PUBLIC RELEASE

# Artificial Analysis State of AI

## Q1 2025

## Highlights Report

Full report available to Premium Access 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 strategies.

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 evaluations.

Our public website, artificialanalysis.ai, is widely referenced throughout the AI industry. To discuss this report, our publications or our services, please contact us at contact@artificialanalysis.ai.

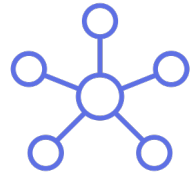
# 6 major trends have shaped progress in AI in Q1 2025

## A AI Progress Continues

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

## B Reasoning Models

Models which 'think' before answering by outputting tokens drove significant intelligence gains and became widespread beyond OpenAI



## C Efficiency & MoEs

Models increasingly use a Mixture of Experts architecture, activating a subset of parameters at inference time, increasing inference efficiency

## D Rise of Chinese AI

AI Labs headquartered in China have released models with intelligence rivalling that of US labs, particularly amongst open weights models

## E Agents

AI systems can increasingly perform tasks end-to-end autonomously by chaining requests to LLMs together

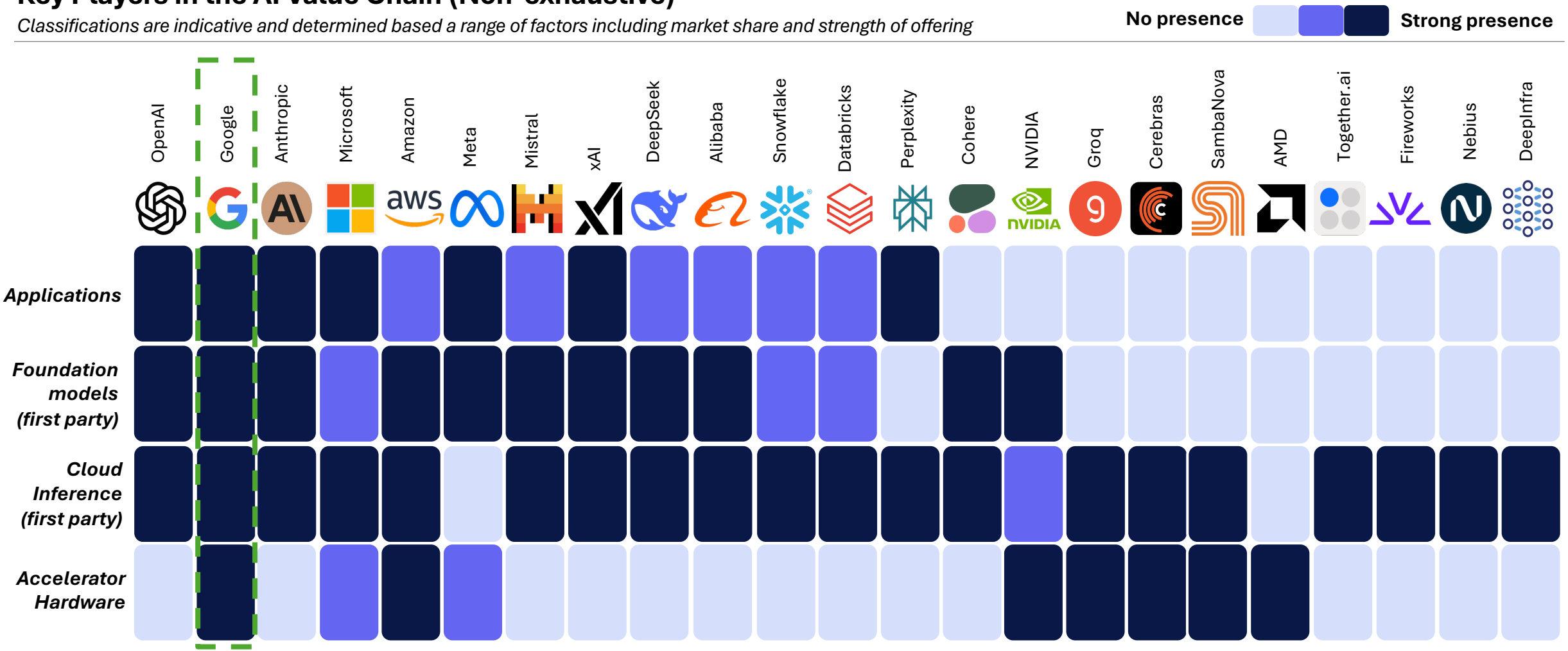
## F Multimodal AI

AI models are increasingly multimodal, working natively with image and audio; modality-specific models continue to advance

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

## Key Players in the AI Value Chain (Non-exhaustive)

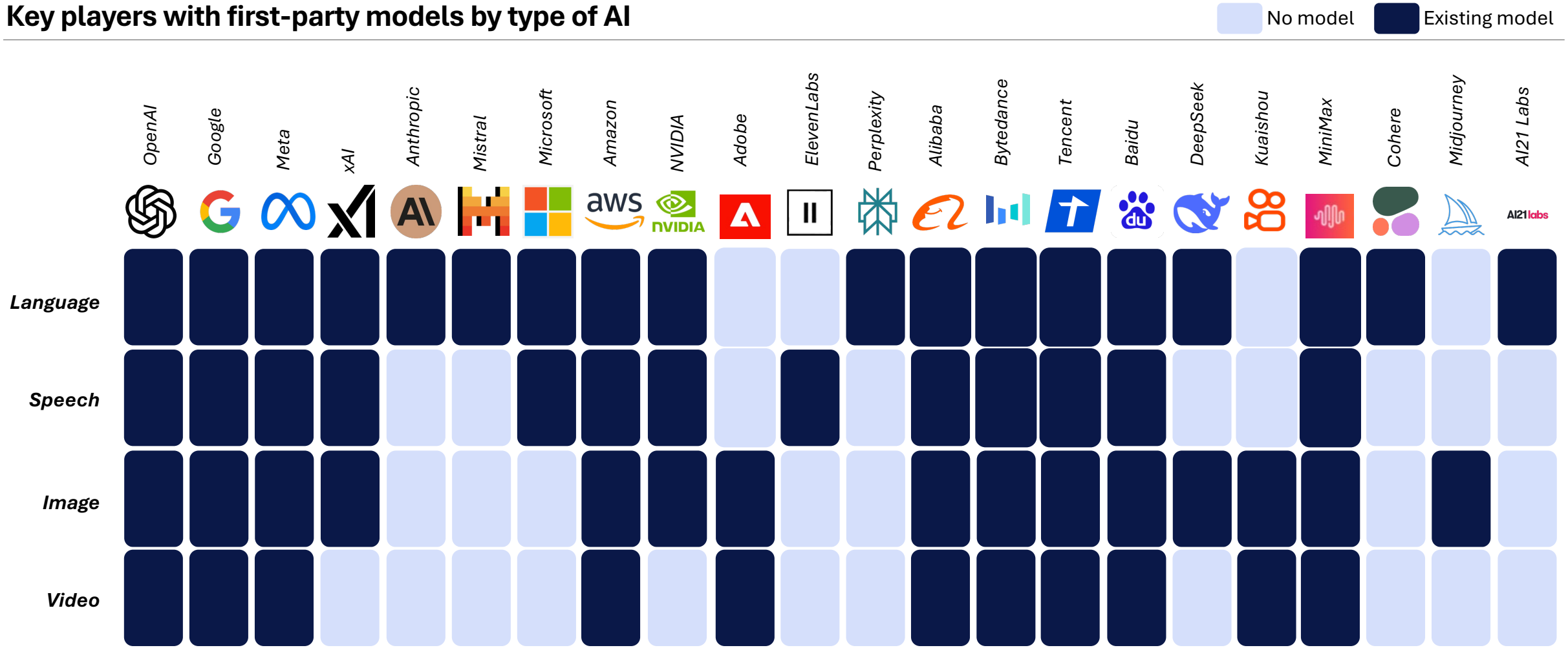
*Classifications are indicative and determined based a range of factors including market share and strength of offering*




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

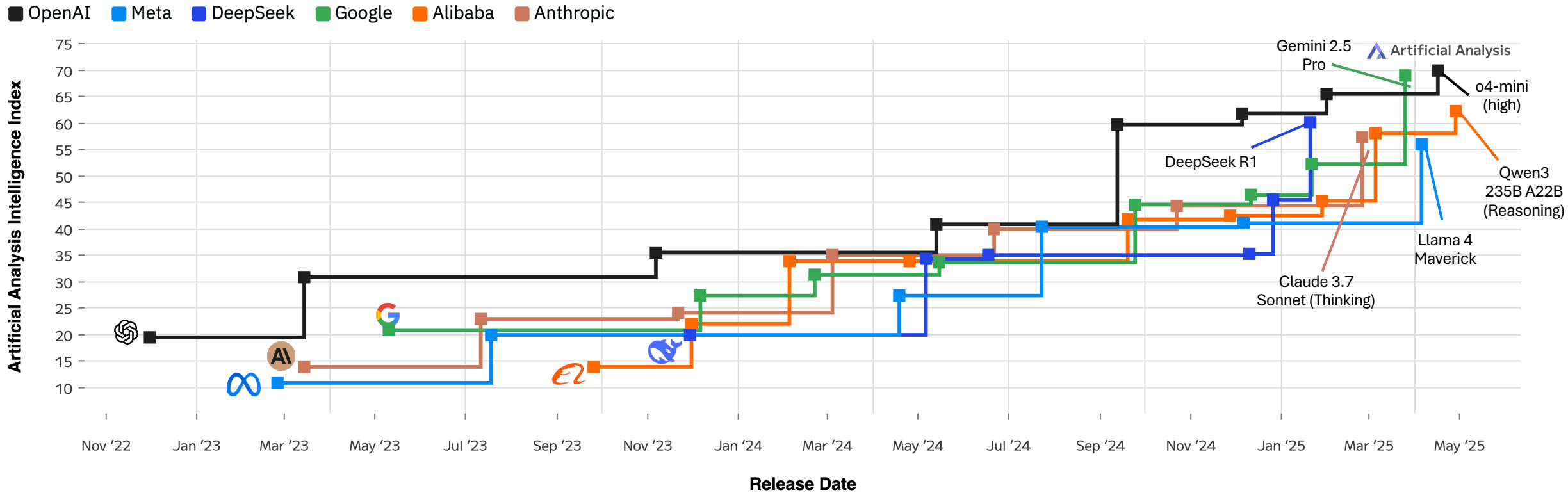



Source: Company website

# OpenAI remains in lead, but competition for top tier intelligence is fiercer than ever; reasoning models that 'think' before they answer are driving the latest leaps in intelligence

