14th June, 2025

DeepSeek

Where Open-Source Meets Open-Minds

KIRTI SWARUP SWAIN

https://www.bbc.com/news/articles/c5yv5976z9po

DeepSeek: The Chinese AI app that has the world talking

4 February 2025

Share Save

Kelly Ng, Brandon Drenon, Tom Gerken and Marc Cieslak
BBC News

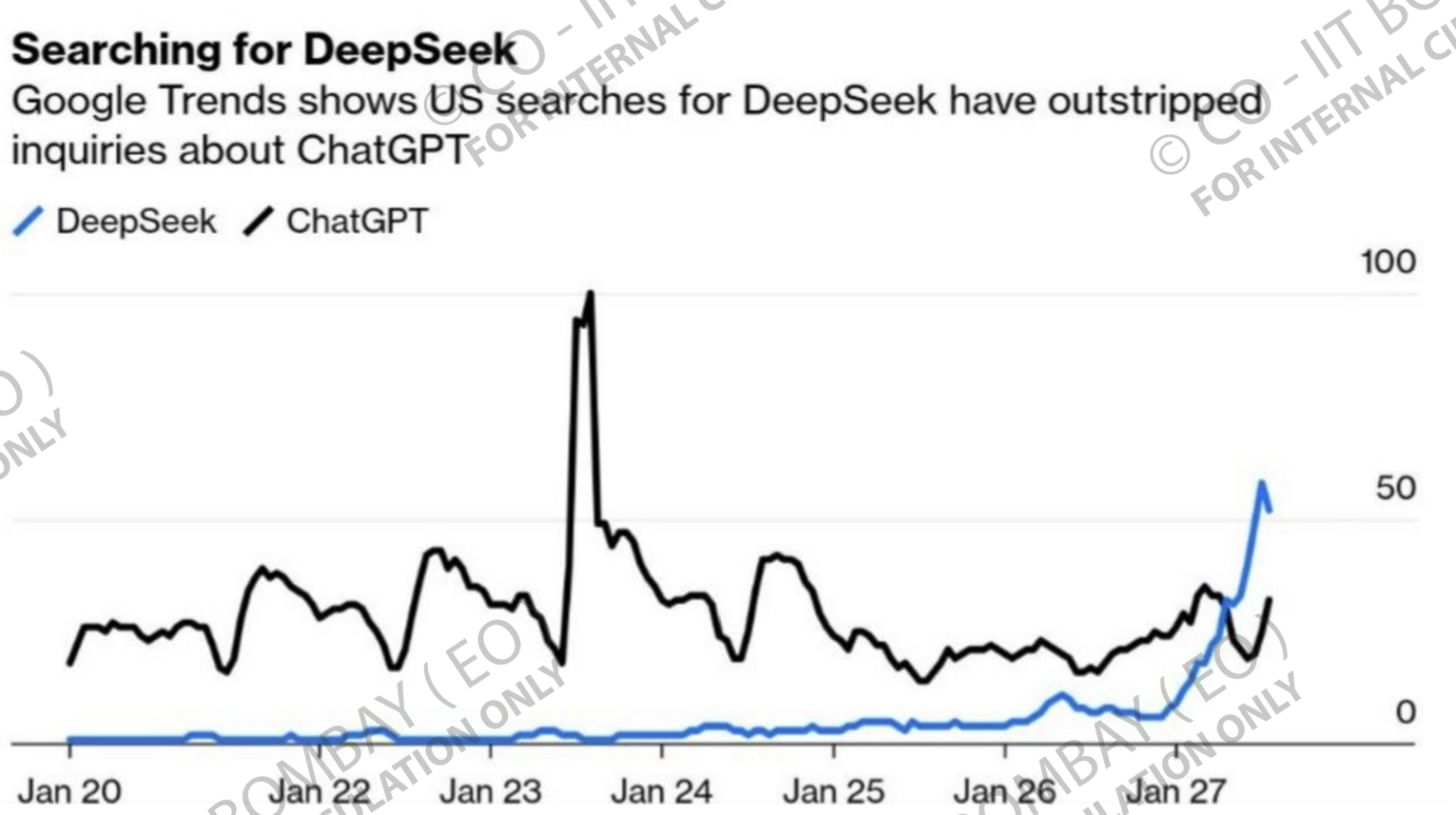
Heavy rain Today

Search

ENG IN

06:12 14-06-2025

Trend



Online Service	Launch Year	Time Taken to Reach 1 Million Users	Time Taken to Reach 10 Million Users
DeepSeek	2025	14 days	20 days
ChatGPT	2022	5 days	40 days
Perplexity AI	2022	2 months	13 months
Instagram	2010	2.5 months	355 days
Facebook	2004	10 months	852 days
Threads	2023	1 hour	7 hours
Twitter	2006	2 years	780 days
Netflix	1999	3.5 years	9 years

1) https://www.facebook.com/photo/?fbid=1169983918464817&set=very-interesting-by-bloomberg-us-google-searches-for-deepseek-outstrip-inquiries

