

How AI Is Changing AI

A look at how the builders of AI have changed, not just the models they ship.



Youssef Elshihy

Regional Presales Consulting Manager

· MENA · iFLYTEK

<https://www.linkedin.com/in/youssef-elshihy/>



Three AI eras in under a decade

Generative AI, as a field anyone outside a research lab could touch, **is younger than most of the phones in this room**. Each era changed the cast of players who could build.



ERA 01

Machine Learning & Fine-Tuning

Anyone with the right data and modest compute could shape a base model.



ERA 02

RAG & Workflow Orchestration

Training moved out of reach; value shifted to retrieving, routing and chaining.



ERA 03

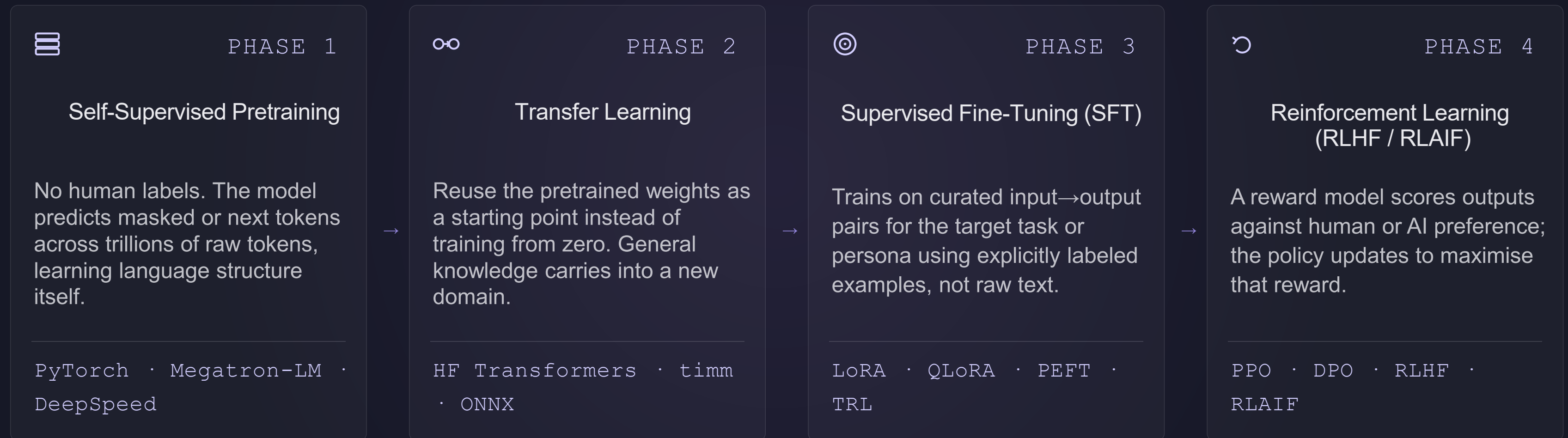
Fully Autonomous Agents

Systems plan and act across steps, sitting on top of everything before it.



Era 01: Machine Learning & Fine Tuning

Take a base model. Shape it with your data. Low barrier to entry. A team with the right data and modest compute could enter. Eras 2 and 3 are what's coming.

