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

Updated on 06/12/2026

Sign up

Small Language Models Training with Hugging Face and Ollama

2 days (14 hours)

Overview

Small Language Models refer to compact language models capable of handling targeted tasks with fewer resources than a large general-purpose LLM. Combined with Hugging Face and Ollama, they enable the creation of AI assistants that are faster, more private, more efficient, and more cost-effective.

Our Small Language Models with Hugging Face and Ollama training will teach you how to select, run, customize, and integrate compact language models into real-world business use cases.

You will learn to explore the Hugging Face Hub, read model cards, compare multiple models, and analyze licenses, evaluation results, and constraints for local or private deployment.

By the end of the course, you will be able to run models with Ollama, create a specialized assistant using a Modelfile, connect an SLM to an application, and set up a basic, cost-effective RAG on internal documents.

This training also covers response evaluation, security, handling sensitive data, documenting model selection, and best practices for scaling SLMs for professional use.

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 uses, benefits, and limitations of Small Language Models
- Select and compare SLMs on Hugging Face Hub
- Run and customize models locally with Ollama
- Create an application connected to a local model via the Ollama API
- Build a lightweight RAG using internal documents, embeddings, and SLMs
- Evaluating, securing, and scaling the use of SLMs in the enterprise

Target Audience

- Developers and tech leads
- AI and ML engineers
- Data scientists and MLOps engineers
- AI architects and cloud architects
- Innovation teams looking to deploy on-premises or private AI models

Prerequisites

- General knowledge of generative AI or language models
- Basic knowledge of Python, JavaScript, or TypeScript is a plus
- Basic understanding of APIs, JSON files, and development environments

Technical prerequisites

- A computer running Linux, macOS, or Windows with WSL2
- Create a Hugging Face account to access models, datasets, and private spaces
- Ensure a stable internet connection to download models and dependencies

Our Small Language Models training program with Hugging Face and Ollama

[Day 1 - Morning]

Understanding Small Language Models and their use cases

- Understand what a Small Language Model is and how it differs from a general-purpose LLM
- Identify the benefits of SLMs: reduced latency, controlled inference costs, privacy, and on-premises deployment
- Understanding suitable use cases: classification, extraction, summarization, business assistant, internal chatbot
  internal and document automation
- Identify the limitations of compact models: complex reasoning, long-form context, hallucinations, robustness, and domain specialization
- Position SLMs within an enterprise architecture: local workstation, internal server, edge, or application API

- Hands-on workshop: identify relevant enterprise use cases for an SLM and define model selection criteria

[Day 1 - Afternoon]

Selecting a model with Hugging Face

- Use Hugging Face Hub to search for compact language models suited to a business need
- Reading a model card: intended uses, limitations, license, training data, evaluation, and deployment constraints
- Compare model families: Phi, Gemma, Qwen, Llama, Mistral, SmolLM, and Instruct models
- Analyze selection criteria: size, license, language, context, format, performance, resources, and local compatibility
- Use benchmarks, leaderboards, and published metrics to compare multiple candidate models
- Hands-on workshop: select several models on Hugging Face and build a for a business use case

Run and customize an SLM with Ollama

- Install and configure Ollama on a local machine or test server
- Download, launch, and test a model compatible with the Ollama CLI
- Use the main commands: pull, run, list, show, remove, and serve
- Understand inference parameters: temperature, top_p, context, seed, repetition, and number of tokens
- Create a specialized assistant using a Modelfile, a system prompt, and tailored parameters
- Hands-on workshop: launch an SLM with Ollama, create a local domain-specific assistant with a Modelfile, and test several prompt strategies

[Day 2 - Morning]

Integrate an SLM into an enterprise application

- Use Ollama’s local API to connect an SLM to a Python, JavaScript, or TypeScript application
- Create an internal API to expose a local model to a business application
- Manage sessions, prompts, parameters, errors, timeouts, and structured responses
- Implement usable output formats: JSON, structured summaries, classification, or information extraction
- Add simple safeguards: input validation, output filtering, logs, usage restrictions, and sensitive data management
- Hands-on workshop: create a mini application service that queries a local SLM via the Ollama API

[Day 2 - Afternoon]

Frugal RAG with Hugging Face, Ollama, and internal documents

- Understanding the role of RAG in connecting an SLM to enterprise documents
- Designing a frugal RAG architecture: ingestion, tokenization, embeddings, vector search, and generation
- Choosing between local models, local embeddings, hosted embeddings, and vector databases based on
  privacy constraints
- Reducing costs and latency: chunk size, number of retrieved documents, useful context, and short prompts
- Limiting hallucinations with grounding, citations, sources, and rejection guidelines
- Hands-on workshop: building a local RAG mini-application with internal documents, embeddings, and SLM running via Ollama

Evaluation, security, and industrialization of SLMs

- Building an evaluation dataset with prompts, expected responses, edge cases, and business examples
- Compare multiple SLMs based on quality, latency, memory consumption, model size, and operational cost
- Identify risks: hallucinations, bias, sensitive data, non-compliant responses, and
  model dependency
- Document the model selection: license, limitations, use cases, benchmarks, parameters, and usage rules
- Define a target architecture: local machine, internal server, container, private API, edge, or
  cloud environment
- Hands-on workshop: designing an enterprise SLM architecture with Hugging Face, Ollama, RAG, evaluation, security, and monitoring

Target Companies

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 domain knowledge or modern methodologies.

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 regarding 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.