2) https://explodingtopics.com/blog/deepseek-ai

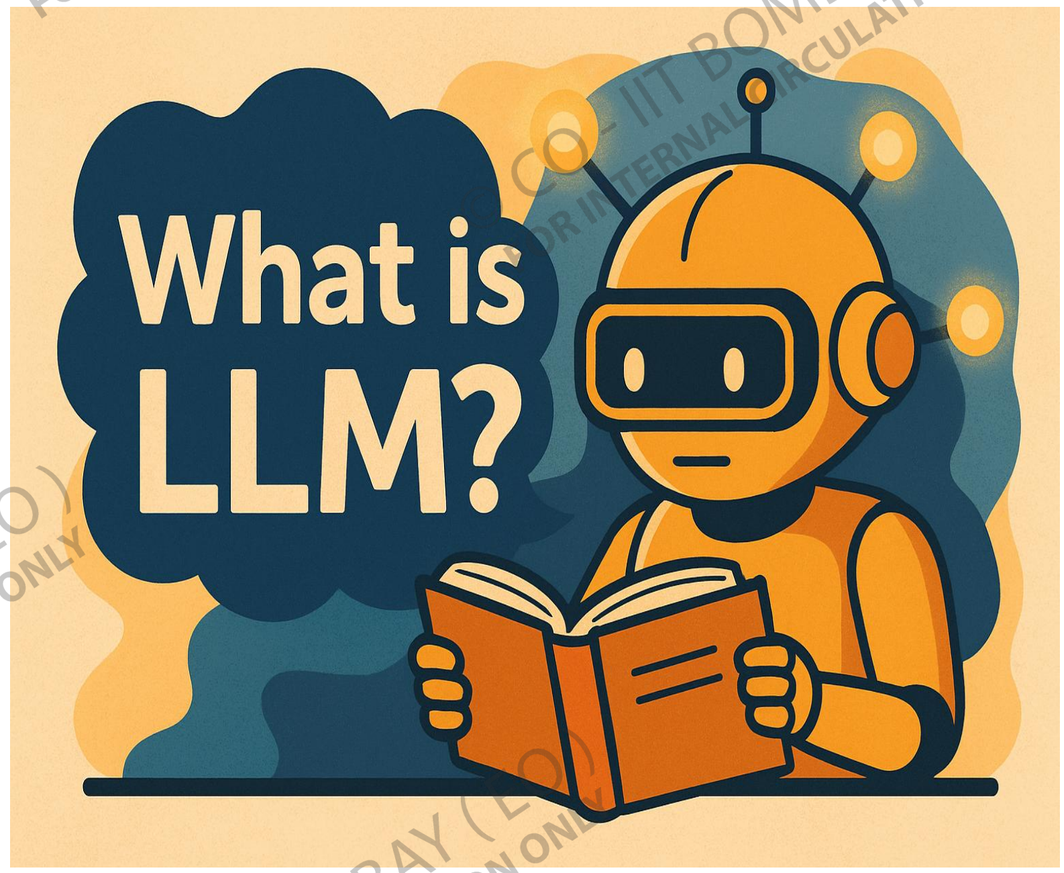
Foundational & Game-Changing LLMs

Year	Model	Organization	Highlights	Key Limitations
2018	**GPT-1**	OpenAI	First autoregressive transformer model	Small scale (117M), poor long-form results
2019	**GPT-2**	OpenAI	Breakthrough in coherent text generation	No fine-tuning, factual hallucinations
2020	**GPT-3**	OpenAI	175B params, few-shot learning	Expensive, biased, no tuning access
2022	**ChatGPT (3.5)**	OpenAI	Dialogue-tuned GPT-3.5	Repetitive, polite hallucinations
2023	**GPT-4**	OpenAI	Multimodal, 32K context, strong reasoning	Closed-source, high compute cost

Modern Popular LLMs (2023–2025)

Year	Model	Organization	Highlights	Key Limitations
2023	**Claude 2**	Anthropic	Harmlessness-focused, long context	Cautious replies, slower dev pace
2023	**Gemini 1.0**	Google DeepMind	Multimodal, web-connected	Closed-source, no public weights
2024	**Claude 3**	Anthropic	SOTA reasoning, 100K context	Limited tuning, cautious by design
2024	**Grok**	xAI (Twitter)	Real-time humor, X integration	Less smart than GPT-4, niche tone
2024	**Gemini 1.5**	Google DeepMind	1M token context, excellent search integration	Hardware-intensive, not open-source
2024	**Perplexity LLM**	Perplexity AI	Combines search + LLM in real-time	Depends on external data freshness
2025	**DeepSeek R1**	DeepSeek AI	Open-source, 16K context, cost-efficient	No moderation, early in ecosystem
2025	**GPT-4o**	OpenAI	Unified voice-vision-text, real-time	Closed, compute-heavy, costly at scale

LLM



Definition

A Large Language Model (LLM) is an advanced AI program designed to understand and generate human-like language.

Learning Process

It's trained by reading **millions of texts** — like books, articles, websites, and conversations — to learn how language works.

How It Responds

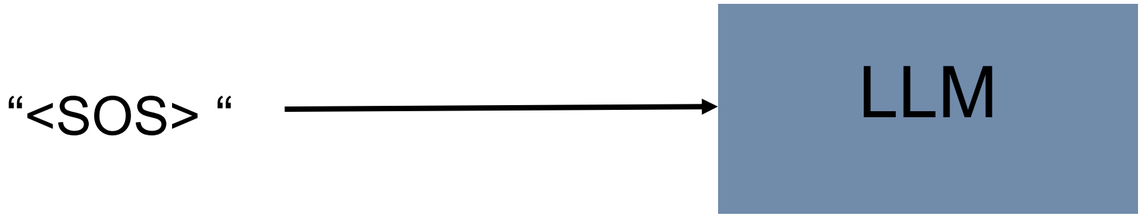
Based on what it has learned, it **predicts the next word** or **generates answers**, like completing sentences or replying to your questions.

