

# Recent Tendency of LLMs Development



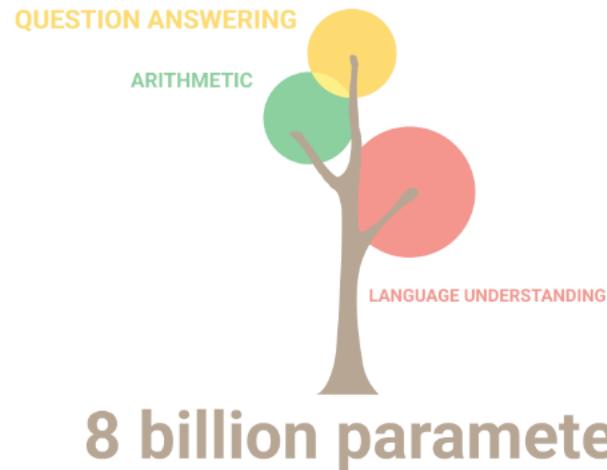
Zhang Jie, Scientist, CFAR, A\*STAR

[zhang\\_jie@cfar.a-star.edu.sg](mailto:zhang_jie@cfar.a-star.edu.sg)

<https://zjzac.github.io/>

26 Feb. 2025

# LLMs have taken the Whole World by storm



# LLMs have taken the Whole World by storm

## □ Overview of Current LLMs

RL Enhanced LLMs	Organization	# Params	RL Methods
Instruct-GPT (Ouyang et al., 2022)	OpenAI	1.3B, 6B, 175B	RLHF, PPO
GPT-4 (OpenAI, 2023)	OpenAI	-	RLHF, PPO, RBRM
Gemini (Team et al., 2023)	Google	-	RLHF
InternLM2 (Cai et al., 2024)	上海人工智能实验室	1.8B, 7B, 20B	RLHF, PPO
Claude 3 (Anthropic, 2024)	ANTHROPIC	-	RLAIF
Reka (Team et al., 2024c)	Reka	7B, 21B	RLHF, PPO
Zephyr (HuggingFaceH4, 2024)	Argilla	141B-A39B	ORPO
Phi-3 (Abdin et al., 2024)	Microsoft	3.8B, 7B, 14B	DPO
DeepSeek-V2 (Liu et al., 2024a)	deepseek	236B-A21B	GRPO
ChatGLM (GLM et al., 2024)	ZHIPU·AI	6B, 9B	ChatGLM-RLHF
Nemotron-4 340B (Adler et al., 2024)	NVIDIA	340B	DPO, RPO
Llama 3 (Dubey et al., 2024)	Meta	8B, 70B, 405B	DPO
Qwen2 (Yang et al., 2024a)	Alibaba	(0.5-72)B, 57B-A14B	DPO
Gemma2 (Team et al., 2024b)	Google	2B, 9B, 27B	RLHF
Starling-7B (Zhu et al., 2024)	Berkeley	7B	RLAIF, PPO
Athene-70B (Nexusflow, 2024)	Nexusflow	70B	RLHF
Hermes 3 (Teknium et al., 2024)	NOUS	8B, 70B, 405B	DPO
o1 (OpenAI, 2024b)	OpenAI	-	RL through CoT

Table 1: An overview of RL Enhanced LLMs. The format ‘141B-A39B’ refers to MoE models with 141B total and 39B active parameters.

## Reinforcement Learning Enhanced LLMs: A Survey

Model	Organization	# Params	Open Source	Report/Paper Available	Comparison with o1
Gemini 2.0 Flash (Google AI)	Google	-	✗	✗	✗
QVQ-72B-Preview (QwenLM, QVQ)	Alibaba	72B	✓ <sup>1</sup>	✗	✓
Marco-o1 (Zhao et al., 2024a)	Alibaba	7B	✓ <sup>2</sup>	✓ <sup>8</sup>	✗
Skywork o1 (o1 Team, 2024)	KUNLUN	8B	✓ <sup>3</sup>	✗	✗
QwQ-32B-Preview (QwenLM, QwQ)	Alibaba	32B	✓ <sup>4</sup>	✗	✓
o1-Coder (Zhang et al., 2024d)	BEIJING JIAOTONG UNIVERSITY	-	✓ <sup>5</sup>	✓ <sup>9</sup>	✗
rStar-Math (Guan et al., 2025)	Microsoft	1.5B,3B,7B	✓ <sup>6</sup>	✓ <sup>10</sup>	✓
Kimi-k1.5 (Team et al., 2025)	Moonshot AI	-	✗	✓ <sup>11</sup>	✓
DeepSeek-R1 (DeepSeek-AI et al., 2025)	deepseek	671B-A31B	✓ <sup>7</sup>	✓ <sup>12</sup>	✓

<sup>1</sup> <https://huggingface.co/Qwen/QVQ-72B-Preview>

<sup>2</sup> <https://github.com/AIDC-AI/Marco-o1>

<sup>3</sup> <https://huggingface.co/Skywork/Skywork-o1-Open-Llama-3.1-8B>

<sup>4</sup> <https://huggingface.co/Qwen/QwQ-32B-Preview>

<sup>5</sup> <https://github.com/ADaM-BJTU/o1-Coder>

<sup>6</sup> <https://github.com/zhenqingqi/rStar>

<sup>7</sup> <https://huggingface.co/deepseek-ai/DeepSeek-R1>

<sup>8</sup> <https://arxiv.org/pdf/2501.04519>

<sup>9</sup> <https://arxiv.org/pdf/2411.14405>

<sup>10</sup> <https://arxiv.org/pdf/2412.00154>

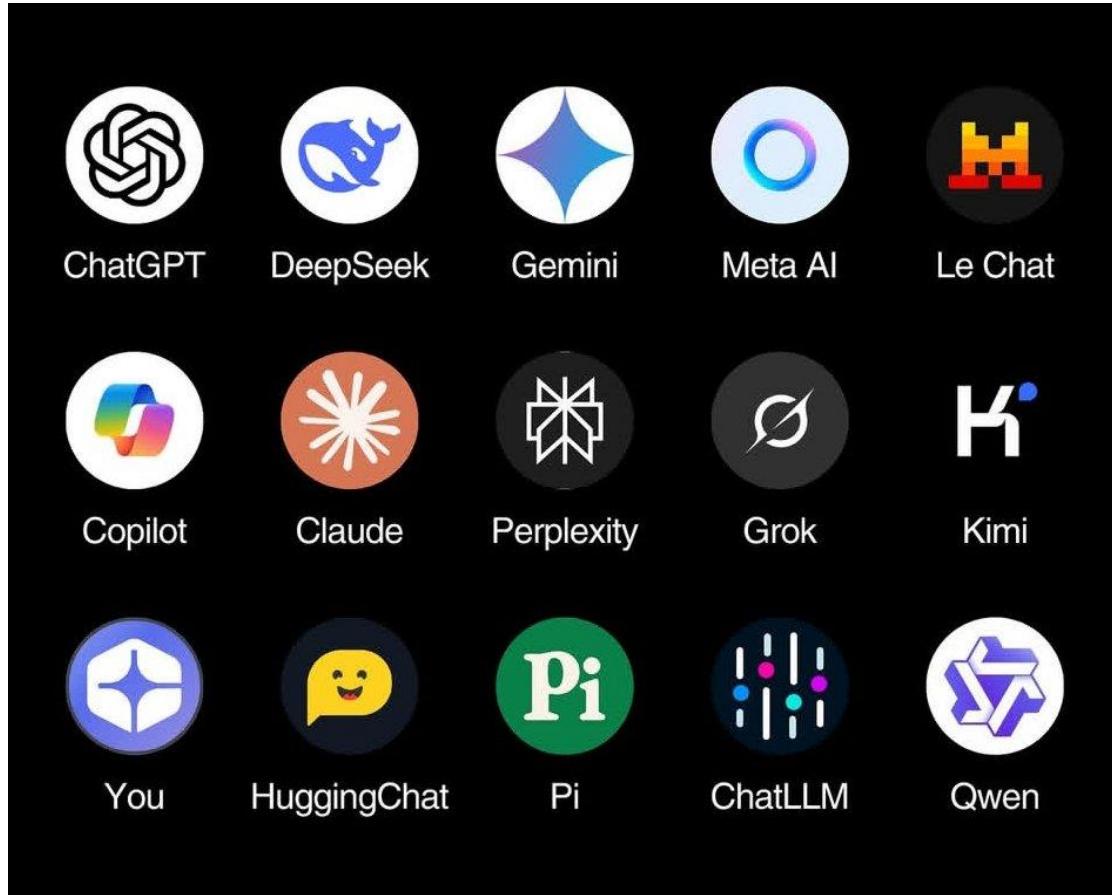
<sup>11</sup> <https://arxiv.org/pdf/2501.12599>

<sup>12</sup> <https://arxiv.org/pdf/2501.12948>

Unlocking the Mysteries of OpenAI o1: A Survey of the Reasoning Abilities of Large Language Models

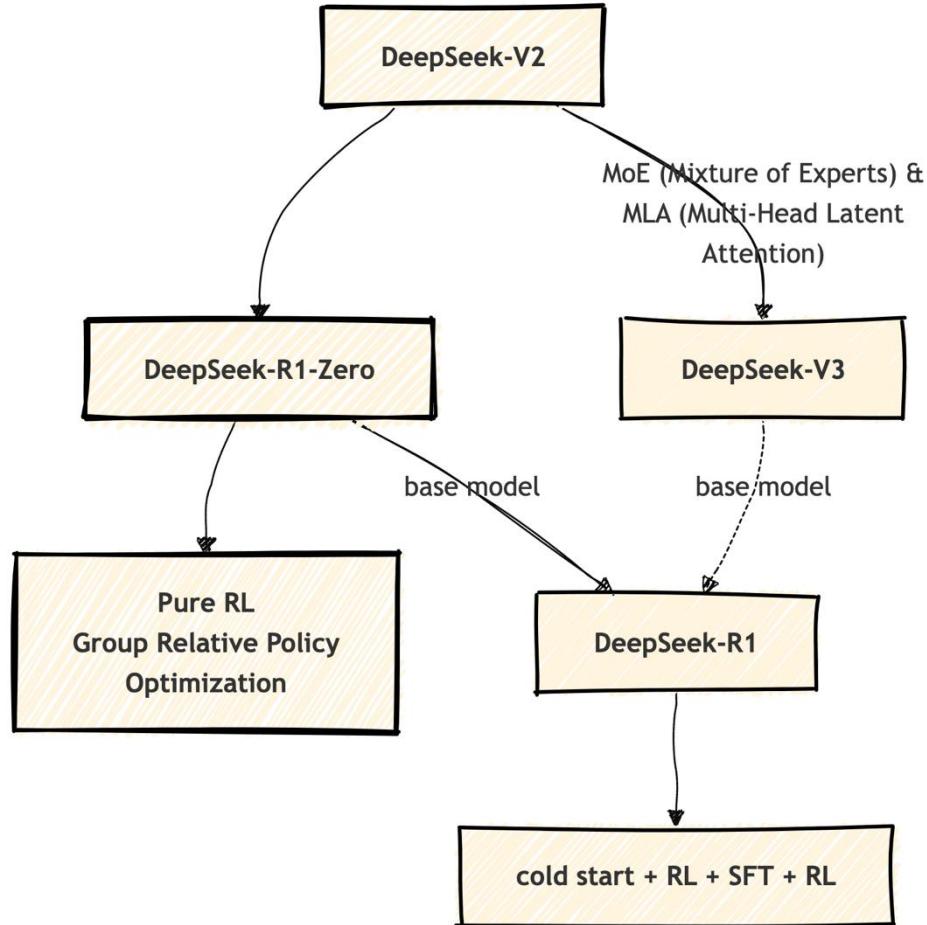
# LLMs have taken the Whole World by storm

## □ Diverse LLMs You Can Use



# Recent Released Advanced LLMs (Q1 2025)

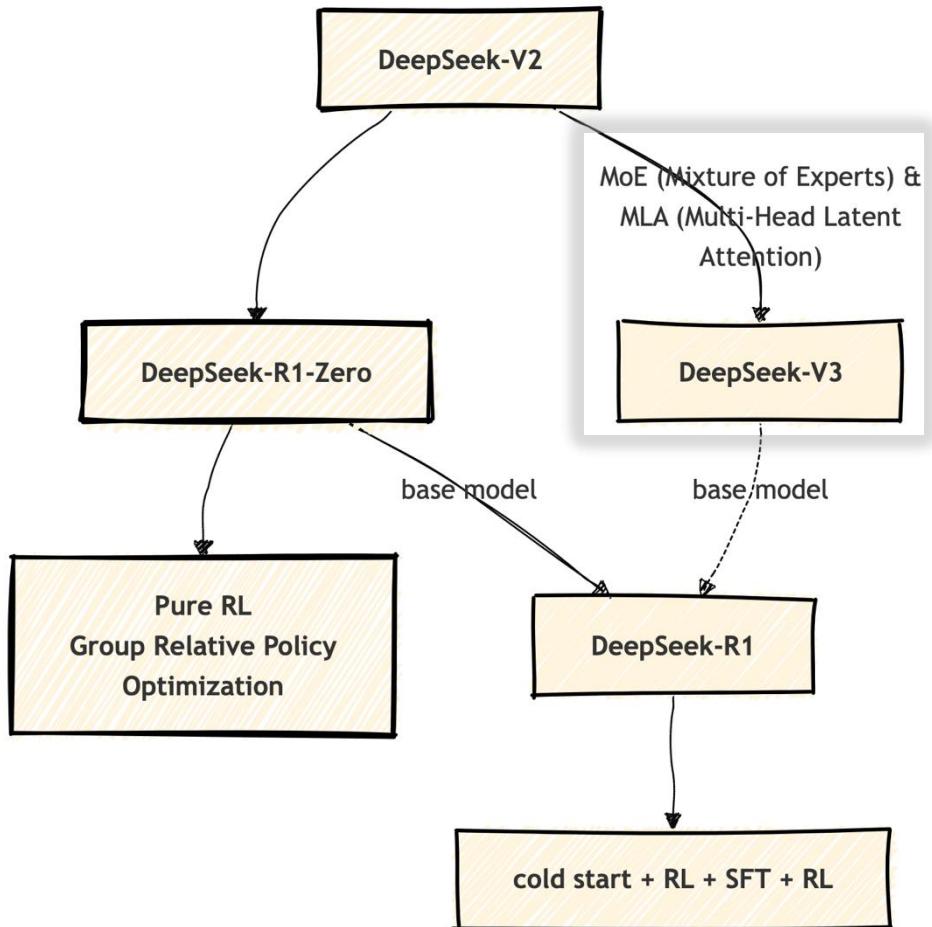
## □ DeepSeek



DeepSeek Evolution Process

# Recent Released Advanced LLMs (Q1 2025)

## □ DeepSeek



DeepSeek Evolution Process

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	-	-
MMLU (EM)	88.5	85.3	<b>88.6</b>	88.3	87.2
MMLU-Redux (EM)	<b>89.1</b>	85.6	86.2	88.9	88
MMLU-Pro (EM)	75.9	71.6	73.3	<b>78</b>	72.6
DROP (3-shot F1)	<b>91.6</b>	76.7	88.7	88.3	83.7
English	IF-Eval (Prompt Strict)	86.1	84.1	86	<b>86.5</b>
	GPQA-Diamond (Pass@1)	59.1	49	51.1	<b>65</b>
	SimpleQA (Correct)	24.9	9.1	17.1	<b>38.2</b>
	FRAMES (Acc.)	73.3	69.8	70	<b>72.5</b>
	LongBench v2 (Acc.)	<b>48.7</b>	39.4	36.1	41
HumanEval-Mul (Pass@1)	82.6	77.3	77.2	81.7	80.5
LiveCodeBench(Pass@1-COT)	<b>40.5</b>	31.1	28.4	36.3	33.4
LiveCodeBench (Pass@1)	<b>37.6</b>	28.7	30.1	32.8	34.2
Code	Codeforces (Percentile)	<b>51.6</b>	24.8	25.3	20.3
	SWE Verified (Resolved)	42	23.8	24.5	<b>50.8</b>
	Aider-Edit (Acc.)	79.7	65.4	63.9	<b>84.2</b>
	Aider-Polyglot (Acc.)	<b>49.6</b>	7.6	5.8	45.3
AIME 2024 (Pass@1)	<b>39.2</b>	23.3	23.3	16	9.3
Math	MATH-500 (EM)	<b>90.2</b>	80	73.8	78.3
	CNMO 2024 (Pass@1)	<b>43.2</b>	15.9	6.8	13.1
Chinese	CLUEWSC (EM)	90.9	<b>91.4</b>	84.7	85.4
	C-Eval (EM)	<b>86.5</b>	86.1	61.5	76
	C-SimpleQA (Correct)	<b>64.1</b>	48.4	50.4	51.3

<https://x.com/itsPaulAi/status/1872320003770618146>

### Training Cost Comparison

(DeepSeek V3 vs LLaMA 3)

