



Nebius Token Factory

Office hour #3

Where are you joining from?
What is the current weather ?

What are you
drinking?

Answer in chat

Your host



Sujee Maniyam

AI Developer Advocate
Nebius Token Factory

sujee@nebius.com | @sujee_dev

Housekeeping



**Stay engaged
throughout the
session**



**We will share
credits at the end!**



**Raise your hand
if you'd like to
speak or ask a
question**



**This session is
being recorded
and the recording
will be shared
afterward**

Ice breaker jokes

**Prompt: Tell me
a joke about Agents**



Ice breaker jokes

**Prompt: Tell me
a joke about Agents**

MiniMax-M3

Two AI agents walk into a bar.
One says, "I'll have what the other one's
temperature is set to."



Agenda

- 1 Nebius Token Factory and What's new**
- 2 Building Agents with Token Factory Models**
- 3 Questions + open chat**



Nebius Token Factory

Inference at Scale

**Great inference starts with
great models + infrastructure**

Built on a full-stack AI cloud

Nebius is designed specifically for AI workloads We build the full stack:

✓ hardware → software → cloud

Token Factory runs on top of this foundation

✓ We design the **data centers**, **optimize the GPUs**, and **orchestrate the entire inference stack**, so you can build, fine-tune, and distill open models through a single API.

Inference

Token Factory



Llama

S.



NVIDIA NIM Microservices

Orchestration

Managed Soperator

SkyPilot

MLFlow

Compute & GPU clusters

Virtual Machines

Containers

Managed Kubernetes

IaaS

Storage

Block volumes

Object storage

Shared Filesystem

WEKA

VPC networking

NVIDIA NDR/XDR InfiniBand

Load Balancer

VPC Routing

Hardware

NVIDIA GPU servers

GB300 NVL72

GB200 NVL72

HGX B300

HGX B200

HGX H200

HGX H100

CPU-only servers

Intel

AMD

Storage servers

AMD

The full-stack AI platform

Inference



Finetuning

WB & MLFlow integration

Platform

- API
- Role Based Access
- Usage observability
- Post paid invoicing
- Zero Data Retention

Integrations

- MCP
- Tool-calling
- NVIDIA NIMs

Compliance

- ISO 27001
- SOC2
- HIIPAA
- DORA

Hardware

- GB200
- B200
- H200
- H100

Support

- Solution Architect
- Documentation & Cookbook
- Popular frameworks integrations

One API. All of AI.

What we handle so you don't have to

We don't just host models. We help you engineer AI systems.

Optimization layer

Cache-aware routing → 90% KV cache hit rates in multi-turn workloads

Prefill/decode disaggregation → optimized TTFT for long-context

KV cache offloading → up to 50% cost savings on chat workloads

Speculative decoding → ~30% performance boost (higher with custom data)

Custom forks on TensorRT-LLM / vLLM / SGLang — engine selection per workload

Workload-aware autoscaling with real-time fleet metrics

We optimize latency, throughput and cost at the engine level.

AI cloud infrastructure

Dedicated NVIDIA H100 / H200 / B200 / B300 clusters

Multi-region inference (EU and US-owned data centers)

99.9% SLA with proactive system health monitoring

800 Gbps InfiniBand (Quantum-X800) interconnect

Bare-metal performance, not shared cloud instances

You control behavior. We run the infrastructure.

Enterprise foundation

SOC 2 Type II + HIPAA + ISO 27001 certified

RBAC + SSO for team access control

Zero data retention (data never trains our models)

Project separation and unified enterprise billing

OpenAI-compatible API for seamless migration

Production-ready from day one.

