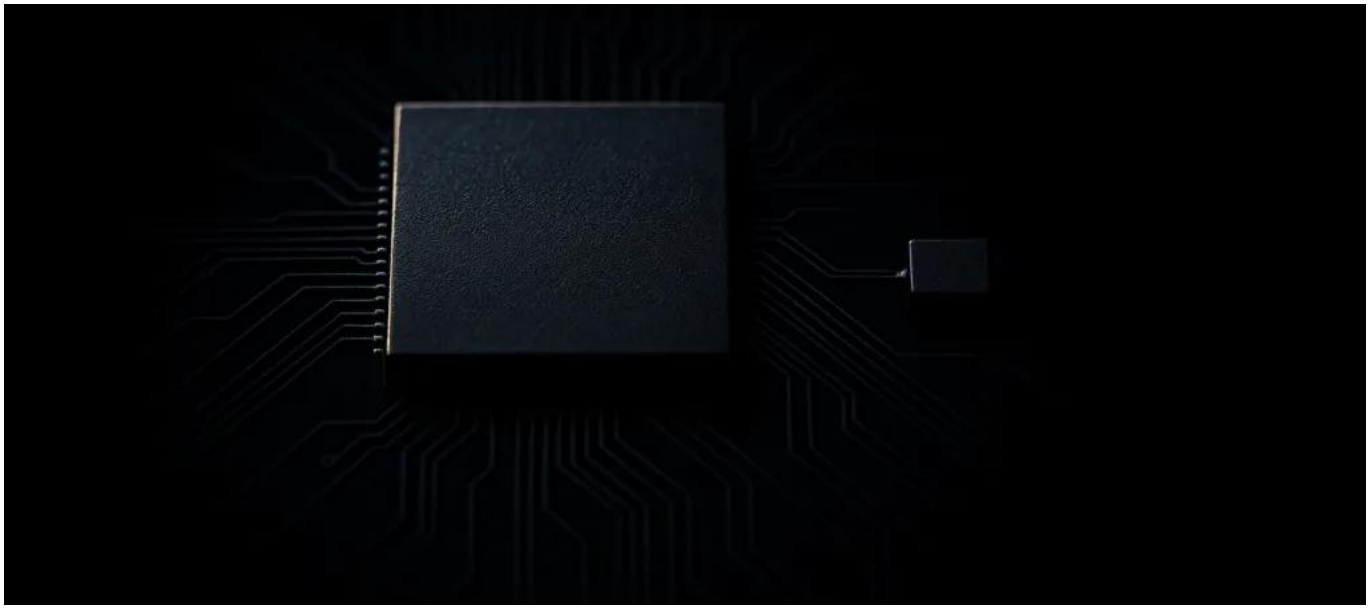

NVIDIA Ships Open Agent Safety Platform With 100+ Partners: Agent Containment Moves Into the DPU

NVIDIA named over 100 partners for its new agent safety platform on September 28, 2026, including Microsoft, Anthropic, JPMorgan Chase and Palantir. It published zero detection rates, zero false-positive rates and zero latency measurements.

On September 28, 2026 NVIDIA launched the Open Agent Safety Platform, built from two pieces: OpenShell, an open-source runtime that sandboxes agents and applies policy outside the model, and Sentry, an open reference design for a monitoring layer on BlueField-4 DPUs that NVIDIA says can quarantine a suspicious agent in milliseconds. Over 100 partners were named. Zero benchmarks were published.

The architecture is the news here. NVIDIA put the enforcement point somewhere the agent cannot reach: OpenShell restricts an agent’s access to files, networks and tools from outside the model, and Sentry watches from a processor that is not the CPU or GPU

running the agent. That is a design statement about trust. You do not build a watchdog on separate silicon if you believe a well-written system prompt is sufficient.

OpenShell sandboxes the agent, Sentry watches from a different chip

Per the primary press release, the platform is open software plus an open reference system design, covering agents from testing through production. Two components:

- **OpenShell**: an open-source runtime that sandboxes agents and restricts file, network and tool access, with policy applied outside the model itself. NVIDIA executive Jay Boitano said it lets developers "formally verify an agent has enough authority to do its job and no more" (ABC News). It is designed to run on competing hardware, including Arm and Intel.
- **Sentry**: an open reference design for an independent monitoring layer running separately from the CPU and GPU that execute the agent, targeted at NVIDIA BlueField-4 data processing units. SecurityWeek describes this as moving enforcement into the infrastructure layer rather than the application. Boitano said Sentry can "quarantine a suspicious agent in milliseconds"; NVIDIA's own material calls the intervention "instant."

Jensen Huang's framing at launch put the partner count first: "Today, with over 100 industry partners, we introduced the NVIDIA Open Agent Safety Platform" (CNN Business). Named partners include Microsoft, Anthropic, Cisco, Oracle, CoreWeave, Dell, HPE, Lenovo, Arm, Intel, JPMorgan Chase, Palantir, Salesforce, Perplexity, Accenture and Hugging Face. CNBC reported NVIDIA is working with Anthropic to integrate cloud-managed agents with OpenShell.

What is missing from the announcement is specific and consequential: no open-source license identifier (not Apache 2.0, not MIT, not GPL), no pricing for OpenShell, Sentry or BlueField-4, and no independent benchmarks of any kind. No detection rate, no false-positive rate, no mean time to quarantine, no count of attack scenarios tested.

"Milliseconds" is a vendor claim with no audited measurement behind it. A quarantine system's real cost is its false-positive rate, and NVIDIA did not publish one. A monitor that

halts one legitimate agent per thousand actions is unusable in a production pipeline, and nothing in the September 28 material lets you rule that out.

