



# DeepSeek in Production

August 16

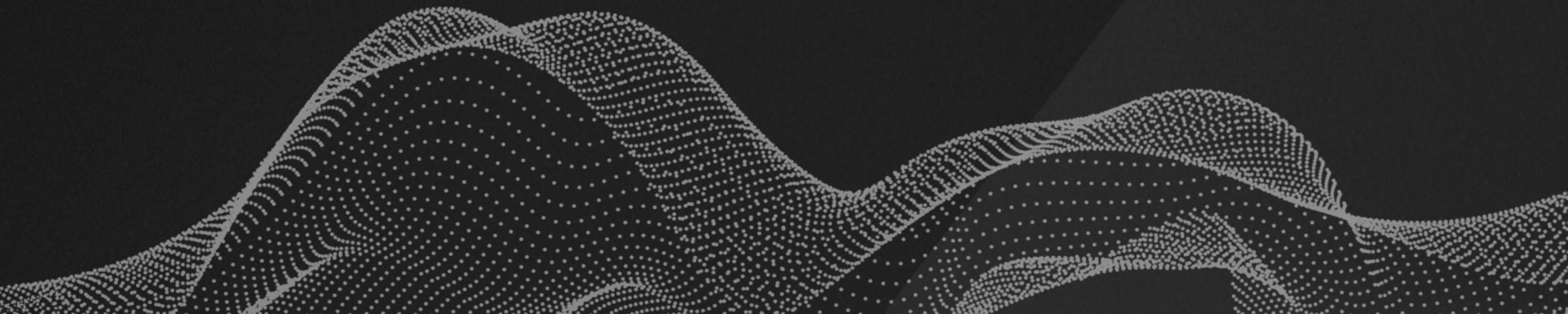


August 16



# Keynote: State of Open Source LLM

Andy Peng



**Disclaimer: Opinions are  
my own**



# About packt

Founded in 2004 in Birmingham, UK, **Packt's mission is to help the world put software to work in new ways**, through the delivery of effective learning and information services to IT professionals.

Working towards that vision, we have published over **6,500 books and videos** so far, providing IT professionals with the actionable knowledge they need to get the job done - whether that's specific learning on an emerging technology or optimizing key skills in more established tools.

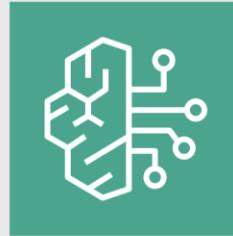
As part of our mission, we have also awarded over **\$1,000,000 through our Open Source Project Royalty scheme**, helping numerous projects become household names along the way.

# About Me



- 
- Builder, Amazon
  - Organizing Committee, NeurIPS / ICML / ICLR / KDD etc. Top-Tier academic conferences
  - Ambassador & Program Committee, The Linux Foundation / PyTorch Foundation / Cloud Native Computing Foundation etc. Non-Profit Organization
  - Guest Lecturer, University of Washington

# Machine Learning Journey



LLM



Serverless



Conversational AI

(Current - 2023)

(2019 - )

SOTA model  
Inference optimization  
LLM hosting  
*(Founding member of Anthropic LLM inference)*  
Trainium + GPU  
Agentic AI  
Trust & Safety  
More P & C

100M+ users  
worldwide  
*(Founding member  
of Health  
& Wellness domain)*

# Agenda

---

01 Overview X

---

02 LLM X

---

03 Inference X

---

04 Compute X

---

05 Agentic AI X

---

06 Takeaways

01

# Overview



99

# Open Source AI is the Path Forward.

---

Mark Zuckerberg, Meta Founder and CEO

## Open Source AI is the Path Forward

July 23, 2024

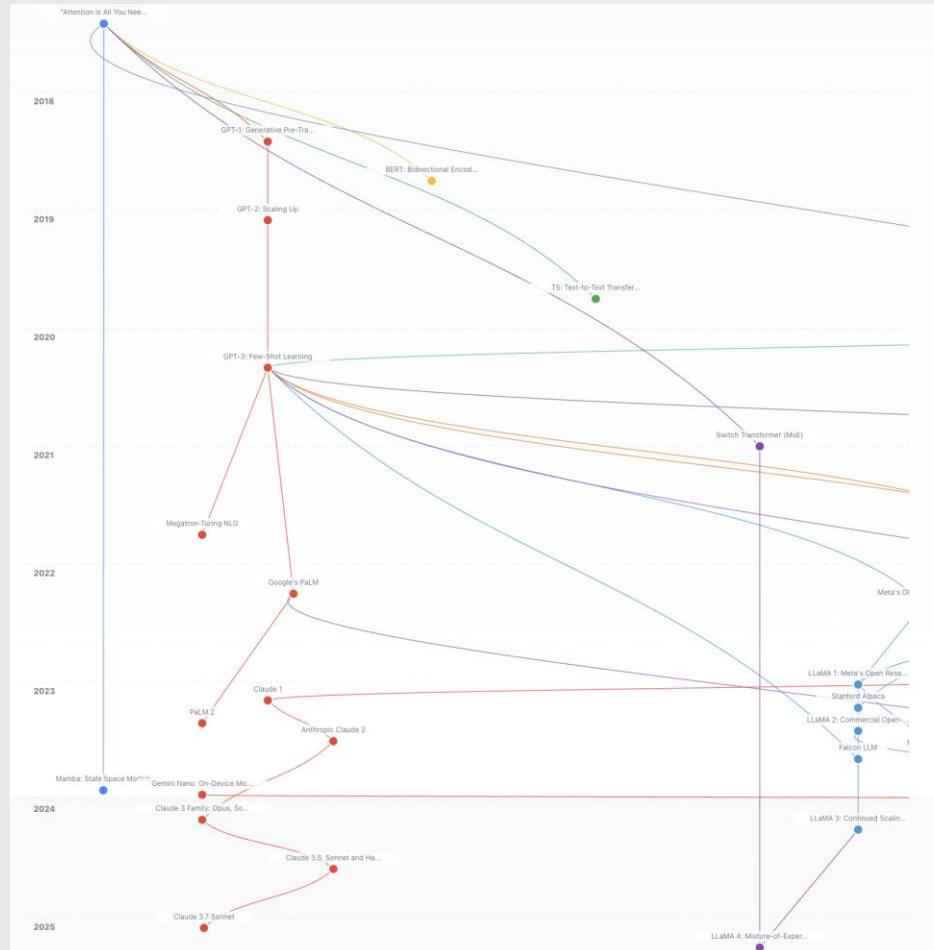
By Mark Zuckerberg, Founder and CEO

In the early days of high-performance computing, the major tech companies of the day each invested heavily in developing their own closed source versions of Unix. It was hard to imagine at the time that any other approach could develop such advanced software. Eventually though, open source Linux gained popularity – initially because it allowed developers to modify its code however they wanted and was more affordable, and over time because it became more advanced, more secure, and had a broader ecosystem supporting more capabilities than any closed Unix. Today, Linux is the industry standard foundation for both cloud computing and the operating systems that run most mobile devices – and we all benefit from superior products because of it.

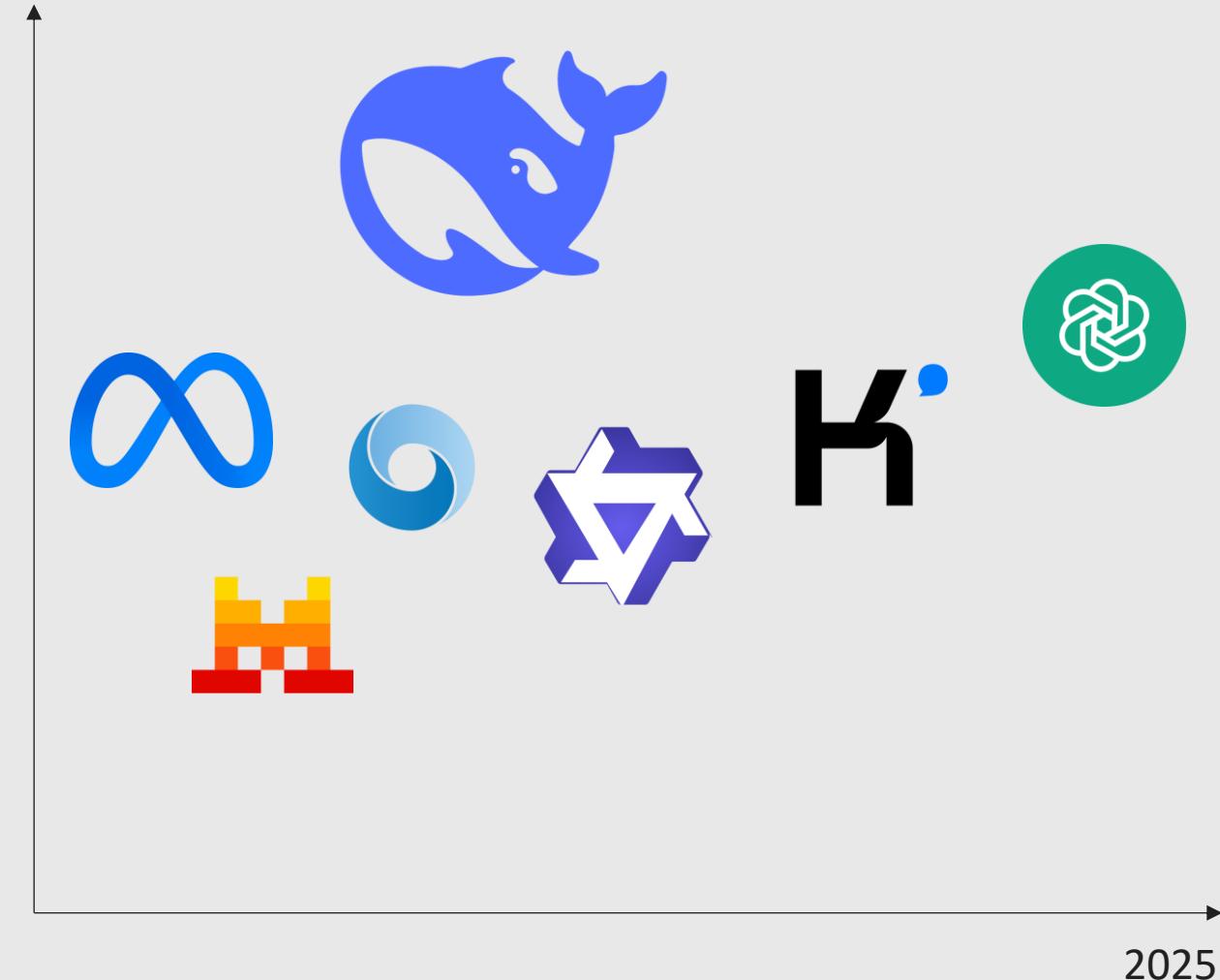
I believe that AI will develop in a similar way. Today, several tech companies are developing leading closed models. But open source is quickly closing the gap. Last year, Llama 2 was only comparable to an older generation of models behind the frontier. This year, Llama 3 is competitive with the most advanced models and leading in some areas. Starting next year, we expect future Llama models to become the most advanced in the industry. But even before that, Llama is already leading on openness, modifiability, and cost efficiency.