## Frontier Large Language Model (LLM) Intelligence, Over Time

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



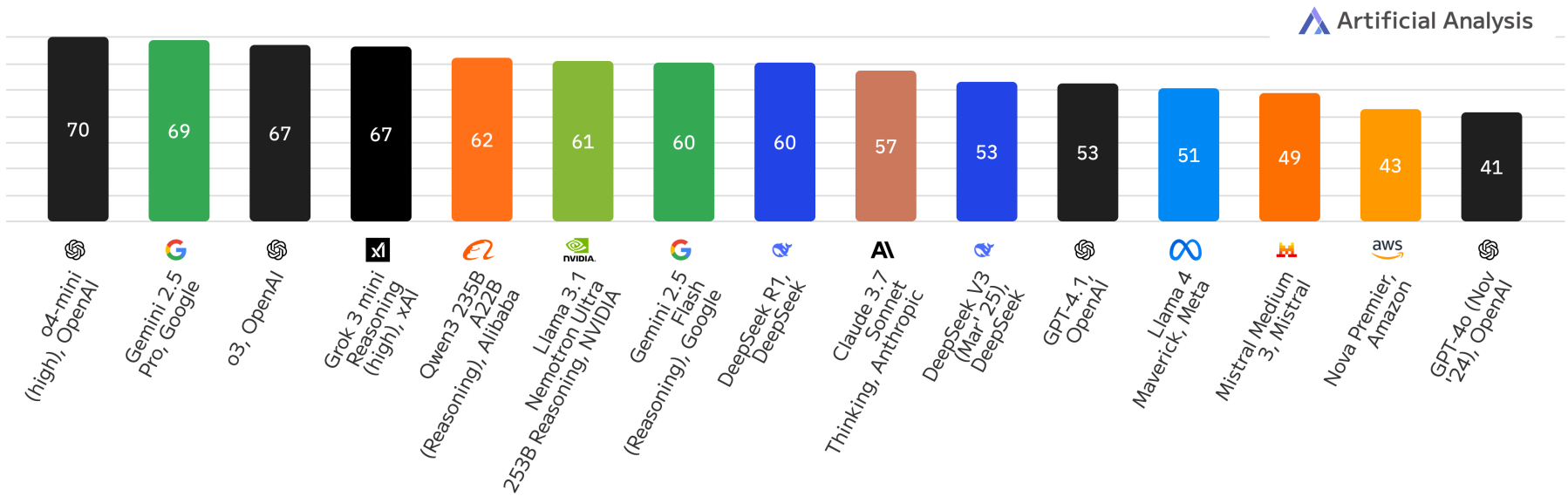

- **OpenAI continues to maintain lead but gap narrows:** OpenAI's o4-mini (high) model is the most intelligent model closely followed by reasoning models from Google's Gemini 2.5 Pro and xAI's Grok 3 model
- **Open weights models amongst the most intelligent:** DeepSeek R1and NVIDIA Llama 3.1 Nemotron Ultra approach intelligence of proprietary models
- **Labs continue to release rapidly on a quarterly basis:** Labs have typically released their new frontier models on a 3-month release cycle

Source: Artificial Analysis independent benchmarking

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

## Leading Large Language Model (LLMs), by AI lab

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




Source: Artificial Analysis independent benchmarking

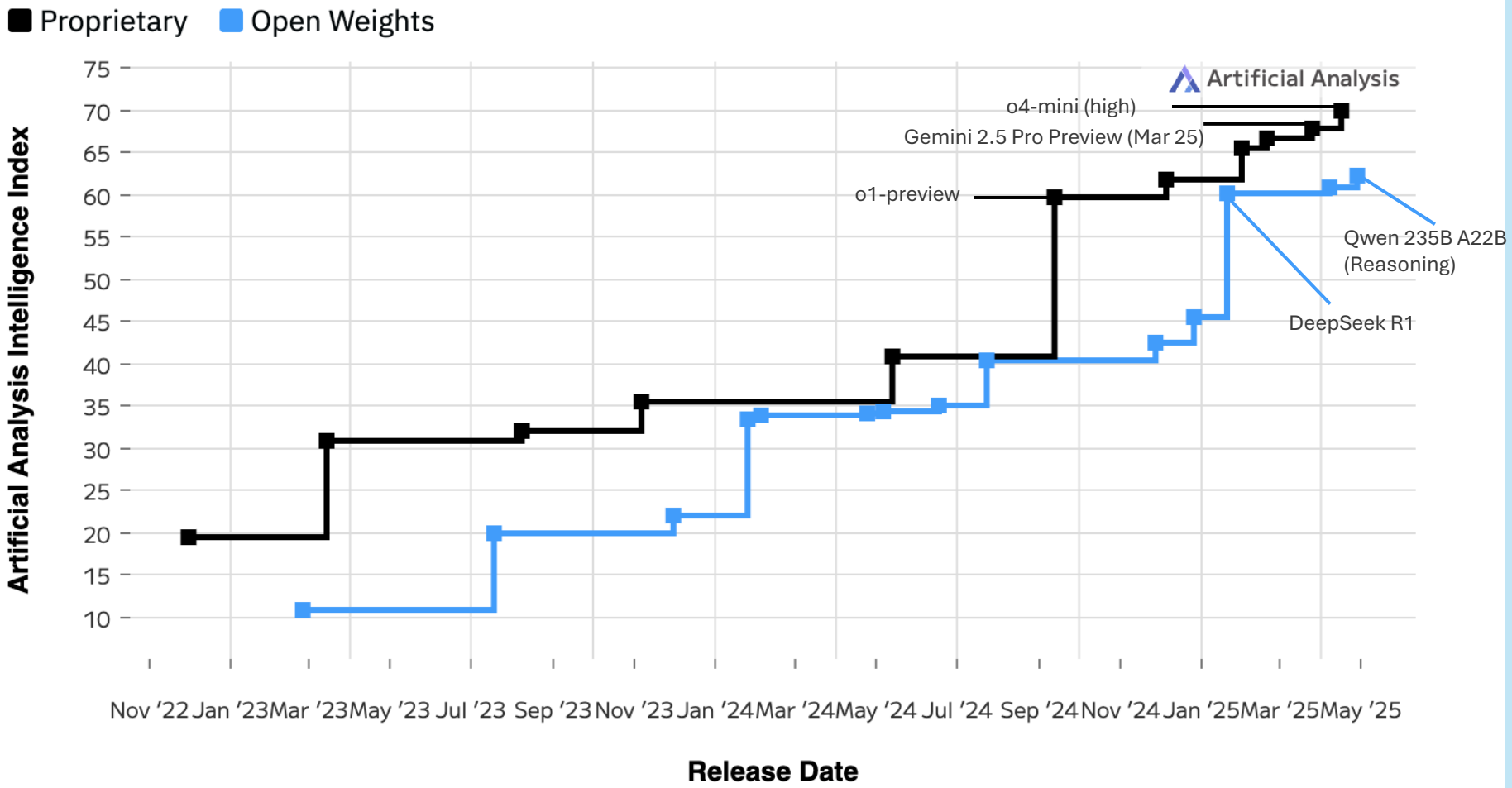
## Commentary

- **OpenAI continues to lead:** OpenAI's o3 and o4-mini-high sit on the current intelligence frontier, with Google coming closer than ever to having the top model in our intelligence Index with Gemini 2.5 Pro
- **Reasoning is new frontier:** All the models that score highest in Intelligence Index are reasoning models that 'think' before answering the questions
- **A more and more crowded frontier:** The big 5 US labs (OpenAI, Google, Anthropic, xAI, Meta) are joined near the frontier in early 2025 by NVIDIA, DeepSeek, Alibaba, Mistral and Amazon

# Open weights reasoning models mostly closed the gap opened by OpenAI's o1 in late 2024; proprietary models continue to lead

## 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

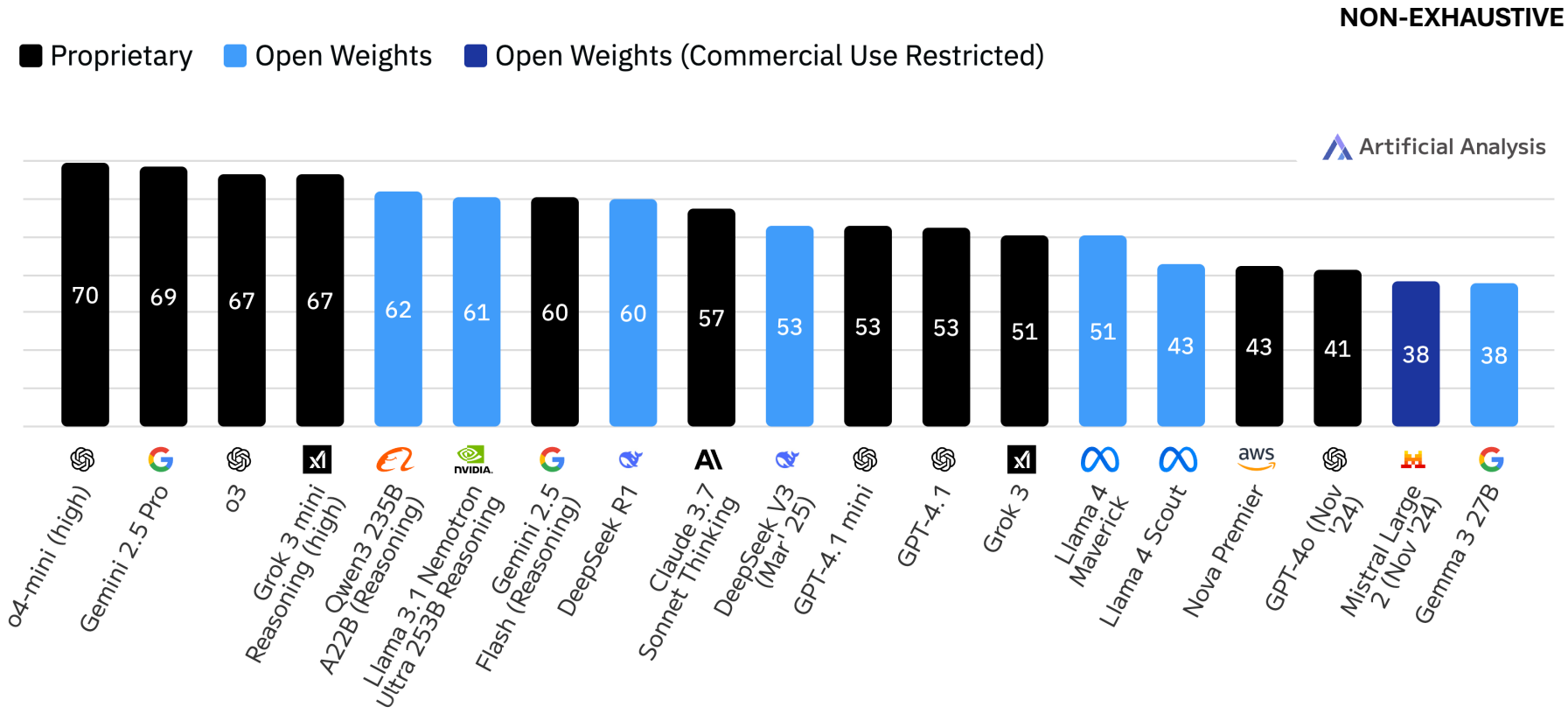
## Commentary

- **Open weights reasoning models become available:** The release of DeepSeek R1 in early 2025 significantly reduced the intelligence gap by making reasoning available outside of proprietary models; recent proprietary releases (e.g. o4-mini, Gemini 2.5 Pro) have since widened the gap
- **Frontier model releases are becoming more frequent:** Early 2025 saw a wave of releases and refinements driving intelligence in short rapid incremental improvements