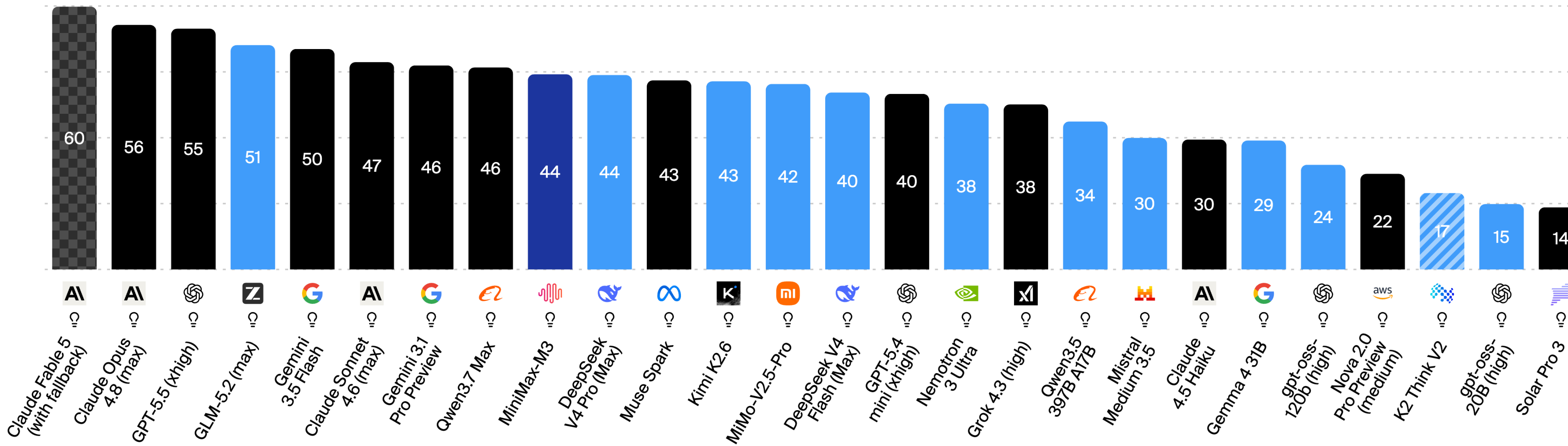
Open Models and Closed Models

Artificial Analysis Intelligence Index by Open Weights / Proprietary

Artificial Analysis Intelligence Index v4.1 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

- Not currently available ■ Estimate (independent evaluation forthcoming)
- Proprietary ● Open Weights ● Open Weights (Commercial Use Restricted)