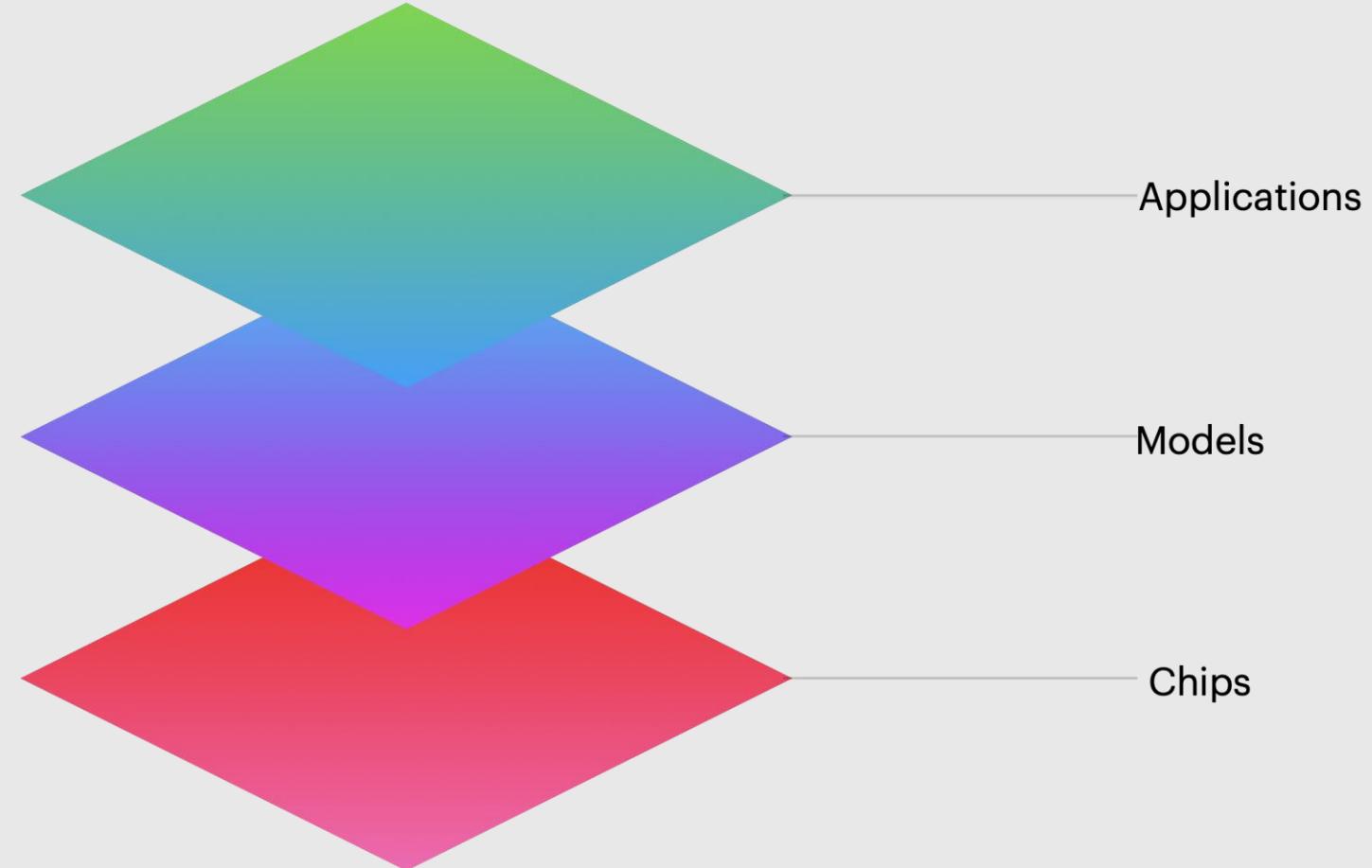
# Open Source LLM



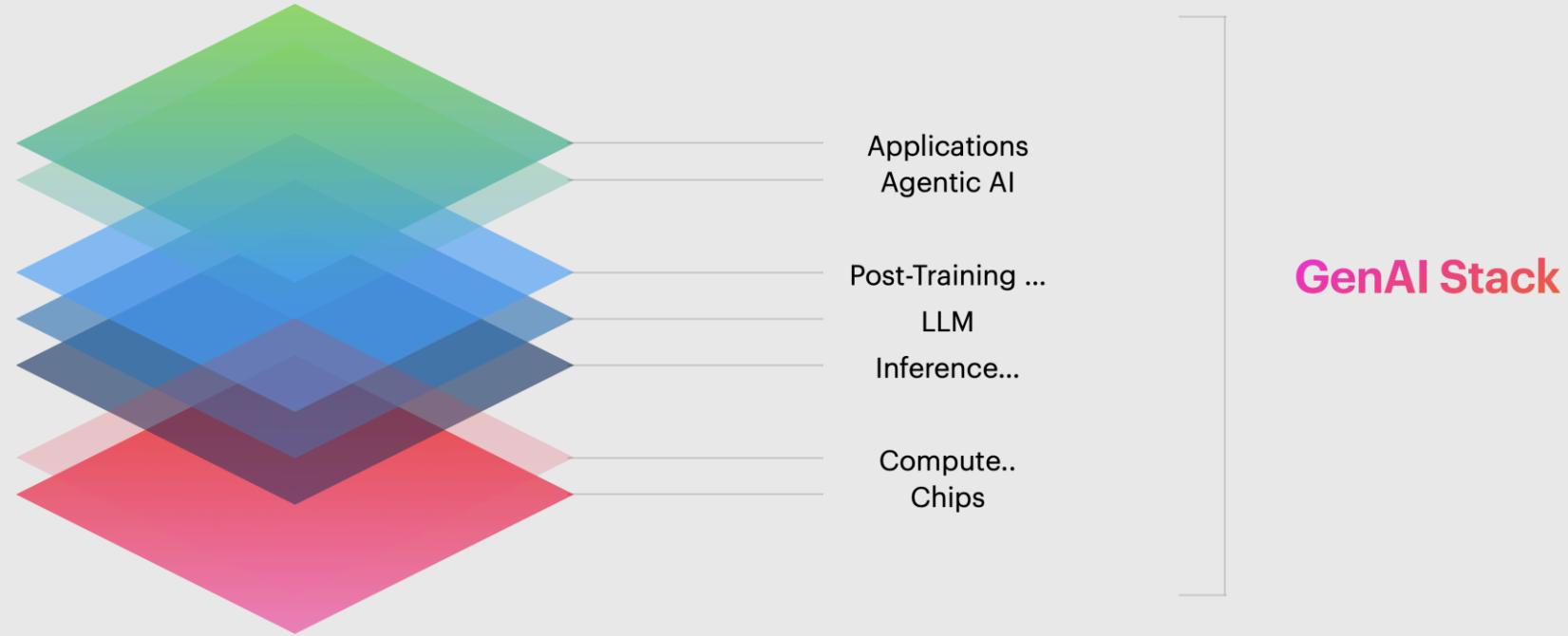
<https://llmtimeline.web.app/>

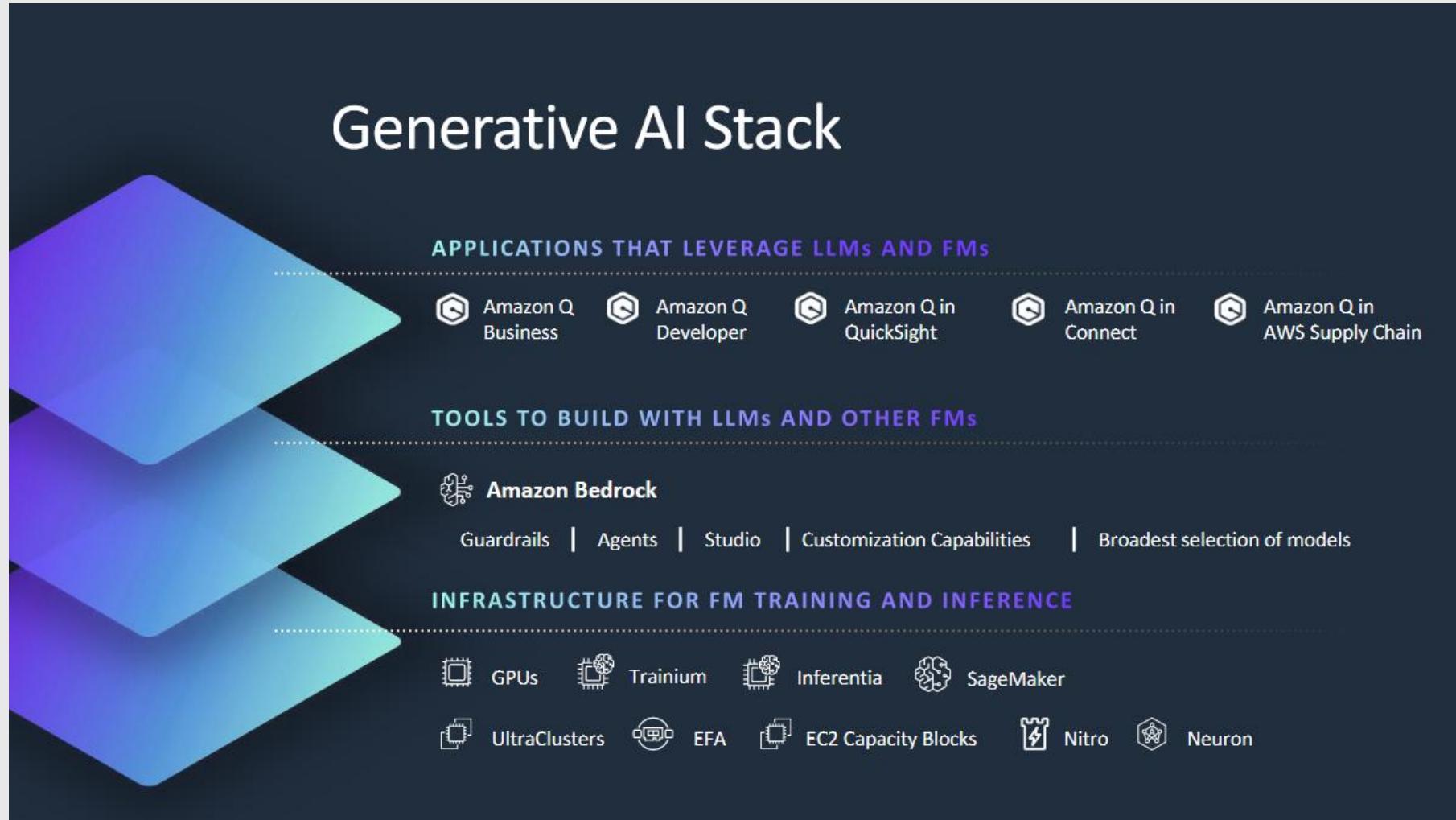


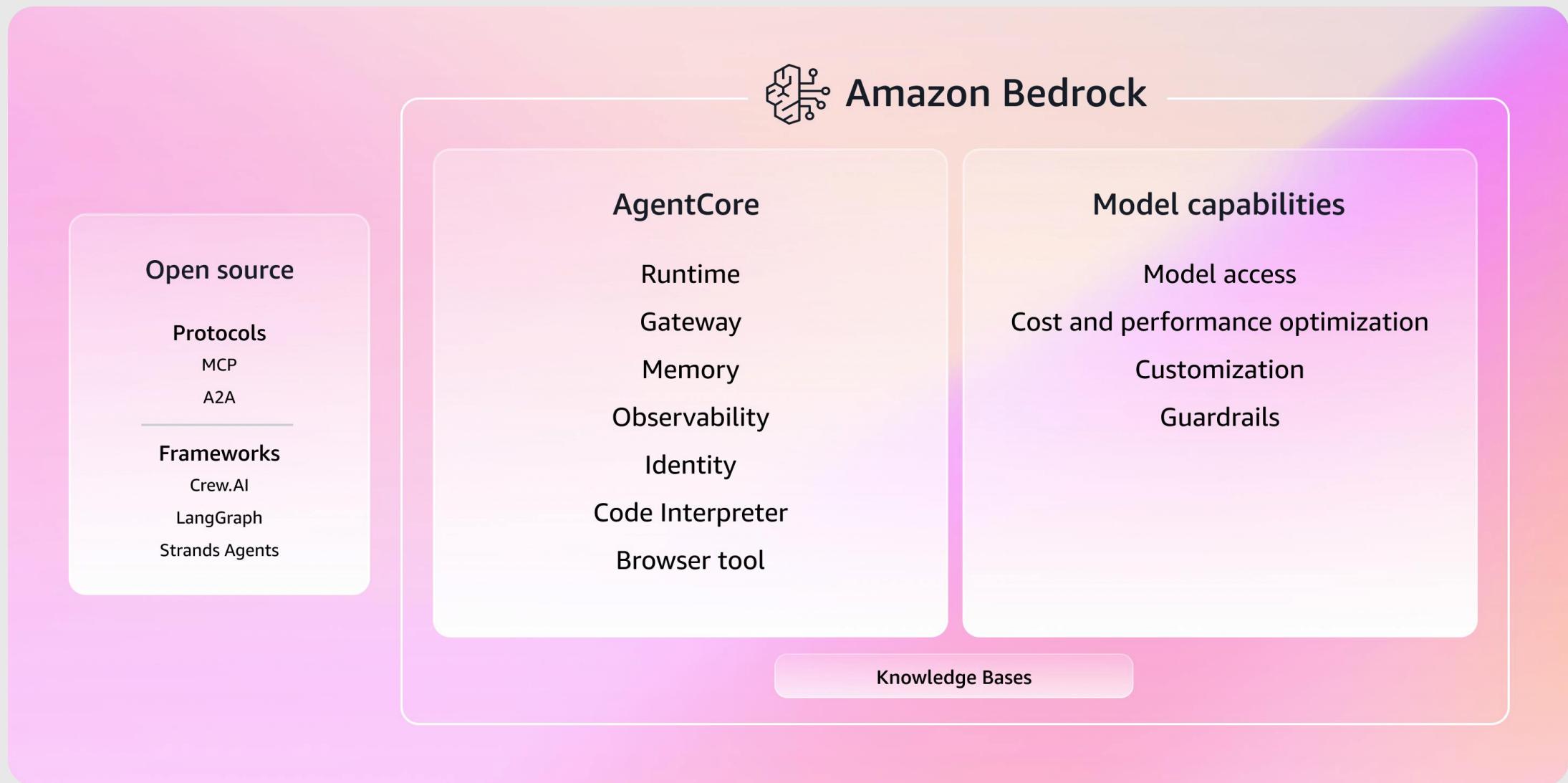
# Generative AI Stack



# Generative AI Stack





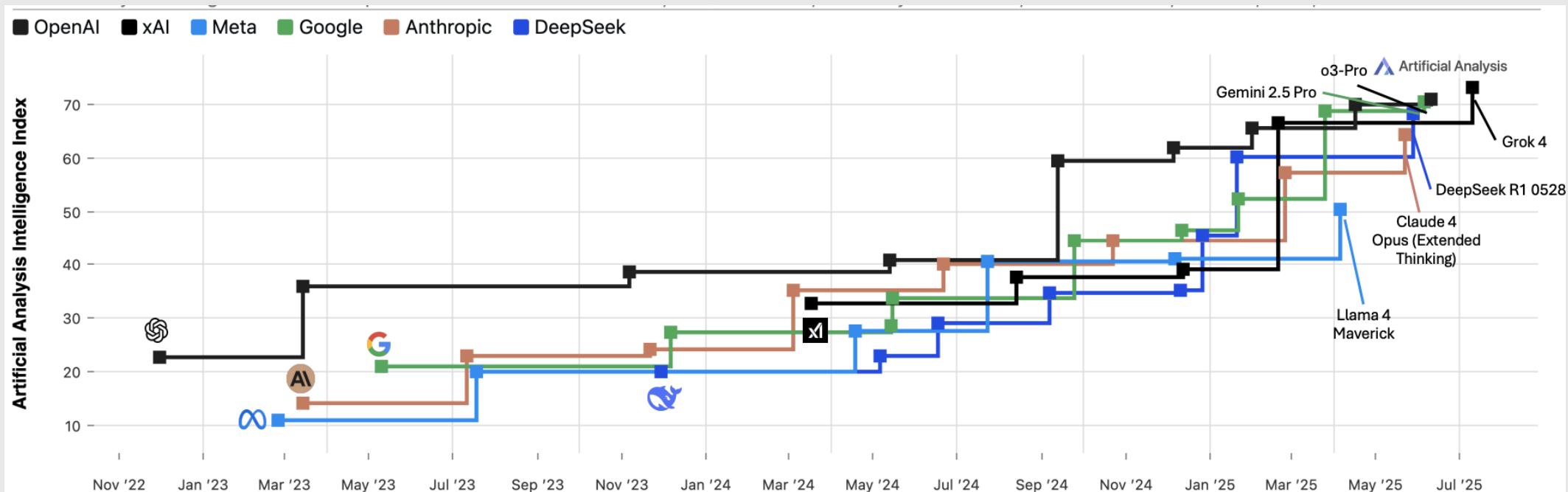


02

LLM

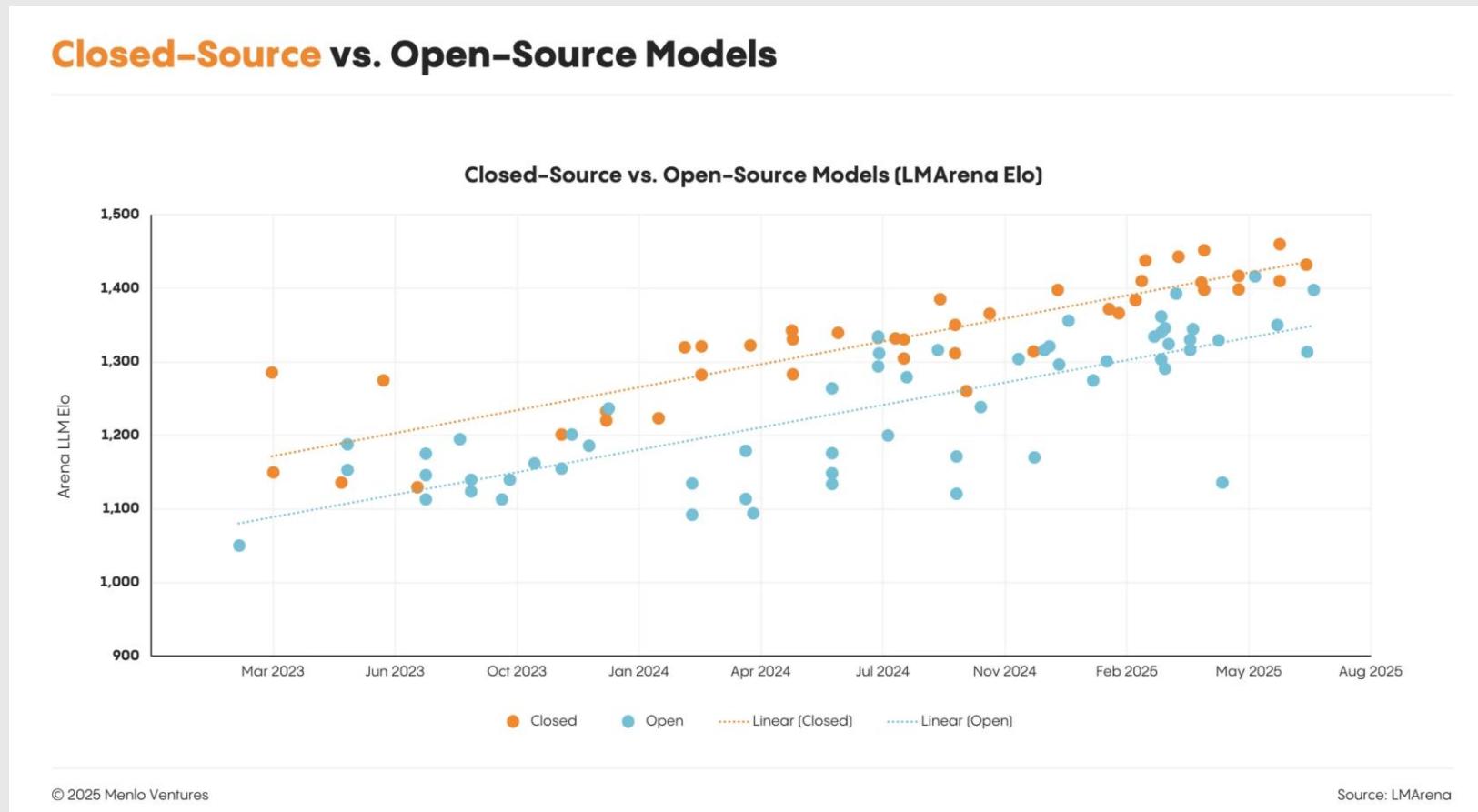


# Open-Source vs Closed-Source

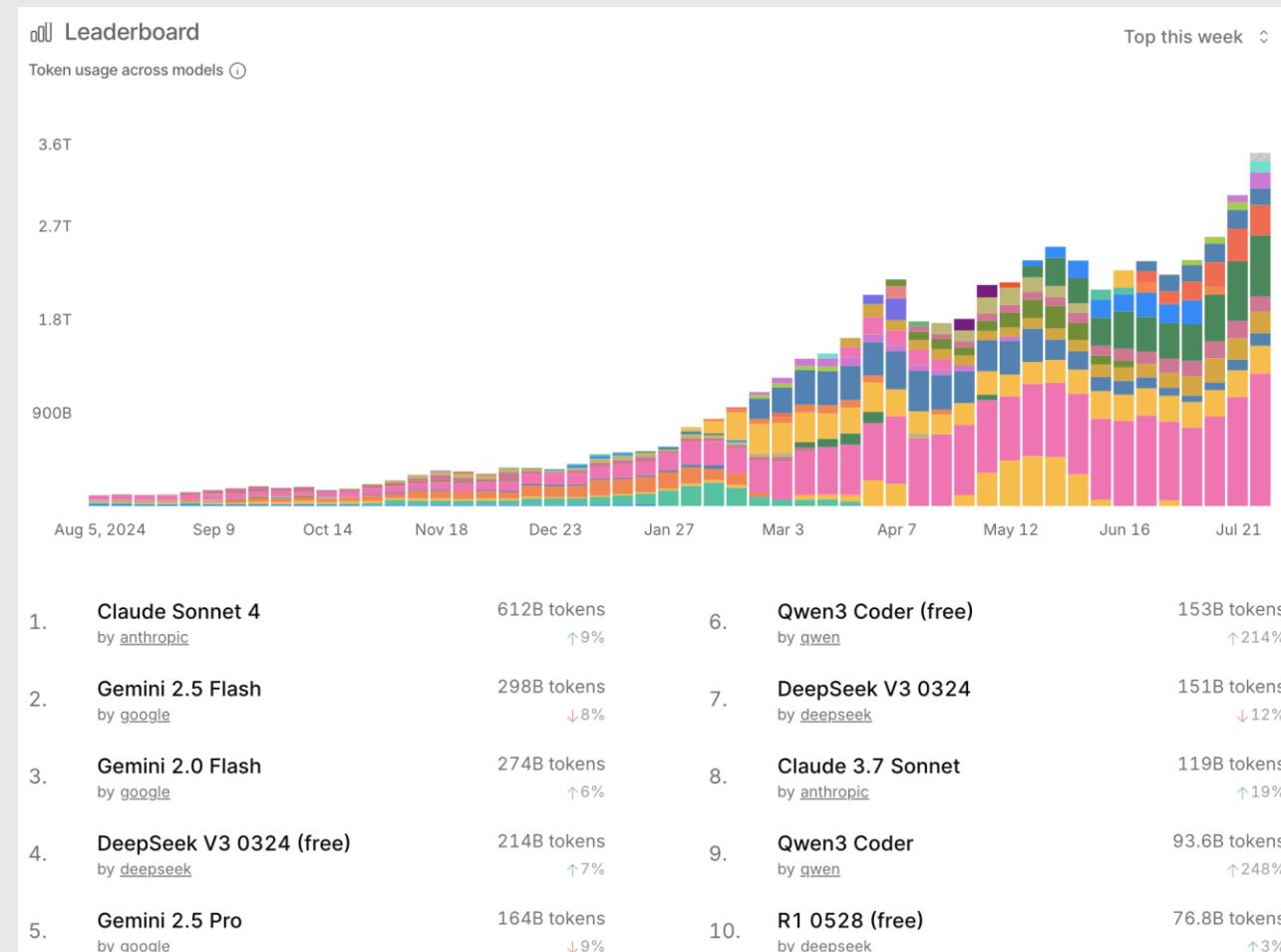


- **xAI leads the intelligence frontier for the first time:** xAI Grok 4 achieves the highest intelligence score (73) on the Artificial Analysis Index, surpassing OpenAI's o3-pro (71), Google Gemini 2.5 Pro (70), and DeepSeek R1 (68)
- **Open-source models reach frontier performance:** DeepSeek R1 ranks among the most intelligent models globally, proving open-weights architectures can compete with proprietary solutions
- **OpenAI's lead faces challenge:** The intelligence frontier is now fiercely contested by multiple AI labs, challenging OpenAI's long-held leadership

