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

Updated on September 16, 2026

Sign Up

Qwen Training: LLM and Generative AI

2 days (14 hours)

Overview

Qwen is a family of artificial intelligence models developed by Alibaba, designed to address a variety of needs in text generation, reasoning, multimodal understanding, programming, and tool usage.

This Qwen training course will help you understand the different model families available and select the most suitable model based on your performance, cost, latency, and privacy constraints.

You'll learn how to use Qwen models with Python, master the key inference parameters, utilize reasoning and tool-calling capabilities, and integrate the models into your technical environments.

The training also covers the use of multimodal models, adapting models to specific use cases, and the main strategies for quantification and
fine-tuning.

By the end of the training, you will be able to select, configure, adapt, and deploy a Qwen model based on a business or technical need, whether via an API or on your own infrastructure.

The curriculum is regularly updated to incorporate the latest generations of Qwen models and best practices recommended by the ecosystem.

Objectives

- Understand the Qwen ecosystem and the main model families
- Select a model based on performance, cost, and infrastructure requirements
- Use Qwen with Python and the main inference frameworks
- Use the capabilities for reasoning, structured generation, and tool calling

- Understand the principles of model quantification and fine-tuning
- Deploy a Qwen model via API or on a private infrastructure

Target Audience

- Python Developer
- AI Engineer
- ML Engineer
- Data Scientist
- AI Architect
- Engineer interested in using open-weight models in a professional setting

Prerequisites

- Proficiency in Python
- Knowledge of the basic principles of LLMs and generative AI
- Basic understanding of how to use APIs and Python environments
- Basic knowledge of Docker is a plus

Technical Requirements

- Computer with Python 3 and a compatible development environment
- Access to an environment that allows you to use Qwen models via API or locally
- A Hugging Face or ModelScope account is recommended for accessing open-weight models
- Access to a machine equipped with a GPU is recommended for local deployment workshops
- Stable internet connection to download the necessary models and dependencies
- Environments, API keys, and resources required for the exercises will be provided in advance or by the instructor

Our Qwen Training Program

[Day 1 - Morning]

Understanding the Qwen ecosystem and choosing the right model

- Qwen’s Position in the Large Language Model Ecosystem
- Overview of the main model families: general-purpose, reasoning, code, and multimodal models
- Understanding the Dense and Mixture of Experts architectures used in the Qwen ecosystem
- Comparing models based on their size, context, performance, and hardware requirements
- Understanding the differences between Instruct, Reasoning, and specialized models
- Choosing between API usage and running an open-weight model on your own infrastructure
- Identify selection criteria: quality, latency, cost, privacy, and available resources

- Hands-on workshop: Compare several Qwen models and select the appropriate model for different use cases

[Day 1 - Afternoon]

Using Qwen models with Python

- Using Qwen models with Hugging Face Transformers and ModelScope
- Configure key generation parameters: temperature, top-p, context length, and generated tokens
- Understanding and using the reasoning capabilities of compatible models
- Generate structured output and work with JSON formats
- Implement tool calling to connect the model to external functions
- Leverage multimodal models to analyze text, images, and documents
- Measure resource consumption and inference performance
- Hands-on workshop: Integrate a Qwen model into a Python environment and compare multiple inference configurations

[Day 2 - Morning]

Adapt and optimize Qwen models

- Understanding when to use prompting, RAG, or fine-tuning
- Principles of supervised fine-tuning to adapt a model to a specific domain
- Introduction to LoRA approaches and lightweight adaptation methods
- Preparing and structuring a training dataset
- Understanding the principles of quantization: trade-offs between memory, speed, and accuracy
- Discover the main quantized formats and variants that can be used with Qwen
- Evaluating model performance before and after adaptation
- Hands-on workshop: Prepare a dataset and test a strategy for adapting or quantizing a Qwen model

[Day 2 - Afternoon]

Deploying and scaling a Qwen model

- Deploy a Qwen model with vLLM or SGLang
- Expose the model as an API compatible with the OpenAI protocol
- Configure reasoning, structured outputs, and server-side tool calling
- Scale resources based on model size and expected load
- Understand the principles of parallelization and inference optimization
- Compare on-premises, cloud, and managed service deployments
- Implement a strategy for testing, monitoring, and updating models
- Define a production architecture tailored to security, cost, and performance constraints
- Hands-on workshop: Deploy a Qwen model behind an API and validate its operation in a production-like environment

Target Audience

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

Pre-training 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

Hands-On Training: 60% Practical, 40% Theoretical. Training materials distributed in digital format to all participants.

Organization

The course alternates between theoretical input from the instructor—supported by examples and reflection sessions—and group work.

Assessment

At the end of the session, a multiple-choice quiz is administered to verify that participants have successfully acquired the skills.

Certification

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

Training organization registered under number 11 75 54743 75. This registration does not constitute state accreditation.
© Ambient IT 2015-2026. All rights reserved. Paris, France - Switzerland - Belgium - Luxembourg