

Nebius launches Nebius Token Factory to deliver production AI inference at scale

- Platform enables production inference using open-source models on Nebius's dedicated, high-capacity AI infrastructure
- Brings the full model lifecycle from fine-tuning to deployment together into a single governed platform with enterprise-grade security
- Teams report major gains in speed, efficiency and performance from using Nebius Token Factory to power real-world AI solutions

Amsterdam, November 5, 2025 — Nebius today unveiled Nebius Token Factory, a production inference platform that enables vertical AI companies and digital enterprises to deploy and optimize open-source and custom models at scale and with enterprise-grade reliability and control.

Built on Nebius's full-stack AI infrastructure, Nebius Token Factory brings together high-performance inference, post-training and fine-grained access management into a single governed platform. It supports all major open models, including DeepSeek, GPT-OSS by OpenAI, Llama, NVIDIA Nemotron and Qwen, and also offers customers the option to host their own models.

As AI moves from experimentation to production, relying on closed models can create scaling bottlenecks. Open-source and custom models can remove those barriers, unlocking both innovation and better economics, but managing and securing them in production has remained complex and resource-intensive for most teams.

Nebius Token Factory empowers teams to realize these advantages by combining the flexibility of open models with the governance, performance and cost-efficiency needed to run AI at scale. It is optimized for efficiency, delivering sub-second latency, autoscaling throughput and 99.9% uptime, even for workloads exceeding hundreds of millions of requests per minute.

“Every team has unique requirements, and they want speed, reliability and cost efficiency without heavy lifting,” said Roman Chernin, co-founder and Chief Business Officer of Nebius. “We built Nebius Token Factory not just to serve models, but to help customers solve real challenges and engineer for scale — optimizing inference pipelines and turning open models into production-ready systems.”

How customers and the community are using Nebius Token Factory

Early adopters of Nebius Token Factory are leveraging the platform to power a wide range of AI solutions from intelligent chatbots and coding copilots to high-performance search, retrieval-augment generation (RAG), document intelligence and automated customer support.

Prosus, the power behind some of the world’s leading lifestyle and e-commerce brands, has achieved up to 26x cost reductions compared to proprietary models.

“We move fast, test and iterate quickly, and the flexibility, products and quick responses from Nebius Token Factory allowed us to keep this pace all the way through production,” said Zülküf Genç, Director of AI at Prosus. “By leveraging Nebius Token Factory’s dedicated endpoints, Prosus was able to secure guaranteed performance and isolation. The addition of autoscaling was the game-changer, allowing us to handle massive workloads of up to 200 billion tokens per day without manual intervention.”

Leading AI video platform **Higgsfield AI** relies on Nebius for on-demand and autoscaling inference.

“Running inference at scale with healthy economics requires efficient on-demand and autoscaling capabilities. Nebius was the only provider that met our requirements — reducing overhead, simplifying management, and enabling us to deliver faster, more cost-efficient AI in production,” said Alex Mashrabov, Founder and CEO at Higgsfield AI.

Open-source leaders like **Hugging Face** are also collaborating with Nebius to improve access and scalability for developers.

“Hugging Face and Nebius share the same mission of making open AI accessible and scalable. By partnering with Nebius Token Factory, we’ve been able to provide faster and more reliable inference for developers building on large open-source models,” said Julien Chaumond, CTO at Hugging Face.

Full-stack AI infrastructure as the foundation

Nebius Token Factory is built on top of [Nebius AI Cloud 3.0 “Aether”](#). This ensures enterprise-grade security, proactive monitoring and consistent performance, validated by benchmarks including [MLPerf® Inference](#). By pairing Nebius’s full-stack infrastructure with a tech stack optimized for inference, Nebius Token Factory helps customers scale their AI applications and solutions faster.

“At SemiAnalysis, we track total cost of ownership for every single GPU Cloud player. Nebius is the only neocloud that uses custom ODM chassis, which translates to massively lower total cost of ownership. We are excited to see their new Inference platform engineered around the tradeoff triangle: cost, output speed per user and model quality,” said Dylan Patel, Chief Analyst and SemiAnalysis.