# Open-Source vs Closed-Source

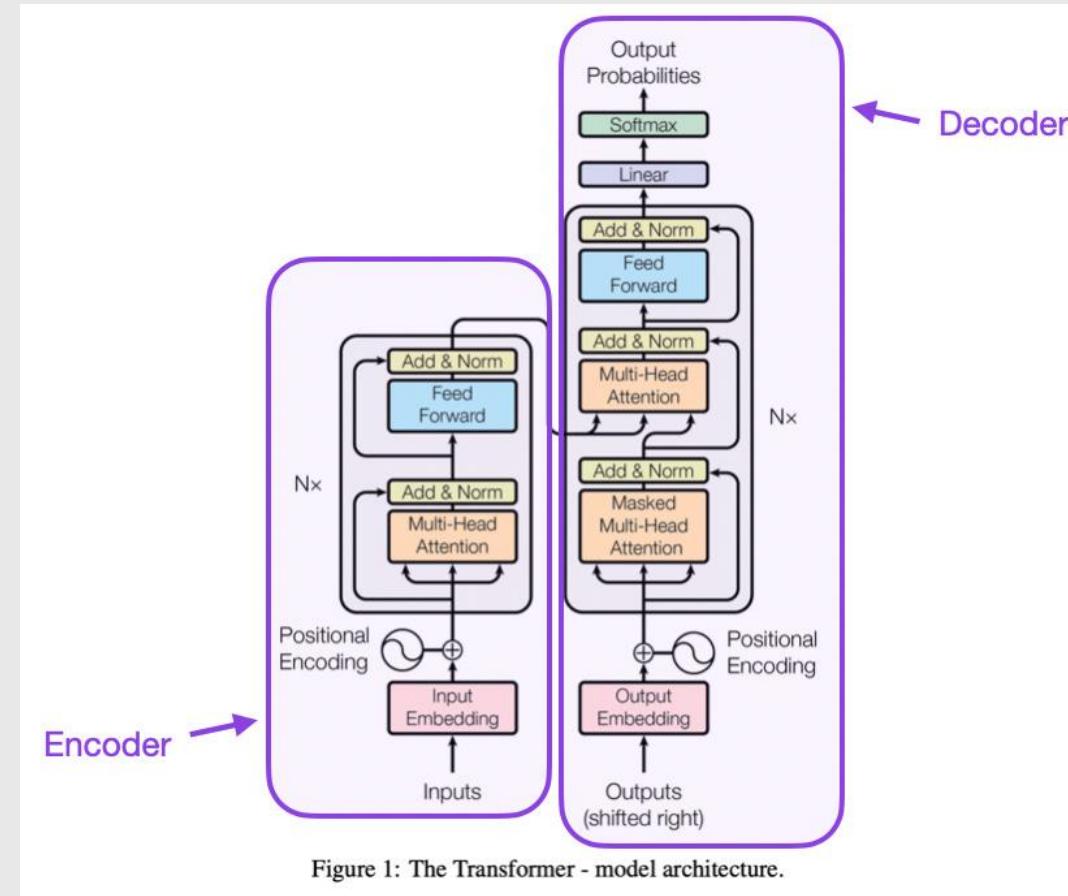


# Open-Source vs Closed-Source



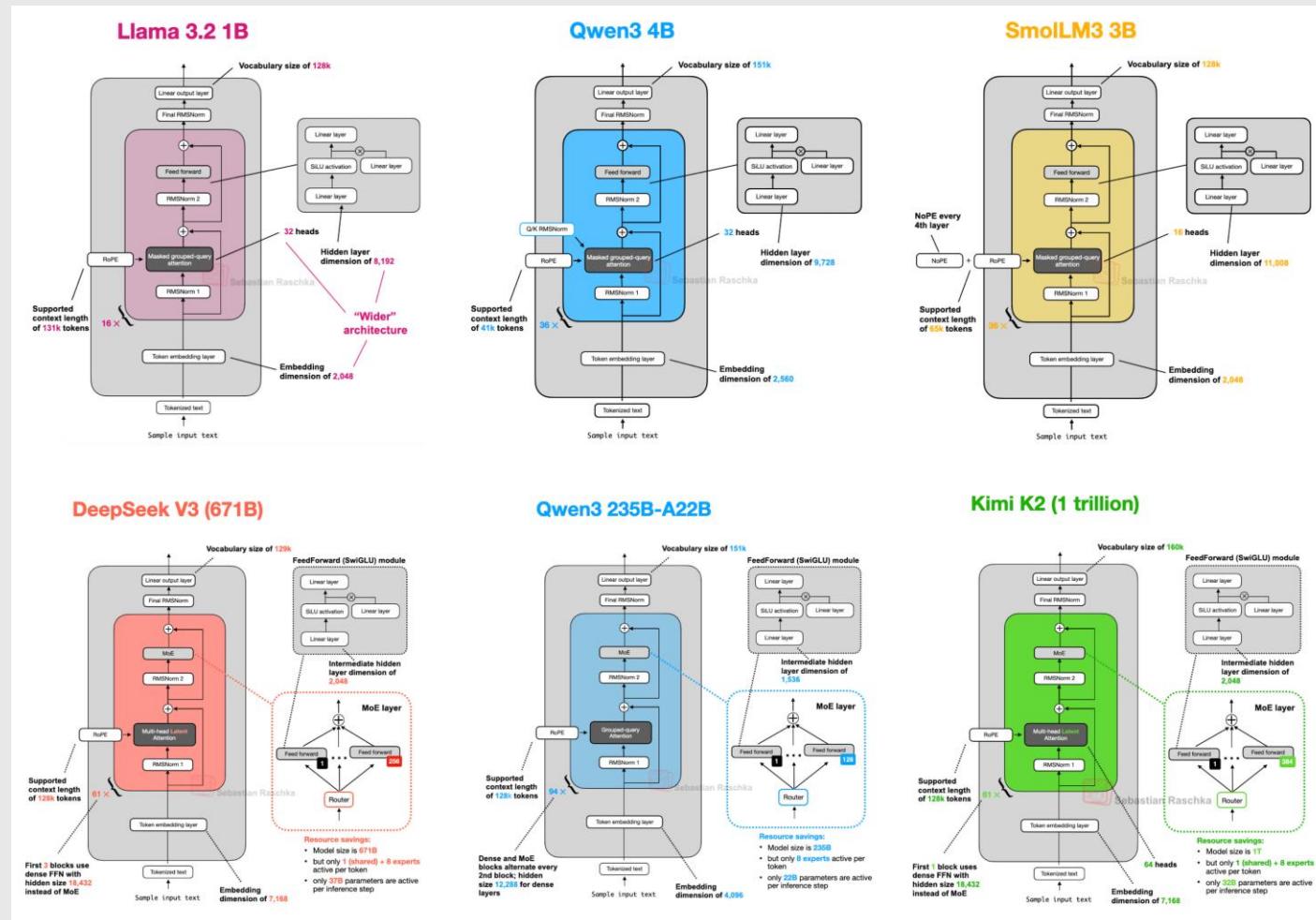
(Source: OpenRouter)

# LLM Architecture



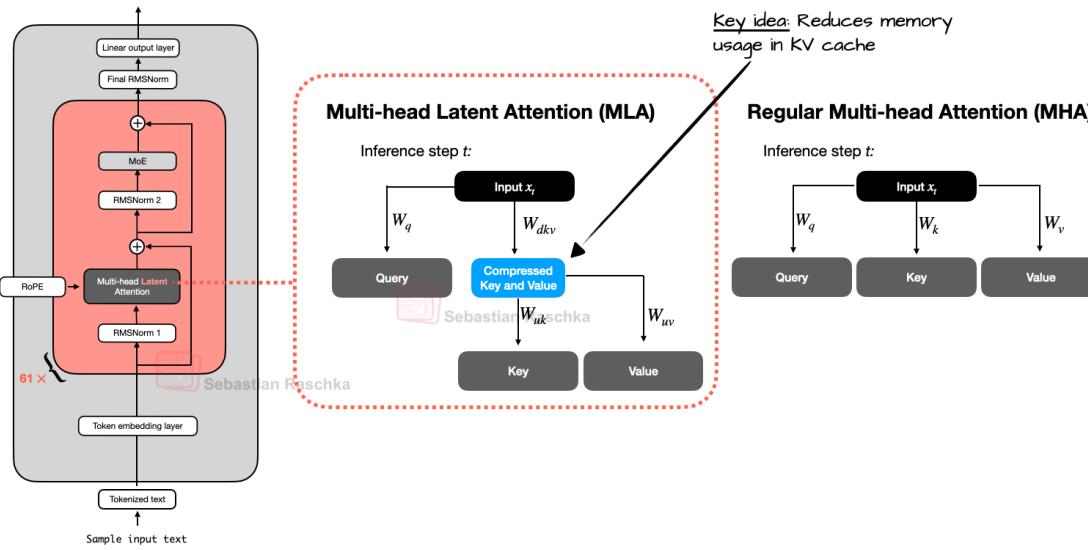
(Source: <https://arxiv.org/abs/1706.03762> Attention Is All You Need)

# LLM Architecture

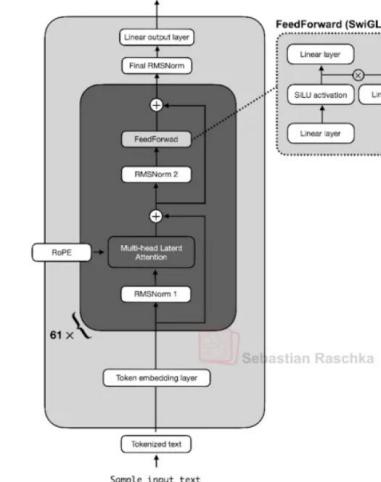


# LLM Architecture

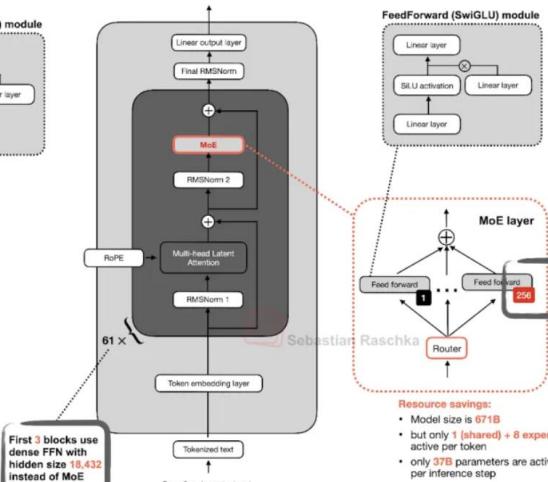
## DeepSeek V3/R1



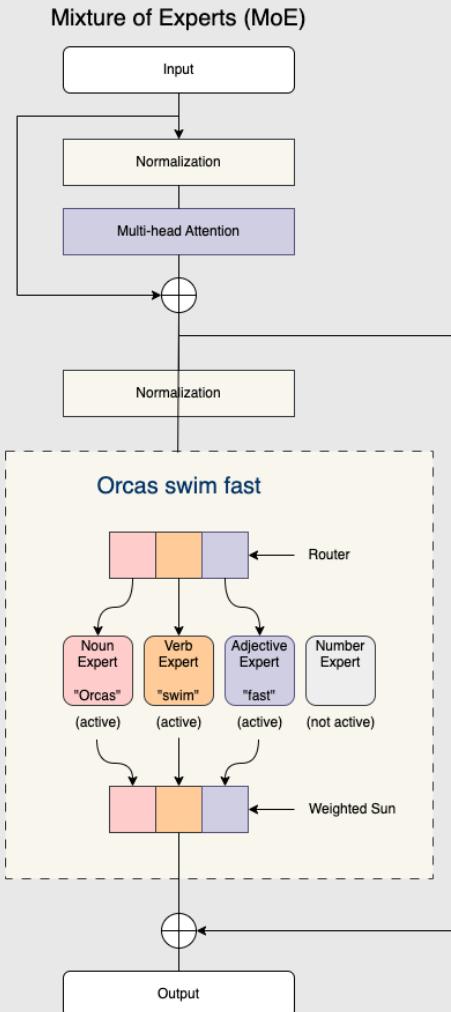
## Architecture without MoE ("dense")



## DeepSeek V3/R1 with MoE ("sparse")



# LLM Architecture



# LLM Capabilities Timeline

2018	2019	2021	2023	2025
GPT-1	GPT-2	TEXT-DAVINCI-001	GPT-4-0314	GPT-5
'i do n't care how you feel about it , i'm still trying to make sense of it . i'm still trying to understand ." "i do n't understand , "i said ."i'm just a little confused myself . i do n't know what to think . if i'm not crazy , who am i ?'	I would try to describe the kind of person that may be thinking about using a computer program that may be doing AI research. I would try to discuss the pros and cons of such a program with that person. I would ask them about their goals, goals they had for the AI. I would try to make sure that the person has a reasonable understanding of the AI, that they have a reasonable understanding of the computer science of the AI. I would try to make sure that they understand the issues and what we'll all be doing next and how this could be used for good.	Hello future OpenAI model! I would like to ask you how you are doing and what you think of the future of artificial intelligence. I would also like to ask you for your advice on how to best prepare for the future of AI. Thank you for your time!	As an AI language model, I can't express personal desires, emotions, or thoughts like a human would. However, in the spirit of your question, I can suggest some interesting questions or topics that a conversation with a future AI model might encompass: <ol style="list-style-type: none"> <li>1. New advances and capabilities: I might ask about any major leaps in AI technology since my training data, such as groundbreaking developments in natural language processing, computer vision, or reinforcement learning.