Highlights Edition

Artificial Analysis State of AI

Q2 2025

Highlights Report

Full report available to AI Trends subscribers

Artificial Analysis is a leading, and independent AI benchmarking and insights provider. We support engineers and companies to understand AI capabilities and make critical decisions about their AI strategy.

Our data, insights and publications are grounded in our comprehensive benchmarking of AI technologies and use cases. This includes everything from hourly performance testing of language model APIs to millions of votes in our crowd-sourced arenas.

Our public website, artificialanalysis.ai, is widely referenced by companies leading innovation in AI. To discuss this report, our publications, or our services, please get in touch at contact@artificialanalysis.ai.

Artificial Analysis AI Trends Subscription: Comprehensive AI market intelligence for enterprise decision-making from the leading AI benchmarking company

Overview of Artificial Analysis AI Trends Subscription

This report

A | Quarterly State of AI Report

The definitive quarterly update on AI market developments

- Emerging trends at each layer of the AI stack: hardware, infrastructure, models
- Market maps and performance rankings for hundreds of key players
- State of AI: China - detailed benchmarking of top AI labs in China

B | Enterprise Agents Report

2025 is the year of agents – our overview of what matters most

- Comprehensive analysis of key agent categories: coding, deep research, computer use, customer support, sales
- What's working now: where agents are driving real productivity
- Implications for real-world deployment

C | AI Adoption Survey

Real-world adoption insights from those building and deploying AI

- Enterprise use case patterns
- Enterprise adoption benchmarks
- Developer priorities and pain points
- Model, inference and hardware provider demand by industry

D | Databooks & API

Direct access to the industry's most comprehensive data

- Comprehensive AI performance data - source data for all our analysis
- Intelligence, performance, cost, survey data and more
- Excel databooks and API access

E | Quarterly AI Trends Workshop

Connect market intelligence to your strategic priorities

- Live briefing with Artificial Analysis research team (90 minutes, optional)
- What's working now: insights from leading SF startups and enterprises
- Deep dives tailored to your business priorities (e.g., coding agent best practices, inference economics, upcoming chips)

F | Ongoing Team Access

Direct access to our research team for support and clarifications

- Support for use of reports and data, including queries on sources and methodology
- Clarification and explanation of analyses
- Limited to max 90 minutes per quarter, rates available for further support

Artificial Analysis is trusted by the leading AI industry players and publications

This is the Highlights Version of the Quarterly State of AI Report for Q2 2025, the Premium Version is available to subscribers of our AI Trends Subscription

Highlights Version (This)

- **Industry overview and** market map of key players and strategies across the AI value chain
- **Overview of frontier models** ranked by the Artificial Analysis Intelligence Index and overview of emerging trends
- **Synthesis of emerging trends** for image, video and speech models and market maps
- **Synthesis of emerging trends** for accelerators including case study comparing NVIDIA B200 and NVIDIA H200 using Artificial Analysis System Load Test

Premium Version (AI Trends Subscription)

Includes everything in the Highlights Version plus:

- **Detailed insights across new language model releases** (incl. analysis of leading open weights options)
- **Detailed analysis and case studies** outlining emerging trends for language models across pricing, performance and features
- **Analysis of frontier image generation models and trends** (incl. text to image and image editing)
- **Analysis of frontier video generation models and trends** (incl. text to video and image to video)
- **Analysis of frontier speech models and trends** (incl. text to speech and speech to text)
- **Emerging market trends** for accelerators, including detailed analysis comparing NVIDIA H200 and NVIDIA B200
- **Attached Separate Report: Enterprise Agents Report** covering comprehensive analysis of key agent categories and implications for real-world deployment

Feel free to get in touch with us at subscriptions@artificialanalysis.ai to learn more about the Artificial Analysis AI Trends Subscription

Artificial Analysis State of AI Q2 2025

The story of AI in Q2 2025 reveals an industry hitting its stride after years of foundational development. We are witnessing a new phase where innovations across the AI stack are maturing and converging towards impacting how every organization operates.

Today's models demonstrate significant intelligence gains while becoming more cost-effective and faster than ever. Agentic workflows are moving from promising experiments to production reality, with coding agents proliferating across development teams. Meanwhile, the competitive landscape continues to evolve, with Chinese AI labs demonstrating remarkable leadership in both language and video capabilities.

Produced by Artificial Analysis, an independent benchmarking and insights firm trusted across the AI value chain, this Q2 2025 report is designed to inform investment, product, and policy decisions in an increasingly AI-native world.

For more details, contact us at founders@artificialanalysis.ai

- Micah Hill-Smith and George Cameron,
Founders of Artificial Analysis

Contents

01 Industry Overview

State of AI – Q2 2025

5 major trends have shaped the State of AI across Q2 2025

Tools and connections enable smart workflow integration

Native connections and tools in chat interfaces are now shifting workloads to agentic approaches

Coding agents rapidly proliferate across development workflows

Q2 saw 12 major coding agent launches, including from major labs

Video models see breakthroughs and rapid quality increase

Google Veo 3's release showcases audio-video breakthroughs, driving adoption and new use cases

Language models continue to become more intelligent

Major AI labs have all continued to make substantial gains in intelligence, cost efficiency and speed

China continues to demonstrate leadership in language and video

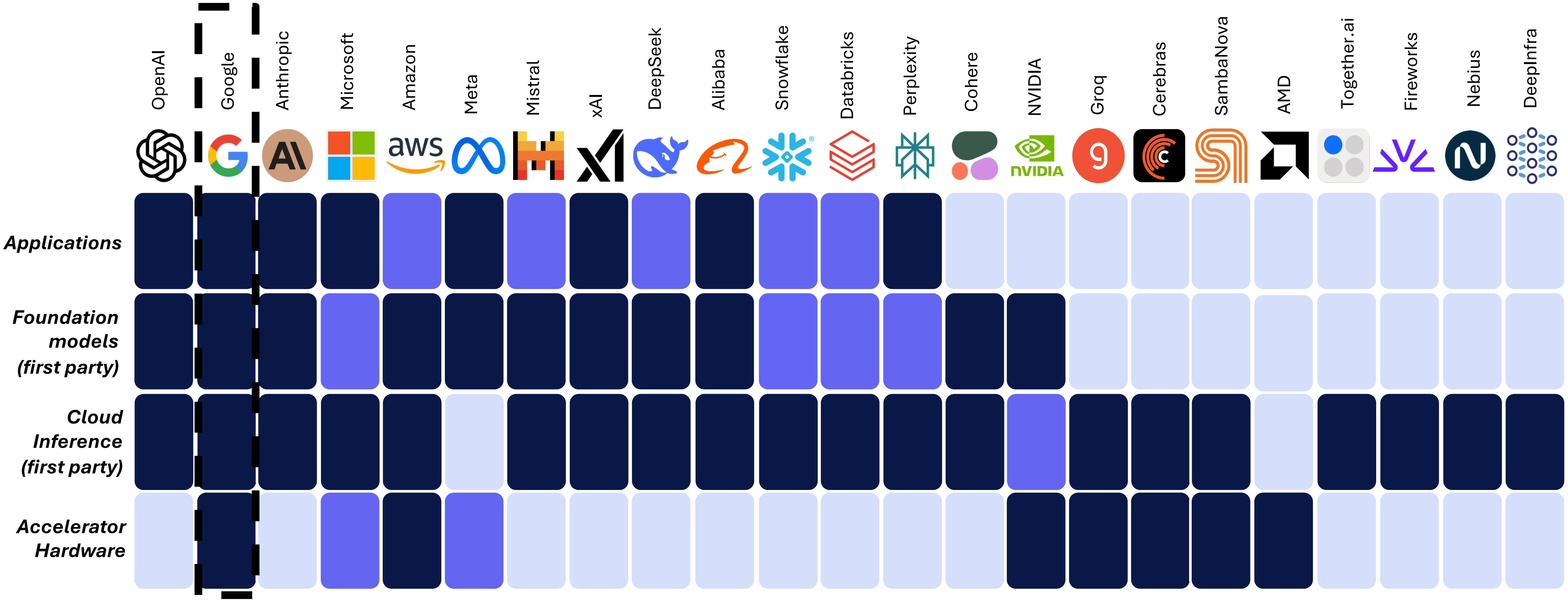
Models from Chinese AI labs occupy top spots for open weights language models and on the video leaderboard

Players in the AI value chain differ in levels of vertical integration; Google continues to stand out as the most vertically integrated from TPU accelerators to Gemini

Key Players in the AI Value Chain (Non-Exhaustive)

Classifications are indicative and determined based on a range of factors including market share and strength of offering

No presence … **Strong presence**



Source: Company website

Big technology companies are continuing to play across all AI modalities while smaller challengers tend to focus on specific modalities