1/100

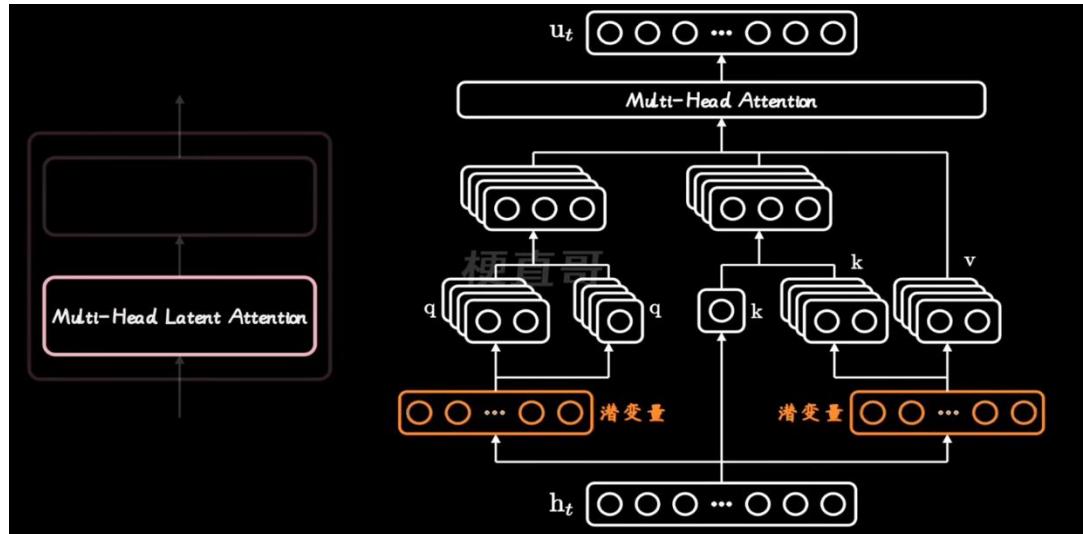
### Inference Cost Comparison

(DeepSeek V3 vs OpenAI o1)

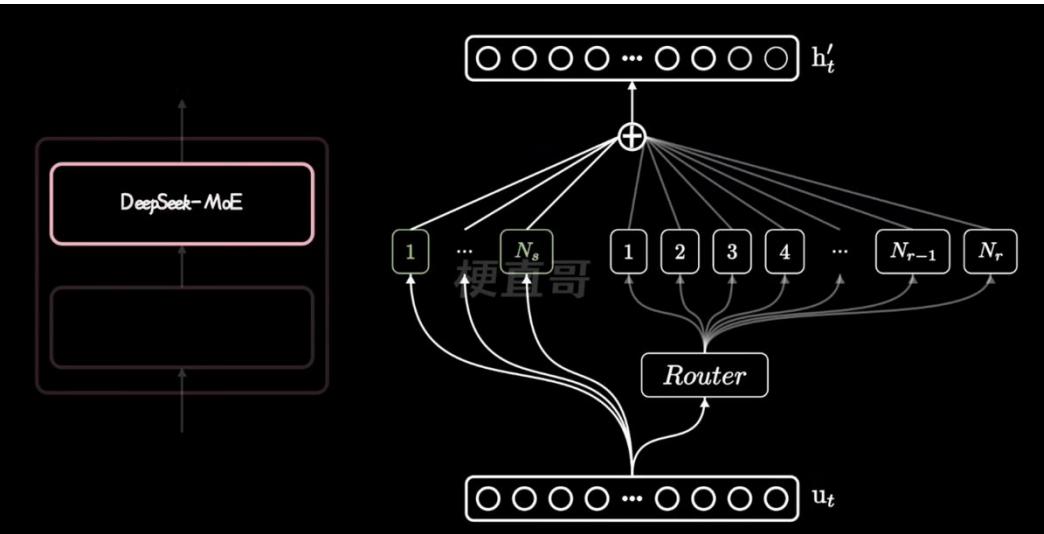
1/30

# Recent Released Advanced LLMs (Q1 2025)

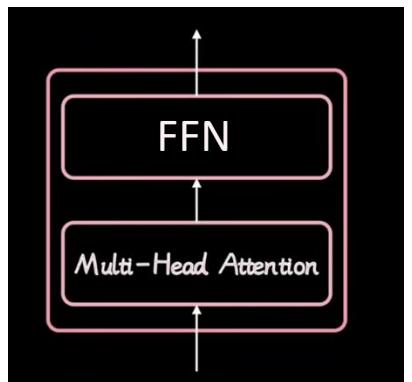
## □ DeepSeek V2&V3



<https://arxiv.org/pdf/2405.04434>



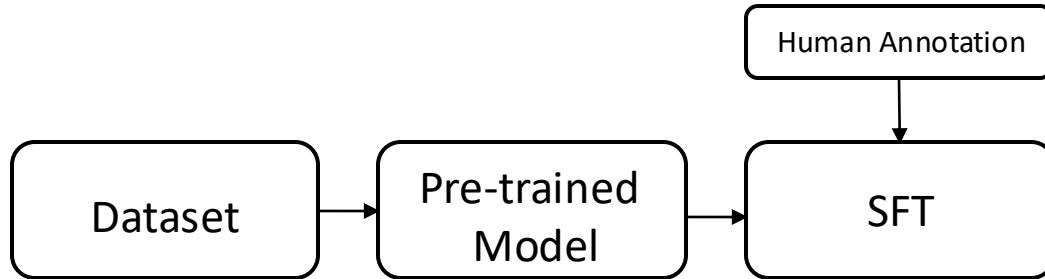
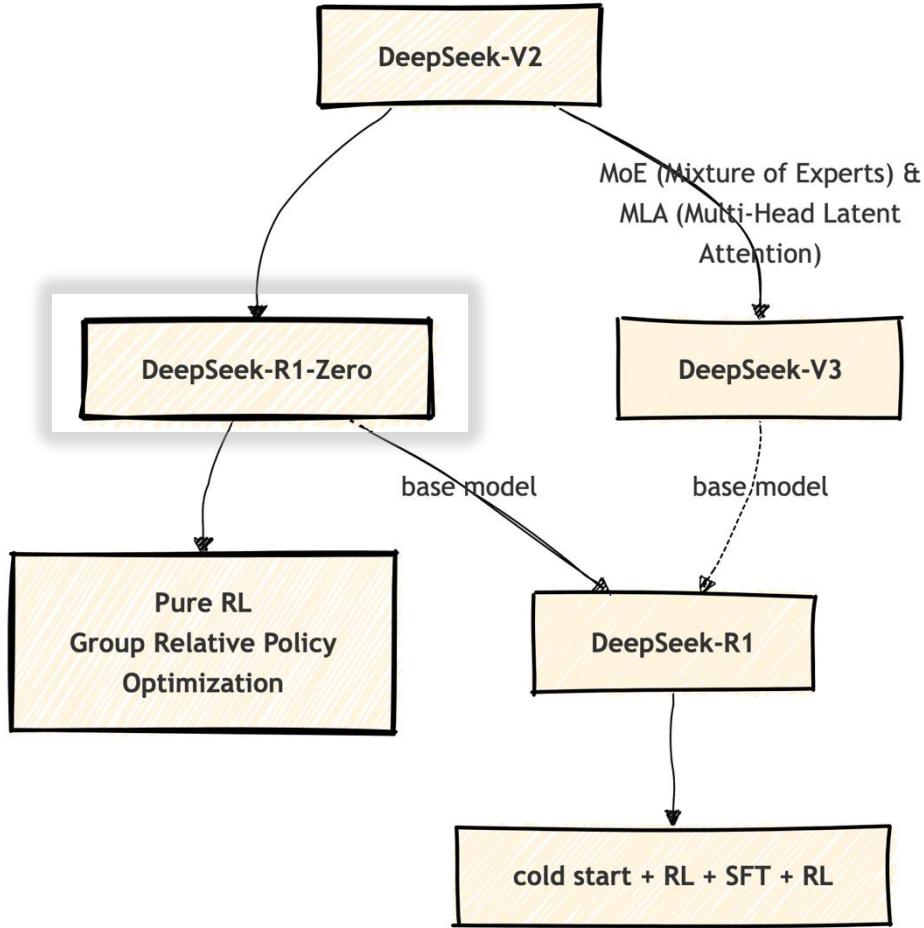
<https://arxiv.org/pdf/2401.06066>



[https://www.bilibili.com/video/BV16dNfeME3S?spm\\_id\\_from=333.788.player.switch&vd\\_source=7345af47d402aec64db3e67607045949](https://www.bilibili.com/video/BV16dNfeME3S?spm_id_from=333.788.player.switch&vd_source=7345af47d402aec64db3e67607045949)

# Recent Released Advanced LLMs (Q1 2025)

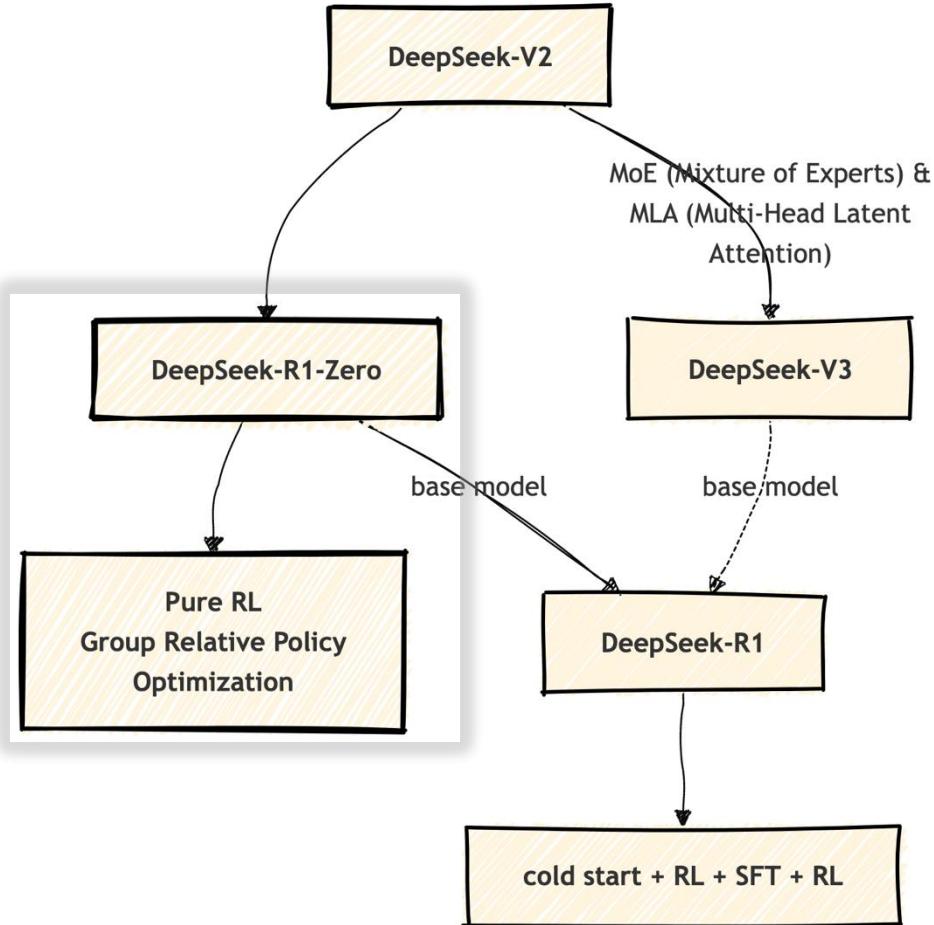
## □ DeepSeek



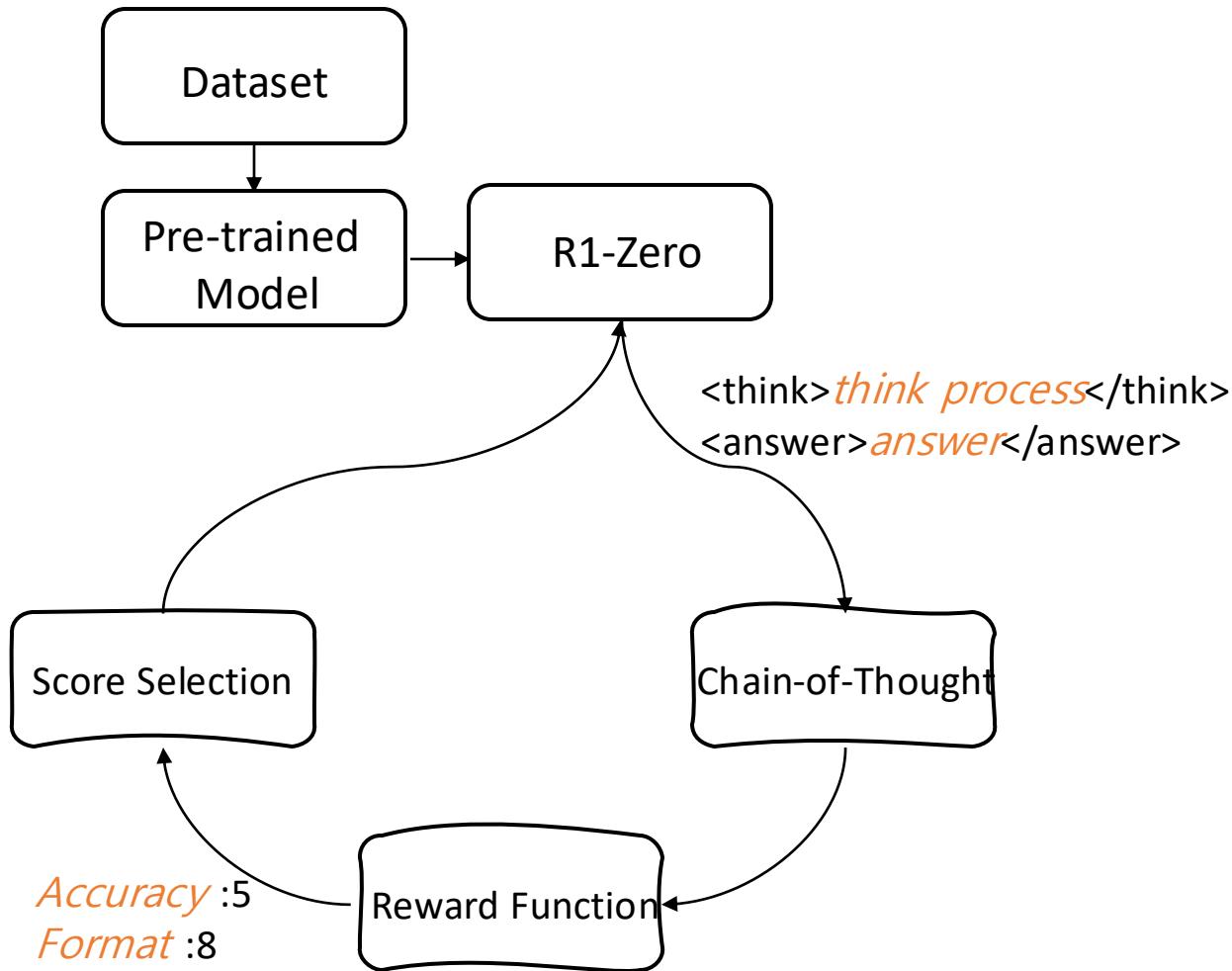
Traditional SFT, heavy cost

# Recent Released Advanced LLMs (Q1 2025)

## □ DeepSeek



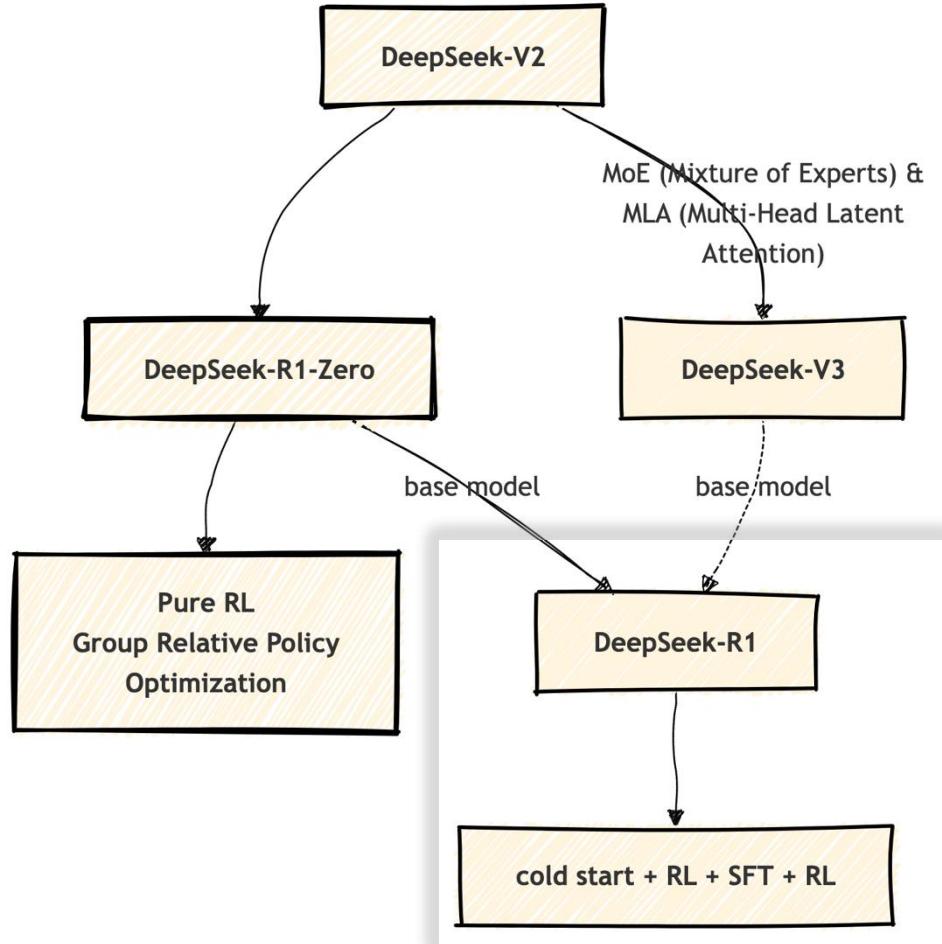
DeepSeek Evolution Process



DeepSeek-R1-Zero encounters challenges such as *poor readability, and language mixing*