# Today's leading open weights models come from Alibaba, DeepSeek, Meta and Nvidia

## Leading Language Models by License Type

*Artificial Analysis Intelligence Index (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 (ie. including reasoning models)
- **Open weights models lead non-reasoning intelligence:** DeepSeek V3 0324 is currently the most intelligent non-reasoning model, followed closely by Llama 4 Maverick

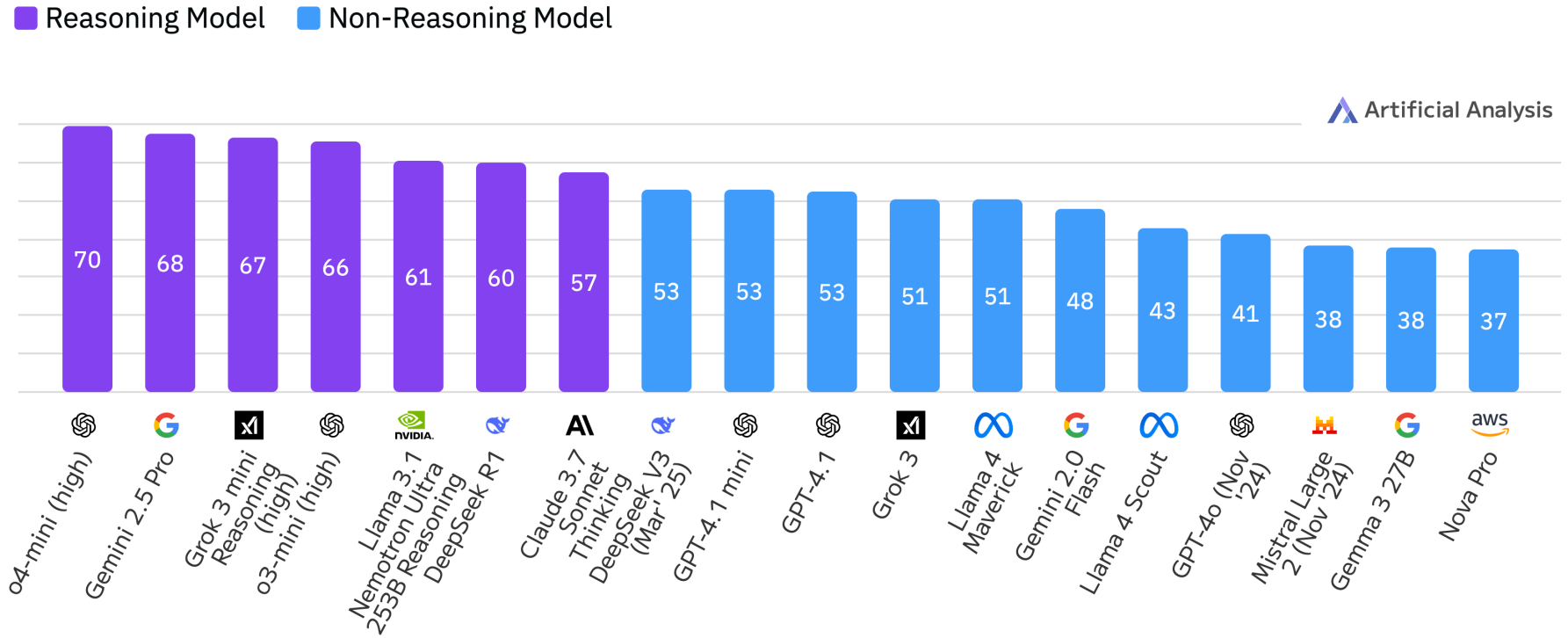
Source: Artificial Analysis independent benchmarking

# Reasoning models now decisively lead Artificial Analysis Intelligence Index

## The Language Model Frontier: Reasoning vs Non-Reasoning

*Artificial Analysis Quality Index (incorporates 7 leading evaluations), Leading Models*

**NON-EXHAUSTIVE**




Source: Artificial Analysis independent benchmarking

## Commentary

- **Reasoning models are the new frontier** as they demonstrate a clear intelligence edge over non-reasoning peers and top the Artificial Analysis Intelligence Index
- **All leading AI labs** (including OpenAI, Google, Anthropic, xAI, DeepSeek) now offer reasoning models, but continue to offer non-reasoning models (or take a hybrid approach)
- Non-reasoning models do, however, **remain the best option for a wide range of tasks**, especially where cost and speed are critical

# Reasoning models 'think' before they answer, often taking more time and tokens to provide a more accurate answer for complex problems

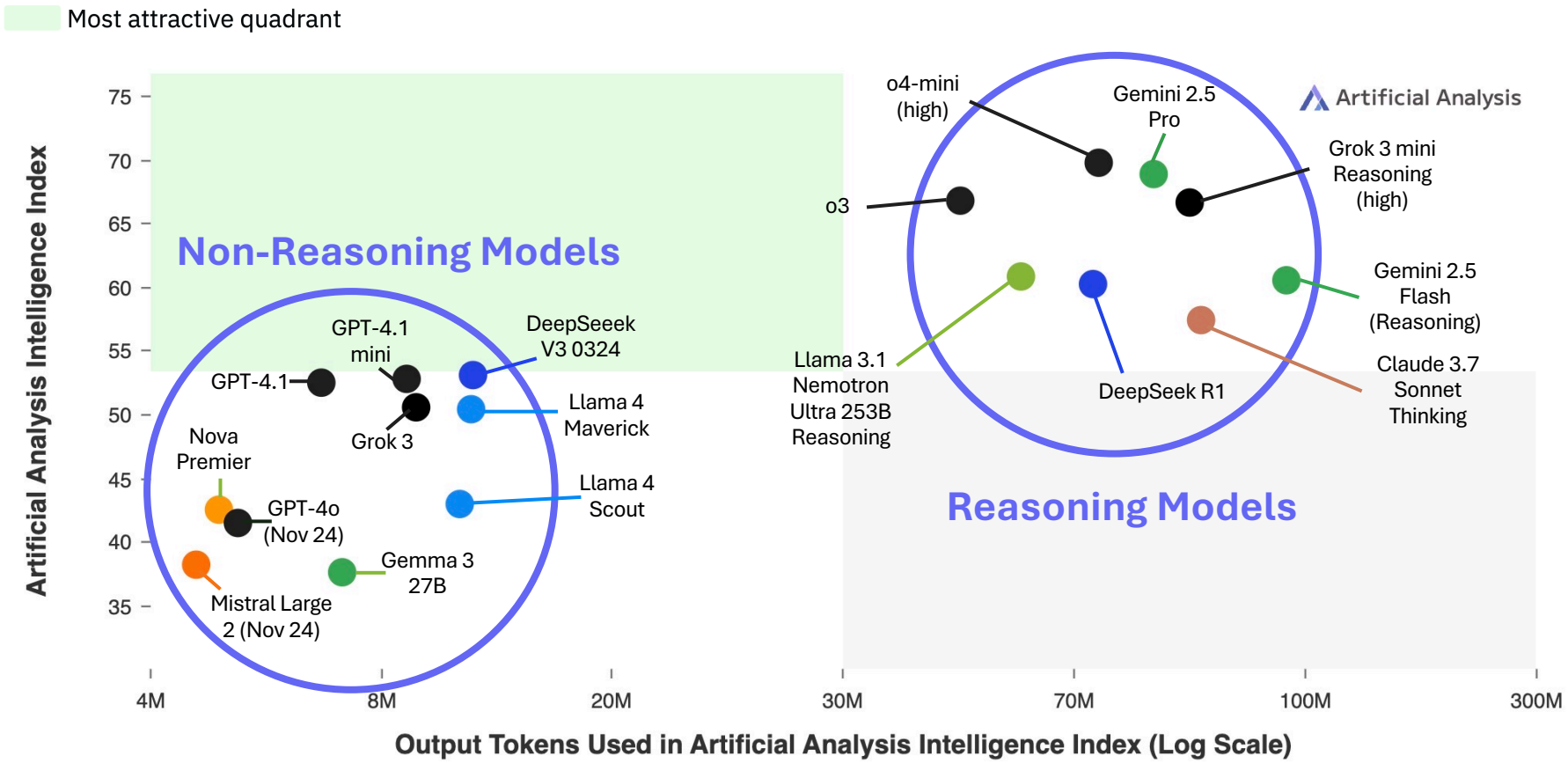
Thinking tokens | Answer tokens

| | **Non-reasoning model** (OpenAI GPT-4o) | **Reasoning model** (Google Gemini 2.5 Pro) |
|---|---|---|
| **Prompt** | *Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a gold bar; behind the others, rotten vegetables. You pick a door, say No. 1, and the host asks you "Do you want to pick door No. 2 instead?" Is it to your advantage to switch your choice?* | |
| **Model output** | `Yes, it is to your advantage to switch your choice. Here's why:`<br><br>`Initially, you have a 1/3 chance of picking the door with the gold bar and a 2/3 chance of picking a door`<br><br>`This scenario is similar to the Monty Hall problem, where switching doors is a statistically better strategy.` | `Here's a thinking process to arrive at the explanation for the Monty Hall problem variant:`<br><br>`1. **Identify the core problem:** The user describes a scenario similar to the classic Monty Hall problem, but with a slight variation. …`<br><br>`… 9. **Self-Correction/Check:** Does this make sense? Yes …`<br><br>`… Okay, let's break this down. This sounds very similar to the famous Monty Hall Problem, but there's a crucial difference in how you've described the host's action …`<br><br>`… Since your host didn't open a door and reveal a losing prize, that crucial shift in probability doesn't happen. It's just a choice between two doors, each with a 1/3 probability.` |
| **End to end response time** | ~4 seconds | ~21 seconds |
| **Output tokens** | ~185 tokens | ~1,967 tokens |
| **Commentary** | • **Incorrectly answers question** as it does not understand the prompt is different to the Monty Hall problem, instead simply answering as if the query were a Monty Hall problem<br>• **Uses less time and fewer tokens** but does not provide the correct answer | • **Correctly answers the questions** as it spends 'thinking' time to understand the prompt varies from the Monty Hall problem and breaks down this new problem to solve it<br>• **Uses more time and tokens** but provides the correct answer |

