+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Updated on 06/08/2026

Sign up

NVIDIA AI Enterprise Training: Deploying and Scaling LLMs

3 days (21 hours)

Overview

NVIDIA AI Enterprise is an enterprise software platform designed to develop, deploy, and operate artificial intelligence applications in production.

It brings together frameworks, microservices, SDKs, Kubernetes operators, and infrastructure components to accelerate the implementation of reliable, secure, and scalable AI services.

Our NVIDIA AI Enterprise training will enable you to master the deployment and industrialization of LLMs in the enterprise by leveraging the key building blocks of the NVIDIA ecosystem: NIM, NeMo, TensorRT-LLM, Kubernetes, RAG, and observability.

You will learn how to deploy inference microservices, optimize GPU performance, design a RAG architecture, secure model usage, and monitor your LLM endpoints.

By the end of the course, you will be able to design, deploy, and monitor an enterprise LLM architecture with NVIDIA AI Enterprise, while addressing security, cost, performance, and governance challenges.

Like all our training courses, this one will introduce you to **the latest stable version** of the technology and its new features.

Objectives

- Understand the fundamentals of NVIDIA AI Enterprise
- Deploy LLM models with NVIDIA NIM

- Optimize inference with TensorRT-LLM
- Structure fine-tuning and evaluation workflows with NeMo
- Design an enterprise RAG architecture
- Industrializing, Securing, and Monitoring LLM Services in Production

Target Audience

- AI and ML engineers
- Data scientists looking to industrialize generative models
- Cloud architects and AI architects
- DevOps, MLOps, and platform engineers
- AI technical leads and innovation teams

Prerequisites

- General knowledge of generative AI and LLM models
- Basic understanding of Python, REST APIs, and containers
- General understanding of Kubernetes or cloud architectures
- Basic knowledge of application deployment and observability

Technical prerequisites

- A computer running Linux, macOS, or Windows with WSL2
- Access to a GPU-compatible environment or a GPU cloud environment
- Install Docker, kubectl, Helm, and a code editor
- Have an NVIDIA Developer or NVIDIA NGC account, depending on the selected workshops
- Ensure a stable internet connection

NVIDIA AI Enterprise Training Program

[Day 1 - Morning]

NVIDIA AI Enterprise Fundamentals and Generative AI in Production

- Understand the role of NVIDIA AI Enterprise in an enterprise AI strategy
- Identify key components: NIM, NeMo, TensorRT-LLM, drivers, GPU Operator, and the NGC registry
- Understand the challenges of LLM deployment: cost, latency, security, scalability, and governance
- Distinguish between AI experimentation, RAG prototypes, and production-ready LLM services
- Identify deployment architectures: bare metal, cloud, Kubernetes, OpenShift, and hybrid environments
- Hands-on workshop: explore the NVIDIA AI Enterprise ecosystem, the NGC Catalog, and the components
  used during the training

[Day 1 - Afternoon]

NVIDIA NIM for serving LLM models

- Understand the role of NVIDIA NIM in the accelerated deployment of foundation models
- Deploy a NIM microservice to expose an inference API compatible with LLM use cases
- Configure environment variables, access keys, container images, and GPU settings
- Test inference with simple prompts, batch requests, and conversational scenarios
- Analyzing throughput, response time, concurrency, and GPU consumption
- Hands-on workshop: Deploy your first NIM LLM and query its inference endpoint

Inference optimization with TensorRT-LLM

- Understand the role of TensorRT-LLM in optimizing LLM inference
- Identify performance levers: batching, quantization, KV cache, tensor parallelism, and paged attention
- Compare model constraints based on size, context, GPU memory, and expected latency
- Understand the benefits of FP8, INT8, and other optimization strategies
- Evaluate trade-offs between quality, cost, throughput, and response time

[Day 2 - Morning]

NeMo for adapting, evaluating, and governing LLMs

- Understand the role of NVIDIA NeMo in the generative model lifecycle
- Identify use cases: fine-tuning, adaptation, evaluation, guardrails, and dataset preparation
- Understand the concepts of LoRA, PEFT, alignment, and controlled personalization
- Set up a model adaptation workflow based on business needs
- Establishing evaluation criteria: response quality, hallucinations, safety, and relevance
- Hands-on workshop: structuring a dataset and defining a strategy for adapting an LLM

[Day 2 - Afternoon]

Building an enterprise RAG architecture

- Understand the role of RAG in connecting LLMs to internal data
- Design an architecture with document ingestion, embeddings, a vector database, and a generation engine
- Define chunking, metadata, filtering, and source access control strategies
- Connect a NIM endpoint to an application RAG pipeline
- Evaluate the relevance of responses, source traceability, and the risk of hallucinations

Security, compliance, and governance of LLMs

- Identify risks associated with LLMs: data leakage, prompt injection, hallucination, and unauthorized access
- Implement security controls for prompts, responses, and internal documents
- Define a strategy for guardrails, filtering, auditing, and output validation
- Understand compliance requirements: confidentiality, logs, traceability, and data retention
- Apply best practices for governance of models, datasets, and endpoints
- Hands-on workshop: auditing a RAG use case and identifying security checkpoints to add

[Day 3 - Morning]

Kubernetes deployment and MLOps industrialization

- Deploying AI workloads with Kubernetes, GPU Operator, and NIM Operator
- Understanding GPU constraints: scheduling, resources, quotas, isolation, and multi-tenancy
- Set up a versioned deployment workflow for LLM endpoints
- Automate promotion between development, staging, and production
- Integrating NVIDIA AI Enterprise into an MLOps or LLMOps approach

[Day 3 - Afternoon]

Observability, performance, and operations

- Monitor key metrics: latency, throughput, GPU utilization, errors, and queues
- Set up dashboards to track usage, costs, and performance of LLM endpoints
- Analyze application logs, traces, prompts, responses, and inference events
- Define alert thresholds for service degradation or usage deviations
- Optimize resources based on the model, workload, context, and business criticality

Final case: Industrializing an enterprise LLM service

- Assemble the components: NIM, TensorRT-LLM, RAG, security, and observability
- Define a target architecture tailored to an enterprise business use case
- Implement best practices for deployment, monitoring, governance, and operations
- Prepare a production deployment checklist for an LLM service
- Identify next steps: scaling up, model adaptation, costs, and multi-project governance
- Hands-on workshop: Design and present an industrialized LLM architecture.

To learn more

Target Audience

This training is intended for both individuals and companies, large or small,

wishing to train their teams in a new advanced IT technology or to acquire specific business knowledge or modern methods.

Entry-level assessment

The pre-training assessment complies with Qualiopi quality standards. Upon final registration, the learner receives a self-assessment questionnaire that allows us to evaluate their estimated proficiency in various types of technologies, as well as their expectations and personal goals for the upcoming training, within the limits imposed by the selected format. This questionnaire also allows us to anticipate certain connection or internal security issues within the company (intra-company or virtual classroom) that could pose challenges for monitoring and ensuring the smooth running of the training session.

Teaching Methods

Practical Course: 60% Practical, 40% Theory. Training materials distributed in digital format to all participants.

Organization

The course alternates between theoretical input from the trainer, supported by examples and reflection sessions, and group work.

Assessment

At the end of the session, a multiple-choice questionnaire is used to verify that the skills have been properly acquired.

Certification

A certificate will be issued to each trainee who has completed the entire training program.