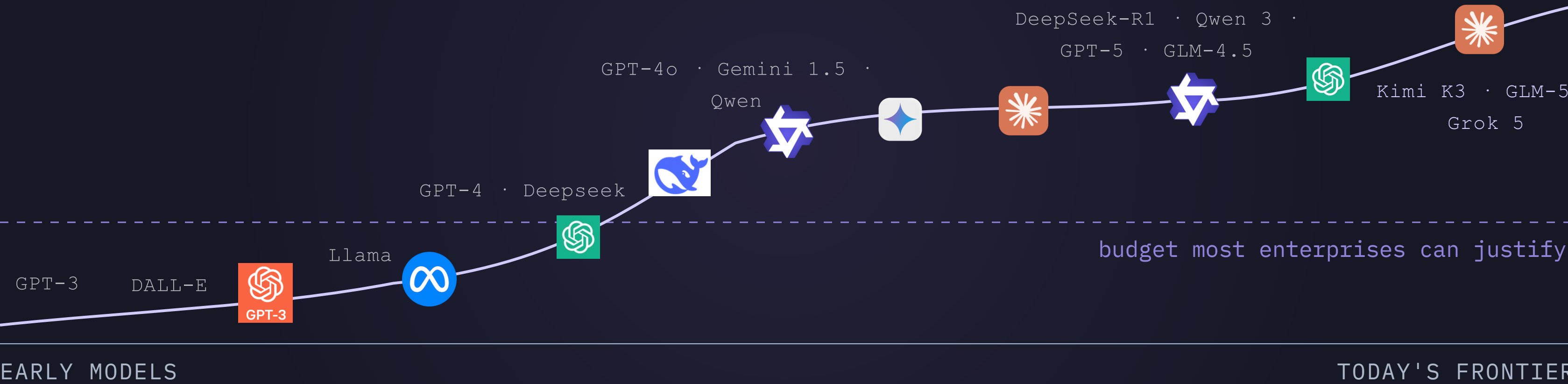


These aren't competing approaches — almost every deployed model runs all four in sequence, each stage narrowing general capability into aligned, task-specific behaviour.



Era 02: RAG & Workflow Orchestration

Frontier training now needs an R&D budget



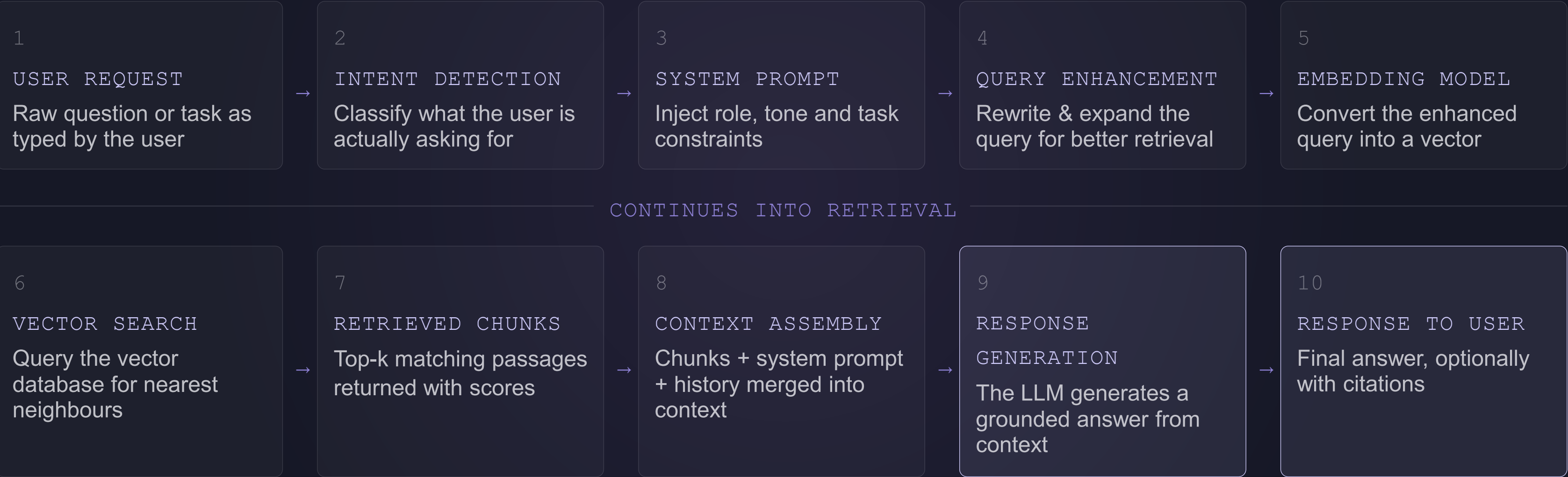
EARLY MODELS		TODAY'S FRONTIER		
GPT-3	GPT-4	Gemini 1.5	Claude Opus 4.5	Frontier
2020 · ~\$4.6M	2023 · ~\$100M	2024 · ~\$190M	2025 · ~\$500M*	2026 · >\$1B*
				

Building and training a foundation model now takes resources only a handful of enterprises can justify. *Estimated, undisclosed frontier runs.



