

NEBIUS



TOKEN
FACTORY

Coding with Open Models & Coding Agents

Sujee Maniyam

Developer Advocate @ Nebius

sujee@nebius.com | [@sujee_dev](https://twitter.com/sujee_dev) | [W: sujee.dev](https://www.sujee.dev)

"Tell me a joke about Coding Agents"

I asked my coding agent for a small refactor.

It replied:

“Done. I migrated your monolith to Kubernetes.”

Audience Poll

What Coding agents
are you using now?



Nebius Token Factory

Inference at scale

Full-stack GenAI platform



Inference

- 60+ models • Dedicated endpoints • Fast and base flavors
- Batch API • Structured output • Function calling



Data Lab (Beta)

- Inference log import • SQL-like dataset filtering
- Dataset versioning • One-click export • Batch evaluation



Post-training

- LoRA and full fine-tuning • Model distillation • Custom spec. decoding training
- Quantization and calibration • Preference optimization (RL/GRPO)



Deployment

- Custom Weights Hub • GPU type selection • Region control
- Autoscaling config • Observability surfaces • FT dashboard

Security and compliance

Reliability and scale

Model ops and innovation

Rise of open models

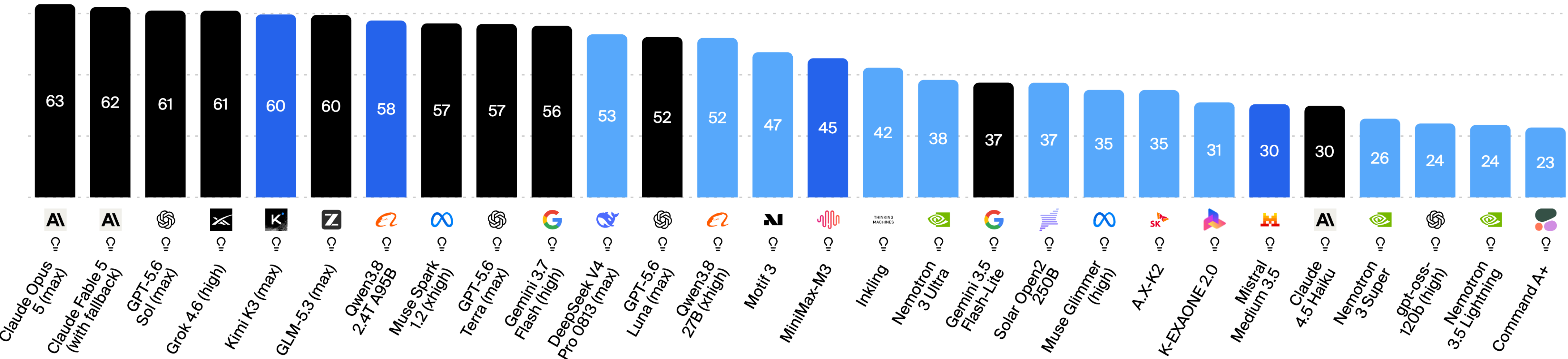
Open vs Closed Models

Artificial Analysis Intelligence Index by Open Weights / Proprietary

Artificial Analysis Intelligence Index v4.1.1 incorporates 9 evaluations: GDPval-AA v2, τ^3 -Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPt, AA-Omniscience, AA-LCR

