DeepSeek R1

Anoop Kunchukuttan

28th Jan 2025

Paper: https://arxiv.org/abs/2501.12948

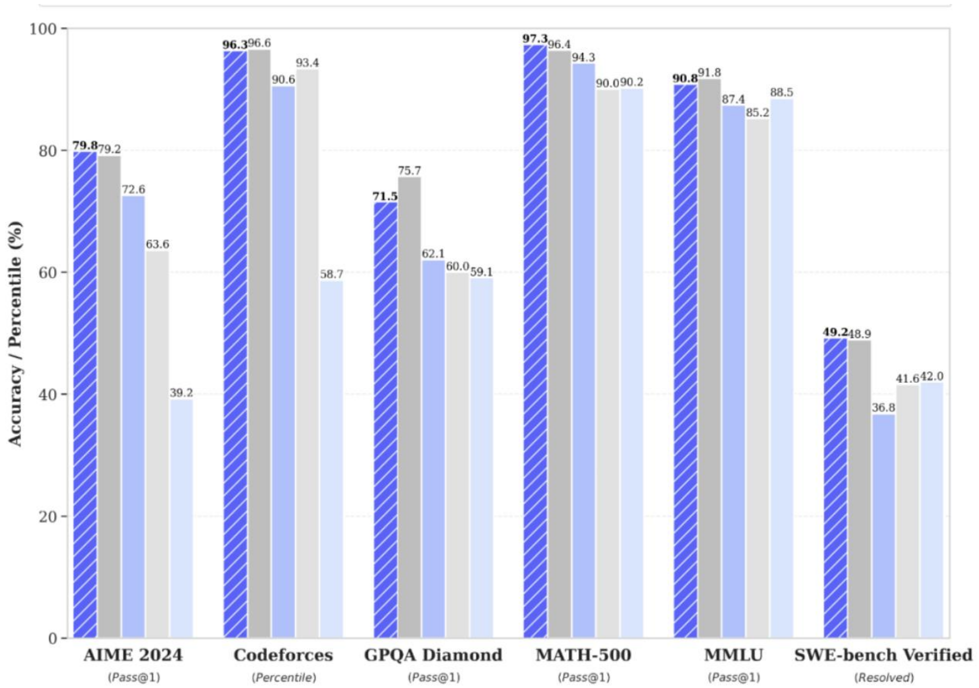
DeepSeek Impact

- **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!

