



Artificial Analysis State of AI

Artificial Analysis State of AI: 2025 Year-End Edition

This report includes model releases up to the end of 2025. For the latest benchmarking results, visit the live Artificial Analysis website at artificialanalysis.ai.

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 Premium Insights: Comprehensive AI market intelligence and insights for enterprise decision making from the leading independent benchmarking company



AI Market Intelligence

Quarterly State of AI Reports

Stay ahead of AI market developments with the definitive quarterly update, incl. China report

[This Report](#)

AI Adoption Survey

Gain real-world adoption insights from those building and deploying AI

Quarterly AI Webinars

Connect the latest AI market intelligence to your strategic priorities



AI Capability Guides

Enterprise Agents Guide

Discover how agents are reshaping productivity and deployment across industries

Model Deployment Guide

Compare models, inference providers, and hardware with specific benchmarks

Additional guides launching soon

New guides are added regularly, with a focus on high priority capabilities



AI Strategy Support

Leaders AI Strategy Guide

Equip your leadership team to harness AI effectively at organizational scale

Applied AI Trends Workshops

Engage your teams in an interactive 90-minute deep-dive on the most important AI trends

Bespoke Support

Accelerate your AI strategy with expert support on planning, architecture, and implementation



AI Benchmarking Support

AI Databooks & API Access

Access the industry's most comprehensive AI performance and cost data

AI Custom Benchmarking

Evaluate and compare models, chips, and providers through our custom independent benchmarking

AI Launch Support

Strengthen your AI launch with trusted performance metrics, brand assets, and independent validation

Trusted by the leading AI industry players, media and institutions



Entities that have publicly referenced Artificial Analysis

Join the world's leading AI labs and enterprises with subscriptions starting from \$3K per quarter
subscriptions@artificialanalysis.ai

This Highlights version of the Quarterly State of AI Report is a limited version. The full report is available to subscribers of our Premium Insights Subscription

Highlights version (this version)

- ✓ **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 H100, H200 and B200 using Artificial Analysis System Load Test

Full Version (Premium Insights Subscription)

Includes everything in the Highlights Version plus:

- ✓ **New language model release coverage and analysis** (incl. analysis of leading open weights options)
- ✓ **Model trends analysis** outlining emerging trends for language models across pricing, performance and features
- ✓ **Agents coverage** including analysis of key agent categories, use-cases and implications for real-world deployment
- ✓ **Image generation models and trends** (incl. text to image and image editing)
- ✓ **Video generation models and trends** (incl. text to video and image to video)
- ✓ **Speech models and trends** (incl. text to speech, speech to text and native speech to speech models)
- ✓ **Emerging market trends** for accelerators, including detailed analysis comparing NVIDIA H100, H200 and B200

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

Artificial Analysis State of AI: 2025 Year-End Edition

Just several months ago, in a letter just like this one, we proclaimed that rumors of AI progress slowing had been greatly exaggerated. In early 2026, the idea that we would start a letter like that seems ridiculous.

At the start of 2025, coding agents didn't exist. By the end of the year, the profession of software engineering had changed forever - from copy-pasting code into ChatGPT and Cursor Chat to instructing agents that work autonomously for several minutes at a time. We expect 2026 to be the year of agents for everything else.

There was no consolidation of the race in 2025 – competition only intensified, contributing to the cost of every level of intelligence continuing to fall consistently. Progress was driven by labs scaling reinforcement learning, focusing on large sparse mixture-of-expert architectures, the arrival of Blackwell hardware and more.

Produced by Artificial Analysis, the leading independent AI benchmarking and insights provider, this 2025 State of AI Report is designed to inform product, engineering and investment 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

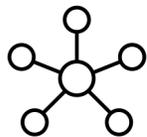
-
- | | |
|----------------------|---|
| 1. Industry Overview | Overview of market movements and trends by key players in the AI industry |
| 2. Language Models | Trends in frontier language models, including increasing agentic intelligence, cost and efficiency improvements |
| 3. Image and Video | Trends in frontier image and Video including an overview of the leading models in Artificial Analysis Image and Video Arenas |
| 4. Speech and Audio | Trends across new speech and music models and an overview of new and leading models in the Artificial Analysis Speech Arena |
| 5. Accelerators | Overview of the AI accelerator market including market trends, available accelerators and vertical integration by select chip makers |
-



01

Industry Overview

Artificial Analysis State of AI: 2025 Year-End Edition



The AI industry becomes more contested

The AI landscape diversified significantly in 2025, with an expanding list of companies releasing models. In 2026, we expect the race will continue to become more competitive, not less



Reasoning models become the status quo

At the start of 2025, OpenAI's o1 was the only 'reasoning' model, however 2025 saw all AI labs develop reasoning models that now occupy the spots for the most intelligent models



Agents take off

2025 marked the shift from single-query workloads to multi-turn agentic tasks. Coding agents led early adoption; 2026 will likely expand agents into broader enterprise workloads



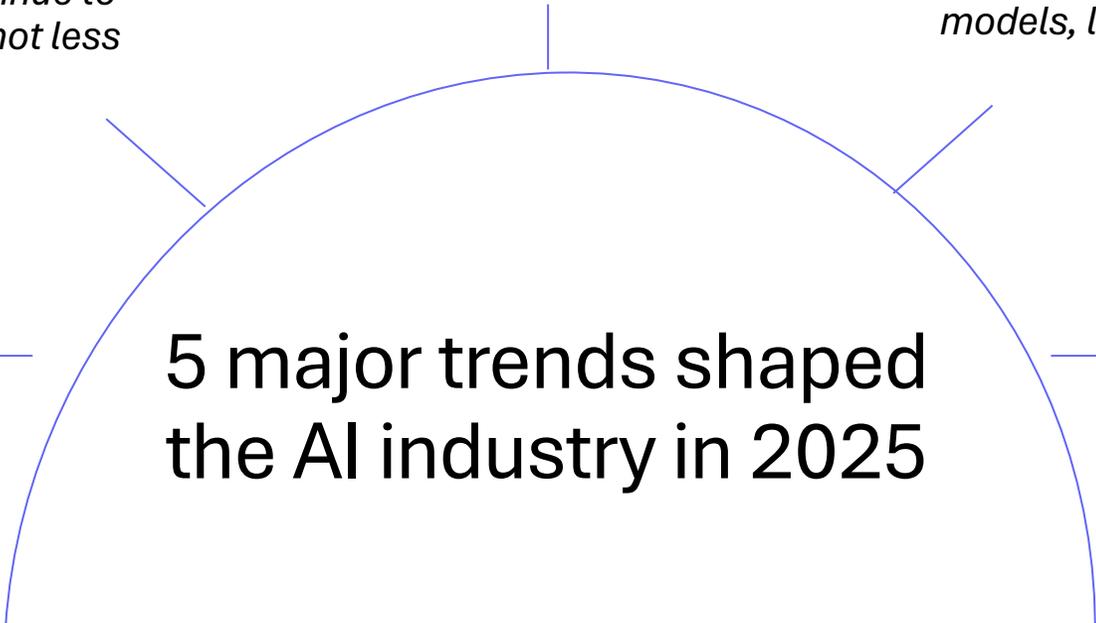
Native speech to speech models give rise to voice agents

2025 saw a massive improvement in speech to speech quality with the development of native audio reasoning models, laying the foundations for voice agents



Image editing and video generation go mainstream

Image editing and video generation reached mainstream viability, with models like Gemini 2.5 Flash (Nano Banana) delivering step-change quality improvements



