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

Updated on 05/12/2026

Sign up

LLaMA-Factory and Unsloth Training

3 days (21 hours)

Overview

LLaMA-Factory and Unsloth are now two essential tools for industrializing the fine-tuning of open-source language models. Thanks to their advanced optimizations, they enable the rapid training of high-performance LLMs while significantly reducing hardware requirements and GPU costs.

Our LLaMA-Factory and Unsloth training will enable you to master the entire LLM fine-tuning pipeline: dataset preparation, optimized training, quantization, advanced tuning, model evaluation, and deployment.

You will learn to effectively leverage modern techniques such as QLoRA, PEFT, and 4-bit training to optimize your AI infrastructure.

By the end of this course, you will be able to develop complete generative AI fine-tuning pipelines, industrialize open-source models, and optimize training and inference costs.

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

Objectives

- Understand open-source LLM architectures
- Master modern fine-tuning techniques
- Effectively use LLaMA-Factory and Unsloth
- Optimize GPU training with QLoRA
- Build domain-specific conversational datasets
- Deploy generative AI models in production

Target Audience

- Data Scientists
- AI / Machine Learning Engineers
- Python Developers
- MLOps Engineers
- AI Architects

Requirements

- Strong knowledge of Python
- General knowledge of Machine Learning
- Basic understanding of PyTorch
- Experience with GPU environments recommended

Technical requirements:

- Laptop with at least 16 GB of RAM; 32 GB recommended
- Access to a CUDA-compatible NVIDIA GPU environment, ideally with at least 16 GB of VRAM
- At least 50 GB of available disk space for models, datasets, and checkpoints
- Stable internet connection to download models, dependencies, and datasets

Our LLaMA-Factory and Unsloth training program: LLM Fine-Tuning

[Day 1 - Morning]

Open-source LLM environment and fine-tuning strategies

- Understanding open-source LLM architectures
- Overview of the LLaMA, Mistral, Qwen, and Gemma models
- Differences between pre-training, instruction tuning, and fine-tuning
- Introduction to LoRA, QLoRA, and PEFT techniques
- Setting up a GPU environment with CUDA, PyTorch, and Transformers
- Hands-on workshop: Complete installation of LLaMA-Factory and GPU validation.

[Day 1 - Afternoon]

Getting started with LLaMA-Factory

- Architecture and operation of LLaMA-Factory
- Managing conversational datasets
- Alpaca, ShareGPT, and custom dataset formats

- Configuring fine-tuning pipelines
- Monitoring training and managing checkpoints
- Hands-on workshop: Fine-tuning your first conversational model.

[Day 2 - Morning]

Memory and performance optimization with Unsloth

- Understanding GPU and VRAM limitations
- Introduction to Unsloth and kernel optimizations
- Accelerating fine-tuning with Flash Attention
- Memory reduction with 4-bit quantization
- Comparison of standard vs. optimized performance
- Hands-on workshop: Optimizing a QLoRA training run with Unsloth.

[Day 2 - Afternoon]

Advanced fine-tuning and domain-specific datasets

- Building specialized domain-specific datasets
- Cleaning and validating training data
- Data augmentation and synthetic generation techniques
- Handling bias and hallucinations
- Adjustment of critical hyperparameters
- Hands-on workshop: Creating a domain-specific dataset and specialized training.

[Day 3 - Morning]

Model evaluation, testing, and alignment

- Qualitative and quantitative evaluation methods
- LLM benchmarks and metrics
- Introduction to RLHF and alignment
- Detection of drift and hallucinations
- Optimization of system prompts
- Hands-on workshop: Comparative evaluation of several fine-tuned models.

[Day 3 - Afternoon]

Industrialization and deployment of LLMs

- Exporting fine-tuned models
- Deployment with vLLM, Ollama, and Text Generation Inference
- GPU Cost Management and Inference Strategy

- Introduction to LLM MLOps Pipelines
- AI Model Security and Governance
- Hands-on Workshop: Deploying a Fine-Tuned Model as a Local API

Target Audience

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

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 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.