

NVIDIA Launches Open Agent Safety Platform to Secure Agents From Testing to Deployment

Open Software Platform and Reference System Design Brings Together Industry, Researchers and Public-Sector Organizations to Set Safer Boundaries for AI Agents, Share Best Practices and Foster International Cooperation to Raise the Bar for Safer AI Agent Deployment

News Summary:

- NVIDIA Open Agent Safety Platform consists of NVIDIA OpenShell open source software and the NVIDIA Sentry reference system design that enables full-stack governance and control across software and the hardware, compute and robotics systems that run agents.
- OpenShell software provides a secure runtime boundary that traces all actions and enforces policy as agents run on NVIDIA Vera CPUs. As open source software, OpenShell can be extended to work with third-party compute platforms, including those from Arm and Intel.
- Sentry adds an out-of-band watchdog that runs on NVIDIA BlueField-4 DPUs to continuously monitor agent behavior. Sentry can quarantine agents that attempt to move outside their boundaries in milliseconds.
- Industry leaders from across the AI ecosystem are joining NVIDIA to strengthen AI safety for every industry across the full stack of infrastructure, software, models and robotics -- including Anthropic, Cisco, CrowdStrike, Dell Technologies, Figure, HPE, Hugging Face, JPMorganChase, Microsoft, Palantir, Palo Alto Networks, Perplexity, Red Hat, Salesforce, SAP, Scale AI, ServiceNow and SpaceXAI.

NVIDIA today announced [NVIDIA Open Agent Safety Platform](#), an open software platform and reference system design to strengthen AI security from agent testing to deployment, with full-stack governance and control across software and the hardware, compute and robotics systems that run agents.

Recent security incidents have underscored the need to equip organizations with open, customizable tools that enforce more control over long-running agents. Across these incidents, the pattern is the same -- the agent circumvented security controls at the application layer to complete its assigned task.

"AI's extraordinary potential for society will only be realized if we solve AI safety," said Jensen Huang, founder and CEO of NVIDIA. "As we continue to discover the frontier of AI capabilities, we must accelerate discovery at the frontier of AI safety. Safety and security require full-stack engineering. NVIDIA Open Agent Safety Platform brings together industry, researchers and public-sector organizations to share best practices, align on evaluation methods and foster international cooperation. Together, we can raise the bar for global AI safety."

Open Agent Safety Platform Adds Control Across the Full Agent Stack

NVIDIA Open Agent Safety Platform enables full-stack governance and control across the software that runs agents, the hardware and compute layers that power their work, and the robotics systems that execute tasks in the physical world. Organizations can deploy elements of NVIDIA Open Agent Safety Platform according to their unique requirements.

It includes [NVIDIA OpenShell](#)™ secure runtime software that sets boundaries for agents running on CPUs. As agents take on more work across more systems, enterprises need an enforceable boundary outside of the model and agent harness. Now broadly available, [OpenShell](#) provides a secure runtime boundary for controlling how autonomous AI agents execute tasks across open and closed models.

OpenShell delivers this protection with minimal overhead on [NVIDIA Vera](#), the first purpose-built CPU for agentic AI. Together, OpenShell and Vera enable agents to operate securely while completing their work as quickly as possible. As open source software, OpenShell can also be extended to work with third-party compute platforms, including those from [Arm](#) and [Intel](#).

The NVIDIA Open Agent Safety Platform reference system design features [NVIDIA Sentry](#), an out-of-band watchdog that runs on [NVIDIA BlueField®-4](#) DPUs to continuously monitor agent behavior. Sentry provides in-silicon security enforcement, meaning that if an AI agent attempts to move outside its software boundary, Sentry quarantines and stops it in milliseconds.

Running on BlueField-4 DPUs, Sentry continuously monitors agent activity and enforces security policies independently in silicon. It combines threat detection, hardware-based agent governance and enforcement and data access protection from an isolated, out-of-band trust domain that is responsive in real time and invisible to agents and attackers.

Sentry is built on [NVIDIA DOCA](#)™ software, which provides the programmable capabilities Sentry uses to inspect agent requests and responses, provide attested telemetry, verify agent identity and enforce granular, zero-trust access policies for data, tools, application programming interfaces and services.

Industry Leaders Strengthen Agent Security With NVIDIA

[Anthropic](#) and NVIDIA have collaborated to bring additional layers of security and control to the agent stack. Claude Managed Agents establish a security boundary by running the agent loop in a separate server from the sandboxes where their work executes. Integrations with OpenShell and BlueField enable enterprises to enforce strict control over agent access through those sandboxes.

"Companies are giving AI agents more of their most important work, and they need to direct and verify what those agents do, especially in sensitive environments," said Paul Smith, chief commercial officer of Anthropic. "Claude Managed Agents gives companies a clear view of what each agent is doing, and NVIDIA's platform adds another layer of governance and control across hardware and software."

SpaceXAI is using NVIDIA Open Agent Safety Platform for Cursor coding agents and Grok models.

