FriendliAI Expands Ultra-Fast AI Inference Platform with Nebius AI Cloud Integration

REDWOOD CITY, CA, UNITED STATES, October 28, 2025 /EINPresswire.com/ -- FriendliAI, a rapidly growing AI inference platform company, today announces a collaboration with Nebius to deliver faster, more efficient inference for enterprises and startups. The collaboration combines FriendliAI's optimized inference stack with Nebius's AI cloud infrastructure, enabling customers to scale AI workloads instantly for maximum speed and reliability.

Organizations powering customer support bots, coding assistants, and AI agents can now achieve ultra-low latency and cost-efficient inference through FriendliAI APIs running on Nebius infrastructure.

"Our goal is to make world-class AI inference accessible to every company," said Byung-Gon Chun, Founder and CEO of FriendliAI. "By combining our inference optimization technology with Nebius's AI cloud, customers can deploy advanced AI models with the best latency, reliability, and cost efficiency, without any infrastructure complexity."

FriendliAI's technology delivers up to 90% GPU cost savings and the fastest AI inference speeds on the market. Supporting more than 460,000 Hugging Face models, the platform helps teams move from prototyping to production seamlessly, accelerating product launches while reducing infrastructure spend. This combination of cost efficiency and speed has positioned FriendliAI as a compelling solution for enterprises seeking to optimize their AI infrastructure investments. With trillions of tokens served monthly, FriendliAI is redefining how organizations scale AI inference.

FriendliAI's platform optimizes AI inference workloads by addressing the critical challenges faced when deploying AI at scale: prohibitively high infrastructure costs; slow inference speeds that impact the user experience; reliability challenges at scale; and the complexity of managing AI models in production environments. As nearly 90% of a model's cost is due to inference, FriendliAI's optimizations directly address the most resource-intensive phase of AI, enabling sustained performance at scale.

About FriendliAI
FriendliAI is a leading AI inference platform company helping enterprises deploy and scale AI models efficiently, cost-effectively, and reliably. Its platform delivers superior performance while reducing infrastructure costs, making it easy to move from AI experimentation to large-scale

production. FriendliAI supports over 460,000 Hugging Face models through its Dedicated Endpoints, Serverless API, and Container solutions. Learn more at www.friendli.ai.

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 gives AI builders the compute, storage, managed services, and tools they need to build, tune, and run their models.

Lisa Langsdorf
GoodEye PR
+1 347-645-0484
email us here