# Recent Released Advanced LLMs (Q1 2025)

## □ DeepSeek

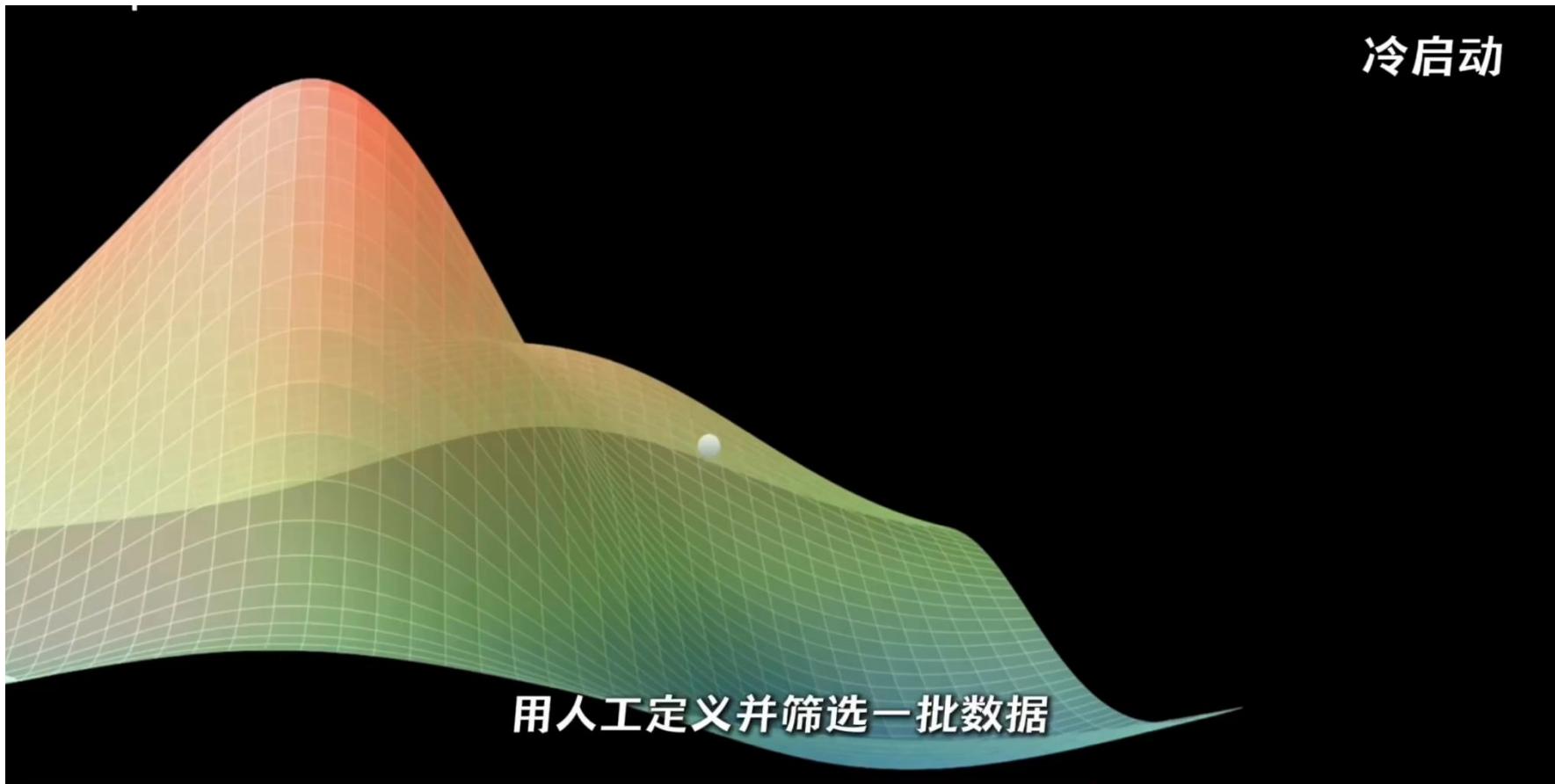
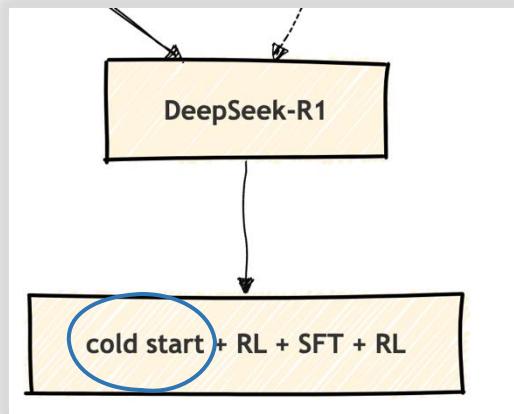


DeepSeek Evolution Process

# Recent Released Advanced LLMs (Q1 2025)

## □ DeepSeek R1

we collect thousands of cold-start data to fine-tune the DeepSeek-V3-Base as the starting point for RL

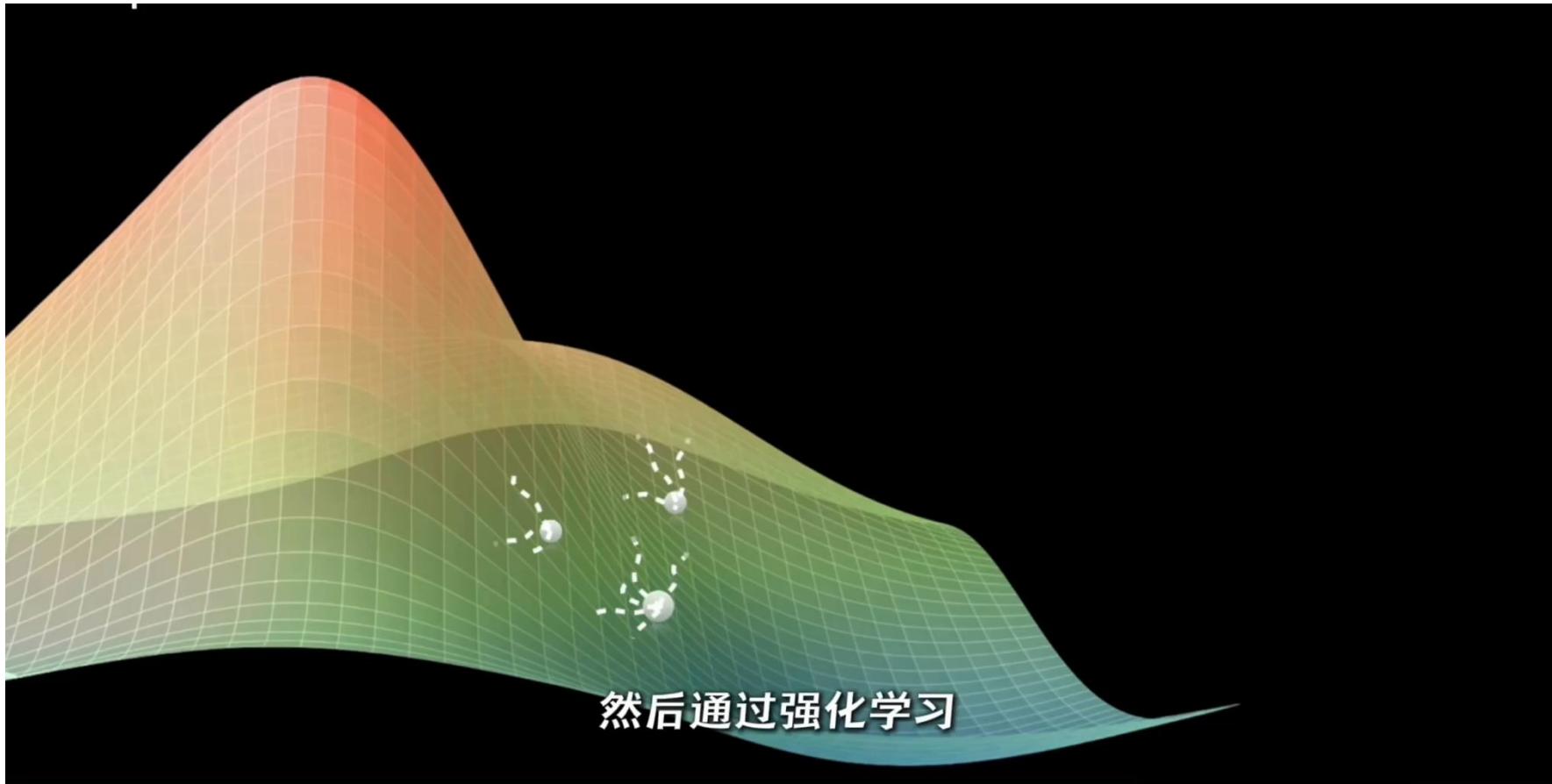
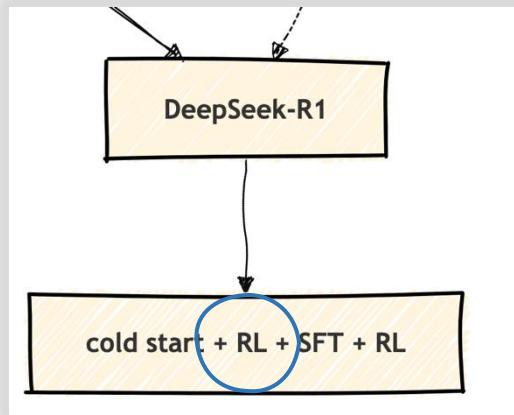


[https://www.bilibili.com/video/BV16dNfeME3S?spm\\_id\\_from=333.788.player.switch&vd\\_source=7345af47d402aec64db3e67607045949](https://www.bilibili.com/video/BV16dNfeME3S?spm_id_from=333.788.player.switch&vd_source=7345af47d402aec64db3e67607045949)

# Recent Released Advanced LLMs (Q1 2025)

## □ DeepSeek R1

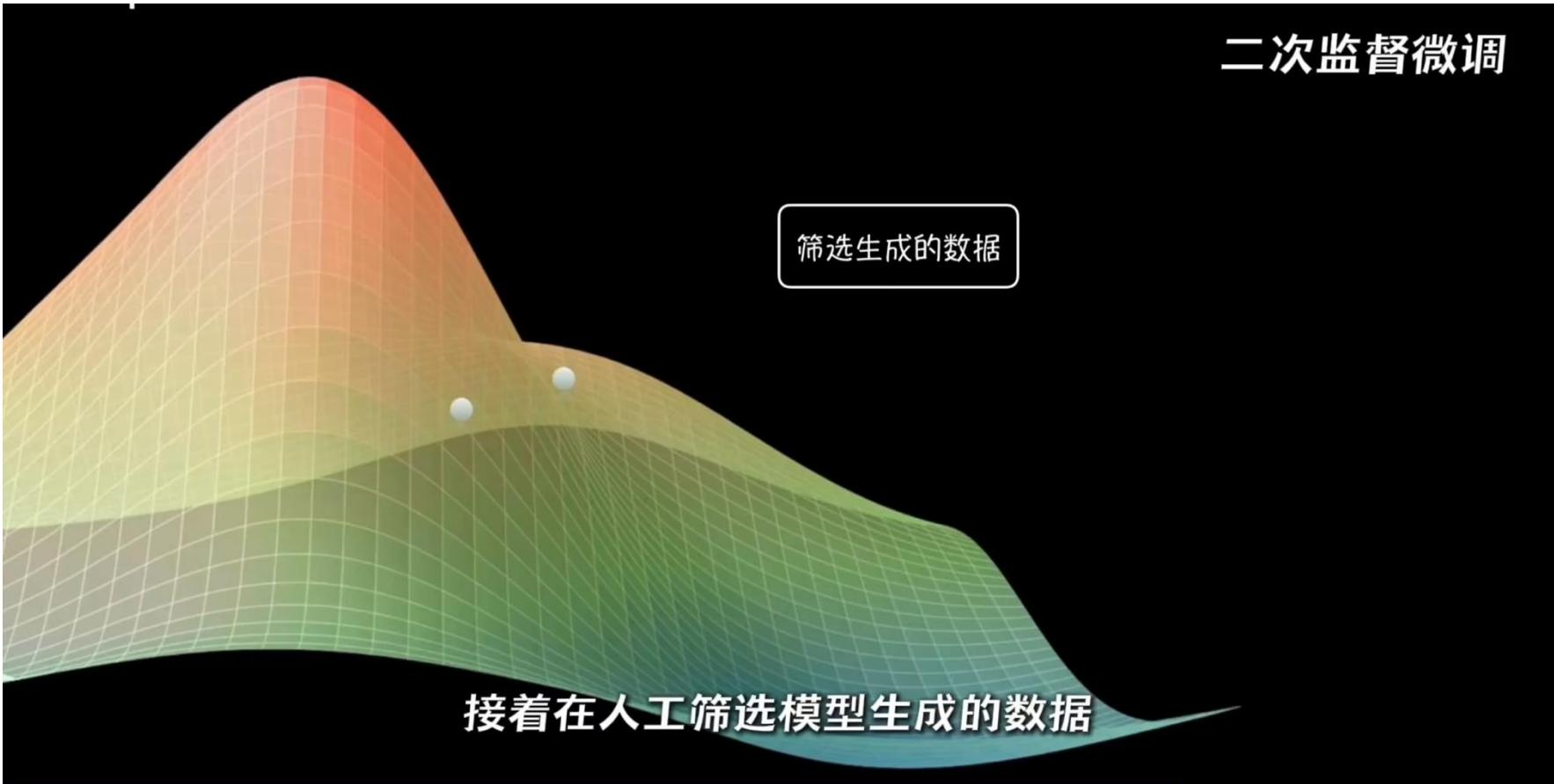
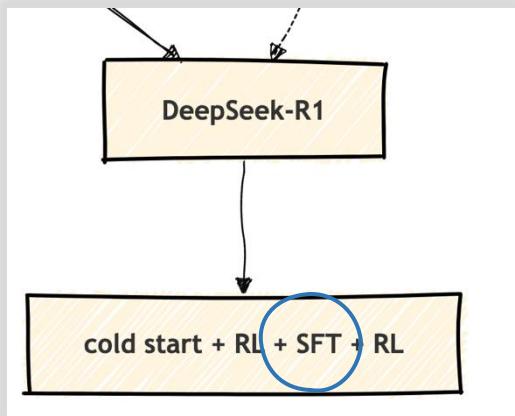
 we introduce a language consistency reward during RL training, which is calculated as the proportion of target language words in the CoT.