5 major trends shaped the AI industry in 2025

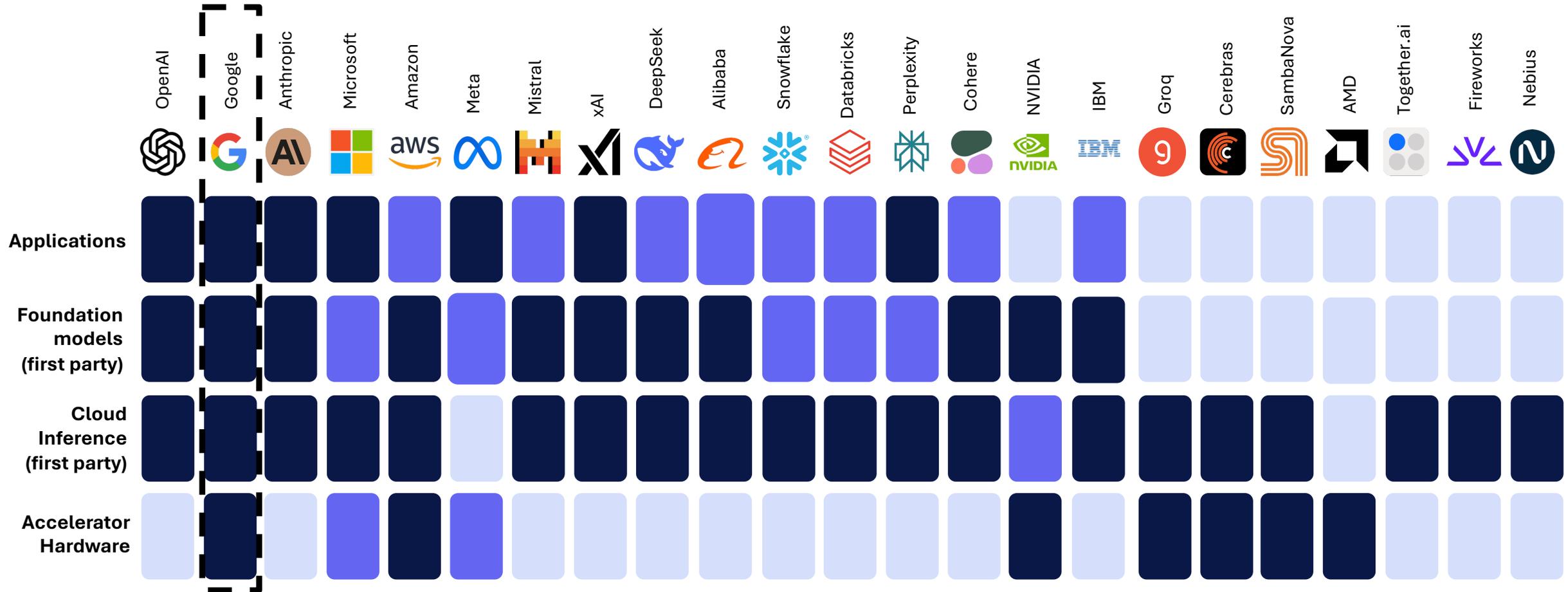
Google continues to be positioned as AI's most vertically integrated player from TPU accelerators to Gemini applications, spanning across the entire AI value chain

Key Players in the AI Value Chain

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

NON-EXHAUSTIVE

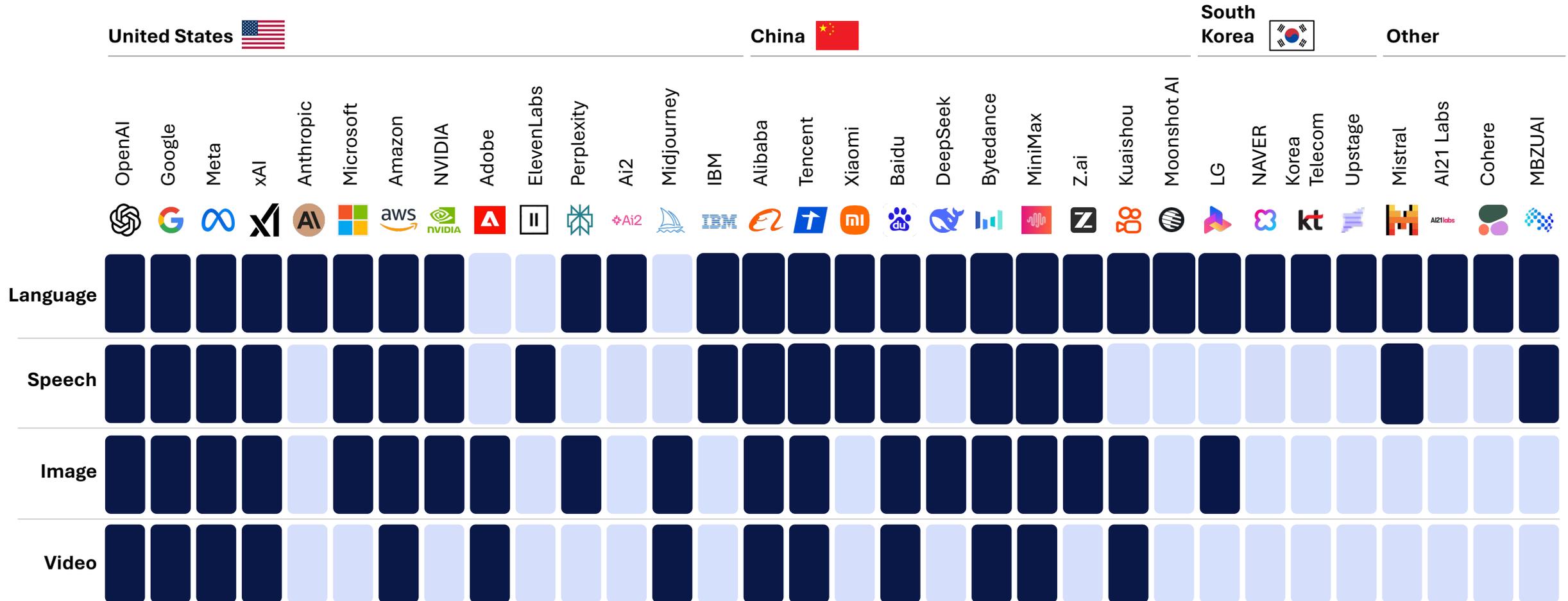
No presence    Strong presence



The AI landscape is becoming increasingly competitive, with new international labs entering the race in 2025, though US and China firmly lead

Key players with first-party models by modality

NON-EXHAUSTIVE No model Existing model



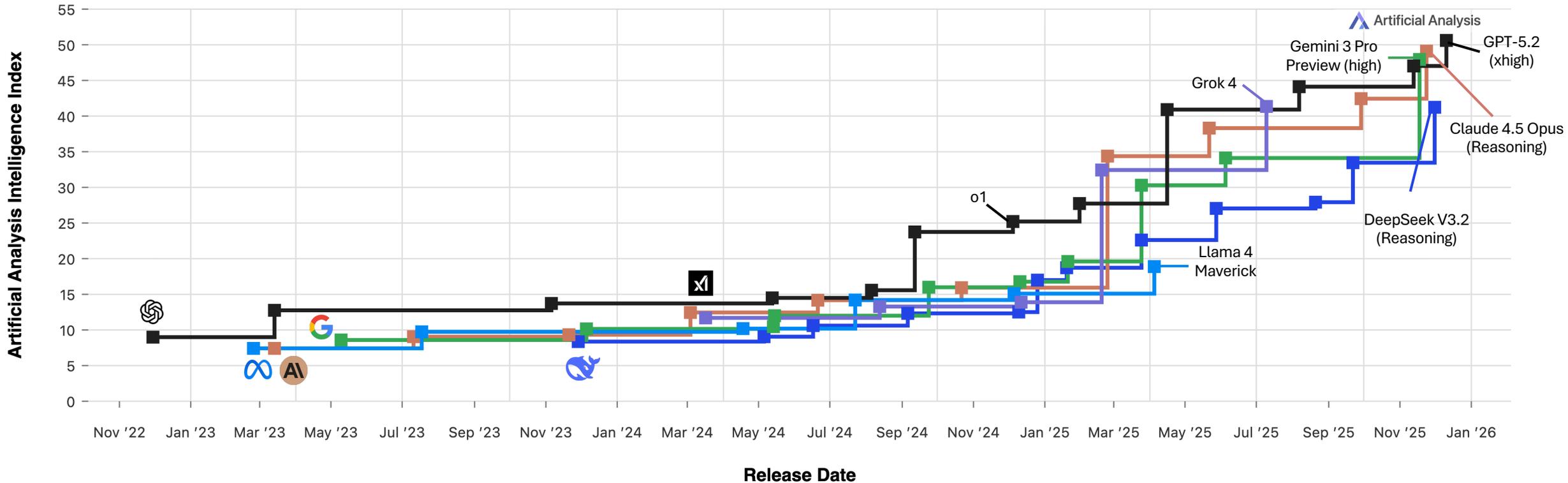
Source: Company website

OpenAI began and ended 2025 with the most capable language model, but their lead is narrower than ever

