EN

GTC 2026 — Post B3

NVIDIA Agent Toolkit

The Enterprise Agentic Stack · GTC 2026 Series · March 2026

March 2026

Table of Contents

Introduction — GTC 2026, the Agent Convergence

GTC 2026 marked a turning point: agentic AI is no longer a concept. It is infrastructure. Jensen Huang devoted a significant portion of his keynote to the NVIDIA Agent Toolkit — a complete ecosystem to deploy, secure and orchestrate AI agents at enterprise scale.

This B3 post of our GTC 2026 series breaks down the Agent Toolkit: what it contains, why it changes the game, and how your organization can adopt it.

---

1. Agent Toolkit = NemoClaw + AI-Q Blueprint + cuOpt

The NVIDIA Agent Toolkit is not a single product. It is a 4-layer complementary stack:

NemoClaw — The Security and Governance Layer

NemoClaw is the agentic security runtime built on OpenClaw. It provides:

- Native sandboxing: each agent runs in an isolated environment
- Least-privilege model: agents receive only the minimum permissions needed
- Built-in Privacy Router: automatic filtering of sensitive data before transmission to LLMs
- Complete audit trail: every action of every agent is logged and auditable

i Security partners: Cisco, CrowdStrike, Google Security, Microsoft Security, TrendAI

AI-Q Blueprint — The Intelligence Layer

AI-Q Blueprint is the reference architecture for agents with broad access to enterprise data. Key points:

- Hybrid frontier + Nemotron architecture: dynamically routes between powerful and lightweight models
- 50% reduction in inference costs by using Nemotron for repetitive tasks
- Native connectors: SharePoint, Salesforce, SAP, ServiceNow, SQL/NoSQL databases
- Long context: indexing and semantic search across large knowledge bases

cuOpt — The Optimization Layer

cuOpt is NVIDIA's GPU-native mathematical optimization library. Enterprise use cases:

- Logistics route optimization (supply chain, deliveries, field service tours)
- Resource planning: team allocation, machine scheduling, cloud capacity
- Workflow scheduling: maximize throughput of agentic pipelines

Nemotron — The Model Layer

Nemotron is NVIDIA's family of open-source models, optimized for:

- Multi-step reasoning (chain-of-thought, tree-of-thought)
- Tool use (function calling, API calls, codebases)
- On-premise deployment with privacy guarantees

2. The 31,000 Enterprise Partners

NVIDIA is not building the Agent Toolkit alone. 31,000 companies have integrated their systems into the NVIDIA ecosystem. The most advanced use cases:

Partner	Agent Use Case
Adobe	Creative agents — multimedia content generation and revision
Salesforce	CRM agents — lead qualification, customer follow-up, auto outreach
SAP	ERP agents — invoice approval, inventory management, financial analysis
ServiceNow	ITSM agents — incident triage, L1/L2 resolution, SLA tracking
Siemens	Industrial agents — predictive maintenance, digital twins
Atlassian	DevOps agents — code review, sprint management, documentation
Box	Document agents — classification, extraction, compliance
Palantir	Analytics agents — decision ops, intelligence, risk management

i Strong signal: when Adobe, SAP, Salesforce and ServiceNow all align on the same stack, the ecosystem has reached enterprise production maturity.

3. Why a Toolkit vs Isolated Agents

Before the Agent Toolkit, companies built agents case by case. One agent for invoicing. One for CRM. One for IT incidents. Every team reinvented the wheel. The problems:

- Inconsistent security: each team implemented its own permission management
- Data silos: agents did not share common context
- High cost: each project paid the full LLM infrastructure overhead
- Impossible auditability: no unified trace of agentic decisions

The Agent Toolkit solves these 4 problems with a platform approach:

- A single security layer (NemoClaw) — the entire organization shares the same policies
- A shared context bus (AI-Q) — agents can collaborate and enrich each other
- A centralized inference router — cost optimization at enterprise scale
- A unified audit registry — GDPR/SOC2 compliance by design

4. Architect Adoption Guide — How to Implement

The question is no longer "should we adopt agentic AI?" — it's "in what order?". Here is the sequence recommended by BOTUM architects:

Phase 1 — Foundations (Weeks 1-4)

- Deploy OpenClaw: local dev environment with Docker Compose
- Configure NemoClaw: access policies, sandbox isolation, Privacy Router
- Identify the pilot use case: choose a high-volume, low-risk process
- Benchmark models: test Nemotron vs GPT-4o vs Claude on your real data

Phase 2 — Pilot (Weeks 5-12)

- Build the pilot agent with AI-Q Blueprint as reference architecture
- Integrate internal data sources via native connectors (SharePoint, Salesforce, SAP...)
- Implement agentic monitoring: latency, accuracy, human escalation rate
- Validate ROI: measure productivity gain vs inference cost

Phase 3 — Scale (Months 4-12)

- Migrate to Vera Rubin if volume > 10,000 requests/day
- Deploy cuOpt for complex workflow optimization
- Build an 'Agent Ops' team: hybrid SRE and data science profile
- Establish the Agentic Center of Excellence: governance, standards, knowledge sharing

Criterion	BOTUM Recommendation
Base model	Nemotron for repetitive tasks, frontier for complex reasoning
Infrastructure	Cloud for pilot, on-premise for sensitive production
Security	NemoClaw mandatory, Privacy Router for all customer data
Inference budget	AI-Q Blueprint to reduce costs by 50% from day one
Minimum team	1 AI architect + 1 backend dev + 1 ops for the pilot

-> Online version: blog.botum.ca/gtc2026-b3-agent-toolkit-nvidia-stack/