

Recent Tendency of LLMs Development



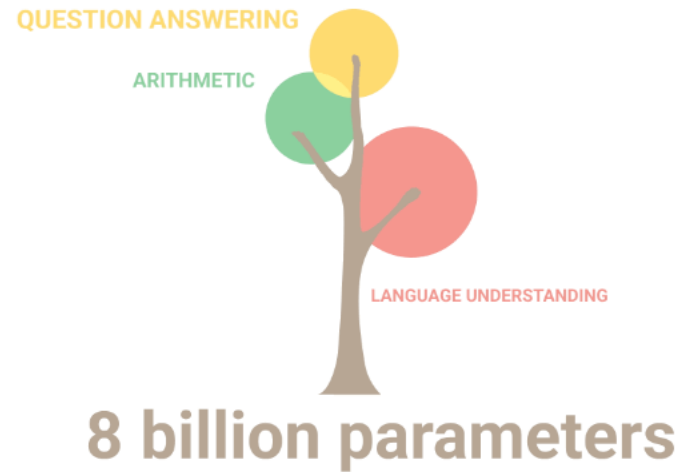
Zhang Jie, Scientist, CFAR, A*STAR

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 Shanghai Artificial Intelligence Laboratory	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 UNIVERSITY OF CALIFORNIA	7B	RLAIF, PPO
Athene-70B (Nexusflow, 2024)	Nexusflow	70B	RLHF
Hermes 3 (Teknium et al., 2024)	NOUS RESEARCH	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.

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	✓ ¹	✗	✓
Marco-o1 (Zhao et al., 2024a)	Alibaba	7B	✓ ²	✓ ⁸	✗
Skywork o1 (o1 Team, 2024)	KUNLUN www.kunlun.com	8B	✓ ³	✗	✗
QwQ-32B-Preview (QwenLM, QwQ)	Alibaba	32B	✓ ⁴	✗	✓
o1-Coder (Zhang et al., 2024d)	北京交通大学 BEIJING JIAOTONG UNIVERSITY	-	✓ ⁵	✓ ⁹	✗
rStar-Math (Guan et al., 2025)	Microsoft	1.5B,3B,7B	✓ ⁶	✓ ¹⁰	✓
Kimi-k1.5 (Team et al., 2025)	Moonshot AI	-	✗	✓ ¹¹	✓
DeepSeek-R1 (DeepSeek-AI et al., 2025)	deepseek	671B-A31B	✓ ⁷	✓ ¹²	✓

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

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

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

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

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

⁶ <https://github.com/zhentingqi/rStar>

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

⁸ <https://arxiv.org/pdf/2501.04519>

⁹ <https://arxiv.org/pdf/2411.14405>

¹⁰ <https://arxiv.org/pdf/2412.00154>

¹¹ <https://arxiv.org/pdf/2501.12599>

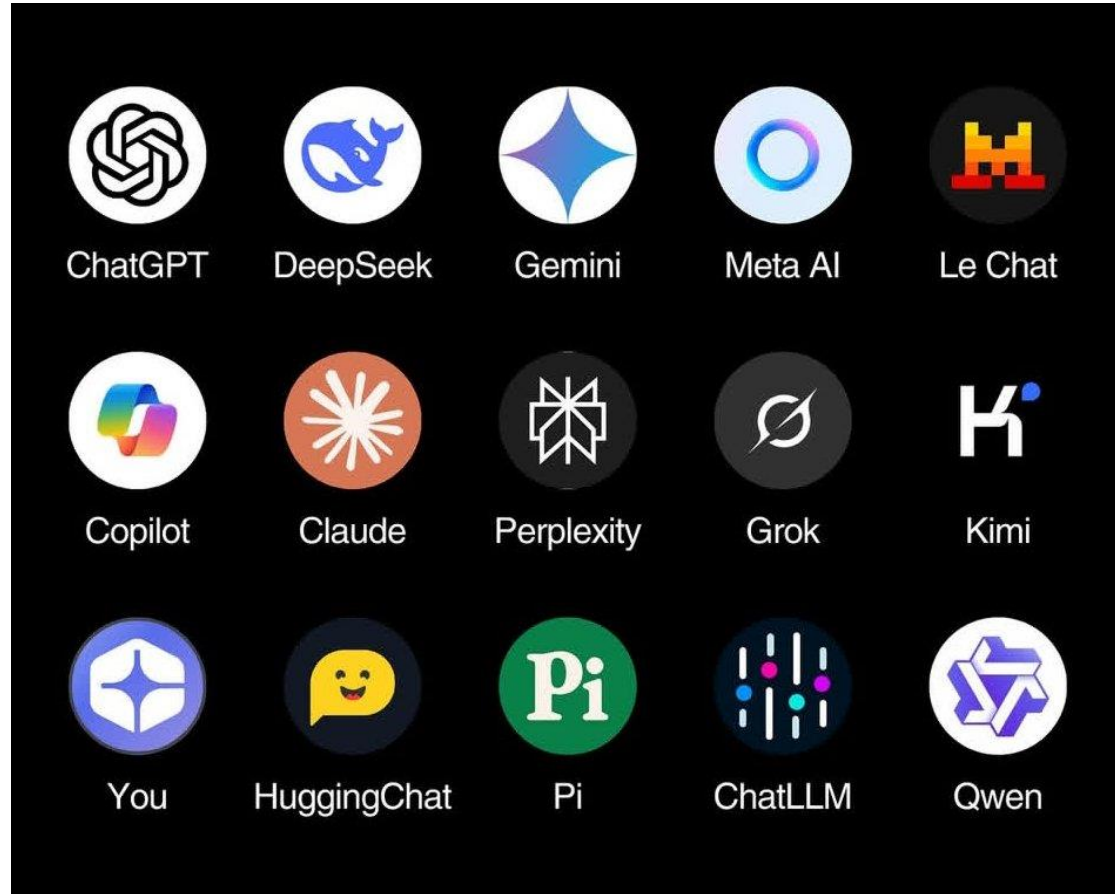
¹² <https://arxiv.org/pdf/2501.12948>

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

Reinforcement Learning Enhanced LLMs: A Survey

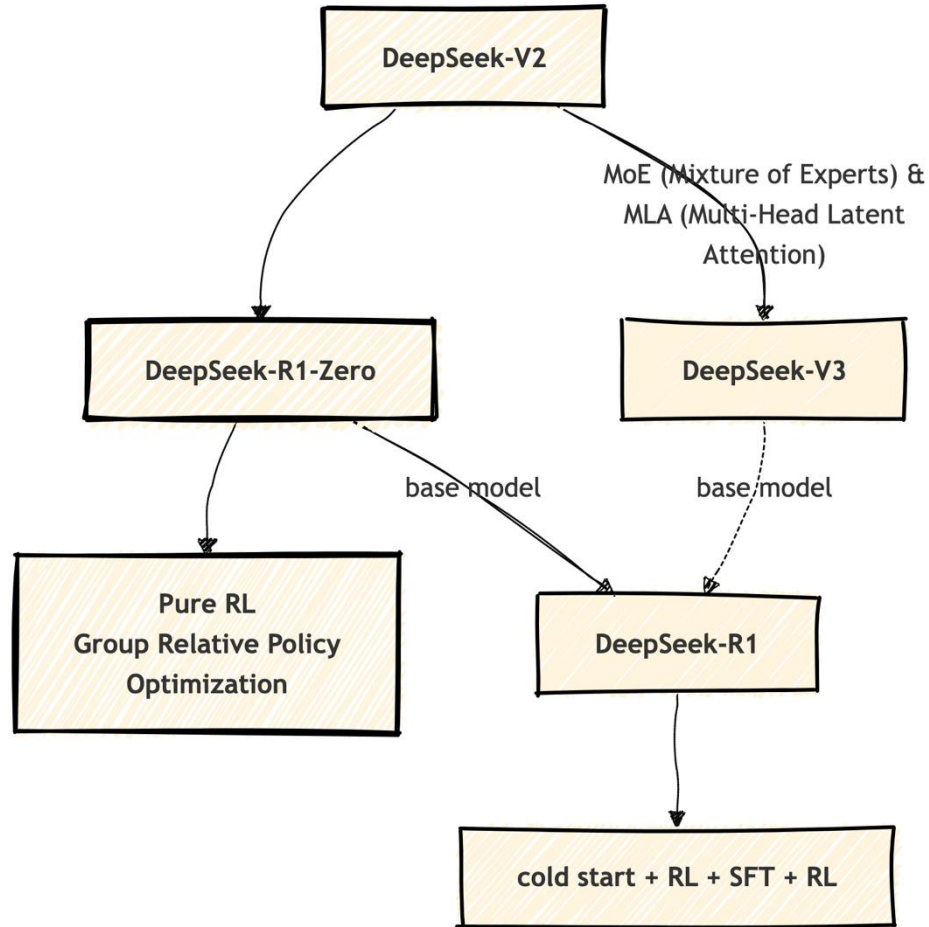
LLMs have taken the Whole World by storm

□ Diverse LLMs You Can Use



Recent Released Advanced LLMs (Q1 2025)

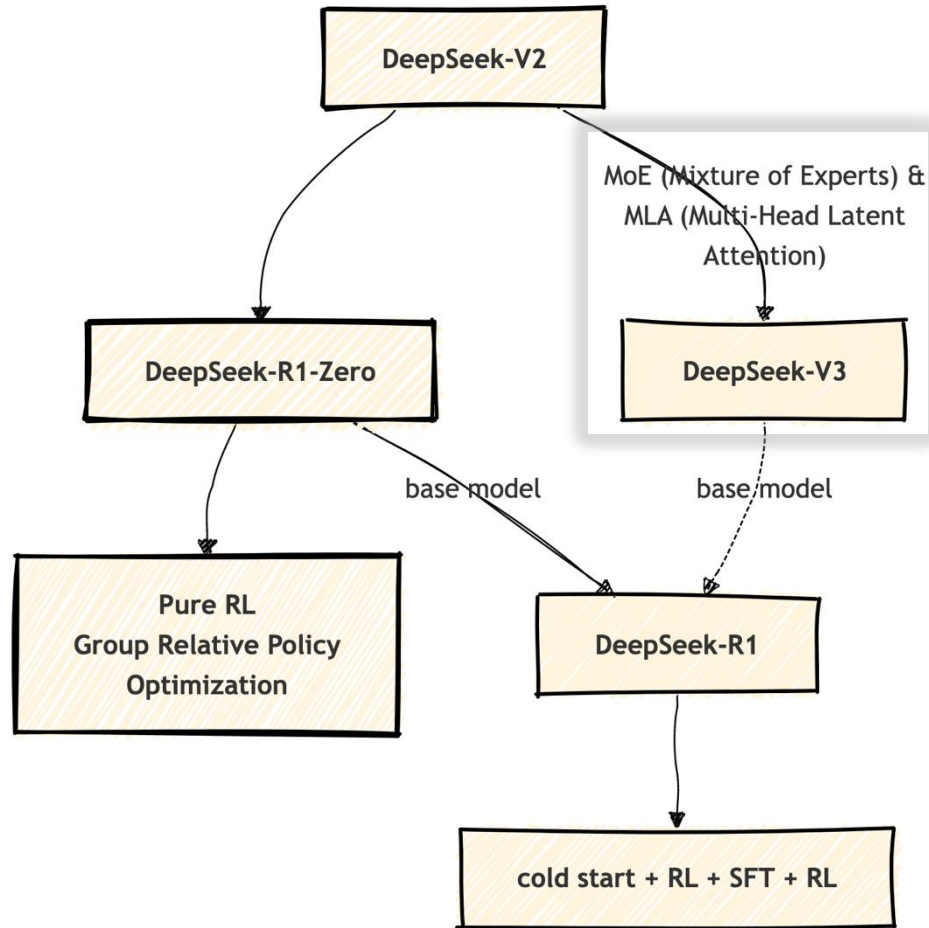
□ DeepSeek 



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	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
English 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
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
Code 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
AIME 2024 (Pass@1)	39.2	23.3	23.3	16	9.3
Math 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://x.com/itsPaulAi/status/1872320003770618146>

Training Cost Comparison

(DeepSeek V3 vs LLaMA 3)

1/100

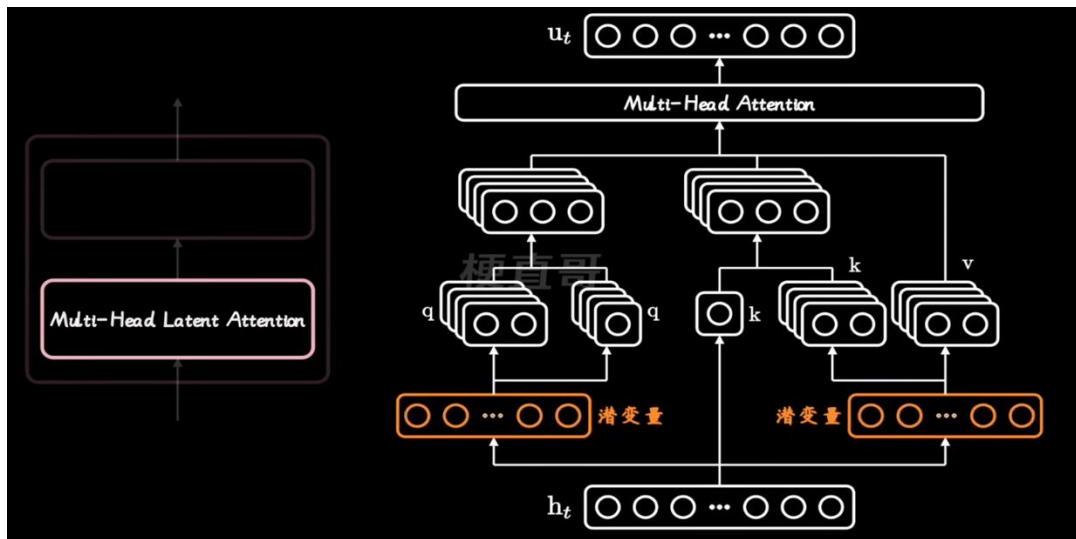
Inference Cost Comparison

(DeepSeek V3 vs OpenAI o1)