Frontier Large Language Model (LLM) Intelligence till Jan 2026

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

Anthropic DeepSeek Google Meta OpenAI xAI



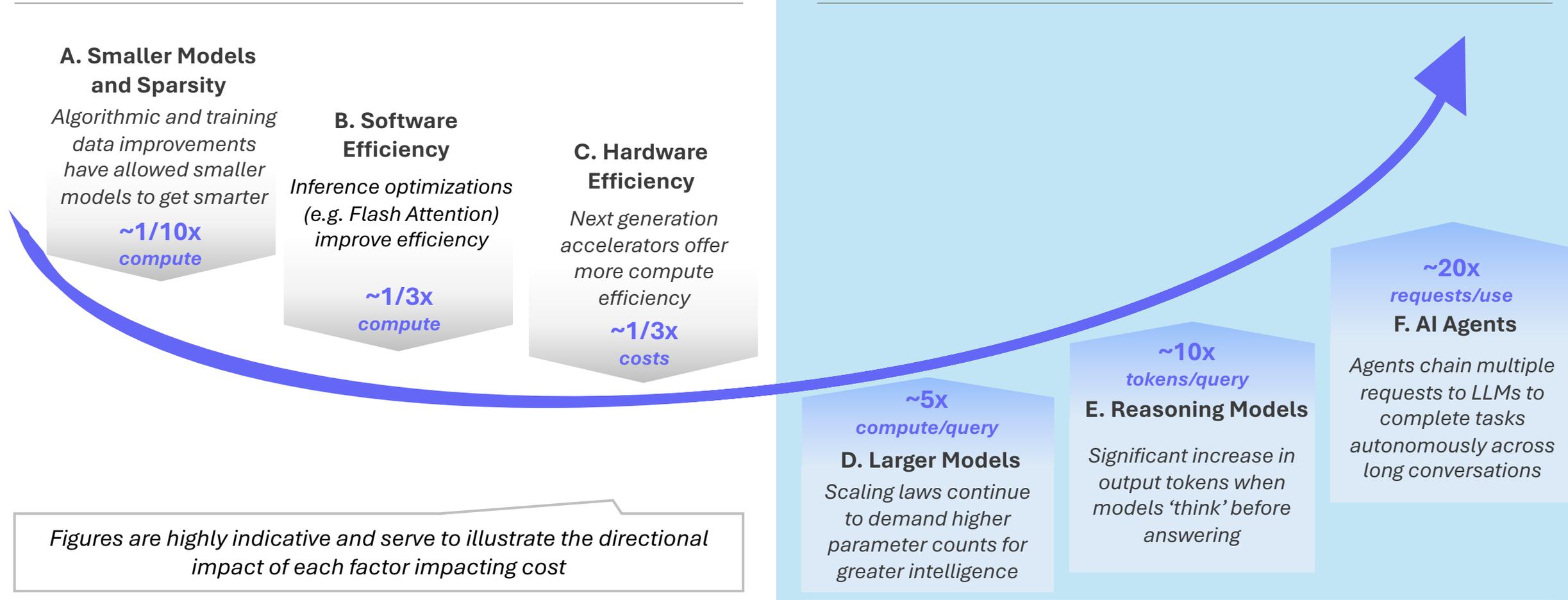
The intelligence frontier is now fiercely contested between OpenAI, Anthropic, and Google with increasing competition from labs from China. Meta has restructured its AI efforts and has not released a new model since April 2025

Efficiency improved through model scaling and hardware/software optimizations ...

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

... however larger reasoning models and more agentic workloads mean compute demand continued to increase

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





02

Language Models

Artificial Analysis State of AI: 2025 Year-End Edition

2025 was defined by the reasoning paradigm, driving significant intelligence gains, falling costs, and the rise of agentic AI, as open weights and global labs narrowed the gap with US frontier

Key themes

A. 2025 saw a significant increase in model intelligence, driven by a paradigm shift towards reasoning models that ‘think’ before answering

- By end-2025, OpenAI, Anthropic, and Google led the intelligence frontier with reasoning-first models that ‘think’ before answering - marking a clear break from early 2025, when non-reasoning models held the top spots as the most intelligent models
- At the same time, the reasoning paradigm materially expanded average workload size as models generated far more output tokens when ‘thinking’, while driving higher performance across general/scientific reasoning, long-horizon agentic tasks, and coding

B. 2025 marked the rise of agentic AI, with models increasingly executing long-horizon tasks end-to-end

- Agents evolved from targeted use cases (e.g. deep research) to generalized tools, with frontier models now reliably orchestrating multi-step workflows across domains
- Tool calling training is now universal with most models released in 2025 having been pre-trained and RL-optimized for agentic task execution
- Long horizon coding tasks were the largest beneficiaries of improvements in agentic workflow with a clear proliferation of coding agents being released in 2025 by small players and incumbents

C. 2025 witnessed a democratization of foundation models, though the US and China maintain a significant lead

- AI labs from across the world, including Europe, Middle East, and Asia continued to release competitive foundation models, however frontier capabilities remain concentrated around US (OpenAI, Anthropic, Google) and China (Moonshot AI, Z.ai, DeepSeek, Minimax)
- While the US labs continue to lead the development of proprietary frontier models, Chinese labs continue to release frontier open weights models

D. 2025 saw new open weights models continue to keep pace with proprietary models in intelligence, however the frontier is held by proprietary models

- In 2025, the open weights ecosystem expanded and by the end of 2025, the most capable open weights models were increasingly from Chinese labs
- Throughout 2025, open weights models broadly kept pace with proprietary models, but proprietary models still led overall intelligence

E. The cost of intelligence fell significantly for o1-level intelligence

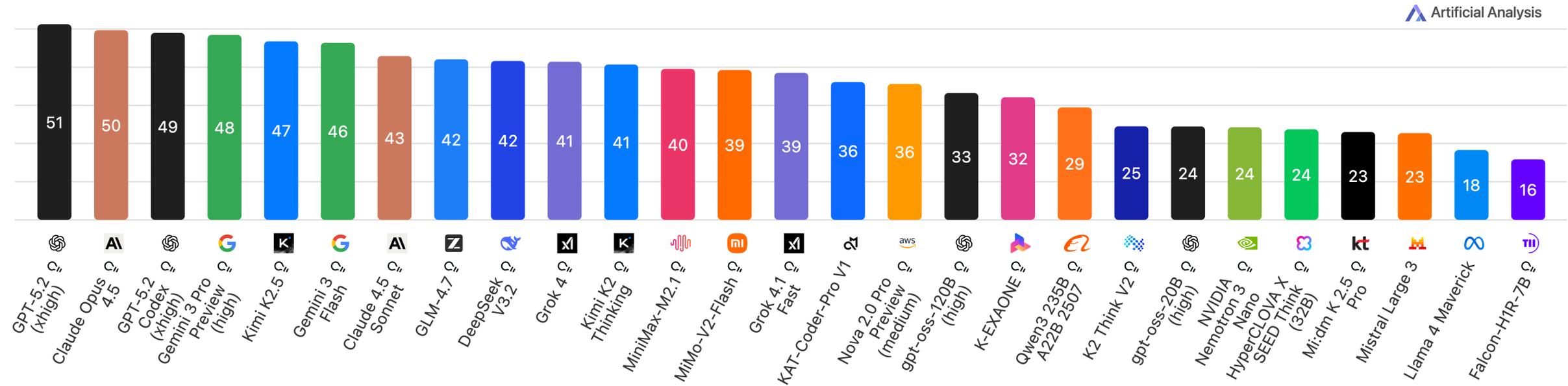
- Price per token fell 128x for the o1-level intelligence that 2025 began with – driven primarily by smaller models achieving higher level of intelligence, as well as software and hardware optimizations