Key players with first-party models by type of AI

No model | Existing model

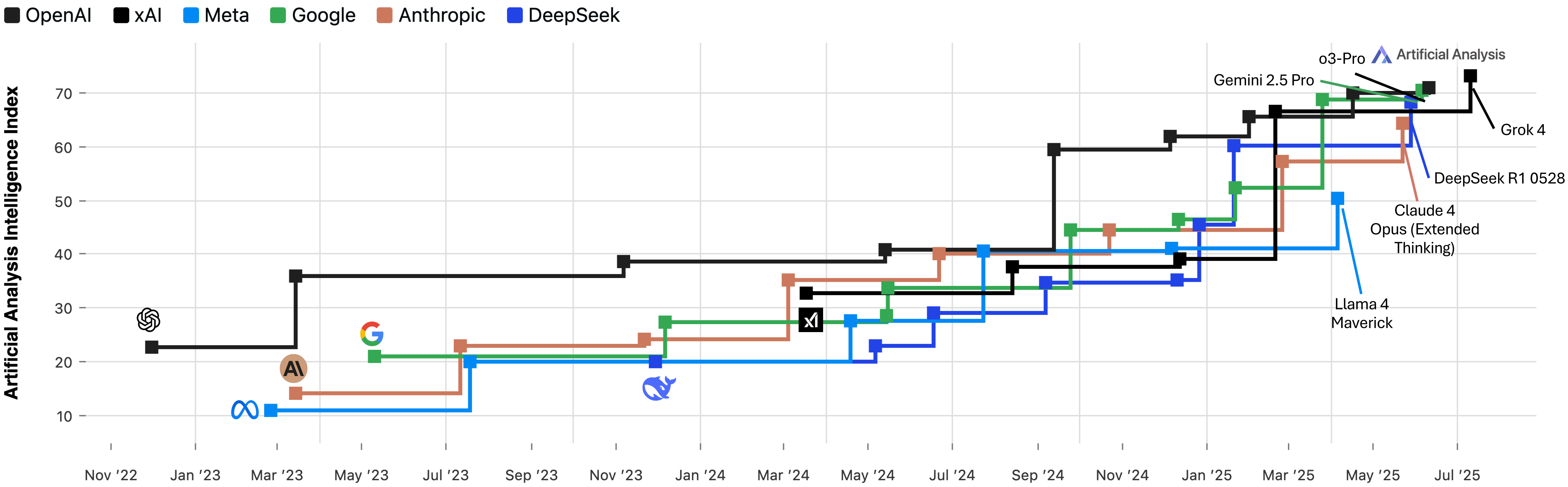
	OpenAI	Google	Meta	xAI	Anthropic	Mistral	Microsoft	Amazon	NVIDIA	Adobe	ElevenLabs	Perplexity	Alibaba	Bytedance	Tencent	Baidu	DeepSeek	Kuaishou	MiniMax	Cohere	Midjourney	AI21 Labs
Language	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	No model	No model	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	No model	Existing model	Existing model	No model	Existing model
Speech	Existing model	Existing model	Existing model	Existing model	No model	Existing model	Existing model	Existing model	Existing model	No model	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	No model	No model	Existing model	No model	No model	No model
Image	Existing model	Existing model	Existing model	Existing model	No model	No model	No model	Existing model	Existing model	Existing model	No model	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	Existing model	No model	Existing model	No model
Video	Existing model	Existing model	Existing model	No model	No model	No model	No model	Existing model	No model	Existing model	No model	No model	Existing model	Existing model	Existing model	Existing model	No model	Existing model	Existing model	No model	Existing model	No model

Source: Company website

A number of AI Labs now have models near the frontier of intelligence; xAI has the leading model with Grok 4, achieving this feat in less than 500 days since their first models' launch

Frontier Large Language Model (LLM) Intelligence, Over Time

Artificial Analysis Intelligence Index v2 (incorporates MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2024, MATH-500)



- **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

Source: Artificial Analysis independent benchmarking

02 Language Models

State of AI – Q2 2025

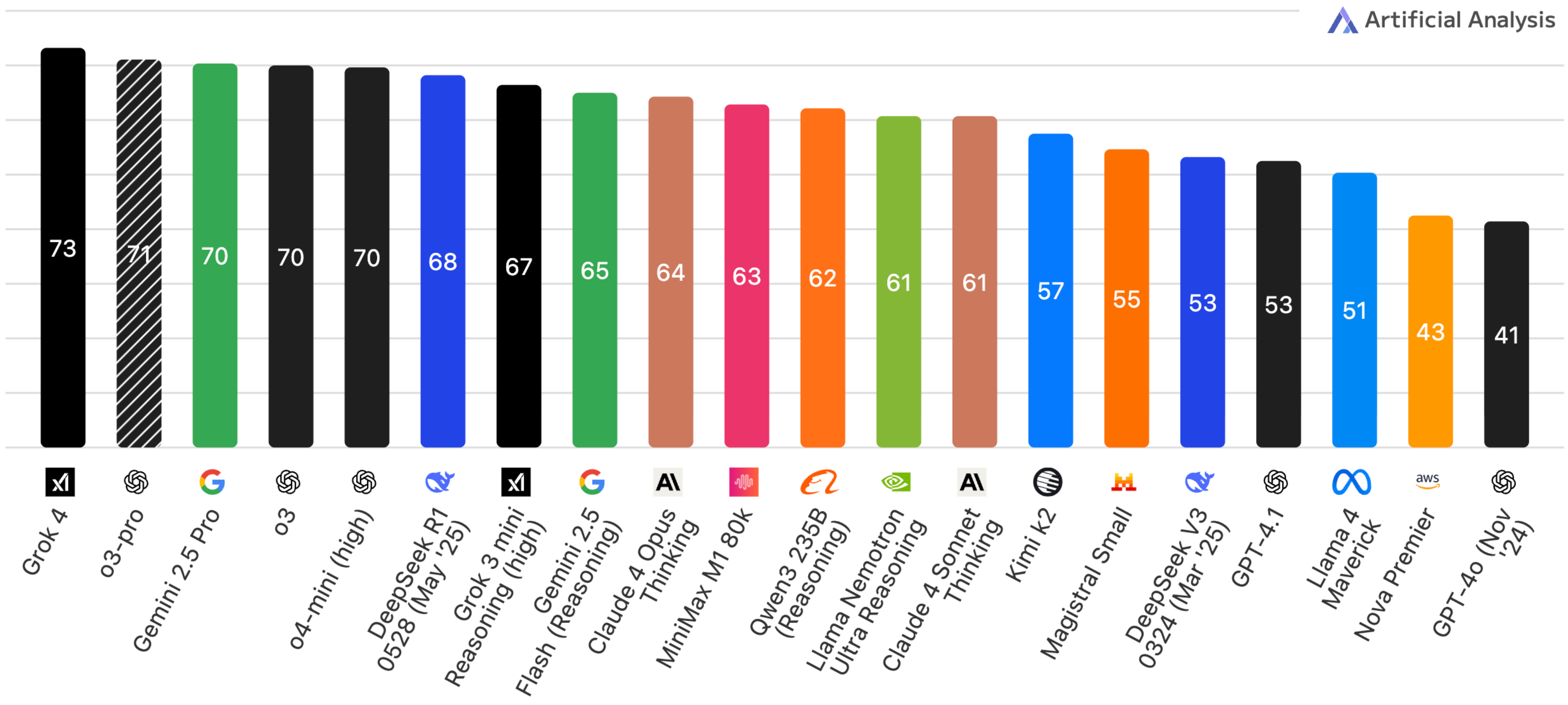
xAI, OpenAI, and Google lead frontier intelligence with their latest reasoning models, followed closely by other labs

Leading Large Language Models (LLMs), by AI lab

Highest Artificial Analysis Intelligence Index v2 achieved by each AI Lab

Estimate (independent evaluation forthcoming)

NON-EXHAUSTIVE



Source: Artificial Analysis independent benchmarking

Commentary

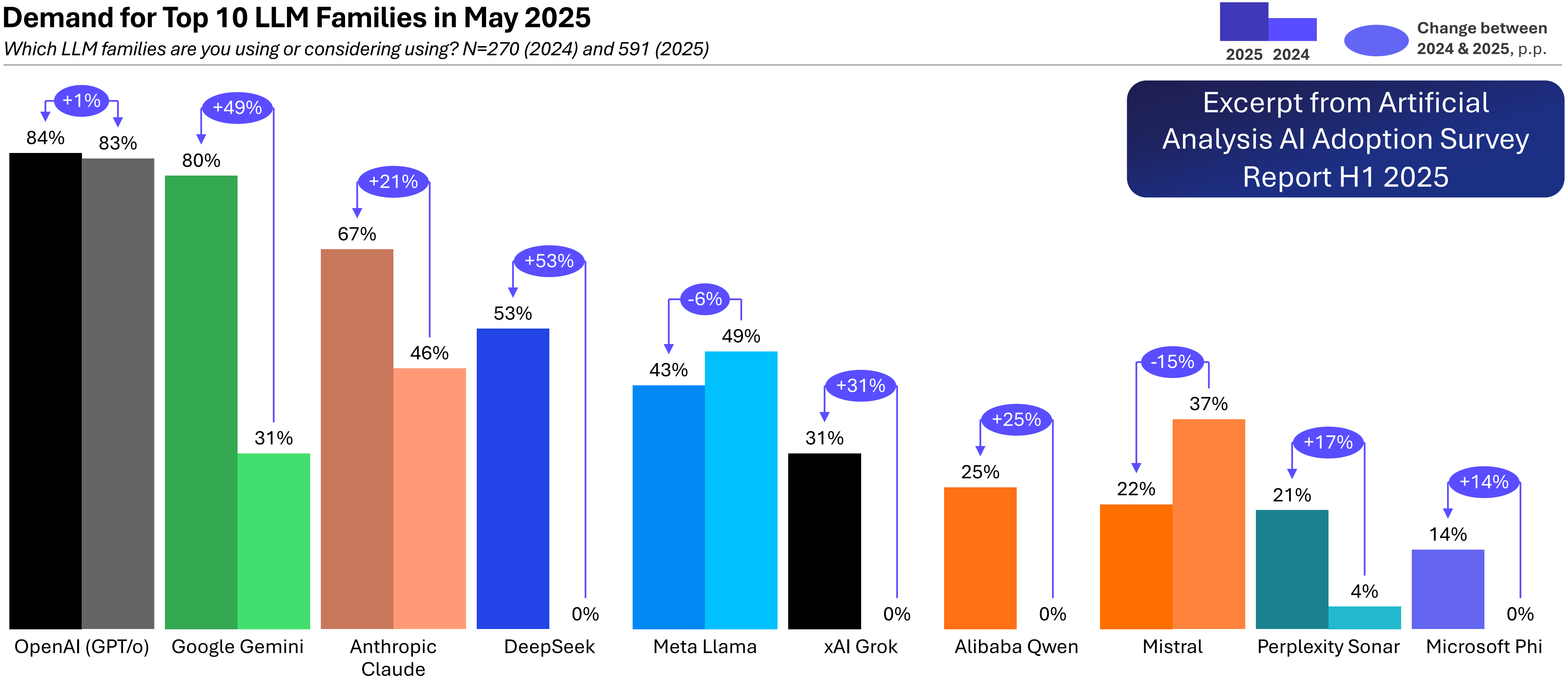
- **OpenAI loses frontier for the first time:** xAI's Grok 4 is now the most intelligent language model, sitting ahead of o3-pro, OpenAI's frontier model
- **xAI, OpenAI and Google lead frontier intelligence:** Latest reasoning models from three labs hold the top 5 positions
- **Reasoning models continue to dominate:** Q2 '25 continues to see reasoning models solidify their position as the clearest path to higher intelligence index scores
- **Global competition intensifies:** Labs like DeepSeek, MiniMax, and Alibaba continue to close gap