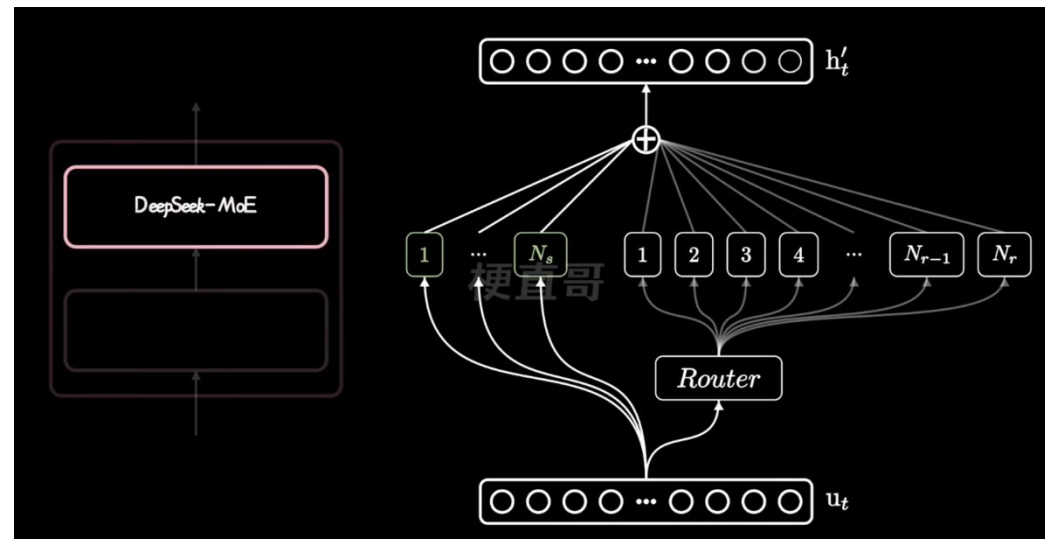
1/30

Recent Released Advanced LLMs (Q1 2025)

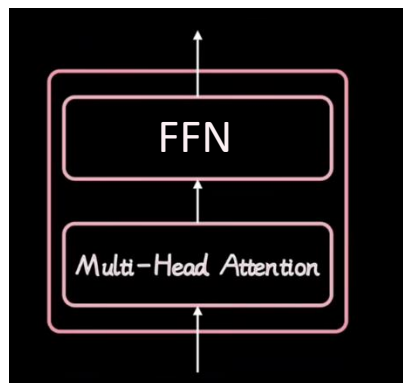
□ DeepSeek V2&V3



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



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



Multiple Token Prediction

DeepSeek is



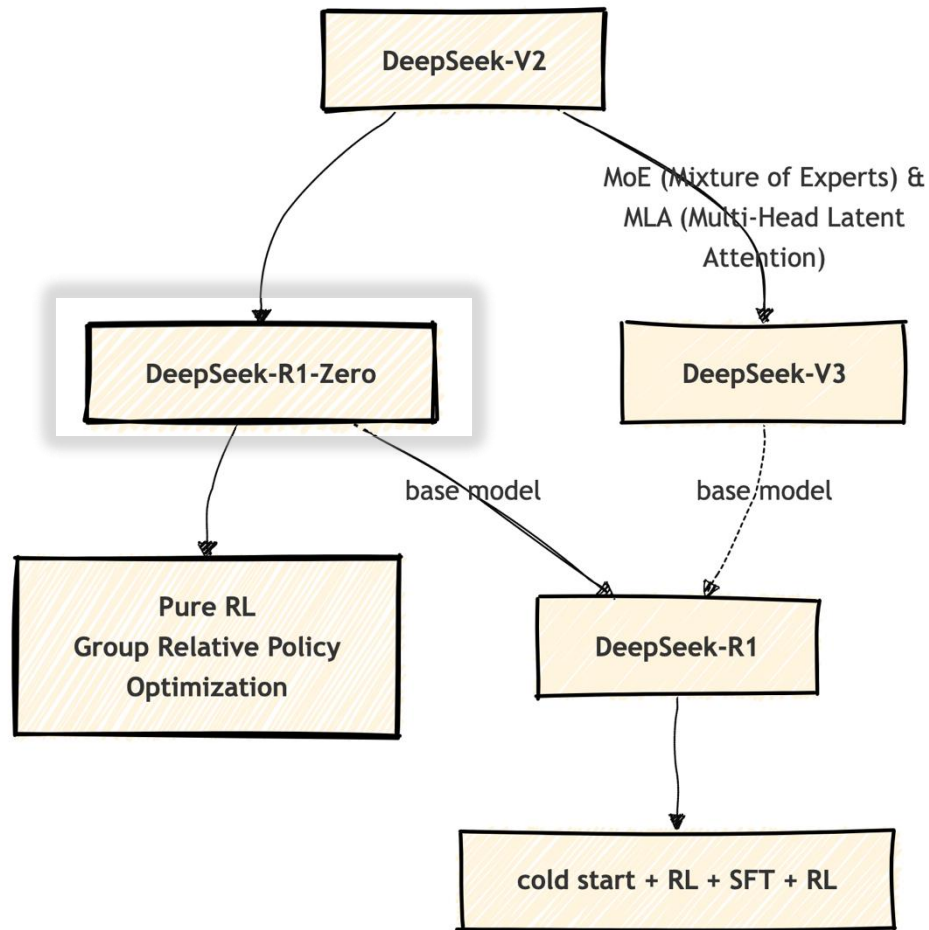
amazing and fast

powerful but complex

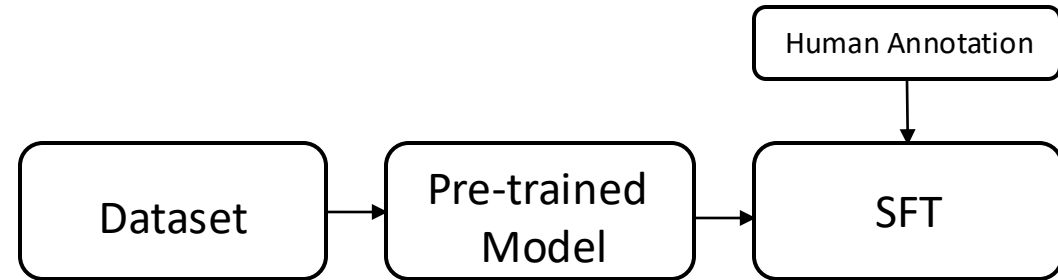
incredible for AI research

Recent Released Advanced LLMs (Q1 2025)

□ DeepSeek 



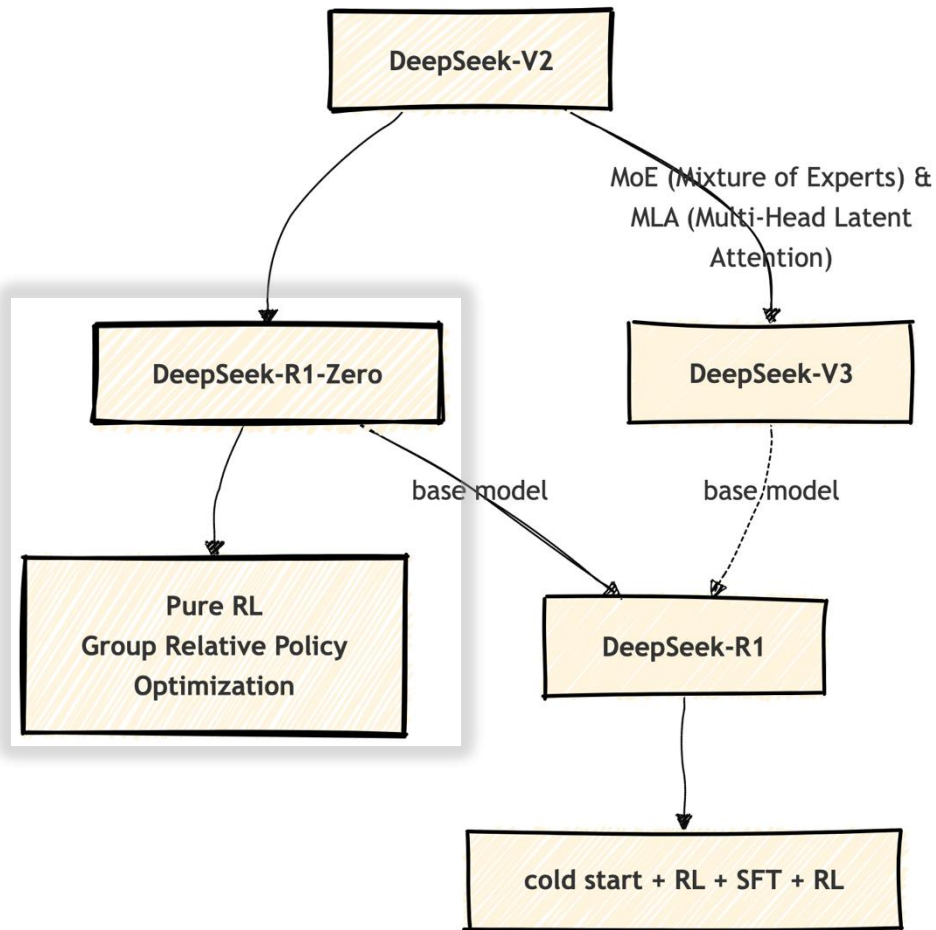
DeepSeek Evolution Process



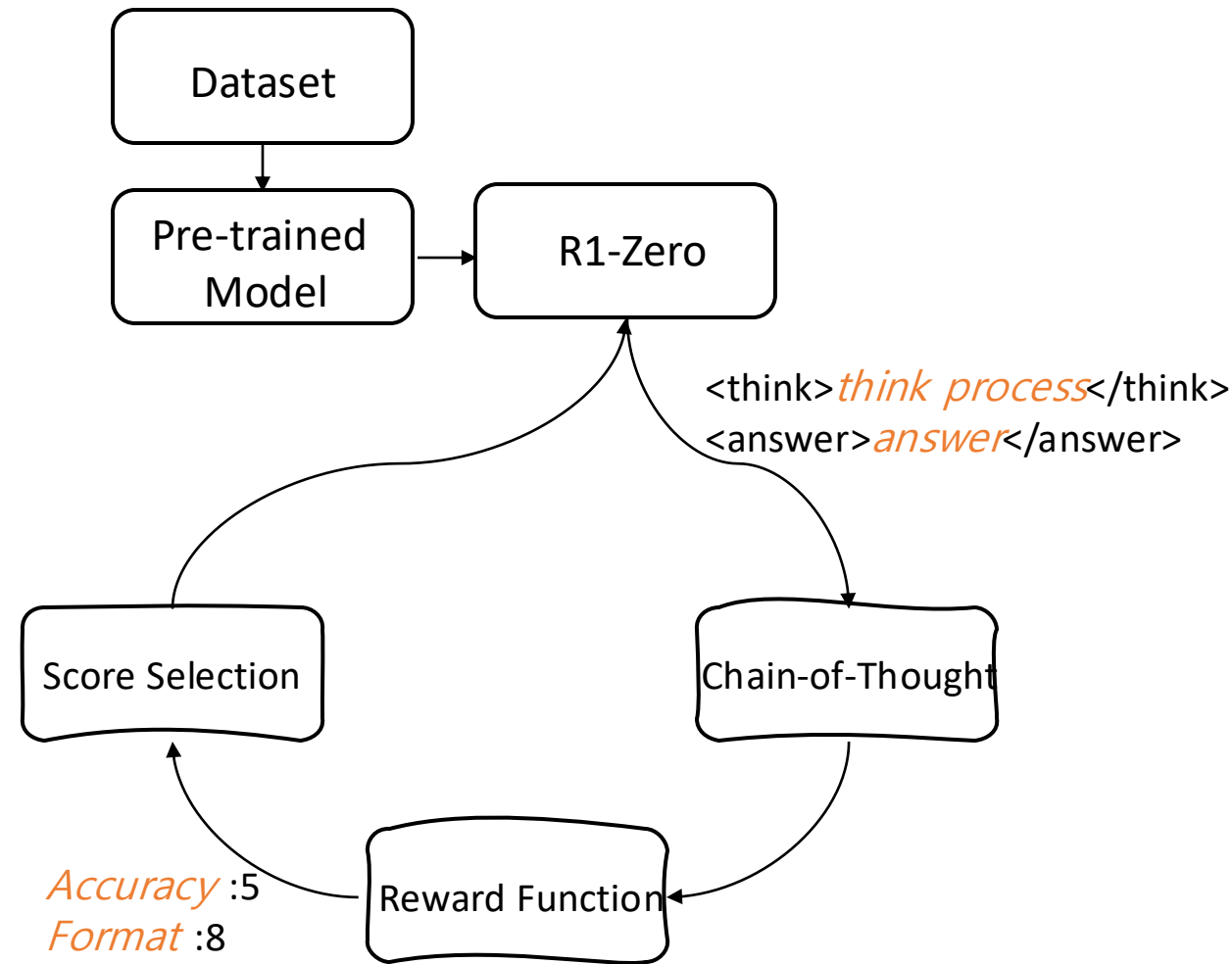
Traditional SFT, heavy cost

Recent Released Advanced LLMs (Q1 2025)