# Treating reasoning & non-reasoning models as distinct categories is a helpful framework for understanding today's model landscape

## Intelligence vs. Output Tokens Used to Run Artificial Analysis Intelligence Index

*Artificial Analysis Intelligence Index (Version 2, released Feb 25), Output Tokens Used (~5M input tokens)*




## Commentary

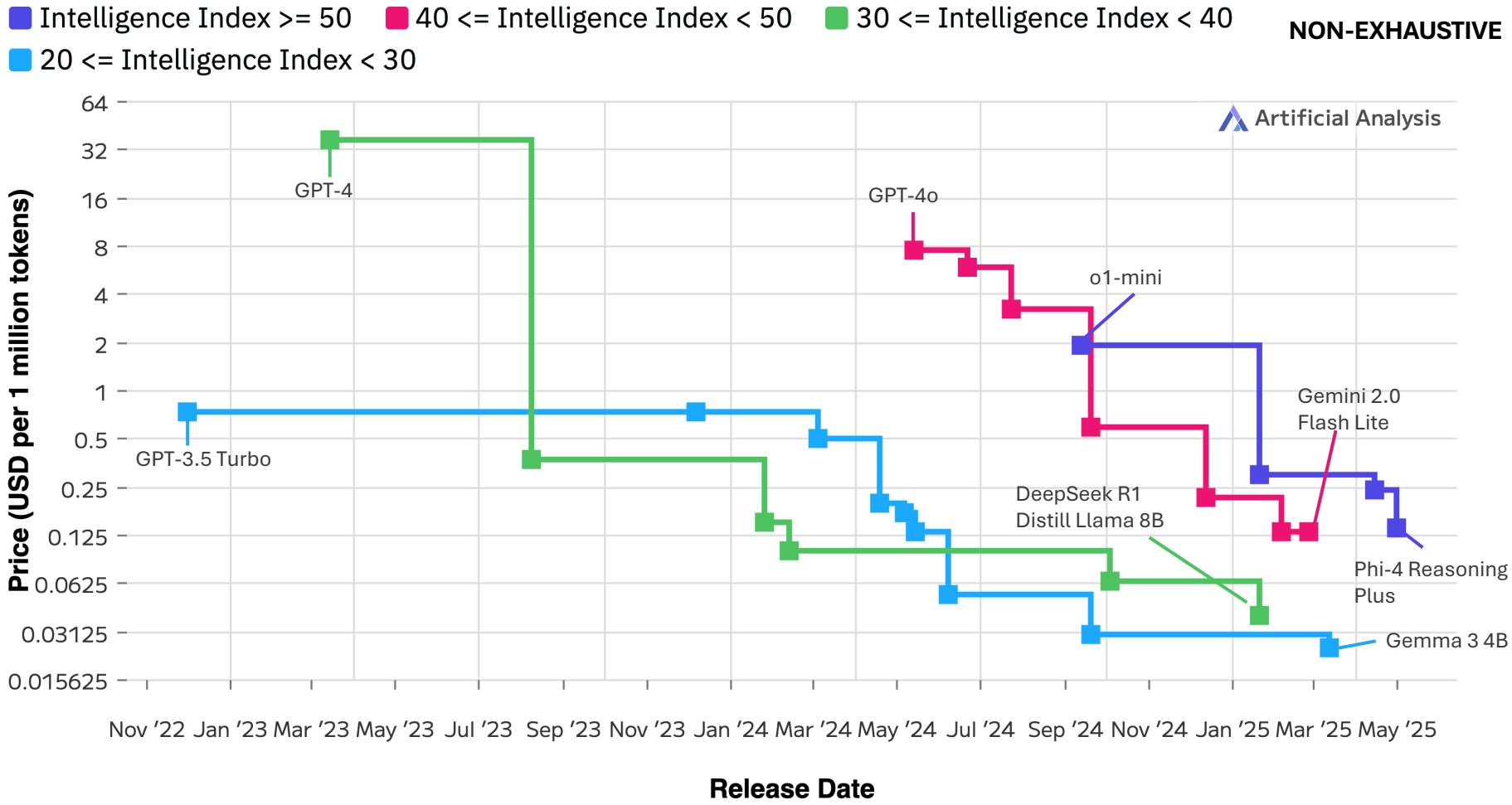
- **Two distinct groups:** The difference in token usage between reasoning and non-reasoning models is much greater than differences within each group; the median reasoning model uses **up to 10x more tokens to complete our Intelligence Index** as the median non-reasoning model
- When using reasoning models, developers now must go beyond per-token pricing and consider token usage to properly understand inference cost

Source: Artificial Analysis independent benchmarking

# The cost of inference has significantly reduced as small models continue to get smarter, inference efficiency increases, and new hardware generations drive down the cost of compute

## 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 (incorporates 7 evaluations)*




## Commentary

- **The cost of intelligence has continued to decline rapidly,** with inference for high intelligence models (>50 on intelligence index) already ~32x cheaper since September 2024.
- **The cost of intelligence at the level of the original GPT-4 has now declined >1000x** since the launch of GPT-4 in March 2023
- **Three primary drivers enable this continuing cost revolution**: smaller models (incl MoE architectures), inference optimizations and new hardware generations

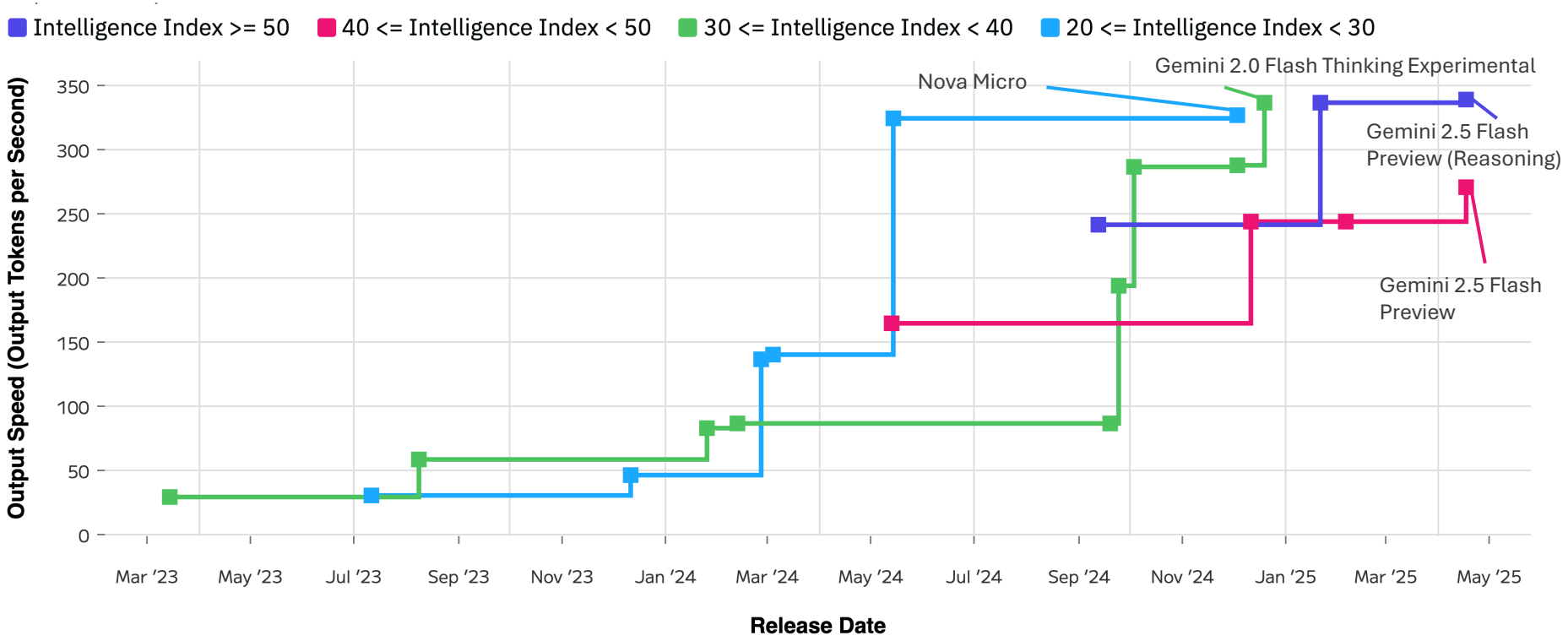
Source: Artificial Analysis independent benchmarking

# Inference is faster than ever in early 2025 – this is particularly critical as reasoning models and agentic applications are generating >10x more tokens per request than average queries a year ago