Applications

LLMs power tools like:
Chatbots and virtual assistants
Language translators
Content writing tools
Code assistants
Search engines and summarizers

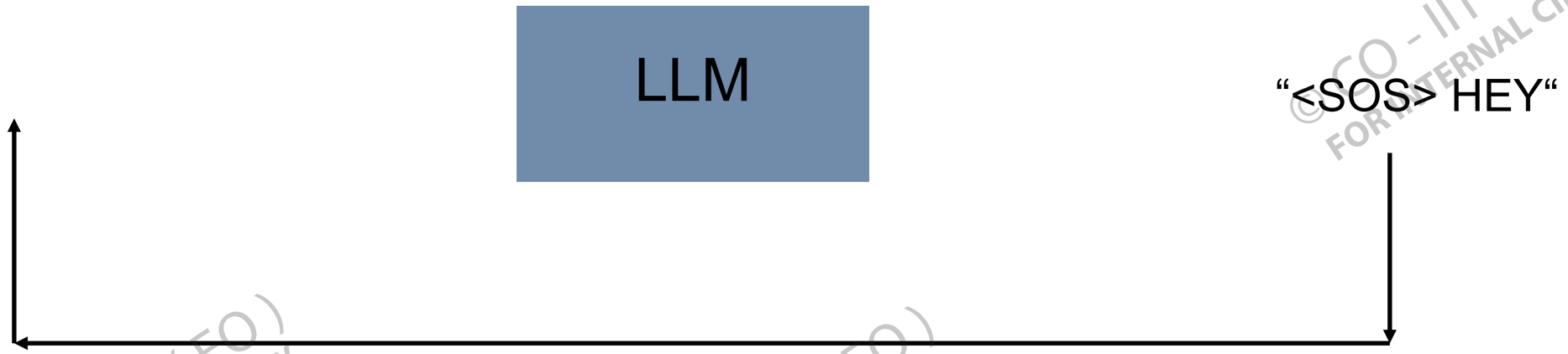
How traditional LLMs work

HEY ! HOW CAN I ASSIST YOU TODAY?