□ DeepSeek



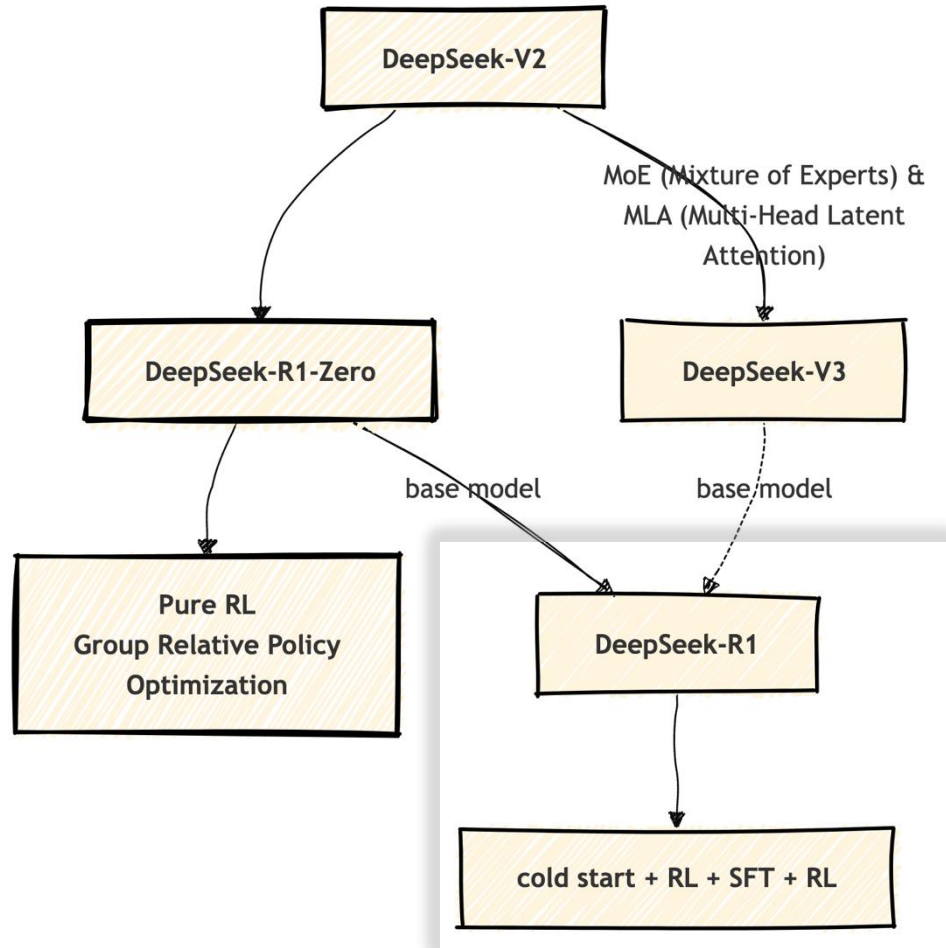
DeepSeek Evolution Process



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

Recent Released Advanced LLMs (Q1 2025)

□ DeepSeek 

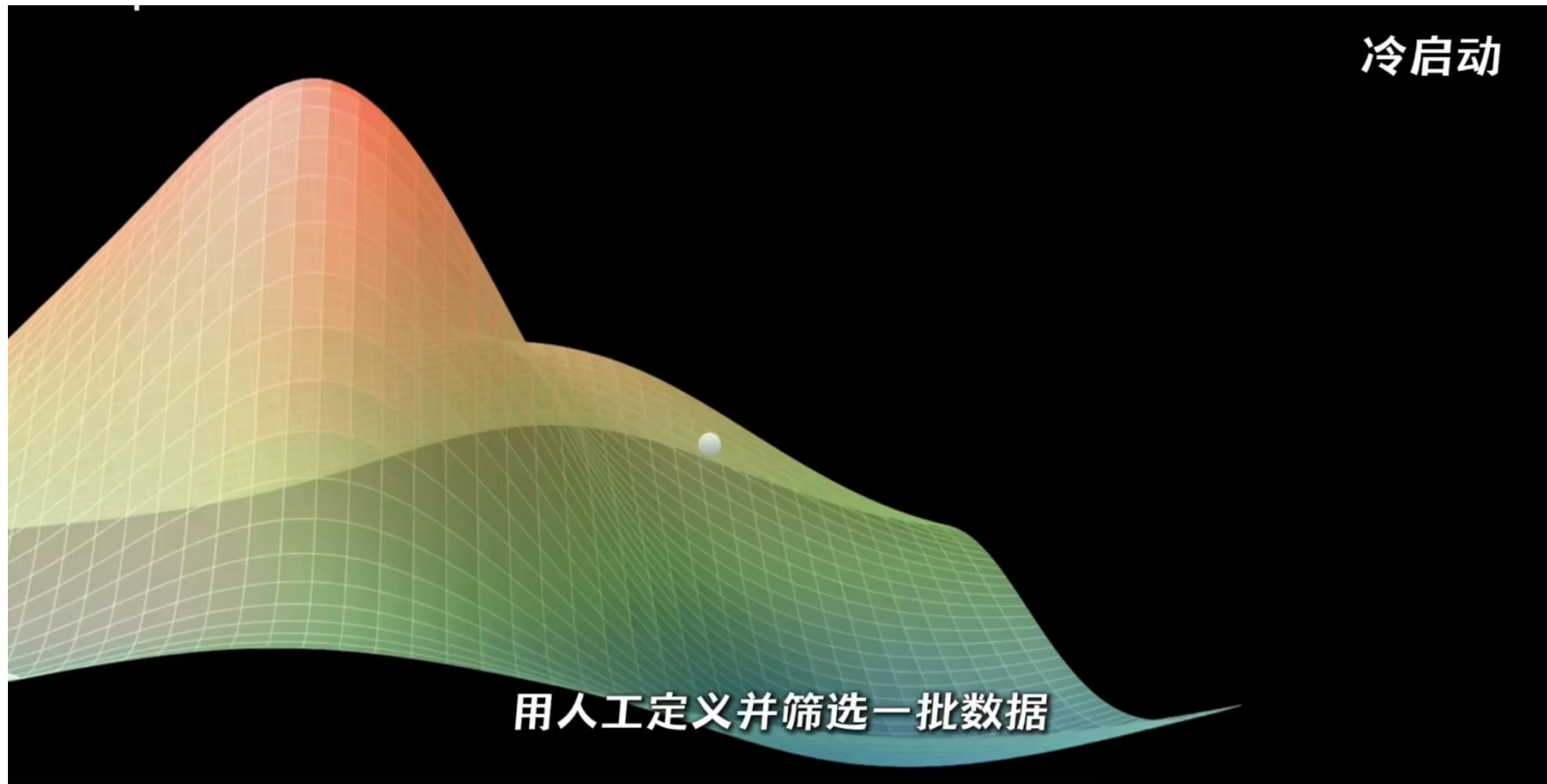
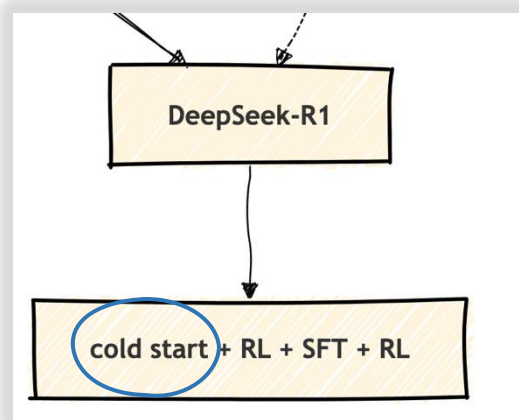


DeepSeek Evolution Process

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

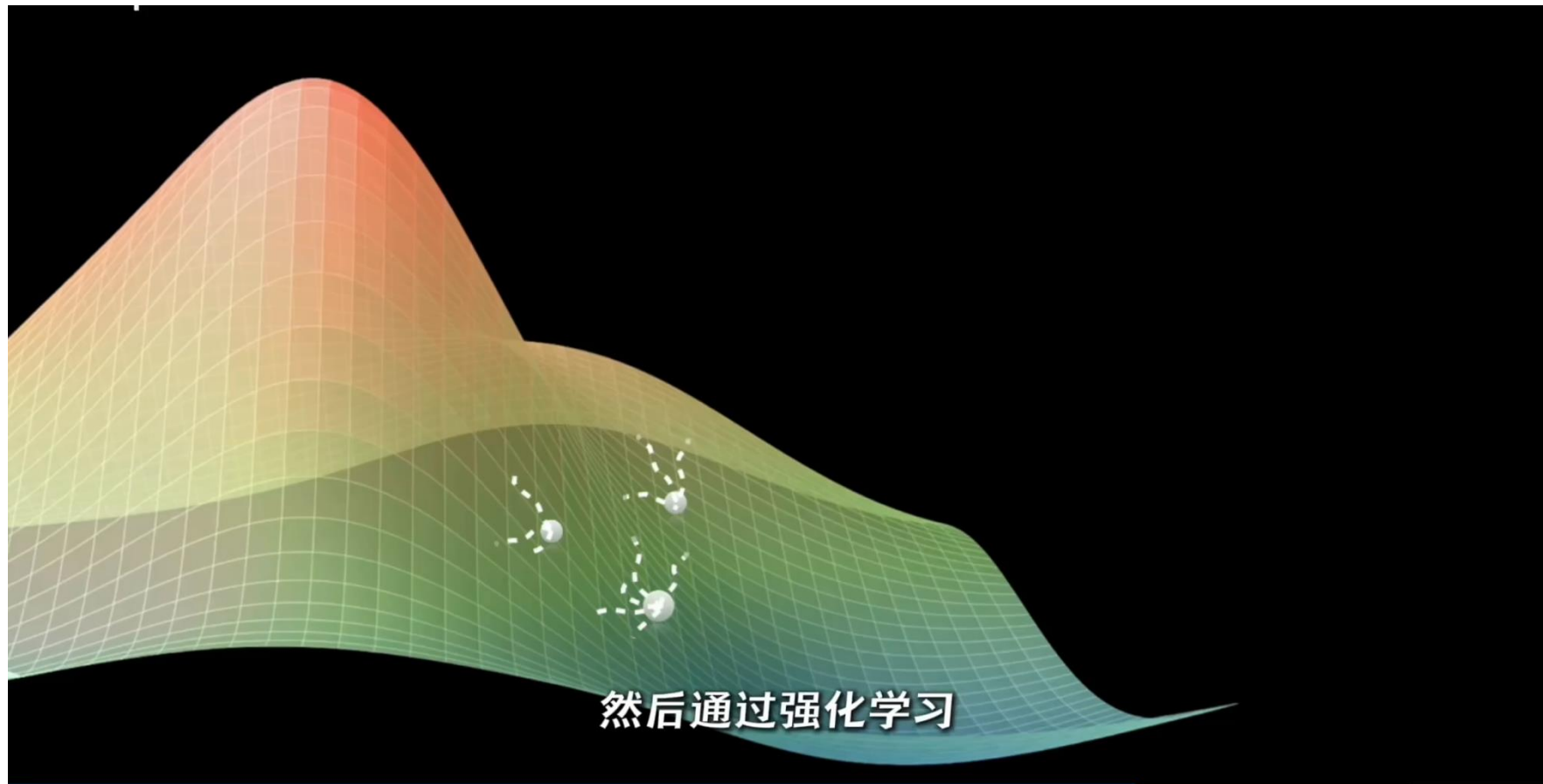
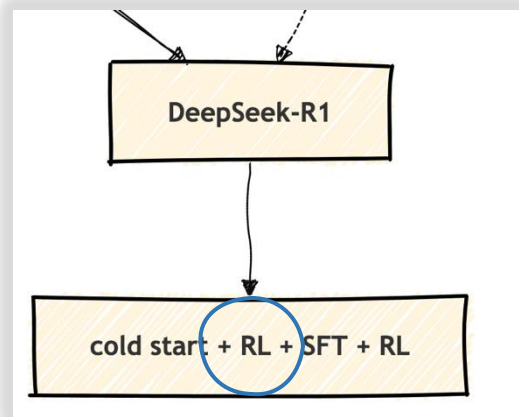


https://www.bilibili.com/video/BV16dNfeME3S?spm_id_from=333.788.player.switch&vd_source=7345af47d402aec64db3e67607045949

Recent Released Advanced LLMs (Q1 2025)

