

Updated on 08/26/2026

Sign Up

Ollama Enterprise Training: Deploying Private LLMs in Production

3 days (21 hours)

Overview

Ollama enables you to deploy private LLMs by managing local execution, data security, and the industrialization of AI use cases in production.

Our Ollama Enterprise training goes beyond basic installation to help you understand how to design, secure, and operate language models in a professional setting.

You'll learn to structure a deployment pipeline around Modelfiles, the Ollama API, inference settings, and best practices for model management to build reliable and reproducible AI services.

The training also covers performance, monitoring, governance, and application integration challenges, enabling you to transition from a demonstration prototype to robust deployment in a real-world environment.

By the end of the course, you'll be able to deploy private LLMs, optimize their resource consumption, manage their lifecycle, and design an architecture tailored to production and privacy constraints.

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

Objectives

- Understand Ollama's architecture and its business applications.

- Create and customize models using Modelfiles and runtime parameters.
- Deploy private LLMs with production and privacy in mind.
- Integrate Ollama into applications and services via its API.
- Implement security, monitoring, and optimization practices.
- Standardize the model lifecycle, from testing to deployment.

Target Audience

- DevOps Engineer
- Cloud Architect
- Backend Developer
- MLOps Engineer
- System Administrator

Requirements

- Basic knowledge of Linux and server environment administration.
- Experience with REST APIs and JSON exchanges.
- Understanding of language models and the uses of generative AI.
- Experience with Docker, Git, or application deployment tools.
- Understanding of security, privacy, and production deployment challenges.

Technical Prerequisites

- A recent computer with sufficient resources to run Ollama locally.
- A compatible operating system, such as Windows 11, macOS, or a recent version of Linux.
- Stable Wi-Fi or Ethernet network connection for data transfers and downloads.
- Modern browsers and a development environment with a terminal, code editor, and API client.
- A microphone and headset are recommended for remote sessions, as well as local administrator access if necessary.

Agenda for our Ollama Enterprise training: Deploying Private LLMs in Production

[Day 1 - Morning] Ollama

Fundamentals

- Understand Ollama's role in a private LLM strategy.
- Identify relevant use cases for the enterprise, the workstation, and servers.
- Explore the model lifecycle, from download to local execution.
- Analyze constraints related to privacy, sovereignty, and cost control.

- Present the key components: Modelfile, runtime, API, and model library.
- Hands-on workshop: Installing Ollama, launching your first model, and getting started with essential commands.

[Day 1 - Afternoon]

Models and Configuration

- Explore the concepts of models, variants, and quantification.
- Understand the structure of a Modelfile and its role in reproducibility.
- Configure inference behavior using Ollama's main settings.
- Customize system prompts and messages to guide the model's responses.
- Organize a library of models for different business use cases.
- Hands-on workshop: Create a custom Modelfile and compare several inference configurations.

[Day 2 - Morning]

API and Application Integration

- Explore the capabilities of the Ollama API for data generation and exchange.
- Understand request and response formats, including JSON and generation parameters.
- Set up simple calls from a testing tool or script.
- Integrate a model into an internal application or a backend service.
- Manage response times, context memory, and usage limits.
- Hands-on workshop: developing your first API client to query a local model and utilize its response.

[Day 2 - Afternoon] Performance

and Deployment

- Evaluate the impact of model sizes on CPU and GPU resources.
- Understand optimization strategies related to memory, cache, and context.
- Identify best practices for scaling in shared-use scenarios.
- Set up a framework for load testing and functional validation.
- Prepare operational scenarios for pre-production and production environments.
- Hands-on workshop: Comparative performance measurement of multiple models and tuning of execution parameters.

[Day 3 - Morning]

Security and Governance

- Define a framework for the use of private LLMs within the company.
- Analyze the risks associated with sensitive data, prompts, and generated output.
- Organize access controls, logs, and audit policies.
- Implement validation procedures before releasing data to business units.
- Address compliance, accountability, and traceability issues.
- Hands-on workshop: designing a secure usage policy for an internal AI service.

[Day 3 - Afternoon]

Deployment and Production Rollout

- Design a target architecture for an Ollama service in production.
- Choose between on-premises deployment, a dedicated server, or a containerized environment.
- Set up monitoring, alerts, and operational metrics.
- Organize the deployment, updates, and lifecycle management of models.
- Leverage user feedback to improve the AI platform.
- Hands-on workshop: Build a mini-plan for production deployment, including architecture, monitoring, and operations.

Target Audience

This training is intended for both individuals and companies—large or small—that wish to train their teams in a new advanced IT technology or to acquire specific business knowledge or modern methods.

Assessment upon enrollment

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 with 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

Hands-On Training: 60% Practical, 40% Theoretical. Training materials distributed in digital format to all participants.

Organization

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

Assessment

At the end of the session, a multiple-choice quiz is administered to verify that participants have successfully acquired

of the skills.

Certification

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