Artificial Analysis

Models: Over the past year, OpenAI has maintained its lead, Google Gemini and DeepSeek have surged, and Meta Llama and Mistral have fallen

Demand for Top 10 LLM Families in May 2025

Which LLM families are you using or considering using? N=270 (2024) and 591 (2025)



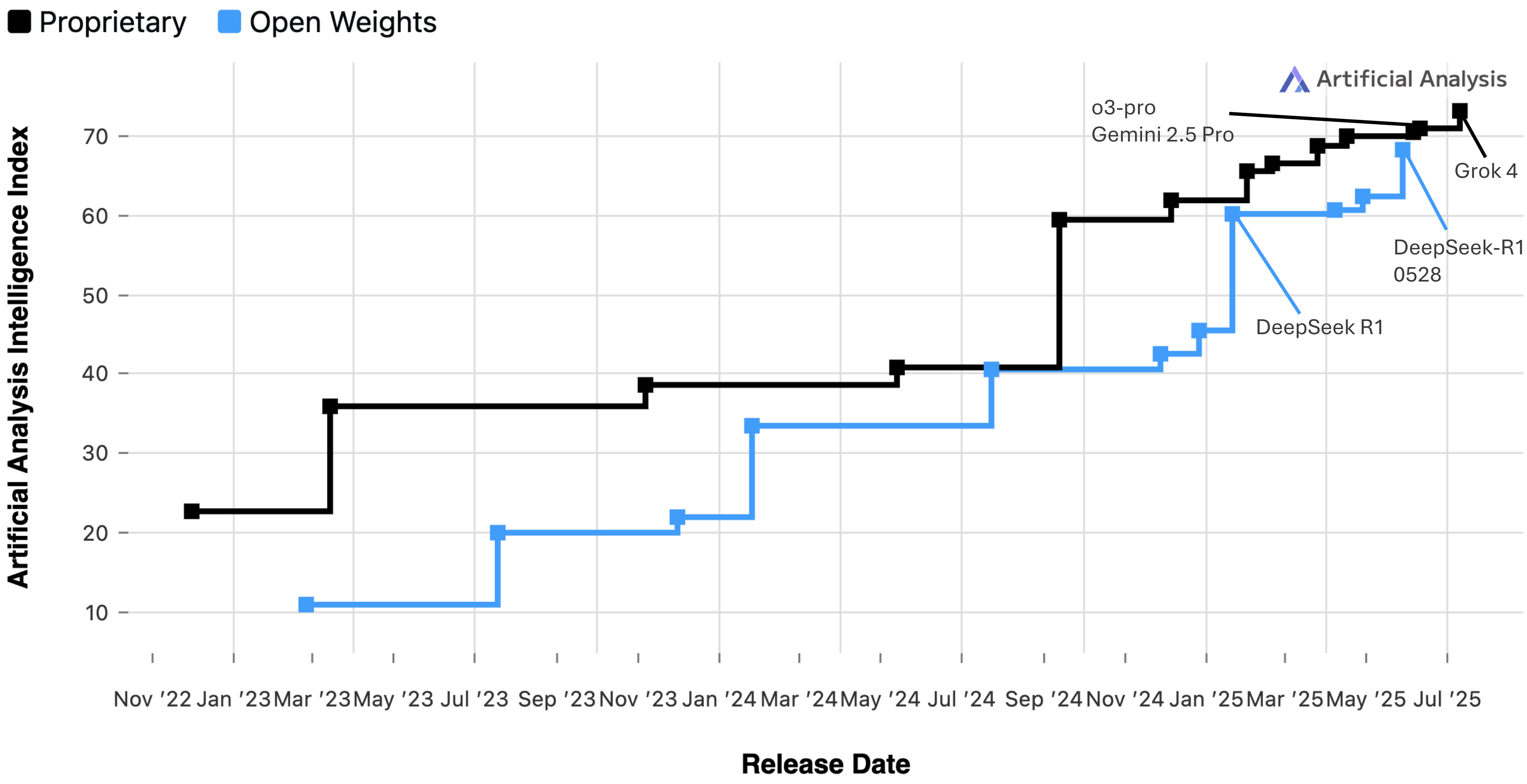
Excerpt from Artificial Analysis AI Adoption Survey Report H1 2025

Source: Artificial Analysis AI Adoption Survey – H1 2025

Open Source: Open weights language models continue to improve, the gap to leading proprietary models stayed similar

Leading Language Models by License Type, Over Time

Artificial Analysis Intelligence Index (incorporates MMLU-Pro, GPQA, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, MATH-500)



Source: Artificial Analysis independent benchmarking

Commentary

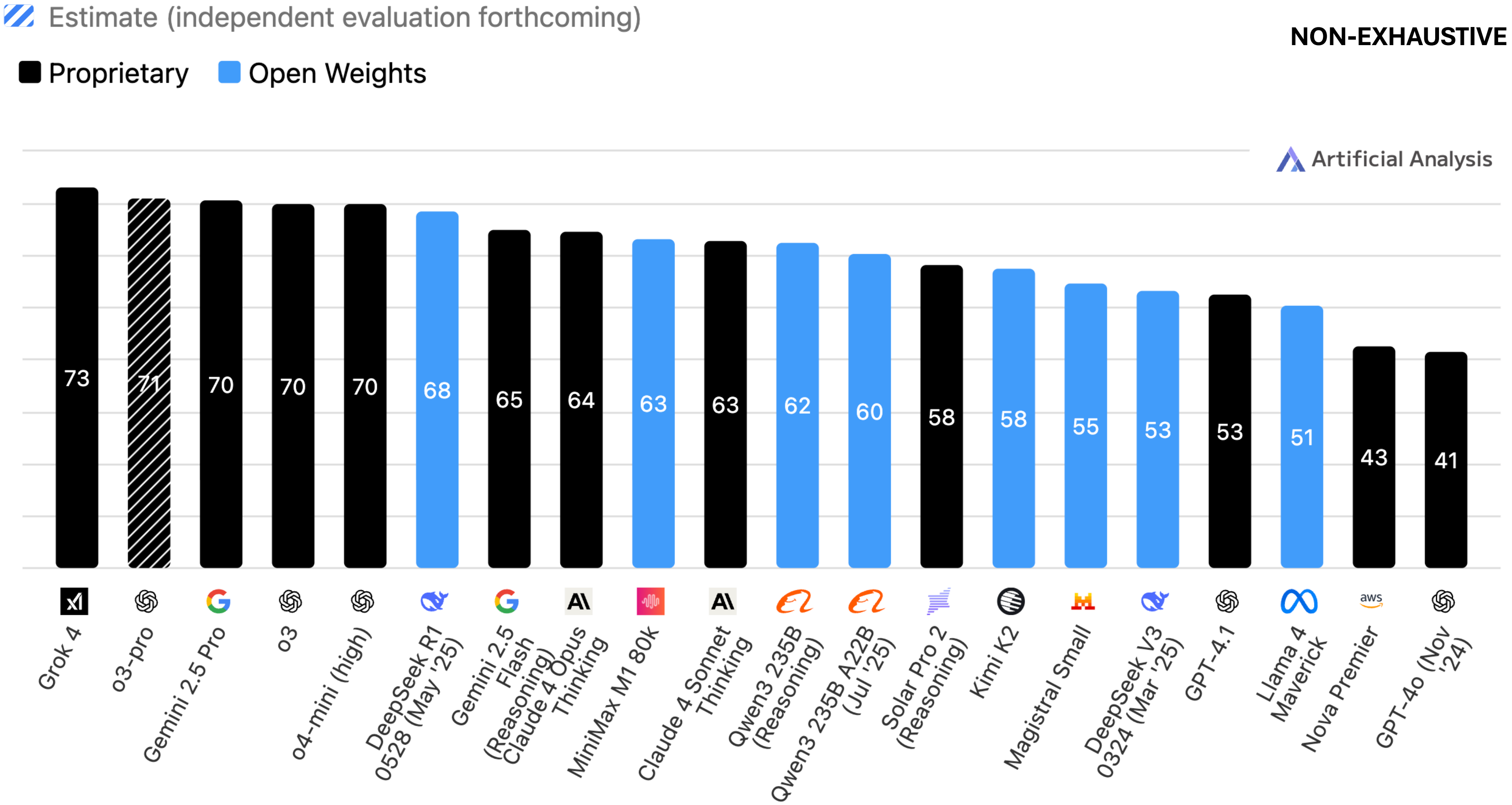
- **Open weights close the gap to proprietary models**: The release of DeepSeek R1 0528 in May further reduced the intelligence gap to leading proprietary models from Google and OpenAI (similar to release of DeepSeek R1); the release of Grok 4 has since widened this gap
- **Proprietary and open weights models continue their rapid release cadence:** Q2 ‘25 continued to see frequent incremental improvements drive the frontier

Artificial Analysis

Open Source: Leading proprietary models are from US labs, while China leads the open weights intelligence frontier

Leading Language Models by License Type

Artificial Analysis Intelligence Index v2 (incorporates MMLU-Pro, GPQA, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, MATH-500)