Era 02: RAG & Workflow Orchestration

Everyone aside from R&D companies are moving to orchestration



STACK

LANGCHAIN

LLAMAINDEX

PINECONE / PGVECTOR

MCP

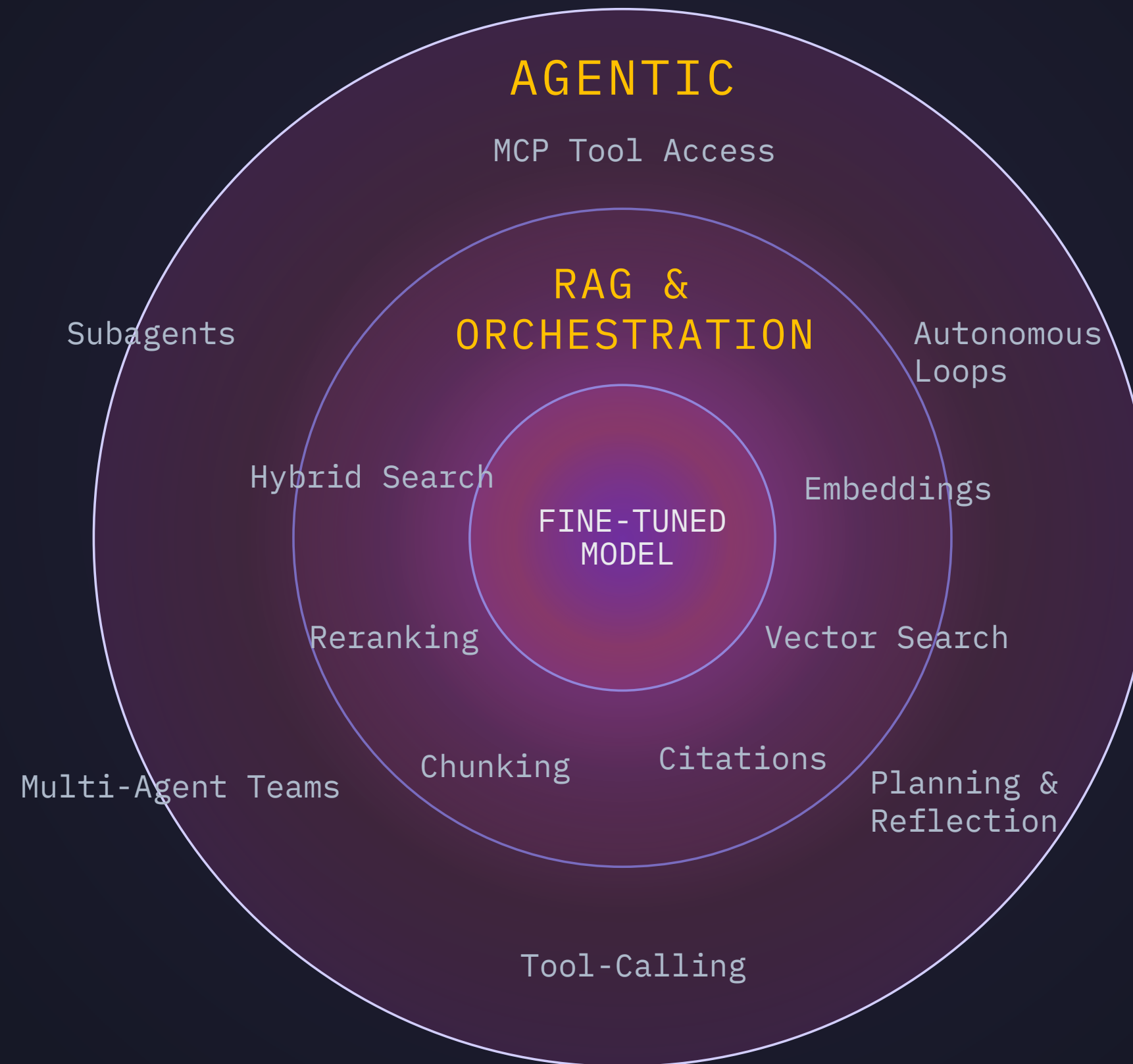


Era 03: Fully Autonomous Agents

The barrier to entry has risen sharply. Era 3 is still an outline. Most organizations are only just arriving at era 2.



