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

Updated on 01/09/2026

Register

vLLM Training: Deployment and Optimization

3 days (21 hours)

Overview

vLLM is a high-performance language model serving engine designed for production environments. Thanks to innovations such as PagedAttention and continuous batching, it allows models such as Llama or Mistral to be run with optimal efficiency and controlled memory consumption.

This vLLM: Deployment & Optimization training course guides you step by step through the process of setting up a reliable, scalable, and monitored serving infrastructure. You will learn how to deploy vLLM on Kubernetes and Cloud environments, monitor its performance, and automate your deployments with modern CI/CD tools.

The hands-on teaching approach alternates between technical input and practical workshops to enable you to understand, configure, and effectively operate an LLM-based generation service.

You will learn how to integrate vLLM into your existing applications, optimize performance, reduce operating costs, and ensure the security and compliance of your production environment.

By the end of the training, you will be able to design, deploy, and monitor a complete vLLM serving environment, while adopting a professional approach to industrialization and performance management.

Like all our training courses, this one is based on the latest version of vLLM.

Objectives

- Deploy a robust, production-ready vLLM service.
- Optimize latency, throughput, and GPU consumption.
- Monitor the service using Prometheus and Grafana.

- Automate deployments with CI/CD and GitOps.
- Apply security and compliance best practices.

Target audience

- DevOps and MLOps engineers
- Cloud architects
- AI/ML engineers
- SRE

Prerequisites

- Good understanding of Docker and Kubernetes
- Basic knowledge of Python and Linux administration
- Access to a cloud environment or GPU machine for practical work

vLLM Training Program: Deployment and Optimization

[Day 1 - Morning]

Discover vLLM and its ecosystem

- Presentation of the role of vLLM in language model serving architecture
- Operating principles: PagedAttention, continuous batching, and KV cache management
- Compatibilities and integrations: OpenAI-compatible API, Llama and Mistral models
- Professional uses: internal assistants, chatbots, conversational engines, and text analysis
- Technical constraints: GPU management, drivers, runtime (CUDA or equivalent), and dependencies
  Python
- Hands-on workshop: Install vLLM and run your first inference query.

[Day 1 - Afternoon]

Install and configure a reliable environment

- Creating a stable and maintainable Docker or virtual environment
- Configuring key settings
- Managing templates: downloading, licensing, and efficient storage
- Securing network configuration and deployment
- Validation testing and initial performance verification

Mastering architecture and execution modes

- Understanding single and multi-GPU modes
- Distributing computation and parallelizing queries
- Reducing latency with prefix caching and model preloading
- Performance measurement: response time, error rate, and memory consumption
- Best practices for continuous operation and update management
- Hands-on workshop: Observing and analyzing execution metrics with a controlled set of queries.

[Day 2 - Morning]

Deploying vLLM in Kubernetes

- Creating and configuring Deployment, Service, and HorizontalPodAutoscaler manifests
- Configuring GPU nodes, managing selectors and tolerances
- Storing models in persistent volumes (S3, GCS, PVC)
- Deployment strategies: Rolling Update, Blue/Green, and Canary
- Managing restarts and monitoring critical pods
- Hands-on workshop: Deploy vLLM on a Kubernetes cluster and verify its scalability.

[Day 2 - Afternoon]

Monitoring, reliability, and cost control

- Collecting and visualizing metrics with Prometheus and Grafana
- Configuring alerts for latency, errors, and GPU usage
- Tracking SLOs and SLIs to assess service reliability
- Financial optimization: adjusting sizing and batching strategies
- Building management dashboards

Optimizing performance and scalability

- Balancing throughput and latency for responsive service
- Optimizing execution parameters: batching, cache, and number of workers
- Use quantization to reduce costs without significant loss
- Performing comparative benchmarks and interpreting the results
- Hands-on workshop: Implement a load testing session and adjust settings accordingly.

[Day 3 - Morning]

Connect vLLM to existing applications

- Creating a facade API for internal or external consumption
- Develop robust Python and JavaScript clients
- Managing request context and conversational memory
- Streaming responses for progressive token display

- Monitoring calls via distributed tracing

Automating deployments with CI/CD and GitOps

- Designing a complete CI/CD pipeline for vLLM
- Using Terraform and Helm for infrastructure
- Implementing a GitOps approach with ArgoCD or Flux
- Integrating security audits and automatic dependency checks
- Hands-on workshop: Create an automated deployment pipeline and validate a deployment in simulated production.

[Day 3 - Afternoon]

Ensuring security and compliance

- Managing secrets and access rights using RBAC
- Applying the principle of least privilege and separation of environments
- Processing and protecting sensitive data (PII)
- Control of generated content and filtering policies
- Validation of the secure production checklist

Continuous operation and improvement

- Monitoring performance, usage trends, and resource consumption
- Preventive maintenance and updating of models and dependencies
- Incident management and implementation of corrective actions
- Communication of results and availability indicators
- Hands-on workshop: Review an operational runbook and adjust monitoring thresholds.

Target companies

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

Positioning at the start of training

The positioning at the start of the training complies with Qualiopi quality criteria. Upon final registration, the learner receives a self-assessment questionnaire that allows us to assess their estimated level of knowledge of different types of technologies, their expectations and personal objectives 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 be problematic for the monitoring and smooth running of the training session.

Teaching methods

Practical training: 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 discussion sessions, and group work.

Assessment

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

Certification

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