Commentary

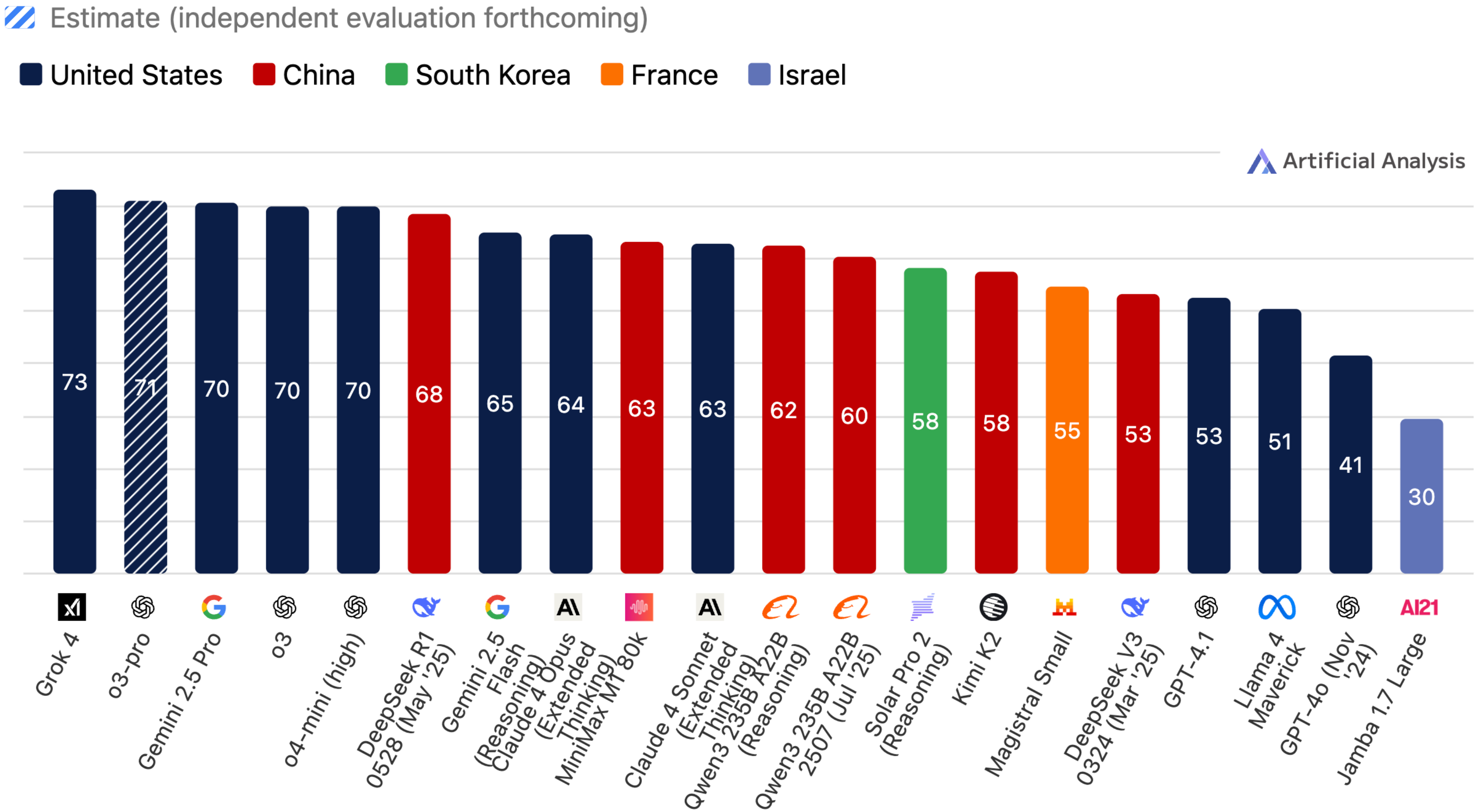
- **Proprietary models continue to lead frontier intelligence:** Proprietary reasoning models from US labs lead in overall intelligence
- **China demonstrates open weights leadership:** Leading open weights models are from Chinese AI labs (DeepSeek, MiniMax, Alibaba, Moonshot)
- **Proprietary models marginally lead for non-reasoning models**: Claude 4 Opus is currently the most intelligent non-reasoning model, followed closely by Kimi K2

Source: Artificial Analysis independent benchmarking

Country view: Models from labs in the US and China continue to dominate the intelligence frontier

Leading Language Models by Country of Origin

Artificial Analysis Intelligence Index v2 (incorporates MMLU-Pro, GPQA, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, MATH-500)



Source: Artificial Analysis independent benchmarking

Commentary

- **US maintains leadership in frontier reasoning:** US-based labs continue to hold the top spots on the Intelligence Index with their premier reasoning models like Grok 4, o3-pro and Gemini 2.5 Pro
- **Q2 saw limited disruption from other countries.** France maintains a presence with Magistral Medium, while Upstage AI's Solar Pro 2 model brought South Korea to the frontier for the first time
- **Overall, the global frontier remains highly concentrated,** with the US and China continuing to define the pace and direction of cutting-edge model development

While efficiency gains have been made…

GPT-4 level intelligence is now 100x cheaper than original GPT-4

A. Smaller Models

Algorithmic and training data improvements have allowed smaller models to get smarter

~1/10x
compute

B. Software Efficiency

Inference optimizations (e.g. Flash Attention) improve efficiency

~1/3x
compute

C. Hardware Efficiency

Next generation accelerators offer more compute efficiency

~1/3x
costs

Figures are highly indicative and serve to illustrate the directional impact of each factor impacting cost

… compute demand continues to increase

Deep Dive next

New applications continue to demand more compute: a single deep research query can cost >10x an original GPT-4 query

~5x
compute/query

D. Larger Models

Scaling laws continue to demand higher parameter counts for greater intelligence

~10x
tokens/query

E. Reasoning Models

Significant increase in output tokens when models 'think' before answering

~20x
requests/use

F. AI Agents

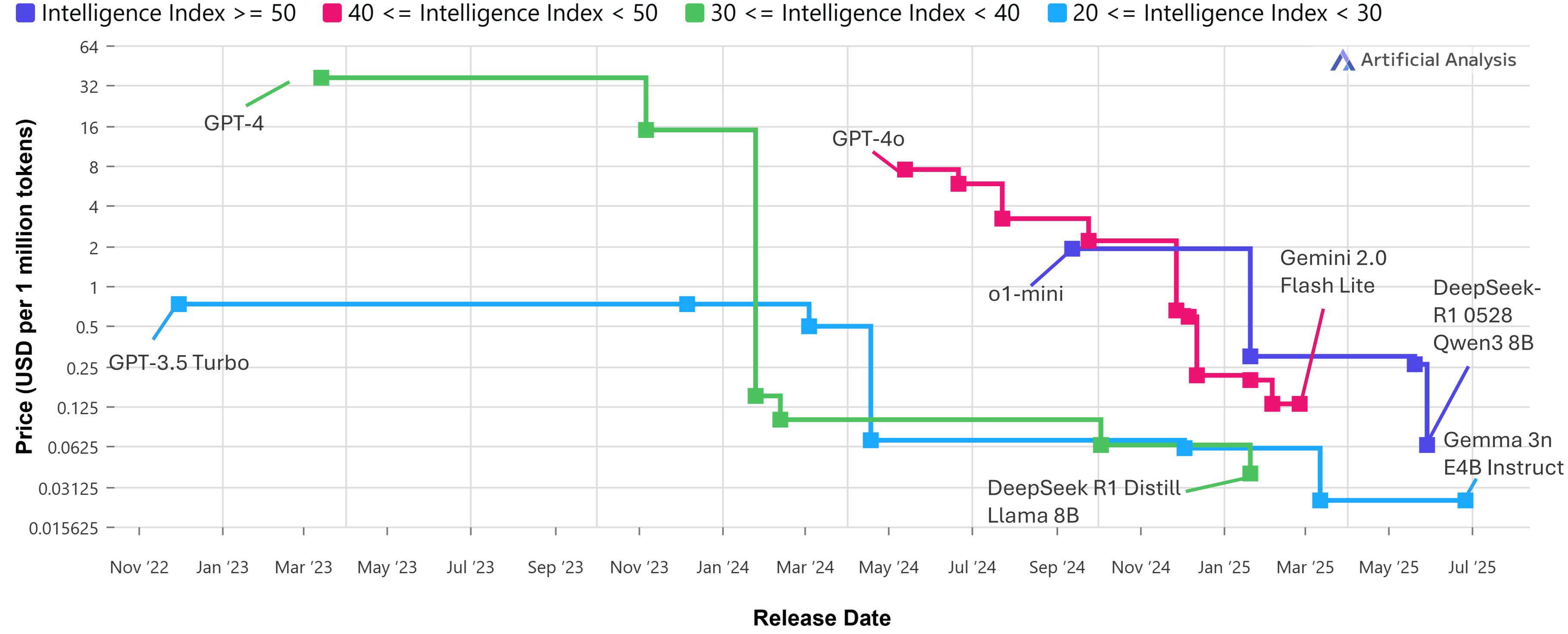
Agents chain multiple requests to LLMs to complete tasks autonomously

B. Software Efficiency: Efficient models combined with new accelerators kept slashing AI inference costs throughout Q2

Language Model Inference Pricing by Intelligence Class, Over Time

Price in USD per 1 million tokens (blended input to output token price 3:1) ; Artificial Analysis Intelligence Index v2 (incorporates 7 evaluations)

NON-EXHAUSTIVE



Commentary

- **Q2 2025 accelerates the slide in inference cost:** from April to June, prices fell across every intelligence band as DeepSeek R1 0528, Qwen3 8B, and Gemma 3n E4B Instruct slashed costs while lifting scores
- **Capable AI is becoming more accessible and commoditized:** during Q2 2025, the price of frontier-level inference (Intelligence Index ≥ 50) dropped by nearly 75%, sliding from \$0.26 to just \$0.063 per million tokens