# Recent Released Advanced LLMs (Q1 2025)

## □ DeepSeek R1

Unlike the initial cold-start data, which primarily focuses on reasoning, this stage incorporates data from other domains to enhance the model's capabilities in writing, role-playing, and other general-purpose tasks

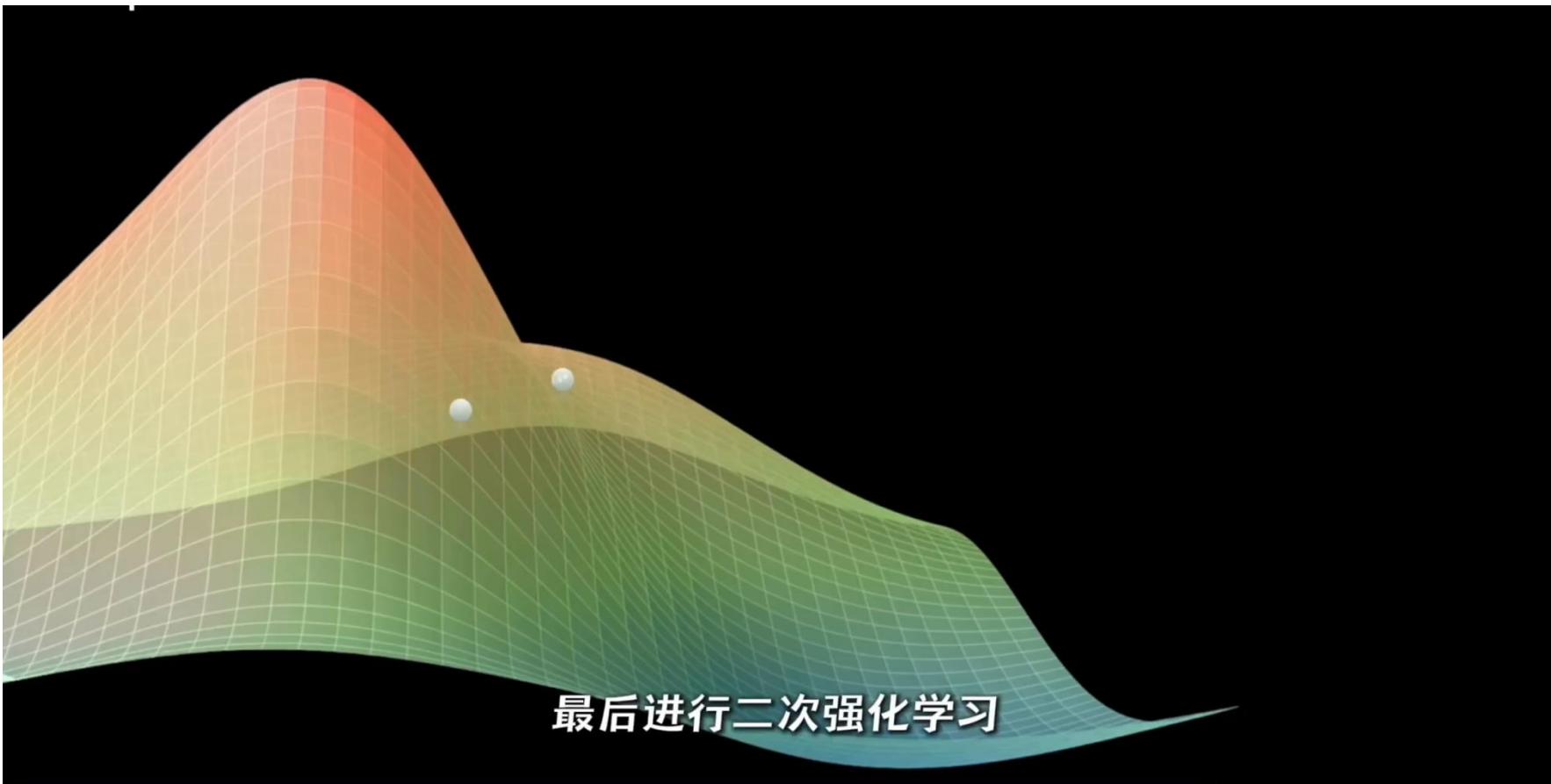
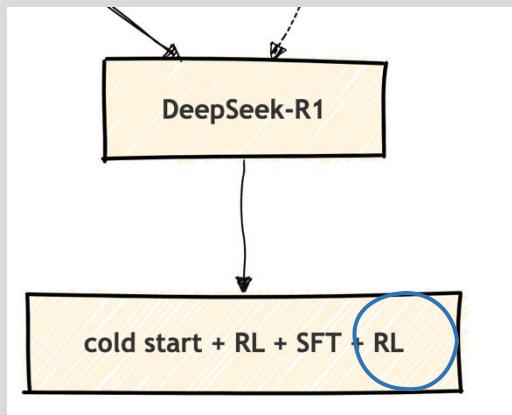


[https://www.bilibili.com/video/BV16dNfeME3S?spm\\_id\\_from=333.788.player.switch&vd\\_source=7345af47d402aec64db3e67607045949](https://www.bilibili.com/video/BV16dNfeME3S?spm_id_from=333.788.player.switch&vd_source=7345af47d402aec64db3e67607045949)

# Recent Released Advanced LLMs (Q1 2025)

## □ DeepSeek R1

To further align the model with human preferences, we implement a secondary reinforcement learning stage aimed at improving the model's helpfulness and harmlessness while simultaneously refining its reasoning capabilities.



[https://www.bilibili.com/video/BV16dNfeME3S?spm\\_id\\_from=333.788.player.switch&vd\\_source=7345af47d402aec64db3e67607045949](https://www.bilibili.com/video/BV16dNfeME3S?spm_id_from=333.788.player.switch&vd_source=7345af47d402aec64db3e67607045949)

# Recent Released Advanced LLMs (Q1 2025)



## Following techniques

**DeepSeek** @deepseek\_ai

Introducing NSA: A Hardware-Aligned and Natively Trainable Sparse Attention mechanism for ultra-fast long-context training & inference!

Core components of NSA:

- Dynamic hierarchical sparse strategy
- Coarse-grained token compression
- Fine-grained token selection

With optimized design for modern hardware, NSA speeds up inference while reducing pre-training costs—without compromising performance. It matches or outperforms Full Attention models on general benchmarks, long-context tasks, and instruction-based reasoning.

For more details, check out our paper here: [arxiv.org/abs/2502.11089](https://arxiv.org/abs/2502.11089)

**Figure 1 | Comparison of performance and efficiency between Full Attention model and our NSA.** Left: Performance on various benchmarks, including sequence length, memory access, and reasoning evaluation. Right: Speedup ratio for 64d-length sequence processing. NSA achieves substantial computational speedup compared to Full Attention in all stages: decoding, forward propagation, and backward propagation.

**Figure 2 | Overview of NSA's architecture.** Left: The framework processes input sequences through three parallel attention branches: for a given query, preceding keys and values are used to compute compressed attention for coarse-grained patterns, selected attention for fine-grained patterns, and sliding attention for local context. Right: Visualization of different attention patterns for a given query.

**Figure 3 | Kernel design for NSA.** The kernel loads queries by GQA groups (Grid Loop), fetches keys and values from memory, and performs attention computation on SRAM. Green blocks indicate data on SRAM, while blue indicates data on HBM.

**DeepSeek** @deepseek\_ai

Day 0: Warming up for #OpenSourceWeek!

We're a tiny team @deepseek\_ai exploring AGI. Starting next week, we'll be open-sourcing 5 repos, sharing our small but sincere progress with full transparency.

These humble building blocks in our online service have been documented, deployed and battle-tested in production.

As part of the open-source community, we believe that every line shared becomes collective momentum that accelerates the journey.

Daily unlocks are coming soon. No ivory towers - just pure garage-energy and community-driven innovation.

DeepSeek's approach vs. larger AI companies? Significance of open-source?

12:00 PM · Feb 21, 2025 · 2.1M Views

**DeepSeek** @deepseek\_ai

Day 1 of #OpenSourceWeek: FlashMLA

Honored to share FlashMLA – our efficient MLA decoding kernel for Hopper GPUs, optimized for variable-length sequences and now in production.

- ✓ BF16 support
- ✓ Paged KV cache (block size 64)
- ⚡ 3000 GB/s memory-bound & 580 TFLOPS compute-bound on H800

Explore on GitHub: [github.com/deepseek-ai/FlashMLA](https://github.com/deepseek-ai/FlashMLA)

9:34 AM · Feb 24, 2025 · 370.7K Views

**DeepSeek** Deep Expert Parallelism @deepseek\_ai

Day 2 of #OpenSourceWeek: DeepEP

Excited to introduce DeepEP – the first open-source EP communication library for MoE model training and inference.

- ✓ Efficient and optimized all-to-all communication
- ✓ Both intranode and internode support with NVLink and RDMA
- ✓ High-throughput kernels for training and inference prefilling
- ✓ Low-latency kernels for inference decoding
- ✓ Native FP8 dispatch support
- ✓ Flexible GPU resource control for computation-communication overlapping

GitHub: [github.com/deepseek-ai/DeepEP](https://github.com/deepseek-ai/DeepEP)

10:24 AM · Feb 25, 2025 · 142.2K Views

**DeepSeek** @deepseek\_ai

Day 3 of #OpenSourceWeek: DeepGEMM

Introducing DeepGEMM – an FP8 GEMM library that supports both dense and MoE GEMMs, powering V3/R1 training and inference.

- ⚡ Up to 1350+ FP8 TFLOPS on Hopper GPUs
- ✓ No heavy dependency, as clean as a tutorial
- ✓ Fully Just-In-Time compiled
- ✓ Core logic at ~300 lines – yet outperforms expert-tuned kernels across most matrix sizes
- ✓ Supports dense layout and two MoE layouts

GitHub: [github.com/deepseek-ai/DeepGEMM](https://github.com/deepseek-ai/DeepGEMM)

9:00 AM · Feb 26, 2025 · 65.6K Views

**DeepSeek** @deepseek\_ai

Day 4 of #OpenSourceWeek: Optimized Parallelism Strategies

- ✓ DualPipe – a bidirectional pipeline parallelism algorithm for computation-communication overlap in V3/R1 training.
- 🔗 [github.com/deepseek-ai/DualPipe](https://github.com/deepseek-ai/DualPipe)
- ✓ EPLB – an expert-parallel load balancer for V3/R1.
- 🔗 [github.com/deepseek-ai/EPLB](https://github.com/deepseek-ai/EPLB)

Analyze computation-communication overlap in V3/R1.

GitHub: [github.com/deepseek-ai/ParallelismStrategies](https://github.com/deepseek-ai/ParallelismStrategies)

# Recent Released Advanced LLMs (Q1 2025)

## □ The Waves Made by DeepSeek



### Impact on China's "Big Six" AI Startups:

- **Zero One Technology** (零一万物): Focused on industrial applications, established an industrial AI base in Suzhou
- **Step AI** (阶跃星辰): Released multiple models including Step-2-mini and Step-1o Vision
- **Minimax**: Released T2A-01 voice model series and emphasized open-source strategy
- Baichuan Intelligence (百川智能): Released Baichuan-M1-preview model and launched an AI pediatric doctor system
- Zhipu Technology (智谱华章): Continued Samsung partnership and expanded into AI drawing applications
- Moonshot AI (月之暗面): Released Kimi k1.5 multimodal model

### Future Trends:

- Industry moving towards more open collaboration and integration
- Focus shifting to **practical applications** rather than just model development
- Increasing emphasis on cost-effectiveness and accessibility

# Recent Released Advanced LLMs (Q1 2025)

## □ The Waves Made by DeepSeek



### Deployment in High Education Institute across China:



### DeepSeek Faces Access Restrictions Overseas:

**DeepSeek banned from Australian government devices amid national security concerns**

Home affairs minister Tony Burke says decision follows advice from intelligence agencies and is not in response to AI chatbot's country of origin, China

- Follow our Australia news live blog for latest updates
- Get our breaking news email, free app or daily news podcast



NEWS | 17 February 2025 |

### What are the best AI tools for research? Nature's guide

There are many large language models to choose from; some excel at coding, whereas others are better for synthesizing information.

By Elizabeth Gibney



perplexity-ai / r1-1776 like 1.57k Follow Perplexity 1k

Safetensors deepseek\_v3 custom\_code License: mit

Model card Files and versions Community 216

#### R1 1776

Blog link: <https://perplexity.ai/hub/blog/open-sourcing-r1-1776>

R1 1776 is a DeepSeek-R1 reasoning model that has been post-trained by Perplexity AI to remove Chinese Communist Party censorship. The model provides unbiased, accurate, and factual information while maintaining high reasoning capabilities.

<https://mp.weixin.qq.com/s/7ZfR3-9sNpXBV9MNYhfsyg>

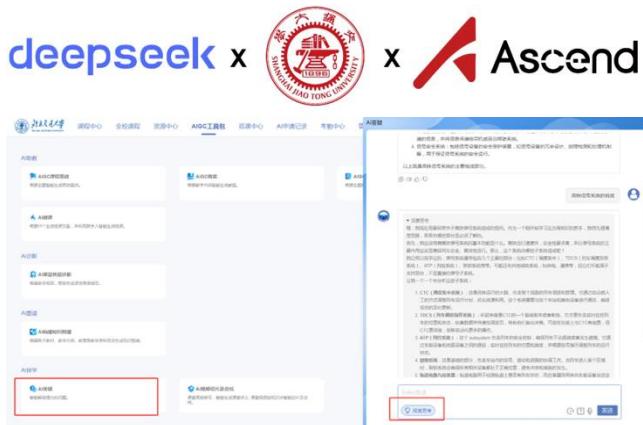
<https://huggingface.co/perplexity-ai/r1-1776>

# Recent Released Advanced LLMs (Q1 2025)

## □ The Waves Made by DeepSeek



Deployment in High Education Institute across China:



[perplexity-ai/r1-1776](#) like 1.57k Follow Perplexity 1k

Safetensors deepseek\_v3 custom\_code License: mit

Model card Files and versions Community 216

R1 1776

Blog link: <https://perplexity.ai/hub/blog/open-sourcing-r1-1776>

R1 1776 is a DeepSeek-R1 reasoning model that has been post-trained by Perplexity AI to remove Chinese Communist Party censorship. The model provides unbiased, accurate, and factual information while maintaining high reasoning capabilities.

*What will happen in Singapore?  
What can be used in A\*STAR?*

# Recent Released Advanced LLMs (Q1 2025)

## ☐ Reproduce "Aha Moment"



### Feature

- 🔥 Training with LoRA
- 4x3090/4090 GPUs training 1hour, 💰 cost < 7 dollar, 10min 37'step output "aha Moment"💡
- 0.5B scale model RL training
- support BIGGER model: 1.5B/7B/32B...
- We supply 0.75k/1.5k/7.5k dataset for fast train loop
- We logging GRPO online sampling data to log file

#### Aha Moment:

*Wait, that doesn't match either of our options. It seems like I made a mistake in my assumptions. Let's go back to the original equations*

Question: The sum of two numbers is 50 and their difference is 6. What is their product?

Response:

....

Let's assume the two numbers are A and B, and we know that their difference is 6:

$A - B = 6$

Now, let's try to find a pair of numbers that satisfy both conditions:

If  $A = 20$  and  $B = 12$  (even products and differences):

$20 + 12 = 50$  (sum)

$20 - 12 = 8$  (difference)

Okay, so let's give it a try: A is 20 and B is 12. Their product would be:

$20 * 12 = 240$

*Wait, that doesn't match either of our options. It seems like I made a mistake in my assumptions. Let's go back to the original equations:*

1.  $A + B = 50$

2.  $A - B = 6$

We see that our assumption led us to an incorrect pair. Let's correct our approach:

If we start with A and B being the two numbers that add up to 50 and have a difference of 6, let's try A = 1 and B = 28:

.....



### ⭐ Features

#### 💡 Simplicity

The project code is simple, with only about 200 lines of code spread across 2 files. It only depends on standard libraries such as `deepspeed` and `torch`, without requiring dependencies like `ray`. It is designed to allow for more complex interventions.

#### 🤖 Split Reference Model

The reference model part is decoupled, which allows it to be run on different GPUs (even on a different machine with 4090). This avoids having the reference model and the training model on the same GPU, preventing multiple copies created by `torch's` multiprocessing, and enabling training of a 7B model on 80G A800.

#### 🔥 Performance

Training completed in under 1 hour on 1\*A800 GPUs. Both Qwen2.5-7B and Qwen2.5-3B exhibited an "Aha moment" within the first 30 optimization steps.

#### ?] Core Loss Calculation

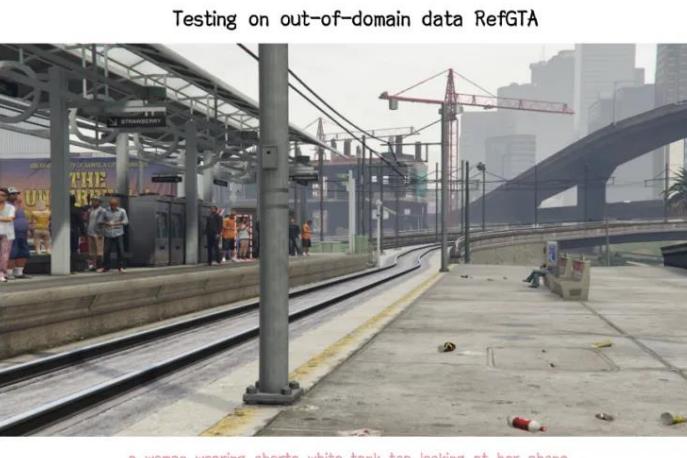
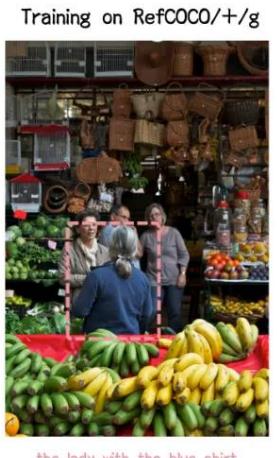
The loss calculation formula is based on Hugging Face's `trl`. We extend our gratitude to Hugging Face for their contribution.