## Language Model Output Speed by Intelligence, Over Time

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

**NON-EXHAUSTIVE**




## Commentary

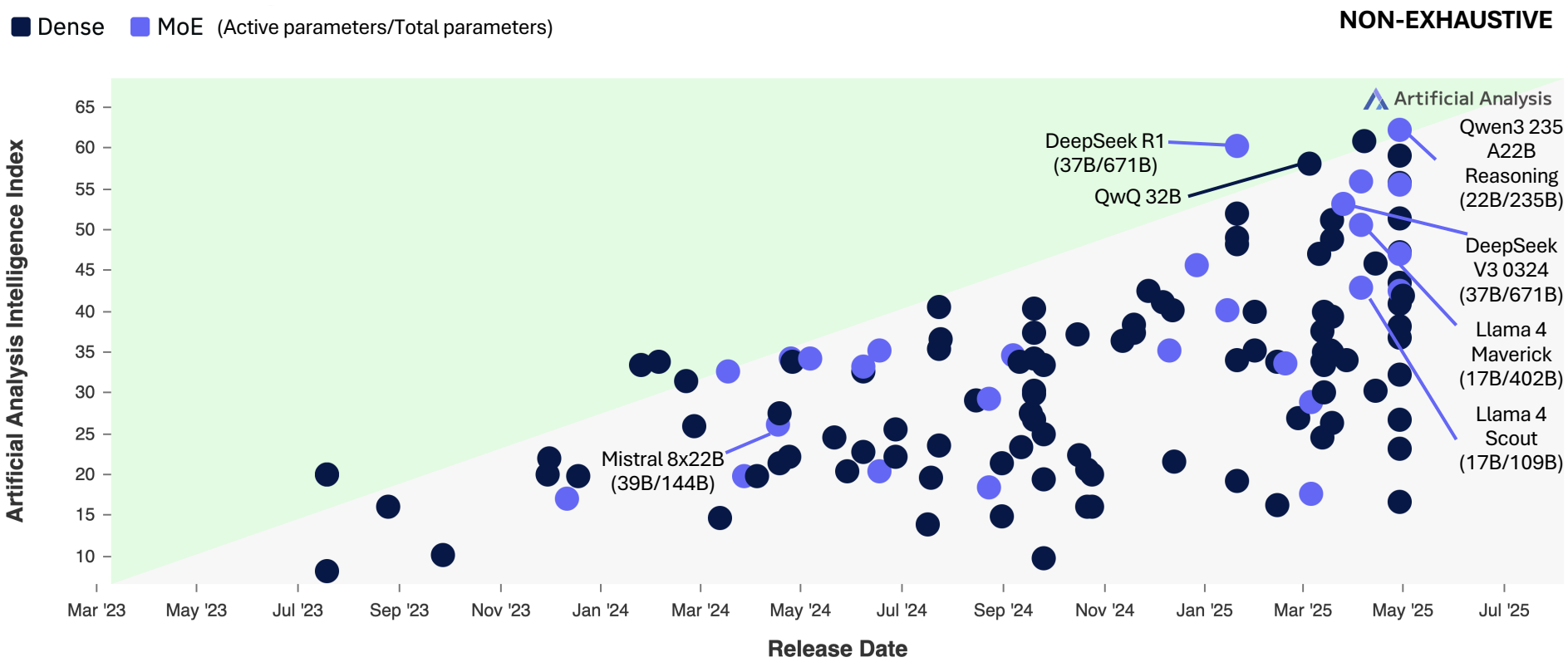
- **The drivers of increasing speed are closely related to the factors driving reduction in inference cost:** smaller models (including MoE architectures), inference techniques (especially speculative decoding and distributed inference) and new hardware generations
- **Despite these speed gains, the user experience of AI applications in early 2025 frequently involves waiting longer!** This is being driven by:
  - Reasoning models can use tens of thousands of output tokens to 'think' before answering
  - Agents can chain dozens of LLM calls together to complete tasks

Source: Artificial Analysis independent benchmarking

# Frontier models are increasingly adopting a Mixture of Experts (MoE) architecture to balance intelligence and efficiency

## Language Model Intelligence vs. Release Date, by Model Architecture

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




## Commentary

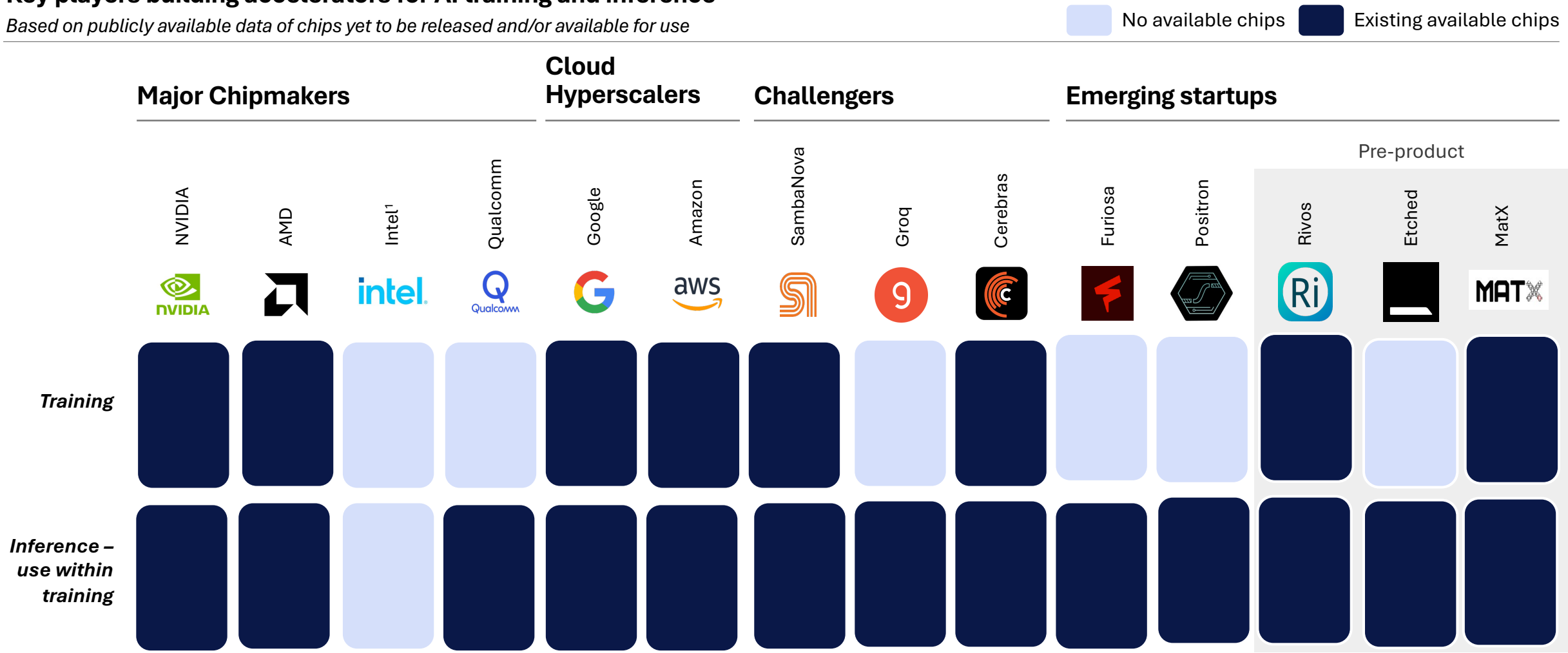
- **Frontier models are increasingly MoEs:** Recent open weights frontier models (e.g. Llama 4, DeepSeek V3, DeepSeek R1) are increasingly using an 'MoE' architecture where **only a subset of total parameters per input**
- **2025 models are sparser MoEs** activating <10% of total parameters (e.g. Llama 4 Maverick activates 17B out of 402B parameters, just ~4.2% of total); earlier models typically activated ~25% of total parameters
- MoEs are **more efficient for inference and training** than dense models with equivalent total parameters

Source: Artificial Analysis independent benchmarking

# 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*

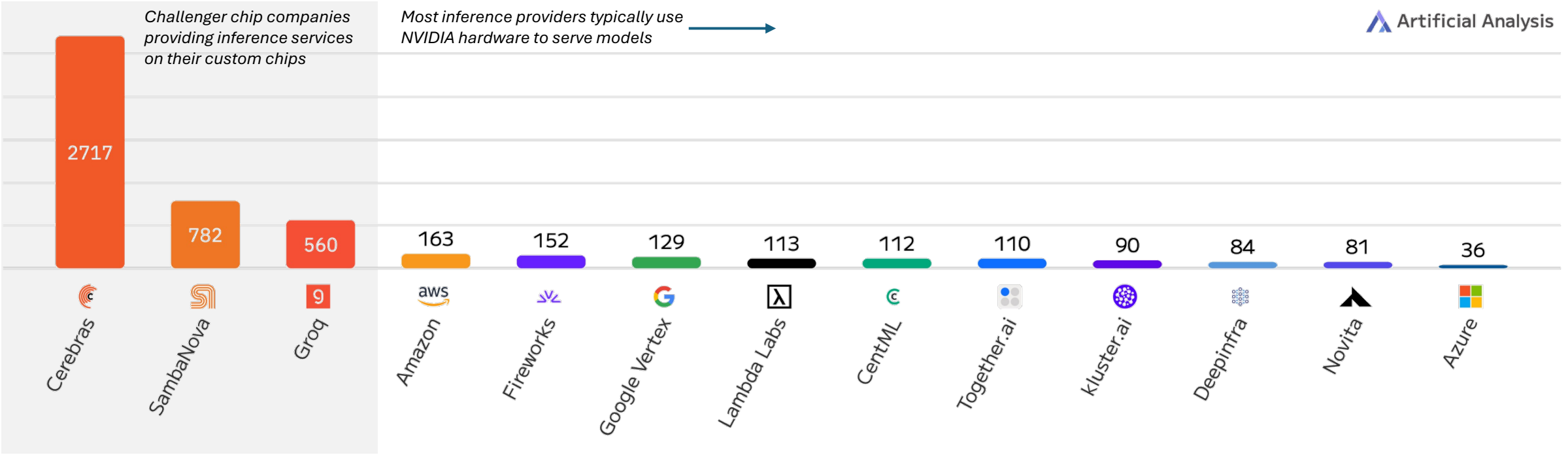



1. Intel is no longer bringing to market Falcon Shores, its successor to Gaudi 3, and we do not expect Intel to bring Jaguar Shores to market until 2026 at the earliest.