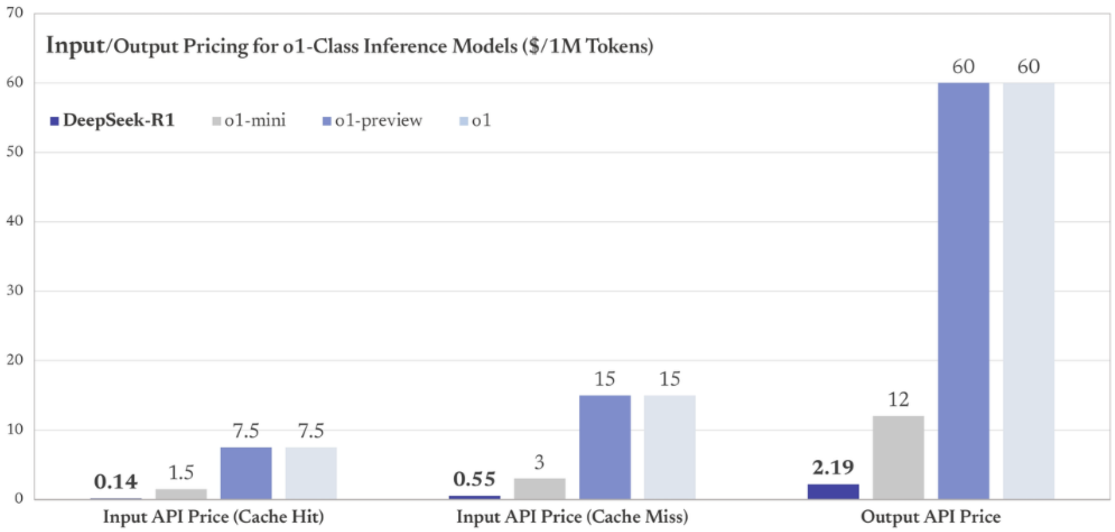
DeepSeek-R1 at a glance

- ⚡ 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

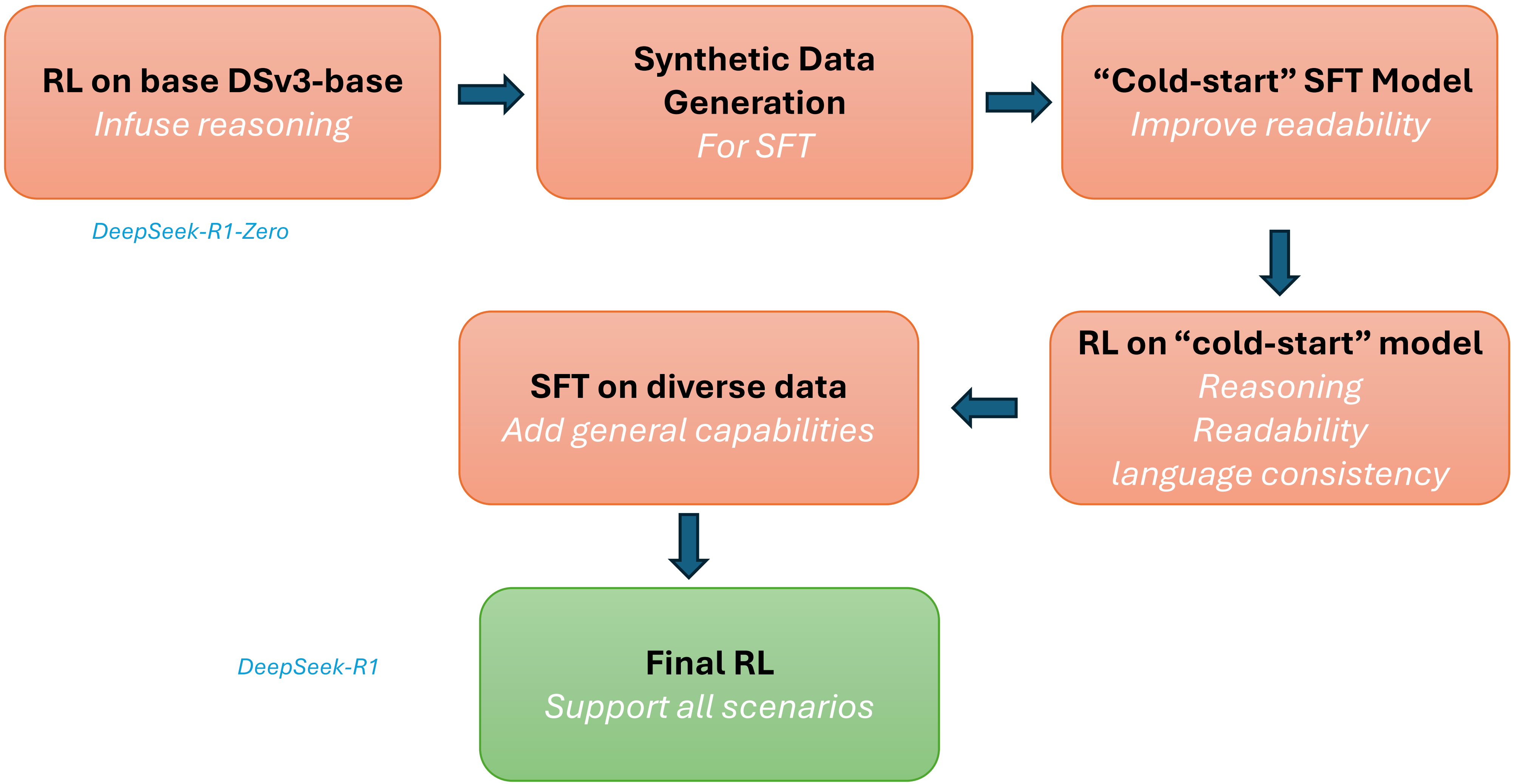


What does DeepSeek R1 release provide?