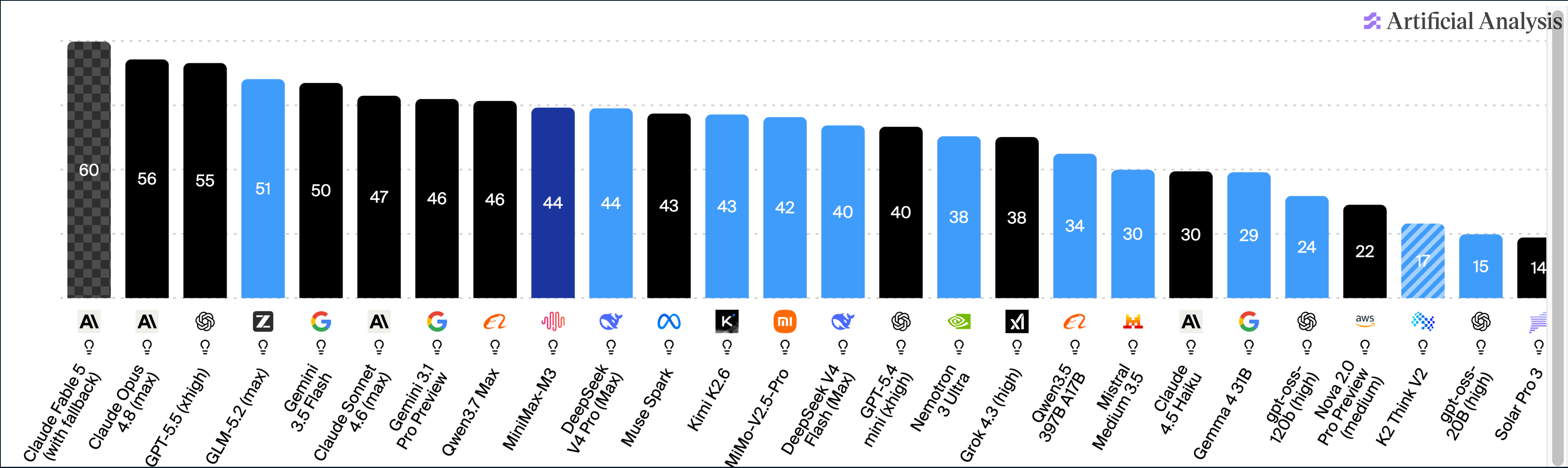
Artificial Analysis



💡 Reasoning models are indicated by a lightbulb icon

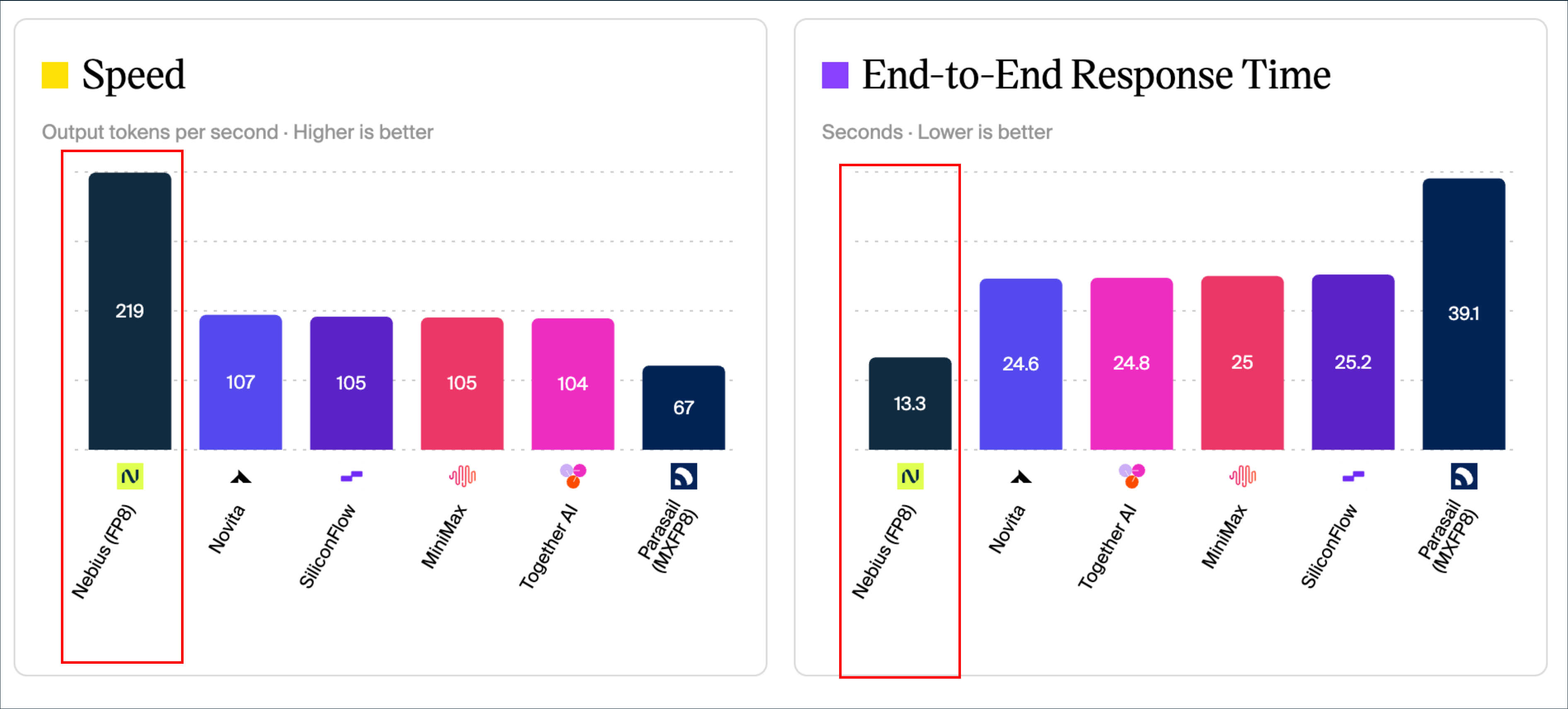
SOTA Open Models on Token Factory

Model	AA Intelligence Score ↓	Size	Context Window
GLM 5.2	51	735 B	1 M
Minimax M3	44	428 B	1 M
Kimi 2.7 Code	42	1000 B (1 T)	256 k
Nemotron-3-Ultra-550b	38	550 B	262 k



