Artificial Analysis State of AI

Q3 2025 – Highlights Edition

Artificial Analysis is the leading 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 platform, **artificialanalysis.ai**, is trusted by companies driving the frontier of AI as well as leading corporations, media, research and government institutions. 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
Full Version of 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

NVIDIA, Meta, cerebras, TechCrunch, Google, xAI, amazon, FT, The Economist, CNBC, OpenAI, AMD, WSJ, VentureBeat

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 Q3 2025

The models are smarter, they're using more tools and uptake is faster than ever.

Innovation has continued at pace across the entire AI stack. We make no comment on whether valuations have entered bubble territory, but we can confirm that any suggestion of progress stalling has been greatly exaggerated.

The level of competition across AI landscape has continued to only increase, with no signs of consolidation or clear winners. xAI joined OpenAI, Google and Anthropic at the top of our Intelligence Index, and over a dozen others (mostly from China) are mere months behind. Innovation across the stack, from hardware to models to products, is enabling agentic experiences that are getting more done than ever before.

Produced by Artificial Analysis, the leading independent AI benchmarking and insights provider, this Q3 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

5 major trends shaped AI in Q3 2025

Agentic capabilities become the focal point for AI labs

AI labs are increasingly focusing on agentic capabilities, including longer horizon tool use and agentic multi-step workloads

Open weights models released at their fastest rate yet

OpenAI released their first open weights model since GPT-2 – competing with dozens leading open weights models from Chinese labs

Native Speech to Speech models reach viability for production use

Speech capabilities are maturing across transcription, generation, and native Speech to Speech, enabling more production-ready voice agents

Competition intensifies across all modalities

Major labs continue advancing intelligence, efficiency, and speed while the number of competing labs is growing across all modalities

Image Editing and Video Generation go mainstream

Image editing saw significant improvements in quality and popularity with the release of Gemini 2.5 Flash (Nano Banana)

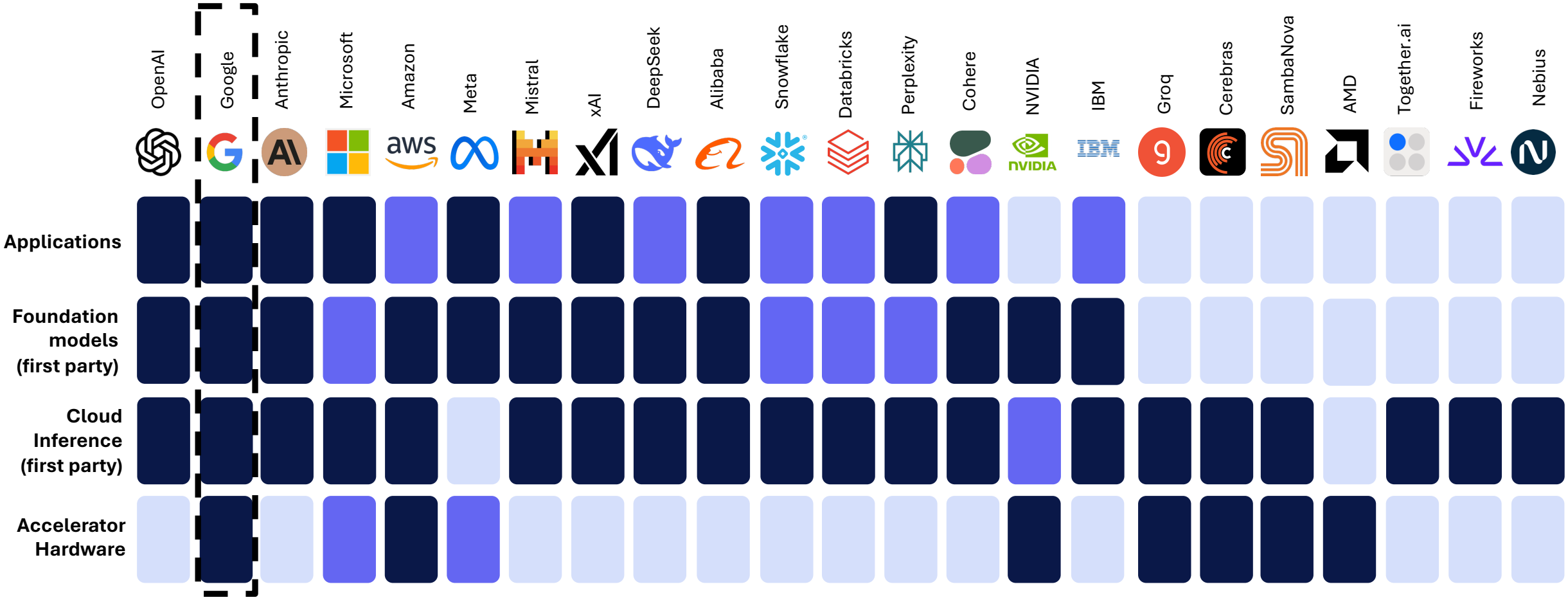
Players in the AI value chain differ in levels of vertical integration; Google continues to stand out as the most vertically integrated from TPU accelerators to Gemini application

NON-EXHAUSTIVE

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

No presence — Strong presence



Source: Company website

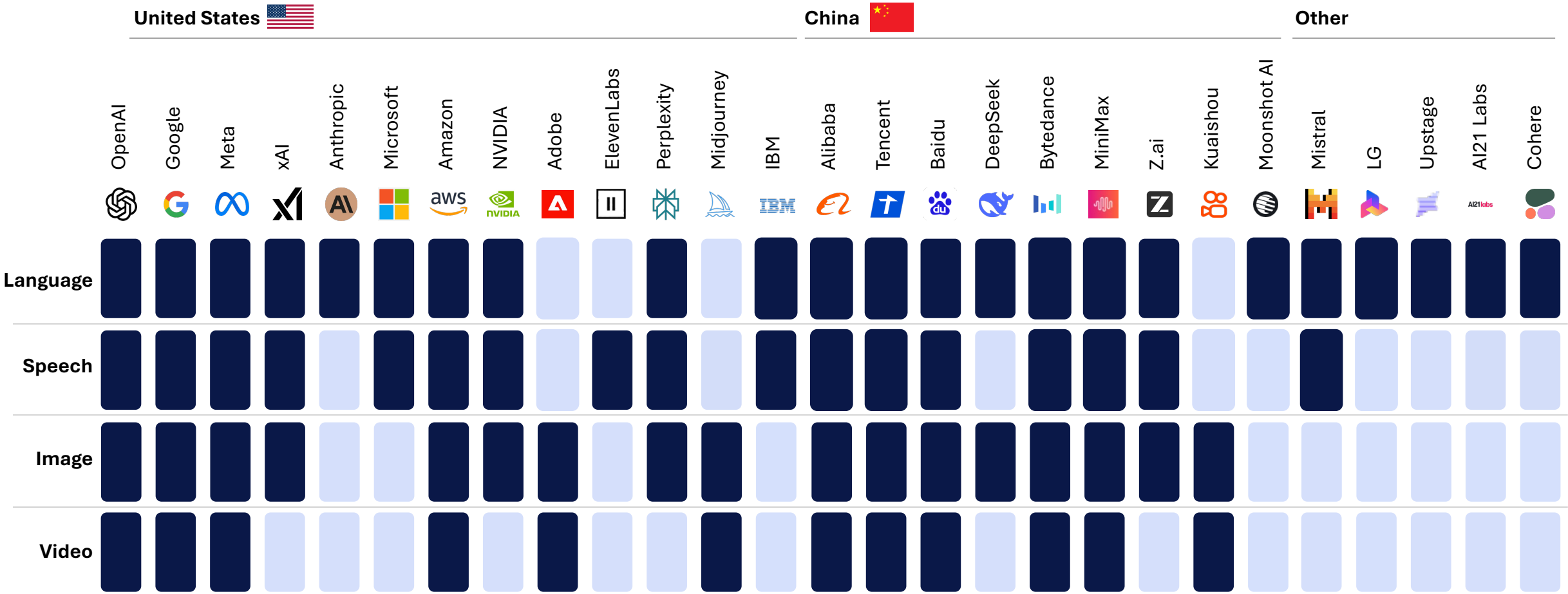
Big technology companies in the US and China are continuing to play across all AI modalities while smaller challengers tend to focus on specific modalities

Key players with first-party models by modality

NON-EXHAUSTIVE

No model

Existing model

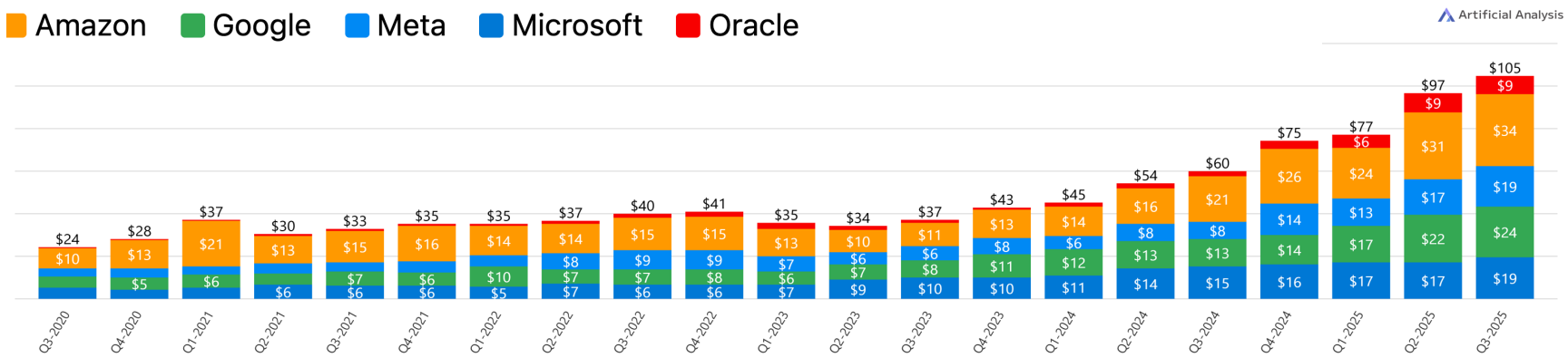


Source: Company website

Investment in AI infrastructure continues to drive large tech company capital expenditure materially higher into the second half of 2025 and likely to continue into 2026