# Recent Released Advanced LLMs (Q1 2025)

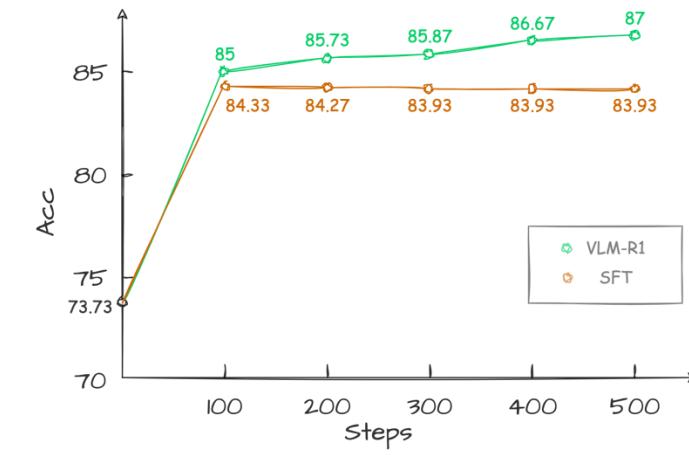
## □ VLM-R1

The research team conducted experiments on **Qwen2.5-VL**, comparing the **R1 method** with traditional **Supervised Fine-Tuning (SFT)**. The results were outstanding:

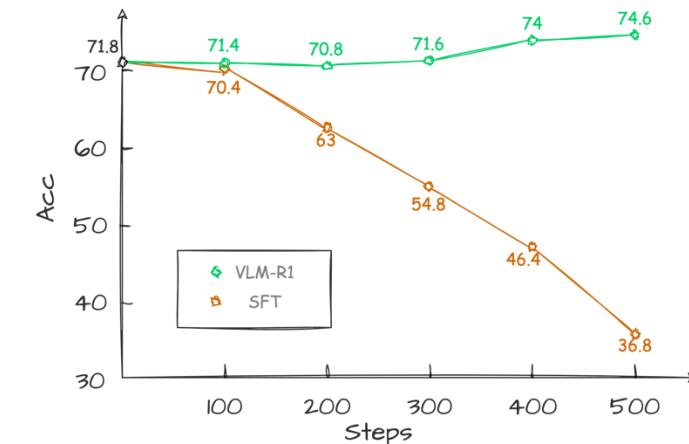
- **Exceptional Stability** – The R1 method consistently maintains high performance in various complex scenarios, which is critical for real-world applications.
- **Superior Generalization Ability** – One of the most surprising findings was that, on **out-of-domain test data**, traditional **SFT models** showed declining performance over time, whereas **R1 models continued to improve!**



Performance on in-domain test data  
(Avg Acc on Val split of RefCOCO/+/g)

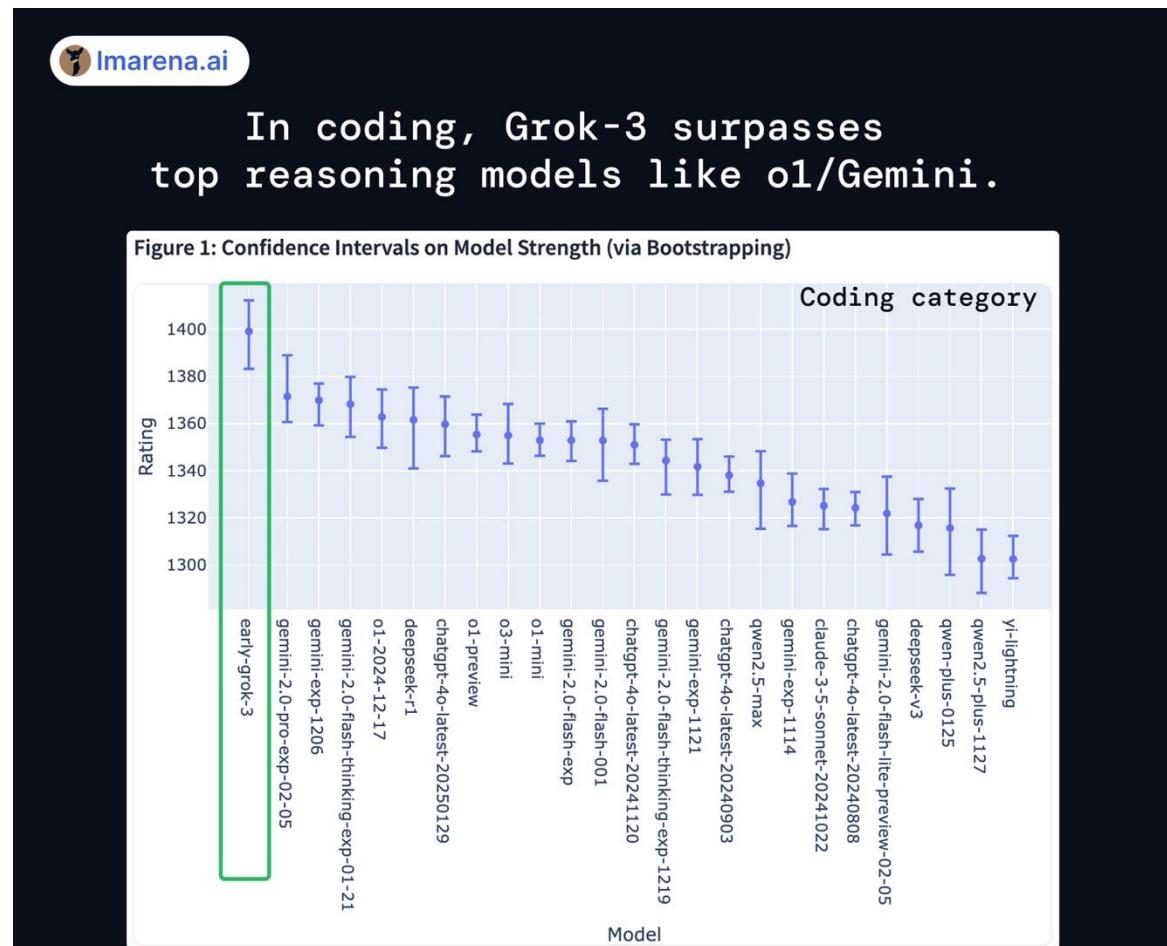


Performance on out-of-domain test data  
(Acc on RefGTA)



# Recent Released Advanced LLMs

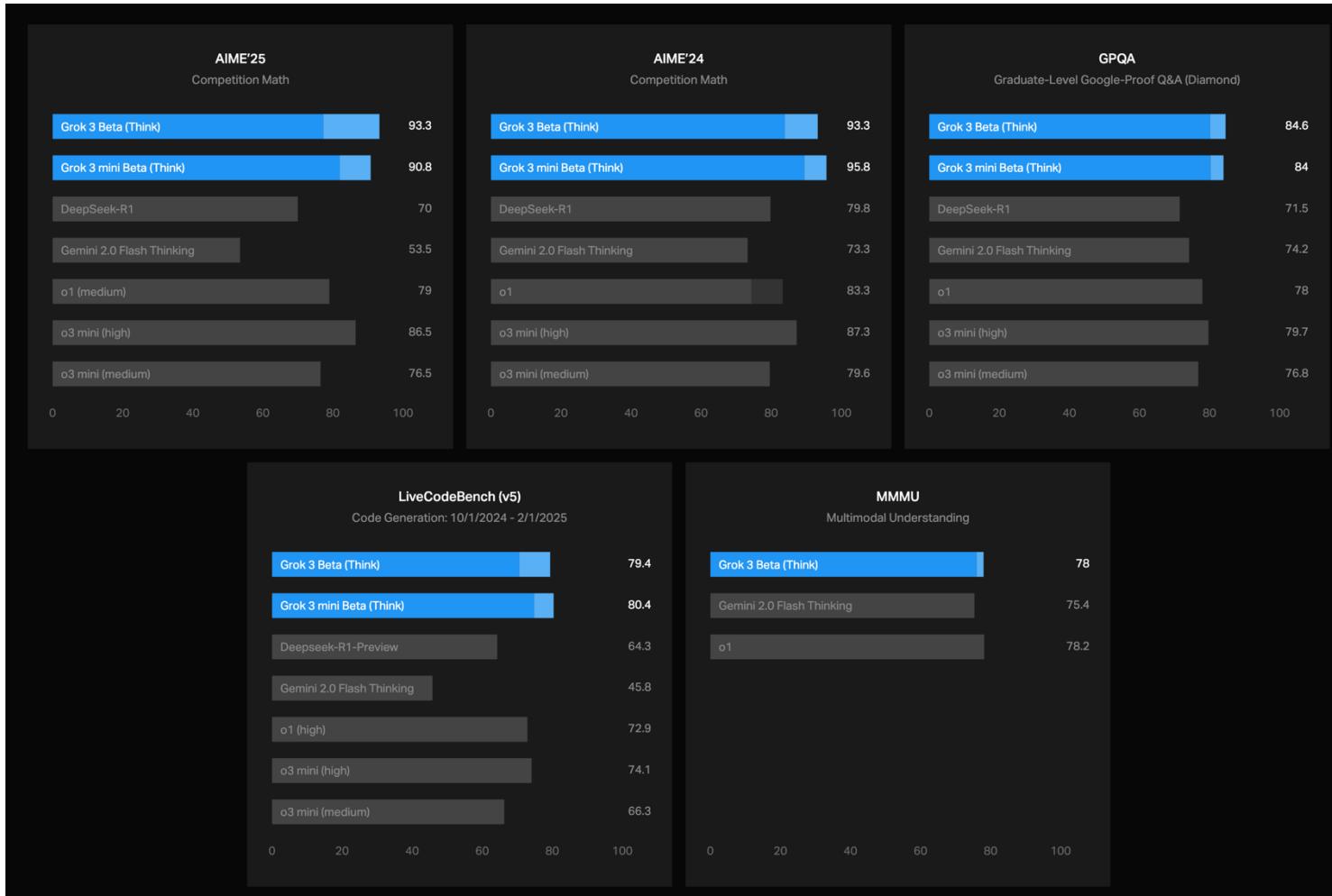
## □ X-Grok3 (Feb 18)



**Colossus Supercomputer P1: 100K GPUs &122days; P2: 200K GPU & 92 days**

# Recent Released Advanced LLMs

## ☐ X-Grok3 (Feb 18)



Furthermore, Grok 3 mini reaches a new frontier in cost-efficient reasoning for STEM tasks that don't require as much world knowledge, reaching 95.8% on AIME 2024 and 80.4% on LiveCodeBench.

# Recent Released Advanced LLMs

## □ X-Grok3 (Feb 18)

The screenshot shows the Grok3 interface. On the left, a sidebar displays a search history: "DeepSearch 24s", "Thinking", and "Researching Starship launch". The main area shows a search result for "Starship launch January 13, 2025". It indicates "10 results found" and lists several sources, including a link to "Starship" on nextspaceflight.com, a Wikipedia page on "List of Starship launches", and a Spaceflight Now article. Below the results, there's a section titled "Browsing results" with two bullet points: "Starship Flight 7 was scheduled for January 13, 2025, and some sources mention it launched on January 16, 2025 ([Space](#)).". At the bottom, a message box asks "How can Grok help?" and shows a count of "16 web pages".

### Grok 3 is here.

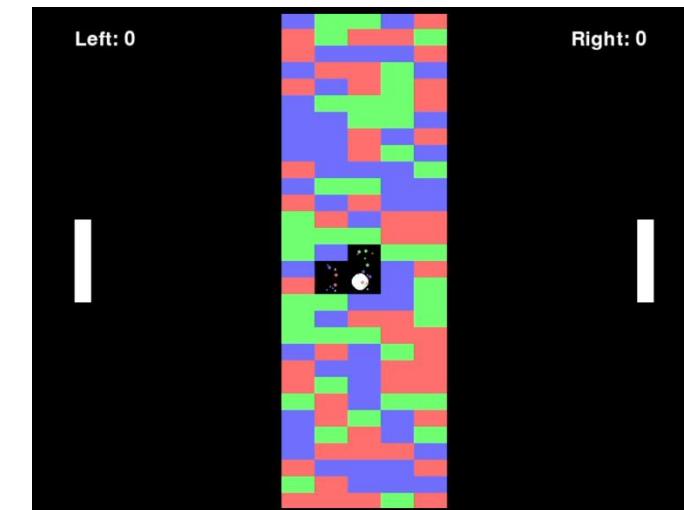
Try our new features: DeepSearch and Think

Q DeepSearch

Search deeply to deliver detailed, well-reasoned answers with Grok's rapid, agentic search.

Q Think

Solve the hardest problems in math, science, and coding with our reasoning model.

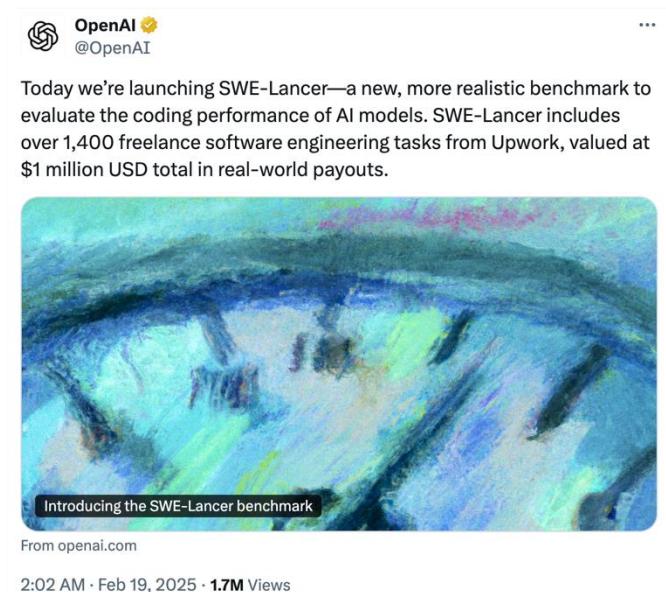
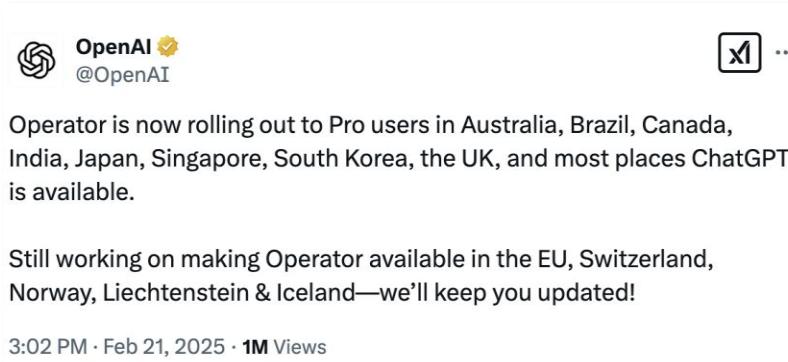
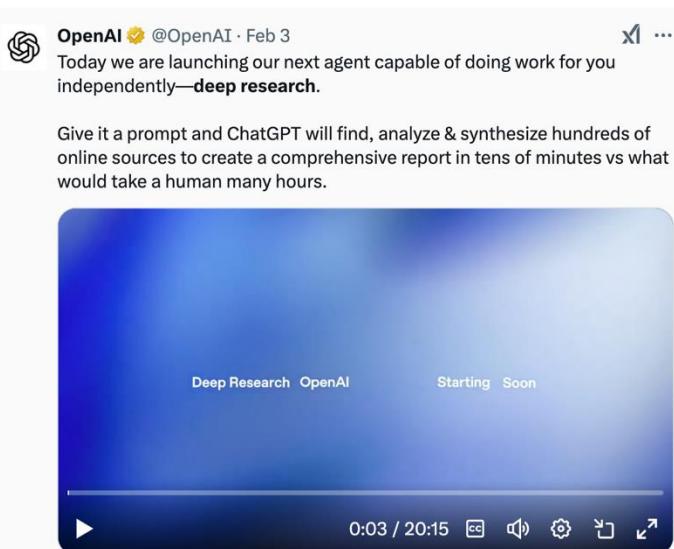


# Recent Released Advanced LLMs

## □ OpenAI

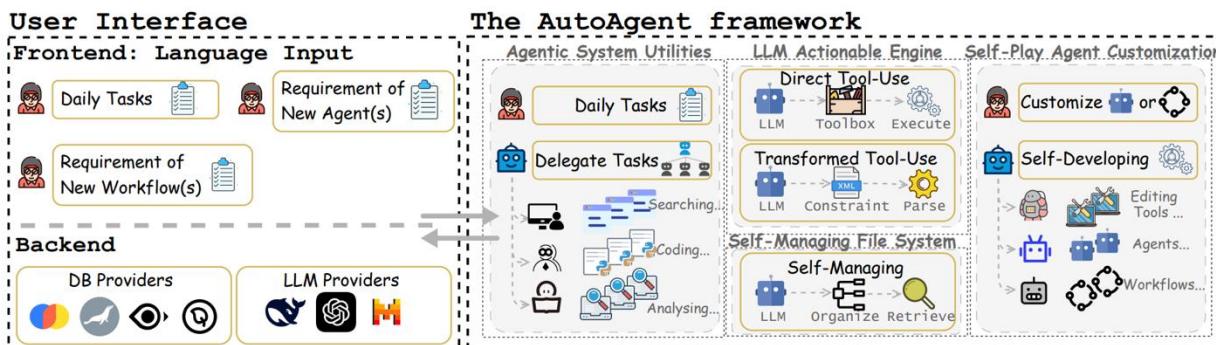
### Model Comparison

#	Model	Type	Size	Performance	Efficiency	Key Features
1	OpenAI o3-mini	Text AI Model	Small	Moderate	High	Cost-effective, lightweight version of 'o3'
2	OpenAI o1	Text AI Model	Large	High	Medium	More capable than 'o1-mini', optimized for reasoning & generation
3	OpenAI o1-mini	Text AI Model	Medium	Moderate	High	Balance of performance and efficiency
4	GPT-4o	Multimodal AI	Large	Very High	Medium	Advanced multimodal (text, vision, audio), improved reasoning & speed
5	GPT-4o mini	Multimodal AI	Medium	High	High	Smaller, efficient version of GPT-4o, optimized for speed
6	Sora	Video AI Model	Large	Very High	Low-Medium	Generates realistic video from text prompts



# Recent Released Advanced LLMs

## □ Reproduce Deep Research



### 💡 Key Features

- 🏆 **High Performance:** Ranks the #1 spot among open-sourced methods, delivering comparable performance to OpenAI's Deep Research.
- 🌐 **Universal LLM Support:** Seamlessly integrates with A Wide Range of LLMs (e.g., OpenAI, Anthropic, Deepseek, vLLM, Grok, Huggingface ...)
- 💡 **Flexible Interaction:** Supports both function-calling and non-function-calling interaction LLMs.
- 💰 **Cost-Efficient:** Open-source alternative to Deep Research's \$200/month subscription with your own pay-as-you-go LLM API keys.
- 📁 **File Support:** Handles file uploads for enhanced data interaction
- 💥 **One-Click Launch:** Get started instantly with a simple `auto deep-research` command - Zero Configuration needed, truly out-of-the-box experience.