Era 03: Fully Autonomous Agents



Era 03: Fully Autonomous Agents

New roles emerge as others shrink. Not purely a loss.

EMERGING

- ↑ AI Engineer
- ↑ MLOps Engineer
- ↑ Prompt / Context Engineer
- ↑ AI Solutions Architect
- ↑ Agent Orchestration Engineer
- ↑ AI Governance & Safety Lead

SHRINKING

- ↓ Traditional software engineer (routine CRUD)
- ↓ Manual QA tester
- ↓ Junior support / ops roles
- ↓ Basic data entry & reporting analyst

Hardware is regaining the value software lost. Compute is the one layer every era in this talk still depends on.



The New Emerging Age of AI

A new model ships every **1.5 days**

24

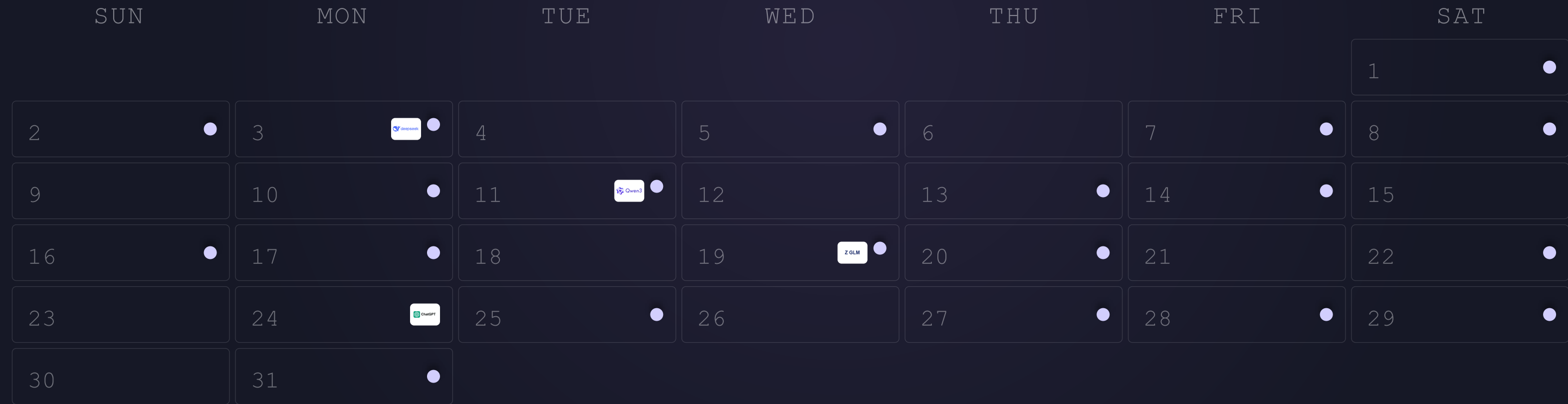
18

1.5

RELEASES, AUG 2026

DIFFERENT PROVIDERS

DAYS BETWEEN RELEASES



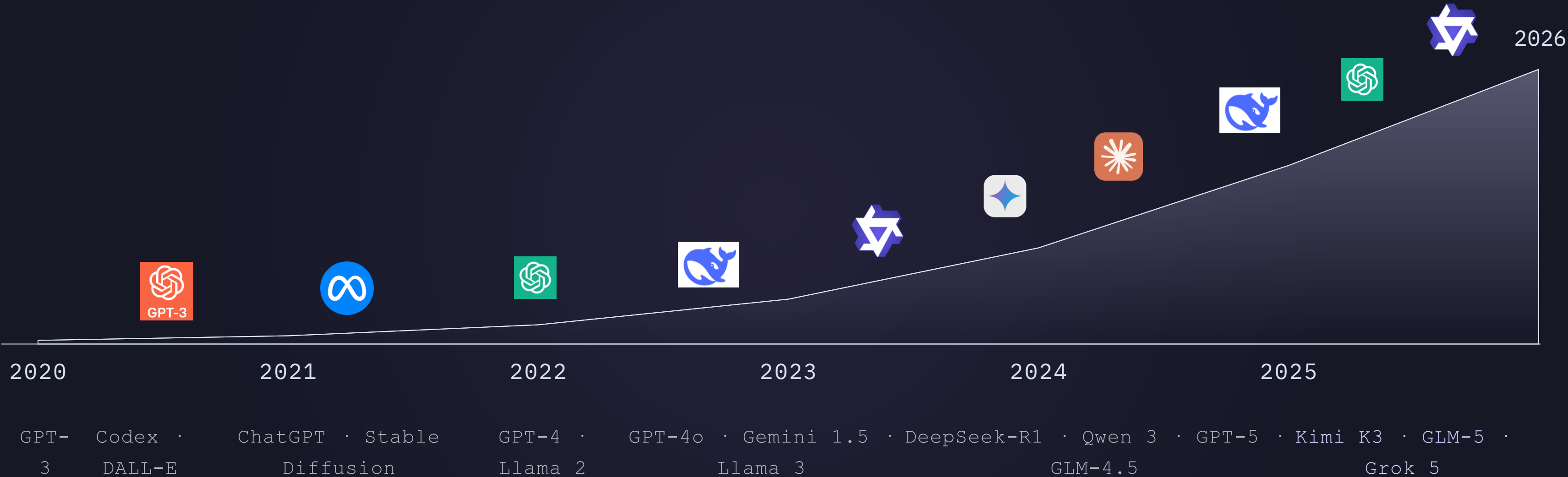
Aug 3 — DeepSeek-V4 · Aug 11 — Qwen3.5 · Aug 19 — GLM-5 · Aug 24 — GPT-5.2.

Most organisations are still catching up to what era 2 required. Era 3, and this pace of releases, are already reshaping the field.



The New Emerging Age of AI

How many models are actually out there?

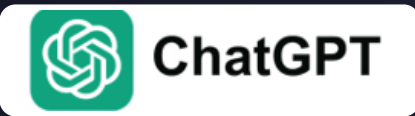
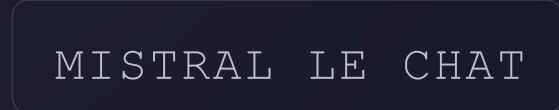


Illustrative growth in notable model releases tracked publicly. The count keeps climbing even as the barrier to enter at the frontier keeps rising.



What Open Source Means Now