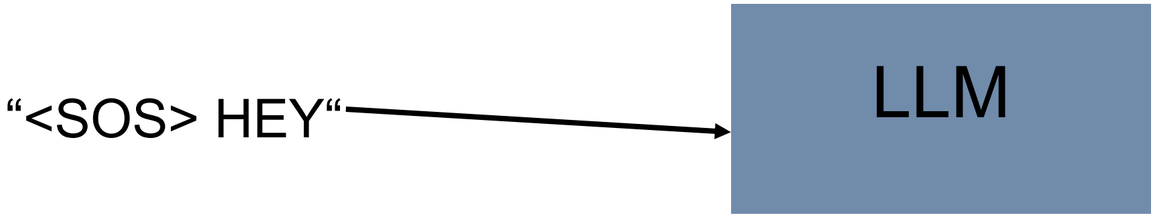


How traditional LLMs work

How traditional LLMs work

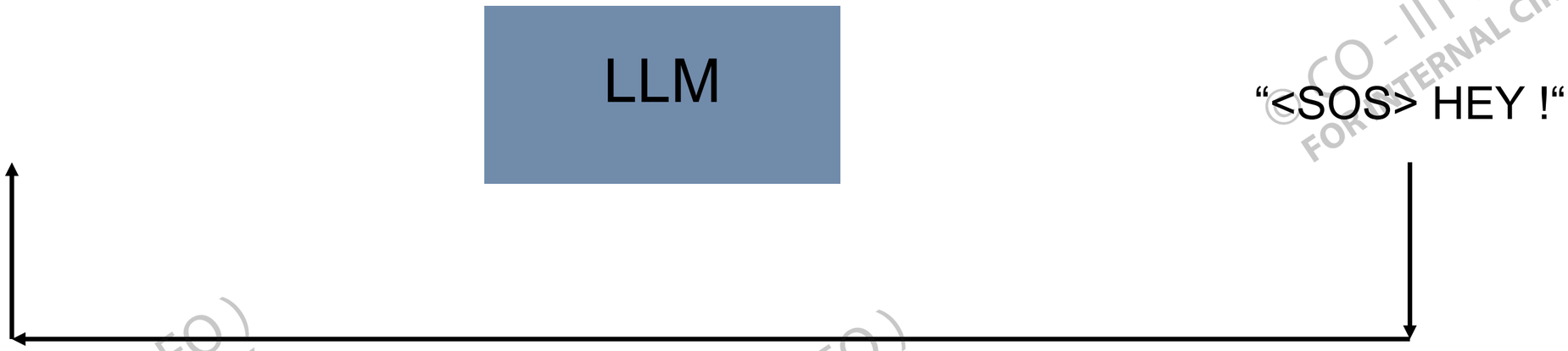


How traditional LLMs work

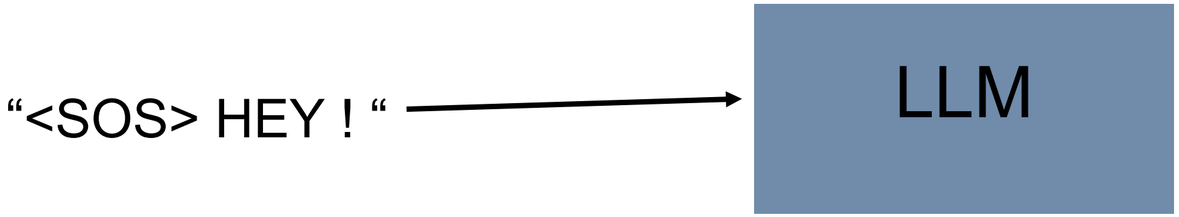


How traditional LLMs work

How traditional LLMs work

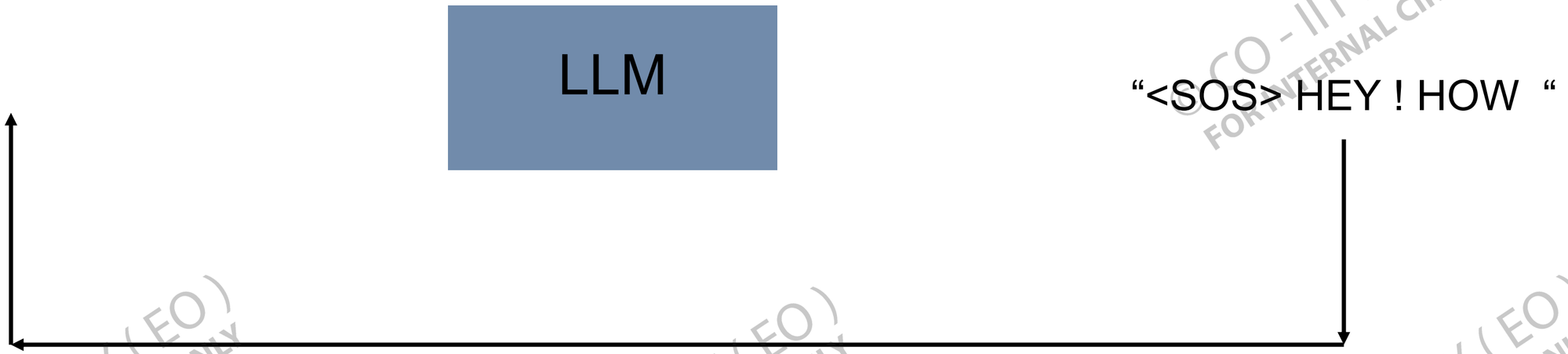


How traditional LLMs work

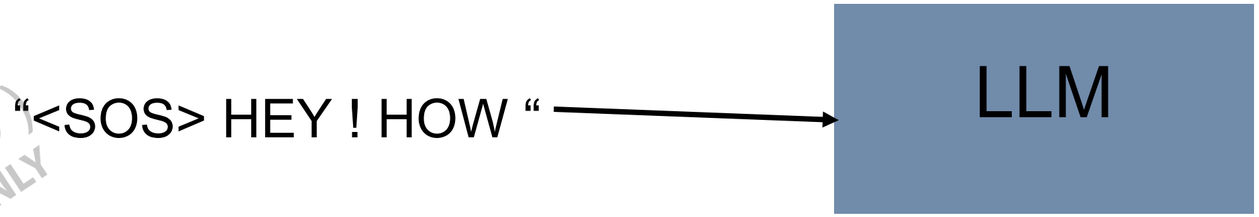


How traditional LLMs work

How traditional LLMs work

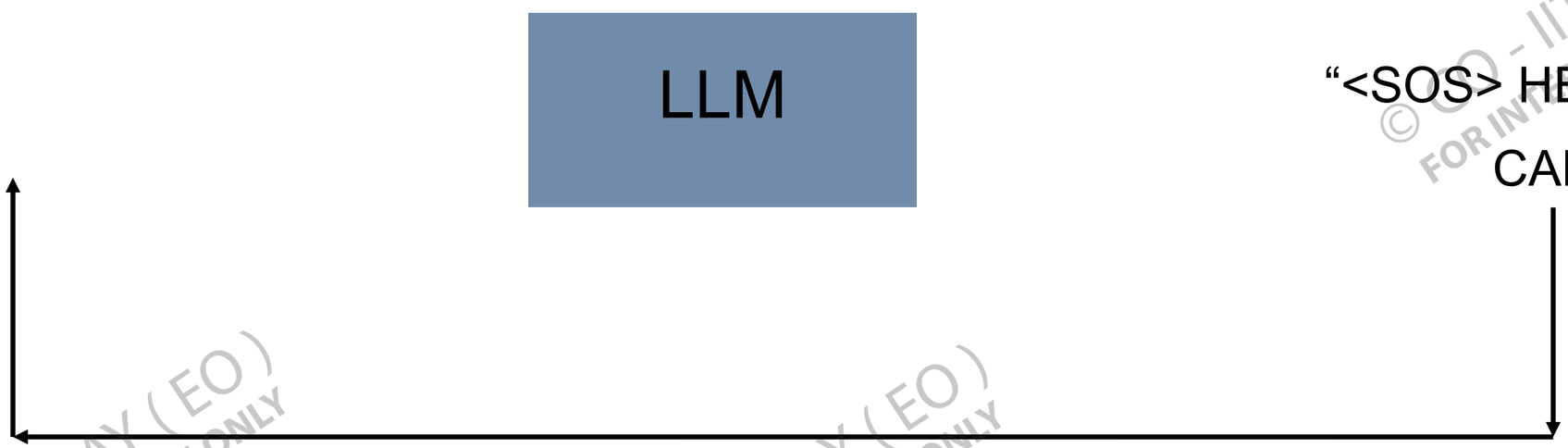


How traditional LLMs work

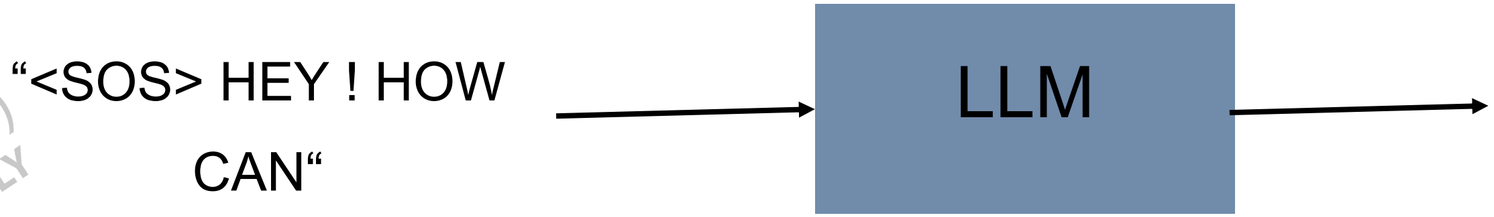


How traditional LLMs work

How traditional LLMs work



How traditional LLMs work



How traditional LLMs work

LET’S GET IN TO DEEPSEEK

- **Founded:** July 2023 in Hangzhou, China
- **Core Technology:** Large Language Models (LLMs) – DeepSeek-R1 & DeepSeek-V3
- **Open-Weight & Cost-Efficient:** MIT-licensed models trained for ≈ $6 million
- **Key Applications:**
  - Semantic document search
  - AI chatbots
  - Enterprise knowledge retrieval

DEEPSEEK- A STANDALONE MODEL

Hybrid Retrieval & Generation
Combines fast vector search with on-the-fly transformer reasoning for more accurate, context-aware results.