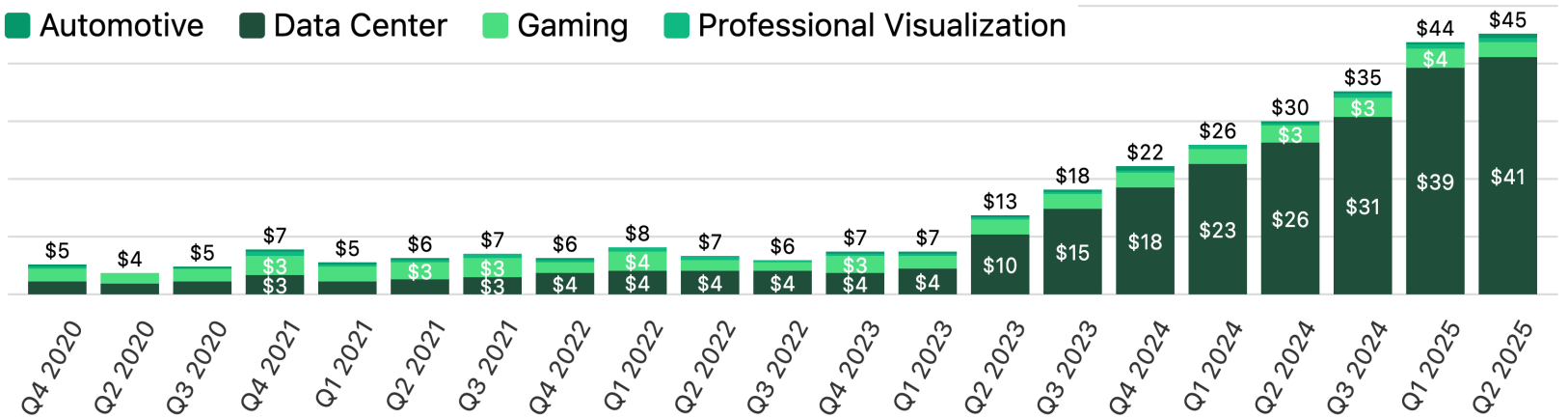
Total Capital Expenditure of US Public Big Tech Firms (Q4 '20 – Q3 '25)

USD Billions



NVIDIA Revenue Breakdown, Over Time

USD Billions



NVIDIA will report Q2 FY26 (Q3 CY25) in late Nov

Commentary

- **Public tech companies maintain aggressive AI investment:** AI infrastructure investment continues to drive big tech capex higher, mostly to support growth in frontier lab workloads
- **US big tech is likely driving the majority of AI infrastructure build-out, but private spend that is not tracked on the chart to the left is growing:** xAI is looking to purchase 300,000 Nvidia GPUs for its Colossus 2 data center, Neocloud capex is increasing quickly
- **Caution is required when interpreting forward-looking claims:** AI compute deals are being announced weekly, but totals frequently include overlapping numbers and spend over many years

Source: Securities and Exchange Commission, Press Search. Calendar year quarters.

While efficiency gains have been made...

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

A. Smaller Models and Sparsity

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***

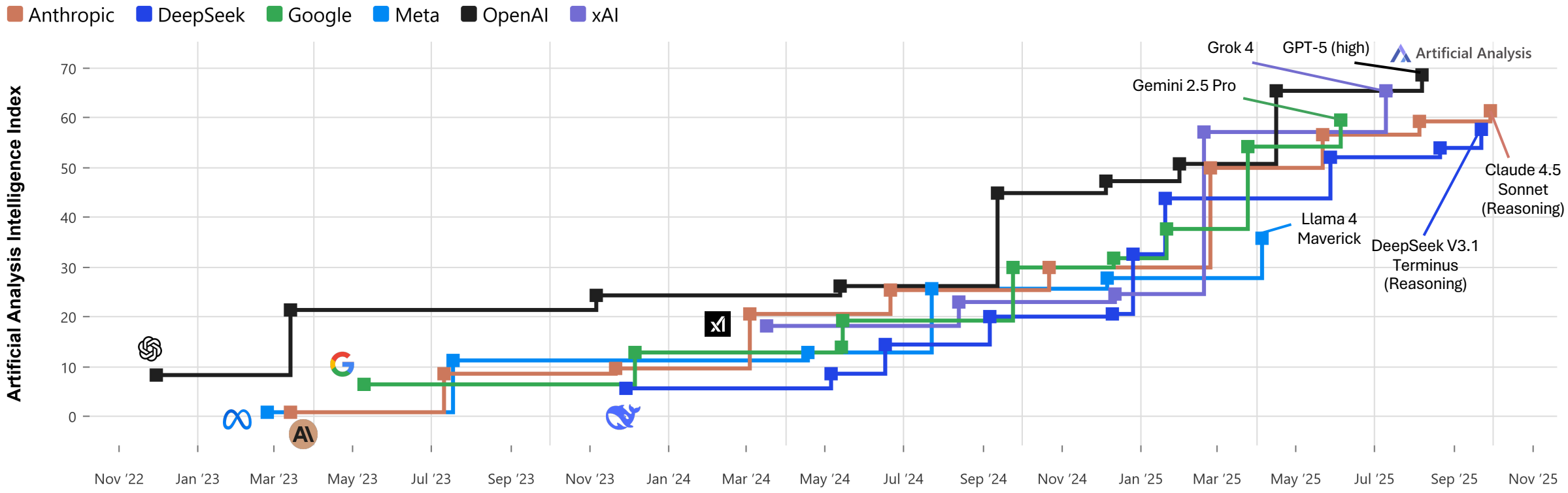
F. AI Agents

Agents chain multiple requests to LLMs to complete tasks autonomously across long conversations

OpenAI has reclaimed the top spot for the most intelligent model with the release of GPT-5, however with frequent model releases by all frontier labs, the race is closer than ever

Frontier Large Language Model (LLM) Intelligence, Over Time

Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, τ^2-Bench Telecom



- **OpenAI regains the lead:** GPT-5 (high) is the most intelligent model scoring 68 on the Artificial Analysis Intelligence Index, and places ahead of Grok 4 (65), Claude 4.5 Sonnet (Thinking, 63) and Gemini 2.5 Pro (60)
- **Competition between labs for frontier intelligence continues to increase:** The intelligence frontier is now fiercely contested by between OpenAI, xAI, Anthropic, and Google. Meta has restructured their AI efforts and has not released a new model since April.

Source: Artificial Analysis independent benchmarking

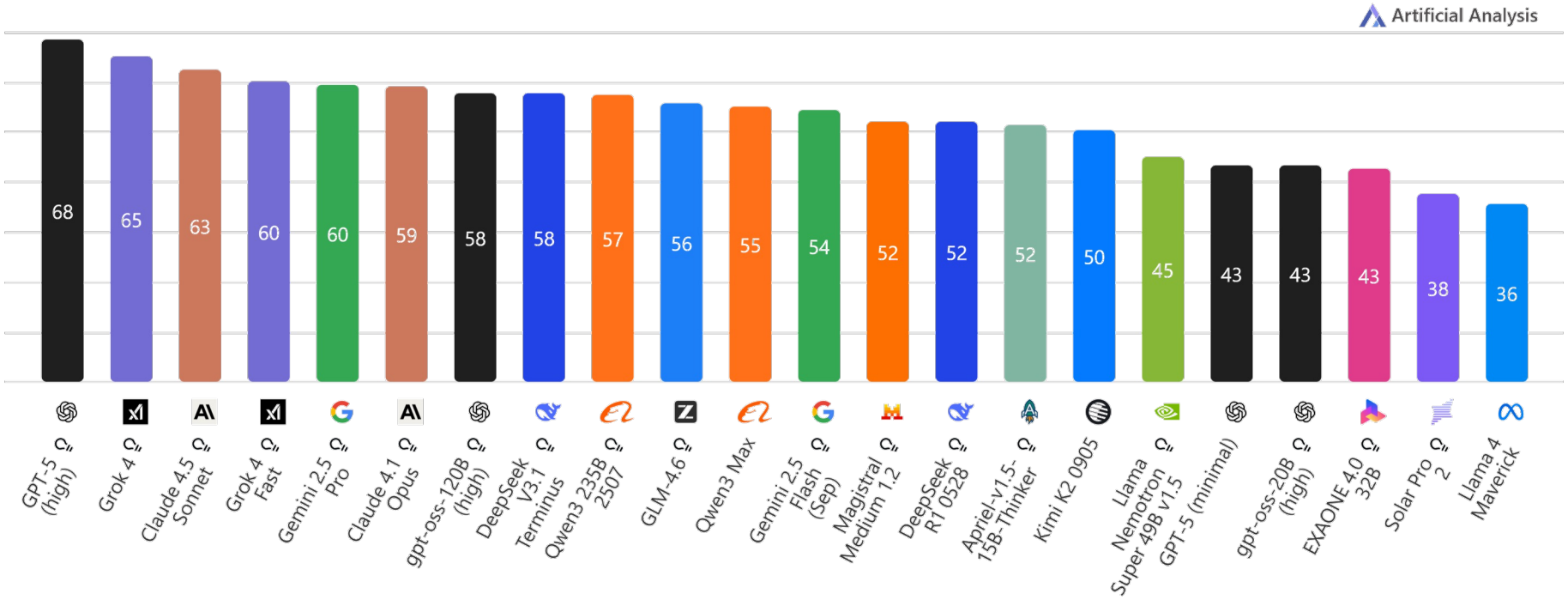
OpenAI, xAI and Anthropic lead frontier intelligence with their latest reasoning models, followed closely by other labs

Leading Large Language Models (LLMs), by AI lab

Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, τ^2-Bench Telecom

Reasoning Model

NON-EXHAUSTIVE



Commentary

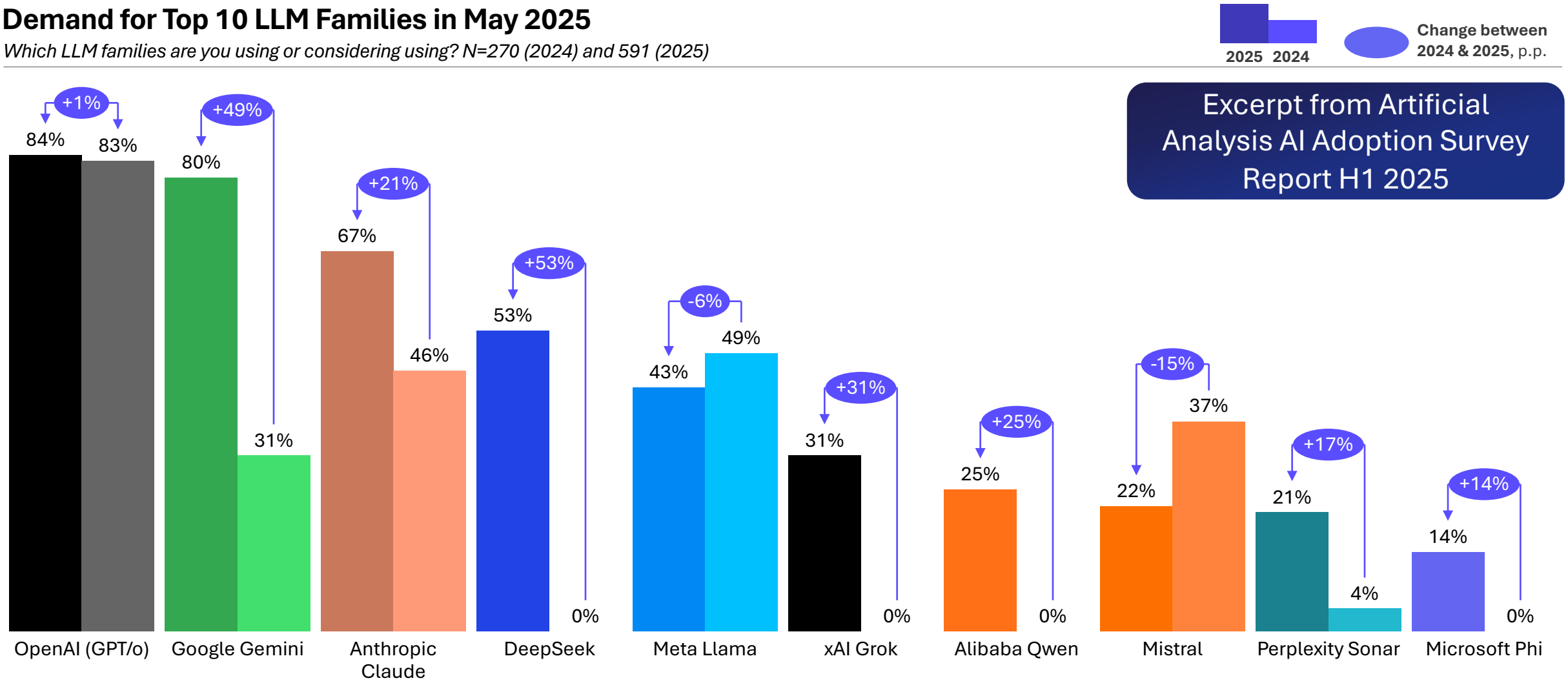
- **OpenAI regains frontier leadership:** GPT-5 (High) takes the lead with an intelligence index of 68, moving ahead of xAI's Grok 4 and reclaiming the lead that OpenAI briefly lost last quarter
- **U.S. labs dominate the frontier**: OpenAI, xAI, Anthropic, and Google models hold the top 7 positions, underscoring continued U.S. leadership in frontier models
- **Global peers keep pace**: Beyond the top U.S. labs, players like Alibaba (Qwen3), DeepSeek, Mistral, Z.ai are keeping up to date with advances in new models