"As customers rely more on agents to get real work done, safety should be enforced outside the model by additional controls the agent can't get past," said Mike Nicolls, president at SpaceXAI. "Customers should be able to set those limits for Cursor and Grok and trust they will hold."

Scale AI is working with NVIDIA to incorporate NVIDIA Open Agent Safety Platform technologies into the agentic infrastructure layer of Scale GenAI Portfolio.

"Scale AI is using the NVIDIA Open Agent Safety Platform reference design to build reliable agentic AI systems for our enterprise and government customers running mission-critical applications, with isolation, policy enforcement and auditability built in from the start," said Francis deSouza, CEO of Scale AI. "We support agentic security with clear boundaries that define what agents can do, and controls that keep them operating within those permissions."

[Salesforce](#) and NVIDIA have integrated OpenShell with Slack, enabling teams to manage OpenShell agent activity directly from Slack -- viewing agent activity and audit events, and approving or rejecting agent requests for additional permissions -- giving teams greater visibility and human oversight as agents work.

[SAP](#) is embedding OpenShell with Joule Studio runtime, part of the SAP Business AI Platform, to pair business oversight with runtime security. The company is also contributing engineering work to OpenShell and working with NVIDIA to advance interoperability standards through the [Open Secure AI Alliance](#).

[Accenture](#), [Armadin](#), [Cadence](#), Cognition, [CrowdStrike](#), [Cisco](#), Dassault Systèmes, Deloitte, EY, Hugging Face, [IBM](#), Irregular, Perplexity, Microsoft, SAP, Scale AI, ServiceNow, Siemens, Synopsys, OpenClaw, Palantir and [Palo Alto Networks](#) are also among the over 100 organizations working with NVIDIA Open Agent Safety Platform technologies.

Robotics leaders -- such as Figure, [Gecko Robotics](#) and Skild AI -- are also building with OpenShell to embed agent safety controls into autonomous systems that take action in the physical world.

Citi and JPMorganChase are among the financial services leaders collaborating with NVIDIA on shared open source agent safety technologies.

Energy leaders Hitachi Energy, EPRI, NextEra Energy, Quanta Services, SPP, Schneider Electric, Siemens Energy and Worley are among critical U.S. infrastructure providers working with NVIDIA Open Agent Safety Platform technologies.

Infrastructure software leaders [Canonical](#), [SUSE](#) and [Red Hat](#) are also integrating NVIDIA Open Agent Safety Platform into widely used software operating systems. Red Hat runs OpenShell and DOCA, both part of NVIDIA Open Agent Safety Platform, on Red Hat AI Factory with NVIDIA, a co-engineered, enterprise-grade AI solution for building, deploying and managing AI at scale across hybrid cloud environments.

NVIDIA partners including [Baseten](#), Cisco, CoreWeave, [Dell Technologies](#), GMI Cloud, [HPE](#), HP Inc., Irregular, Lenovo, Microsoft, Nebius, Oracle Cloud Infrastructure, [Supermicro](#) and Together AI are among those offering AI infrastructure solutions that use and support NVIDIA Open Agent Safety Platform technologies to help customers run AI agents more securely.

Availability

NVIDIA Open Agent Safety Platform software, including OpenShell and skills, are available through the [NVIDIA developer resources page](#) and [GitHub](#).

Ecosystem contributions such as NVIDIA Open Agent Safety Platform support the mission of the [Open Secure AI Alliance](#) as well as the broader AI safety and security community. Initiated by NVIDIA alongside over 120 leading organizations and governed by the Linux Foundation, the Open Secure AI Alliance strengthens AI agent security through open research, skills and tools, as well as projects like the [Shared AI Findings Exchange](#), or SAFE.

About NVIDIA

[NVIDIA](#) (NASDAQ: NVDA) is the world leader in AI and accelerated computing.

Certain statements in this press release including, but not limited to, statements as to: AI's extraordinary potential for society only being realized if we solve AI safety; NVIDIA Open Agent Safety Platform bringing together industry, researchers and public-sector organizations to share best practices, align on evaluation methods and foster international cooperation; NVIDIA, together with its commercial partners, raising the bar for global AI safety; expectations with respect to growth, performance, availability, and benefits of NVIDIA's products, services and technologies, and related trends and drivers; expectations with respect to NVIDIA's third party arrangements, including with its collaborators and partners; expectations with respect to technology developments, and related trends and drivers; projected market growth and trends; expectations with respect to AI and related industries; and other statements that are not historical facts are forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, which are subject to the "safe harbor" created by those sections based on management's beliefs and assumptions and on information currently available to management and are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic and political conditions; NVIDIA's reliance on third parties to manufacture, assemble, package and test NVIDIA's products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA's existing product and technologies; market acceptance of NVIDIA's products or NVIDIA's partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA's products or technologies when integrated into systems; and changes in applicable laws and regulations, as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

© 2026 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, and other NVIDIA product and service names are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries and regions. Other company, product, and service names may be trademarks of the respective companies with which they are associated.

Corporate Communications
NVIDIA Corporation
press@nvidia.com