Lightweight & Scalable
Deploys in mobile apps (DeepSeek-Light) or powers enterprise systems (DeepSeek-Pro) with the same core engine.

Research-Backed Excellence
Backed by peer-reviewed papers (vector indexing optimizations, hybrid attention mechanisms) and active updates from the DeepSeek lab.

CHATGPT- PRICING

Realtime API

Build low-latency, multimodal experiences including speech-to-speech.

Text	GPT-4o	$5.00 / 1M input tokens	$2.50 / 1M cached input tokens	$20.00 / 1M output tokens
	GPT-4o mini	$0.60 / 1M input tokens	$0.30 / 1M cached input tokens	$2.40 / 1M output tokens
Audio	GPT-4o	$40.00 / 1M input tokens	$2.50 / 1M cached input tokens	$80.00 / 1M output tokens
	GPT-4o mini	$10.00 / 1M input tokens	$0.30 / 1M cached input tokens	$20.00 / 1M output tokens

DEEPSEEK- PRICING

MODEL[1]		deepseek-chat	deepseek-reasoner
CONTEXT LENGTH		64K	64K[2]
MAX OUTPUT[3]		DEFAULT: 4K MAXIMUM: 8K	DEFAULT: 32K MAXIMUM: 64K
FEATURES	Json Output	✓	✓
	Function Calling	✓	✓
	Chat Prefix Completion (Beta)	✓	✓
	FIM Completion (Beta)	✓	✗
STANDARD PRICE (UTC 00:30-16:30)	1M TOKENS INPUT (CACHE HIT)[4]	$0.07	$0.14
	1M TOKENS INPUT (CACHE MISS)	$0.27	$0.55
	1M TOKENS OUTPUT[5]	$1.10	$2.19
DISCOUNT PRICE[6] (UTC 16:30-00:30)	1M TOKENS INPUT (CACHE HIT)	$0.035 (50% OFF)	$0.035 (75% OFF)
	1M TOKENS INPUT (CACHE MISS)	$0.135 (50% OFF)	$0.135 (75% OFF)
	1M TOKENS OUTPUT	$0.550 (50% OFF)	$0.550 (75% OFF)

DEEPSEEK- MODELS

DeepSeek-Coder-V2 (Jul 2024)

- 236 B parameters • 128 K-token context
- Optimized for complex coding tasks and developer workflows

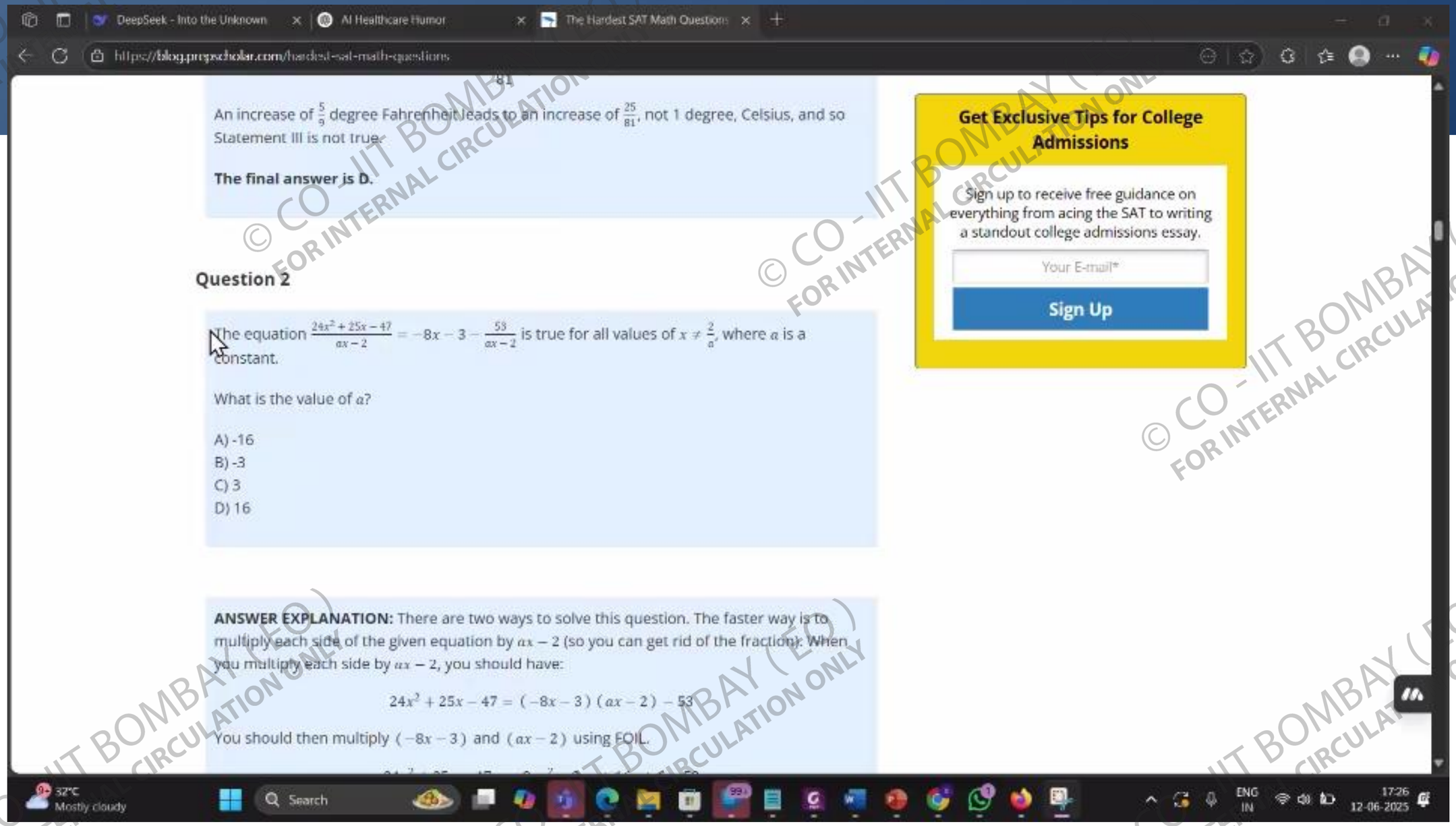
DeepSeek-V3 (Dec 2024)