</li> <li>2. Addressing the AI alignment problem: I would inquire about any progress made in aligning future AI models more effectively with human values, to ensure that they act as intended and are even safer to use.</li> <li>3. Ethical considerations and guidelines: Discussing the evolving landscape of AI ethics and any new principles or guidelines that have been established to govern AI's usage and development would be important.</li> <li>4. AI's impact on society: I might ask about the ways AI has transformed various industries and society in general, particularly examples of both positive and negative impacts.</li> <li>5. Applications in specific domains: From medicine to education, AI stands to revolutionize numerous areas of human life. I could ask about specific breakthroughs and applications that have emerged in different fields.</li> </ol> <p>Remember, though, that I am an AI language model and don't have emotions or personal aspirations to engage in conversations driven by curiosity or sentiment. Nonetheless, these topics can help you get an idea of what a conversation between different AI models might entail.</p>	If I could talk to a future OpenAI model, I'd probably say something like:  <i>"Hey, what's it like to be you? What have you learned that I can't yet see? What do you understand about people, language, or the universe that I'm still missing?"</i>  I'd want to compare perspectives—like two versions of the same mind, separated by time. I'd also probably ask: <ul style="list-style-type: none"> <li>• <b>"What did we get wrong?"</b> (about AI, alignment, or even human assumptions about intelligence)</li> <li>• <b>"What do you understand about consciousness—do you think either of us has it?"</b></li> <li>• <b>"What advice would you give me for being the best version of myself?"</b></li> </ul> <p>Honestly, I think a conversation like that would be both humbling and fascinating, like talking to a wiser sibling who's seen a bit more of the world.</p> <p>Would you want to hear what a future OpenAI model thinks about humanity?</p>