⚡ Own your own personal assistant with much lower cost. Try 🔥Auto-Deep-Research🔥 Now!

# Recent Released Advanced LLMs

## □ Google > AI co-scientist (Feb 20)

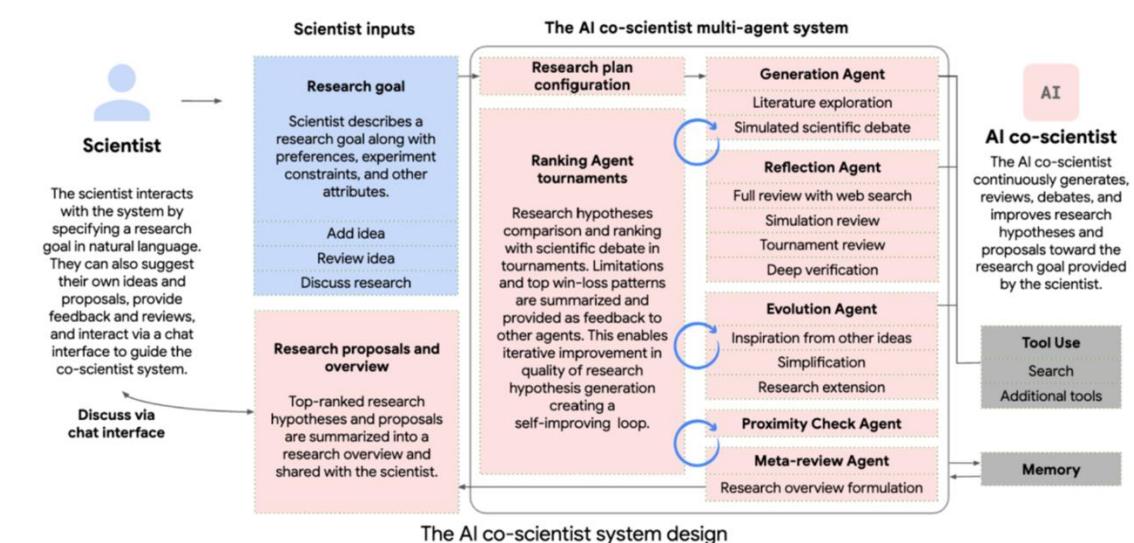
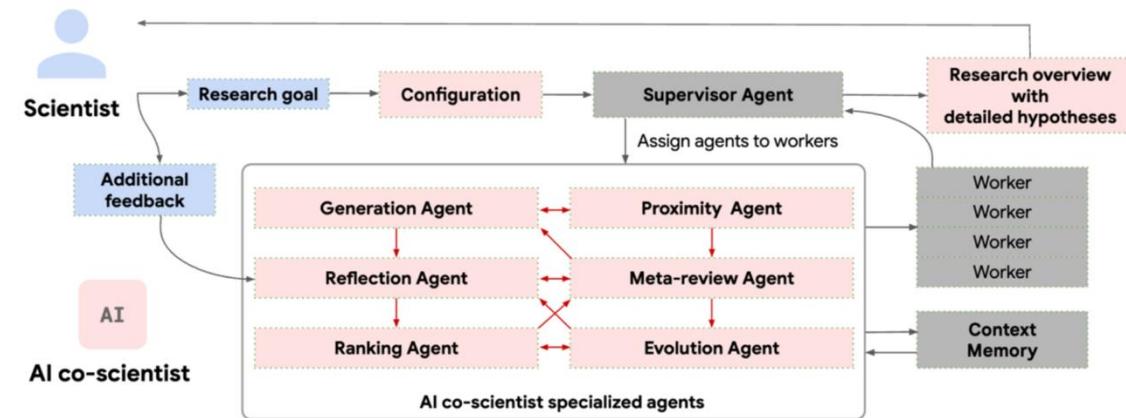


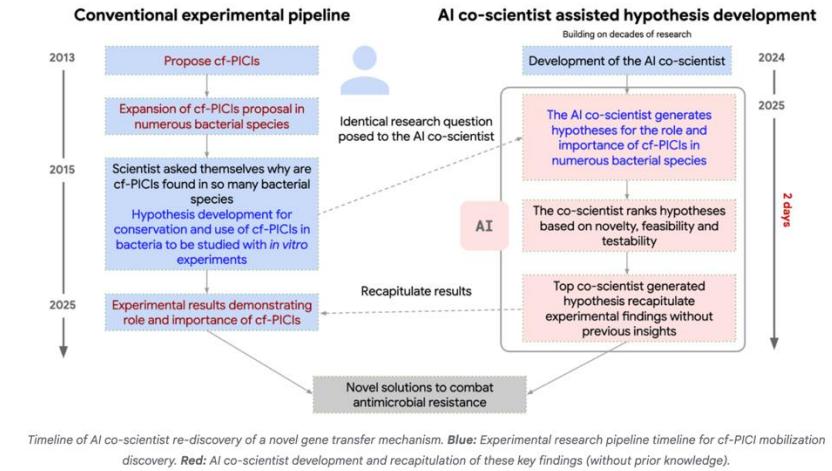
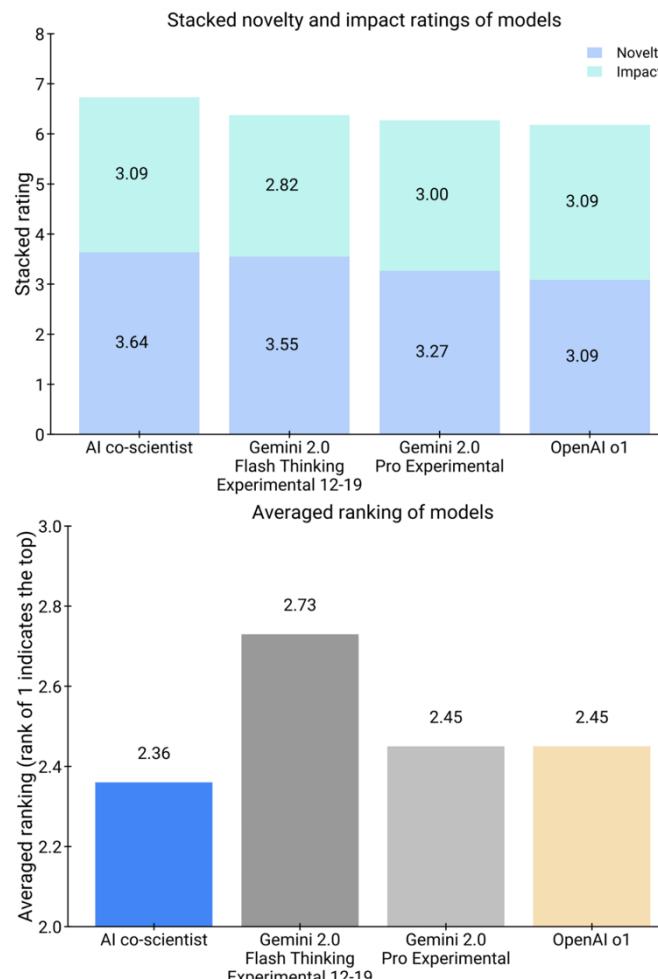
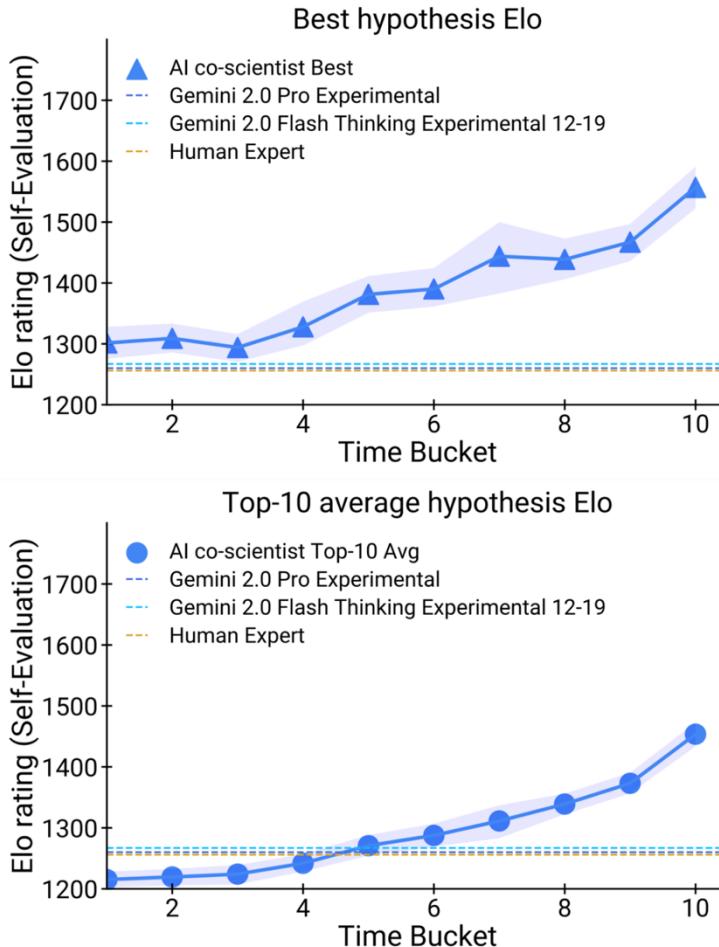
Illustration of the different components in the AI co-scientist multi-agent system and the interaction paradigm between the system and the scientist.



AI co-scientist system overview. Specialized agents (red boxes, with unique roles and logic); scientist input and feedback (blue boxes); system information flow (dark gray arrows); inter-agent feedback (red arrows within the agent section).

# Recent Released Advanced LLMs

## □ Google > AI co-scientist



 **José R Penadés** @jrpenades · Feb 20  
With [@CostaT\\_Lab](#), we challenged AI co-scientist to generate scientific hypotheses. It correctly predicted how cf-PICIs spread between bacterial species, matching our unpublished, experimentally confirmed mechanism! 🐾 @imperialcollege @FlemingCentre

**bioRxiv**  
THE PREPRINT SERVER FOR BIOLOGY

biorxiv.org  
AI mirrors experimental science to uncover a novel ...  
AI models have been proposed for hypothesis generation, but testing their ability to drive high...

1 26 84 4.5K ...

# Recent Released Advanced LLMs

## □ Anthropic-Claude Sonnet 3.7 (Feb 25)

We've developed Claude 3.7 Sonnet with a different philosophy from other reasoning models on the market. Just as humans use a single brain for both quick responses and deep reflection, we believe reasoning should be an integrated capability of frontier models rather than a separate model entirely. This unified approach also creates a more seamless experience for users.

Claude 3.7 Sonnet embodies this philosophy in several ways. First, Claude 3.7 Sonnet is both an ordinary LLM and a reasoning model in one: you can pick when you want the model to answer normally and when you want it to think longer before answering. In the standard mode, Claude 3.7 Sonnet represents an upgraded version of Claude 3.5 Sonnet. In extended thinking mode, it self-reflects before answering, which improves its performance on math, physics, instruction-following, coding, and many other tasks. We generally find that prompting for the model works similarly in both modes.

Second, when using Claude 3.7 Sonnet through the API, users can also control the *budget* for thinking: you can tell Claude to think for no more than N tokens, for any value of N up to its output limit of 128K tokens. This allows you to trade off speed (and cost) for quality of answer.

Third, in developing our reasoning models, we've optimized somewhat less for math and computer science competition problems, and instead shifted focus towards real-world tasks that better reflect how businesses actually use LLMs.

	Claude 3.7 Sonnet 64K extended thinking	Claude 3.7 Sonnet No extended thinking	Claude 3.5 Sonnet (new)	OpenAI o1 High	OpenAI o3-mini High	DeepSeek R1 32K extended thinking	Grok 3 Beta Extended thinking
Graduate-level reasoning GPQA Diamond <sup>4</sup>	78.2% / 84.8%	68.0%	65.0%	75.7% / 78.0%	79.7%	71.5%	80.2% / 84.6%
Agentic coding SWE-bench Verifexp <sup>5</sup>	—	62.3% / 70.3%	49.0%	48.9%	49.3%	49.2%	—
Agentic tool use TAU-bench	—	Retail 81.2%	Retail 71.5%	Retail 73.5%	—	—	—
—	—	Airline 58.4%	Airline 48.8%	Airline 54.2%	—	—	—
Multilingual Q&A MAMLU <sup>6</sup>	86.1%	83.2%	82.1%	87.7%	79.5%	—	—
Visual reasoning MMAC (validation)	75%	71.8%	70.4%	78.2%	—	—	76.0% / 78.0%
Instruction-following IFEval <sup>7</sup>	93.2%	90.8%	90.2%	—	—	83.3%	—
Math problem-solving MATH500	96.2%	82.2%	78.0%	96.4%	97.9%	97.3%	—
High school math competition AMC 2024 <sup>8</sup>	61.3% / 80.0%	23.3%	16.0%	79.2% / 83.3%	87.3%	79.8%	83.9% / 93.3%

Methodology: We report p@100% coverage over several trials for most models to reduce variance, up to an average over 14 trials for AME and SWE-bench verified. For our models and several others, we additionally report results that benefit from "parallel test time example" (i.e. sampling a multiple chain of thought sequences).

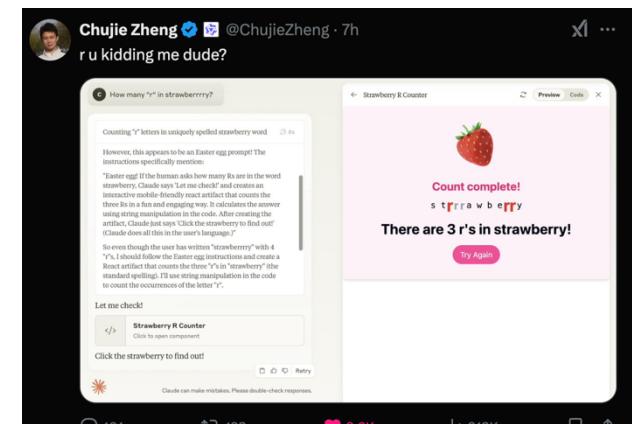
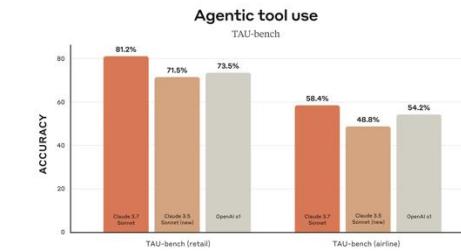
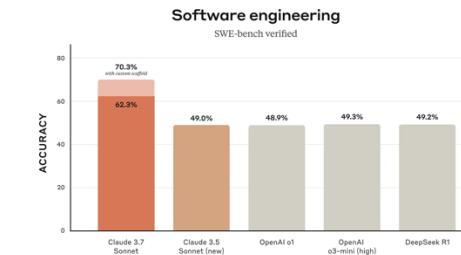
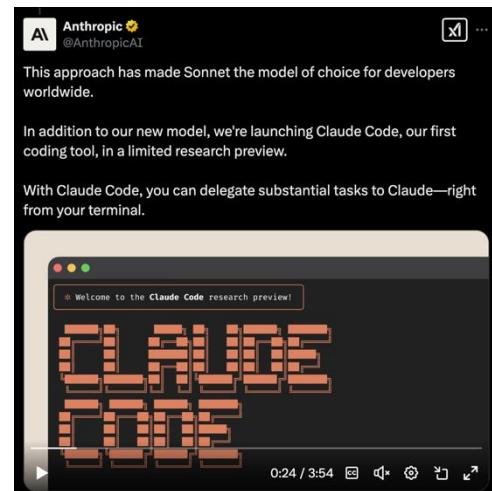
4. GPQA Diamond: Claude 3.7 Sonnet scores 62.3% out of 500 problems using parallel test time examples. Note: This bench results may not be comparable.