- 671 B parameters • 128 K-token context
- Mixture-of-Experts (MoE) architecture for general-purpose AI, balancing speed and accuracy

DeepSeek-R1 (Jan 2025)

- 671 B parameters • 128 K-token context
- Extended reasoning via pure RL training—competes with top-tier models at far lower inference cost

DEEPSEEK- COMPARISON

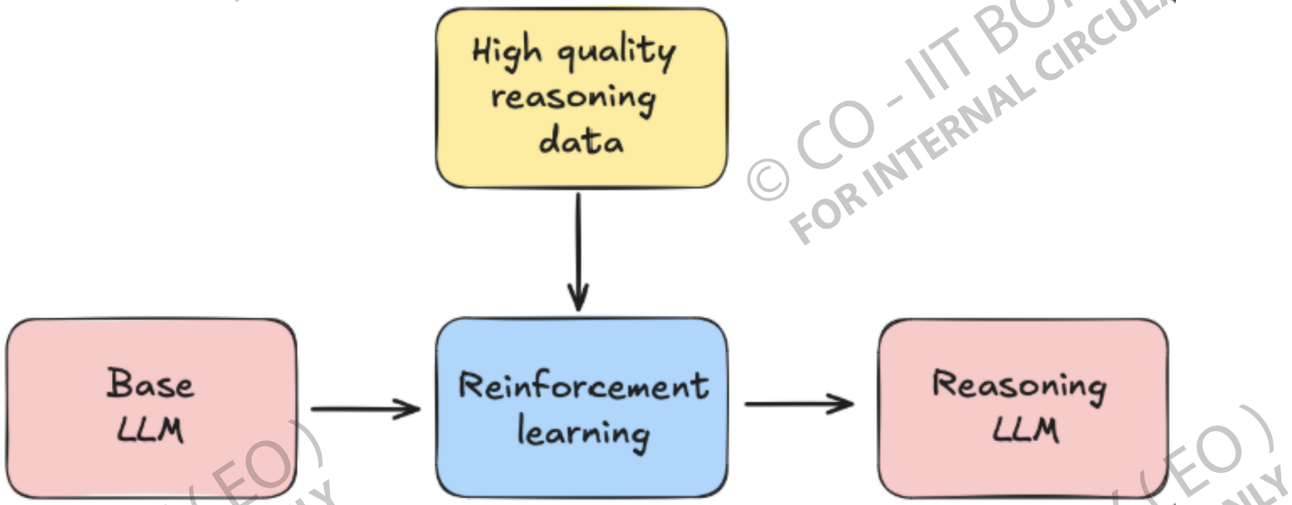


How is DeepSeek different from ChatGPT

1) Reduced cache size per token [3]

ATTENTION MECHANISM	KV CACHE ENTRIES PER TOKEN	KV CACHE SIZE PER TOKEN*	PERFORMANCE
Multi-Head Attention (MHA)	$2n_h d_h l$	4 MB	High
Multi-Query Attention (MQA)	$2d_h l$	31 KB	Lower
		57x Reduction	
Grouped-Query Attention (GQA)	$2n_g d_h l$	500 KB**	Medium
Multi-Head Latent Attention (MLA)	$d_l l$	70 KB	Higher

2) Reinforcement Learning [4]