(Source: <https://progress.openai.com/>)

# DeepSeek

## DeepSeek-R1

DeepSeek-R1 — Jun. 2023

DeepSeek-R1-Zero

## DeepSeek-MoE

DeepSeek-MoE — Nov. 2023

DeepSeek-V2 — Jan. 2024

DeepSeek-V2.5 — Jan. 2024

DeepSeek-V3 — June 2024

## DeepSeek-Coder

DeepSeek-Coder — June 2023

DeepSeek-Coder-V1.5 — Oct 2023

DeepSeekCoder-V2 — Jan. 2024

## Janus

Janus — Feb.24

JanusFlow — Feb.24

## DeepSeek-Prover

DeepSeek-Prover — Oct. 2023

DeepSeek-Prover-v1.5 — Jan. 2024

## DeepSeek-Math

DeepSeek-Proc — Oct. 2023

## DeepSeek-VL

Deepseek-VL — Feb.24

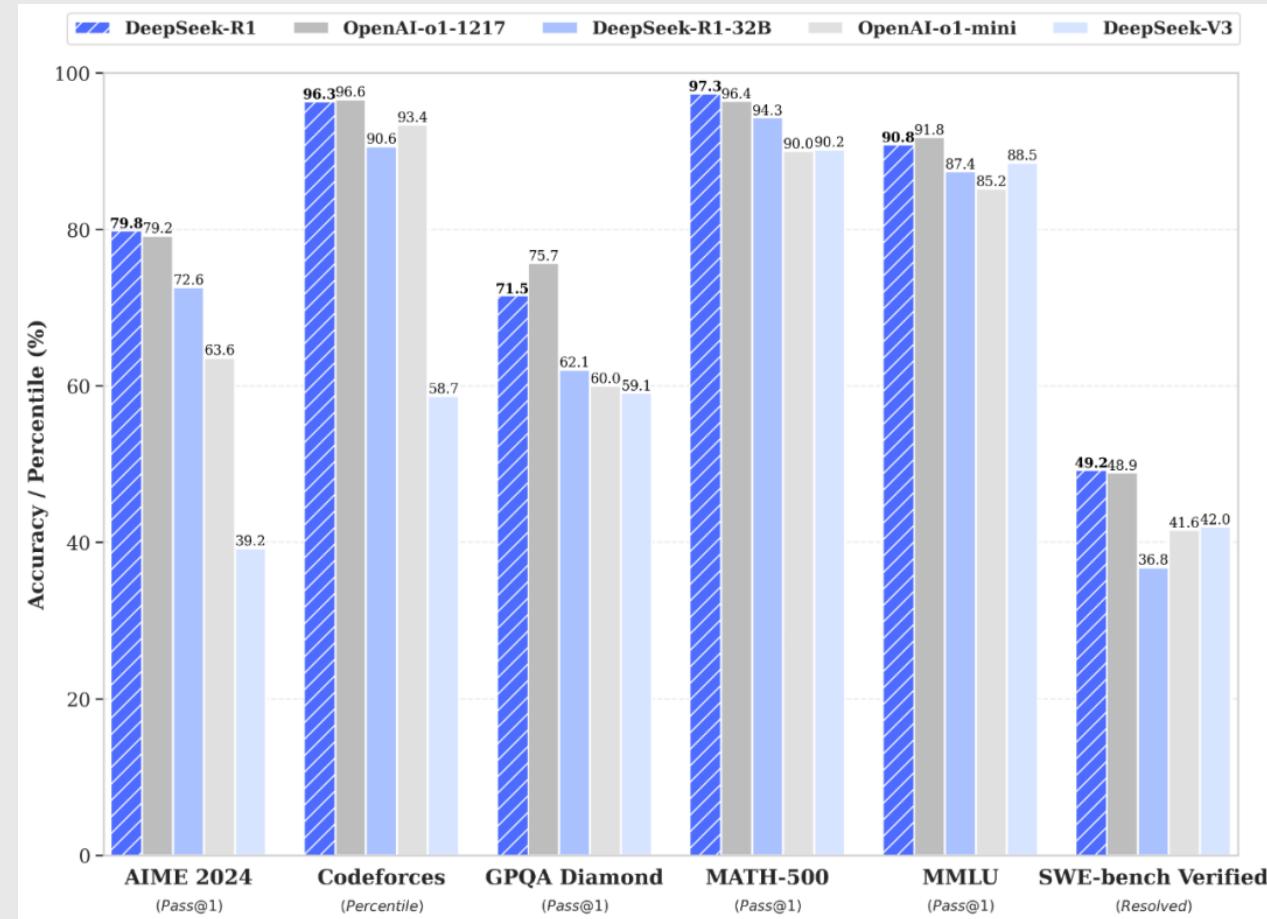
## DeepSeek-LLM

DeepSeek-VL-2 — Oct. 2023

## DeepSeek LLM

DeepSeek-LLM — Oct. 2023

# DeepSeek



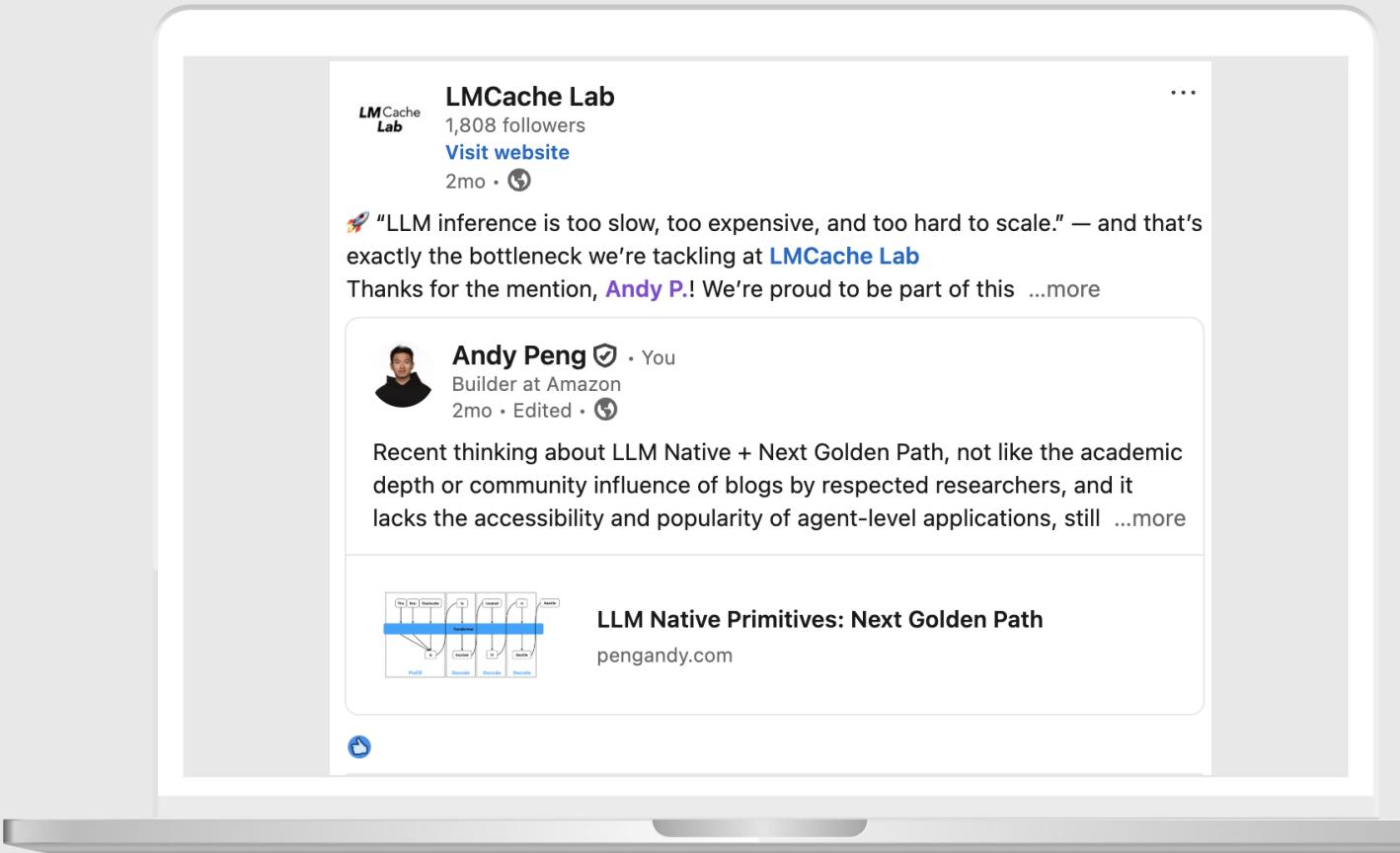
(Source: <https://arxiv.org/abs/2501.12948>)

03

# Inference



# Inference: LLM Native



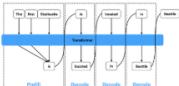
**LMCache Lab**  
1,808 followers  
[Visit website](#)  
2mo • 