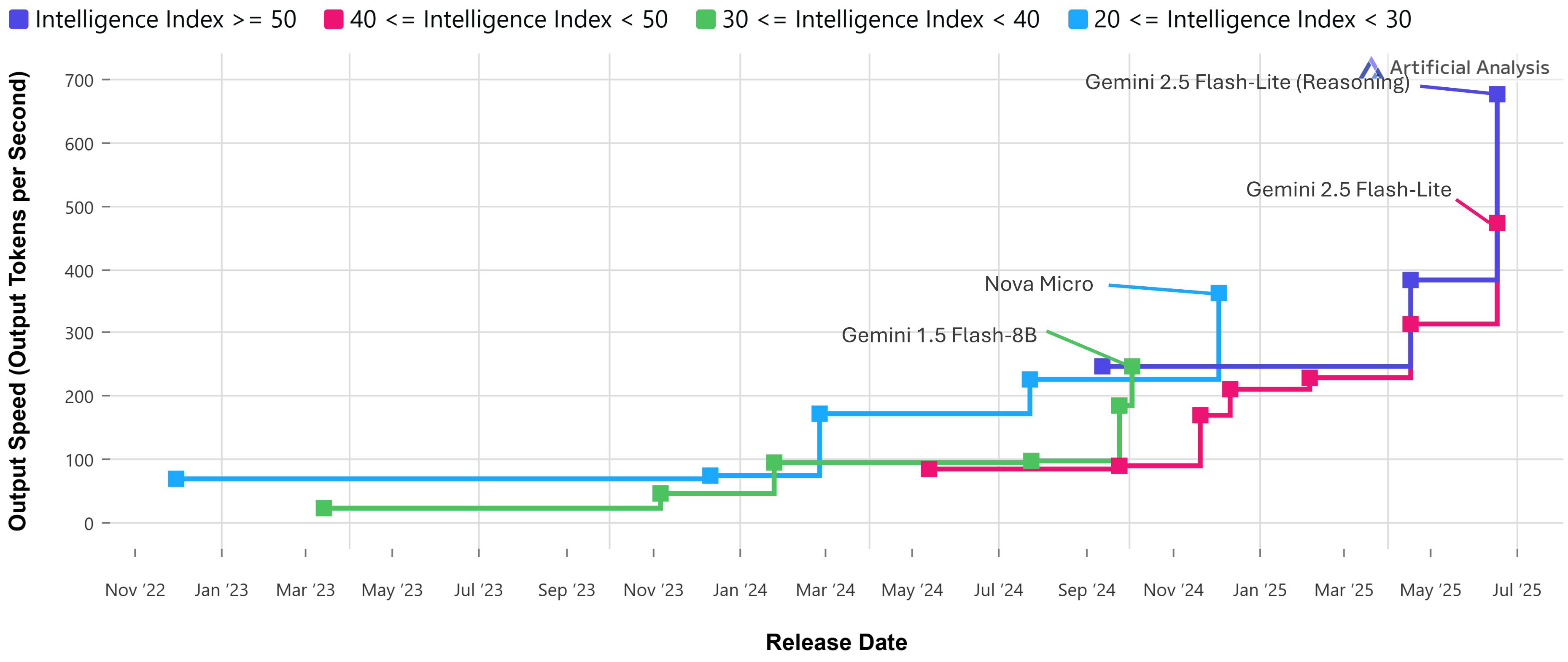
Source: Artificial Analysis independent benchmarking

B. Software Efficiency: Throughput significantly increased in Q2 2025 across model classes, but end-user wait times are sometimes growing due to long reasoning chains

Language Model Output Speed by Intelligence, Over Time

Total output tokens per second, Artificial Analysis Intelligence Index v2 (incorporates 7 leading evaluations)

NON-EXHAUSTIVE



Commentary

- **A Q2 2025 speed surge overcame the trade-off against intelligence:** a significant leap in inference performance occurred in the second quarter of 2025, and new releases made highly-intelligent models (Index >= 50) the fastest category for the first time
- **Latency paradox:** despite higher throughput, end-to-end use can be slower as reasoning and agentic tasks generate tens of thousands of tokens and chain multiple calls, fully offsetting speed gains

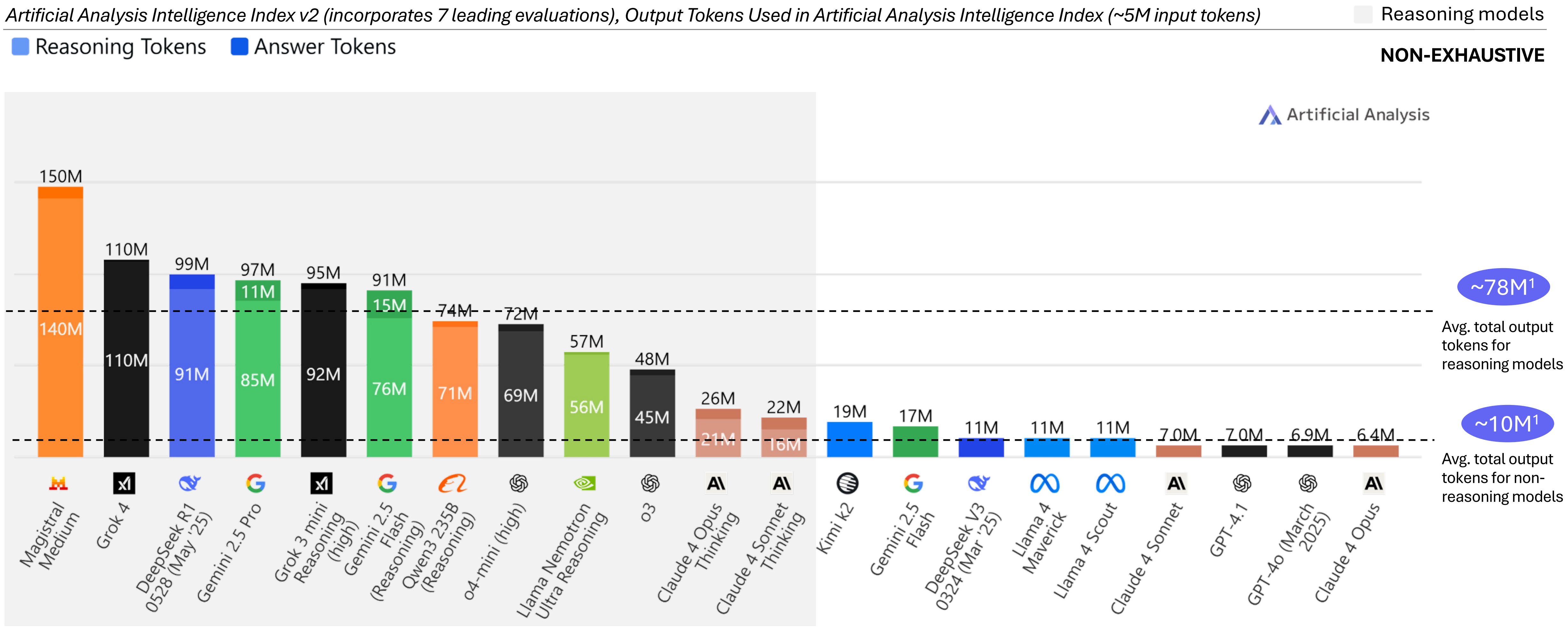
Source: Artificial Analysis independent benchmarking

Artificial Analysis

E. Reasoning Models: Reasoning costs time and compute: reasoning models use up to 10x more tokens to respond to the same prompts as non-reasoning models

Output Tokens Used to Run Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v2 (incorporates 7 leading evaluations), Output Tokens Used in Artificial Analysis Intelligence Index (~5M input tokens)



1. Based on representative models included in the chart

Source: Artificial Analysis independent benchmarking

In Q2 2025 we saw increased use of agentic workflows and explosive growth in coding agents, both enabled by a connection ecosystem and new model training approaches

Key Themes in Q2 '25

Applications move towards 'agentic by default'	• **Agentic workflows continue to become embedded** in a wide range of AI applications that previously used linear execution and minimal tool use, such as chatbots, terminals, and data analysis tools • **Deep research agents became table stakes** for major chatbots and some smaller Chinese lab entrants
Ecosystem of connections continues to grow and enable new functionality	• Applications such as ChatGPT and Claude **expanded their suite of integrations**, both with internally developed tools and increasing Model Context Protocol (MCP) compatibility • First and third party **MCP servers proliferated** across a range of APIs and businesses, with strong usage in developer and consumer applications
Coding agents see rapid growth	• **Q2 saw an unprecedented volume of new coding agent products**, with 12 major coding agent launches within the quarter, including frontier lab products like OpenAI Codex and Gemini CLI • **Coding agent usage has rapidly grown;** approximately half of the respondents to the Artificial Analysis AI Adoption Survey use or are considering using Cursor
Agentic model use will drive up LLM usage costs	• **Agents incur additional usage** of tokens and tools, driving increased costs; **deep research APIs** launched in Q2 have demonstrated costs of **up to $28 for a single complex query** in testing
Training focuses on agents and long-horizon tool use	• Reasoning models and **reinforcement learning has enabled more effective tool use**, including interleaved with model thinking before producing responses to users • Model creators increased the **emphasis in training on long-running tasks and agentic workflows** for models such as the Claude 4 family and Kimi K2

Agents are autonomous systems driven by LLMs…

What are agents?