Source: Artificial Analysis independent benchmarking

Model families: 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)

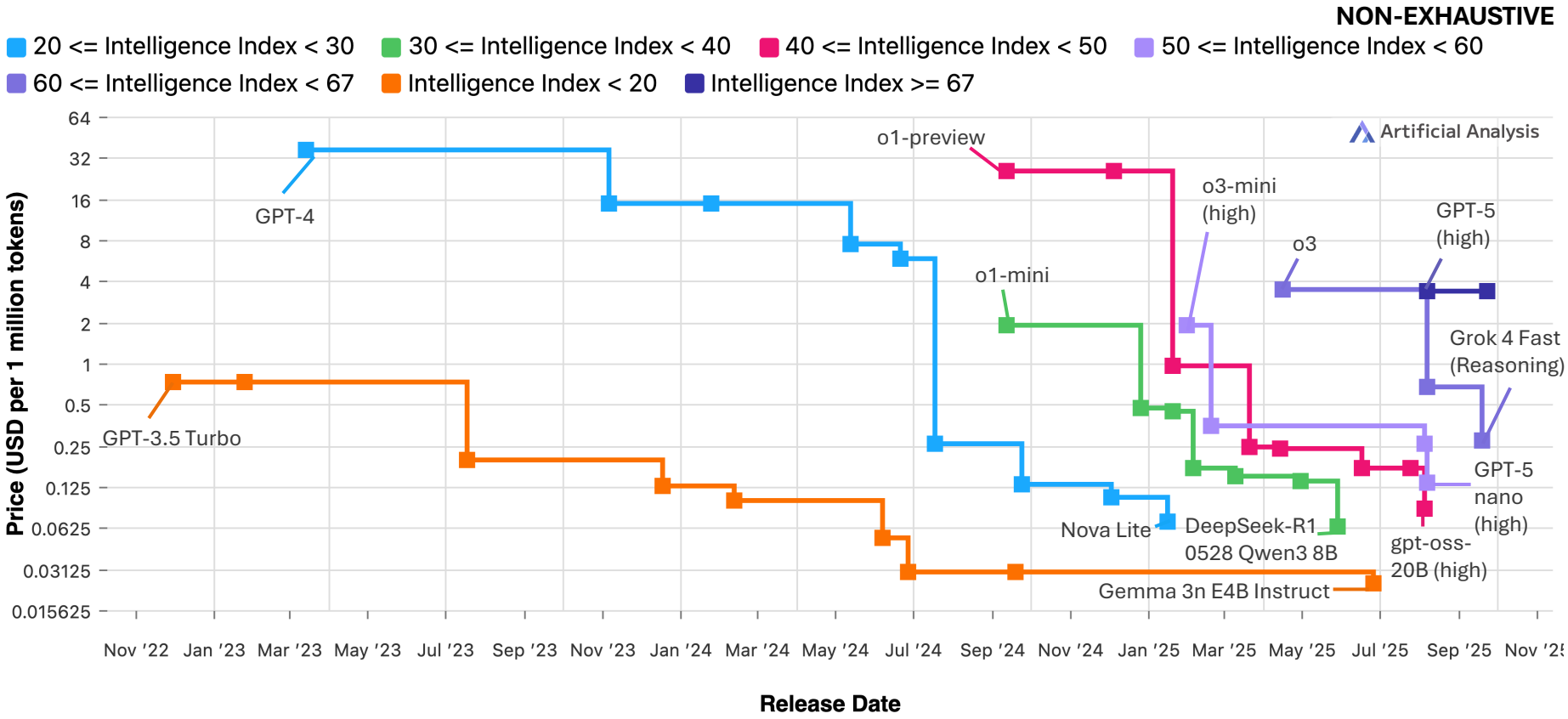


Source: Artificial Analysis AI Adoption Survey – H1 2025

Pricing: Inference pricing continues to fall across all intelligence classes

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 v3 (includes 10 evaluations)



Commentary

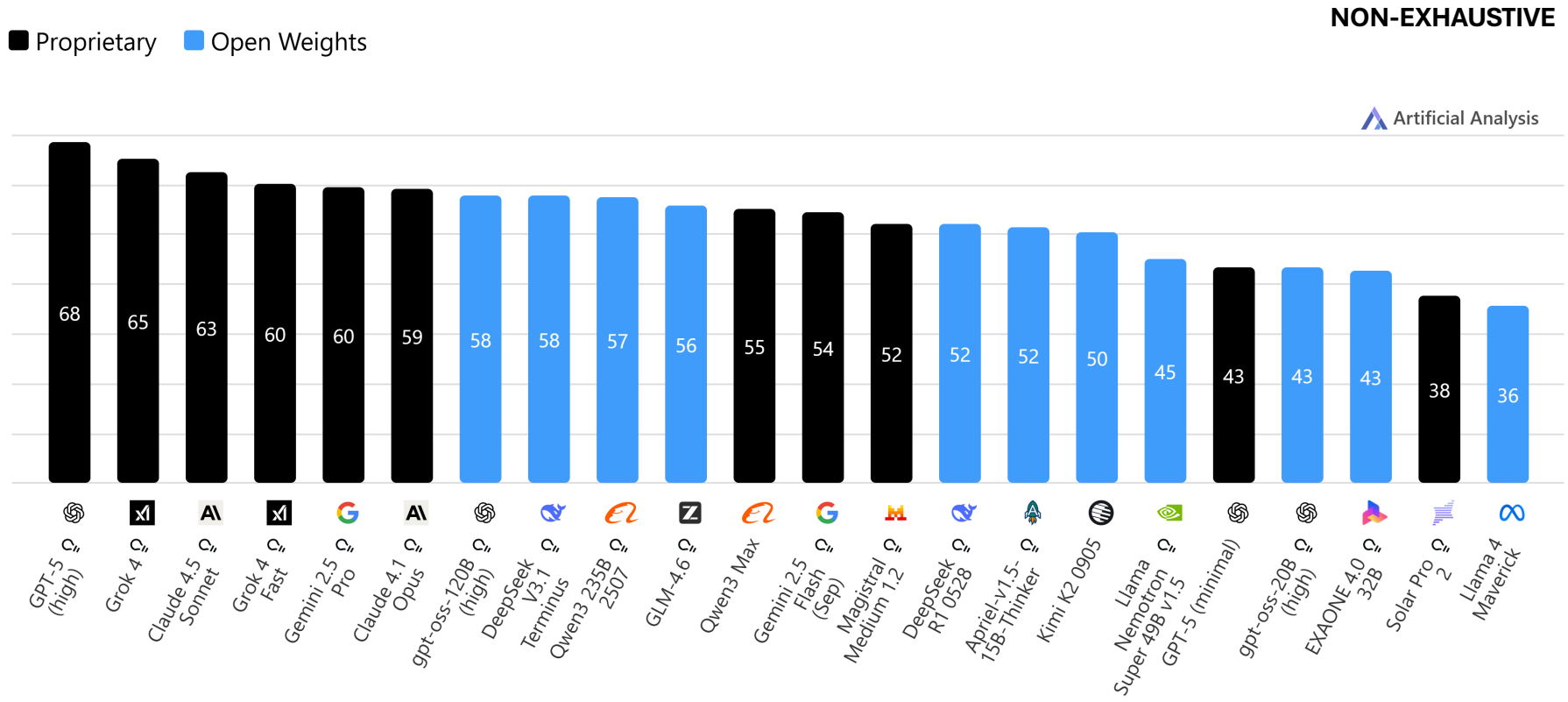
- **Q3 2025 saw steep price declines among frontier models** as the release of Grok 4 Fast, GPT-5 nano, and gpt-oss-20B drove inference prices down by around 50% or more across models scoring 40+ on the Artificial Analysis Intelligence Index
- **While pricing declines within each intelligence class,** continuous intelligence gains unlock new intelligence classes, resulting in smaller price changes at the highest intelligence levels

Source: Artificial Analysis independent benchmarking

Open weights: Leading models are proprietary, however there are select open weights models near the intelligence frontier

Leading Language Models by License Type

Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, τ^2-Bench Telecom



Commentary

- **Proprietary models lead the frontier:** GPT-5 (High), Grok 4, and Claude 4.5 anchor the top of the index, reinforcing U.S. dominance in proprietary reasoning systems
- **OpenAI leads open weights frontier**: OpenAI's gpt-oss-120B pushes U.S. labs back into the frontier of open weights models

Source: Artificial Analysis independent benchmarking

Reasoning models: Treating reasoning & non-reasoning models as distinct categories remains a helpful framework for understanding today's landscape, but lines are blurring