A. As at the end of 2025, OpenAI, xAI and Anthropic lead frontier intelligence with their latest reasoning models with a significant gap to the next set of AI labs

Leading Large Language Models (LLMs)

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, r^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

C, Reasoning Model

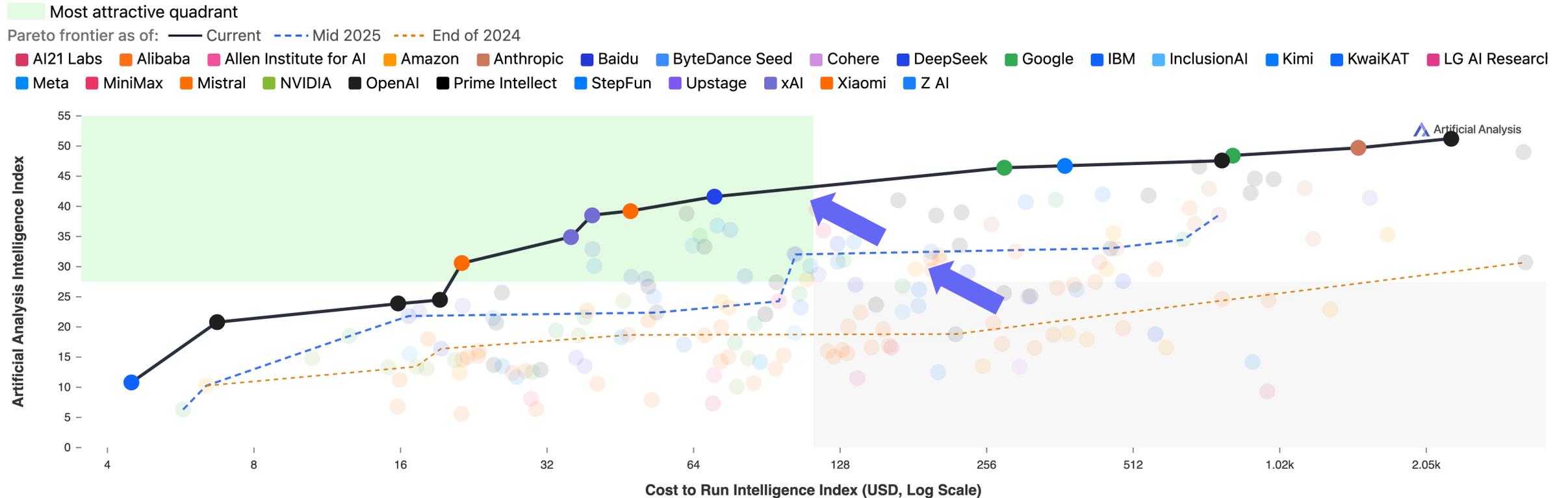


- **The reasoning paradigm that gave OpenAI a decisive lead at the start of 2025** has now been adopted by nearly every major lab, compressing OpenAI's lead in overall model intelligence
- **DeepSeek R1 released at the start of 2025, marked a turning point as the first open weights reasoning model challenging OpenAI's lead**, trained using novel techniques for pre-training and reinforcement learning
- **OpenAI still leads with GPT-5.2 (xhigh), but competes in an increasingly crowded frontier** where Anthropic, Google, xAI, and Chinese labs have all released competitive reasoning models

A. Models released in 2025 pushed the intelligence-cost frontier: organizations can now access higher intelligence at equivalent price points, or equivalent intelligence at materially lower cost

Intelligence vs. Cost to Run Intelligence Index (Pareto Frontiers)

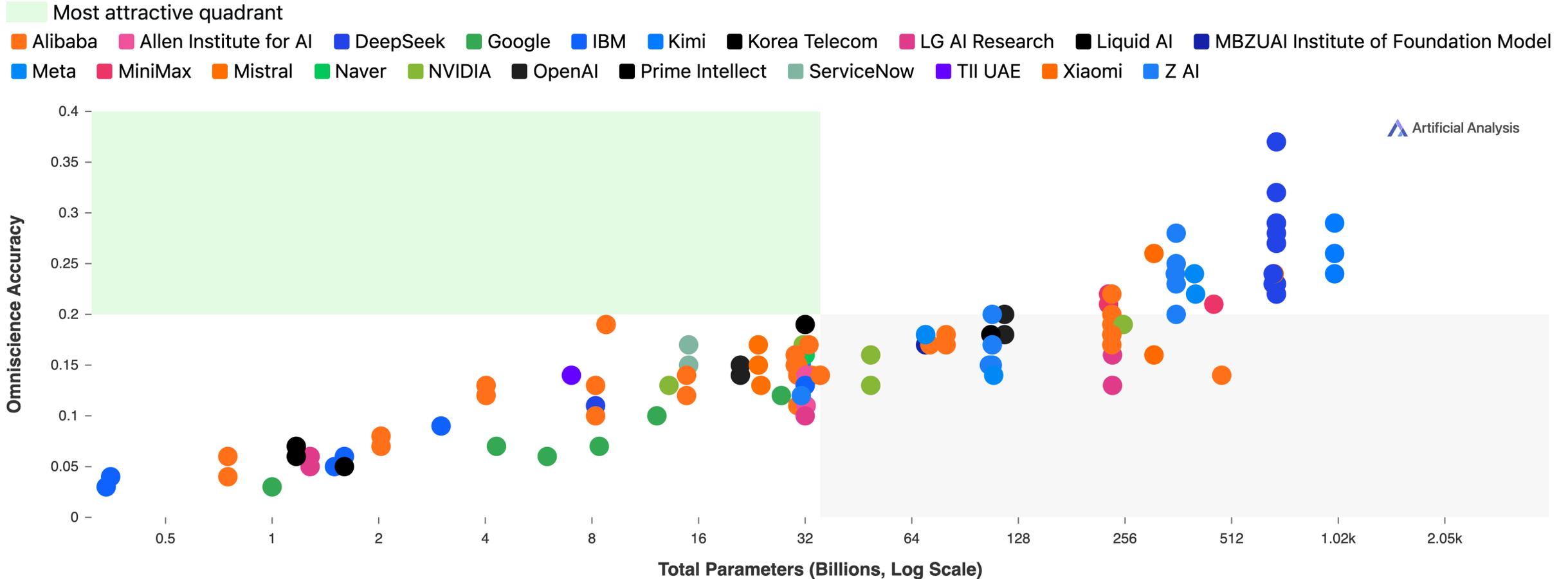
Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt; Cost to Run Intelligence Index; Pareto frontier over time



A. Deep Dive: Larger models (measured by total parameter count) achieve reliably higher AA-Omniscience accuracy scores...

AA-Omniscience Accuracy vs. Total Parameters (Open Weights Models)