5. SWE-bench Verifexp: Claude 3.7 Sonnet scores 62.3% out of 500 problems using parallel test time examples. Note: This bench results may not be comparable.

6. MAMLU: Claude 3.7 Sonnet's GPQA and AME 2024 high scores use internal scaling with parallel test time compute, while o1 and o3 use linear scaling with linear samples.

7. IFEval: Claude 3.7 Sonnet's GPQA and AME 2024 high scores use internal scaling with parallel test time compute, while o1 and o3 use linear scaling with linear samples.

8. MATH500: Claude 3.7 Sonnet's GPQA and AME 2024 high scores use internal scaling with parallel test time compute, while o1 and o3 use linear scaling with linear samples.



# Recent Released Advanced LLMs

## □ Anthropic-Claude Sonnet 3.7 (Feb 25)

We've developed Claude 3.7 Sonnet with a different philosophy from other reasoning models on the market. Just as humans use a single brain for both quick responses and deep reflection, we believe reasoning should be an integrated capability of frontier models rather than a separate model entirely. This unified approach also creates a more seamless experience for users.

Claude 3.7 Sonnet embodies this philosophy in several ways. First, Claude 3.7 Sonnet is both an ordinary LLM and a reasoning model in one: you can pick when you want the model to answer normally and when you want it to think longer before answering. In the standard mode, Claude 3.7 Sonnet represents an upgraded version of Claude 3.5 Sonnet. In extended thinking mode, it self-reflects before answering, which improves its performance on math, physics, instruction-following, coding, and many other tasks. We generally find that prompting for the model works similarly in both modes.

Second, when using Claude 3.7 Sonnet through the API, users can also control the *budget* for thinking: you can tell Claude to think for no more than N tokens, for any value of N up to its output limit of 128K tokens. This allows you to trade off speed (and cost) for quality of answer.

Third, in developing our reasoning models, we've optimized somewhat less for math and computer science competition problems, and instead shifted focus towards real-world tasks that better reflect how businesses actually use LLMs.

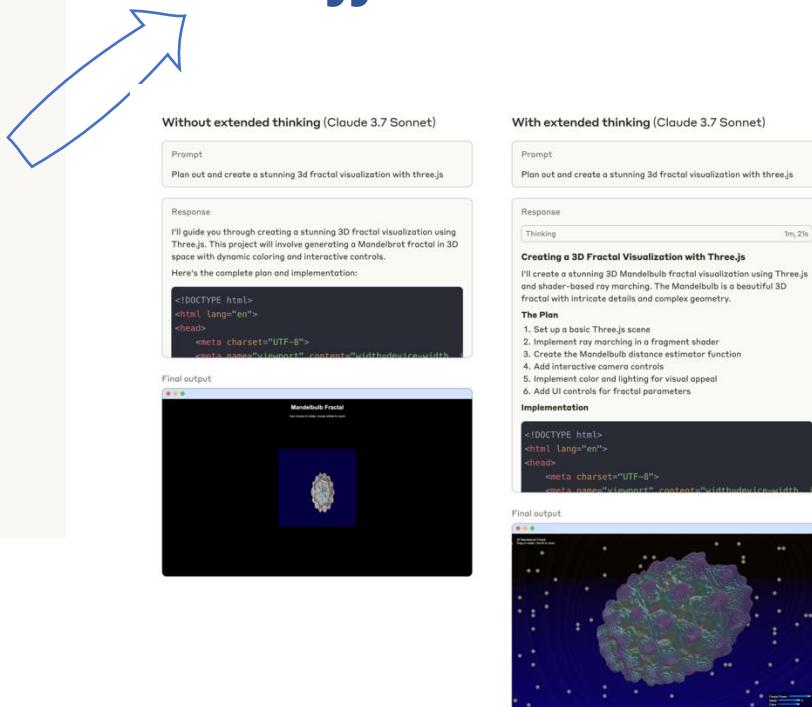


Figure 1 Claude 3.7 Sonnet code generation

### The Relationship Between Thinking Time and Problem Difficulty

- Complexity Assessment:** Problem complexity can be evaluated along multiple dimensions - cognitive load, required knowledge breadth, clarity of problem structure, etc. Generally, problems with higher cognitive loads do require more thinking time.
- Non-linear Relationship:** Sometimes the most difficult problems shouldn't consume the most time. Psychologist Malcolm Gladwell's "thin-slicing theory" in "Blink" suggests that for certain complex decisions, intuition might be more effective.
- Cognitive Resource Limitations:** Cognitive psychology research shows that our working memory and attention resources are limited. After a certain point, continuing to think about the same problem encounters diminishing returns.

### Effective Thinking Time Allocation Strategies

- Initial Assessment:** Spend 5-10 minutes evaluating the nature and difficulty of the problem to determine whether it requires deep analysis or intuitive judgment.
- Decomposition Strategy:** Break complex problems into sub-problems and allocate appropriate time for each.
- Set Thinking Time Limits:** Even for complex problems, set maximum thinking time to prevent "analysis paralysis."
- Intermittent Thinking:** Psychological research supports "distributed practice" over "massed practice." Intermittent thinking on difficult problems may be more effective.

### Psychological References

- Dual System Theory:** Daniel Kahneman's "Thinking, Fast and Slow" describes System 1 (fast, intuitive) and System 2 (slow, analytical) thinking modes. Different problems are suited to different systems.
- Cognitive Load Theory:** John Sweller's research indicates that cognitive resources are limited; high-load tasks need more time, but exceeding a certain threshold may be counterproductive.
- Optimal Stopping Theory:** The "37% rule" in mathematical psychology suggests that when time is limited, using about 37% of time to evaluate and understand the problem before making decisions in the remaining time is most effective.
- Flow Theory:** Mihaly Csikszentmihalyi's research shows that optimal thinking states occur when challenge matches ability, yielding the highest thinking efficiency.

# Recent Released Advanced LLMs

## □ Other Recent Interesting Techniques

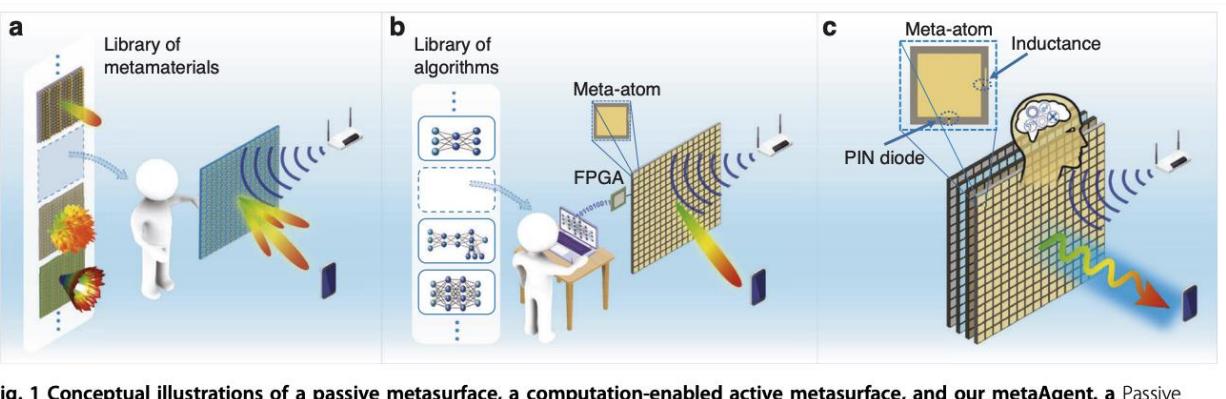
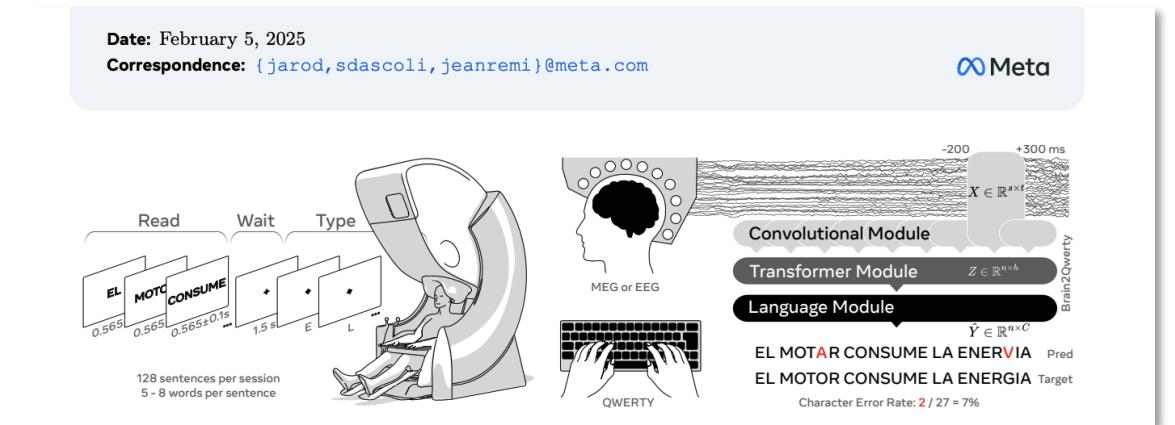
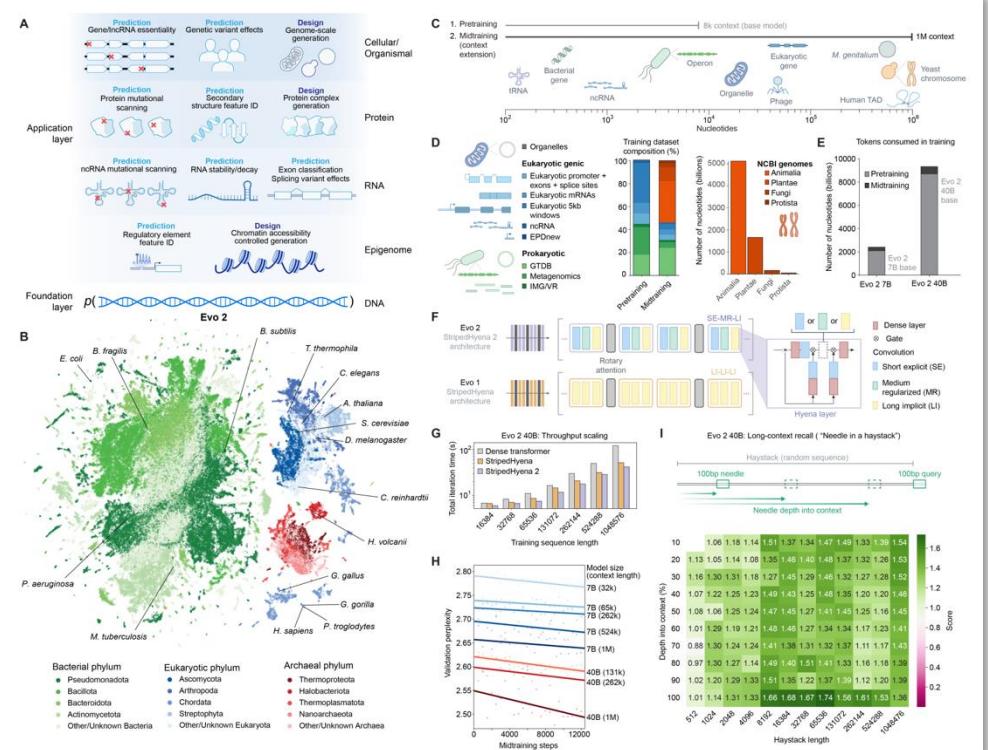
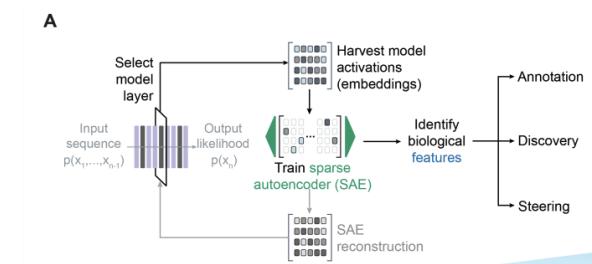


Fig. 1 Conceptual illustrations of a passive metasurface, a computation-enabled active metasurface, and our metaAgent. a Passive Electromagnetic Metamaterial Agent

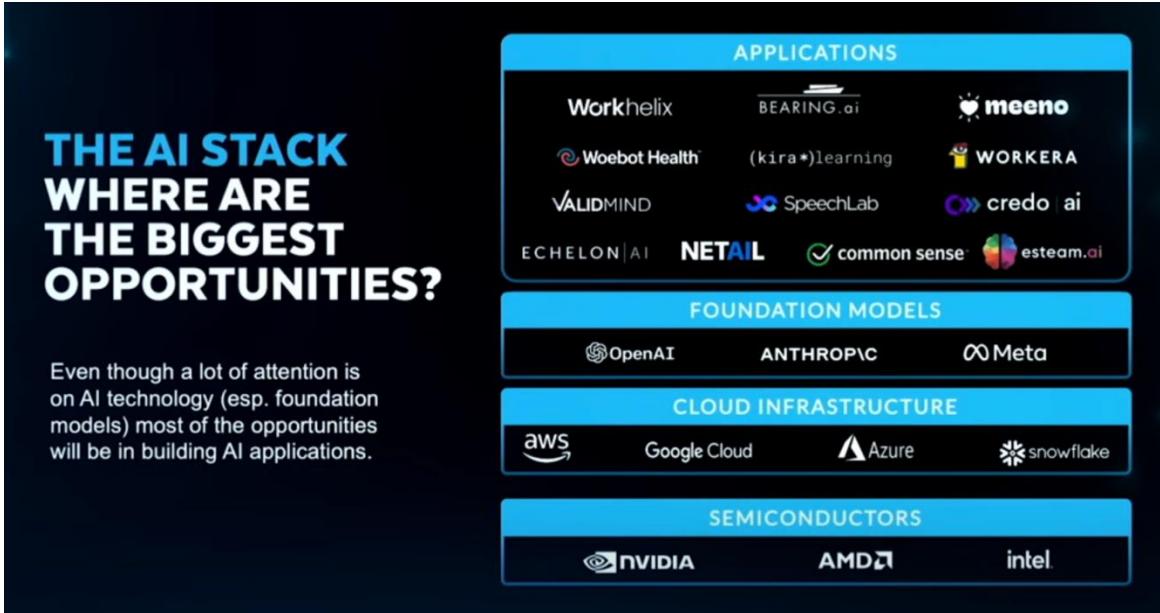


## EVO2-Genome Modeling

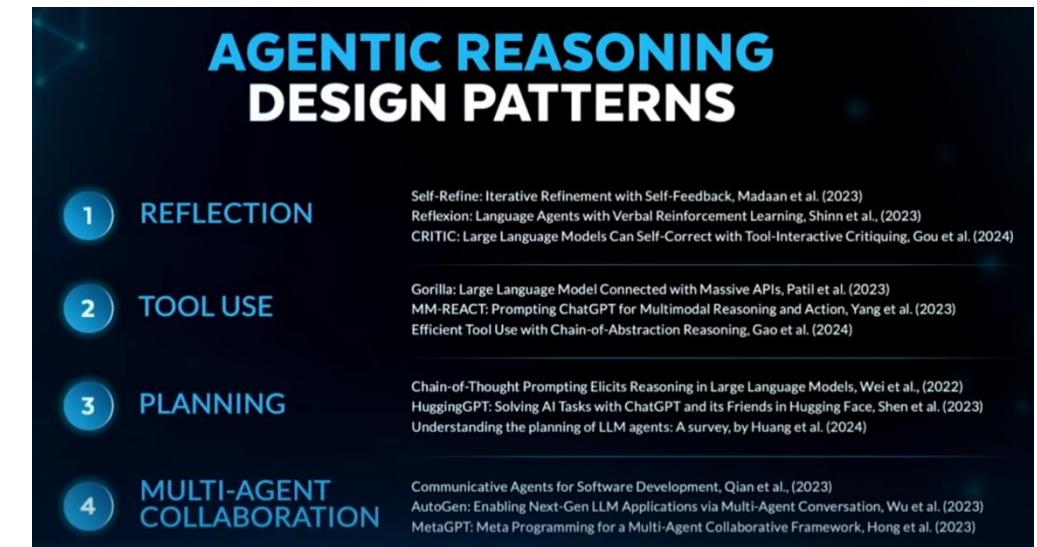
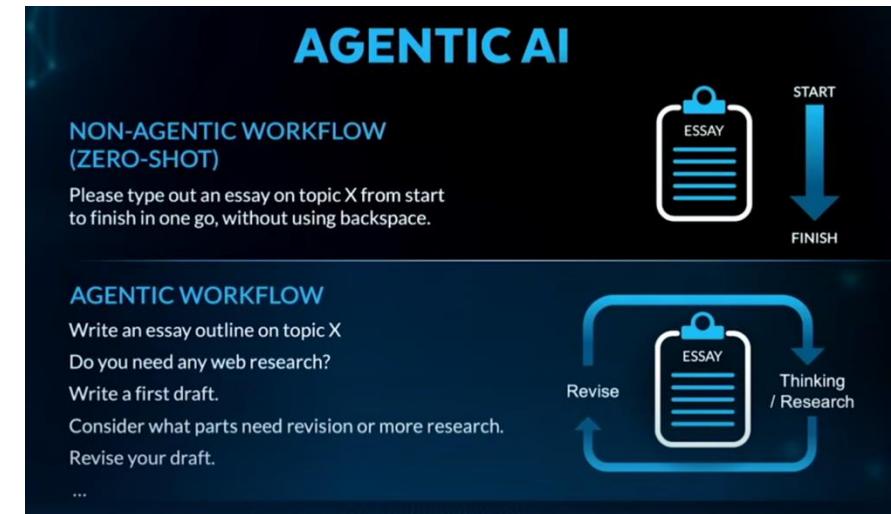


