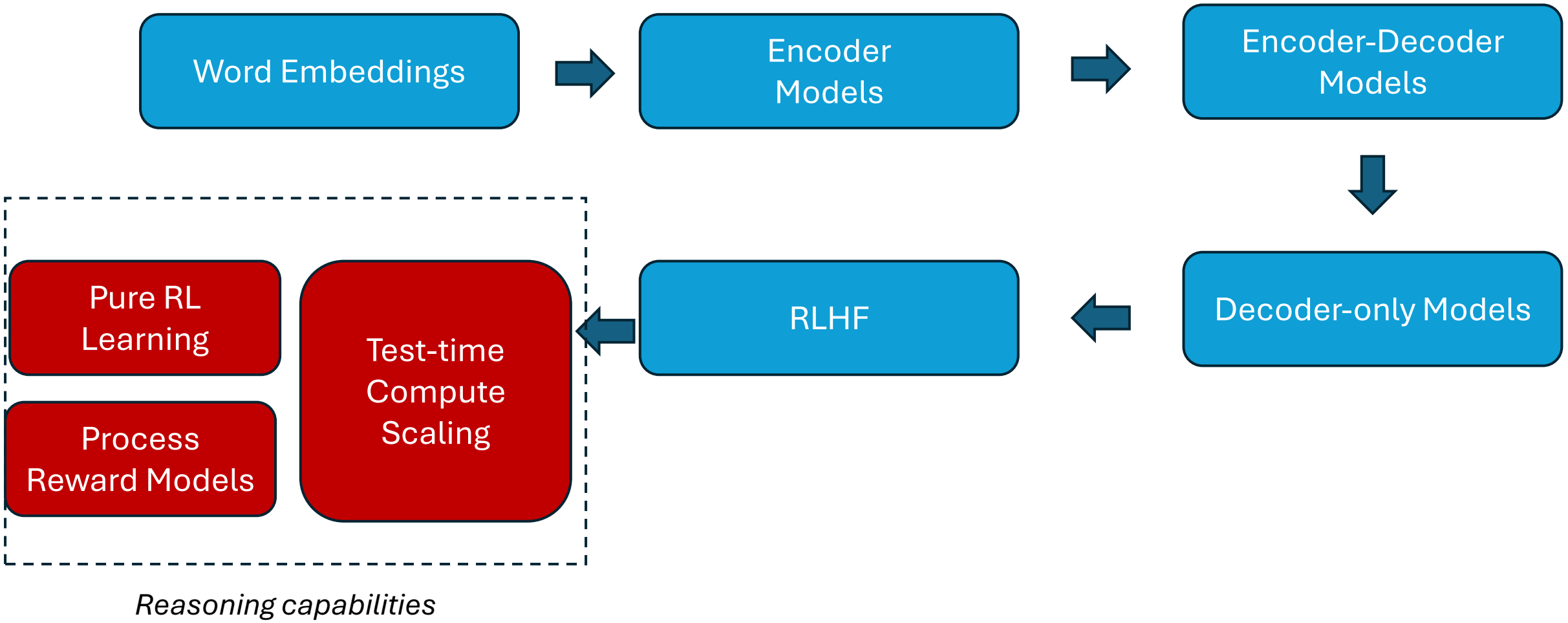
Anoop Kunchukuttan

Microsoft Translator
Azure AI Platform

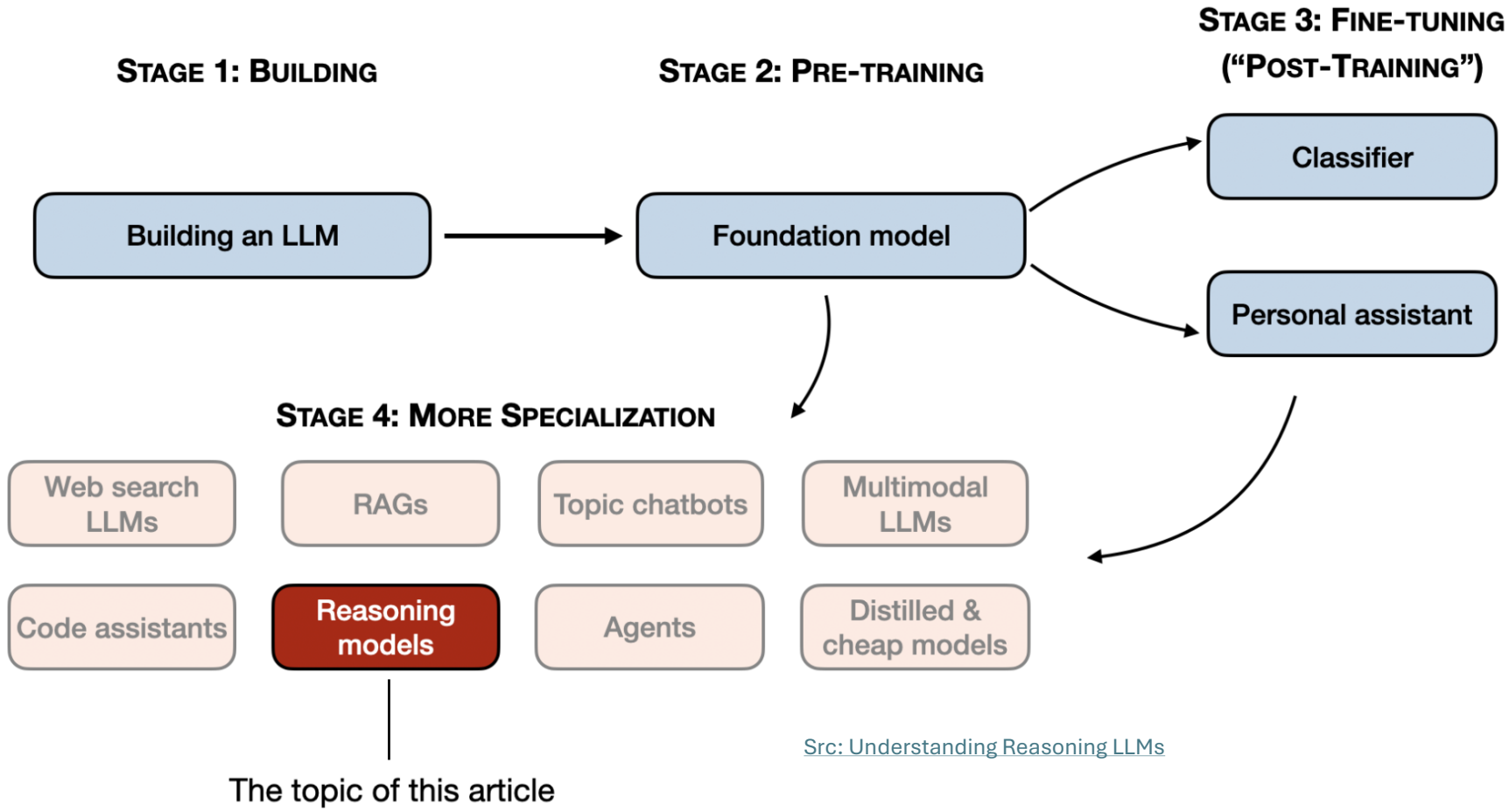
AI4Bharat

CSE Department Day
IIT Hyderabad
February 8, 2025

Language Model Evolution



LLM Specialization



Src: Understanding Reasoning LLMs

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>

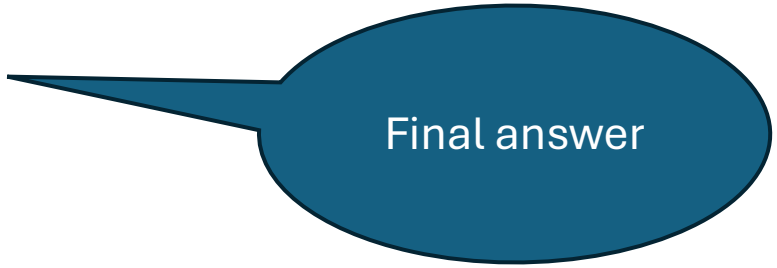
To find the distance between the points ((2, -6)) and ((-4, 3)), we use the distance formula:

[\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

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

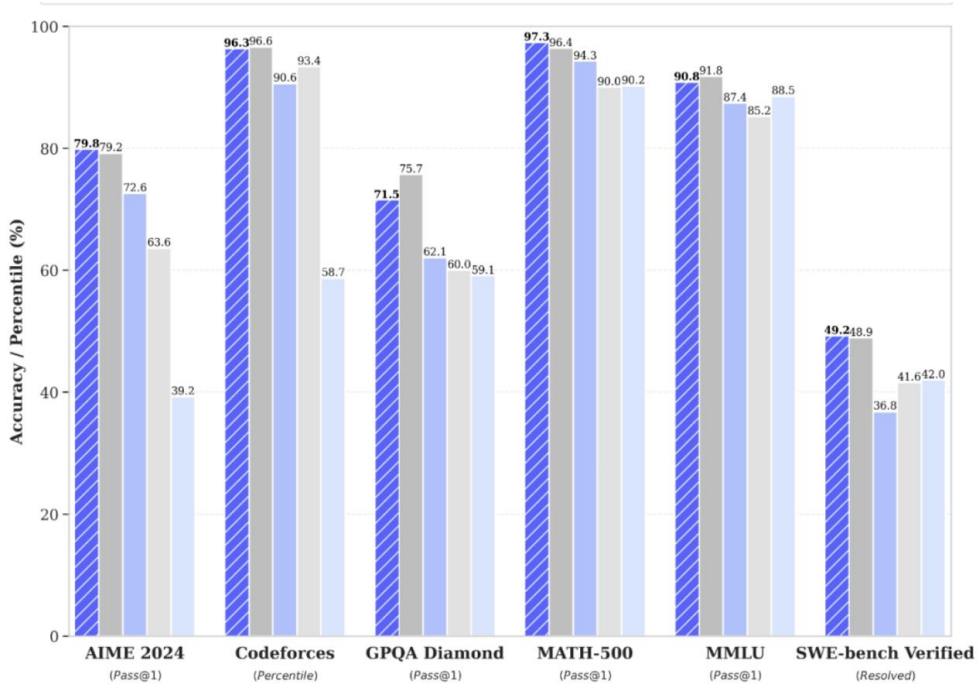
- 671B parameters, MoE, 37B active parameters
- Trained on 15T tokens
- Trained on 2048 GPUs for 2 months, $6m
- Efficiency through techniques like
  - FP8 training
  - Improved quantization
  - Multi-head latent attention
  - Aux loss free load balancing
  - MoE optimizations
  - Multi-token predictions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