AI projects often scale faster than the teams around them. Nebius Token Factory streamlines the post-training lifecycle, turning open-source model weights into optimized, production-ready systems with guaranteed performance and transparent cost per token. Integrated fine-tuning and distillation pipelines allow teams to adapt large open models to their own data while cutting inference costs and latency by up to 70%.

Optimized models can be deployed to production endpoints instantly, without manual infrastructure setup. This approach allows AI builders and enterprises to iterate faster, manage costs predictably and maintain full transparency over every token served.

Token Factory introduces Teams and Access Management, Single Sign-On (SSO), project separation and enterprise-focused billing to simplify collaboration and ensure compliance. Administrators can set granular roles, enforce least-privilege access and maintain clear audit trails across all deployments, from early experimentation to mission-critical workloads.

Nebius Token Factory — key features

- **Dedicated endpoints with guaranteed performance and isolation**, with 99.9 % SLA, predictable latency, and autoscaling throughput
- **Zero-retention inference** in EU or US datacenters, supporting strict data-residency requirements
- **Security certifications** including SOC 2 Type II including HIPAA, ISO 27001 and ISO 27799 certifications
- **Comprehensive fine-tuning capabilities** supporting both LoRA and full model training, with seamless one-click deployment and hosting.

- **Support for 40+ open-source models**, including the latest Deep Seek, Llama, OpenAI and Qwen, optimized for the latest chips
- **Governance by design**, with Team and Access management, SSO, unified billing and audit-friendly workspaces
- **OpenAI-compatible APIs** for seamless migration from proprietary endpoints

Availability

Nebius Token Factory is the next evolution of Nebius AI Studio, redesigned for enterprise readiness and full model-lifecycle management. It's available today, supporting over 60 open-source models across text, code, and vision. Current AI Studio users will upgrade automatically to Token Factory.

Visit <http://tokenfactory.nebius.com/> to get started.

About Nebius

Nebius is a technology company building full-stack cloud infrastructure for the global AI industry. Headquartered in Amsterdam and listed on Nasdaq (NASDAQ: NBIS), the company has a global footprint with R&D hubs across Europe, North America, and Israel.

Nebius AI Cloud has been built from the ground up for intensive AI workloads. With proprietary software and hardware designed in-house, Nebius AI Cloud gives AI builders the compute, storage, managed services, and tools they need to build, tune, and run their models.

Contacts

Investor Relations: askIR@nebius.com

Media Relations: media@nebius.com

Disclaimer

Forward Looking Statements

This press release contains forward-looking statements within the meaning of the Private Securities Litigation Reform Act of 1995, which involve risks and uncertainties. All statements contained in this press release other than statements of historical fact, including, without limitation, statements regarding the anticipated technical performance, market adoption and commercial prospects of Token Factory, are forward-looking statements. The words “anticipate,” “believe,” “continue,” “estimate,” “expect,” “guide,” “intend,” “likely,” “may,” “will” and similar expressions and their negatives are intended to identify forward-looking statements.

These forward-looking statements are subject to risks, uncertainties and assumptions, some of which are beyond our control. Actual results may differ materially from the results predicted or implied by such statements, and our reported results should not be considered as an indication of future performance. The potential risks and uncertainties that could cause actual results to differ from the results predicted or implied by such statements include, among others: market, macroeconomic and geopolitical conditions; ; competitive pressures; technological developments; our ability to secure and retain clients; ; unpredictable sales cycles; and potential pricing pressures; as well as those risks and uncertainties related to our continuing businesses included under the captions “Risk Factors” and “Operating and Financial Review and Prospects” in our Annual Report on Form 20-F for the year ended December 31, 2024, filed with the Securities and Exchange Commission (“SEC”) on April 30, 2025.

All information in this press release is as of the date hereof (unless stated otherwise). Except as required by law, we undertake no obligation to update or revise publicly any forward-looking statements, whether as a result of new information, future events or otherwise, after the date on which the statements are made or to reflect the occurrence of unanticipated events.

In addition, statements that “we believe” and similar statements reflect our beliefs and opinions on the relevant subject. These statements are based upon information available to us as of the date hereof and, while we believe such information forms a reasonable basis for such statements, such information may be limited or incomplete, and our statements should not be read to indicate that we have conducted an exhaustive inquiry into, or review of, all potentially available relevant information. These statements are inherently uncertain, and investors are cautioned not to unduly rely upon these statements.