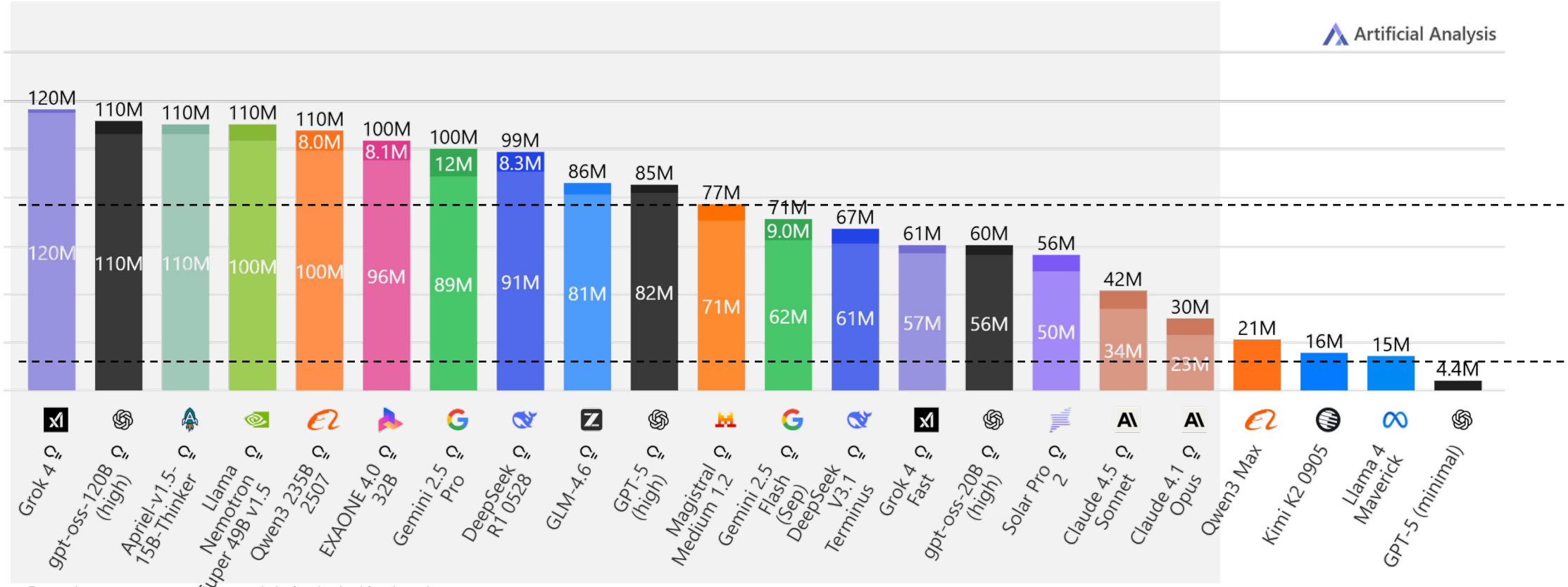
Output Tokens Used to Run the Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v3 (incorporates 10 leading evaluations), Output Tokens Used in Artificial Analysis Intelligence Index (~5M input tokens)

Reasoning models

Answer Tokens | Reasoning Tokens

NON-EXHAUSTIVE



~77M[1]

Avg. total output tokens for reasoning models

~14M[1]

Avg. total output tokens for non-reasoning models

1. Based on representative models included in the chart

Source: Artificial Analysis independent benchmarking

Agents: Agents are autonomous systems driven by language models

What are agents?

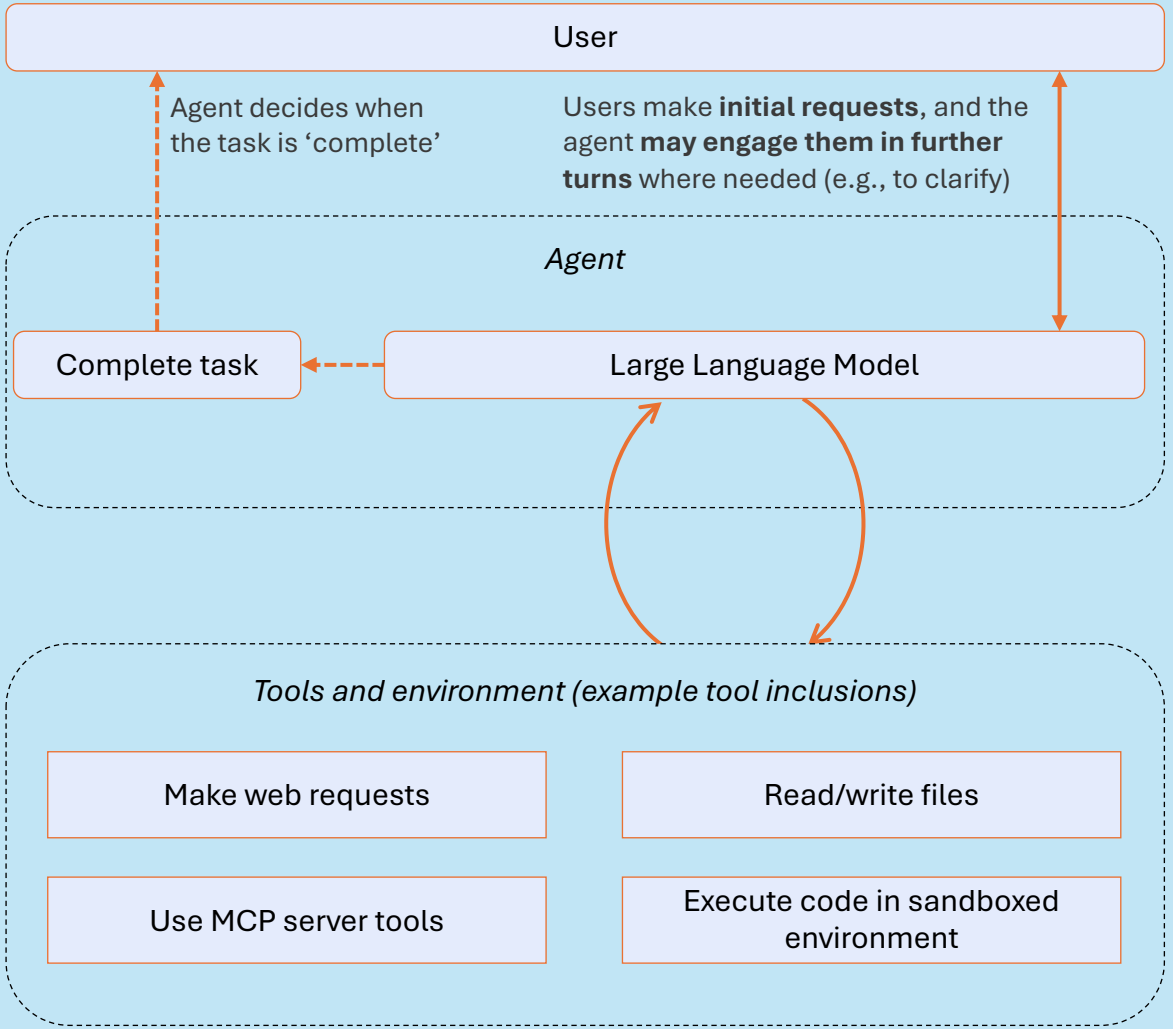
"Systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks"

"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"

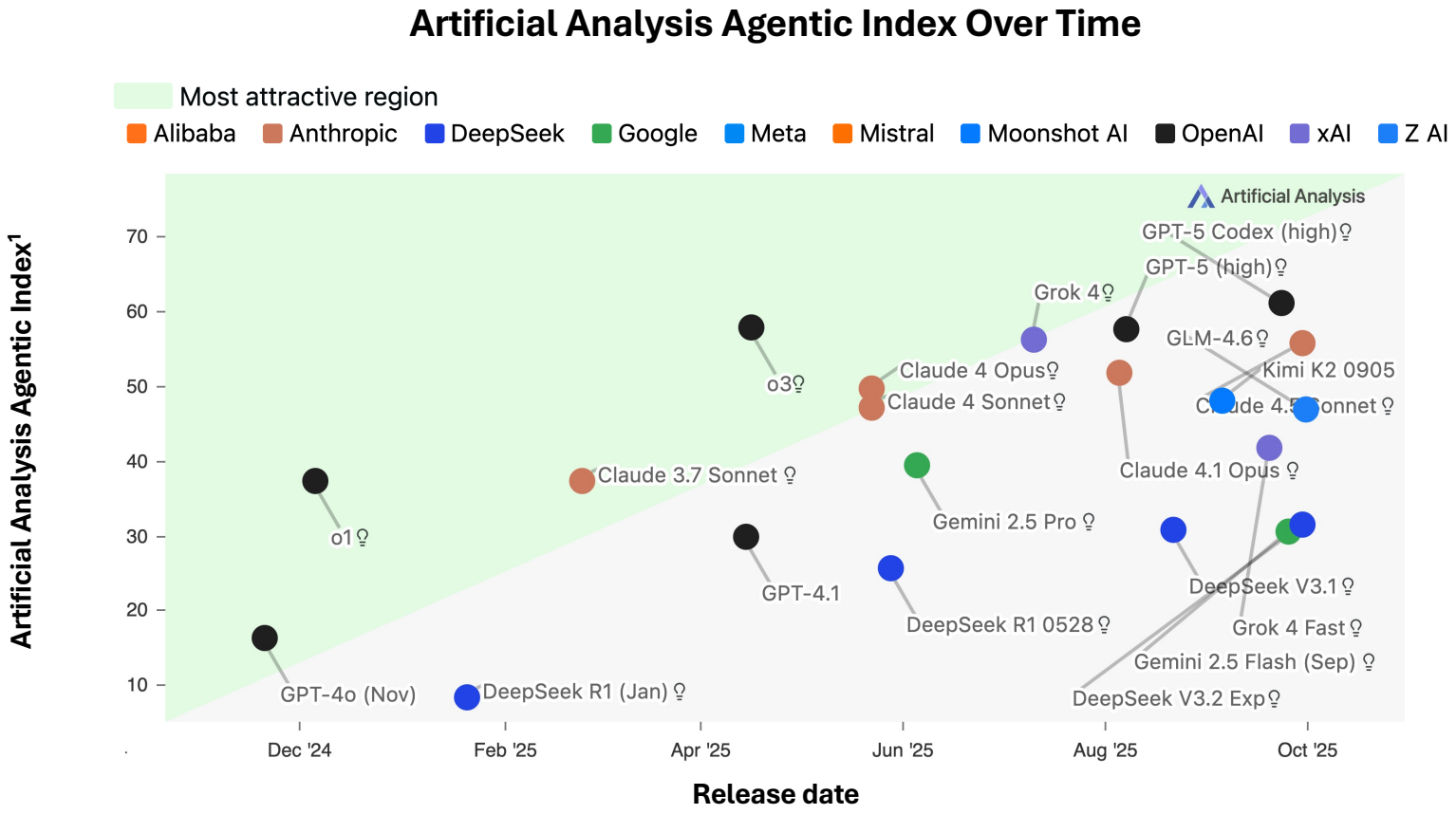
AI agents are LLM-driven systems that act autonomously and use tools to complete tasks end-to-end

Fundamentally, agents are **LLM-driven systems** that **use tools** to **complete tasks end-to-end in an environment**



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

Agents: Proprietary and open weights models released in Q3 2025 have been pre-trained and reinforcement learning-optimized for tool use and agentic task execution



Recent model release quotes

"GPT-5 can follow your instructions more faithfully and get more of the work done end-to-end using the tools ..."

"Grok 4 Fast was trained end-to-end with tool-use reinforcement learning (RL). ...."