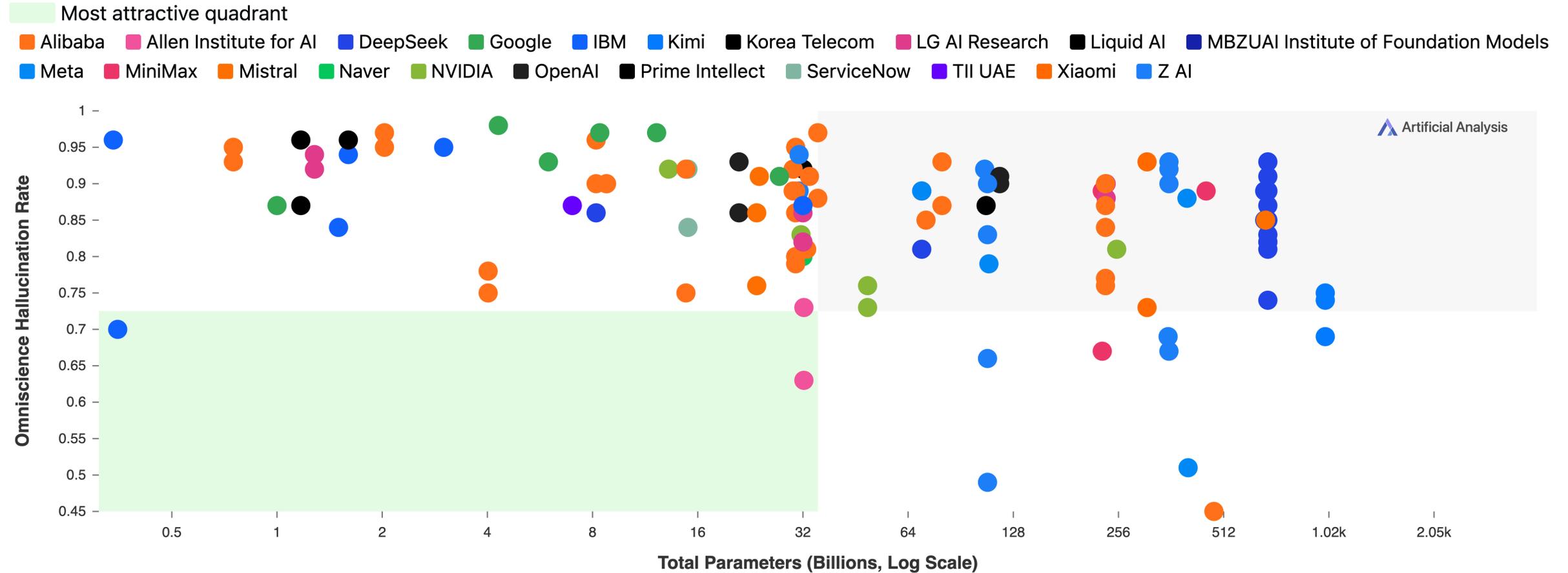
AA-Omniscience Accuracy; Size in Parameters (Billions)



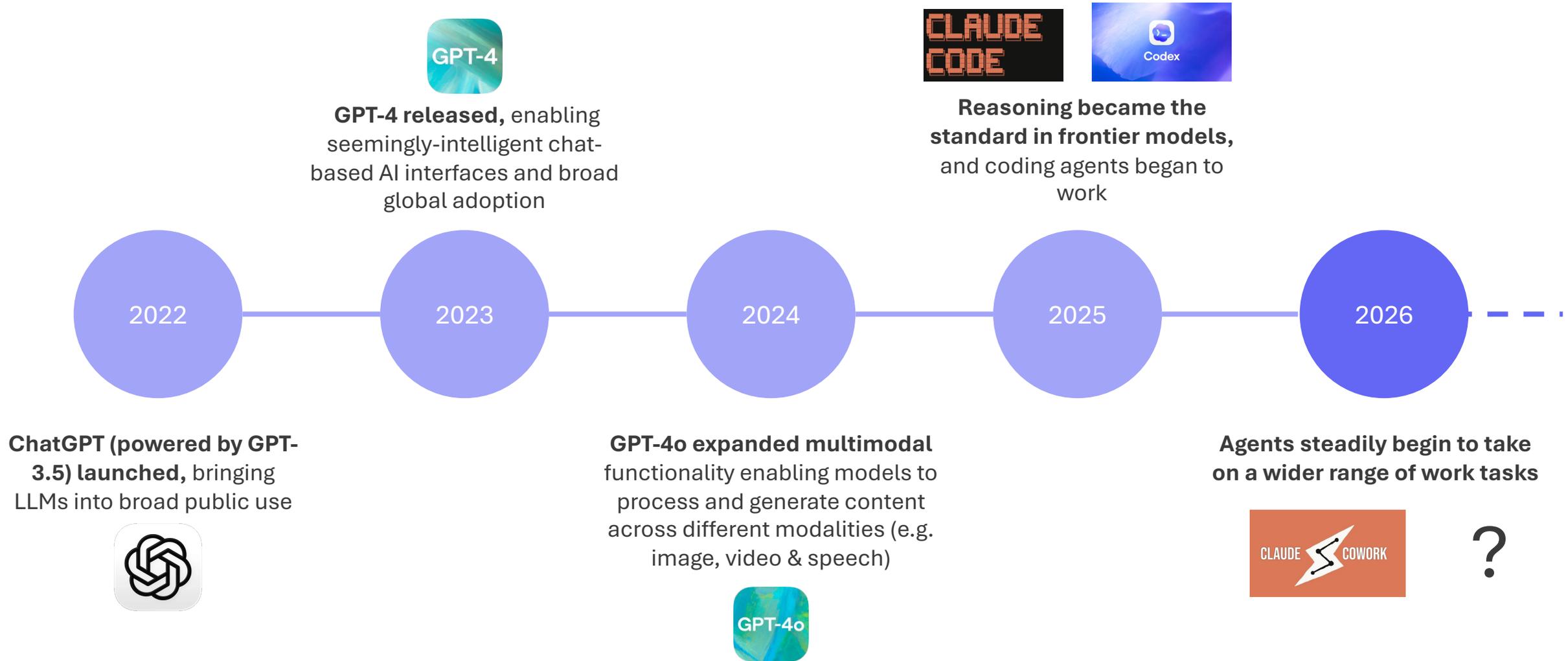
A. Deep Dive: ...but hallucination rate is less correlated with the size of the model, indicating greater impact of other training decisions

AA-Omniscience Hallucination Rate vs. Total Parameters (Open Weights Models)

AA-Omniscience Hallucination Rate; Size in Parameters (Billions)



B. 2025 was the year coding agents began to work; 2026 will be the year 'agents for everything' begin to work



C. Beijing is emerging as a hub of frontier AI startup activity; Established “Big Tech” companies are more geographically dispersed with no single nexus of tech innovation



1 Beijing

China's leading AI research center, combining top universities (Tsinghua, Peking), the Beijing Academy of Artificial Intelligence, the world's largest concentration of AI scientists, and Zhongguancun Science Park to dominate foundational research



2 Shanghai

AI hub featuring Shanghai Foundation Model Innovation Center (China's first and largest foundation model incubator), the Shanghai AI Laboratory, and government initiatives targeting a \$55 billion AI industry with strong semiconductor manufacturing support



2 Hangzhou

Rising AI hub pioneering smart city applications through Alibaba's City Brain platform and a dedicated 3.43 sq km AI Town with full 5G coverage for research and development



4 Shenzhen

China's AI hardware and robotics manufacturing capital, leveraging its world-class electronics supply chain and giants like Huawei, Tencent, and DJI for AI development



C. Korea's government-backed Sovereign AI Initiative has catalyzed the domestic AI ecosystem, producing multiple near-frontier AI labs