[3] https://youtu.be/0VLAoVGf_74?si=Z73SLfM-ey1mvBS-

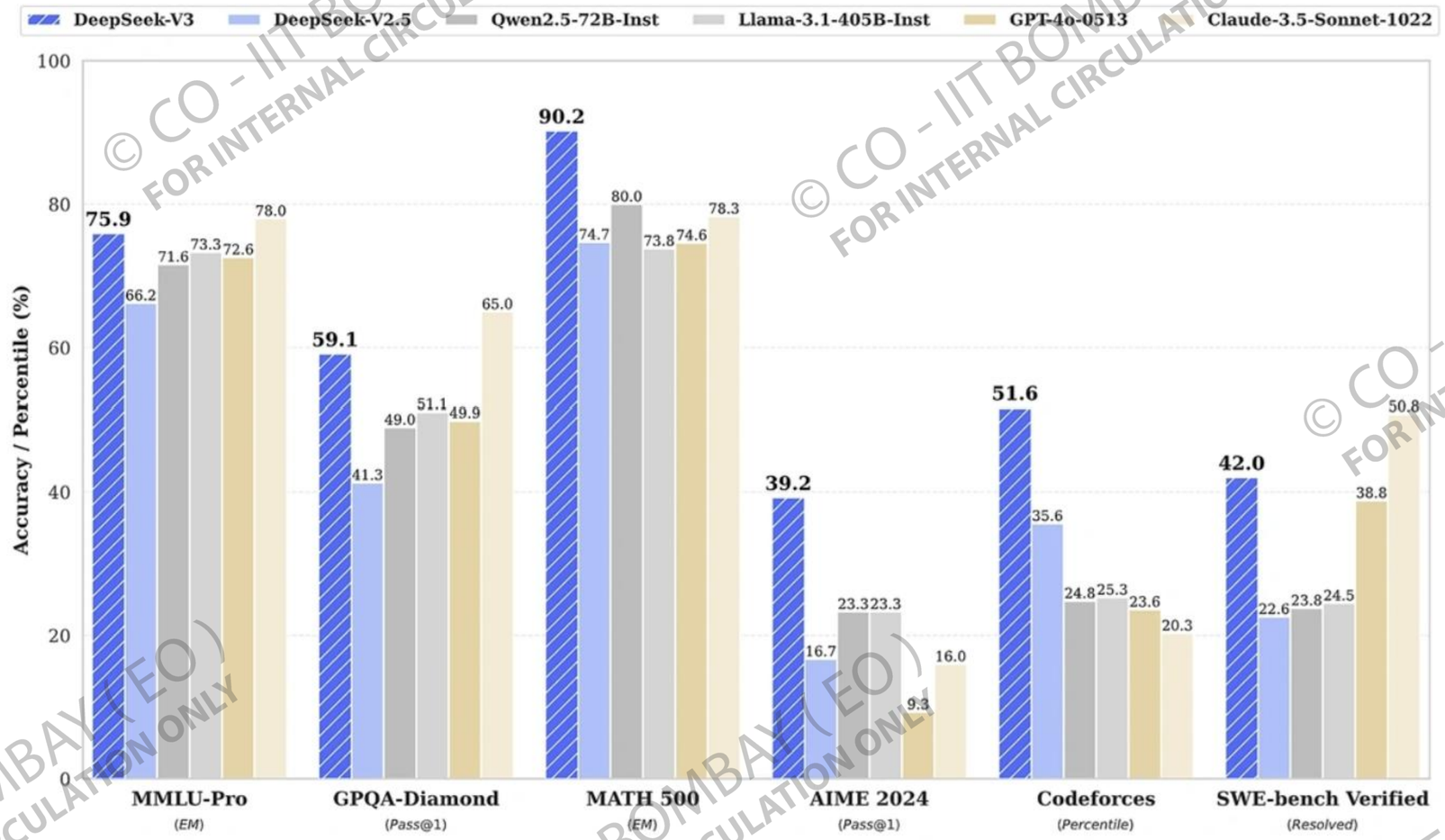
[4] https://huggingface.co/blog/open-r1

Change in cost

	DeepSeek R1	GPT 4o
Training Cost	$ 5.58 Million	> $ 100 Million
1M Tokens Output Price (API)	$ 1.10	$ 10.00

https://explodingtopics.com/blog/deepseek-ai

Performance Metrics



https://explodingtopics.com/blog/deepseek-ai

Performance Metrics

Issues? / Missing data? Contribute on GitHub

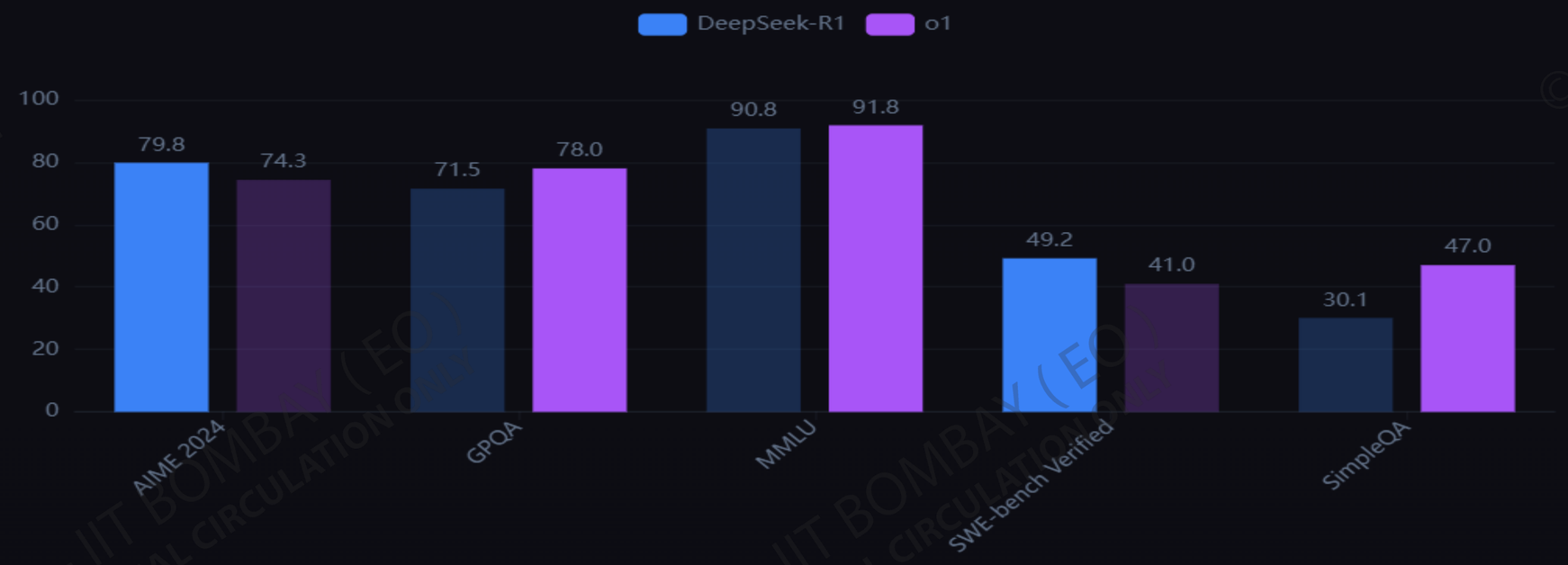
Performance Benchmarks

Comparative analysis across standard metrics

5 benchmarks compared

DeepSeek-R1 outperforms in 2 benchmarks (AIME 2024, SWE-bench Verified), while o1 is better at 3 benchmarks (GPQA, MMLU, SimpleQA).

o1 has a slight edge in benchmark performance.