"Smarter tool calling: Through post-training optimization, [DeepSeek V3.1 Terminus'] performance in tool usage and agent tasks has significantly improved"

1. *Average of Terminal-Bench Hard and τ^2-Bench Telecom*

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

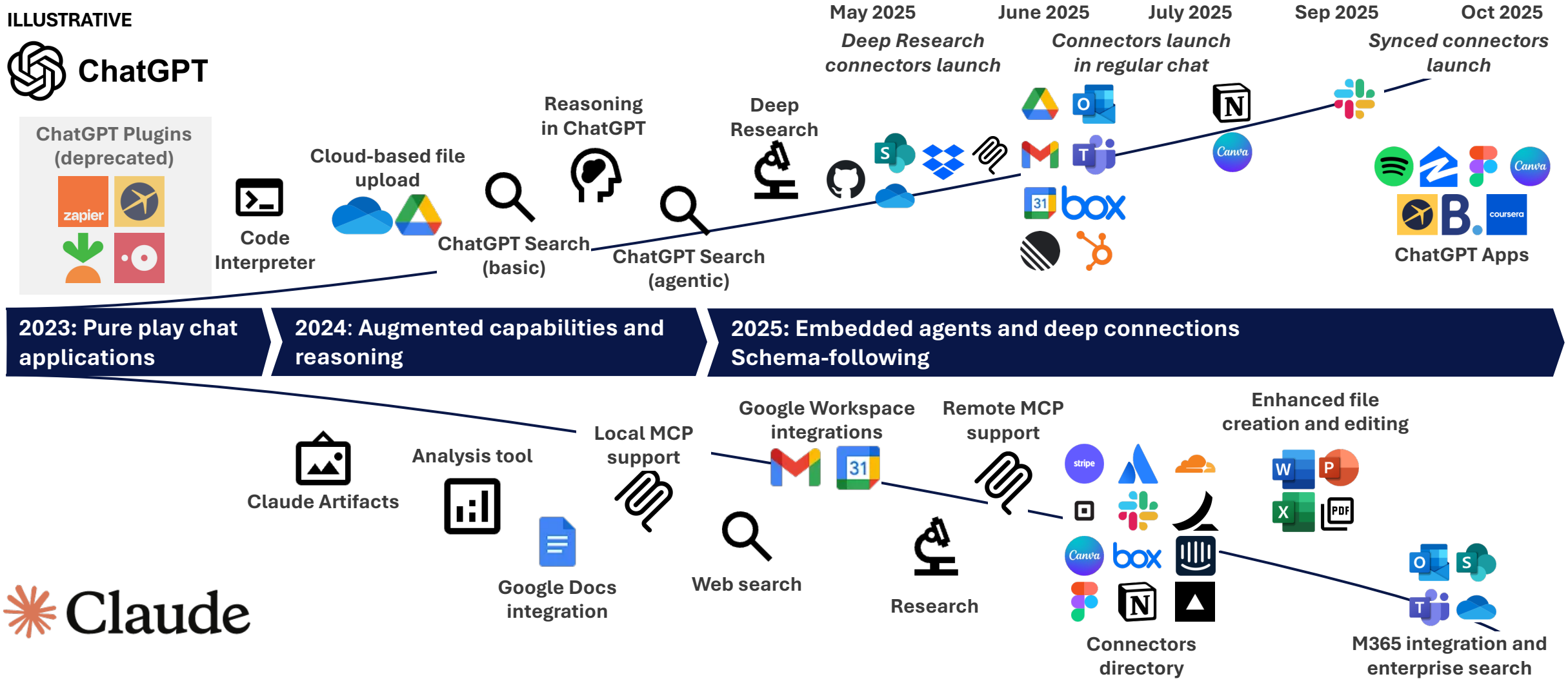
NON-EXHAUSTIVE

Frontier lab product

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

Agents: Chat applications are steadily expanding integrations and tool availability to enable multi-step workflows and broader task execution

ILLUSTRATIVE



03 Image and Video Models

State of AI – Q3 2025

Q3 ‘25 saw a continued shift in progress to video models, with audio support and quality improvements, while open weights model progress slowed in both image and video

Key Themes in Q3 ‘25

Video models gain audio support	• **OpenAI’s Sora 2 joins Google’s Veo 3 as a high quality, mainstream model that natively supports audio generation,** driving increased adoption and hype • **Audio inputs begin to be supported for video generation**, with Wan 2.5 Preview accepting input audio clips • **Video models with audio support are more expensive,** with Sora 2 at $0.5/s of 1080p video with audio, and Veo 3 at $0.40/s, surpassing comparable non-audio models such as Hailuo 2 Pro at ~$0.08/s of 1080p video.
Video models see rapid progress in quality	• **Video models see rapid progress in quality,** with short time at the top of the leaderboard: Runway Gen 3, the leader in Q1 2025 for Image to Video, is now ranked 23rd in the leaderboard • **Open weights video models lag behind proprietary alternatives,** with Alibaba Wan 2.2 A14B representing the SOTA for open weights video generation, ranking 11th in Text to Video and 20th in Image to Video overall
Image editing models gain popularity	• **Instruction based image editing models gain popularity**, with Gemini 2.5 Flash (Nano Banana)’s launch driving strong adoption of Google’s iOS Gemini app, and GPT Image 1 maintaining popularity in ChatGPT • **Multi-image input for image editing continues to become standard**, with models such as Gemini 2.5 Flash and Qwen Image Edit 2509 enabling more precise control of output images • **Image Generation models become increasingly generalized, supporting both text to image and image editing** e.g., Seedream 4.0, and FLUX.1 Kontext [max] support both text to image, and image editing modalities • **Open weights image models keep pace in Image Editing,** with Qwen Image Edit 2509, the leading open weights image editing model, coming in 3rd in the Artificial Analysis Image Editing Leaderboard
China maintains parity with US in images, and leads in video generation	• **Chinese and US labs continue to be at parity for image generation models** with Bytedance’s Seedream 4.0 leading text to image and Google’s Gemini 2.5 Flash leading in image editing • **Chinese labs lead in video generation** with Kling 2.5 Turbo leading both text and image to video leaderboards; Google Veo 3 and Luma Labs Ray 3 are the only Western models in the image to video top 10
Text to image models improve incrementally	• **Text to image models have improved incrementally,** with Seedream 4.0 improving by ~30 Elo on Imagen 4 Ultra • **Progress in open weights text to image models has slowed**, with HunyuanImage 2.1 the only text to image model to surpass HiDream-I1-Dev (the open weights leader in Q1)

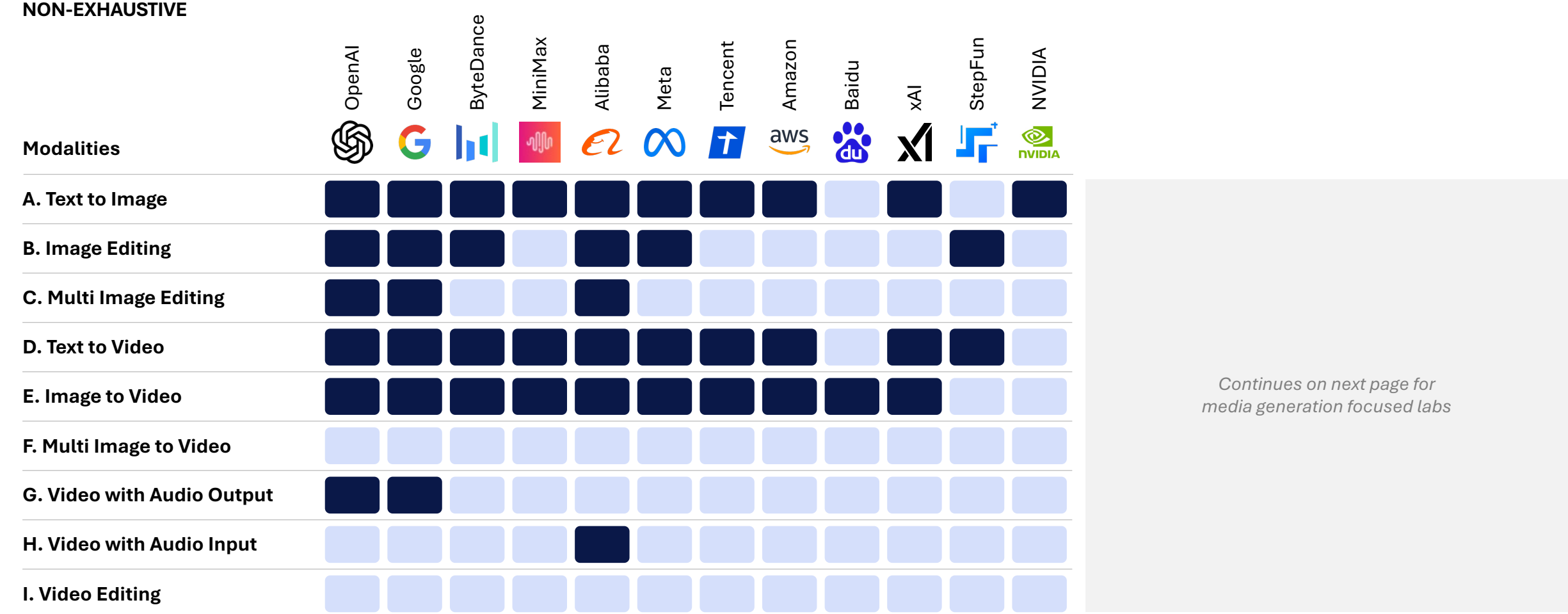
Unlike in language models, smaller organizations focused on media generation continue to compete with larger organizations with a broad focus across various modalities (1/2)

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



Continues on next page for media generation focused labs

Unlike in language models, smaller organizations focused on media generation continue to compete with larger organizations with a broad focus across various modalities (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	■	■	■	■	■		■				■	■	■	■	■	■	■		■	
B. Image Editing				■											■	■				
C. Multi Image Editing																				
D. Text to Video	■	■	■		■	■	■	■				■	■		■			■		■
E. Image to Video	■	■	■		■	■	■	■	■	■		■	■		■			■		
F. Multi Image to Video	■	■					■													
G. Video with Audio Output																				
H. Video with Audio Input																				
I. Video Editing		■								■										

Deep Dive: Text to Image and Image Editing leaderboards continue to be dominated by proprietary models, but open weights models are improving in parallel