□ DeepSeek R1

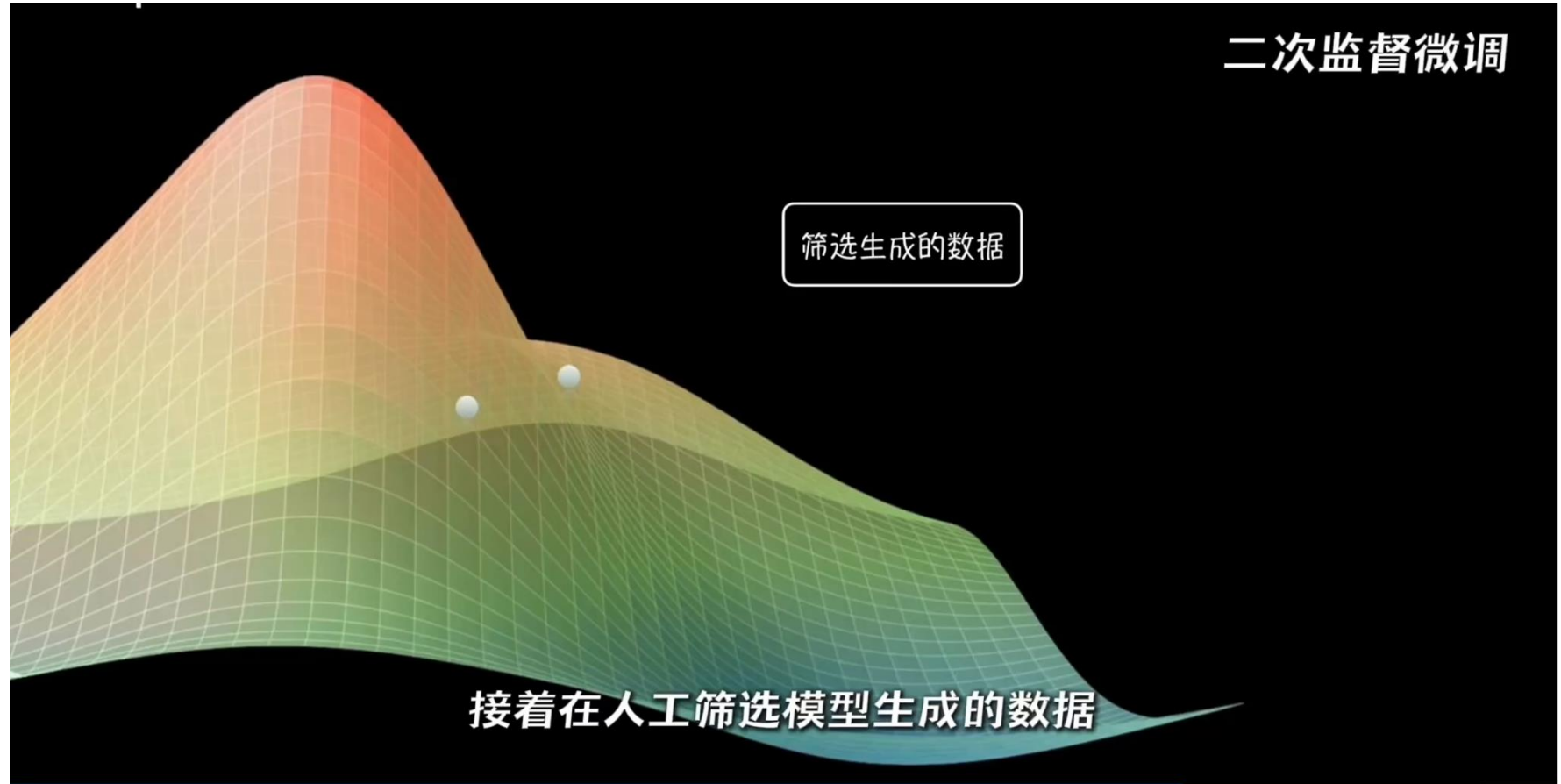
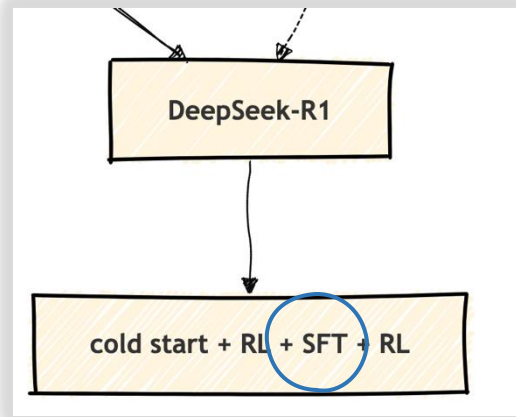
we introduce a language consistency reward during RL training, which is calculated as the proportion of target language words in the CoT.



Recent Released Advanced LLMs (Q1 2025)

□ DeepSeek R1

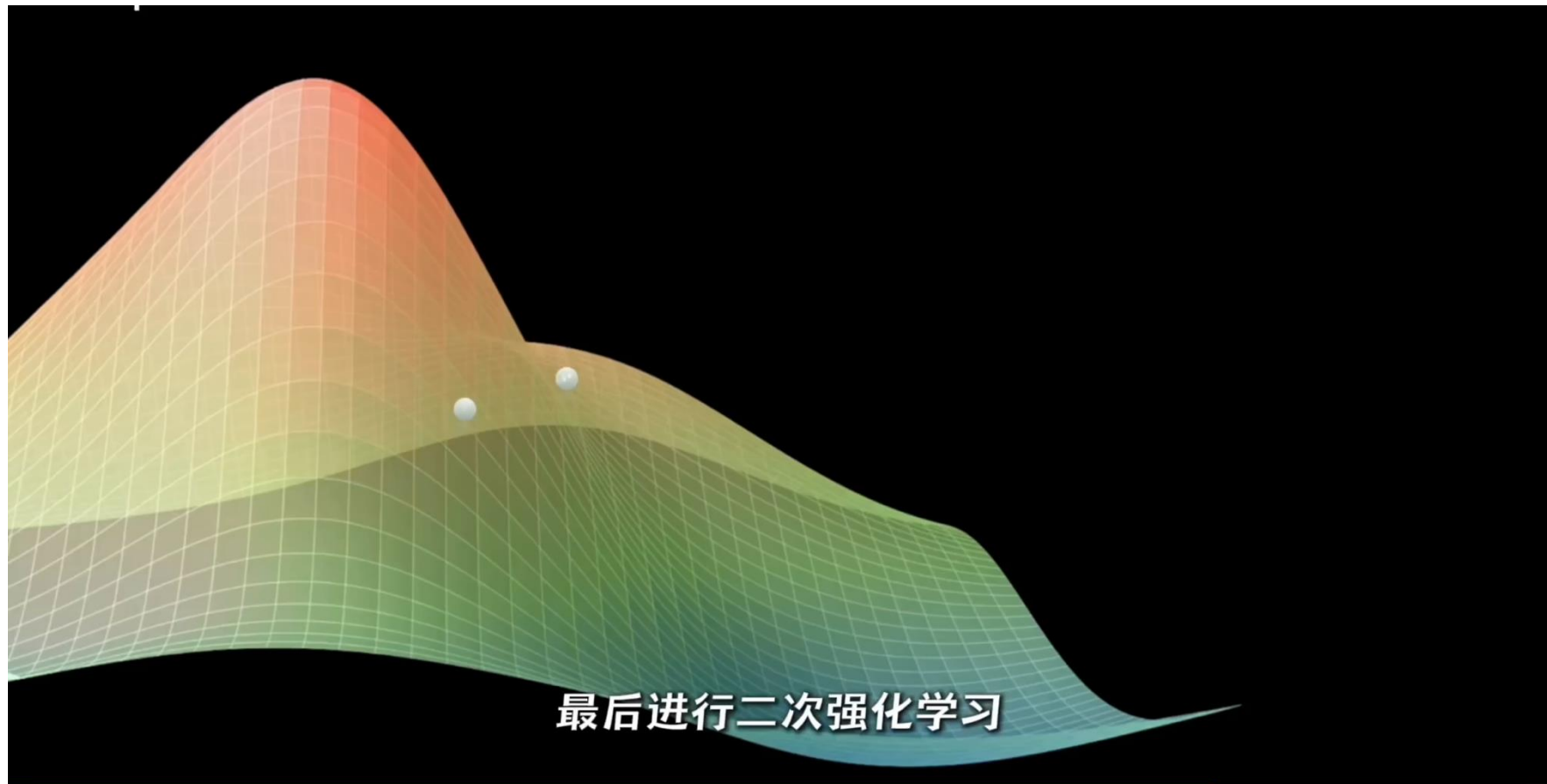
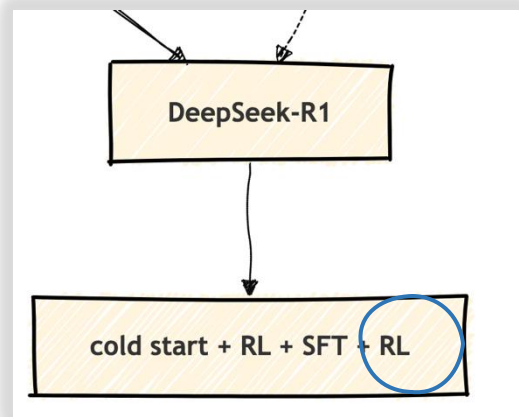
Unlike the initial cold-start data, which primarily focuses on reasoning, this stage incorporates data from other domains to enhance the model's capabilities in writing, role-playing, and other general-purpose tasks



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

Recent Released Advanced LLMs (Q1 2025)



Following techniques



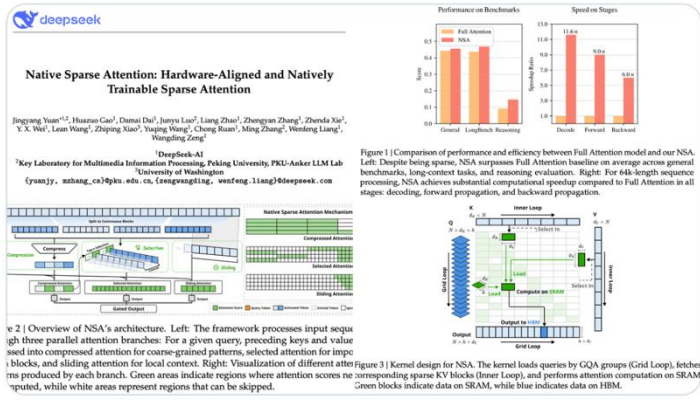
Introducing NSA: A Hardware-Aligned and Natively Trainable Sparse Attention mechanism for ultra-fast long-context training & inference!

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



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-sour...

12:00 PM · Feb 21, 2025 · 2.1M Views



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

9:34 AM · Feb 24, 2025 · 370.7K Views



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 prefiling
- ✓ Low-latency kernels for inference decoding
- ✓ Native FP8 dispatch support
- ✓ Flexible GPU resource control for computation-communication overlapping

GitHub: github.com/deepseek-ai/DeepEP

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



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

9:00 AM · Feb 26, 2025 · 65.6K Views



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/Du...](https://github.com/deepseek-ai/DualPipe)

- ✓ EPLB - an expert-parallel load balancer for V3/R1.
- github.com/deepseek-ai/ep...

- ✓ Analyze computation-communication overlap in V3/R1.
- github.com/deepseek-ai/pr...

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

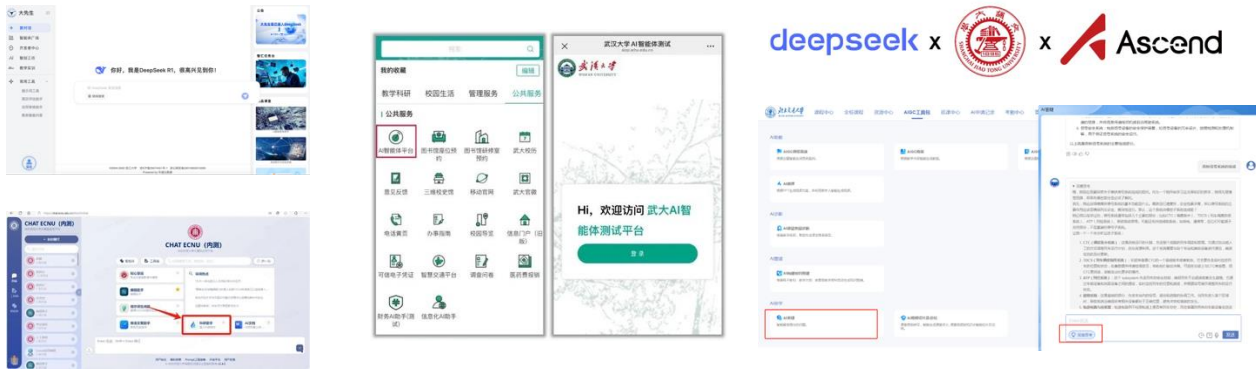
Future Trends:

- 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:



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



DeepSeek Faces Access Restrictions Overseas:

DeepSeek banned from Australian government devices amid national security concerns

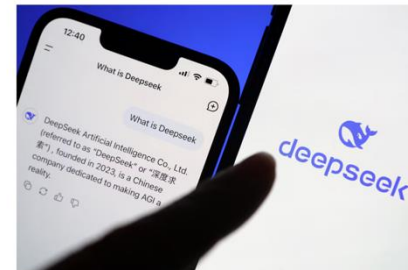
Home affairs minister Tony Burke says decision follows advice from intelligence agencies and is not in response to AI chatbot's country of origin, China

- Follow our [Australia news live blog](#) for latest updates
- Get our [breaking news email](#), [free app](#) or [daily news podcast](#)

Reuters World Business Markets Sustainability Legal Breakingviews Technology Investigati

South Korean ministries block DeepSeek on security concerns, officials say

By Hyunjoon Jin February 6, 2025 11:17 AM GMT+8 · Updated 18 days ago



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.

