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

Updated on 05/11/2026

Sign up

Distilabel Training

3 days (21 hours)

Overview

Distilabel is an open-source framework developed by Argilla that enables the design of pipelines for generating synthetic datasets for next-generation language models.

Our Distilabel training will enable you to master the design of data generation pipelines for LLMs, the automation of dataset creation workflows, and integration with the Argilla ecosystem, Hugging Face, and leading AI model providers.

You will learn to produce conversational datasets, create preference sets for model alignment, monitor the quality of generated data, and set up collaborative human validation workflows.

By the end of this training, you will be able to develop complete Distilabel pipelines, generate usable datasets for LLM fine-tuning and evaluation, integrate Argilla into your AI workflows, and scale synthetic data generation in a modern LLMOps-oriented environment.

Like all our training courses, this one covers the latest stable version of Distilabel and emphasizes a resolutely practical and operational approach.

Objectives

- Understand the fundamental concepts of Distilabel and Argilla
- Create synthetic dataset generation pipelines for LLMs
- Generate data for fine-tuning, RLHF, and DPO
- Evaluate and validate the quality of generative AI datasets
- Integrate Distilabel into LLMOps workflows
- Scale up AI data generation pipelines

Target audience

- LLM Engineers
- AI Engineers
- Data Scientists focused on generative AI
- ML Engineers
- MLOps Engineers
- Python Developers Specializing in AI

Prerequisites

- Basic knowledge of Python
- Basic understanding of machine learning and language models
- General knowledge of AI APIs and the Hugging Face ecosystem is a plus

Distilabel Training Program with Argilla

[Day 1 - Morning]

Introduction to Distilabel and the Argilla ecosystem

- Understanding Distilabel and its role in generating synthetic datasets for LLMs
- Exploring the Argilla, Hugging Face, and Transformers ecosystem
- Identifying use cases: fine-tuning, RLHF, DPO, RAG
- Install and configure a Python environment for Distilabel
- Understanding the concepts of Steps, Tasks, and Pipelines
- Hands-on workshop: Creating your first conversational data generation pipeline.

[Day 1 - Afternoon]

Building LLM generation pipelines

- Designing modular AI data generation pipelines
- Using prompt engineering to produce coherent datasets
- Generating multi-turn data and conversational scenarios
- Leveraging different model providers: OpenAI, Mistral, vLLM
- Structuring reusable workflows for AI projects
- Hands-on workshop: Creating an advanced instruction generation pipeline.

Preparing datasets for fine-tuning

- Understanding data formats for supervised fine-tuning
- Creating datasets tailored to conversational models
- Generating domain-specific datasets
- Cleaning and normalizing synthetic data

- Organizing datasets for training and evaluation
- Hands-on workshop: Creating a dataset ready for LLM fine-tuning.

[Day 2 - Morning]

Distilabel and LLM model alignment

- Understanding the concepts of RLHF and DPO
- Generating preference and ranking datasets
- Creating data for comparative evaluation of responses
- Structuring alignment datasets
- Automating the generation of preference data
- Hands-on workshop: Creating a preference dataset for model alignment.

[Day 2 - Afternoon]

Integration with Argilla and human validation

- Exploring Argilla's features for AI annotation
- Importing and versioning datasets in Argilla
- Organizing human validation workflows
- Annotate and correct outputs generated by LLMs
- Build a collaborative data review workflow
- Hands-on workshop: Validating and annotating a dataset in Argilla.

Evaluation and quality of AI datasets

- Measuring the quality of automatically generated data
- Detecting model hallucinations and inconsistencies
- Evaluating the diversity and relevance of responses
- Implementing LLM evaluation metrics
- Building benchmark datasets
- Hands-on workshop: Quality audit of a synthetic dataset generated with Distilabel.

[Day 3 - Morning]

Distilabel for RAG and generative AI workflows

- Generate data for chatbots and RAG systems
- Produce contextualized question-and-answer sets
- Create multilingual and specialized datasets
- Structure document generation pipelines
- Leveraging Distilabel in generative AI architectures
- Hands-on workshop: Generating an RAG dataset for a business assistant.

[Day 3 - Afternoon]

Industrialization and automation of pipelines

- Automating the execution of Distilabel pipelines
- Using Docker to standardize environments
- Versioning AI datasets and pipelines
- Monitor workflows and inference costs
- Deploying generation pipelines in production
- Hands-on workshop: Industrializing a complete AI data generation pipeline.

Best practices and overarching project

- Defining best practices for AI data generation
- Optimize pipeline performance and quality
- Securing API access and generated data
- Developing a dataset maintenance strategy
- Preparing workflows for large-scale AI projects
- Hands-on workshop: Final project on generating, validating, and utilizing a complete LLM dataset.

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 industry 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 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 the correct acquisition of the skills.

Certification

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