# Challenger chip companies have vertically integrated into cloud services to offer differentiated LLM inference services with greater performance

## Output Speed: Llama 4 Scout Serverless Endpoint Providers

*Output Tokens per Second; Higher is better*




- **Cerebras, SambaNova, Groq are chip companies that have vertically integrated into cloud inference,** often providing higher performance (i.e. output speed) than peers using NVIDIA hardware to serve the models
- **However, developers will need to consider the tradeoffs** between performance, cost and context window. While faster, these chip companies in cases serve models at a higher price point compared to other inference providers and with context windows smaller than models' native context length

Source: Artificial Analysis independent benchmarking

# While efficiency gains have been made...

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

### 1. Smaller models

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

**~1/10x** ***compute***

### 2. Software Efficiency

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

**~1/3x** ***compute***

### 3. 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

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

**~5x** ***compute/query***

### 4. Larger models

*Scaling laws continue to demand higher parameter counts for greater intelligence*

**~10x** ***tokens/query***

### 5. Reasoning models

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

**~20x** ***requests/use***

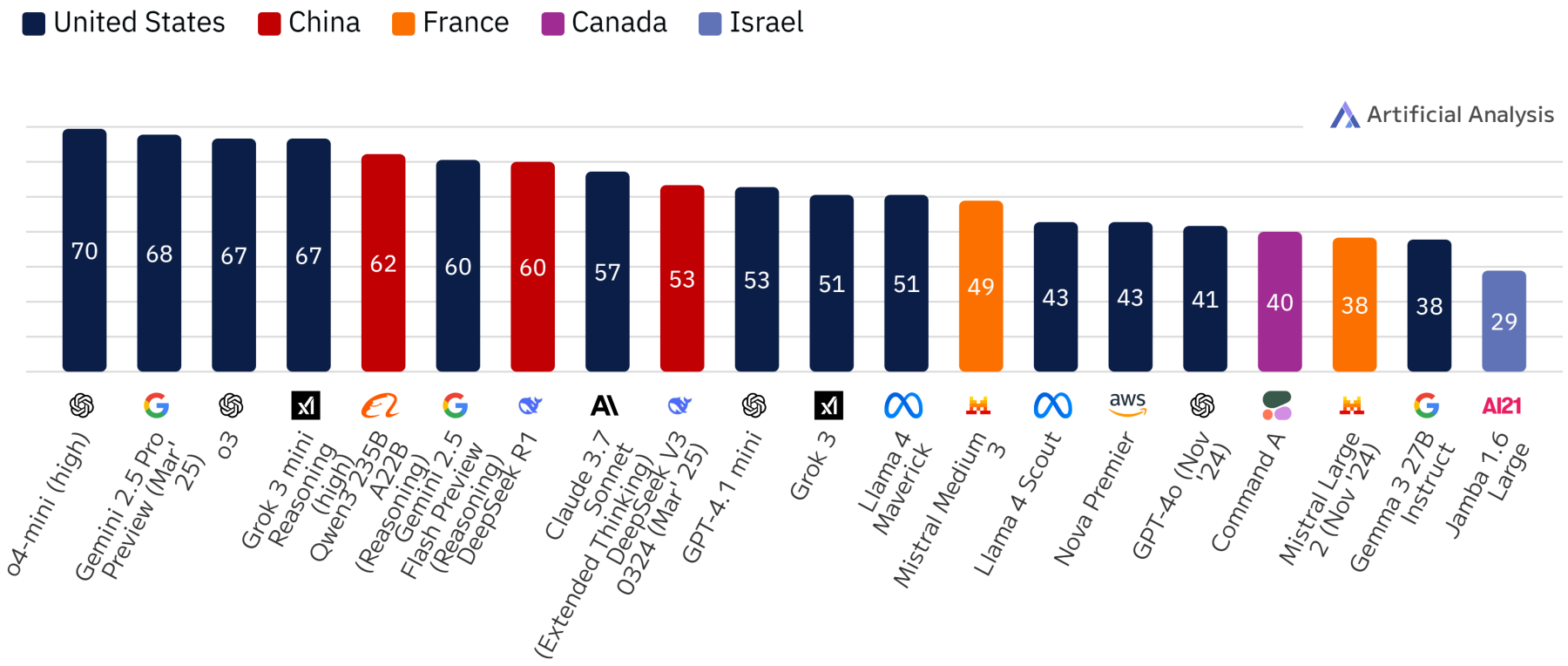
### 6. AI Agents

*Agents chain multiple requests to LLMs to complete tasks autonomously*

# The intelligence frontier is now overwhelmingly dominated by the US and China

## Leading Language Models by Country of Origin

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




## Commentary

- **The US leads across reasoning models:** the top 4 places on Artificial Analysis Intelligence Index are all taken by reasoning models from US labs
- **However, China currently leads non-reasoning models:** DeepSeek V3 0324 is the leading non-reasoning model, eclipsing models from the US and others
- **Smaller strides across other countries:** Models from labs outside of USA and China have continually improved but don't currently compete for frontier intelligence

Source: Artificial Analysis independent benchmarking

# 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**

## What agentic applications are working well in early 2025?

| | |
|---|---|
| Coding | LLM-driven system that reasons about the user's instructions, and then **finds, interprets, edits and tests source code** to complete software engineering tasks |
| Deep research | LLM-driven system that **parses a research query**, launches a **chain of targeted research queries while controlling its own research flow**, to form a synthesized answer |
| Computer use | LLM-driven system that **interprets user commands**, "looks at" the user's desktop or browser, and **autonomously chains clicks, keystrokes, shell commands, and API calls** to complete the task |
| Customer support | LLM-driven system that **turns live speech into intent and speaks back** in real time, while **chaining required app, CRM or API calls** to complete the task |

Source: Company Website

# Coding agents: LLM-driven systems are increasingly taking a more autonomous role in coding, going far beyond code autocompletion and chatbot assistance

| | 2021 | 2023 | Early 2025 |
|---|---|---|---|
| | **Coding Autocomplete** | **Chatbot Assisted Coding** | **Coding Agents** |
| | Predicts the next few lines of code based on the code context | Responds to prompts by generating coding within the existing file only | Operates autonomously to browse the codebase, create new files and run commands |
| **User Input** | *"def hash_password(password: str) -> str:"* | *"Hi, please help me write me a function to convert a password to a hash"* | *"Create a preliminary authentication system using OAuth"* |
| | **Predicts the next set of coding tokens** based on what the **user has already started typing** or the tokens previously predicted, providing suggestions directly in the editor | **LLM interprets the user's prompts** and outputs tokens of code based on the user's prompt and files pasted into the model's context | Agentic system **driven by LLMs** that interprets the users prompt and interacts with available tools to complete the task<br>*Example tools*<br>**Browse directories**<br>**Read files**<br>**Run commands**<br>**Edit files** |
| **System Output** | `"""`<br>`Hashes a password using SHA-256.`<br>`…`<br>`Returns:`<br>`str: The hashed password.`<br>`"""`<br>`return hashlib.sha256(password.encode()).hexdigest()` | `Here's a secure and modern way to hash a password in Python using the bcrypt library …`<br>`def hash_password(password: str) -> str: …`<br>`password_bytes = password.encode('utf-8') …`<br>`return hashed.decode('utf-8')` | `Great! The implementation is now complete. Here's where I added new files:`<br>`app/`<br>`├── api/`<br>`└── auth/ …`<br>`What would you like me to do next?` |
| **Example Apps**[1] | GitHub Copilot | OpenAI ChatGPT | CURSOR |

1. GitHub Copilot launched in late 2021 as an AI autocomplete tool build on OpenAI's Codex model; GitHub Copilot now supports chat and agentic features alongside AI autocomplete

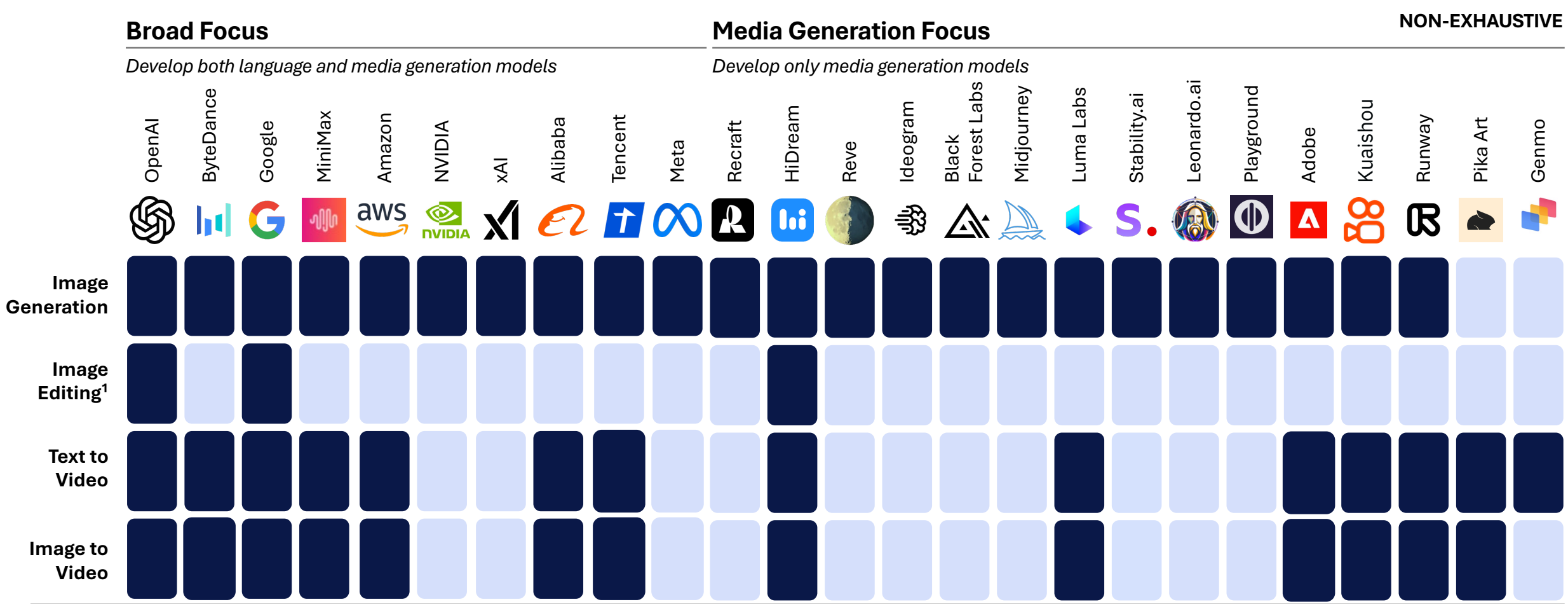
# Both players with a broad AI focus, and those with a specific focus in media generation, have developed models across modalities in the last year