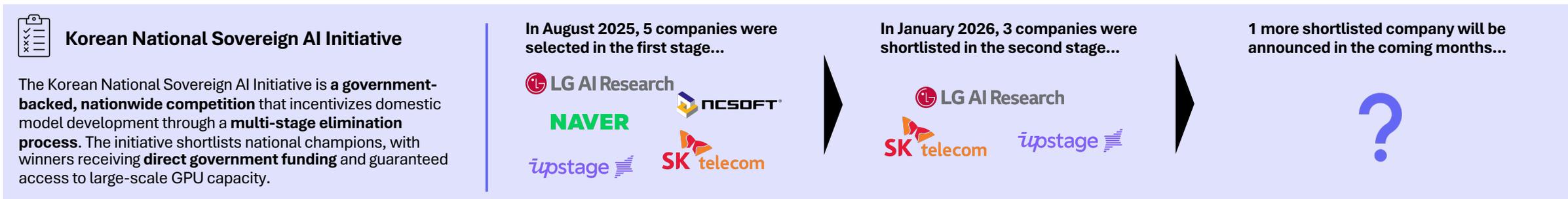


Description of key Korean model labs

NON-EXHAUSTIVE

Shortlisted in the National Initiative

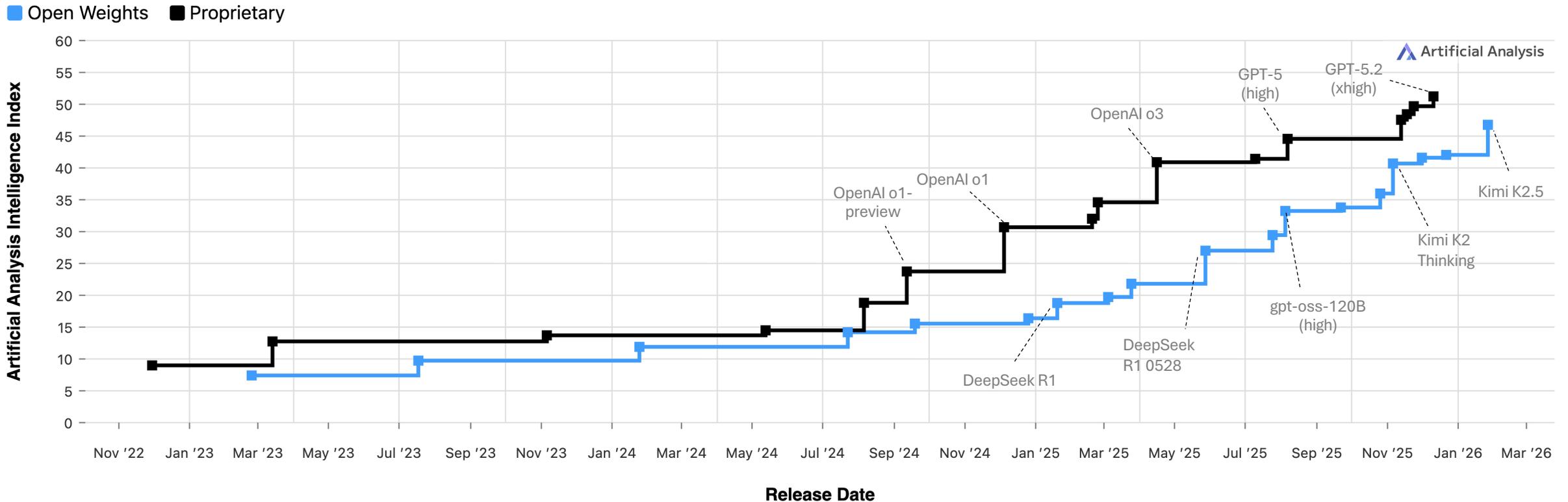
	LG AI Research	SK telecom	upstage	NAVER	ncsoft	Korea Telecom	MOTIF
Description	AI research arm of LG Group. LG AI Research has been the clear front-runner in the Korean AI race for the past year, consistently leading in intelligence benchmarks	Korean telecom giant founded in 1984 that leads the local wireless market. Developing language models as part of broader digital transformation efforts	AI startup founded in 2020 by Sung Kim (ex-Naver AI lead) that has quickly become a powerhouse in Korea's sovereign AI race with strong funding	South Korea's most widely used search-engine and internet conglomerate founded in 1999. Naver is now applying AI across search, cloud, and consumer products	Established game developer behind hits like Lineage and Guild Wars, whose AI arm focuses on in-game and industrial AI use cases	Founded in 1981, KT is one of the country's largest telecom players and primarily intends to integrate its AI strategy into existing products for enterprise clients	Seoul-based AI startup developing proprietary LLM and multimodal models. Motif emphasizes fully homegrown architecture
Company Type	Large conglomerate	Large conglomerate	Startup	Large conglomerate	Medium sized public company	Large conglomerate	Startup
Latest model	K-EXAONE, a 236B open weights reasoning model	A.X K1, a 500B open weights hyperscale reasoning model	Solar Open, a 100B open weights reasoning model	HyperCLOVA X SEED Think, a 32B open weights reasoning model	Varco-vision-2.0, a 1.7B image generation model	Mi:dm K 2.5 Pro, a proprietary reasoning model	Motif-2-12.7B, a small open weights model



D. OpenAI's first open weights language model since GPT-2 pushed the frontier for open weights models, but the open-to-proprietary gap remains steady

Leading Language Models by License Type, Over Time

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

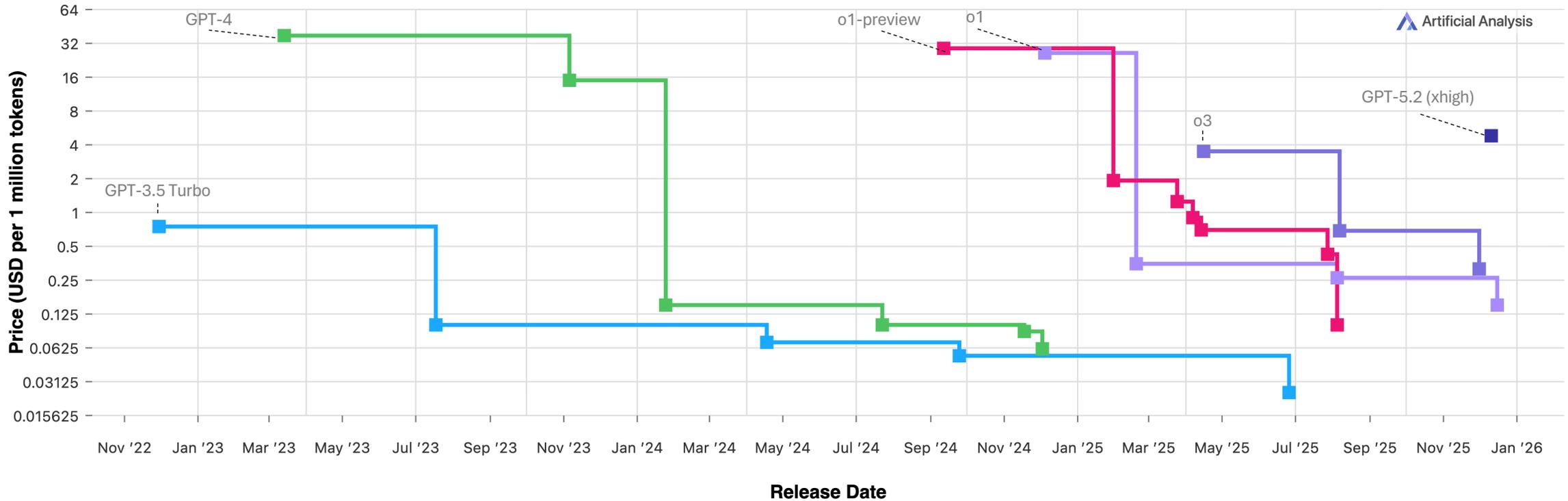


E. More efficient model architecture combined with software and hardware efficiencies helped drive down model costs - per token pricing fell 128x for o1-level intelligence

Language Model Inference Price by Intelligence Category, Over Time

Blended Input / Output Token Price (USD per M Tokens), Artificial Analysis Intelligence Index v4.0

■ Intelligence Index < 10
 ■ 10 <= Intelligence Index < 20
 ■ 20 <= Intelligence Index < 30
 ■ 30 <= Intelligence Index < 40
 ■ 40 <= Intelligence Index < 50
 ■ Intelligence Index >= 50



While frontier pricing has declined across successive intelligence categories (from GPT-4 to o1 to GPT-5.2), these reductions are gradual and stepwise, in contrast to the cost decline at equivalent intelligence levels



03

Image and Video

Artificial Analysis State of AI: 2025 Year-End Edition

Major improvements to both Image and Video came in 2025, including support for multi-modal inputs (image to video, image editing) and outputs (video with audio)