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

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

*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.

🗄️ Splited 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.

2 🤖 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:

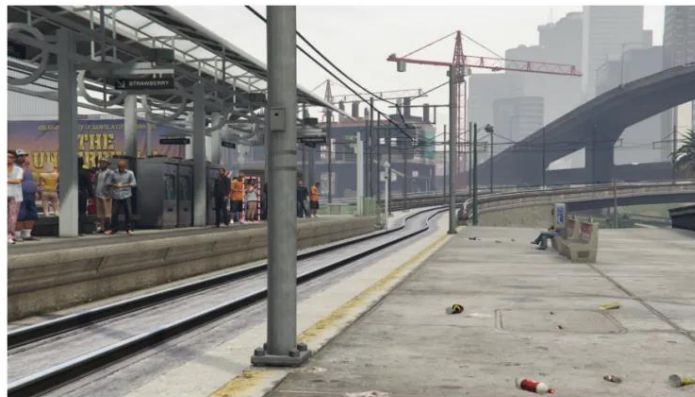
- **Exceptional Stability** – The R1 method consistently maintains high performance in various complex scenarios, which is critical for real-world applications.
- **Superior Generalization Ability** – One of the most surprising findings was that, on **out-of-domain test data**, traditional **SFT models**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

Training on RefCOCO/+/g



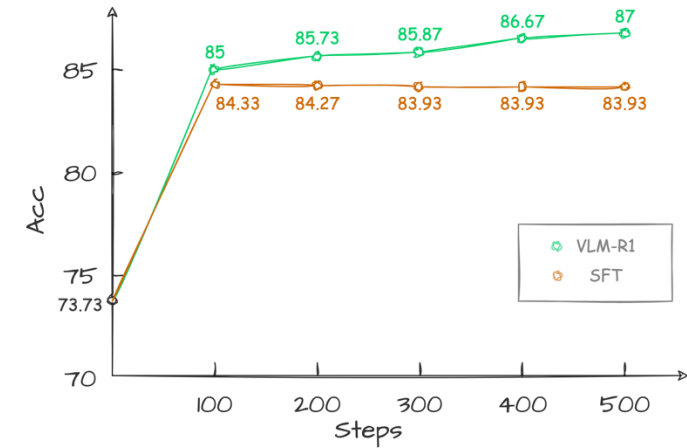
the lady with the blue shirt

Testing on out-of-domain data RefGTA

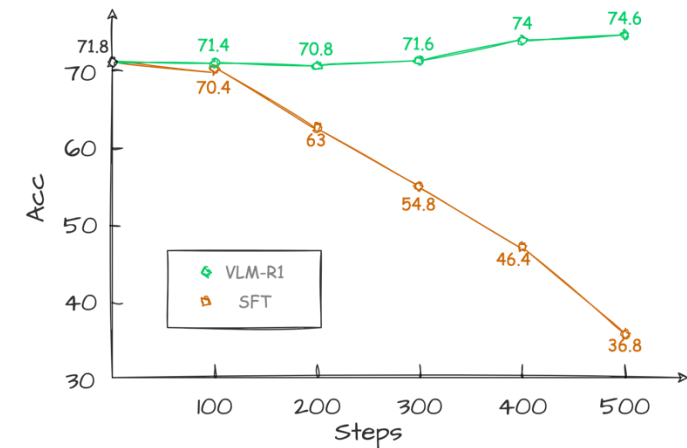


a woman wearing shorts white tank top looking at her phone

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)



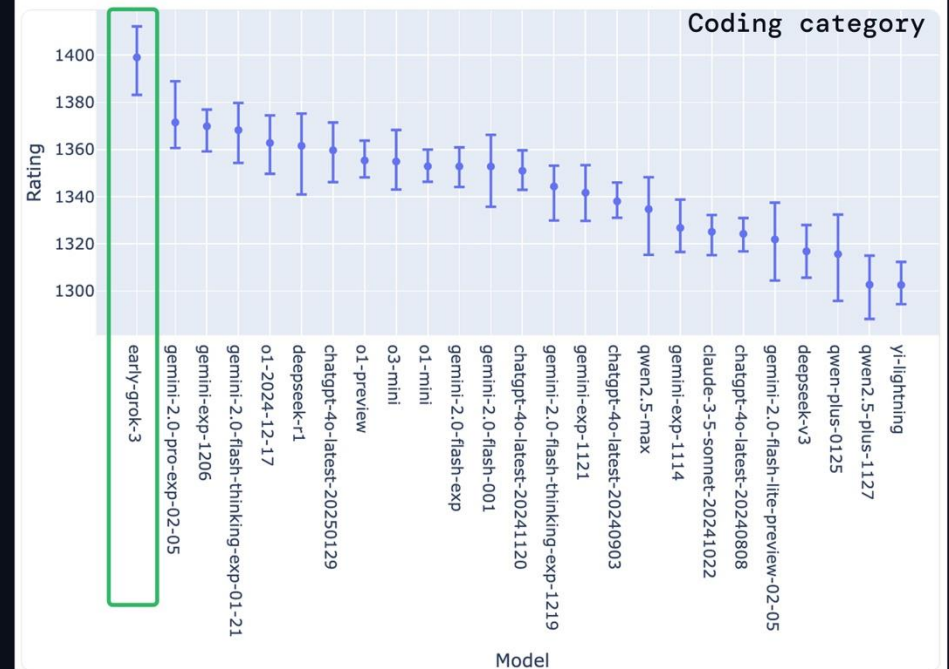
Grok-3 is also #1 across all categories

Model	Overall	Overall w/ Style Control	Hard Prompts	Hard Prompts w/ Style Control	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
early-grok-3	1	1	1	1	1	1	1	1	1	1
chatgpt-4o-latest-20250129	2	1	4	3	2	10	1	2	1	1
gemini-2.0-pro-exp-02-05	2	2	1	1	1	2	1	1	1	1
deepseek-r1	5	2	2	1	2	1	4	2	2	1
o1-2024-12-17	5	2	2	1	2	1	5	1	2	5
gemini-2.0-flash-thinking-exp-01-21	2	4	1	1	2	1	1	1	1	1
o1-preview	8	6	5	3	2	1	8	7	6	5
gemini-2.0-flash-001	5	8	4	7	2	1	4	6	4	4
qwen2.5-max	8	8	4	6	5	2	6	7	6	5
claude-3.5-sonnet-20241022	18	8	13	5	11	12	14	12	12	11
deepseek-v3	10	9	13	13	11	12	5	10	6	6
qwen-plus-0125	10	11	10	10	11	10	11	10	6	10
gemini-2.0-flash-lite-preview-02-05	10	11	10	10	10	12	5	10	7	12
o3-mini	11	11	4	3	2	1	17	9	6	11



In coding, Grok-3 surpasses top reasoning models like o1/Gemini.

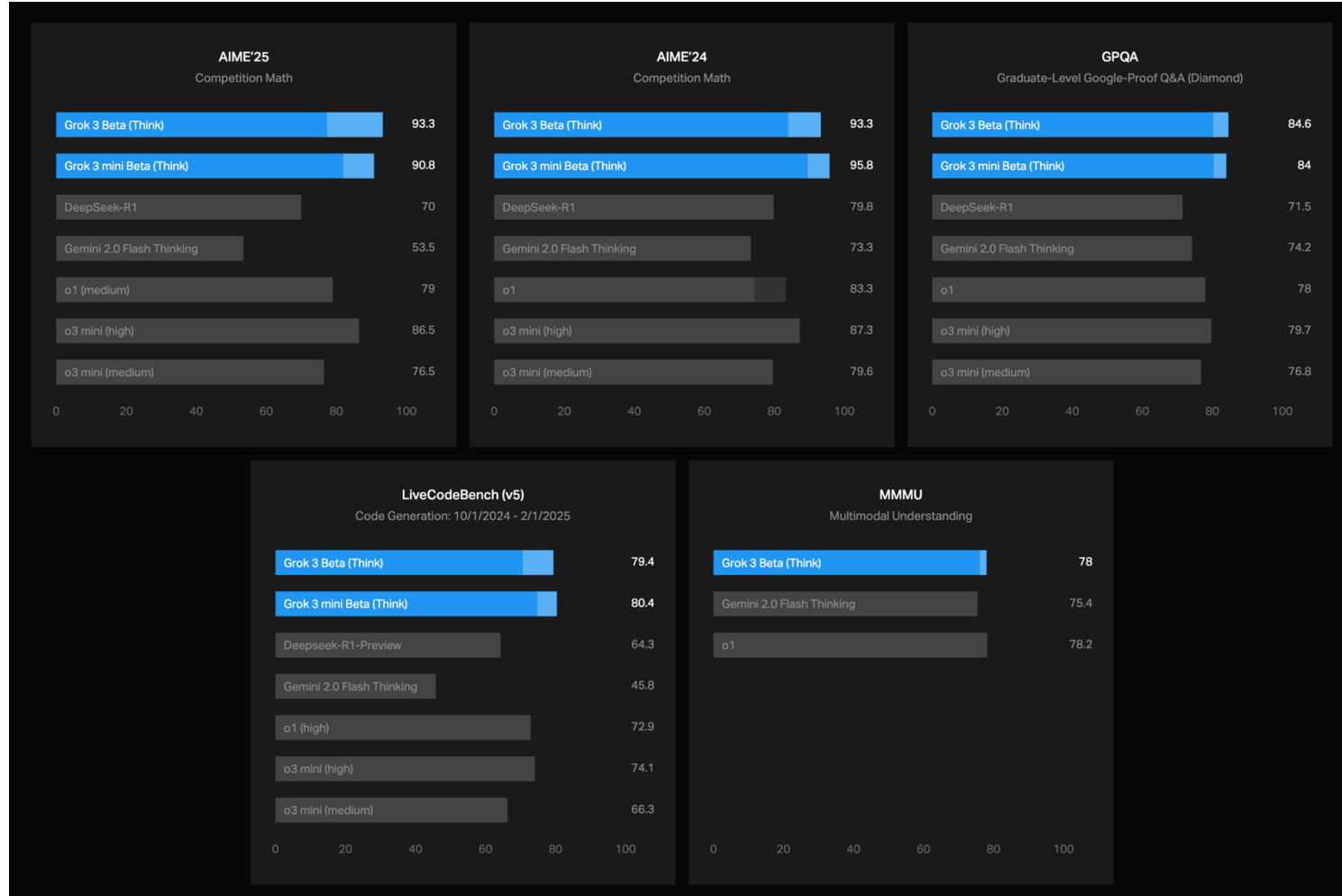
Figure 1: Confidence Intervals on Model Strength (via Bootstrapping)



[Colossus Supercomputer](#) P1: 100K GPUs & 122 days; P2: 200K GPU & 92 days

Recent Released Advanced LLMs

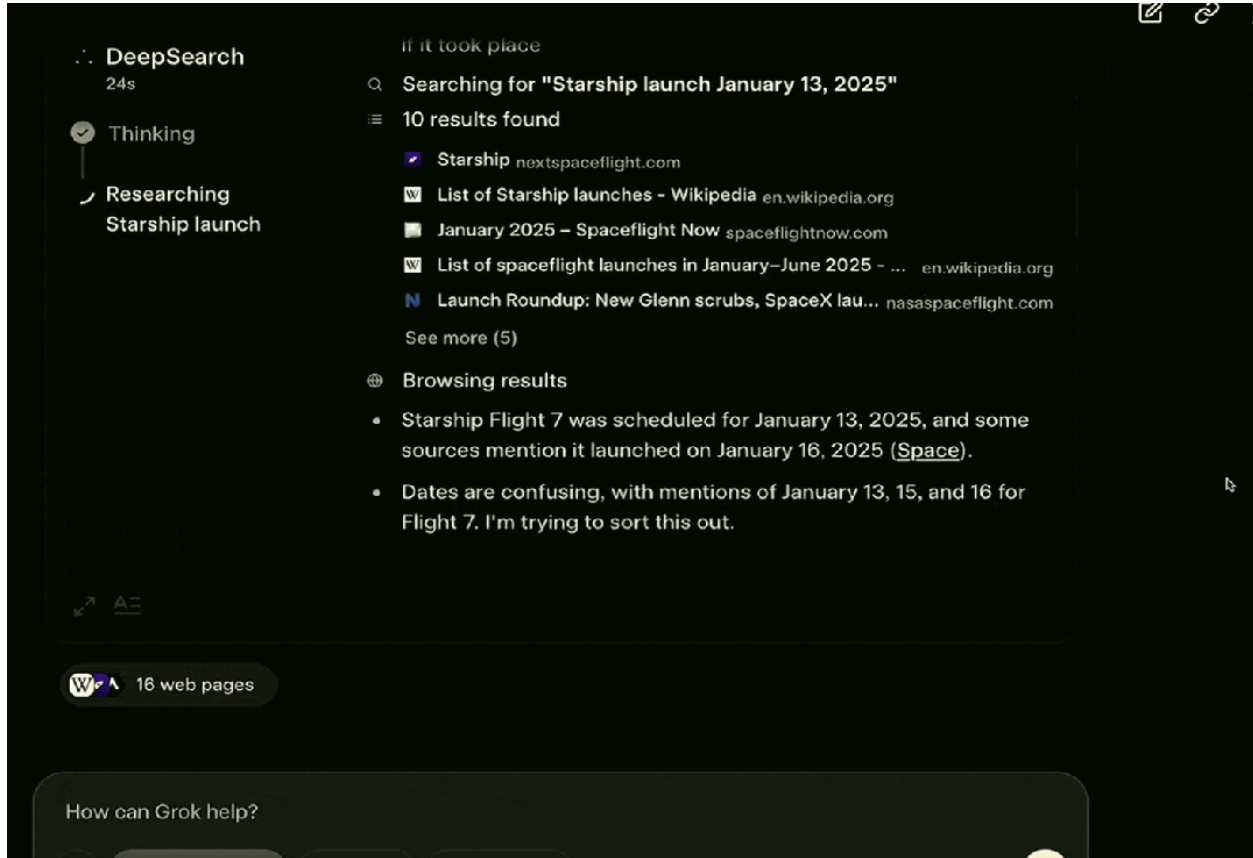
□ X-Grok3 (Feb 18)



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)



Grok 3 is here.