Leading Text to Image and Image Editing Models by License Type, Over Time

Text to Image

Elo Score of proprietary and open weights text to image models over time

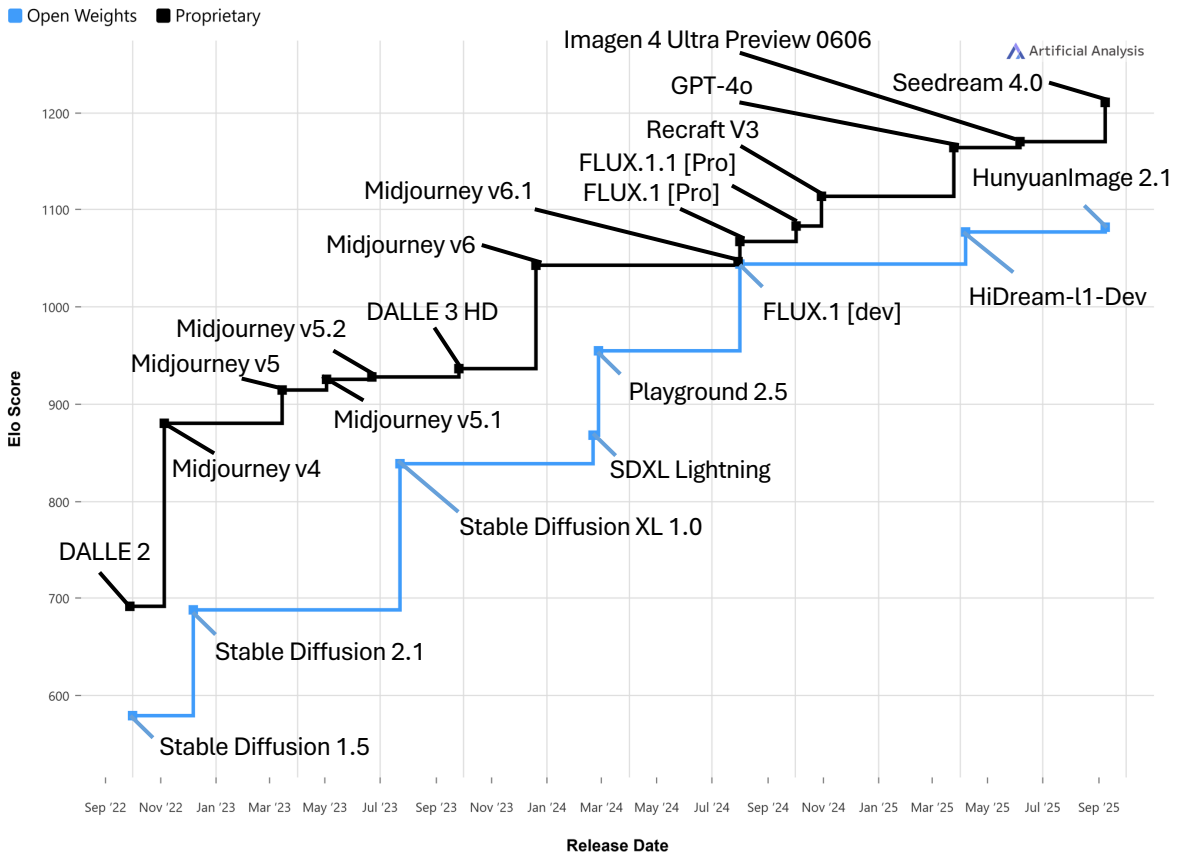
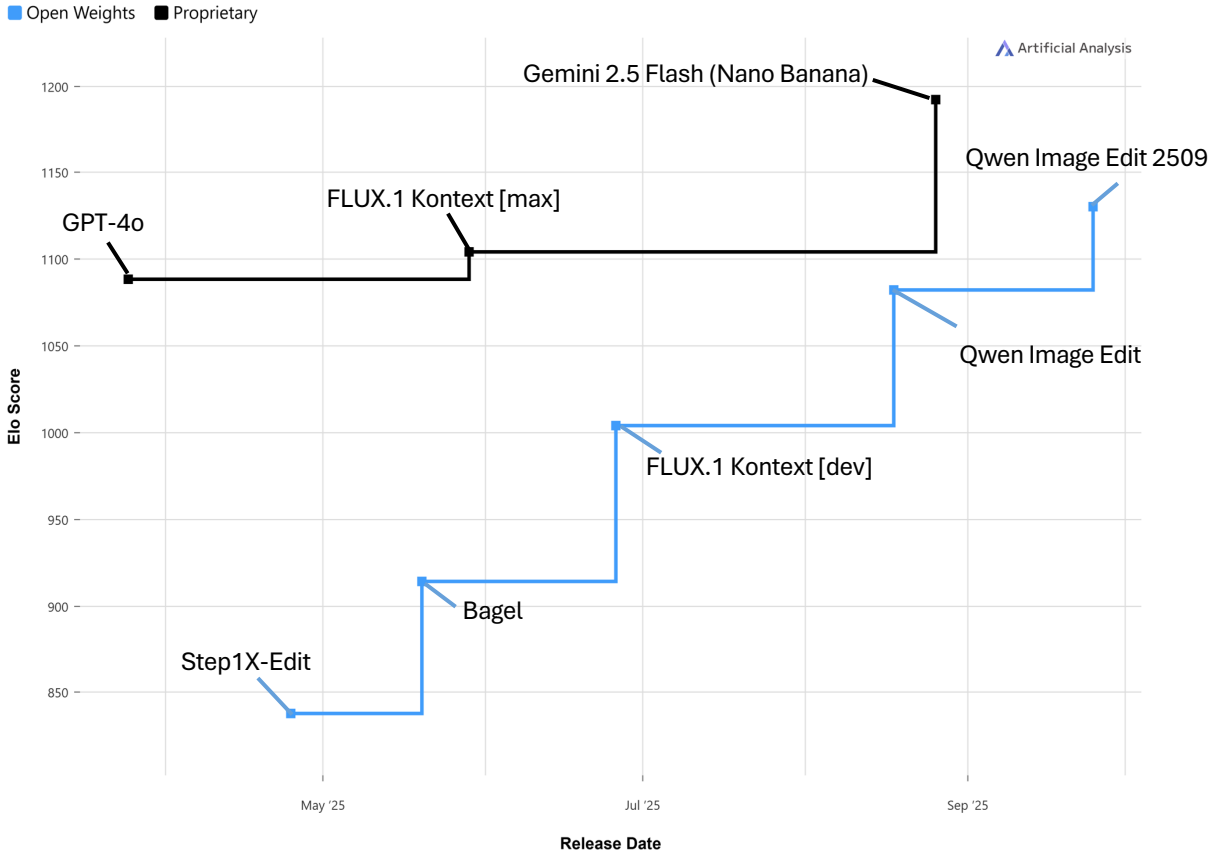


Image Editing

Elo Score of proprietary and open weights image editing models over time



Source: Artificial Analysis independent benchmarking

Deep Dive: Proprietary models lead frontier performance for video generation models as open weights peers consistently lag

Leading Text to Video and Image to Video Models by License Type, Over Time

Text to Video

Elo of proprietary and open weights text to video models over time

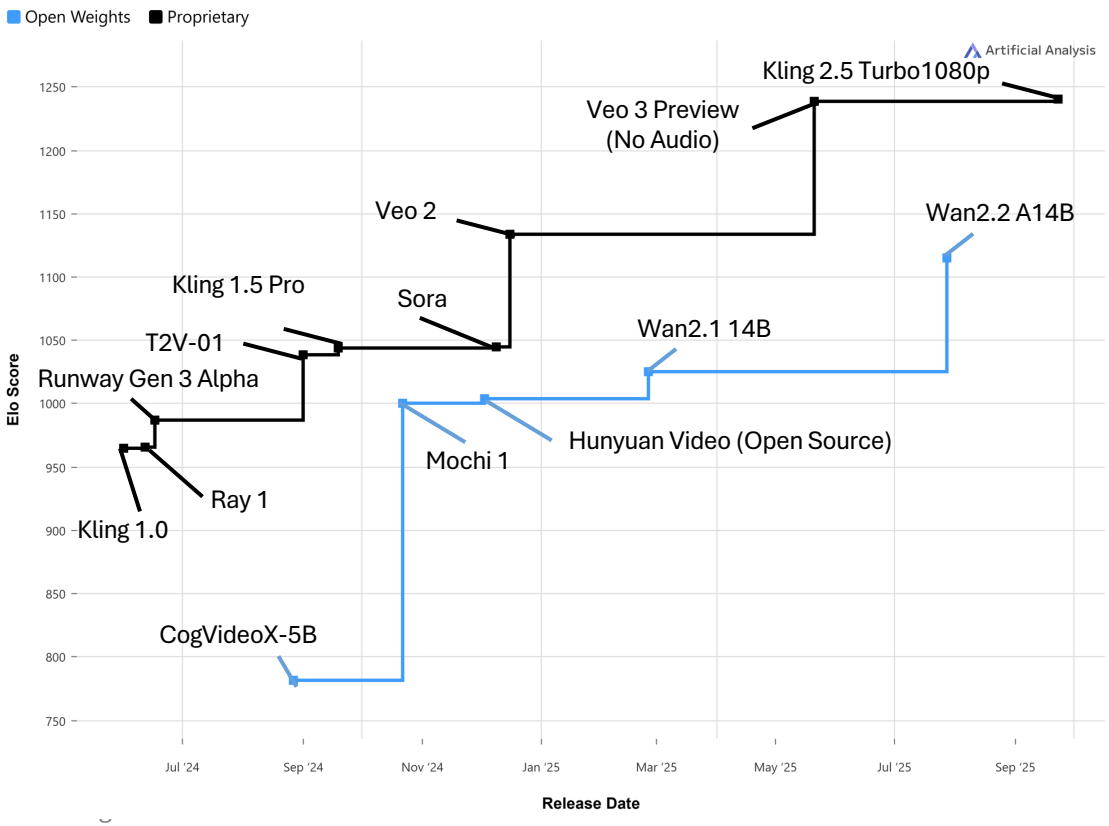
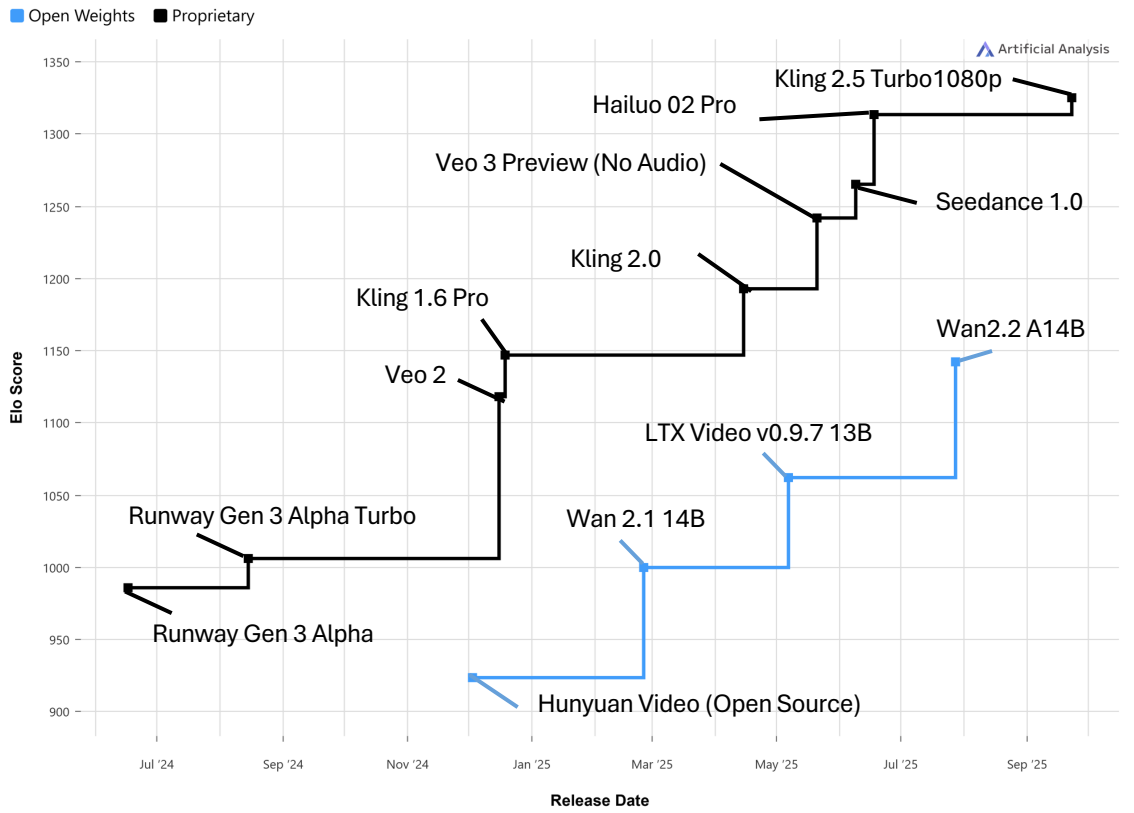


