Tripolis Park,
Burgerweeshuispad 101,
1076 ES Amsterdam

Nebius agrees to acquire Eigen AI, strengthening Nebius Token Factory as a frontier inference platform

- Combines Eigen AI's industry-leading inference stack with Nebius's global capacity
- Jointly optimized endpoints achieved top rankings on Artificial Analysis across multiple models
- Eigen AI's founding team, including MIT HAN Lab researchers, will establish Nebius's Bay Area engineering and research presence

Amsterdam, May 1, 2026 — (NASDAQ: NBIS), the AI cloud company, today announced an agreement to acquire Eigen AI, a leading inference and model optimization company.

The acquisition will strengthen Nebius Token Factory as a frontier managed inference platform for production AI, combining a battle-tested optimization stack with Nebius's global compute capacity and AI cloud platform, and will add elite inference research talent to the company's established in-house AI R&D capabilities.

Following close, Eigen AI's inference and post-training optimization layers will be integrated directly into Nebius Token Factory, which provides enterprise-grade autoscaling endpoints and fine-tuning pipelines across all major open-source models. The two companies have already delivered jointly optimized implementations of leading open source models that ranked among the fastest on Artificial Analysis.

The acquisition also accelerates Nebius's expansion in the US. Eigen AI's founding team – researchers who have developed optimization techniques and tools the

industry runs on – will join Nebius to establish a Nebius engineering and research presence in the San Francisco Bay Area.

Roman Chernin, co-founder and Chief Business Officer of Nebius, said:

“We are operating in a capacity-scarcity world where AI builders need optimized inference and infrastructure scale. The integration of Eigen AI’s optimization capabilities and founding team will establish Nebius Token Factory at the frontier of inference, offering customers market-leading model performance and unit economics with massive compute capacity to back it at scale.”

Eigen AI’s founding team brings deep expertise from research that shapes how the industry deploys inference today. Co-founders Ryan Hanrui Wang and Wei-Chen Wang are alumni of MIT’s HAN Lab, led by Professor Song Han, a pioneering researcher in AI computing and model efficiency.

Ryan’s pioneering Sparse Attention (SpAtten) work is the most-cited HPCA paper since 2020, while Wei-Chen received the MLSys 2024 Best Paper Award for Activation-aware Weight Quantization (AWQ) quantization – now the standard for 4-bit model serving in production deployments. Co-founder Di Jin, an MIT CSAIL PhD, brings deep expertise in post-training and large-scale model alignment, having contributed to Meta's Llama 3 and Llama 4 post-training and co-authored the CGPO RLHF framework.

Ryan Hanrui Wang, co-founder and CEO of Eigen AI, said:

“We’re proud to join Nebius and work alongside the Token Factory team to push the boundaries of inference performance. Nebius has built a world-class AI cloud with a deep engineering culture that perfectly aligns with our own. Together, we are removing the friction of AI model customization and deployment so developers can run models reliably in production without managing the underlying infrastructure.”

Inference is now the fastest-growing segment of AI, forecast to account for about two-thirds of compute demand this year. Open-source model usage is rising alongside it. With more workloads moving into production, the system optimization layer is becoming critical infrastructure.

Running inference efficiently in production is inherently complex and requires deep expertise across the entire execution stack, from how models are represented, to how GPU kernels execute them, to how workloads are scheduled in real time.

Open-source models typically ship unoptimized, and newer architectures such as Mixture-of-Experts (MoE), Compressed Sparse Attention (CSA), reasoning, and long-context models introduce additional challenges around memory, routing, and

compute efficiency. Most teams do not have the capacity to solve these problems in-house.

Eigen AI addresses this challenge with a full-stack optimization approach that spans the entire model lifecycle. From post-training and fine-tuning to production inference optimization, across all major open-source models in production demand, including GPT-OSS, Gemma, Qwen, Llama, Nemotron, DeepSeek, GLM, Kimi and MiniMax.

By integrating Eigen AI's optimization layer directly into Nebius Token Factory, Nebius removes this bottleneck across the lifecycle. The system-, model-, and kernel-level techniques developed by the Eigen team are designed to extract materially better performance from hardware out of the box, delivering higher throughput and lower cost per inference without additional engineering overhead.

As a result, Nebius Token Factory customers will benefit from faster time to production, significantly better unit economics, and the ability to adopt new models more quickly. Existing Eigen AI customers will gain access to Nebius's global AI infrastructure and platform capabilities.

The deal consideration will be paid in a combination of cash and Nebius's Class A shares with aggregate value as of signing, based on Nebius's 30-day weighted average stock price, of approximately $643 million, subject to adjustments. The transaction is expected to close in the coming weeks, subject to certain customary conditions, including antitrust clearance.

About Nebius

Nebius, the AI cloud company, is building the full-stack platform for developers and companies to take charge of their AI future — from data and model training to production deployment. Founded on deep in-house technological expertise and operating at scale with a rapidly expanding global footprint, Nebius serves startups and enterprises building AI products, agents and services worldwide.

Nebius is listed on Nasdaq (NASDAQ: NBIS) and headquartered in Amsterdam.

For more information please visit www.nebius.com

Contacts

Media relations: media@nebius.com

Investor relations: askIR@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 our ability to complete the Eigen AI acquisition and our ability to integrate the Eigen AI team and to achieve the synergies and other benefits anticipated, 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: risks associated with acquisitions and the integration of businesses and teams; market, macroeconomic and geopolitical conditions; technological developments; our ability to secure and retain clients; our ability to secure additional capital to enable the growth of the business; 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, 2025, filed with the Securities and Exchange Commission on April 30, 2026.

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.