Nvidia's AI agent play is here with new models, orchestration blueprints | VentureBeat

Nvidia’s AI agent play is here with new models, orchestration blueprints

January 6, 2025 8:30 PM

Join our daily and weekly newsletters for the latest updates and exclusive content on industry-leading AI coverage. Learn More

The industry's push into agentic AI continues, with Nvidia announcing several new services and models to facilitate the creation and deployment of AI agents.

Today, Nvidia launched Nemotron, a family of models based on Meta's Llama and trained on the company's techniques and datasets. The company also announced new AI orchestration blueprints to guide AI agents. These latest releases bring Nvidia, a company more known for the hardware that powers the generative AI revolution, to the forefront of agentic AI development.

Nemotron comes in three sizes: Nano, Super and Ultra. It also comes in two flavors: the Llama Nemotron for language tasks and the Cosmos Nemotron vision model for physical AI projects. The Llama Nemotron Nano has 4B parameters, the Super 49B parameters and the Ultra 253B parameters.

All three work best for agentic tasks including "instruction following, chat, function calling, coding and math," according to the company.

Rev Lebaredian, VP of Omniverse and simulation technology at Nvidia, said in a briefing with reporters that the three sizes are optimized for different Nvidia computing resources. Nano is for cost-efficient low latency applications on PC and edge devices, Super is for high accuracy and throughput on a single GPU and Ultra is for highest accuracy at data center scale.

"AI agents are the digital workforce that will work for us and work with us, and so the Nemotron model family is for agentic AI," said Lebaredian.

The Nemotron models are available as hosted APIs on Hugging Face and Nvidia's website. Nvidia said enterprises can access the models through its AI Enterprise software platform.

Nvidia is no stranger to foundation models. Last year, it quietly released a version of Nemotron, Llama-3.1-Nemotron-70B-Instruct, that outperformed similar models from OpenAI and Anthropic. It also unveiled NVLM 1.0, a family of multimodal language models.

More support for agents

AI agents became a big trend in 2024 as enterprises began exploring how to deploy agentic systems in their workflow. Many believe that momentum will continue this year.

Companies like Salesforce, ServiceNow, AWS and Microsoft have all called agents the next wave of gen AI in enterprises. AWS has added multi-agent orchestration to Bedrock, while Salesforce released its Agentforce 2.0, bringing more agents to its customers.

However, agentic workflows still need other infrastructure to work efficiently. One such infrastructure revolves around orchestration, or managing multiple agents crossing different systems.

Orchestration blueprints

Nvidia has also entered the emerging field of AI orchestration with its blueprints that guide agents through specific tasks.

The company has partnered with several orchestration companies, including LangChain, LlamaIndex, CrewAI, Daily and Weights and Biases, to build blueprints on Nvidia AI Enterprise. Each orchestration framework has developed its own blueprint with Nvidia. For example, CrewAI created a blueprint for code documentation to ensure code repositories are easy to navigate. LangChain added Nvidia NIM microservices to its structured report generation blueprint to help agents return internet searches in different formats.

“Making multiple agents work together smoothly or orchestration is key to deploying agentic AI,” said Lebaredian. “These leading AI orchestration companies are integrating every Nvidia agentic building block, NIM, Nemo and Blueprints with their open-source agentic orchestration platforms.”

Nvidia’s new PDF-to-podcast blueprint aims to compete with Google’s NotebookLM by converting information from PDFs to audio. Another new blueprint will help build agents to search for and summarize videos.

Lebaredian said Blueprints aims to help developers quickly deploy AI agents. To that end, Nvidia unveiled Nvidia Launchables, a platform that lets developers test, prototype and run blueprints in one click.

Orchestration could be one of the bigger stories of 2025 as enterprises grapple with multi-agent production.