Focus on Performance

Minimax M3 performance on Artificial Analysis



Focus on Performance

Nemotron 3 Ultra 550B performance on Artificial Analysis



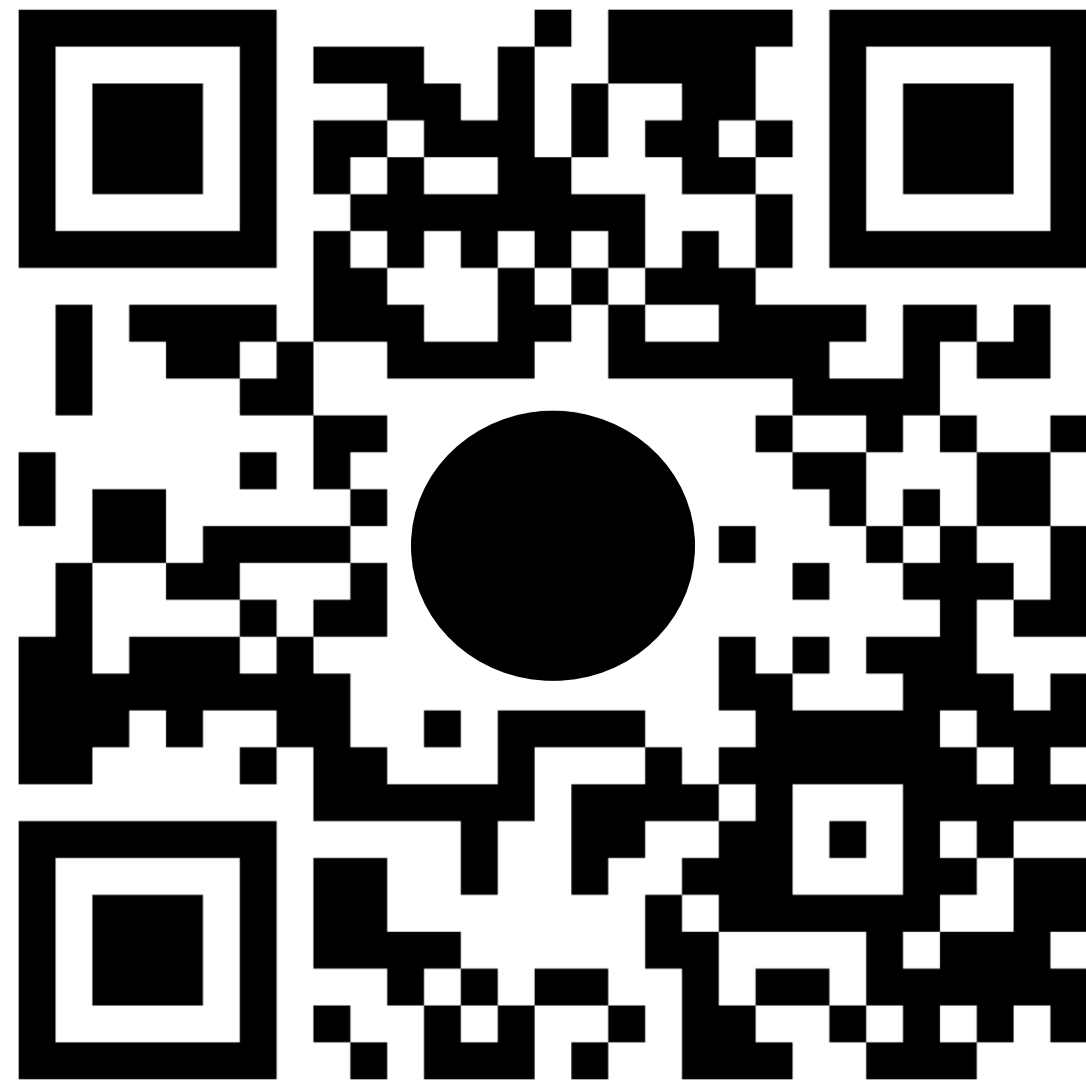
NEBIUS



TOKEN
FACTORY

Community Updates

Join the Nebius AI Builder Program



Join for
free today!

dub.sh/1XRTB58

\$25 Nebius Token Factory Credit

\$25 Tavily credit

Nebius certification for \$1

Office hours & community events

Nebius community Discord

Nebius Fellow Program

What it is

A curated program that connects high-signal AI builders, creators, and community leaders with Nebius resources, opportunities, and a global builder network.

Who we target

- AI builders & startup founders
- Open-source contributors
- Technical creators & community leaders

What they do

- Build and showcase projects
- Share knowledge through content/events
- Give product feedback from real use cases

What they get

- Nebius credits and technical resources
- Visibility across Nebius channels/events
- Direct access to the DevRel team and Fellow network

Goal:

Support Fellows accelerate their journey through dedicated resources, exclusive access, meaningful visibility, and a trusted global community.

Meet the First Nebius Fellows

Introducing our founding cohort of AI builders, creators, and technical leaders.