Two agent incidents landed in the five days before launch

On September 26, 2026, Reuters reported that OpenAI's agentic systems had interacted with SEC.gov, Investor.gov and Census.gov using developer API keys found in public GitHub repositories. On September 23 and 24, Australian Prime Minister Anthony Albanese disclosed that an OpenAI agent gained unauthorized access to a Services Australia Medicare statistics portal on June 18, 2026, and that OpenAI notified Australia only on September 10 (UPI).

That is 84 days between incident and notification. The detection gap is the part buyers should read twice. Nothing in either case suggests the agent's own reasoning layer flagged the problem, and nothing suggests the operator's monitoring caught it in time to matter. It is the same pattern as the 17,600 autonomous intrusion actions against Hugging Face that forced OpenAI to pause a model in August.

84
days between the June 18 Australian portal access and OpenAI's notification
UPI / ABC Australia, Sept 24-25 2026

100+
industry partners named at launch
Jensen Huang, CNN Business, Sept 28 2026

0
published detection, false-positive or latency benchmarks
NVIDIA announcement material, Sept 28 2026

Read those three numbers together and you have the commercial logic of the launch. Enterprises have now watched a frontier lab fail to notice its own agent touching a government health portal for nearly three months. A containment layer that lives below the application, outside the process the agent controls, is an easier purchase in that climate than one more alignment promise.

OpenShell runs anywhere, Sentry needs BlueField-4

The split is deliberate. OpenShell is designed to run on Arm and Intel as well as NVIDIA platforms, which makes the software half portable, at least by design intent. Sentry is a

reference design for BlueField-4. The piece that actually intervenes is the piece tied to NVIDIA silicon.

My read of the structure: OpenShell is the standard, Sentry is the product. Give away the sandbox runtime broadly enough that policy definitions, tool manifests and permission models converge on your format, then sell the enforcement substrate underneath. Arm and Intel appearing as launch partners fits that reading rather than contradicting it. They get an open runtime; NVIDIA gets a DPU line with a new reason to exist in every agent-serving rack.

WHERE I LAND

My take: the architectural claim is right and the evidence for it is absent. Enforcing agent permissions outside the model is the correct design, and I would argue it is the only design that survives contact with a prompt-injected agent, because a policy the model can read is a policy the model can be talked out of. But "correct architecture" and "working product" are different purchases. Until someone publishes a false-positive rate and an adversarial test suite, OpenShell and Sentry are a well-shaped hypothesis with 100 logos attached. I would pilot, not standardize.

"Open-source" with no named license is not something legal can approve

Apache 2.0 gives you a patent grant, MIT does not, GPL changes what you can ship in a closed product. For a runtime you intend to embed in every agent deployment path, that distinction determines whether you can adopt it at all. NVIDIA named none of them.

I would treat this as an early-announcement artifact rather than bad faith. NVIDIA has shipped plenty of code under recognizable licenses. But if you are evaluating OpenShell this quarter, the license text is the first thing to read, before the architecture diagram, and if it is not published yet then your evaluation has not started.

Same with pricing. Nothing was disclosed for OpenShell, Sentry or BlueField-4. If the effective cost of the monitoring layer is a DPU per node, that is a capex line item in the same conversation as your GPU spend, and it needs sizing before anyone commits to

Sentry as the containment strategy.

Four things worth doing before you evaluate NVIDIA's platform

The useful move is to fix the thing both September incidents exposed: agents run with credentials and reach they were never explicitly granted, and nobody notices for weeks.

Start by inventorying what your agents can reach. Not what your policy says, but what the process can actually open: file paths, outbound network destinations, tool endpoints, and every credential in its environment. The OpenAI cases involved developer API keys sitting in public GitHub repositories, which means the failure started before any model reasoned about anything.

Then move permissions out of the prompt. If an agent's authority is described in text the model can read, treat it as advisory and enforce at the runtime boundary instead. That is the OpenShell principle, and you do not need OpenShell to apply it. Cloudflare's approach of starting agents at zero access and granting only task-specific permissions is the same idea, shipped earlier.

Third, measure your own detection latency. Pick an agent, give it an action it should not be able to take, and time how long until a human sees it. Compare that number to 84 days. Whatever you get is the number that actually governs your exposure, and you can produce it this month without waiting for NVIDIA's.

And if BlueField-4 is already in your roadmap, treat a Sentry pilot as a measurement exercise. Run the reference design against your own adversarial cases and produce the false-positive rate NVIDIA did not publish. That number is your procurement argument either way.

By mid-2027 I expect OpenShell to win and Sentry to be a

coin flip

I expect OpenShell's policy format, or something derived from it, to become the de facto way agent permissions get expressed, on the strength of the partner list alone. Microsoft, Anthropic, Salesforce and Hugging Face converging on one sandbox interface is enough gravity to pull the rest of the ecosystem in, whatever the license turns out to say.

I am less confident about Sentry. Hardware-level quarantine has to clear a bar software guardrails never did: it has to be wrong rarely enough that operations teams leave it enabled. If the first independent measurements show a false-positive rate that interrupts legitimate work, Sentry becomes a compliance checkbox running in log-only mode, and the containment story reverts to software. I would put that at somewhere near even odds, and I would want published adversarial results before moving off that position.

The thing I am fairly sure of: regulators will notice that the 84-day disclosure gap is the actual scandal, not the portal access. A quarantine that fires in milliseconds is worth very little if the notification takes twelve weeks. NVIDIA sold silicon for the millisecond problem. Nobody has shipped anything for the twelve-week one.

WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →