NVIDIA’s Nemotron Nano 3 Now Available on FriendliAI’s Inference Platform

FriendliAI’s platform achieves 90% GPU cost savings, delivers the fastest LLM inference performance, for agentic AI

REDWOOD CITY, CA, UNITED STATES, December 16, 2025 /EINPresswire.com/ -- FriendliAI, a rapidly growing AI inference platform company, today announced an official partnership with NVIDIA to launch the Nemotron 3 model family, now available on FriendliAI’s Dedicated Endpoints.

Starting today, developers worldwide can deploy Nemotron 3 models on FriendliAI's high-performance inference platform, with remarkable speed, cost-efficiency, reliability, and scalability.

“The combination of NVIDIA’s Nemotron 3 Nano and FriendliAI’s platform represents a milestone in unlocking the promise of agentic AI,” said Byung-Gon Chun, Founder and CEO of FriendliAI. “Efficient, affordable inference is fundamental to deploying agentic AI at scale, and our commitment to performance and scalability makes that possible.”

NVIDIA’s Nemotron 3 is a new family of highly efficient, state-of-the-art reasoning models designed for next-generation agentic AI and real-world, reasoning-intensive applications in fields, such as software development, retail, finance, and cybersecurity. The fully open, small language MoE model is purpose-built to deliver exceptional reasoning performance while maintaining the efficiency required for production use. Key highlights include:

- Up to 13× faster token generation via hybrid Mamba-Transformer MoE architecture and multi-token prediction (MTP) technique
- MoE routing for reduced compute load and real-time latency
- Leading accuracy on SWE Bench, GPQA Diamond, AIME 2025, Humanity Last Exam, IFBench, RULER, and Arena Hard
- Fully open weights, datasets, and recipes for maximum transparency and control

Inference speed is crucial for agentic AI because it enables real time interaction, scalability and cost efficiency, allowing agents to process complex tasks, make fast decisions, and provide seamless user experiences without prohibitive costs. Running Nemotron 3 Nanoon FriendliAI deliver:

- Faster performance with optimized GPU kernels
- More efficient MoE serving with Online Quantization + Speculative Decoding
- Predictable latency and autoscaling for traffic spikes
- 50%+ GPU cost savings on Dedicated Endpoints
- OpenAI-compatible APIs for easy integration

“The combination of cost efficiency and speed has positioned FriendliAI as a compelling solution for enterprises seeking to optimize their AI infrastructure investments,” added Chun.

About FriendliAI
FriendliAI is a leading AI inference platform company helping startups and 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 484,000 Hugging Face models through its Dedicated Endpoints, Serverless API, and Container solutions. Learn more at www.friendli.ai.

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