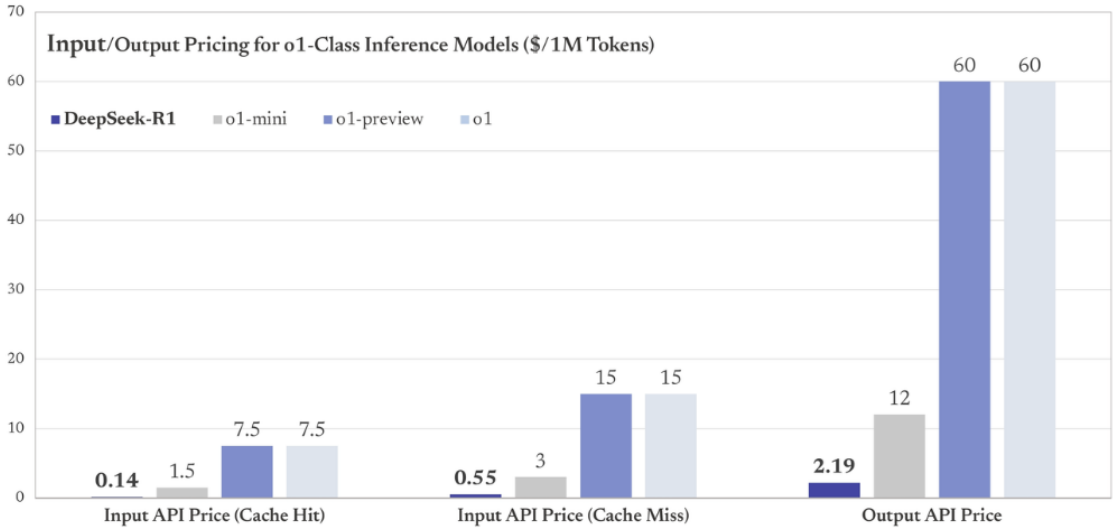
DeepSeek-R1 at a glance

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

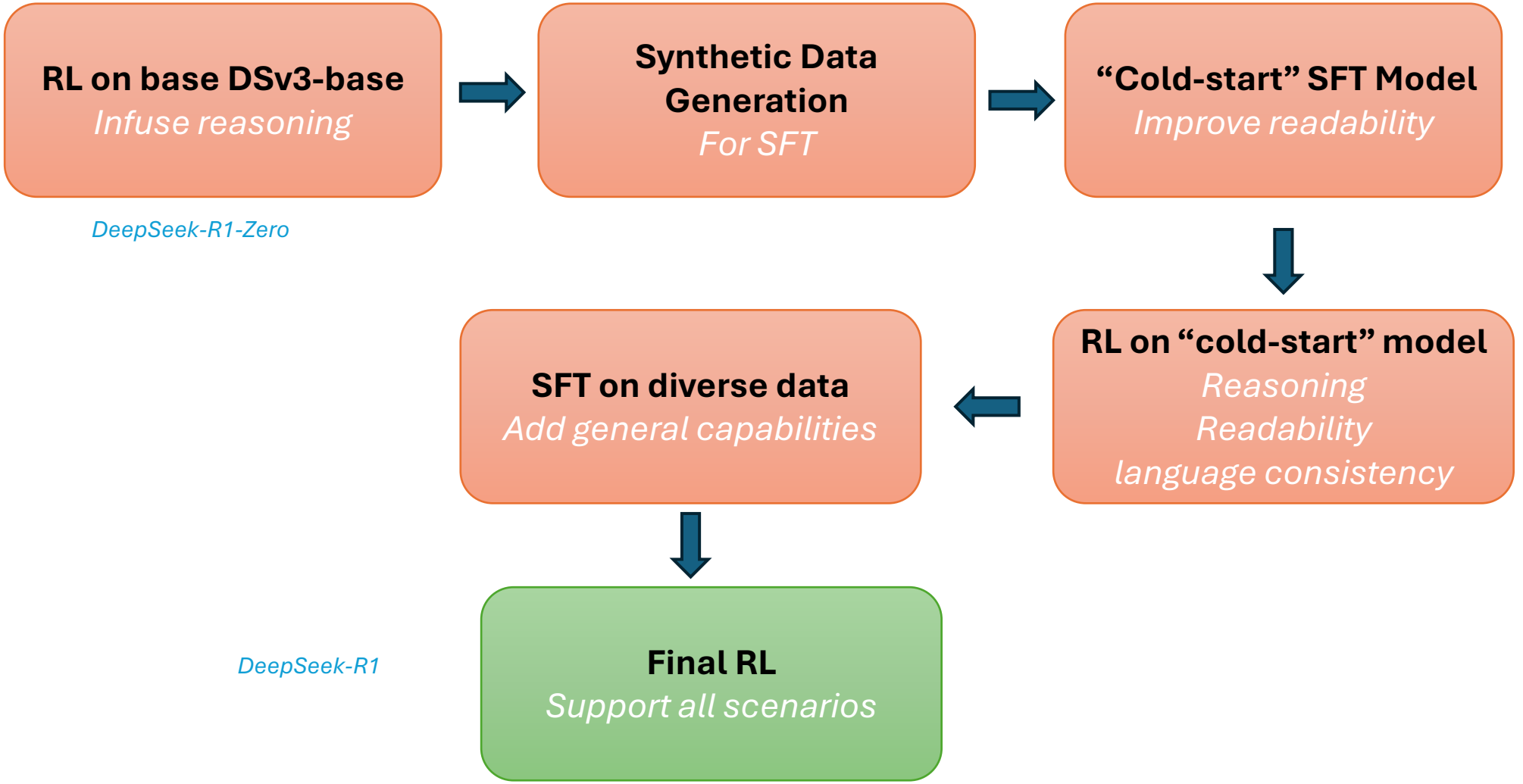


What does DeepSeek R1 release provide?