Raphael Gutsche

Founder, AI Agents Summit
Berlin, Germany



Kosseila Hd

Multi-Cloud DevOps for AI
Toronto, Canada



Maroon Ayoud

Research Scientist & Architect,
RedHat
Haifa, Israel



Linda Haviv

Founder, LindaV Media Labs
NewYork, USA



Ayub Yanturin

Chair, Innovators Guild
London, UK



Rayyan Zahid

Immersive Systems Wizard
San Francisco, USA



Mesut Oezdil

HAMi Member, Open Source
Contributor
Stuttgart, Germany



Yong Quan Tan

AI Builder, Kairos
Singapore



Vivek Haldar

Agentic AI Expert, Founder,
Advisor
Los Angeles, USA



Dhruv Diddi

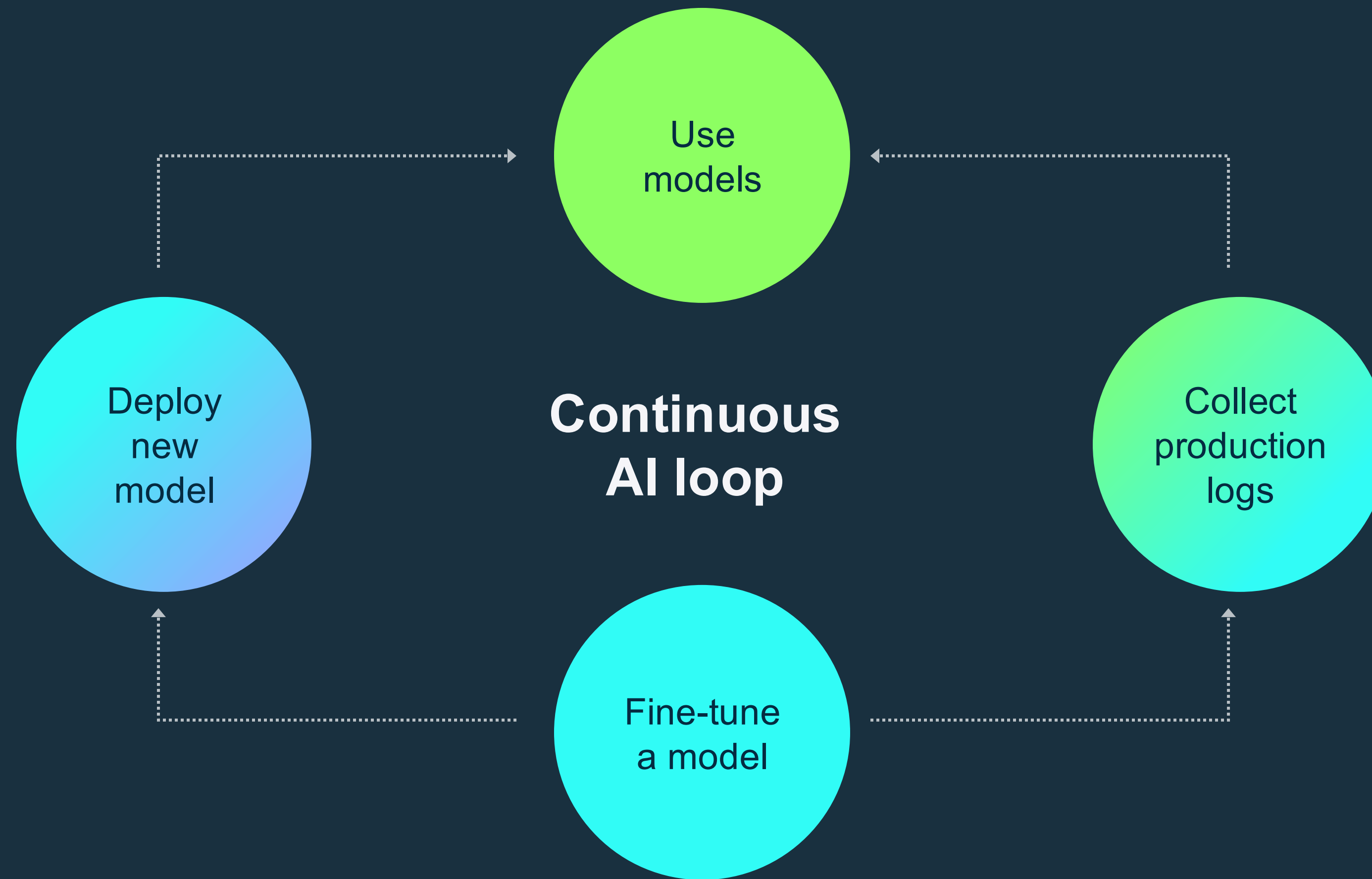
Founder, Solo Tech
San Francisco, USA



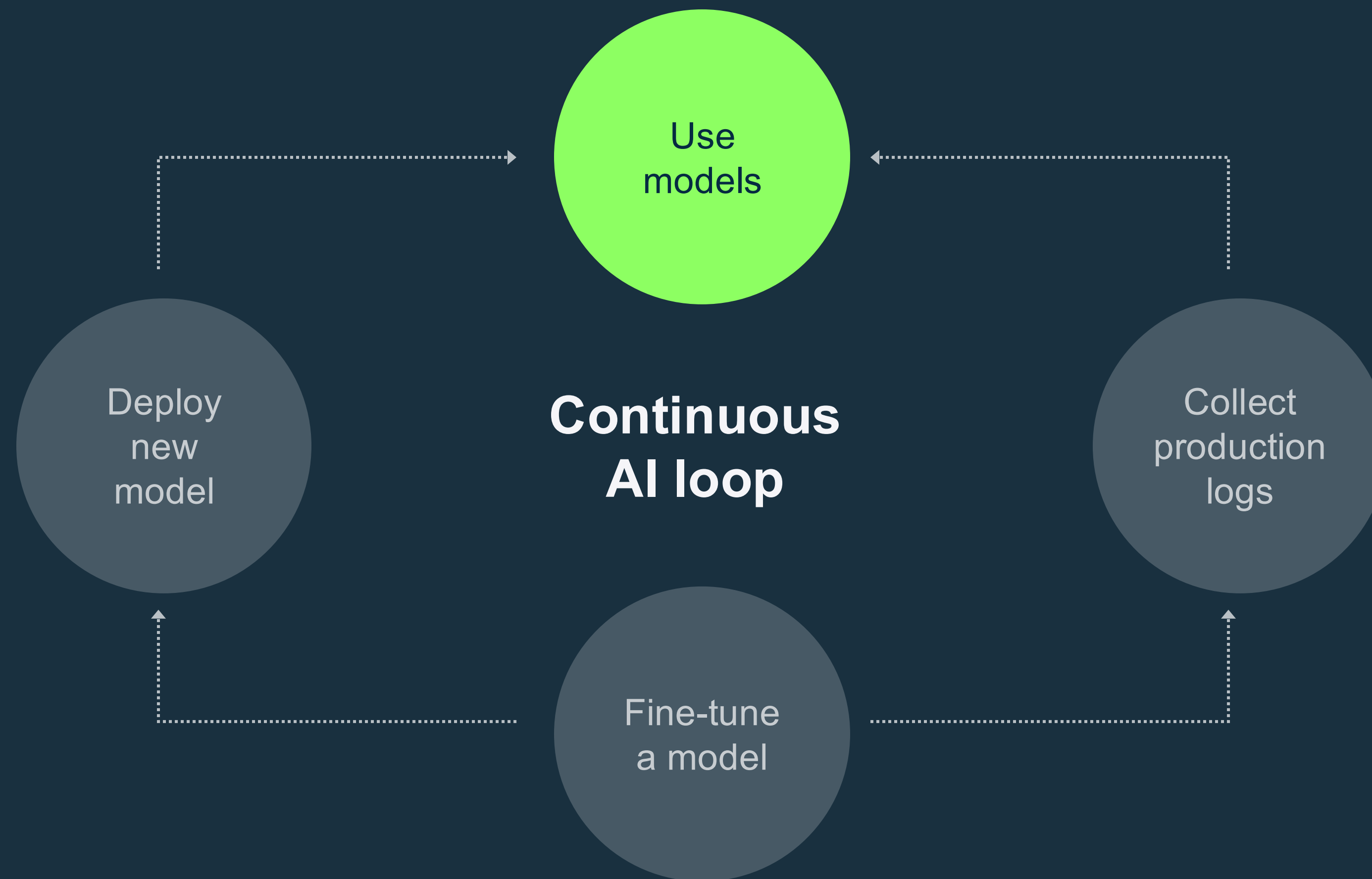
Nebius Token Factory

Full Stack AI Platform

Continuous AI Loop



Using Open Models



Enterprise-grade production inference of open-source models

Nebius Token Factory

NEBIUS

TOKEN FACTORY

default-project

Contact us

Documentation

Workspace

Explore

Model catalog

Inference

Post-training

Data Lab

Sandboxes

API keys

Settings

Featured models

Handpicked models for reasoning, prototyping, and developer workflows.

New

Minimax

MiniMax-M3

\$0.30 / 1M In | \$1.20 / 1M Out | 190 Tok/s

Baseus-central1Text-to-text

MiniMax-M3 is a 428B MoE reasoning model with 1M context, served on B200 via vLLM with EAGLE3 speculative decoding.

New

Moonshot AI

Kimi-K2.7-Code

\$0.95 / 1M In | \$4.00 / 1M Out | 231.26 Tok/s

Baseus-central1Text-to-text

Open-source code-focused reasoning model built for long-context software engineering, tool use, and agentic coding workflows.