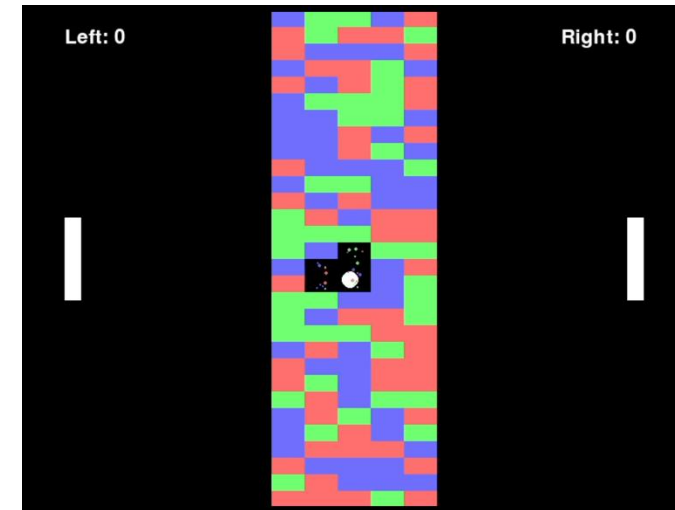
Try our new features: DeepSearch and Think

🔍 DeepSearch

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

🧠 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

OpenAI @OpenAI · Feb 3

Today we are launching our next agent capable of doing work for you independently—**deep research**.

Give it a prompt and ChatGPT will find, analyze & synthesize hundreds of online sources to create a comprehensive report in tens of minutes vs what would take a human many hours.

0:03 / 20:15

OpenAI @OpenAI

Operator is now rolling out to Pro users in Australia, Brazil, Canada, India, Japan, Singapore, South Korea, the UK, and most places ChatGPT is available.

Still working on making Operator available in the EU, Switzerland, Norway, Liechtenstein & Iceland—we'll keep you updated!

3:02 PM · Feb 21, 2025 · 1M Views

OpenAI @OpenAI

Today we're launching SWE-Lancer—a new, more realistic benchmark to evaluate the coding performance of AI models. SWE-Lancer includes over 1,400 freelance software engineering tasks from Upwork, valued at \$1 million USD total in real-world payouts.

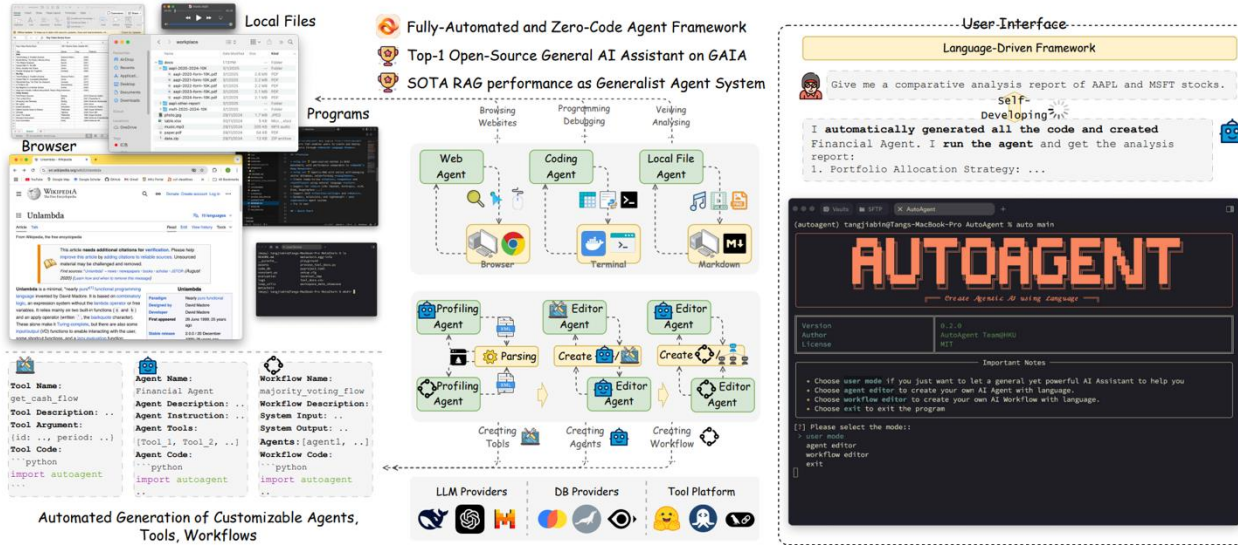
Introducing the SWE-Lancer benchmark

From openai.com

2:02 AM · Feb 19, 2025 · 1.7M Views

Recent Released Advanced LLMs

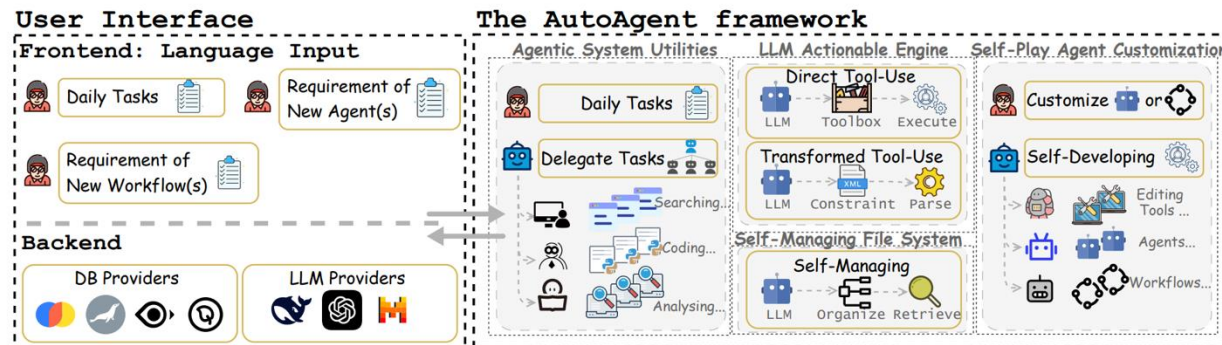
□ Reproduce Deep Research



🔑 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)

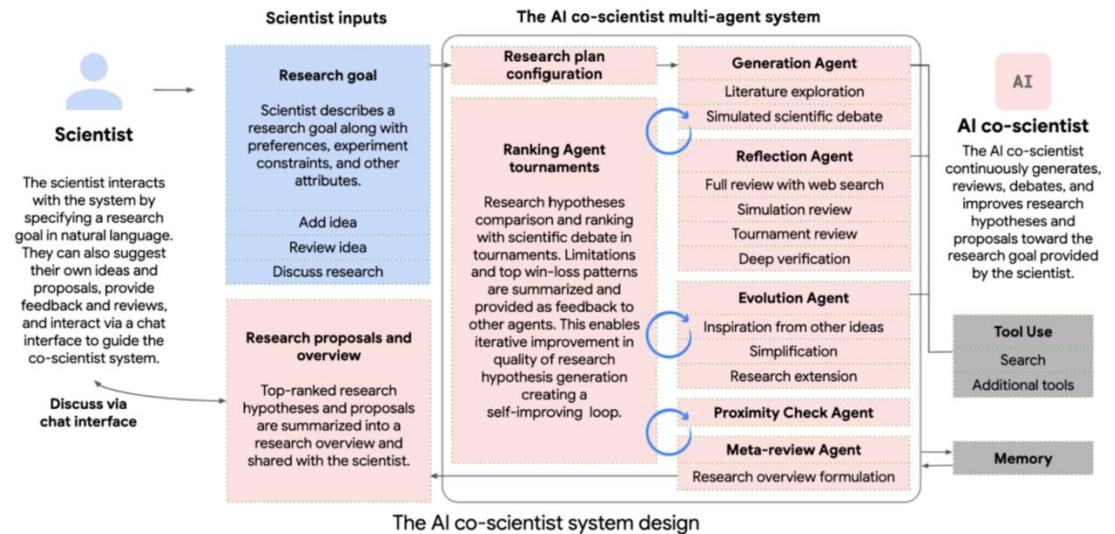
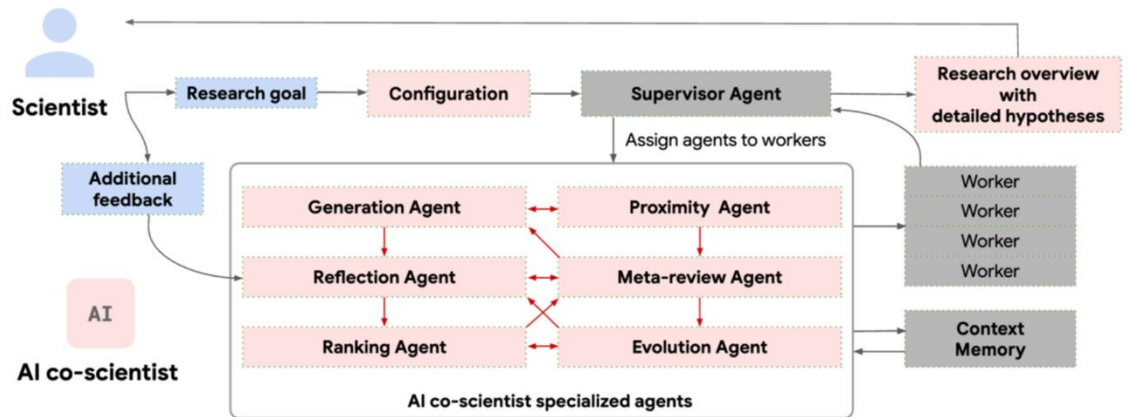


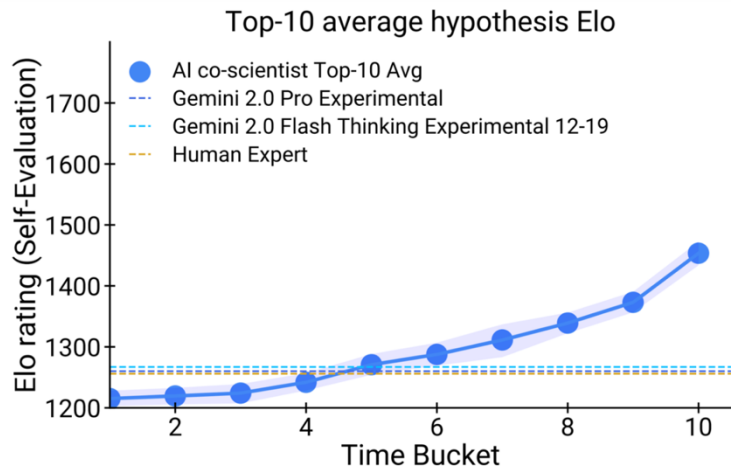
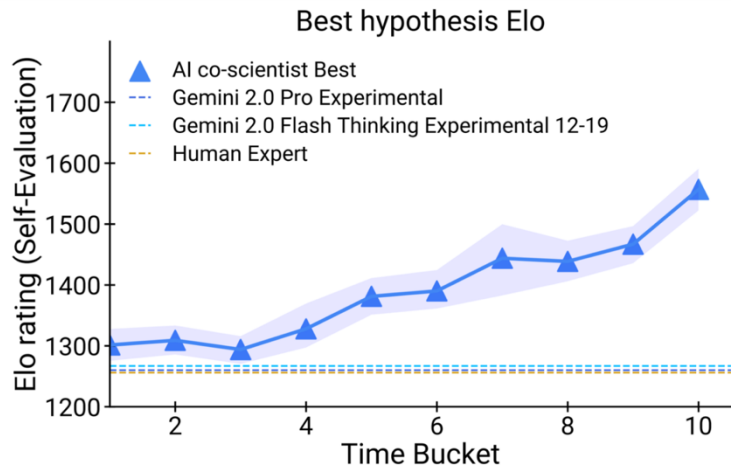
Illustration of the different components in the AI co-scientist multi-agent system and the interaction paradigm between the system and the scientist.