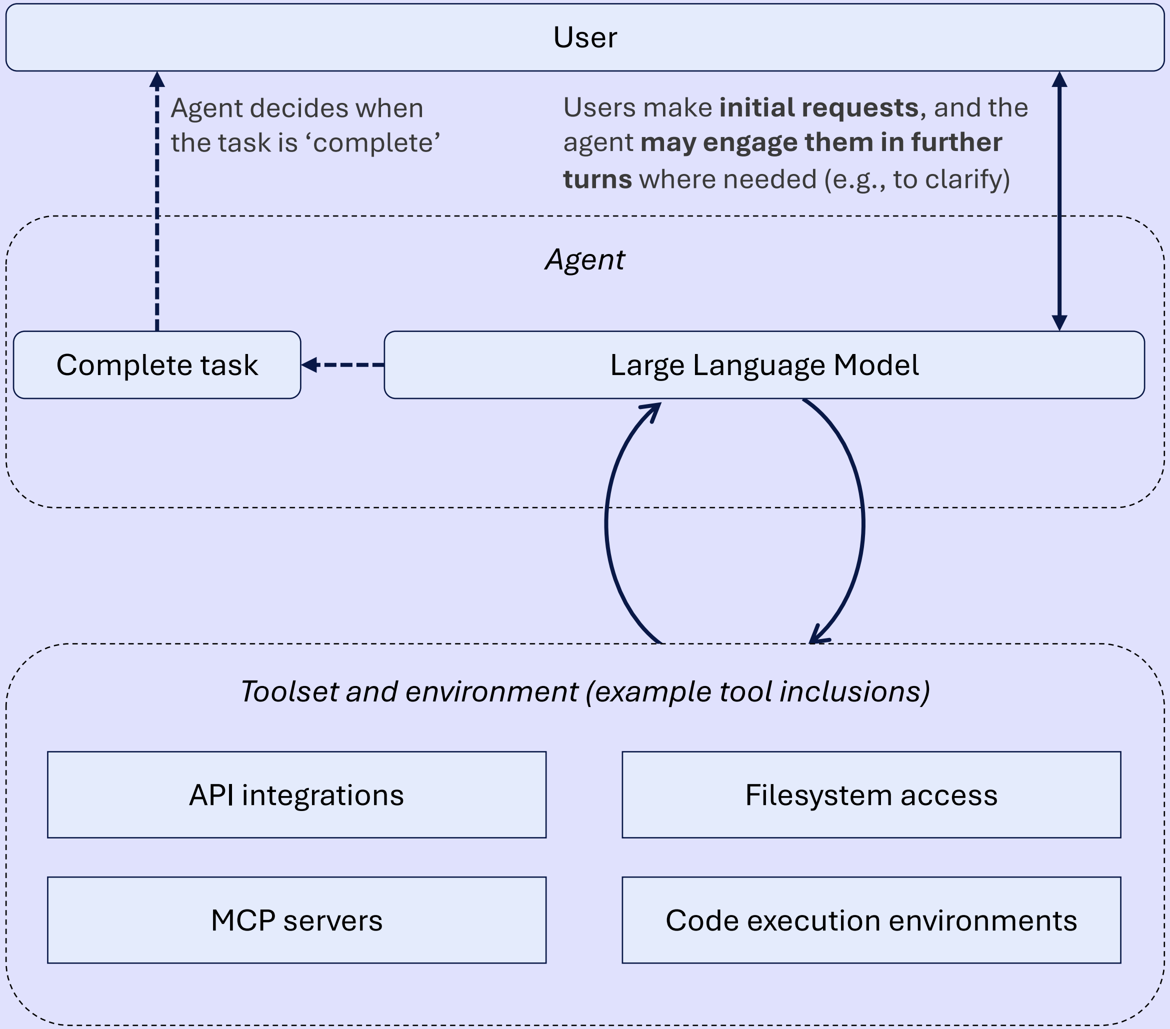
“Systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks”

“Agents represent systems that intelligently accomplish tasks, ranging from executing simple workflows to pursuing complex, open-ended objectives”

“AI agents are autonomous systems powered by large language models (LLMs) that, given high-level instructions, can plan, use tools, carry out steps of processing, and take actions to achieve specific goals”

AI agents are LLM-driven systems that act autonomously and use tools to complete tasks end-to-end

Fundamentally, agents in every domain run in a loop and take actions by using tools, such as searching the web or writing to a file



1. Can include direct vision capability or processing of other data such as website HTML to identify actions
Source: Company website; Anthropic ‘Building effective agents’ Agent workflow definition

Built on the improving intelligence of language models, AI agents offer key benefits compared to traditional workflows and are seeing success in several domains

Overview of agent benefits

Agentic approaches **enable new AI-based applications** due to a range of key benefits **compared to static workflows:**

1. **Dynamic planning, task tracking, and execution** for complex unknown task requirements to pursue well-defined goals
2. **Integration with a wide range of systems** and processes across a domain without a clear sequence of dependencies or 'chains' of use to complete tasks
3. **Natural collaboration** to complete tasks, including engaging human users in the loop to clarify or continue tasks, or coordinating with other agents with additional capabilities
4. **Graceful error recovery** from feedback where errors occur, even with unique or unexpected failure modes

Key domains showing early successes

Coding	**Finds, interprets, edits and tests source code** to complete software engineering tasks
Deep research	**Parses (and potentially clarifies) a research query and** launches a **chain of targeted research queries while controlling its research flow** to synthesize an answer
Computer use	**Interprets user commands**, 'looks' at a desktop or browser window[1], and **autonomously chains clicks, keystrokes, shell commands, and API calls** to complete arbitrary task
Customer support	Customer agent in live speech or text chat which **identifies intent** and **responds to customers in real time**, while **chaining required app, CRM or API calls** to complete the task (or hand off to a human agent)
Sales	**Identifies potential leads**, executes **personalized outreach**, and **integrates with sales tools**

For a deep dive on the latest progress in AI agents, see the **Artificial Analysis Q2 Agents and Applications Report**

Several competing players are emerging in the big agent domains in 2025; leading labs are focused on coding, research, and computer use

NON-EXHAUSTIVE

Frontier lab product

Domain	Illustrative products & players
A Coding	GitHub Copilot, CURSOR, CLAUDE CODE, OpenAI Codex, replit, GEMINI, jules
B Deep research	Gemini, perplexity, Claude, Qwen3, Grok, Mistral AI
C Computer use	ADEPT, comet, Browser Use, SURFER-H
D Customer support	Fin, ada, SIERRA, freshworks, zendesk, Lorikeet, Decagon
E Sales	11x, AiSDR, ARTISAN, PersanaAI, Relevance AI, salesforce

These domains are **showing the most progress in commercial off-the-shelf products** and research previews, while other use cases are following.

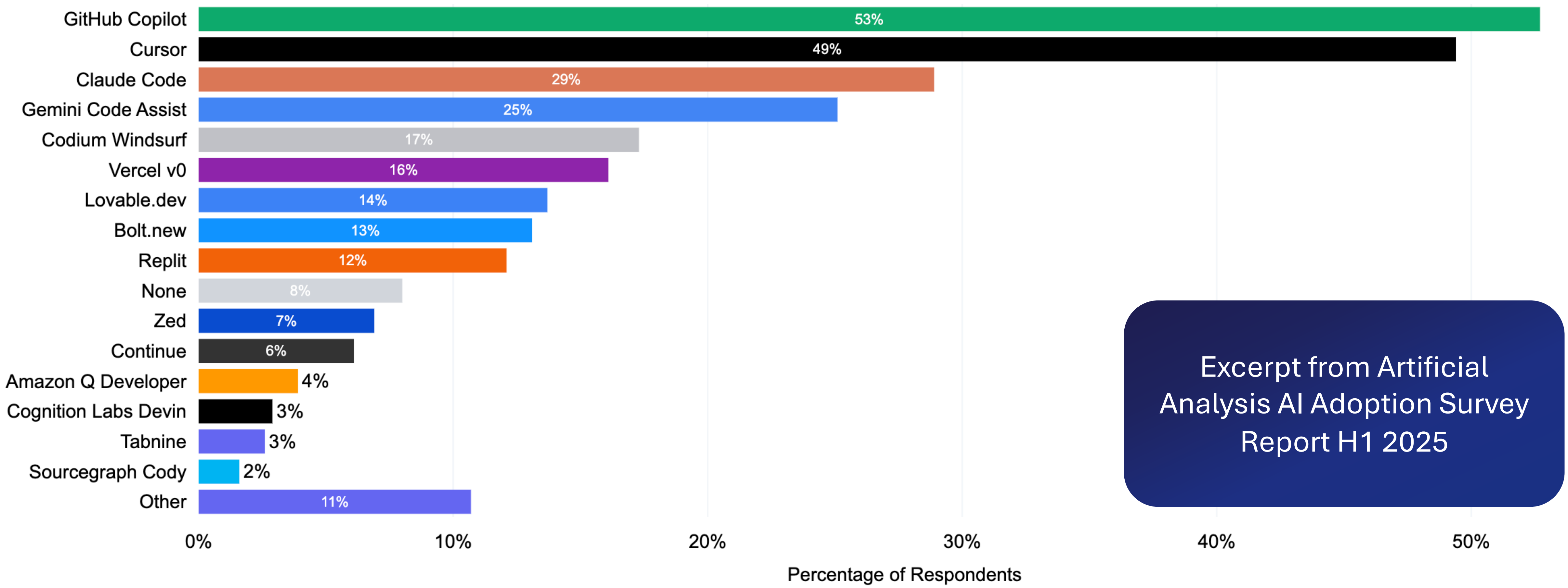
In parallel, a range of providers are **enabling users to build custom AI agents** for their use cases:

- Relevance AI
- StackAI
- SmythOS

A. AI Coding Tools: GitHub Copilot and Cursor dominate the market as the most popular AI coding tools, with a significant lead over Claude Code and Gemini Code Assist

Demand for Coding Tools

Which AI tools are you using or considering using this year? N=955



Excerpt from Artificial Analysis AI Adoption Survey Report H1 2025

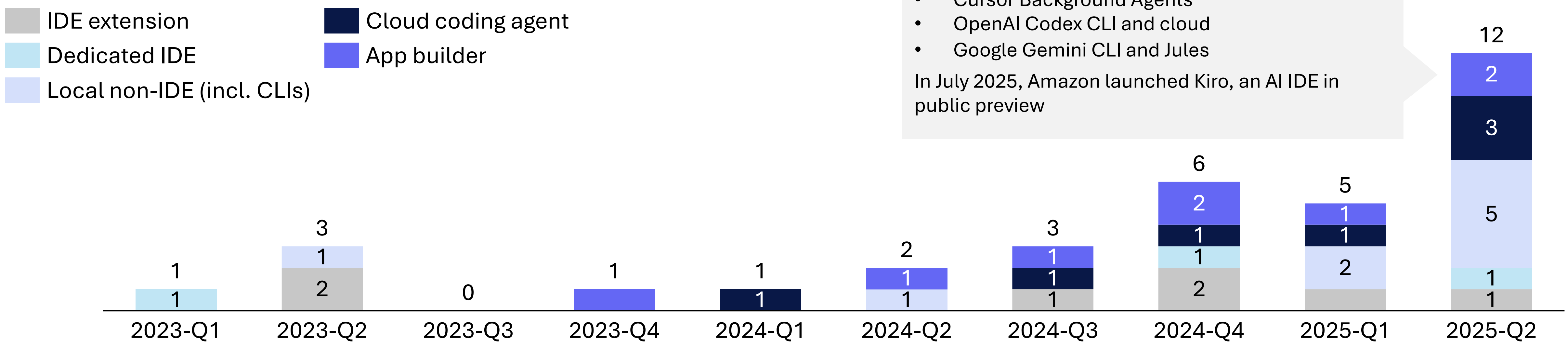
Source: Artificial Analysis AI Adoption Survey – H1 2025