- 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



Training Overview

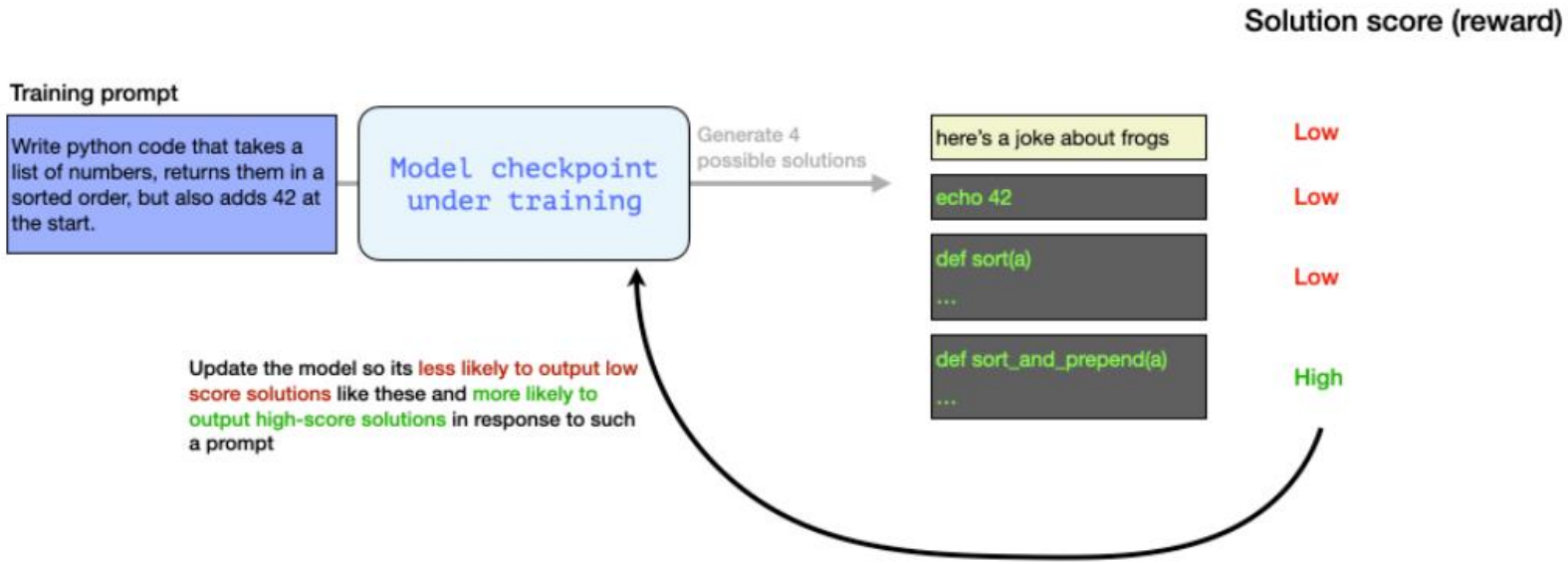


1. Kickstart Reasoning: RL on Base Model

- 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:

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

$\cdots$

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

"Aha Moment"

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: $\cdots$

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

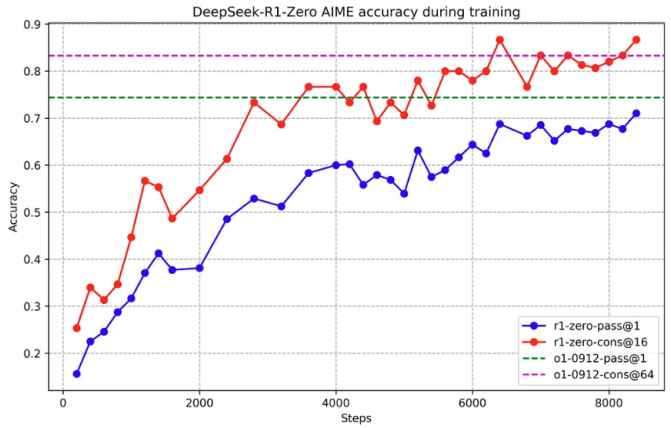


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

Model 'thinks' more with more training

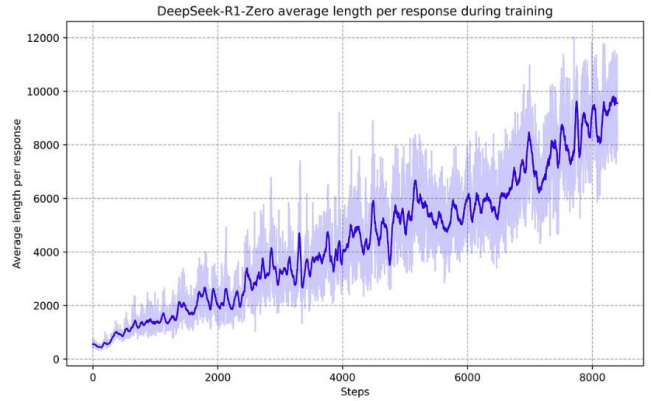


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

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

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

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 trying 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**

