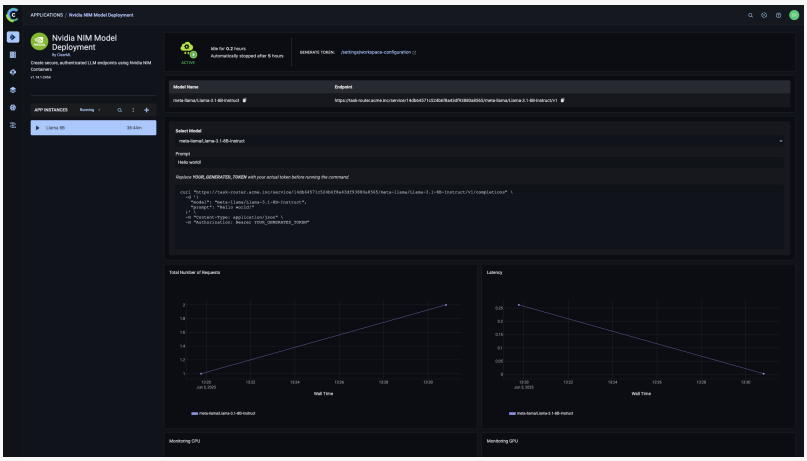
ClearML Integrates NVIDIA NIM to Streamline, Secure, and Scale High-Performance AI Model Deployment

Seamless deployment of NVIDIA-optimized AI inference containers now available through ClearML's end-to-end AI infrastructure platform



NVIDIA NIM Model Deployment with ClearML

SAN FRANCISCO, CA, UNITED STATES, June 12, 2025 /EINPresswire.com/ -- ClearML, the leading end-to-end solution for unleashing AI in the enterprise, today announced full integration with NVIDIA NIM microservices, delivering seamless and secure deployment, scaling, and management of production-grade models in multi-tenant environments. The integration simplifies serving large language models (LLMs) and other AI models by removing the manual complexity of infrastructure setup and scaling. Together, ClearML and NVIDIA remove the operational burden of serving LLMs and other AI models at scale, giving enterprises the flexibility to deploy what they want, where they want, with minimal effort and maximum performance. This integration marks a leap forward for AI builders, infrastructure teams, and platform engineers looking to serve high-performance models without DevOps bottlenecks.

Deploying models at scale has historically been a complex, DevOps-heavy process – from building containers, provisioning GPUs, configuring networking, securing access and authenticating communication with model endpoints, and deploying inference workloads. NVIDIA NIM helps reduce this burden by packaging pre-optimized containers that expose production-ready model endpoints. A new NIM capability takes this even further by decoupling models from their serving infrastructure, offering modularity, performance, and security in a single container.

ClearML bridges the final gap, securely operationalizing NIM microservices with just a few clicks. From within the ClearML platform, users can effortlessly deploy any NIM container on their infrastructure, regardless of whether it's running on bare metal, VMs, or Kubernetes, without needing direct access to the infrastructure. ClearML automatically provisions resources,

manages networking, scales workloads, and enforces secure, authenticated (including role-based access control), tenant-aware access to deployed endpoints.

“NVIDIA NIM makes it easier than ever to serve high-performance AI models,” said Moses Guttmann, Co-founder and CEO of ClearML. “ClearML complements that power by adding authentication, role-based access control, and secure multi-tenancy, as well as making the deployment experience frictionless. With this integration, AI teams can instantly scale inference workloads across their infrastructure – whether on-prem, in the cloud, or hybrid – with full observability, security, control, and automation.”

NVIDIA NIM Expanded Model Coverage
NVIDIA NIM now offers a single container designed to work with a broad range of LLMs. This NIM container decouples the model and runtime; the inference engine (like NVIDIA TensorRT-LLM) is delivered in a continuously maintained container, while the model checkpoint is plugged-in externally. This enables:

- More flexibility across model variants
- Simpler update process and greater flexibility
- Greater security and production-readiness
- Optimal performance on NVIDIA accelerated computing

This modular architecture supports a broad range of LLMs, while improving iteration speed, model provenance, and runtime efficiency.

What ClearML Adds
ClearML brings infrastructure-abstracted deployment and improved observability to NVIDIA NIM deployments. Within the ClearML UI, users simply select the NIM container, assign it to an available resource – whether bare metal, virtual machine, or Kubernetes – and launch and manage the deployment directly from the ClearML user interface.

ClearML’s NIM integration automatically handles:

- Container orchestration on any compute environment
- Networking and endpoint exposure via the ClearML App Gateway
- RBAC-based access control for secure, multi-tenant usage
- Autoscaling and resource management based on workload demand
- Monitoring all endpoints, visualized in a single dashboard
- Enabling multi-tenant deployments on shared compute
- Authenticating access to endpoints for enhanced security

By abstracting away infrastructure complexity, ClearML enhances NIM as a production-ready, secure, and scalable inference engine without requiring users to set up networking, scaling policies, or container orchestration manually – allowing teams to deploy high-performance AI

services, regardless of whether they are on-prem or in the cloud, without custom scripts, manual provisioning, or infrastructure tuning. Combined with ClearML's broader AI infrastructure platform capabilities, such as workload orchestration, resource scheduling, and quotas, this new integration makes enterprise-scale AI both accessible and operationally efficient.

Availability
The NVIDIA NIM integration is now available to all ClearML users, including open-source and enterprise editions. Organizations looking to streamline their inference infrastructure can request a demo at https://clear.ml/demo.

About ClearML
As the leading infrastructure platform for unleashing AI in organizations worldwide, ClearML is used by more than 1,600 customers to manage GPU clusters and optimize utilization, streamline AI/ML workflows, and deploy GenAI models effortlessly. ClearML is trusted by more than 250,000 forward-thinking AI builders and IT teams at leading Fortune 500 companies, enterprises, academia, public sector agencies, and innovative start-ups worldwide. To learn more, visit the company's website at https://clear.ml.

Noam Harel
ClearML
email us here
Visit us on social media:
LinkedIn
YouTube
X