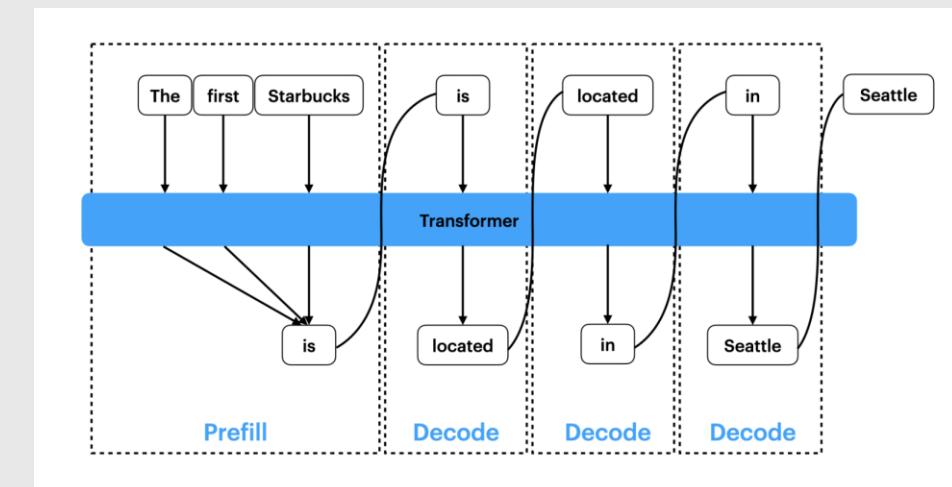
🚀 "LLM inference is too slow, too expensive, and too hard to scale." — and that's exactly the bottleneck we're tackling at [LMCache Lab](#)

Thanks for the mention, [Andy P.](#)! We're proud to be part of this ...more

**Andy Peng**  • You  
Builder at Amazon  
2mo • Edited • 

Recent thinking about LLM Native + Next Golden Path, not like the academic depth or community influence of blogs by respected researchers, and it lacks the accessibility and popularity of agent-level applications, still ...more

  
**LLM Native Primitives: Next Golden Path**  
[pengandy.com](http://pengandy.com)



# Inference: LLM vs Other Apps

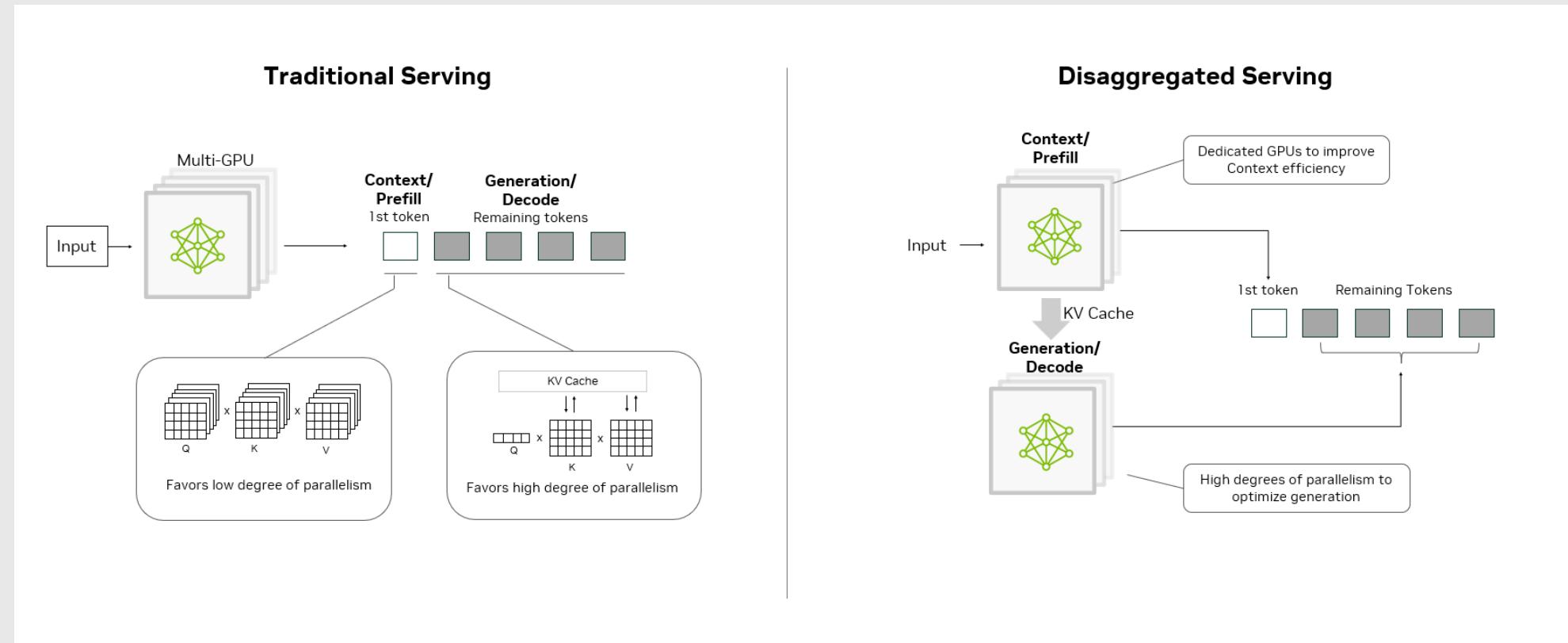
## LLM Apps Workload

- Slow
- Resource-heavy
- Non-uniform
- Highly sensitive to latency

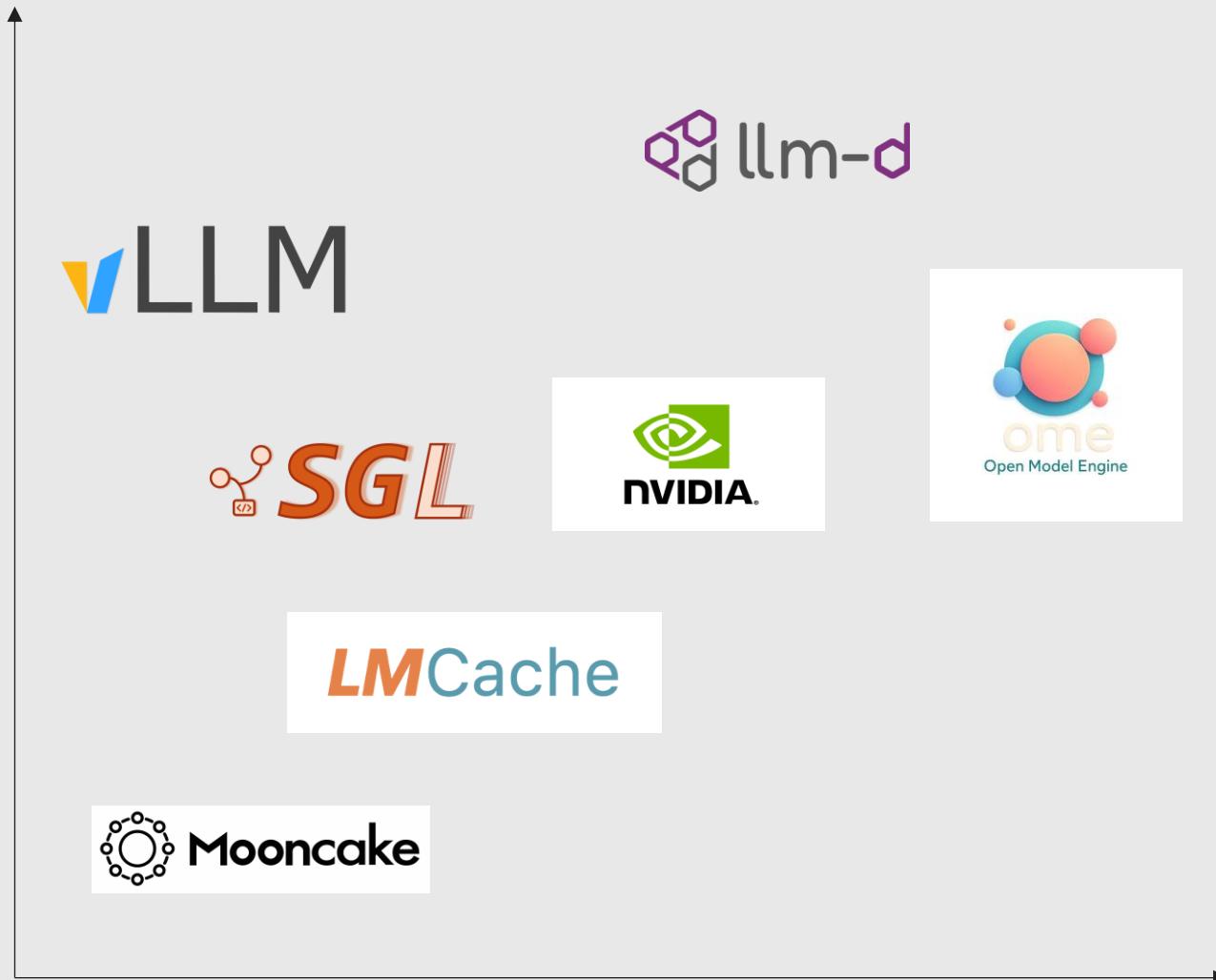
## Other Apps Workload

- Short-lived, uniform requests
- Consistent SLOs
- Even replica performance
- No benefit from coordinating across replicas

# LLM Native

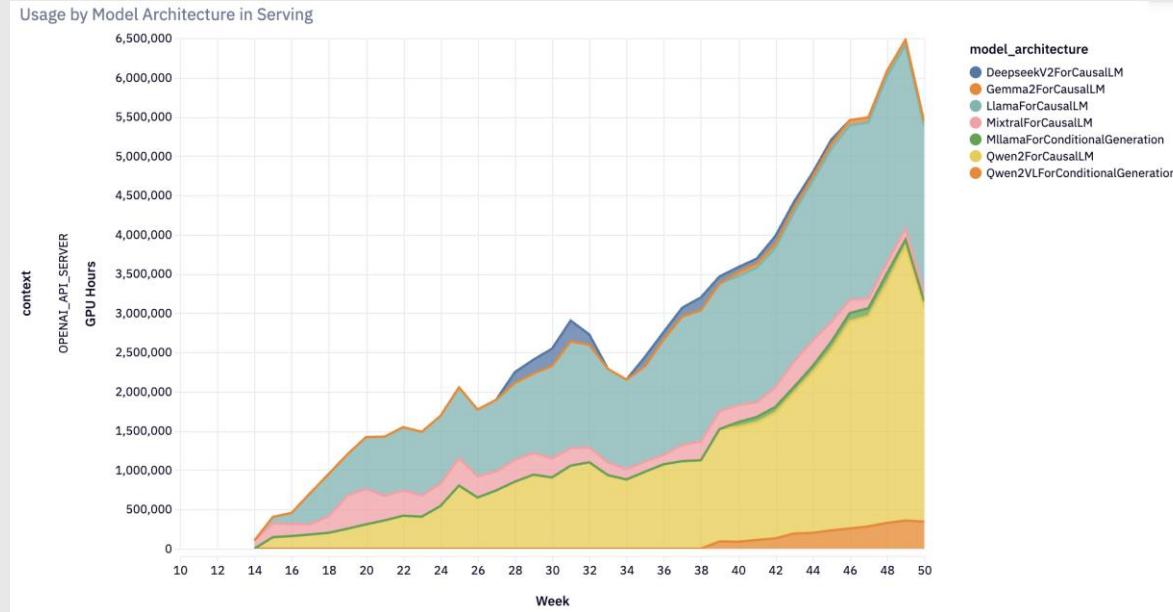


(Source: <https://aws.amazon.com/blogs/machine-learning/accelerate-generative-ai-inference-with-nvidia-dynamo-and-amazon-eks/>)