Key Themes

Text to Image Improves in Quality	<ul style="list-style-type: none"> • Text to image models have improved substantially in quality, with GPT Image 1.5 (leader at EOY 2025) ~150 ELO points higher than FLUX1.1 [pro] Ultra (leader at EOY 2024) • Progress in open weights image models has slowed as major labs such as OpenAI and Google have entered the space. The highest ranking open weights model at EOY was Qwen Image 2512, ranking #12 in the Text to Image Leaderboard
Image Editing Models Launched	<ul style="list-style-type: none"> • Instruction based image editing models gained popularity, with the launches of OpenAI's GPT-4o Image, and Google's Nano Banana (Gemini 2.5 Flash) driving a large increase in usage and mindshare • Multi-image input for image editing became common, with models such as Nano Banana Pro and Qwen Image Edit enabling more precise control of output images • Image Generation models became increasingly generalized, supporting both text to image and image editing e.g., the FLUX.2 family, and Seedream 4.5 support both text to image, and image editing modalities
Video models break into the mainstream	<ul style="list-style-type: none"> • Video models saw a breakthrough in quality, with Runway Gen-4.5 (leader at EOY 2025) ~200 ELO points higher than OpenAI's Sora (leader at EOY 2024) • Focus on Image to Video drove strong adoption, with users able to control video generations with more granularity and able to maintain character references across shots • Open weights video models lagged behind proprietary alternatives, with LTX-2 Pro representing the SOTA for open weights video generation, ranking 29th in Text to Video and 28th in Image to Video overall
Video with Audio starts with Veo 3	<ul style="list-style-type: none"> • Veo 3 released in May 2025 was the first high quality, mainstream model that natively supported audio generation as part of a video model, driving strong adoption • Video labs have quickly followed with their own Video with Audio models, such as OpenAI's Sora 2, Lightricks' LTX-2, Alibaba's Wan 2.6, and ByteDance's Seedance 1.5 pro
China maintains parity with US in media generation models	<ul style="list-style-type: none"> • Chinese and US labs continue to be at parity for image generation models with ByteDance's Seedream 4.5 competitive with Google's Nano Banana Pro, and OpenAI's GPT Image 1.5 • Chinese and US labs continue to be at parity for video generation models with Kling 2.5 Turbo competitive with Veo 3.1 and Runway Gen-4.5

Unlike in language models, smaller media generation focused AI labs have continued to compete with larger labs who have a wider breadth of modality coverage (1/2)

Key players offering image and/or video models (Labs with Broad Focus)

Includes publicly available models in each modality released in the last year by labs that develop both language and media generation models

No model Existing model

NON-EXHAUSTIVE

Modalities	OpenAI	Google	ByteDance	MiniMax	Alibaba	Meta	Tencent	Amazon	Baidu	xAI	StepFun	NVIDIA
A. Text to Image	Existing	Existing	Existing	Existing	Existing	Existing	Existing	Existing	No model	Existing	No model	Existing
B. Image Editing	Existing	Existing	Existing	No model	Existing	Existing	No model	No model	No model	No model	Existing	No model
C. Multi Image Editing	Existing	Existing	No model	No model	Existing	No model						
D. Text to Video	Existing	Existing	Existing	Existing	Existing	Existing	Existing	Existing	No model	Existing	Existing	No model
E. Image to Video	Existing	Existing	Existing	Existing	Existing	Existing	Existing	Existing	Existing	Existing	No model	No model
F. Multi Image to Video	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model
G. Video with Audio Output	Existing	Existing	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model
H. Video with Audio Input	No model	No model	No model	No model	Existing	No model						
I. Video Editing	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model

Continues on next page for media generation focused labs

Unlike in language models, smaller media generation focused AI labs have continued to compete with larger labs who have a wider breadth of modality coverage (2/2)

Key players offering image and/or video models (Labs with Media Generation Focus)

Includes publicly available models in each modality released in the last year by labs that develop only media generation models

No model
 Existing model

NON-EXHAUSTIVE

Modalities	Kuaishou	Runway	Adobe	Black Forest Labs	Luma Labs	Pixverse	Vidu	Pika Art	Lightricks	Decart	Ideogram	Midjourney	Leonardo.ai	Recraft	HiDream	Reve	Stability.ai	MoonValley	Playground	Genmo
A. Text to Image	Existing	Existing	Existing	Existing	Existing	No model	Existing	No model	No model	No model	Existing	Existing	Existing	Existing	Existing	Existing	Existing	No model	Existing	No model
B. Image Editing	No model	No model	No model	Existing	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	Existing	Existing	No model	No model	No model	No model
C. Multi Image Editing	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model
D. Text to Video	Existing	Existing	Existing	No model	Existing	Existing	Existing	Existing	No model	No model	Existing	Existing	Existing	Existing	Existing	No model	No model	Existing	No model	Existing
E. Image to Video	Existing	Existing	Existing	No model	Existing	Existing	Existing	Existing	Existing	Existing	No model	Existing	Existing	Existing	Existing	No model	No model	Existing	No model	No model
F. Multi Image to Video	Existing	Existing	No model	No model	No model	No model	Existing	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model
G. Video with Audio Output	Existing	No model	No model	No model	No model	No model	Existing	No model	Existing	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model
H. Video with Audio Input	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model
I. Video Editing	No model	Existing	No model	No model	No model	No model	No model	No model	No model	Existing	No model	No model	No model	No model	No model	No model	No model	No model	No model	No model



04

Speech and Music

Artificial Analysis State of AI: 2025 Year-End Edition

Speech and Music: Speech and music AI continued to advance in Q4, with especially notable improvements in S2S reasoning and open weights STT accuracy

Key Themes

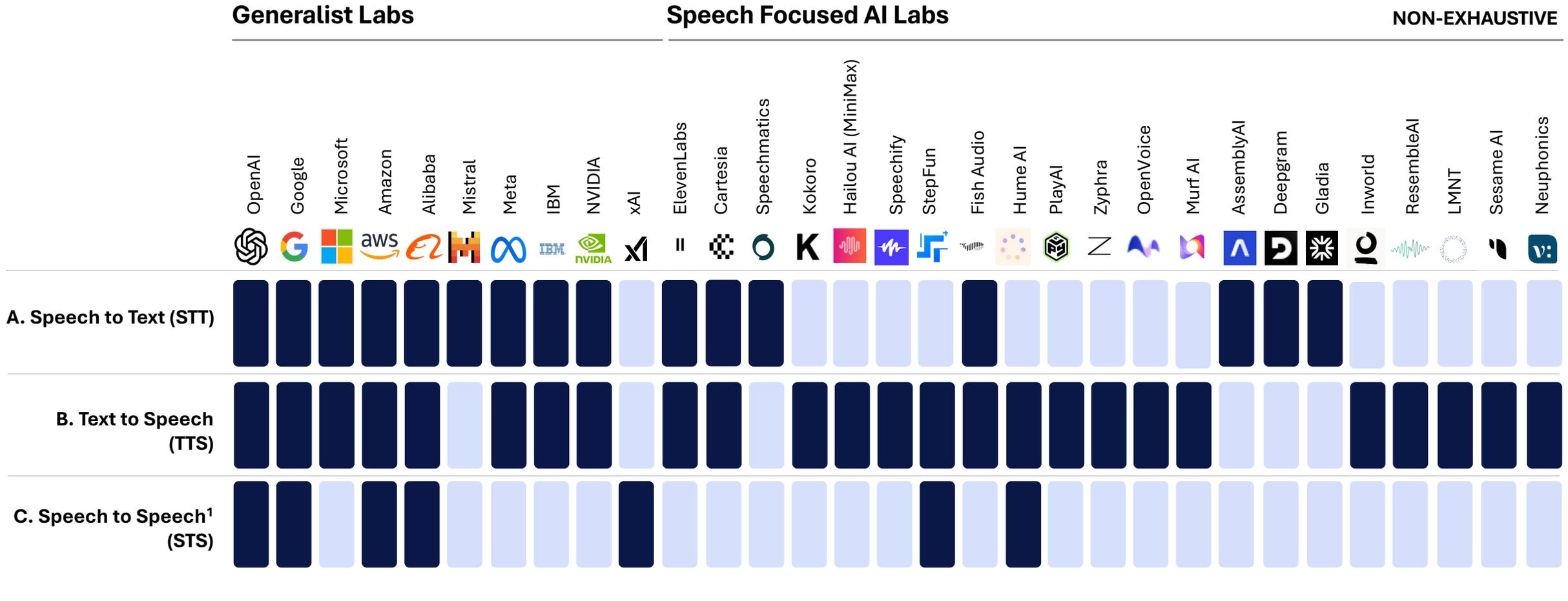
-
- | | |
|---|---|
| Speech to Text (STT) continues to see improvements in word error rates | <ul style="list-style-type: none"> • Multimodal models expand into transcription as a secondary capability, with AWS' Nova 2 Omni delivering competitive accuracy without STT specialization, enabling unified speech, vision, and text processing • Ultra-low latency, real-time variants emerge for voice agent applications, such as ElevenLabs' Scribe v2 Realtime and NVIDIA's Parakeet Realtime |
|---|---|
-
- | | |
|---|--|
| Text to Speech (TTS) models deliver further control of prosody and audio effects | <ul style="list-style-type: none"> • There has been significant improvements in Text to Speech quality, with new models released continually pushing the frontier • Prosody control becoming increasingly prevalent across leading models, supporting emotional tone, pacing, emphasis, and paralinguistic elements (laughter, sighs, breathing) via approaches such as in-text tagging and SSML markup • Voice cloning becoming more popular, including celebrity voice synthesis, driving increased focus on audio authenticity through watermarking technologies and provenance verification systems |
|---|--|
-
- | | |
|---|--|
| STS models mature rapidly as native audio reasoning improves | <ul style="list-style-type: none"> • xAI is now overall leader on Big Bench Audio benchmark while delivering fast speeds, displacing previous leader Google Gemini 2.5 Native Audio Thinking in reasoning – meanwhile Nova 2.0 Sonic emerges as the price-performance leader • While speech pipelines continue to make up vast majority of voice agents, improving native audio reasoning capabilities validate end-to-end audio processing, eliminating LLM intermediaries and enabling models to reason directly with acoustic information for improved context understanding and lower latency |
|---|--|
-
- | | |
|---------------------|---|
| Voice Agents | <ul style="list-style-type: none"> • Performance reaching near-human quality in structured interactions, but significant gaps persist in ambiguous contexts, complex multi-turn reasoning, and degraded audio conditions requiring continued improvement |
|---------------------|---|
-
- | | |
|--------------|---|
| Music | <ul style="list-style-type: none"> • Despite a number of strong releases in 2025, Q4 was relatively quiet for leading releases – nevertheless marketing and adoption grew for leading models such as Suno V4.5, ElevenLabs Music and Producer.ai's Fuzz series |
|--------------|---|
-
- 30
-  Artificial Analysis