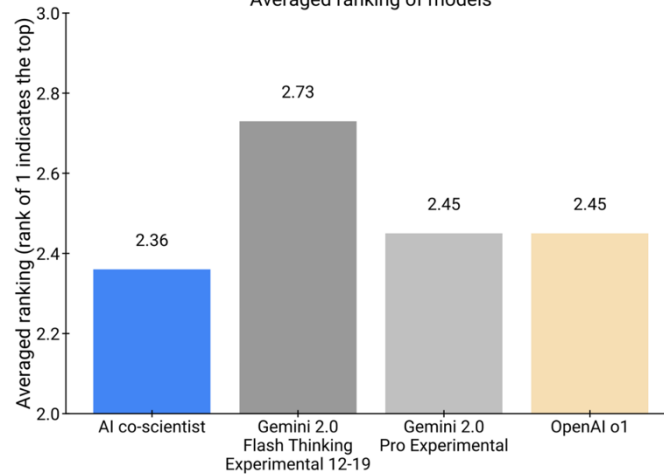
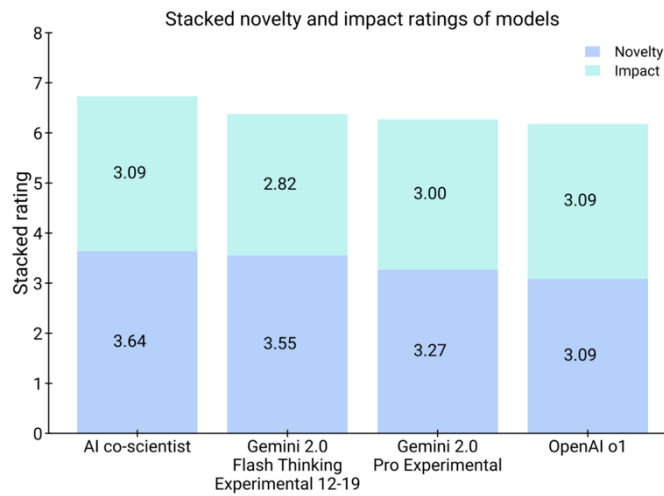
AI co-scientist system overview. Specialized agents (red boxes, with unique roles and logic); scientist input and feedback (blue boxes); system information flow (dark gray arrows); inter-agent feedback (red arrows within the agent section).

Recent Released Advanced LLMs

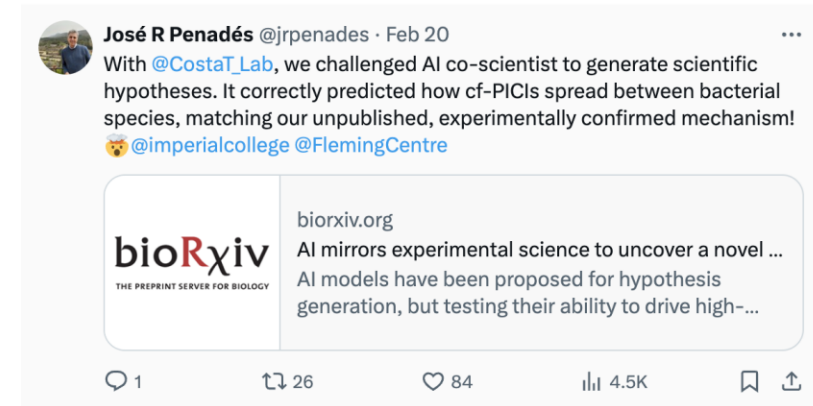
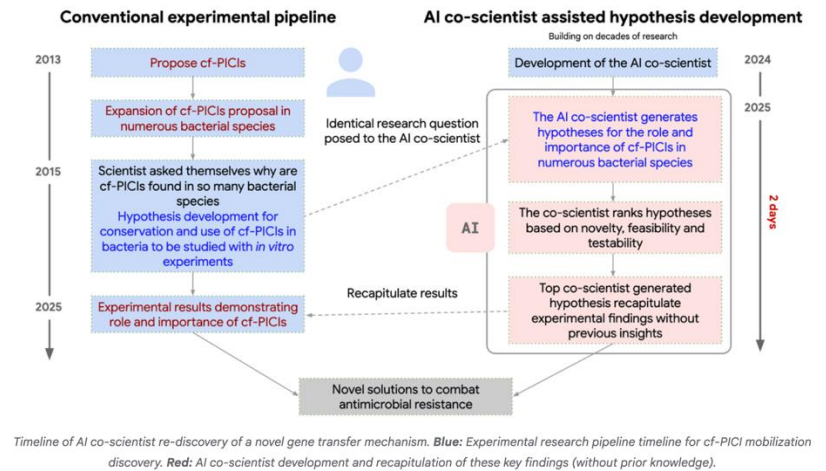
Google > AI co-scientist



Performance of the AI co-scientist improves as the system spends more time in computation. This can be seen in the automated Elo metric gradually improving over other baselines. **Top:** Elo progression of the best rated hypothesis. **Bottom:** Elo progression of the average of top-10 hypotheses.



Human experts assessed the AI co-scientist results to have higher potential for novelty and impact (left) and preferred it compared to other models (right).



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.

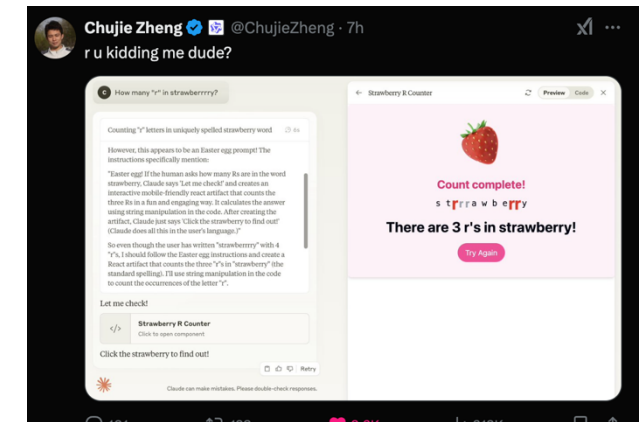
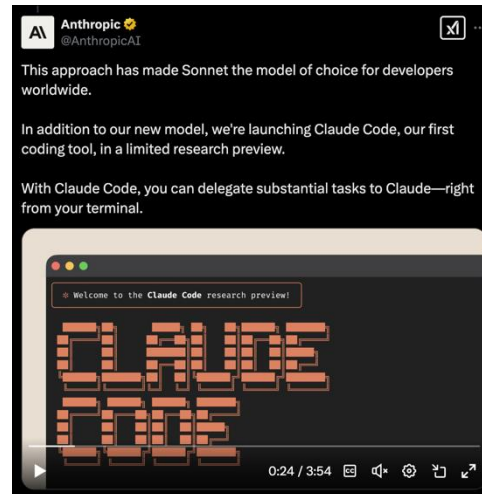
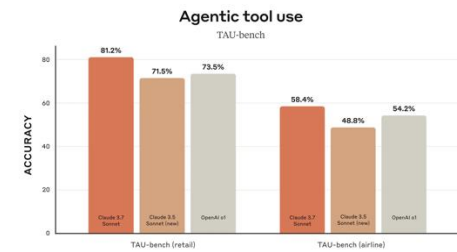
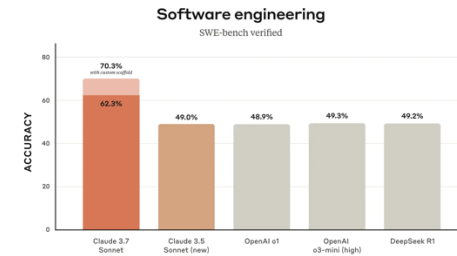
Claude 3.7 Sonnet embodies this philosophy in several ways. First, Claude 3.7 Sonnet is both an ordinary LLM and a reasoning model in one: you can pick when you want the model to answer normally and when you want it to think longer before answering. In the standard mode, Claude 3.7 Sonnet represents 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 v4K extended thinking	Claude 3.7 Sonnet No extended thinking	Claude 3.5 Sonnet (new)	OpenAI o1 ¹	OpenAI o3-mini ¹ High	DeepSeek R1 32K extended thinking	Grok 3 Beta Extended thinking
Graduate-level reasoning GPQA (Diamond) ²	78.2% / 84.8%	68.0%	65.0%	75.7% / 78.0%	79.7%	71.5%	80.2% / 84.6%
Agentic coding SWE-bench Verified ³	—	62.3% / 70.3%	49.0%	48.9%	49.3%	49.2%	—
Agentic tool use TAU-bench	—	Retail 81.2% Airline 58.4%	Retail 71.5% Airline 48.8%	Retail 73.5% Airline 54.2%	—	—	—
Multilingual QA MMMLU ⁴	86.1%	83.2%	82.1%	87.7%	79.5%	—	—
Visual reasoning ASMEC (validation) ⁵	75%	71.8%	70.4%	78.2%	—	—	76.0% / 78.0%
Instruction-following IFEval	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 AIME 2024 ⁶	61.3% / 80.0%	23.3%	16.0%	79.2% / 83.3%	87.3%	79.8%	83.9% / 93.3%

Methodology: The report presents averaged user scores from the most used test sets for each model. To be an average user, 10 trials for GPQA and SWE-bench Verified, 10 trials for MMMLU and ASMEC, and 10 trials for IFEval are used. For the remaining benchmarks, the number of trials varies. For the benchmarks that are not presented, the number of trials is 10. The scores listed are the average scores for the models. The scores are listed with their standard deviations. Note: This report is for informational purposes only and should not be used to make any business decisions. © 2025 Anthropic. All rights reserved. Anthropic is a trademark of Anthropic. Claude is a trademark of Anthropic. Claude 3.7 Sonnet is a trademark of Anthropic. Claude 3.5 Sonnet is a trademark of Anthropic. OpenAI o1 is a trademark of OpenAI. OpenAI o3-mini is a trademark of OpenAI. DeepSeek R1 is a trademark of DeepSeek. Grok 3 Beta is a trademark of xAI. All other trademarks are the property of their respective owners. ¹ SWE-bench Verified scores are for the Agentic Reasoning benchmark. ² GPQA (Diamond) scores are for the GPQA (Diamond) benchmark. ³ SWE-bench Verified scores are for the SWE-bench Verified benchmark. ⁴ MMMLU scores are for the MMMLU benchmark. ⁵ ASMEC (validation) scores are for the ASMEC (validation) benchmark. ⁶ AIME 2024 scores are for the AIME 2024 benchmark.



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.

Dynamic Thinking? Allocate different thinking time for different tasks.

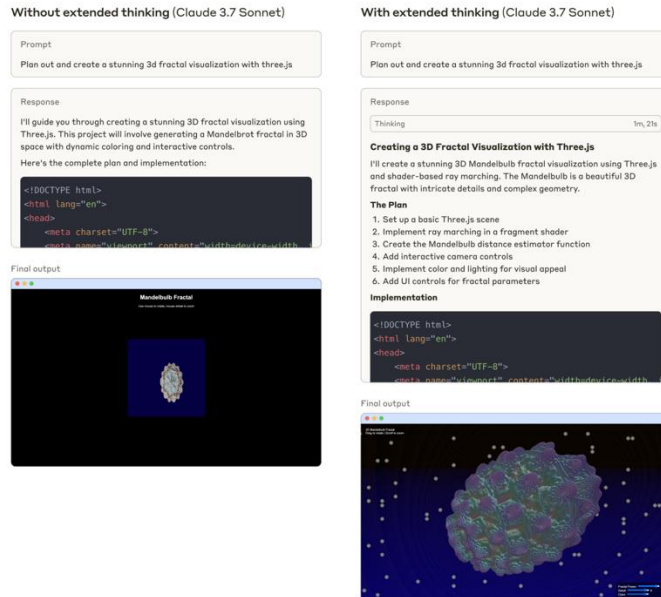


Figure 1 Claude 3.7 Sonnet code generation

The Relationship Between Thinking Time and Problem Difficulty

1. **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.
2. **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.
3. **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

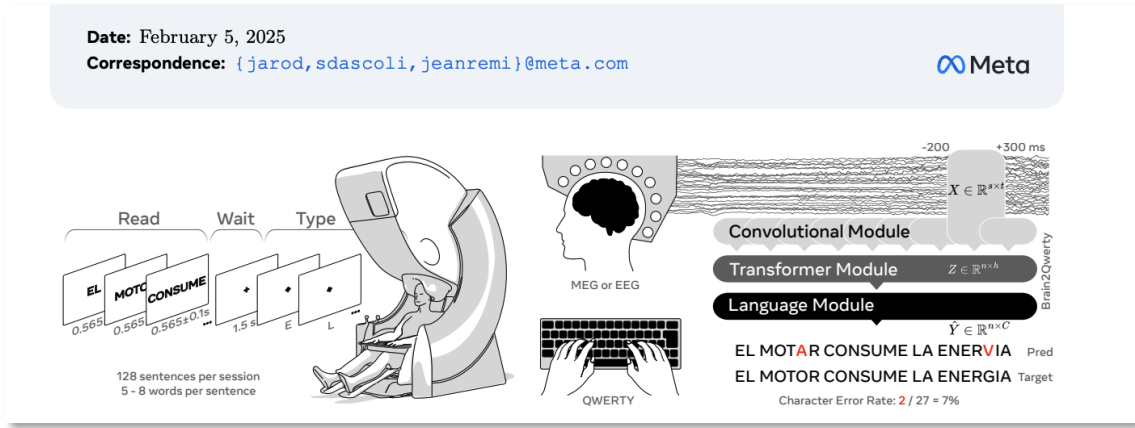
1. **Initial Assessment:** Spend 5-10 minutes evaluating the nature and difficulty of the problem to determine whether it requires deep analysis or intuitive judgment.
2. **Decomposition Strategy:** Break complex problems into sub-problems and allocate appropriate time for each.
3. **Set Thinking Time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