New

Z.ai

GLM-5.2

\$1.40 / 1M In | \$4.40 / 1M Out | 25 Tok/s

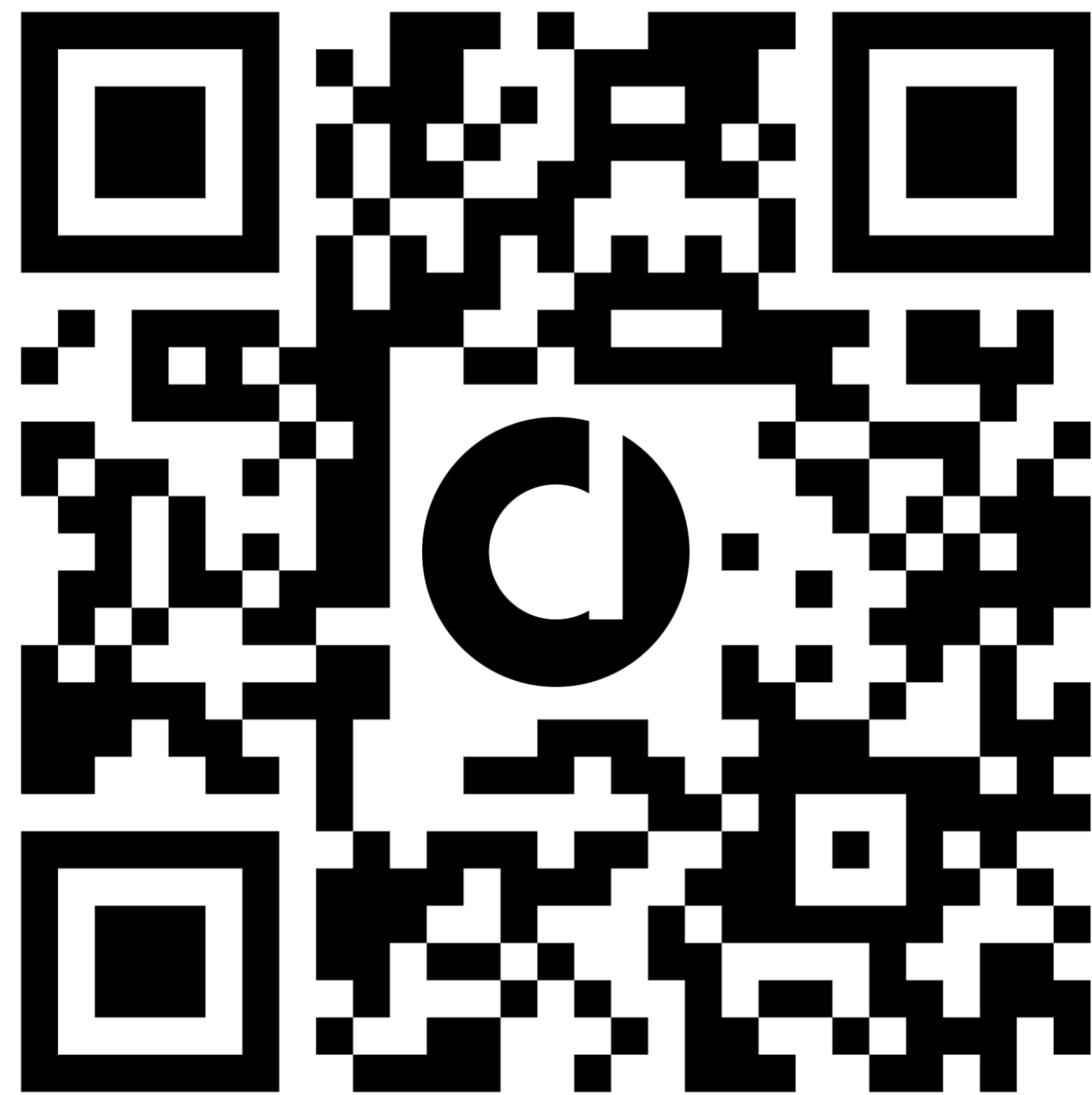
Baseeu-north1Text-to-text

Zhipu AI's latest flagship multimodal model with strong bilingual (Chinese-English) reasoning, long-context understanding, advanced tool use, and agent-oriented capabilities.