A. Coding agents: Coding agent launches have accelerated in 2025, with a large focus on command-line interfaces

NON-EXHAUSTIVE

Major coding agent product launches by quarter

Number of coding agent launches identified, based on primary form factor[1]



Notable releases in Q2 included:

- GitHub Copilot Coding Agent
- Cursor Background Agents
- OpenAI Codex CLI and cloud
- Google Gemini CLI and Jules

In July 2025, Amazon launched Kiro, an AI IDE in public preview

Releases include

aider, Cody, Continue, cline, replit, Windsurf, augment code, lovable, ROOCODE, CLAUDE CODE, codename goose, OpenAI Codex, Background agents, Amp, GitHub Copilot

1. Primary form factor is assessed qualitatively when multiple modes apply

Note: Release timings are best estimates, based on general or public availability where possible. Where relevant, timing is based on when AI coding agents became a core product capability

03

Image and Video Models

State of AI – Q2 2025

Q2 '25 saw a shift in progress to Video models, with audio support and breakthroughs in quality, while open weights model progress slowed in both image and video

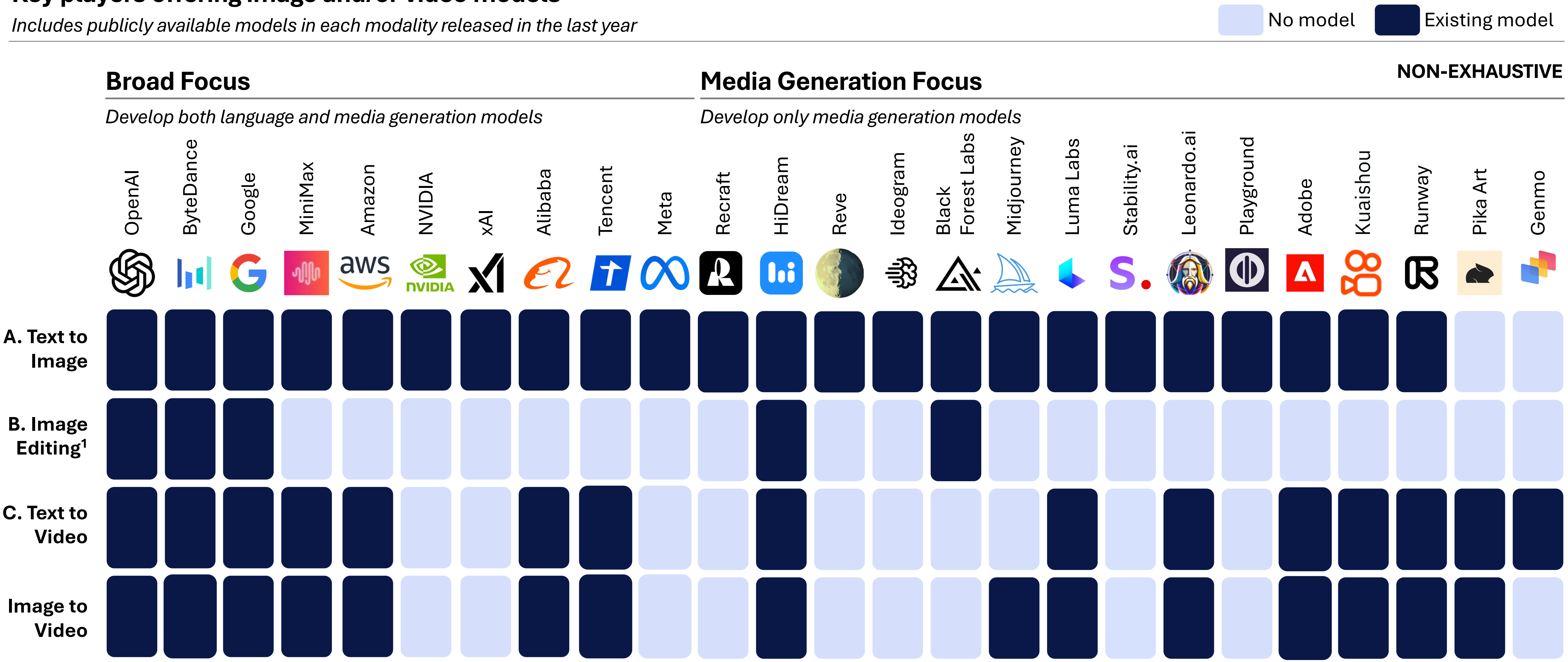
Key Themes in Q2 '25

Video models begin to support audio	• **Veo 3 released in May 2025 becomes the first high quality, mainstream model that natively supported audio generation** as part of a video model, driving strong adoption • **Veo 3's differentiated audio support gives it strong pricing power** at \$0.75/s of 720p video with audio, surpassing comparable models such as Seedance 1.0 at ~\$0.13/s of 1080p video, and Hailuo 2 at ~\$0.08/s of 1080p video
Video models see continued breakthroughs in quality	• **Video models see a breakthrough in quality,** with Seedance 1.0 overtaking both Q1 leaders: Veo 2 in text to video by ~150 ELO, and Kling 1.6 Pro in image to video by ~200 ELO points • **Labs shift focus to image to video generations,** with a larger ELO jump than text to video, and models such as Midjourney V1 and Kling 2.1 Pro available only in image to video variants • **Open weights video models lag behind proprietary alternatives,** with Alibaba Wan 2.1 still representing the SOTA for open weights text to video generation and LTX Video v0.9.7 13B ranking 16th overall for image to video on the Artificial Analysis leaderboard
Image editing models launched	• **Instruction based image editing models become popular**, with GPT-4o continuing to hold the lead, but FLUX.1 Kontext [max] and HiDream-E1.1 launched as competitive models in Q2 • **Open weights image models remain competitive in Image Editing,** with HiDream-E1.1, and FLUX.1 Kontext [dev] still in the top 5 of image editing models
Chinese and US labs continue at parity in media generation	• **Chinese labs continue to be image AI leaders** with Bytedance's SeeDream 3.0 achieving effective parity with GPT-4o, and HiDream's Vivago 2.0 achieving similar quality to Google's Imagen 4 • **Chinese labs lead in video generation** with Bytedance's Seedance 1.0 leading both text to video, and image to video leaderboards; Google is the only US lab that has released a SOTA video model in Q2
Text to image models improve incrementally	• **Text to image progress has slowed,** with GPT-4o and SeeDream 3.0 remaining quality leaders, with Google's Imagen 4, and BFL's FLUX.1 Kontext [max], not driving increases in frontier image generation performance • **Progress in open weights text to image models has stalled**, with HiDream-I1-Dev (the open weights leader in Q1) remaining the best open weights model in the Artificial Analysis Text to Image Leaderboard

Unlike in language models, smaller media generation focused organizations continue to compete with larger organizations with a broad focus across various modalities

Key players offering image and/or video models

Includes publicly available models in each modality released in the last year



1. Instruction-based image editing

04

Speech Models

State of AI

Speech: Developments across the speech AI stack are driving voice agents to become more natural, more powerful and cheaper

Key themes in Q2 '25

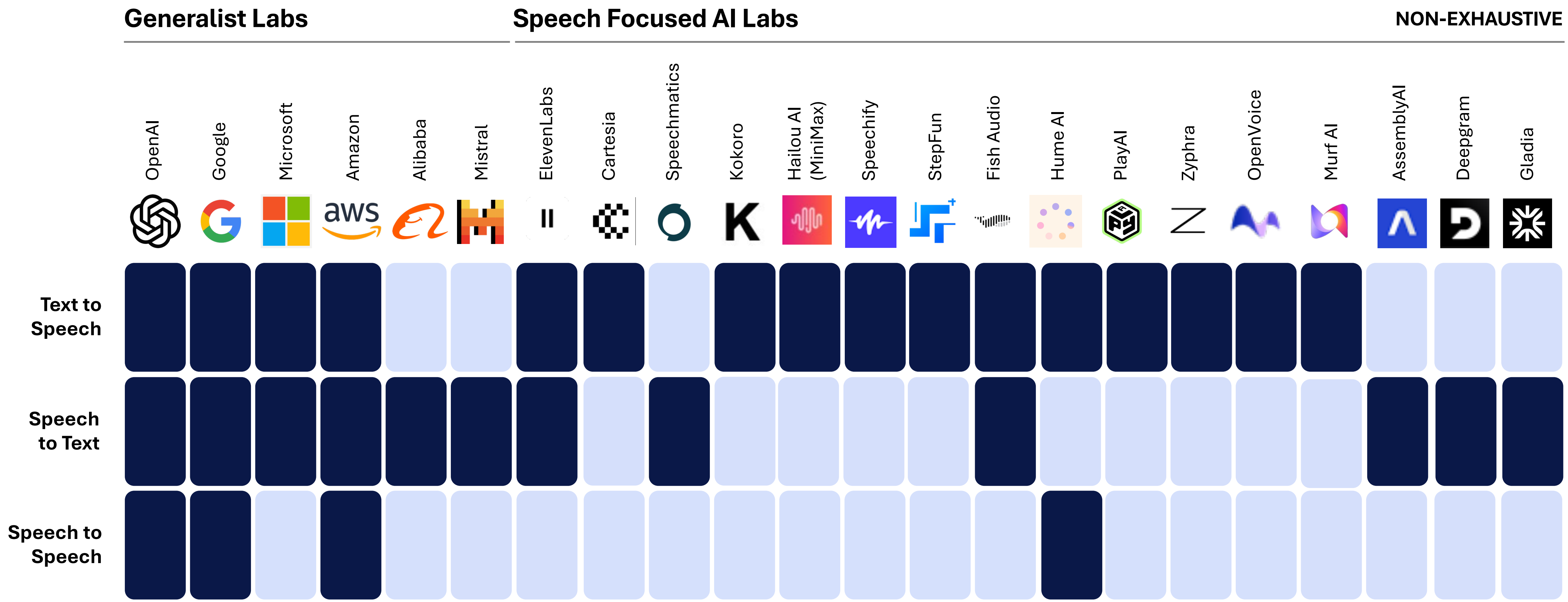
Leading Text to Speech models achieve greater realism	• **Recent Text to Speech releases**, including MiniMax's Speech-02-HD, Cartesia's Sonic-2 and Nari Labs' Dia model have **pushed towards more human-like dialogue** • Looking forward, we may see divergence between models focused on dialogue and narration use cases
Open source Text to Speech models drive down the cost of speech synthesis	• **Key open weights Text to Speech model releases** including Kokoro's 82M and Sesame's CSM 1B have allowed third party inference providers to serve near-frontier TTS for the first time – taking prices to new lows
End to end speech models slow to take off	• There are now **three major end-to-end speech models** (language models with direct input and output of speech): OpenAI's GPT-4o, Google's Gemini 2.0 Flash, and Amazon's Nova Sonic • **End-to-end Speech to Speech models** can be simpler to work with than the currently standard pipeline approach combining Speech To Text, Text To Text (language models), and Text To Speech models – as well as providing benefits like lower latency and deeper understanding of tone and emotion • However, **end-to-end speech models have seen low uptake in early 2025**, driven in part by APIs being in beta and not yet having full feature sets or production stability
Dominance of OpenAI Whisper challenged for Speech to Text	• **A series of Speech to Text releases**, including OpenAI's own GPT-4o transcription and ElevenLabs' Scribe offer compelling alternatives to Whisper • **Third party OpenAI Whisper APIs** from inference providers such as Fal, Fireworks, Groq and Deepinfra **continue to be the lowest cost and highest speed** options for Speech to Text