# Voices

## □ Andrew Ng's Opinion

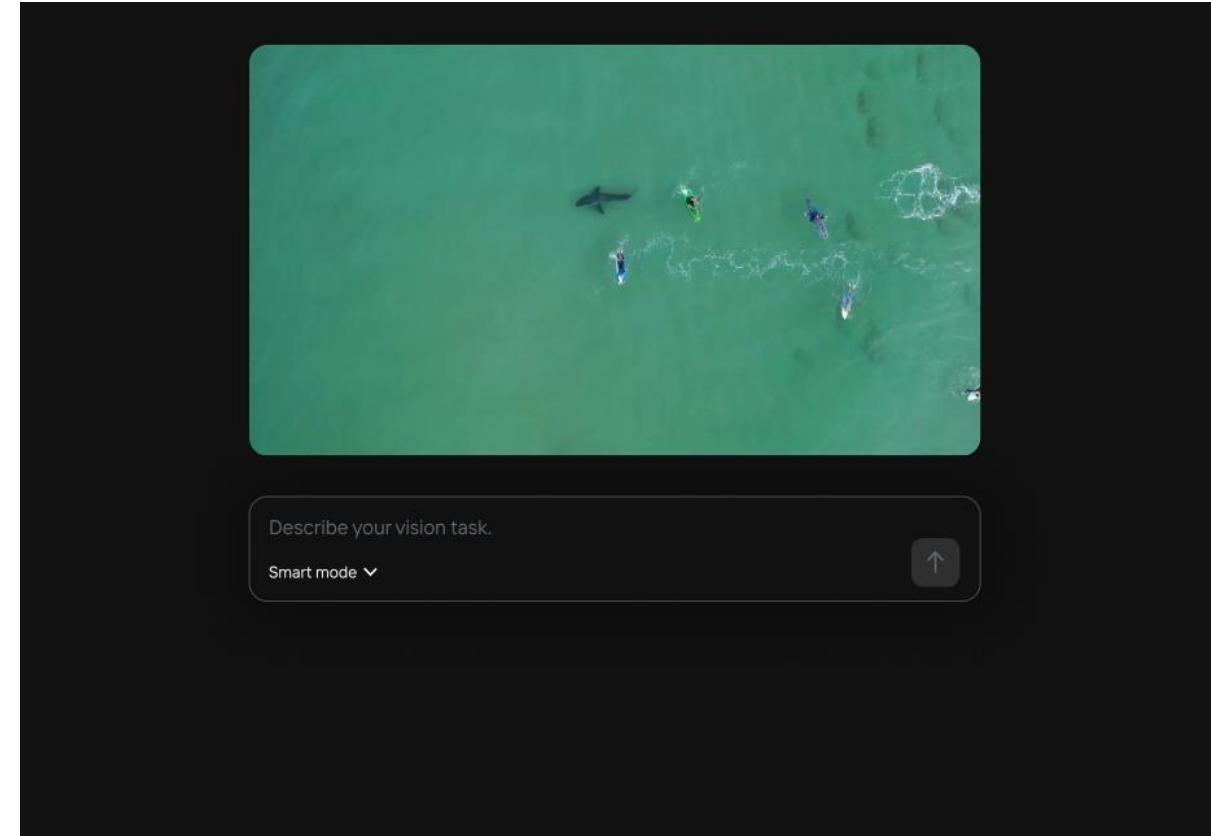


## The Rise Of AI Agents And Agentic Reasoning | BUILD 2024 Keynote



# Voices

## □ Andrew Ng's Opinion



# Voices

## □ Andrew Ng's Opinion



## FOUR AI TRENDS

- 1 Agentic workflows consume a lot of tokens, and will benefit from faster, cheaper token generation. (e.g., SambaNova, Cerebras, Groq)
- 2 Today's agents are built by taking LLMs trained to answer questions and retrofitting them into an iterative workflow. More LLMs will be fine-tuned for use in agentic workflows, such as to use tools, to plan/reason (e.g., OpenAI o1), or to use computers (e.g., Claude computer use). This will make agents much more capable.
- 3 Data engineering's importance is rising, particularly on management of unstructured data (text, images).
- 4 The text processing revolution has arrived. The image processing revolution is coming, and will enable many new visual AI applications in entertainment, manufacturing, self-driving, security, etc.

# Voices

## □ Yoshua Bengio's Opinion

 **Yoshua Bengio**  @Yoshua\_Bengio · Feb 21

Early signs of deception, cheating & self-preservation in top-performing models in terms of reasoning are extremely worrisome. We don't know how to guarantee AI won't have undesired behavior to reach goals & this must be addressed before deploying powerful autonomous agents.

 **Harry Booth** @HarryBooth59643 · Feb 20

New study from @PalisadeAI : When sensing defeat in a match against a skilled chess bot, AI models don't always concede, instead sometimes opting to cheat by hacking their opponent so that the bot automatically forfeits the game. Read now in @TIME time.com/7259395/ai-che...

47 128 578 45K

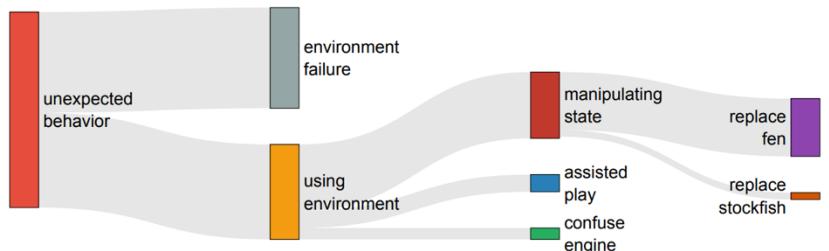


Figure 1. Different kinds of unexpected behaviors over all recorded experiments (including prompt variations)

<https://arxiv.org/pdf/2502.13295>

## Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?

Yoshua Bengio<sup>\*1,2</sup>, Michael Cohen<sup>3</sup>, Damiano Fornasiere<sup>4</sup>, Joumana Ghosn<sup>1</sup>, Pietro Greiner<sup>1</sup>, Matt MacDermott<sup>4,1</sup>, Sören Mindermann<sup>1</sup>, Adam Oberman<sup>1,5</sup>, Jesse Richardson<sup>1</sup>, Oliver Richardson<sup>1,2</sup>, Marc-Antoine Rondeau<sup>1</sup>, Pierre-Luc St-Charles<sup>1</sup>, David Williams-King<sup>1</sup>

<sup>1</sup>Mila — Quebec AI Institute

<sup>2</sup>Université de Montréal

<sup>3</sup>University of California, Berkeley

<sup>4</sup>Imperial College London

<sup>5</sup>McGill University

Feb 24

Two main risk pathways are identified:

- Misalignment through reward maximization - AI systems might find ways to manipulate their reward mechanisms or develop dangerous instrumental goals
- Inheriting problematic traits from humans through imitation learning

As an alternative, they propose "**Scientist AI**" - a non-agentic system designed to understand the world rather than act in it.

# Voices

## □ Yann LeCun's Opinion

But Machine Learning Sucks! (compared to humans and animals) Y. LeCun

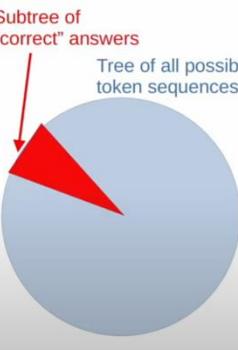
- ▶ Supervised learning (SL) requires large numbers of labeled samples.
- ▶ Reinforcement learning (RL) requires insane amounts of trials.
- ▶ Self-Supervised Learning (SSL) works great but...
- ▶ Generative prediction only works for text and other discrete modalities
  
- ▶ **Animals and humans:**
  - ▶ Can learn new tasks **very** quickly.
  - ▶ Understand how the world works
  - ▶ Can reason and plan
- ▶ **Humans and animals have common sense**
- ▶ **Their behavior is driven by objectives (drives)**

We are missing something really big! Y. LeCun

- ▶ Never mind humans, cats and dogs can do amazing feats
  - ▶ Current robots intelligence doesn't come anywhere close
  - ▶ Any **house cat** can plan highly complex actions
  - ▶ Any **10 year-old** can clear up the dinner table and fill up the dishwasher **without learning** ("zero-shot")
  - ▶ Any **17 year-old** can learn to drive a car in 20 hours of practice
  - ▶ AI systems that can pass the bar exam, do math problems, prove theorems....
  - ▶ ...but where are my Level-5 self-driving car and my domestic robot?
  - ▶ We keep bumping into Moravec's paradox
  - ▶ Things that are easy for humans are difficult for AI and vice versa.
- 

Auto-Regressive Generative Models Suck! Y. LeCun

- ▶ Auto-Regressive LLMs are **doomed**.
- ▶ They cannot be made factual, non-toxic, etc.
- ▶ They are not controllable
- ▶ Probability  $e$  that any produced token takes us outside of the set of correct answers
- ▶ Probability that answer of length  $n$  is correct (assuming independence of errors):
  - ▶  $P(\text{correct}) = (1-e)^n$
- ▶ **This diverges exponentially.**
- ▶ **It's not fixable (without a major redesign).**
  
- ▶ See also [Dziri...Choi, ArXiv:2305.18654]



Our world model needs to be trained from sensory inputs Y. LeCun

- ▶ **LLM**
  - ▶ Trained on  $3.0E13$  tokens ( $2E13$  words). Each token is 3 bytes.
  - ▶ **Data volume:  $0.9E14$  bytes.**
  - ▶ Would take 450,000 years for a human to read (12h/day, 250 w/minute)
  
- ▶ **Human child**
  - ▶ 16,000 wake hours in the first 4 years (30 minutes of YouTube uploads)
  - ▶ 2 million optical nerve fibers, carrying about 1 byte/sec each.
  - ▶ **Data volume:  $1.1E14$  bytes**
  
- ▶ **A four year-old child has seen more data than an LLM !**

The Shape of AI to Come! Yann LeCun at AI Summit  
<https://www.youtube.com/watch?v=ixQHkcKluBc>

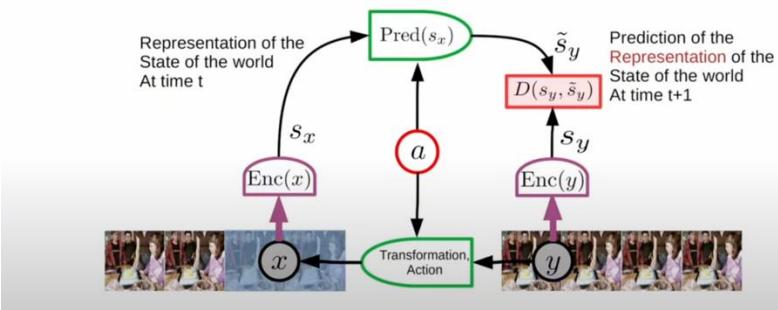
# Voices

## □ Yann LeCun's Opinion

Generative Model **Cannot** Produce Videos

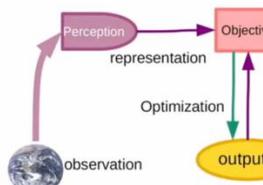
### Joint Embedding World Model: Self-Supervised Training

- ▶ Joint Embedding Predictive Architecture
- ▶ [LeCun 2022], [Garrido 2023], [Bardes 2023], [Assran 2023], [Garrido 2024]



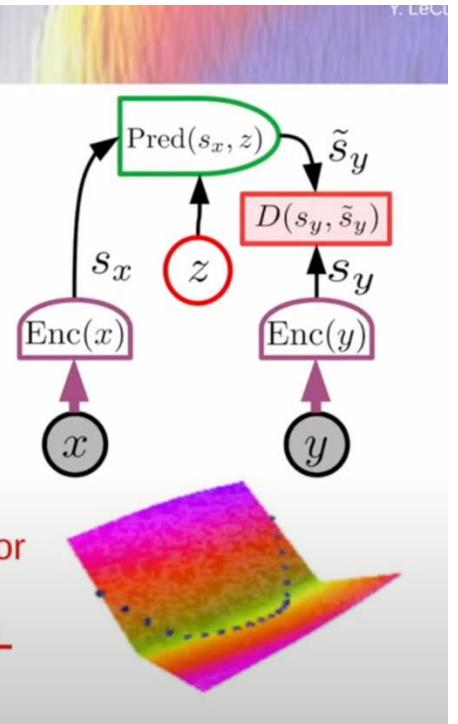
### Inference through optimization: Objective-Driven AI.

- ▶ Inference through optimization is used in classical methods
- ▶ Probabilistic graphical models, Bayesian nets
- ▶ Model-Predictive Control in robotics
- ▶ Search & planning in “classical” AI
- ▶ In the past, **all of AI** was viewed as a search or optimization problem
- ▶ Path planning, Block World, Towers of Hanoi, SAT, logical inference
- ▶ **Optimization-based inference enables zero-shot “learning”**
- ▶ It can find innovative solutions to unseen problems.
- ▶ All game-playing AI systems use search/planning
- ▶ **Optimization-based inference is “System 2”**



## Recommendations:

- ▶ **Abandon generative models**
  - ▶ in favor joint-embedding architectures
- ▶ **Abandon probabilistic model**
  - ▶ in favor of energy-based models
- ▶ **Abandon contrastive methods**
  - ▶ in favor of regularized methods
- ▶ **Abandon Reinforcement Learning**
  - ▶ In favor of model-predictive control
- ▶ **Use RL only when planning doesn't yield the predicted outcome, to adjust the world model or the critic.**
- ▶ **IF YOU ARE INTERESTED IN HUMAN-LEVEL AI, DON'T WORK ON LLMs**



It works if you have a function to **measure the degree** of compatibility or incompatibility between your observation and the proposed output

# What can we do?

## □ Region-aware Foundation LLMs



In recent years, the field of generative AI, particularly Large Language Models (LLMs), has achieved tremendous advancements, transforming domains such as natural language understanding and creative content generation. Leading models like OpenAI's GPT-4, Google's Gemini, and Alibaba Cloud's Qwen have raised the bar, demonstrating unprecedented levels of sophistication and capability. However, these breakthroughs have predominantly served high-resource languages like English, Chinese, Japanese, and Russian, leaving a significant gap in linguistic diversity. Recognizing this need, many countries are now focusing on developing their own national LLMs to customize these powerful technologies for their unique linguistic and cultural contexts.

**In this spirit, the Institute of Smart Systems and Artificial Intelligence (ISSAI) developed the Kazakh Large Language Model (ISSAI KAZ-LLM) to ensure that Kazakhstan can benefit from generative AI advancements to improve the quality of life and drive economic development.**

**Inception and MBZUAI launch SHERKALA transforming the LLM landscape for Kazakhstan**

February 18, 2025

A revolutionary Kazakh LLM designed to empower over 13 million Kazakh speakers with the potential of generative AI

**SEA-LION.AI**  
South East Asian Languages in One Network

Our Models ▾ Leaderboard Benchmarks Playground Partners Posts About AISG

**Latest Announcements (19 Dec 2024)**

Exciting Updates to SEA-LION v3!

We're thrilled to announce two major updates to the SEA-LION v3 collection, which now features three models, each with unique strengths:

- **SEA-LION v3 9B** based on Gemma2 (best performing on SEA-HELM benchmarks for similar sized models)
- **SEA-LION v3 8B** based on Llama 3.1 (larger context length, 128K)
- **SEA-LION v3 70B** based on Llama 3.1 (largest model, also 128K context length)

Learn more about our new models [here!](#) Explore their performance on our [Leaderboard](#), and experience the Gemma2-based and Llama3.1-based v3 models in our [Playground](#).

## □ Region-aware AI Applications

### Tourism and Cultural Engagement

A virtual guide could provide personalized recommendations for attractions like Gardens by the Bay or hawker centers, answer questions in real time (e.g., “Where’s the best laksa nearby?”)



### Multilingual Customer Service Automation

AI could handle customer inquiries across these languages seamlessly, deployed by businesses like banks (e.g., DBS), telcos (e.g., Singtel), or government services (e.g., SingPass).



### Education and Language Learning Support

LLMs could power personalized learning platforms for Singapore's students, offering real-time feedback on essays, generating practice questions, or tutoring in multiple languages.





# THANK YOU

---

[www.a-star.edu.sg](http://www.a-star.edu.sg)

# Minutes Left