- DeepSeek Models
  - R1-Zero and R1
  - 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
- Poor readability and language mixing

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.

<u>Automatically learns to</u>

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

Model improves with more training

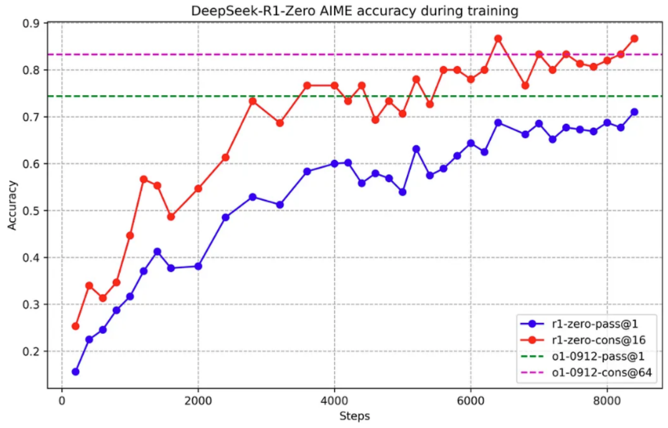


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

Model 'thinks' more with more training

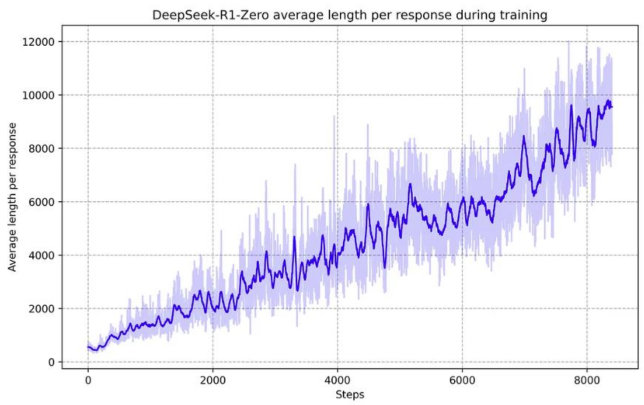


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

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

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.

"Aha Moment"

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
- Helpfulness: evaluate only the final response
- Harmlessness: evaluate reasoning traces as well

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

Post R1 release

- DeepSeek Chat No1 on AppStore
- **Open-R1**: https://huggingface.co/blog/open-r1
  - HuggingFace's community effort to replicate
  - A blueprint for now, not much has happened

Meanwhile other folks are already on the move

- **TinyZero** (from UCB): https://github.com/Jiayi-Pan/TinyZero
  - Reproduction of R1-Zero on countdown and multiplication tasks
  - Code available
  - Initial findings: Choice of RL method doesn't matter, IFT model converges faster
- **Bespoke Labs** generated data from DeepSeek-R1 and created distilled models https://huggingface.co/datasets/bespokelabs/Bespoke-Stratos-17k
- **SimpleRL-Reason**: HKU labs replicated DeepSeek R1 and R1-Zero on small models and resources (https://hkust-nlp.notion.site/simplerl-reason)

Open-source efforts

- **Open-R1**: https://huggingface.co/blog/open-r1
  - HuggingFace's community effort to replicate
  - A blueprint for now, not much has happened

Meanwhile other folks are already on the move

- **TinyZero** (from UCB): https://github.com/Jiayi-Pan/TinyZero
  - Reproduction of R1-Zero on countdown and multiplication tasks
  - Code available
  - Initial findings: Choice of RL method doesn't matter, IFT model converges faster
- **Bespoke Labs** generated data from DeepSeek-R1 and created distilled models https://huggingface.co/datasets/bespokelabs/Bespoke-Stratos-17k
- **SimpleRL-Reason**: HKU labs replicated DeepSeek R1 and R1-Zero on small models and resources (https://hkust-nlp.notion.site/simplerl-reason)

DeepSeek v3

Efficiently trained frontier model

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

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)*

Thank you!