Image to Video

Elo of proprietary and open weights image to video models over time



Source: Artificial Analysis independent benchmarking

04 Speech and Music Models

State of AI – Q3 2025

Speech and Music: Advances in speech and music AI are intensifying competition and enabling more natural, capable, and cost-effective voice agents

Key Themes in Q3 '25

Speech to Text (STT) continues to see new accuracy frontiers	• **Artificial Analysis' new AA-WER (Word Error Rate) Index** measures transcription model accuracy using three datasets comprised of real-world speech with varied accents, domain-specific language, and challenging conditions • **Open weights models continue to demonstrate accuracy gains** with NVIDIA's Canary Qwen (13.2% on AA-WER) and Parakeet TDT (13.7%) placing second and third on the AA-WER, behind Google's Chirp 2 (11.6%) • **Increasing presence of multi-modal / LLM-based models for** transcription with OpenAI now offering a GPT Transcribe STT endpoint and is placing less emphasis on its Whisper-based endpoint
Text to Speech (TTS) models now offer fine-grained control	• **Models released by OpenAI and MiniMax, prior to Q3 '25, continue to lead TTS performance,** with newer models released in Q3 2025 releases not yet surpassing them. ElevenLabs' latest model follows closely behind leaders • **New TTS models (e.g. ElevenLabs v3) now support advanced speech delivery** elements including emotion and tone control through in-text tagging and Speech Synthesis Markup Language (SSML), enabling nuanced & expressive generation
Speech to Speech (STS) models sees sustained growth and new models	• **Google's Gemini 2.5 Native Audio Thinking is the new leading Speech to Speech model for reasoning**, proving the capability of native audio reasoning models • **STS is rapidly expanding with new open weights entrants like Alibaba's Qwen3 Omni Flash**, while OpenAI's GPT Realtime series (August launch, October Mini release) demonstrates continuous iteration with the release of new models
Voice Agents	• **A number of voice agent use cases have gained traction in enterprises** including inbound customer support, tele-service, and employee support/training • Different voice agent platforms take different approaches and include model-focused offerings (Inworld), end-to-end platforms (Decagon) and assembly toolkits (Vapi)
Music	• Q3 '25 saw the emergence of **new proprietary music generation models capable of generating music with instrumentals and vocals,** predominantly by specialist labs such as Suno, ElevenLabs and Producer.ai

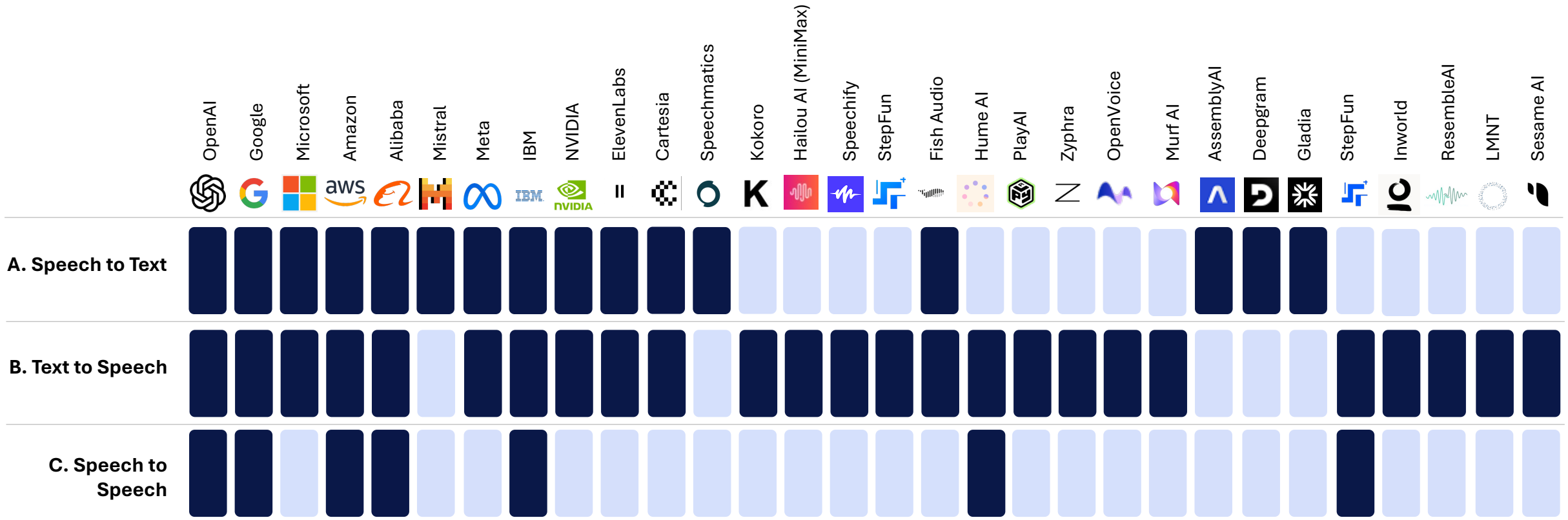
Speech: While the generalist AI labs (e.g. OpenAI, Google) have offerings across speech AI modalities, pure-play speech startups continue to drive innovation

Key players offering speech models

Classifications are indicative and determined based on models available

Low or no presence | Strong presence

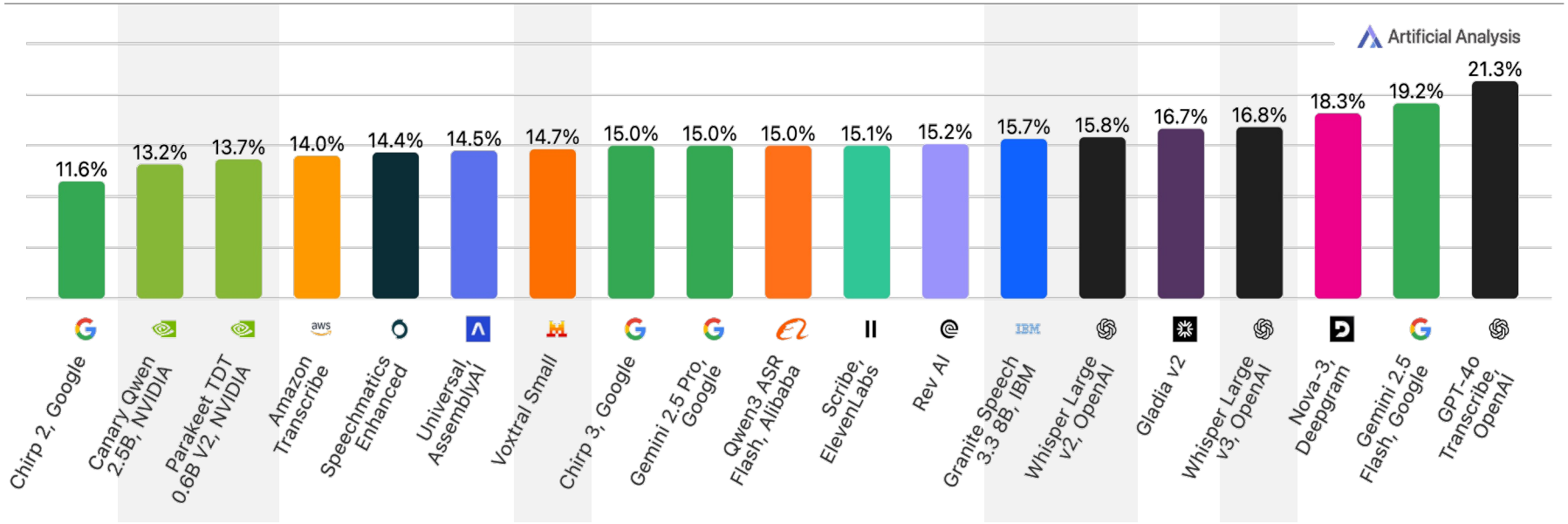
Generalist Labs | **Speech Focused AI Labs** | **NON-EXHAUSTIVE**



A. Speech to Text: Google's Chirp 2 leads the Artificial Analysis Word Error Rate (AA-WER) Index, followed closely by NVIDIA's recent open weights releases

Speech to Text Accuracy

Artificial Analysis Word Error Rate (AA-WER) Index: % of words transcribed incorrectly



- Artificial Analysis Word Error Rate (AA-WER) Index measures transcription accuracy across 3 datasets with diverse accents, domain-specific language, and challenging channel & acoustic conditions
- AA-WER is calculated as an audio-duration-weighted average of WER across ~2 hours each from three datasets: VoxPopuli, Earnings-22, and AMI-SDM

Note: Models that do not support transcription of audio longer than 10 minutes were evaluated on 9-minute chunks of the test set (applies to Whisper (L v2), Deepgram; Granite 3.3 8B; GPT-4o Transcribe; GPT-4o Mini Transcribe; Voxtral Mini; Voxtral Mini, Deepinfra; Gemini 2.5 Flash Lite). For models with even shorter time limits, all files are split into 30-second chunks (applies to Granite Speech 3.3 8B, IBM; Qwen3 ASR Flash).

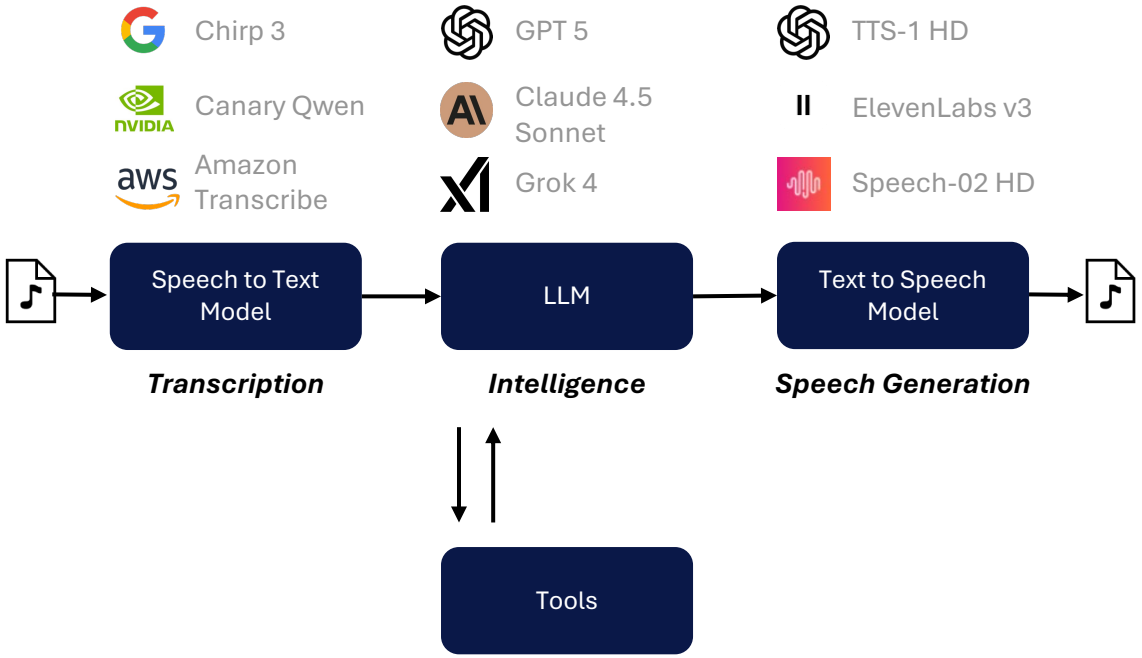
Commentary

- **Google leads in transcription accuracy:** Google Chirp 2 achieves the lowest WER (11.6%), outperforming peers and Google Chirp 3 model
- **Open weights models close the gap**: NVIDIA's Canary and Parakeet deliver strong accuracy at 13–14% WER, as open weights modes rival proprietary peers
- **OpenAI trades literal accuracy for fluency**: GPT-4o Transcribe produces smoother, more natural transcripts, but this results in higher WER (21.3%), especially