Open-weight releases push **baseline capability** up for everyone building on them. Closed models turn a **capability edge** into a shippable product with a price tag. Both are happening across the same industries, at **the same time**.

	CLOSED-SOURCE	PLAY	OPEN-SOURCE	PLAY		
FOUNDATION MODELS	 ChatGPT	 Gemini	 Claude	 deepseek	 Qwen3	 Z GLM
DATABASES	 MICROSOFT SQL SERVER	 ORACLE DB	 POSTGRESQL	 MYSQL		
OPERATING SYSTEMS	 WINDOWS	 MACOS	 LINUX			
PRODUCTIVITY	 MICROSOFT COPILOT		 MISTRAL LE CHAT			



What Open Source Means Now

Open by license only, closed by infrastructure. Actually, running it at this scale needs infrastructure most organizations simply don't have, even though the weights are free to download.



Kimi K3 · Moonshot AI · July 2026 ·
the largest open-weight model released to date

2.8T parameters

\$50K+

EST. MINIMUM SERVER COST, SINGLE CONCURRENT USER

SINGLE GPU

Can download the weights. Can't load a 2.8T-parameter model into memory.

MULTI-GPU SERVER

~\$50K in hardware. Runs it for **one concurrent user**.

ENTERPRISE CLUSTER

The tier that actually serves it at production scale and concurrency.



What Open Source Means Now

The same question, two different markets. Different incentives, different regulation, different buyer psychology. Both sides playing out at the same time. Predicting either one from only one side of the picture is unreliable.