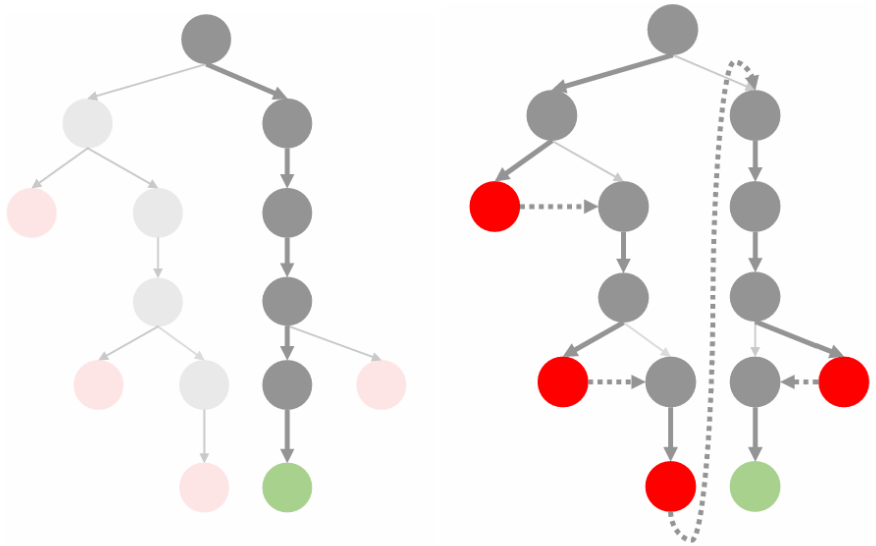
Dataset	Domains	Size
Bespoke	Math, Code	17k
OpenThoughts	Math, Code, Science, Puzzle	114k
Dolphin	Diverse instructions trying to follow R1 distribution	300k

Model	AIME24	MATH500	GPQA-Diamond	LCBv2 Easy	LCBv2 Medium	LCBv2 Hard	LCBv2 All
OpenThinker-7B	43.3	83	42.4	75.3	28.6	6.5	39.9
Bespoke-Stratos-7B	16.6	79.6	38.9	71.4	25.2	0.8	35.8
DeepSeek-R1-Distill-Qwen-7B	60	88.2	46.9	79.7	45.1	14.6	50.1
gpt-4o-0513	10	75.8	46.5	87.4	42.7	8.9	50.5
o1-mini	63	85.6	60	92.8	74.7	39.8	72.8

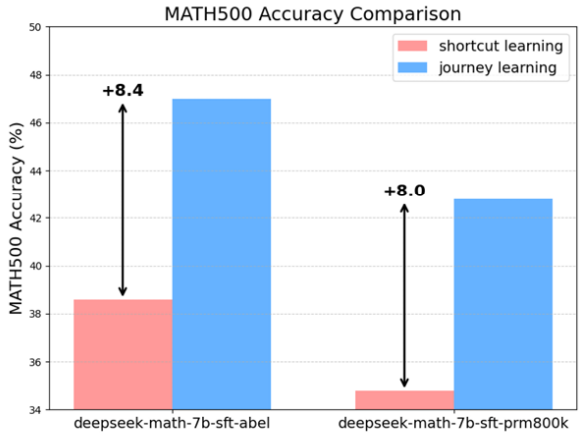
Journey vs. Shortcut Learning

Should you finetune on:

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



(a) Shortcut learning. (b) Journey learning

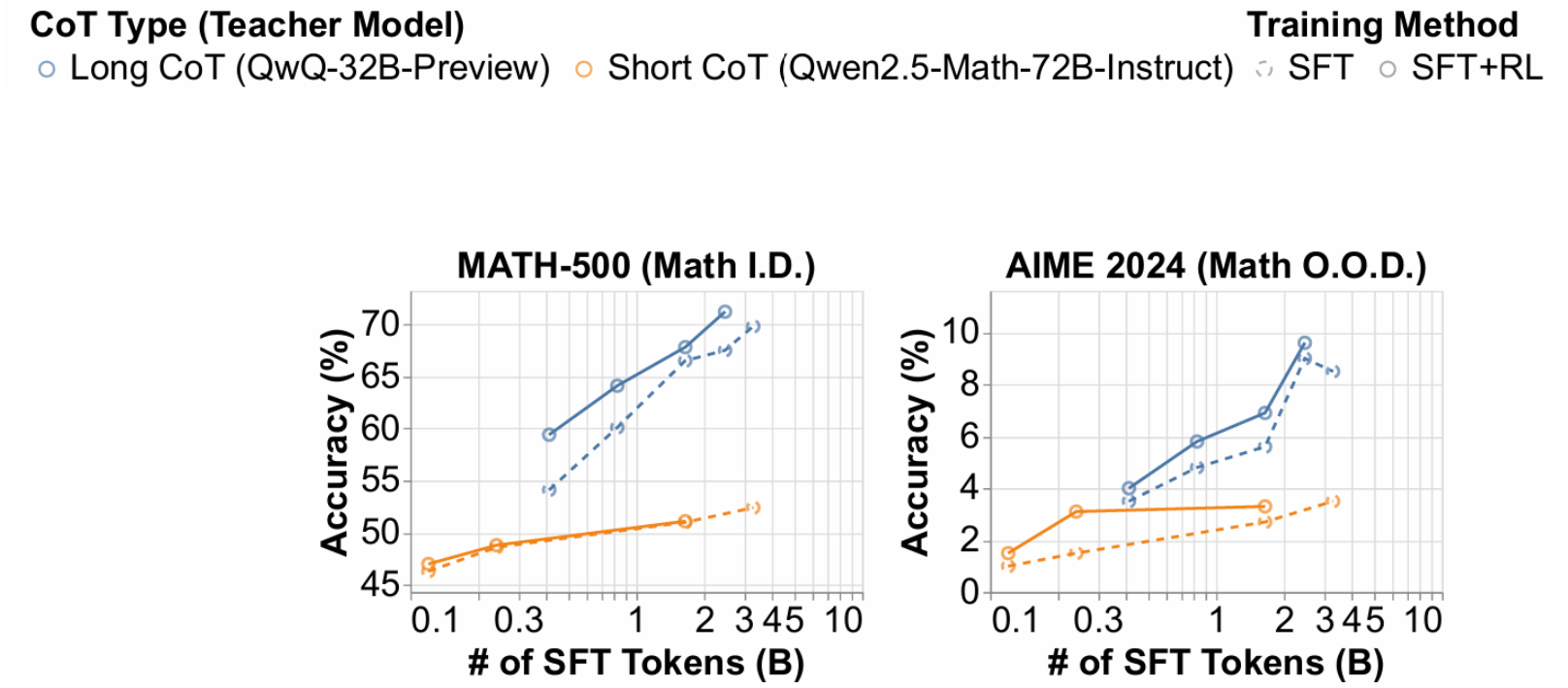


(c) Performance Comparison

(Qin et al. 2024)

Initial evidence that Journey learning can improve reasoning quality

Role of Supervised Finetuning



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)

(Yeo et al. 2025)

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

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)

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
- **Multilingual**
  - How do reasoning models work in non-English settings?
  - Multilingual thinking
  - Multilingual benchmarks

Thank You!

anoop.kunchukuttan@gmail.com

https://anoopkunchukuttan.gitlab.io/

Reading Material

- Jay Al Ammar's "The Illustrated DeepSeek R1" *(of Illustrated Transformer* fame*)*
- Nathan Lambert's "DeepSeek R1's recipe to replicate o1 and the future of reasoning LMs" *(Post-training lead at AI2 for the Tulu project)*
- Nathan Lambert "DeepSeek V3 and the actual cost of training frontier AI models"
- HuggingFace Post on "Scaling Test Time Compute"
- Lightman et al. "Let's Verify Step by Step" *(from OpenAI, ICLR 2024, on process reward models)*
- Phil Schmid https://www.philschmid.de/mini-deepseek-r1
- Qin et al. https://arxiv.org/abs/2410.18982