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

Updated on 02/27/2026

Sign up

Unsloth Training: Mastering LLMOps

3 days (21 hours)

Overview

This LLMOps vLLM Inference and Acceleration with Unsloth training teaches you how to deploy and serve language models in production with reduced latency and optimized GPU costs. You will master the deployment of high-performance inference APIs, throughput optimization, and fine-tuning acceleration to effectively scale your RAG, internal chat, and business assistant use cases.

The goal is to transition from a notebook to a robust service: choosing the runtime, configuring GPUs, managing batching, streaming, quantization, and observability. We compare quality/performance trade-offs and architectural patterns (API, worker, queue) suited to real-world workloads.

The approach is 100% hands-on: guided workshops, reproducible demos, and ready-to-use scripts. Deliverables: an operational vLLM server, an Unsloth pipeline for rapid fine-tuning, and a production deployment checklist (tests, metrics, limits, security).

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

Objectives

- Deploy a vLLM inference server with API and streaming.
- Optimize throughput via batching, KV cache, and GPU settings.
- Implement quantization and evaluate the impact on quality and latency.
- Accelerate fine-tuning with Unsloth (LoRA/QLoRA) and validate the performance gains.
- Scale up the deployment: logs, metrics, load testing, and usage limits.

Target Audience

- ML Engineers / LLM Engineers
- Data Scientists looking to move to production
- DevOps / SREs involved in GPU deployment
- Backend developers integrating LLMs via APIs

Prerequisites

- Python (environments, packages, scripts)
- Basics of Transformers, tokenization, embeddings
- Linux/CLI basics and process management
- Understanding of HTTP/JSON APIs and application logs

Technical prerequisites

- Linux or Windows machine with WSL2, or macOS (NVIDIA GPU recommended)
- Minimum 16 GB RAM, 32 GB recommended
- NVIDIA GPU recommended (CUDA) with at least 12 GB VRAM, 24 GB recommended
- Python 3.10+ and tools: Git, terminal, code editor
- Access to a GPU environment (local or remote) for the workshops

Our Unsloth Training Program: Mastering LLMOps

[Day 1 - Morning]

LLM Ops Fundamentals and Inference Architecture

- Clarifying latency, throughput, cost, and quality objectives (SLO/SLA)
- Understanding the inference pipeline: tokenization, prefill, decode, KV cache
- Choosing a model format: HF Transformers, GGUF, AWQ/GPTQ (performance/quality impacts)
- Set up the GPU environment: drivers, CUDA, PyTorch, checks, and diagnostics
- Hands-on workshop: Measure an inference baseline (p50/p95 latency, tokens/s) on an HF model.

[Day 1 - Afternoon]

Setting up vLLM to deploy an LLM in production

- Installing and configuring vLLM (versions, GPU compatibility, key parameters)
- Starting an OpenAI-compatible server: endpoints, models, limits, and timeouts
- Optimizing throughput with PagedAttention, batching, and KV cache management
- Managing concurrency: queues, quotas, backpressure, and rejection strategies
- Hands-on workshop: Deploy a vLLM server and run a load test (concurrency + tokens/s).

[Day 2 - Morning]

vLLM acceleration: quantization, parallelism, and tuning

- Choosing a quantization strategy: FP16/BF16 vs. INT8/INT4 (quality, VRAM, performance)
- Configuring tensor parallelism and multi-GPU distribution (constraints and benefits)
- Adjusting vLLM parameters: max model length, max number of sequences, block size, swap, and memory limits
- Setting up reproducible benchmarks: prompts, batch sizes, metrics, and comparisons
- Hands-on workshop: Compare two vLLM configurations (quantized vs. non-quantized) and document the
  gains.

[Day 2 - Afternoon]

Unsloth: rapid LoRA fine-tuning and preparation for serving

- Understanding Unsloth: objectives, performance gains, limitations, and compatible models
- Preparing a training dataset: formats, cleaning, splitting, and quality checks
- Launching LoRA/QLoRA fine-tuning: hyperparameters, VRAM, stability, and checkpoints
- Exporting and packaging: LoRA merge, hyperparameter backup, versioning, and artifact traceability
- Hands-on workshop: Fine-tune a model with Unsloth (QLoRA) and then export a ready-to-use artifact.

[Day 3 - Morning]

Integration: serving a fine-tuned model with vLLM

- Loading a fine-tuned model: paths, configurations, tokenizer, and vLLM compatibility
- Configuring generation: temperature, top_p, max_tokens, stop sequences, and safeguards
- Implement a prompt strategy: templates, system prompts, and non-regression tests
- Validating quality: test datasets, simple scoring, and before/after fine-tuning comparisons
- Hands-on workshop: Deploy the Unsloth model in vLLM and run a suite of regression tests.

[Day 3 - Afternoon]

LLM Operations: observability, security, and industrialization

- Implement observability: metrics (latency, tokens/s, errors), structured logs, and traces
- Managing costs: per-client limits, caching, timeout policies, and GPU scaling
- Securing the API: authentication, rate limiting, input filtering, and protection against abuse
- Industrialization: model versioning, rollback, canary deployments, and operations runbooks
- Hands-on workshop: Build a mini-runbook (alerts + actions) and a deployment checklist.

Introduction to Deep Learning

PyTorch Training

TensorFlow training

Relevant companies

This training is designed for both individuals and businesses—large or small—that wish to train their teams in new, advanced IT technologies or to acquire specific industry knowledge or modern methodologies.

Placement 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.

Certification

At the end of the session, a multiple-choice quiz 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.