## LLM Native Open Source

- Prefill Decode Disaggregation: LLM inference is two workloads, Prefill is compute-bound and Decode is memory-bound. Prefill needs a lot of computing resources to convert input into data structure that gets handed off to the next process. Decode (token generation) is that the model generates each token sequentially one at a time.
- Prefix / KV (key-value) Cache
- Context-Aware Routing and AI-Aware Network Routing
- Variant Autoscaling



## LLM Native Open Source

- Prefill Decode Disaggregation: LLM inference is two workloads, Prefill is compute-bound and Decode is memory-bound. Prefill needs a lot of computing resources to convert input into data structure that gets handed off to the next process. Decode (token generation) is that the model generates each token sequentially one at a time.
- Prefix / KV (key-value) Cache
- Context-Aware Routing and AI-Aware Network Routing
- Variant Autoscaling

04

# Compute



## AWS Neuron

185 followers <https://aws.amazon.com/machine-l...>

Overview    Repositories 23    Projects 1    Packages    People

Popular repositories

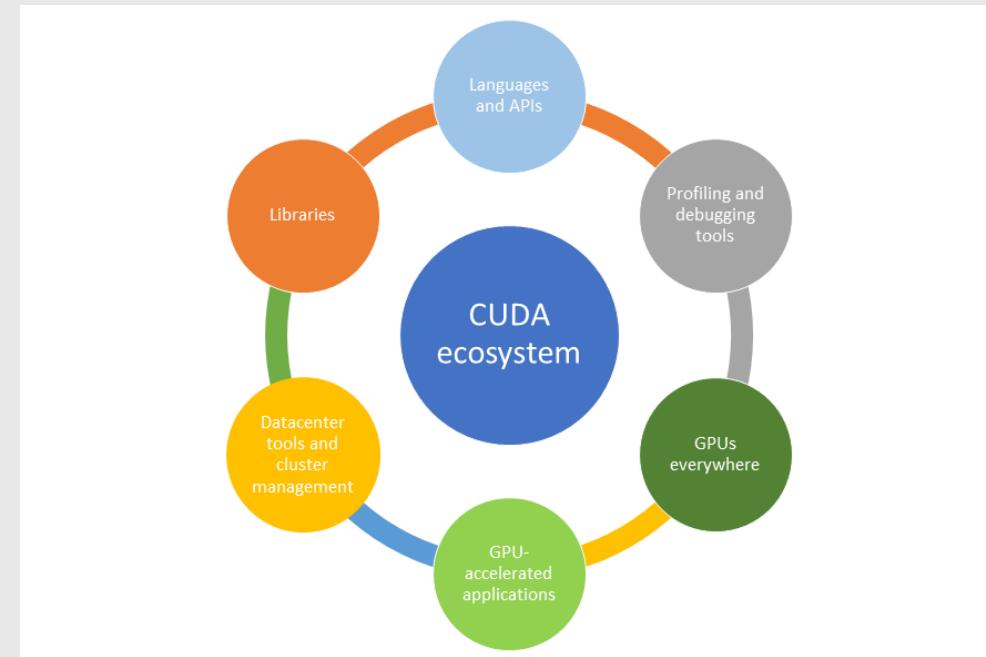
<b>aws-neuron-sdk</b> <span style="float: right;">Public</span> Powering AWS purpose-built machine learning chips. Blazing fast and cost effective, natively integrated into PyTorch and TensorFlow and integrated with your favorite AWS services Python ⭐ 533 ⚡ 170	<b>aws-neuron-samples</b> <span style="float: right;">Public</span> Example code for AWS Neuron SDK developers building inference and training applications Jupyter Notebook ⭐ 148 ⚡ 45
<b>transformers-neuronx</b> <span style="float: right;">Public</span> Python ⭐ 112 ⚡ 32	<b>neuronx-distributed</b> <span style="float: right;">Public</span> Python ⭐ 60 ⚡ 16
<b>neuronx-nemo-megatron</b> <span style="float: right;">Public</span> Python ⭐ 39 ⚡ 13	<b>nki-samples</b> <span style="float: right;">Public</span> Python ⭐ 39 ⚡ 24

People

This organization has no public members. You must be a member to see who's a part of this organization.

Top languages

Python Jupyter Notebook Shell C Smarty



---

05

# Agentic AI



X

# Agentic AI: Protocol Timeline



Claude Computer Use

Oct. 2024



MCP

Nov. 2024



A2A

Apr. 2025

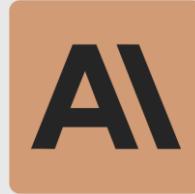
Strands Agent

(an Open Source AI Agent SDK)

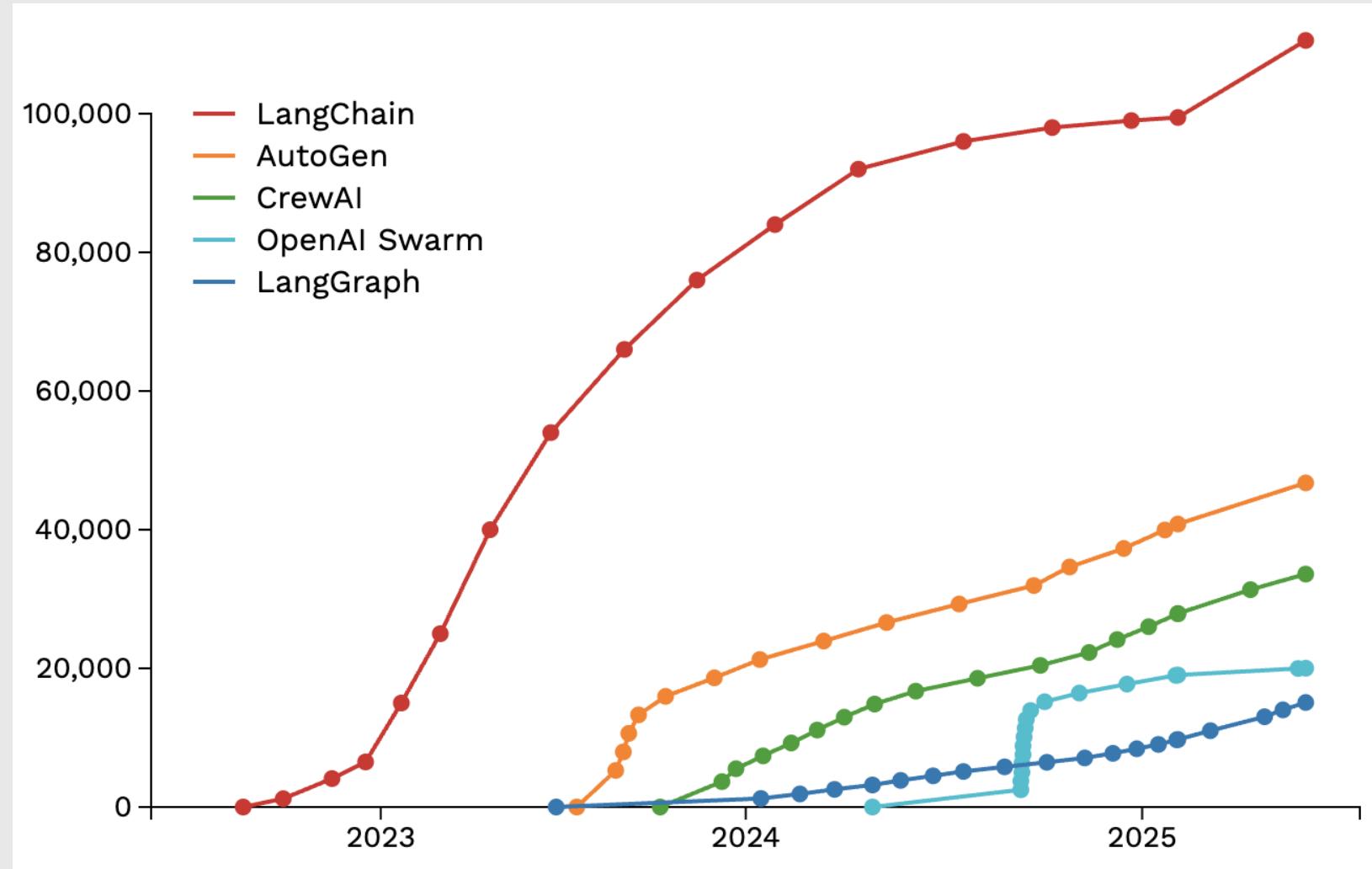
May. 2025

GCP donates to Linux Foundation

June. 2025



# Agentic AI: Frameworks



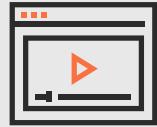
06

# Takeaways



X

# LLM Foundation: Open Source



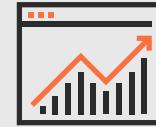
## SOTA Models

DeepSeek, Open AI, Llama, Kimi, Qwen..



## Agentic AI

MCP, A2A, Strands Agent..



## Inference, Compute & Chips

vLLM, Nvidia Dynamo, SGL, llm-d, OME, LMCache, CUDA, Neuron..

00

# Q & A

X

## Q & A



pengandy-us@



pymhq@



00

Thanks