● Proprietary ● Open Weights (Commercial Use Restricted) ● Open Weights

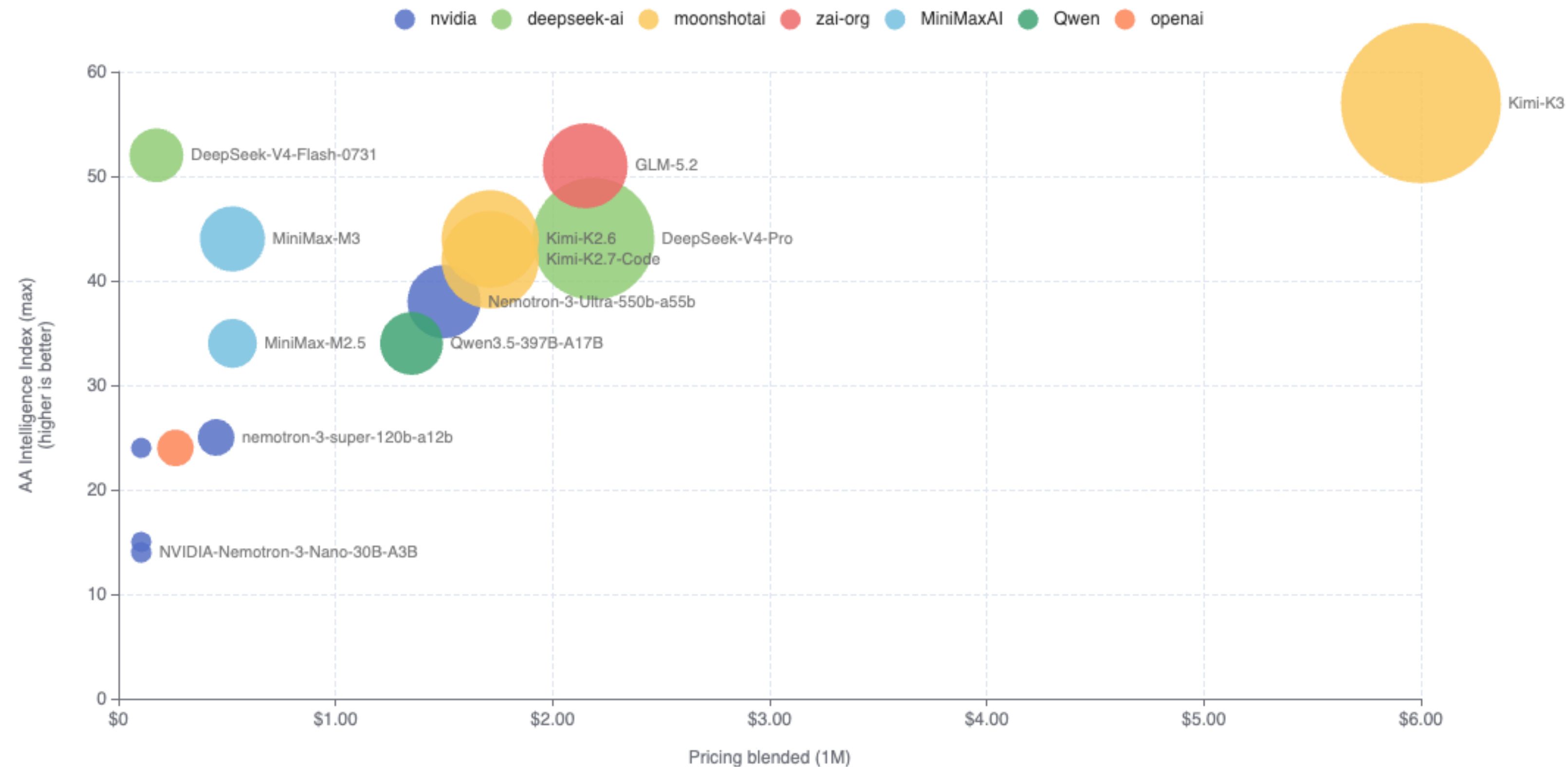
Artificial Analysis



💡 Reasoning models are indicated by a lightbulb icon

As of Aug 24, 2026

Open models on Nebius Token Factory



Try live:

Popular coding agents

Agent	By	Open
Github copilot	Github / Microsoft	NO
Claude Code	Anthropic	NO
Cursor	Cursor	NO
Codex	OpenAI	NO
OpenCode		YES
PyDEV		YES
Cline	Cline.bot	YES

What makes a good Coding Model

Checklist:

✓ Tool calling

✓ Long context

✓ Instruction following

✓ Code generation

✓ Patch editing

✓ Reliability

Benchmarks are not enough - Agent behavior matters

Coding Models



Proprietary Models

Anthropic models

Default for Claude code
+ many others

OpenAI models

Default by codex
+ many others

Gemini

Antigravity + others



Open Model

Deepseek

Qwen

Kimi

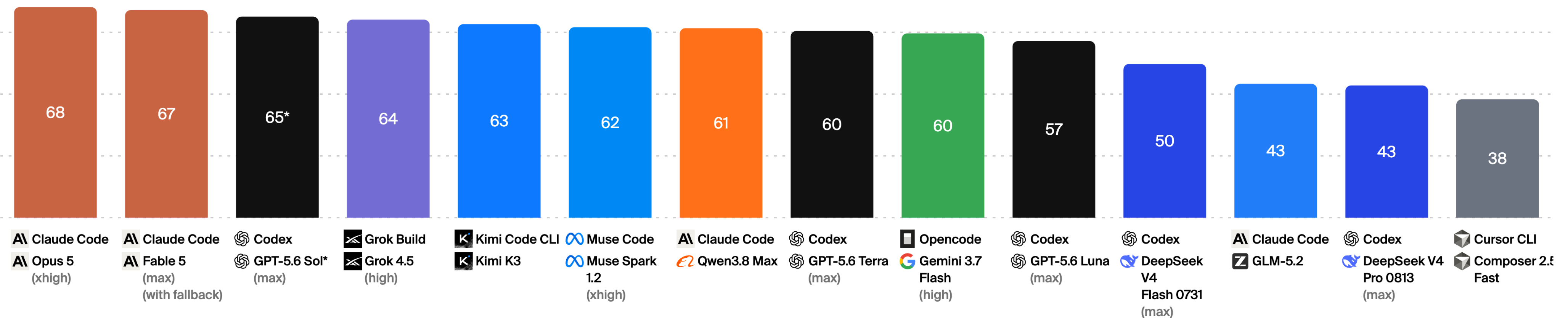
Minimax

Artificial Analysis Coding Agent Index (performance)

Artificial Analysis Coding Agent Index

Artificial Analysis Coding Agent Index v1.4 incorporates 3 benchmarks: DeepSWE, Terminal-Bench v2.1, and SWE-Atlas-QnA · Higher is better

Artificial Analysis

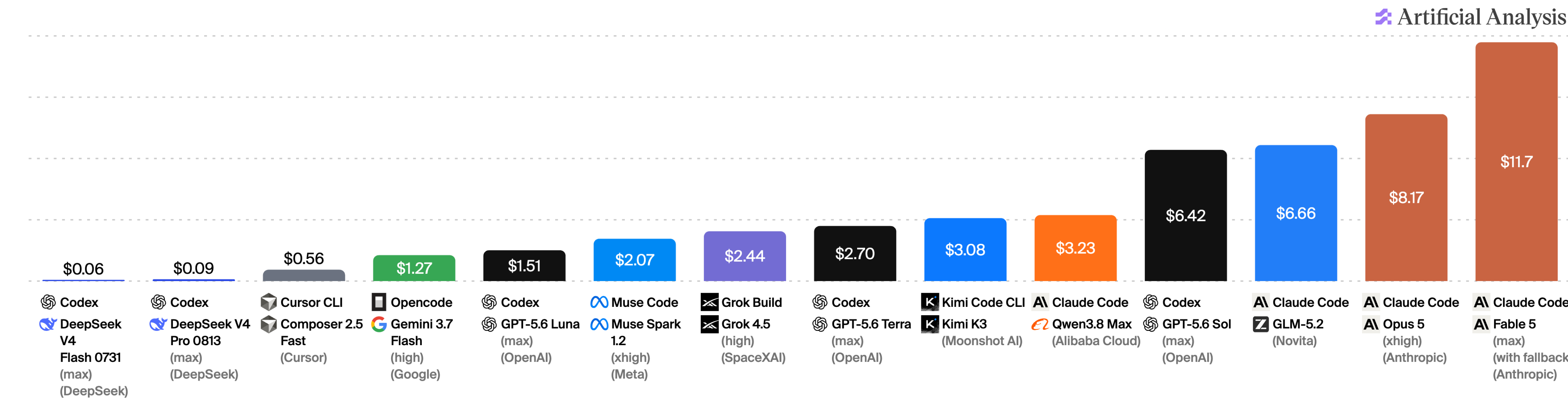


* Since benchmarking, we have observed a higher rate of content safety filtering on this endpoint.