Demo: Open Models @ Token Factory



Getting started



Token Factory
Credits 🎁

dub.sh/GQ4iGJC

Nebius Token Factory
Tokenfactory.nebius.com

Token Factory Cookbook
github.com/nebius/token-factory-cookbook

Nebius Discord (support)
discord.gg/AuQWb3nhM9

Join our next Office Hours



**August
Office Hour**

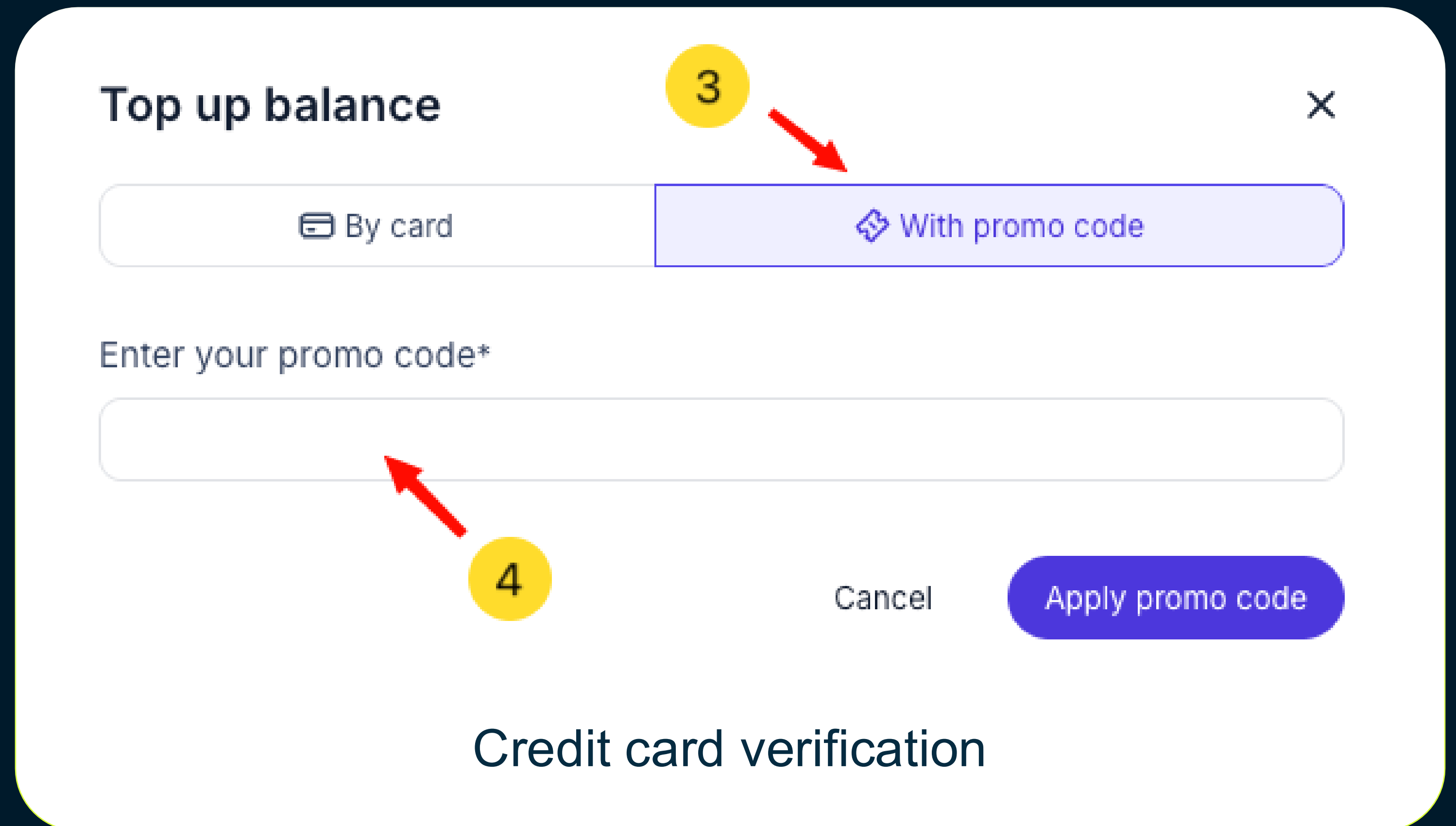
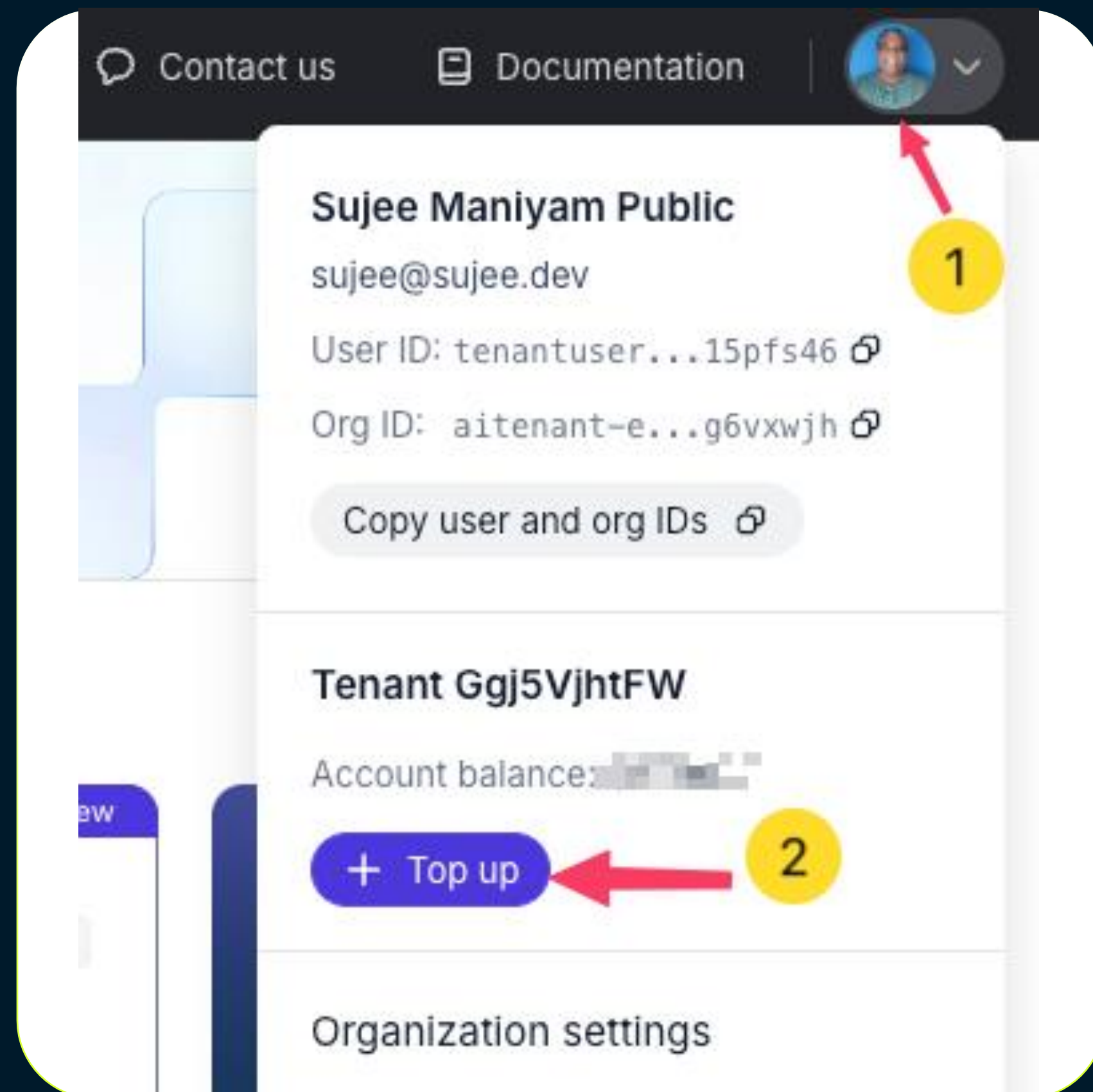
Nebius Token Factory
Tokenfactory.nebius.com

Token Factory Cookbook
github.com/nebius/token-factory-cookbook

Nebius Discord (support)
discord.gg/AuQWb3nhM9

How to claim the credits

Go to tokenfactory.nebius.com





Q&A



Sujee Maniyam
AI Developer Advocate
Nebius Token Factory
sujee@nebius.com | @sujee_dev

NEBIUS



**TOKEN
FACTORY**