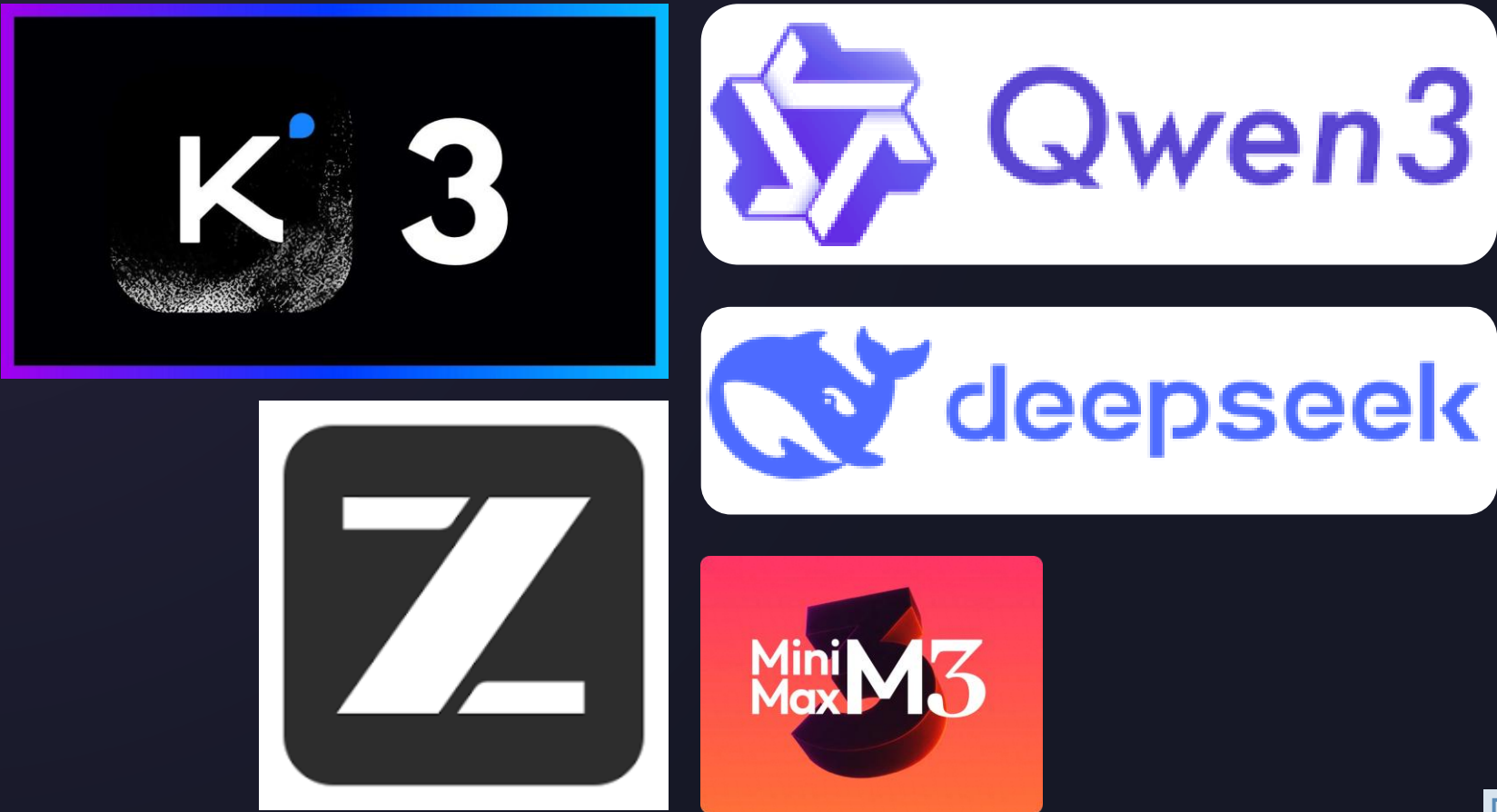
 UNITED STATES

Regulation-first, disclosure-heavy, buyers who want the paper trail before the pilot.



 CHINA

Speed-first, deployment-heavy, buyers who want the pilot before the paper trail.



What Open Source Means Now

Track all three layers, not just the model headlines

LAYER 1

Who can build

The barrier to entry keeps rising, one era at a time.

LAYER 2

How fast it moves

The ground shifts again before most organisations catch up to the last shift.

LAYER 3

What “open” buys you

A licence and the ability to run it at scale are no longer the same thing.

Anyone making AI decisions right now needs to track all three, not just which model tops this week's leaderboard.

What even is “open source” anymore?





Thank you

Questions, pushback, and war stories from your own region welcome.



Youssef Elshihy

Regional Presales Consulting Manager

· MENA · iFLYTEK

<https://www.linkedin.com/in/youssef-elshihy/>