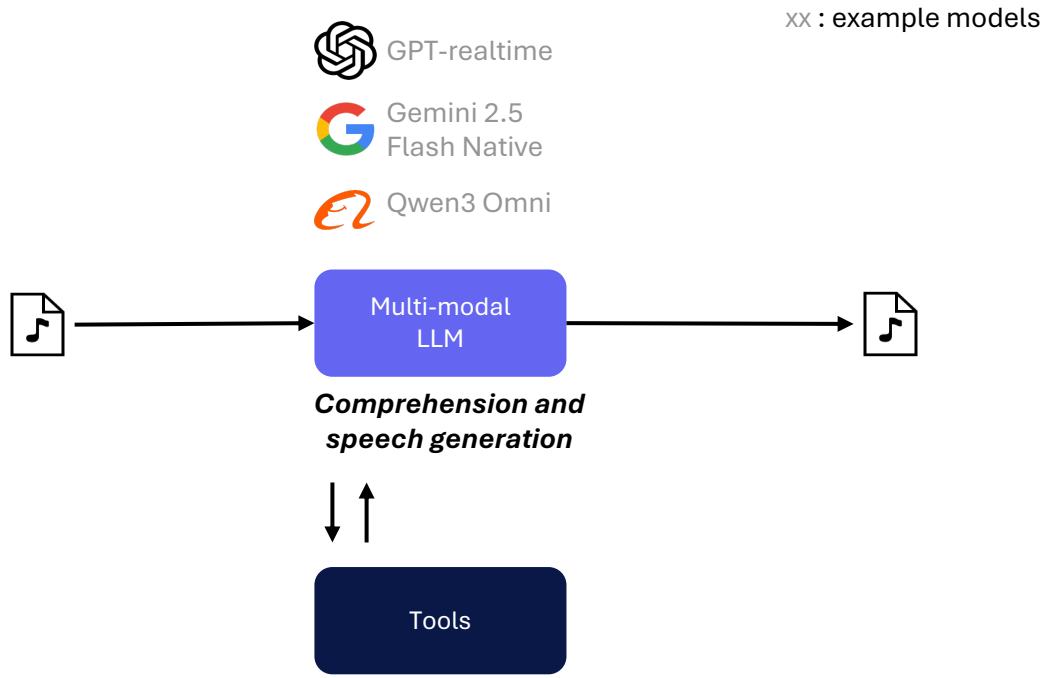
C. Speech to Speech: OpenAI and others have released models with native Speech to Speech capabilities; these can reduce latency and complexity when creating voice agents

Speech to Speech architectures used in voice agents

Pipeline[1]**:** Processes audio sequentially through multiple models



Native Speech to Speech: Directly process audio input and output in a single model



- Traditional implementations, which make up the majority of voice agents, typically leverage at least 3 models in a pipeline (transcription, intelligence, speech generation) which introduces **latency and complexity that native speech-to-speech models avoid**
- **Challenges exist in adoption** of Native Speech to Speech models**,** including the inclusion of information, compliance & monitoring and steerability

1. Also referred to as cascading or chained

05 Accelerators

State of AI – Q3 2025

NVIDIA Blackwell 8xB200 systems now widely available; rack-scale GB200 NVL72 systems are production with limited availability; B300 and GB300 expected before end of year

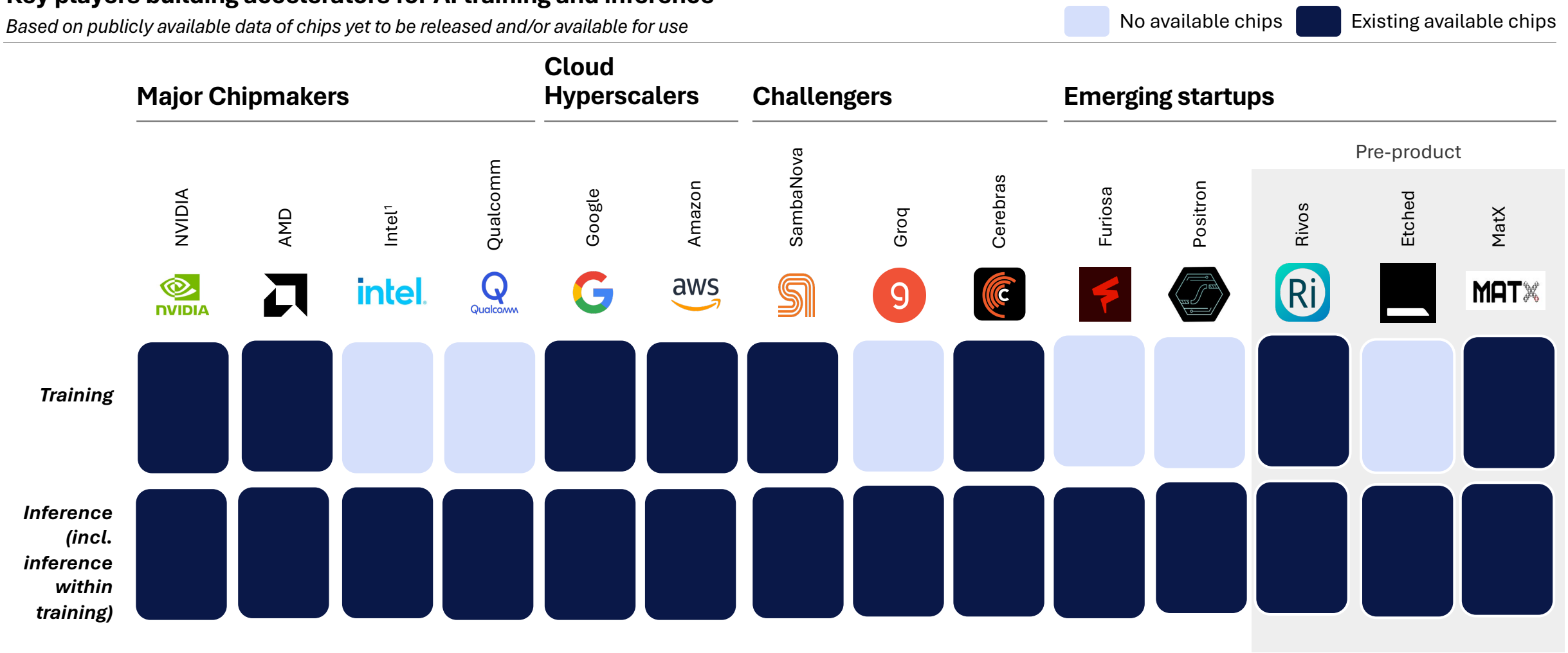
Key themes in Q3 '25

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 models trained on GB200 rack-scale already released	• 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
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 recent 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

Accelerators: 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

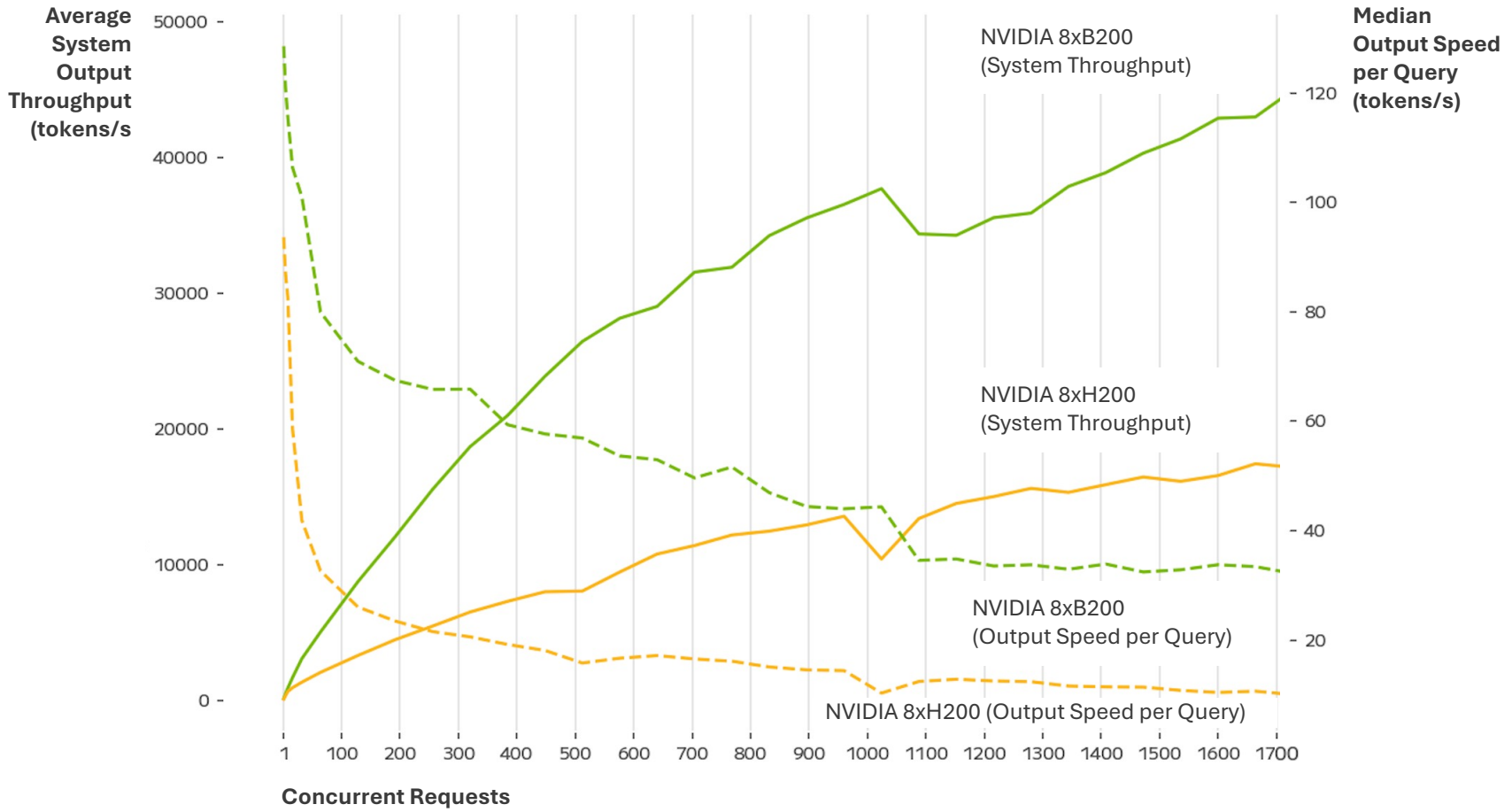


1. Intel is no longer bringing to market Falcon Shores, its successor to Gaudi 3, and we do not expect Intel to bring Jaguar Shores to market until 2026 at the earliest.

Artificial Analysis System Load Test: NVIDIA B200 chips demonstrate significant performance improvements with TensorRT-LLM compared to NVIDIA H200 and H100

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/