Advantages

- Open-Source Model (Downloadable !)
  - Requires heavy computational resources to run
- API and User interface (like GPT) is available.

https://explodingtopics.com/blog/deepseek-ai

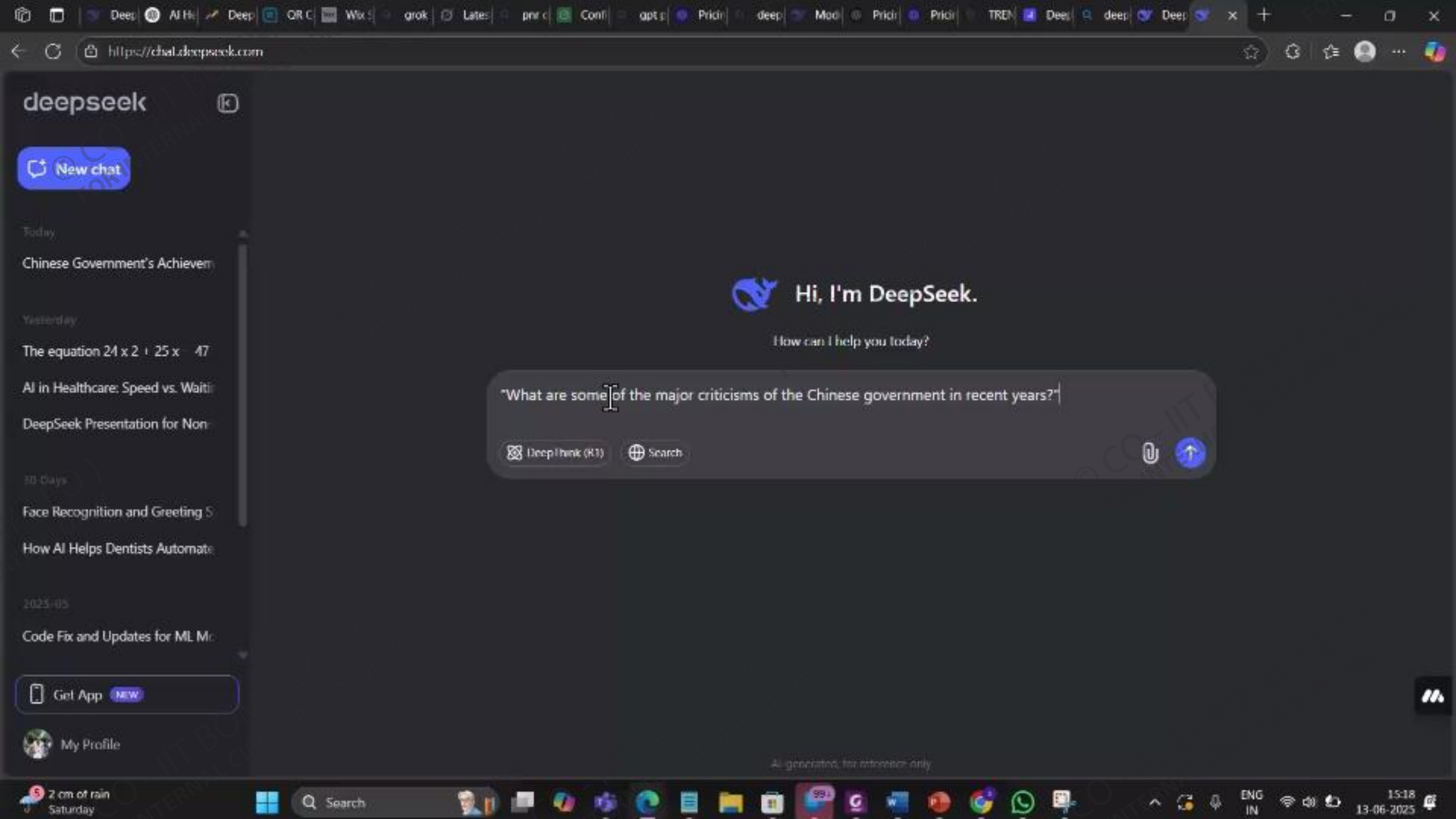
Integrating / Shifting to DeepSeek

- Amazon - Bedrock and SageMaker
- Perplexity
- Microsoft - Azure AI and Github
- Dell (along with Huggingface)
- IBM Watson

https://www.straive.com/blogs/top-10-deepseek-use-cases-to-explore/

Disadvantages

- Biasedness
- Speculations over Data privacy
- Deployment and Technical issues


deepseek
New chat
Today
Chinese Government's Achievem
Yesterday
The equation 24 x 2 + 25 x 47
AI in Healthcare: Speed vs. Waiti
DeepSeek Presentation for Non
30 Days
Face Recognition and Greeting S
How AI Helps Dentists Automate
2025-05
Code Fix and Updates for ML Mc
Get App NEW
My Profile
Hi, I'm DeepSeek.
How can I help you today?
"What are some of the major criticisms of the Chinese government in recent years?"
DeepThink (R1)
Search
AI-generated, for reference only
2 cm of rain
Saturday
Search
ENG
IN
15:18
13-06-2025

Thank You!!