Artificial Analysis Coding Agent Index (Cost)

Cost per Task

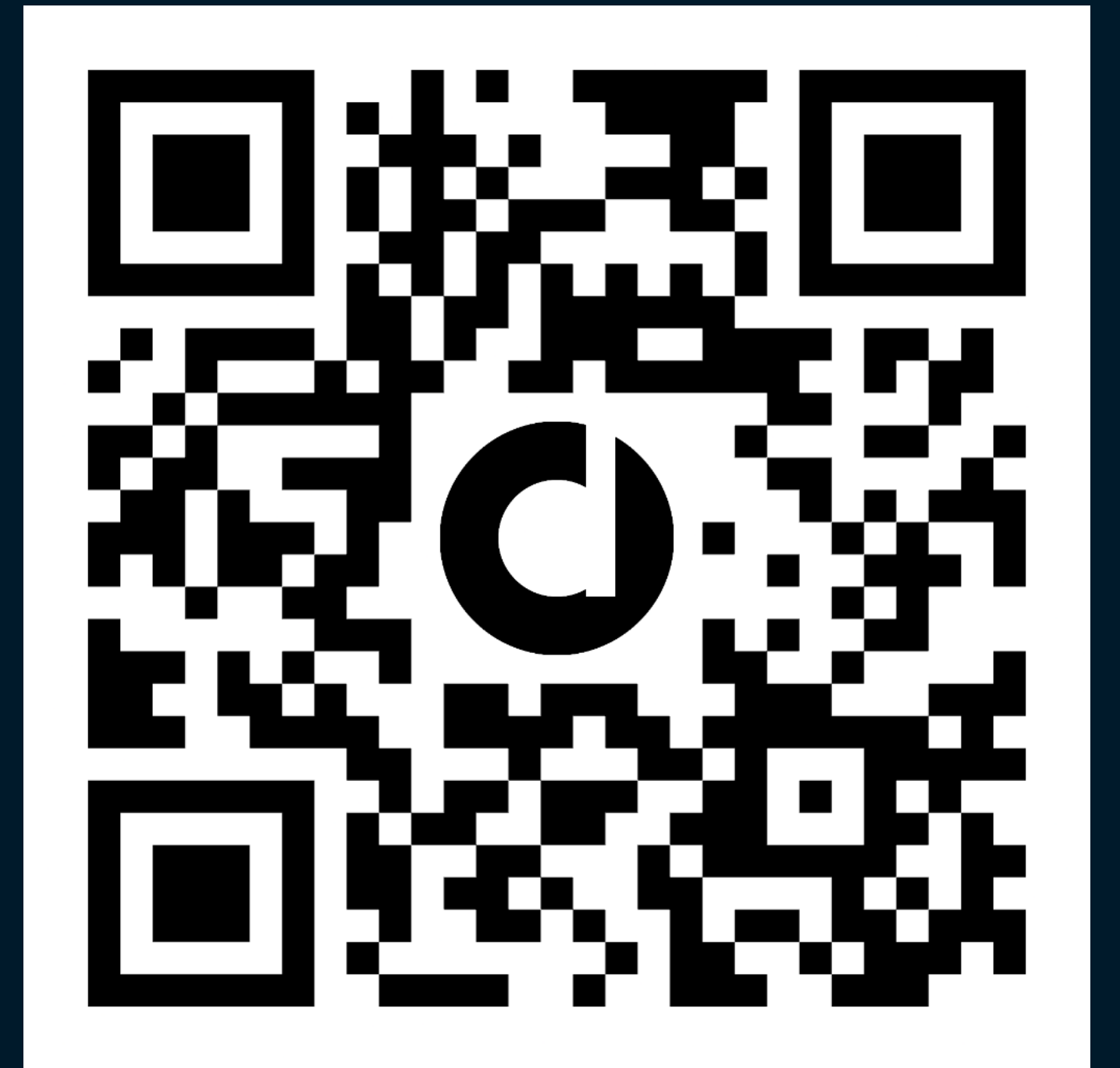
Average pay-per-token API cost per task (USD) · Lower is better



Demo

Getting Started

github.com/sujee/workshops



Thanks and Q&A

Token Factory

Tokenfactory.nebius.com

Token Factory Cookbook (Code + examples)

github.com/nebius/token-factory-cookbook

Nebius Discord (support)

discord.gg/AuQWb3nhM9

Sujee Maniyam Developer Advocate @ Nebius
sujee@nebius.com | [@sujee_dev](https://twitter.com/sujee_dev) | [W: sujee.dev](https://www.sujee.dev)