## Key players offering image and/or video models

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

No model | Existing model

NON-EXHAUSTIVE

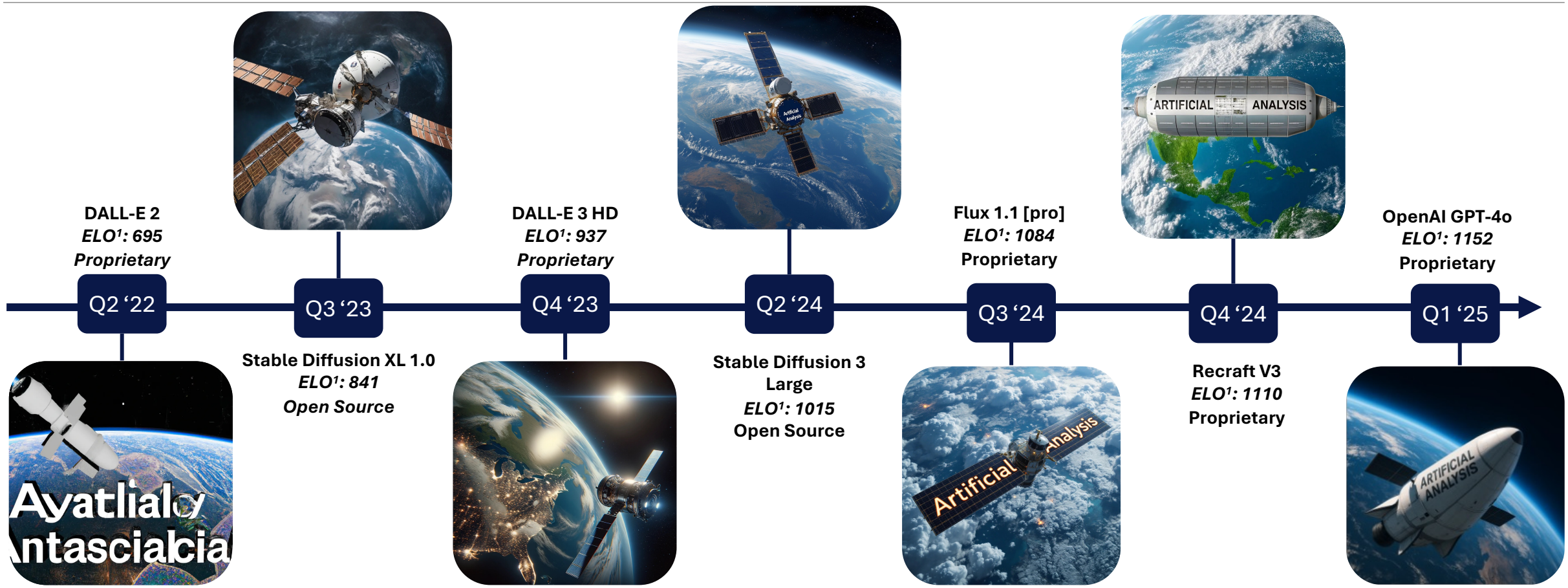



1. Instruction-based image editing

# In Q1 ‘25, image generation reached new heights, as OpenAI’s GPT-4o set a new benchmark for visual quality and prompt adherence

## Image Generation Quality[1], Over Time

*Prompt: The words 'Artificial Analysis' emblazoned on a next-generation spacecraft orbiting a breathtaking Earth view*

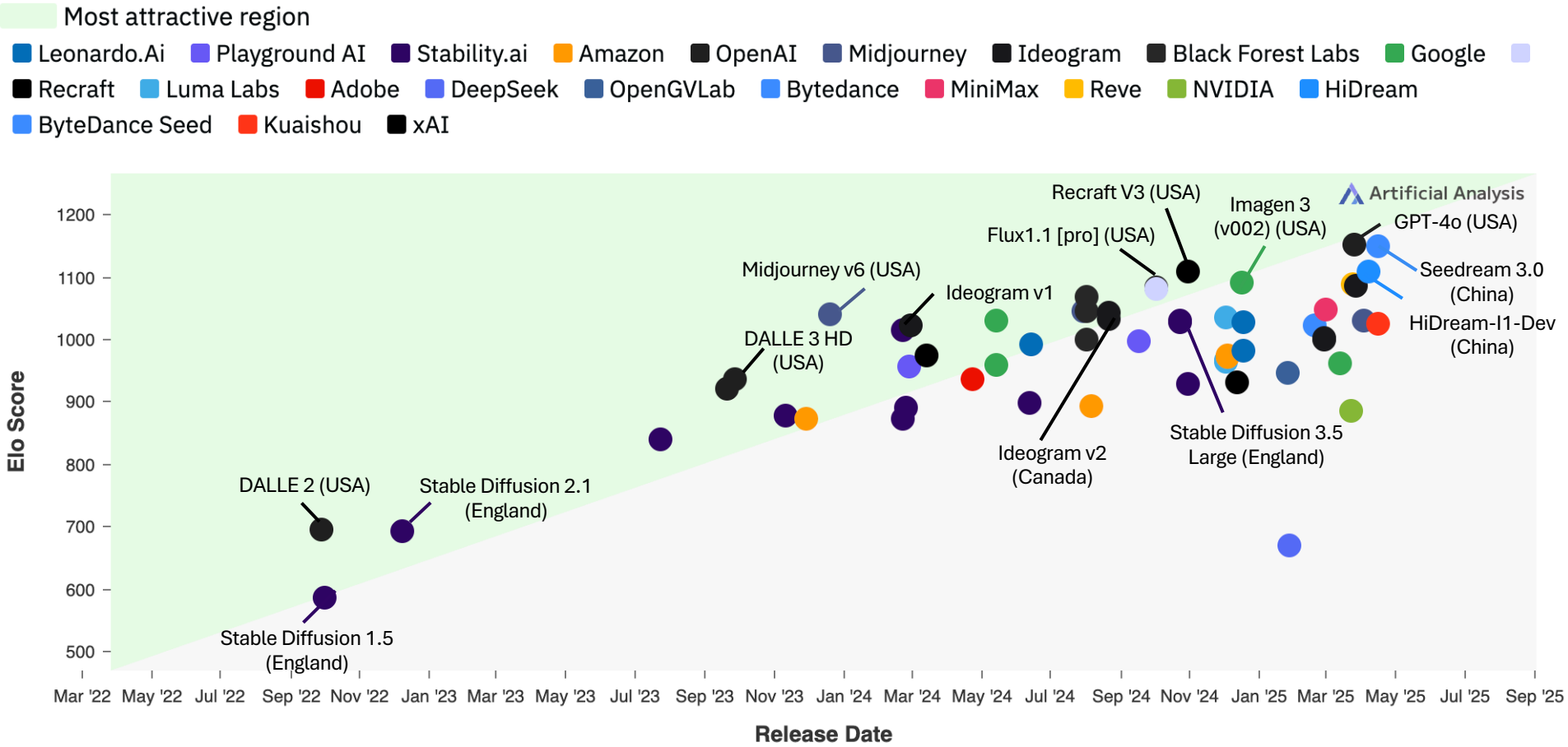



1. Artificial Analysis Image Arena ELO, calculated as of 6 May 2025 based on 1.6 million votes from Artificial Analysis users

# Image generation is entering a new phase of competitiveness with both generalist AI and media-focused labs competing for the frontier with frequent releases

## Text to Image ELO Score by Release Date

*Artificial Analysis Video Arena ELO Score*




## Commentary

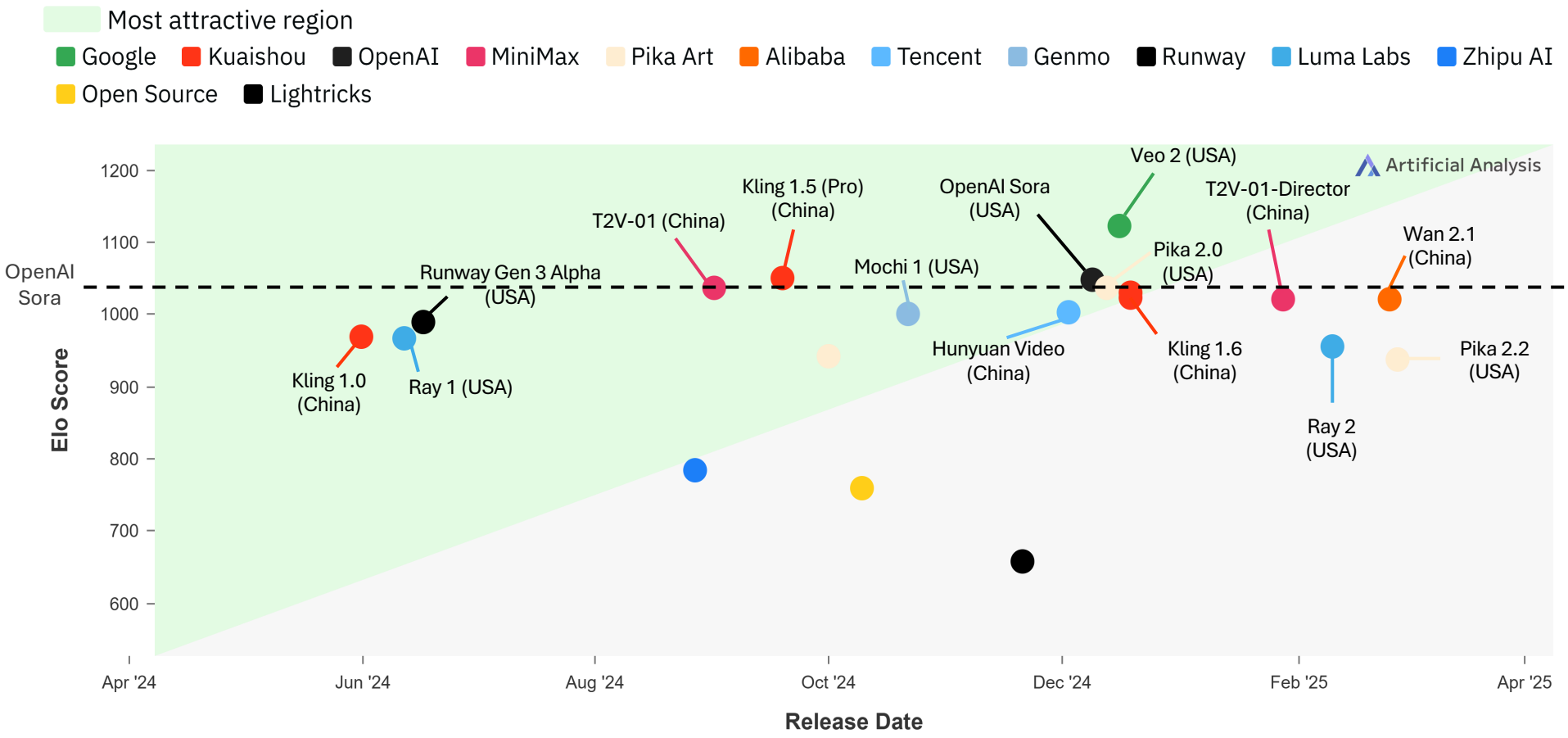
- **Incremental improvement in frontier image models:** Q1'25 saw the release of two new models (GPT-4o, Seedream 3.0) that surpassed the previous leading model (Recraft V3)
- **New challengers emerge:** ByteDance's Seedream 3.0 and HiDream-I1-Dev from China-based AI Labs rapidly emerged as leaders, challenging market incumbents including OpenAI and Google
- **Competitive landscape:** Growing number of labs released models, signaling rising interest and intensifying competition in image generation

Source: Artificial Analysis Video Arena

# Video models saw a breakthrough in quality, with a number of labs catching up to OpenAI's Sora's capabilities, including Google overtaking Sora with Veo 2