Speech: While the generalist AI labs (e.g. OpenAI, Google) have offerings across all speech modalities, pure-play speech labs are more focused, particularly in TTS

Key players offering speech models

Classifications are indicative and determined based on models available

Low or no presence Strong presence



1. Only includes native Speech to Speech models (not pipeline / cascading STT → LLM → TTS approach)



05

Accelerators

Artificial Analysis State of AI: 2025 Year-End Edition

AI infrastructure significantly matured in 2025 - Blackwell systems started shipping, inference software became more mature and challengers continued to evolve

Key themes

Blackwell systems entered production, delivering substantial performance gains over Hopper systems

- B200s became widely available for production workloads in 2025 and GB200 NVL72 rack-scale systems reached full production. IBM's Granite 4 series of models were amongst the first models publicly announced as being trained on GB200 NVL72 clusters, with OpenAI's GPT-5.3 Codex being the first frontier model explicitly disclosed to have been trained on GB200
- NVIDIA announced B300/GB300 in Q3 2025 with shipments beginning later - B300 offers 288GB HBM3e (+50% over B200) and 14 PFLOPs FP4 (versus 9 PFLOPs for B200)
- Software support matured (especially TensorRT-LLM), and Blackwell-series chips now outperform Hopper chips and other AI accelerators across the full Pareto frontier of inference performance

Inference software consolidated around three open source frameworks

- Inference software matured significantly in 2025, consolidating around three frameworks – vLLM, SGLang and NVIDIA TensorRT-LLM

NVIDIA continues to retain dominant market share, but challengers made strategic advances

- NVIDIA acquired Groq for ~\$20 billion (December 2025) and the transaction is structured as IP licensing plus acqui-hire to integrate Groq's LPU technology as an NVIDIA offering
- Google's TPU v6 (Trillium) reached GA in late 2024; TPUs powered training of Gemini 2.5 Pro and Gemini 3 Pro
- Anthropic signed deals with Google and Amazon in 2025 for access to TPU and Trainium for training and inference, while Cerebras joined NVIDIA, AMD and Broadcom to sign a multi-year contract with OpenAI for fast inference

Growing inference demand and evolving workload patterns are driving distributed and disaggregated architectures into 2026

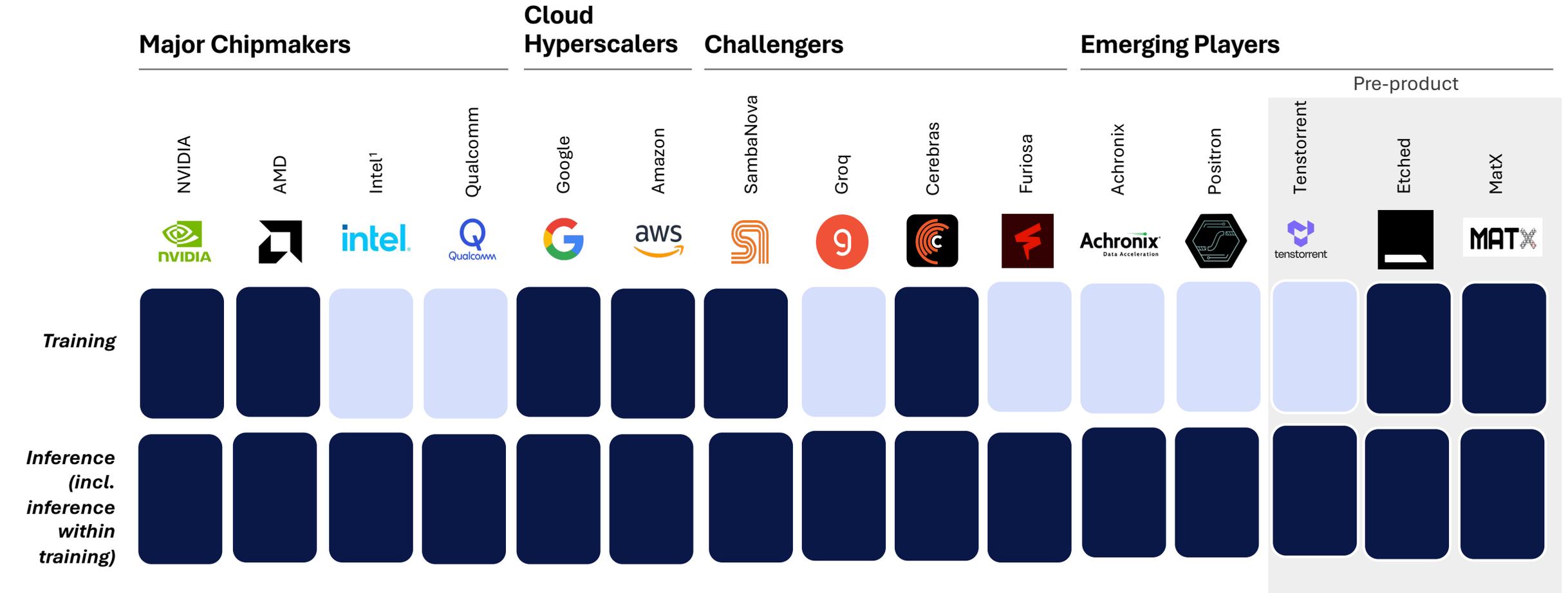
- Distributed inference optimizations previously confined to frontier labs are becoming widely accessible, driven by NVIDIA Dynamo maturation and open-source projects
- Key techniques: prefill/decode disaggregation, expert parallelism across dozens to hundreds of GPUs, and novel load balancing via scaled expert replicas

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



1. Intel is no longer bringing to market Falcon Shores, its successor to Gaudi 3, and Intel's timeline for Jaguar Shores or other future accelerator products is unclear. Rivos was acquired by Meta



Artificial Analysis

contact@artificialanalysis.ai

<https://artificialanalysis.ai/>

Legal notice:

Copyright © 2026 Artificial Analysis, Inc. All rights reserved.

This document, including any data, analysis, and insights contained herein, is provided by Artificial Analysis for informational purposes only. The information is based on data collected through various sources, including but not limited to first party benchmarking and surveys conducted on our website. While Artificial Analysis strives to ensure the accuracy and reliability of the information, it is provided “as is” and may not be complete or up to date. The content should not be construed as professional advice, and recipients are encouraged to conduct their own research and analysis before making any decisions based on this information. By accessing or using this document, you agree to be bound by Artificial Analysis’s Terms of Service, available on our website.