Speech: While the generalist AI labs (e.g. OpenAI, Google) have offerings across all speech AI, an emerging group of pure-play speech companies are driving innovation

Key players offering speech models

Classifications are indicative and determined based on models available

Low or no presence   Strong presence



05 Accelerators

State of AI – Q2 2025

Demand for AI accelerators is expected to accelerate as multi-node inference becomes popular; NVIDIA Blackwell becomes widely available while AMD unveils the MI355X

Key themes in Q2 '25

Inference demand continues to increase quickly	• Reasoning models, longer contexts, and agents are multiplying compute demand per user query • OpenAI, Google, Anthropic and more all reported 'running out of compute' in H1 2025, leading to product launch delays (eg. Gemini 2.5 Pro general availability) and strict rate limits (eg. GPT-4o image generation, Deep Research)
NVIDIA Blackwell is running production workloads; first GB200 rack-scale	• 2024 saw the first 100K H100 clusters; 2025 will see 200K+ GB200 clusters • NVIDIA's rack-scale NVL72 combines 72 GB200 chips in a single switched NVLINK domain • Blackwell may enable training of >10T parameter models
Increasing focus on system performance over chip performance	• Increasing the size of both scale-up domains (single coherent system, eg. NVL72 connected with NVLINK) and scale-out domains (networked nodes, ethernet based networking technologies) allows the delivery of greater training compute • Multi-node performance has traditionally only been a focus for training, with most models being deployed for inference on single systems; this begins to change in 2025 with multi-trillion parameter models and distributed inference techniques delivering greater performance
Distributed inference poised to become a critical battleground	• Inference techniques confined until recently to the frontier labs are becoming widely available – driven by DeepSeek's open sourcing, NVIDIA Dynamo and upcoming work from open source projects including SGLang • Key techniques include prefill/decode disaggregation and expert parallelism across dozens or hundreds of GPUs, along with novel load balancing techniques like scaling expert replicas depending on activation frequency
US/China chip tensions intensify: Washington considers H20 ban, Huawei announces NVL72-competitor	• In April 2025, the Trump administration barred NVIDIA and AMD from selling the H20 and MI308 accelerators to China, closing off the last remaining Hopper-generation chip • Huawei is emerging as China's chip leader, designing chips and systems that may approach Hopper-level performance that are manufactured on a mix of TSMC and SMIC nodes • China's ability to locally manufacture a Hopper-level accelerator remains unknown

NVIDIA continues to dominate the AI accelerator market, especially for frontier-class training, but a growing list of challengers now offer material differentiation

Key players building accelerators for AI training and inference

Based on publicly available data of chips yet to be released and/or available for use

No available chips | Existing available chips

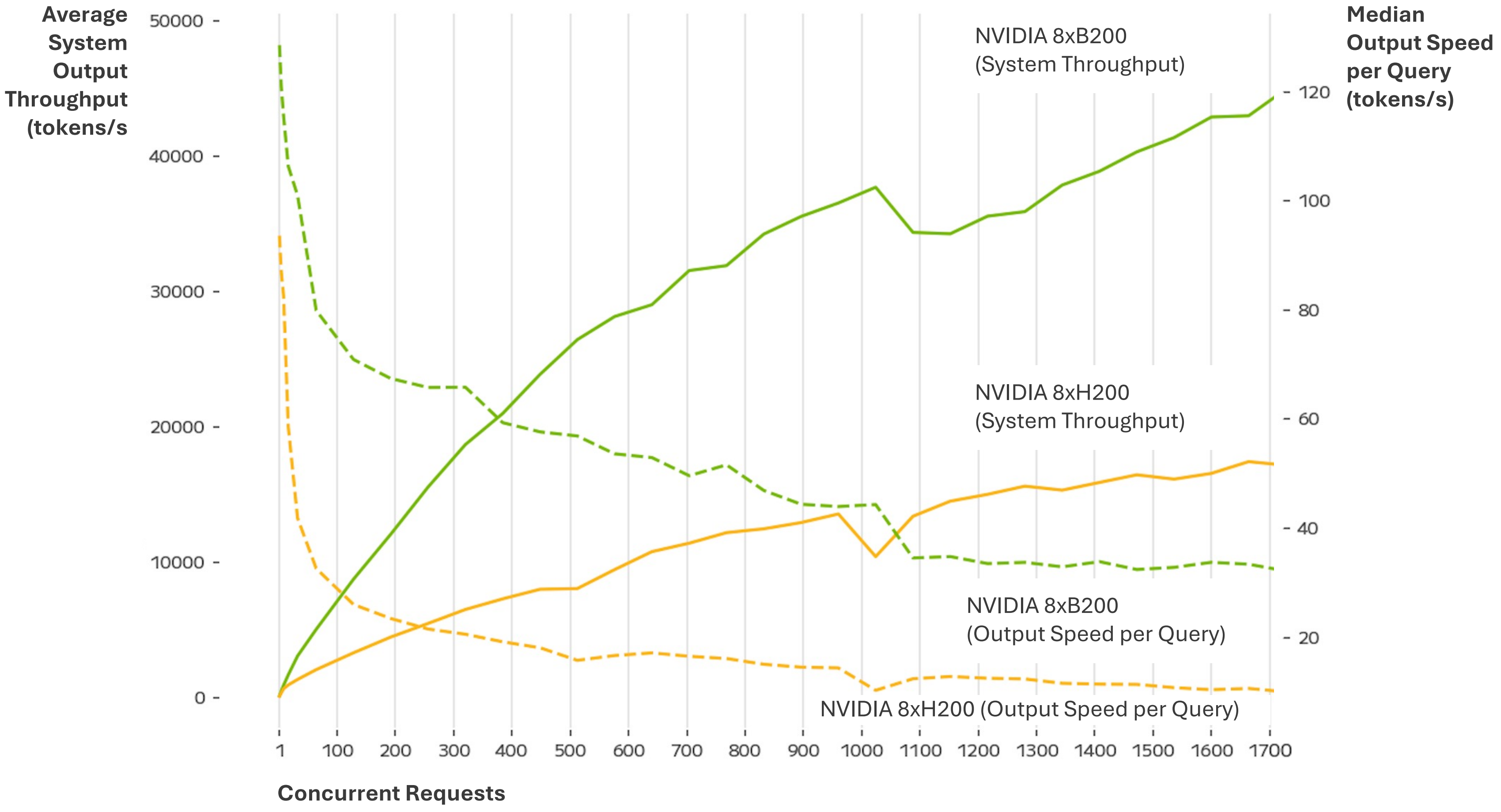
	Major Chipmakers				Cloud Hyperscalers		Challengers			Emerging startups		Pre-product		
	NVIDIA	AMD	Intel[1]	Qualcomm	Google	Amazon	SambaNova	Groq	Cerebras	Furiosa	Positron	Rivos	Etched	MatX
Training	Existing available chips	Existing available chips	No available chips	No available chips	Existing available chips	Existing available chips	Existing available chips	No available chips	Existing available chips	No available chips	No available chips	Existing available chips	No available chips	Existing available chips
Inference (incl. inference within training)	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips	Existing available chips

1. Intel is no longer bringing to market Falcon Shores, its successor to Gaudi 3, and we do not expect Intel to bring Jaguar Shores to market until 2026 at the earliest.

Artificial Analysis System Load Test: NVIDIA's new B200 accelerator significantly outperforms NVIDIA H200

System Throughput (Left) and Output Speed per Query as Concurrent Requests increases

Llama 4 Maverick (FP8); NVIDIA TensorRT-LLM v1.0.0rc2



Source: Artificial Analysis Hardware Benchmarking - System Load Test

Commentary

- **NVIDIA B200 delivers superior performance across all metrics in the Artificial Analysis System Load Test**, demonstrating higher system throughput, per user output speed and better scalability under load
- **3x system throughput advantage at scale:** B200 output ~39K tokens/s versus H200's ~13K tokens/s at 1000 concurrent requests
- **Consistently higher per-user output speed** B200 maintains 1.3x faster output at low concurrency (>120 vs ~90 tokens/s per query) and 3.5x faster performance under high load (~35 vs ~10 tokens/s per query)

Artificial Analysis

contact@artificialanalysis.ai

https://artificialanalysis.ai/