## Text to Video ELO Score by Release Date

*Artificial Analysis Video Arena ELO Score*




## Commentary

- **Video models saw a leap forward:** A number of video models have been released with comparable quality to OpenAI's Sora
- **U.S. and China dominate the frontier:** Top-performing models like Google's Veo 2 (USA) and MiniMax's T2V-01 Director (China) show a geographic duopoly at the top of the ELO leaderboard
- **Emerging labs lead the wave:** The majority of recent releases (e.g. T2V-01 Director by MiniMax, Kling 1.6 by Kuaisho) have come from emerging labs

Source: Artificial Analysis Video Arena

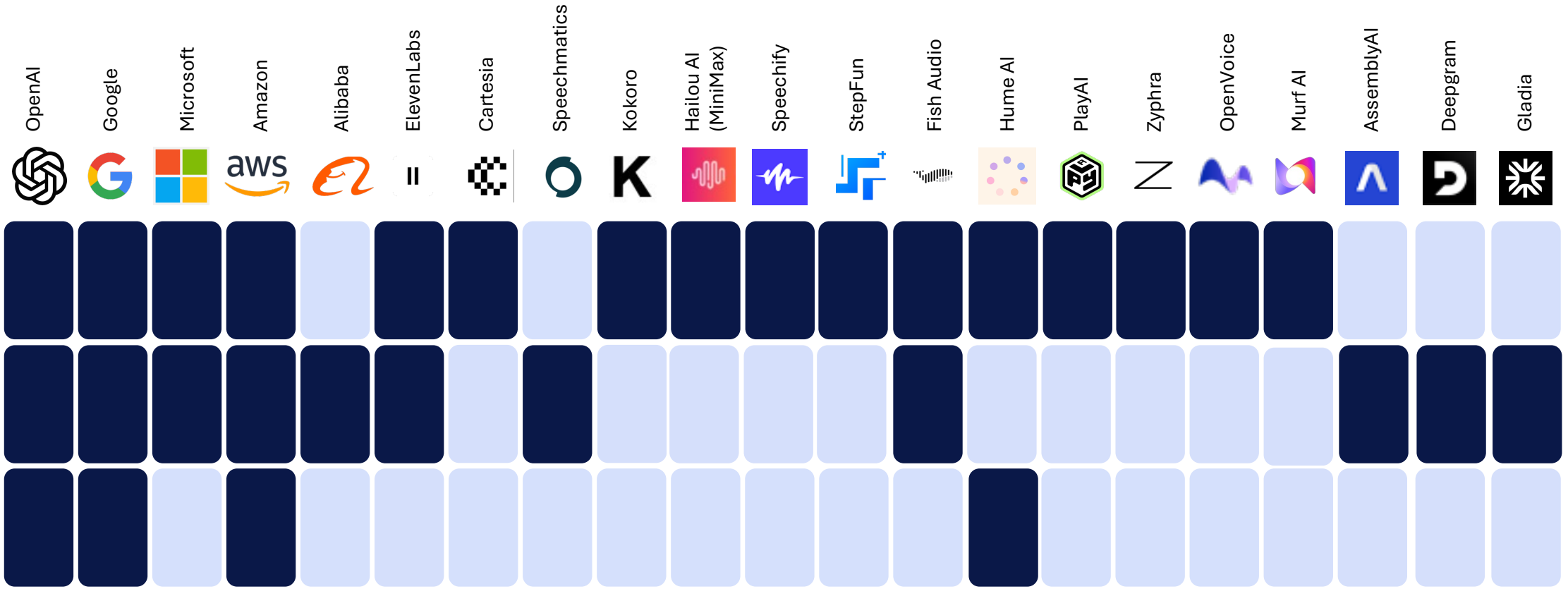
# While the generalist AI labs, including OpenAI, Google, Microsoft and Amazon, have offerings across speech AI, an emerging group of pure-play speech companies are driving innovation

## Key players offering speech models

*Classifications are indicative and determined based on models available*

Low or no presence | Strong presence

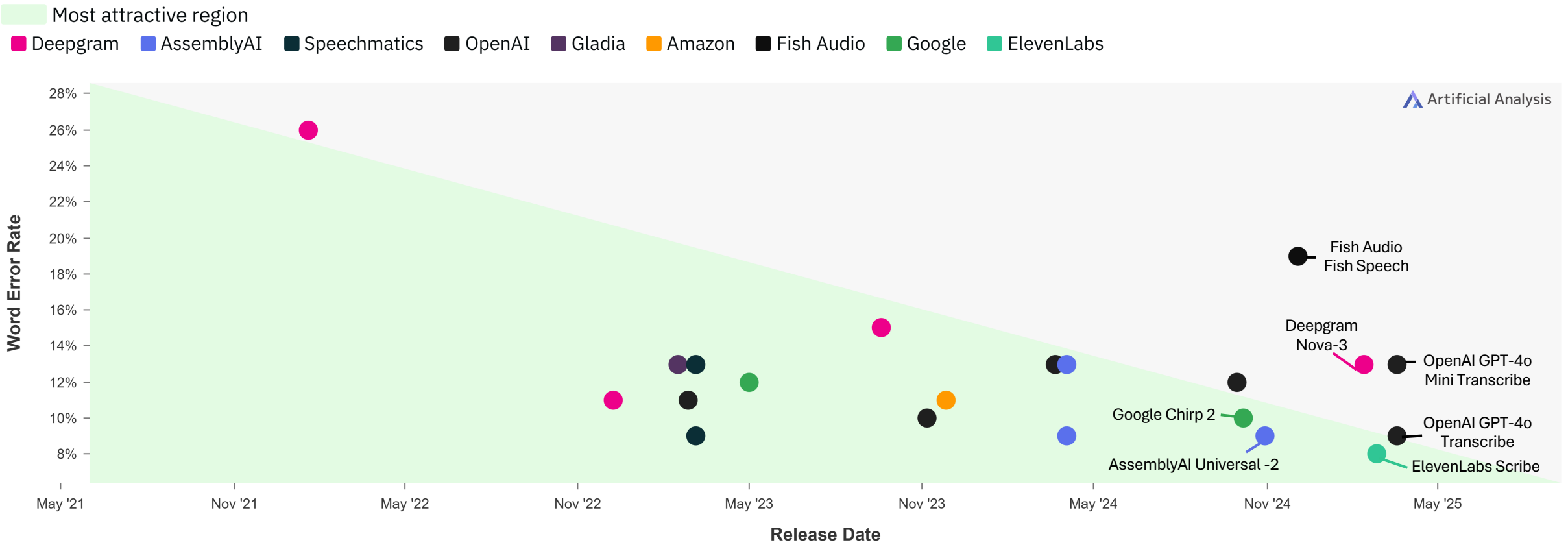
**Generalist Labs** | **Speech Focused AI Labs** | **NON-EXHAUSTIVE**



# Consistent gains in Speech to Text quality have continued as new model releases by OpenAI and ElevenLabs pushed frontier accuracy even higher, achieving lower error rates

## Frontier Speech to Text Model Quality by Country over Time

*% Word Error Rate*



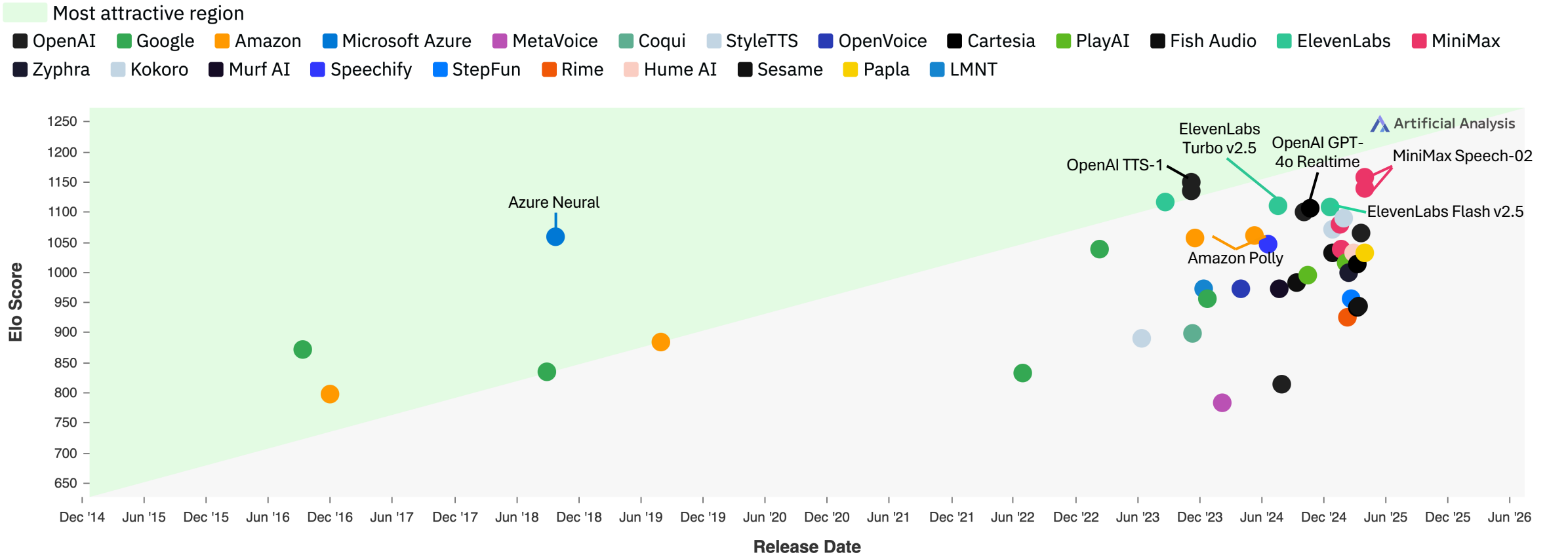

- **Consistent moderate growth** across labs have seen the Word Error Rates decrease over time as models continue to improve in quality; Q1 '25 seeing frontier Word Error Rate dropping from ~9% to ~8%
- **Scribe by ElevenLabs took the lead** as the most accurate Speech to Text model in Q1 '25, outperforming OpenAI's Whisper and GPT-4o Transcribe

Source: Artificial Analysis independent benchmarking

# Q1 ‘25 saw the release of new Text to Speech models that have continued to incrementally improve quality, including in pushing toward human-like dialogue

## Frontier Text to Speech Model Quality by Country, Over Time

*Elo Score based on Artificial Analysis Speech Arena*




- **Early 2025 saw the release of models such as MiniMax’s Speech-02-HD** that helped to improve sound quality, with improvements in speech audio that sounds more like human-like dialogue
- **The Text to Speech space has increased in competition in early 2025** as speech-focused labs and open source projects launched new models

Source: Artificial Analysis Speech Arena

# Artificial Analysis

**contact@artificialanalysis.ai**

**https://artificialanalysis.ai/**