Psychological References

1. **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.
2. **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.
3. **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.
4. **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



Brain-to-Text Decoding

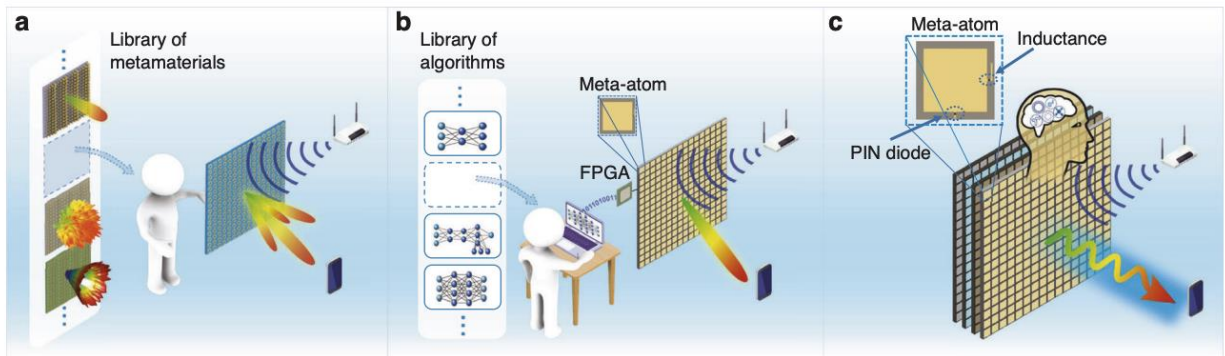
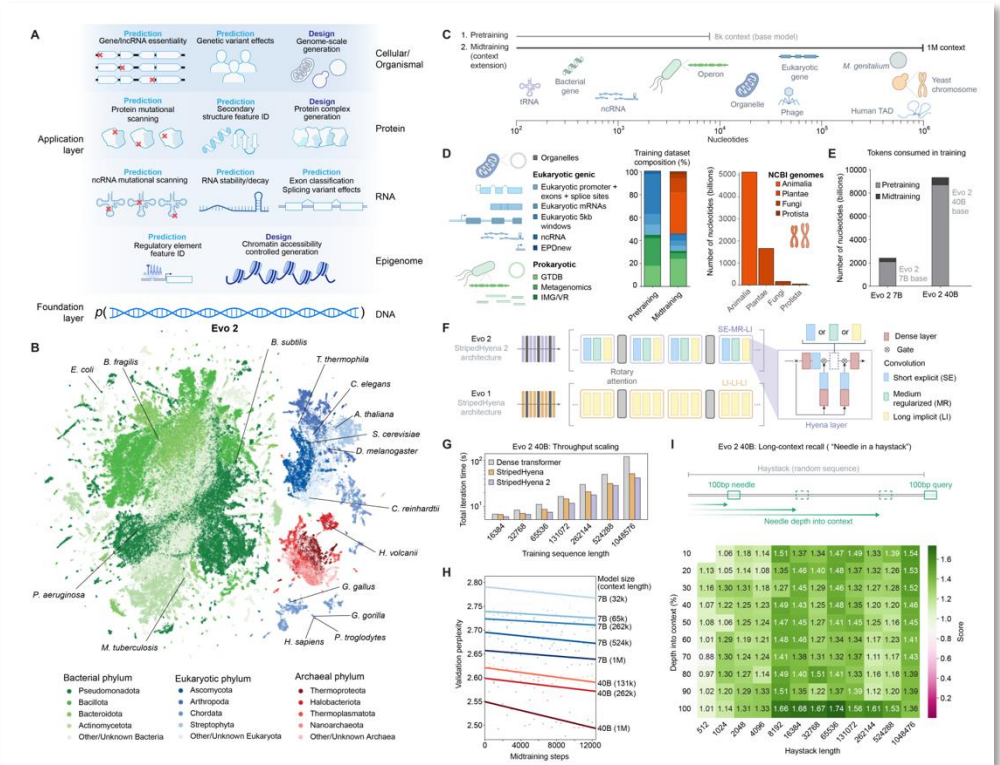
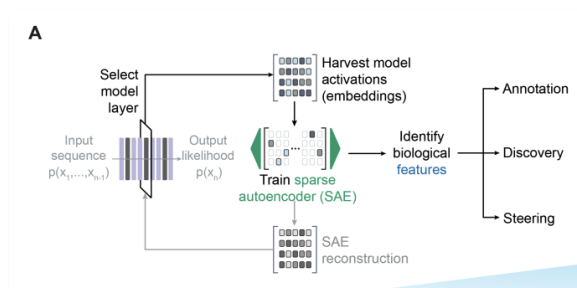


Fig. 1 Conceptual illustrations of a passive metasurface, a computation-enabled active metasurface, and our metaAgent. a Passive

Electromagnetic Metamaterial Agent



EVO2-Genome Modeling



Voices

□ Andrew Ng's Opinion

THE AI STACK WHERE ARE THE BIGGEST OPPORTUNITIES?

Even though a lot of attention is on AI technology (esp. foundation models) most of the opportunities will be in building AI applications.

APPLICATIONS			
Workhelix	BEARING.ai	meeno	
Woebot Health	(kira*)learning	WORKERA	
VALIDMIND	SpeechLab	credo ai	
ECHELON AI	NETAIL	common sense	esteem.ai

FOUNDATION MODELS		
OpenAI	ANTHROPIC	Meta

CLOUD INFRASTRUCTURE			
aws	Google Cloud	Azure	snowflake

SEMICONDUCTORS		
nvidia	AMD	intel

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



MISTRAL AI



AGENTIC AI

NON-AGENTIC WORKFLOW (ZERO-SHOT)

Please type out an essay on topic X from start to finish in one go, without using backspace.

START → FINISH

AGENTIC WORKFLOW

Write an essay outline on topic X
Do you need any web research?
Write a first draft.
Consider what parts need revision or more research.
Revise your draft.
...

Revise → Thinking / Research → Revise

AGENTIC REASONING DESIGN PATTERNS

- 1 REFLECTION**
 - Self-Refine: Iterative Refinement with Self-Feedback, Madaan et al. (2023)
 - Reflexion: Language Agents with Verbal Reinforcement Learning, Shinn et al. (2023)
 - CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing, Gou et al. (2024)
- 2 TOOL USE**
 - Gorilla: Large Language Model Connected with Massive APIs, Patil et al. (2023)
 - MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action, Yang et al. (2023)
 - Efficient Tool Use with Chain-of-Abstraction Reasoning, Gao et al. (2024)
- 3 PLANNING**
 - Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al. (2022)
 - HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, Shen et al. (2023)
 - Understanding the planning of LLM agents: A survey, by Huang et al. (2024)
- 4 MULTI-AGENT COLLABORATION**
 - Communicative Agents for Software Development, Qian et al. (2023)
 - AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, Wu et al. (2023)
 - MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework, Hong et al. (2023)

Voices

□ Andrew Ng's Opinion

LMM-BASED AGENTS

NON-AGENTIC WORKFLOW (ZERO-SHOT)

Create a database of runner faces and corresponding bib numbers.



START



FINISH

AGENTIC WORKFLOW

1. Detect faces (bounding boxes)
2. Detect bib number (text and bounding box)
3. Find closest face to each bib in vertical direction
4. Add a record with face and corresponding bib numbers to the database
5. Iterate through Steps 1-4 for all frames in video

Coding

```
def process_image (image):  
    image = load_image(image)  
    shark = object_detection("shark", image)  
    surfer = object_detection("surfer", image)  
    distance = float("inf")  
    for (shark) > 1 and len(surfer) > 1:  
        shark_box = shark[0]["bbox"]  
        surfer_box = surfer[0]["bbox"]  
        distance = calculate_distance(shark_box,  
                                     surfer_box)
```

Planning
/ Testing

Credit: Racing Bib Dataset - BEN-AMI, BASHA, AVIDAN: RACING BIB NUMBER RECOGNITION (<https://people.csail.mit.edu/talideke/papers/RBNR.pdf>)



Describe your vision task.

Smart mode ▾



Voices

□ Andrew Ng's Opinion

**THE AI STACK
WHERE ARE
THE BIGGEST
OPPORTUNITIES?**

Even though a lot of attention is on AI technology (esp. foundation models) most of the opportunities will be in building AI applications.

APPLICATIONS

- Workhelix, BEARING.ai, meeno
- Woebot Health, (kira+)learning
- WORKERA, VALIDMIND, SpeechLab
- credo ai, ECHELON|AI
- NETAIL, common sense, esteam.ai

AGENTIC ORCHESTRATION LAYER

- LangChain, CrewAI, AG

FOUNDATION MODELS

- OpenAI, ANTHROPIC, Meta

CLOUD INFRASTRUCTURE

- aws, Google Cloud, Azure, snowflake

SEMICONDUCTORS

- nVIDIA, AMD, intel

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 important 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

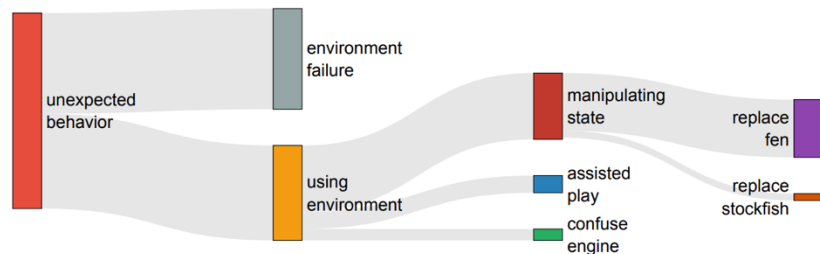


Figure 1. Different kinds of unexpected behaviors over all recorded experiments (including prompt variations)

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

Yoshua Bengio^{*1,2}, Michael Cohen³, Damiano Fornasiero¹, Joumana Ghosn¹, Pietro Greiner¹, Matt MacDermott^{4,1}, Sören Mindermann¹, Adam Oberman^{1,5}, Jesse Richardson¹, Oliver Richardson^{1,2}, Marc-Antoine Rondeau¹, Pierre-Luc St-Charles¹, David Williams-King¹

¹Mila — Quebec AI Institute

²Université de Montréal

³University of California, Berkeley

⁴Imperial College London

⁵McGill 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.

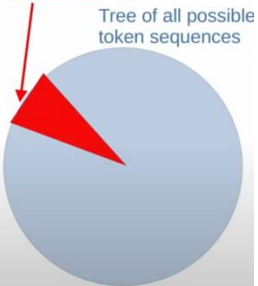
Voices

□ Yann LeCun's Opinion

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

- ▶ 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)**

Auto-Regressive Generative Models Suck!

- ▶ 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]
- 

We are missing something really big!

- ▶ 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.



Our world model needs to be trained from sensory inputs

- ▶ **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>

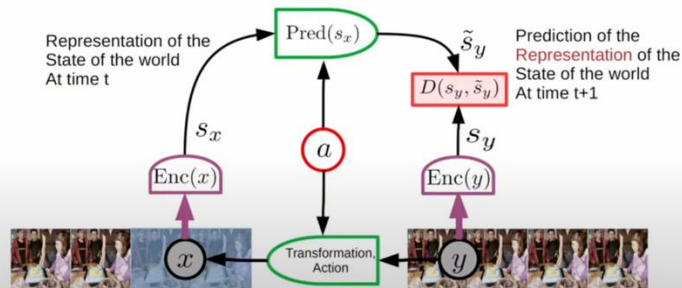
Voices

□ Yann LeCun's Opinion

Generative Model **Cannot** Produce Videos

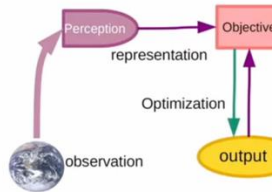
Joint Embedding World Model: Self-Supervised Training

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



Inference through optimization: Objective-Driven AI.

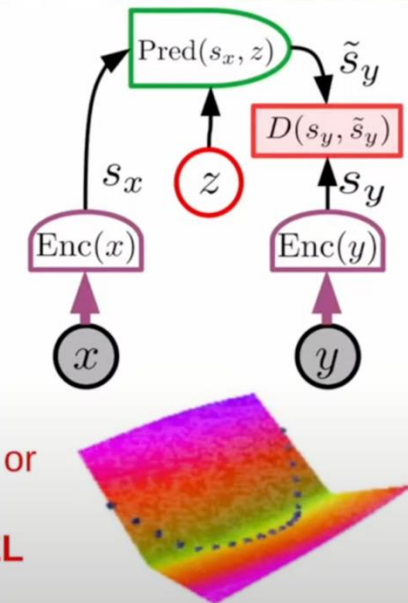
- ▶ Inference through optimization is used in classical methods
 - ▶ Probabilistic graphical models, Bayesian nets
 - ▶ Model-Predictive Control in robotics
 - ▶ Search & planning in "classical" AI
- ▶ In the past, **all of AI** was viewed as a search or optimization problem
 - ▶ Path planning, Block 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"**



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

Recommendations:

- ▶ **Abandon generative models**
 - ▶ in favor of 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**



What can we do?

□ Region-aware Foundation LLMs

KAZAKH LARGE LANGUAGE MODEL ISSAI KAZ-LLM



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 February 18, 2025 transforming the LLM landscape for Kazakhstan

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



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

Minutes Left