

PROGRAMME DE FORMATION

INSTALLER ET EXPLOITER UN LLM EN LOCAL

FICHE SYNTHÉTIQUE DE LA FORMATION

Intitulé	Installer et exploiter un LLM en local — une IA générative confidentielle sur vos propres machines
Référence	IA-LLM-LOCAL-1J
Durée	1 jour (7 heures)
Modalités	Présentiel ou distanciel synchrone (classe virtuelle)
Tarif	700 € HT / participant (inter) — intra sur devis
Délais d'accès	Inscription possible jusqu'à 14 jours avant le début de la session
Lieu	Paris 17e ou distanciel via Zoom
Effectif	4 participants maximum par session
Contact	contact@pythonia.fr – 06 66 06 38 59
Accessibilité PSH	Formation accessible aux PSH. Contactez notre référent handicap pour toute adaptation.
Dernière mise à jour	Octobre 2026

1. Public visé

Cette formation s'adresse aux professionnels qui souhaitent utiliser l'IA générative sans envoyer leurs données à un service cloud : DSI et équipes techniques, développeurs, data analysts, responsables innovation, consultants, ainsi que les métiers soumis à une forte exigence de confidentialité (juridique, santé, finance, assurance, secteur public).

2. Prérequis

- Être à l'aise avec l'environnement Windows, macOS ou Linux et l'installation de logiciels
- Avoir déjà utilisé un assistant IA (ChatGPT, Claude, Mistral Le Chat...)
- Notions de base en Python recommandées pour la partie intégration (scripts fournis, non obligatoire)
- Poste avec 16 Go de RAM minimum, 30 Go d'espace disque libre et droits administrateur (une carte graphique NVIDIA ou un Mac Apple Silicon est un plus, non obligatoire)

Un test de positionnement est proposé avant l'entrée en formation afin de vérifier l'adéquation du parcours avec le profil du participant et la compatibilité de son poste.

3. Objectifs pédagogiques

À l'issue de cette formation, le stagiaire sera capable de :

- Expliquer les enjeux d'un LLM local (confidentialité, RGPD, souveraineté, coûts) et ses limites face aux offres cloud
- Dimensionner le matériel et choisir un modèle open-weight adapté à son besoin (Mistral, Llama, Qwen, Gemma...) et sa quantification
- Installer, configurer et exploiter un LLM local avec Ollama et LM Studio, et le mettre à disposition via Open WebUI
- Interroger le modèle local depuis Python grâce à son API compatible OpenAI
- Connecter le LLM à des documents internes (RAG local) en restant entièrement hors ligne

4. Programme détaillé

Chaque participant repart avec un LLM opérationnel installé sur son propre poste et un assistant documentaire fonctionnant entièrement hors ligne.

JOURNÉE — Installer et exploiter un LLM en local (7h)

Matin (3h30) — Comprendre, choisir et installer

Pourquoi un LLM local ?

- LLM cloud vs LLM local : confidentialité, RGPD, secret professionnel, AI Act, fonctionnement hors ligne
- Coûts comparés : abonnements et API vs investissement matériel
- Panorama des modèles open-weight : Mistral, Llama, Qwen, Gemma, DeepSeek, Phi — forces et usages
- Licences des modèles : ce que l'on peut faire (ou non) en usage professionnel

Dimensionner son matériel

- Paramètres d'un modèle (3B, 8B, 14B, 70B...) et mémoire nécessaire
- Quantification (Q4, Q5, Q8) et format GGUF : compromis qualité / vitesse / mémoire
- CPU, GPU NVIDIA (VRAM), Apple Silicon : quelles performances attendre (tokens/seconde)
- Trouver et évaluer un modèle sur Hugging Face

Installation et prise en main

- Installation d'Ollama sous Windows, macOS ou Linux
- Télécharger, lancer et gérer des modèles : ollama pull, run, list, rm
- Paramétrage : température, taille de contexte, prompt système, Modelfile personnalisé
- Alternative graphique : LM Studio (recherche de modèles, chat, serveur local)

Travaux pratiques :

- TP1 : installer Ollama, télécharger deux modèles de tailles différentes et comparer vitesse et qualité de réponse sur des cas métier
- TP2 : créer un assistant spécialisé avec un Modelfile (rôle, ton, consignes)

Après-midi (3h30) — Exploiter, intégrer et sécuriser

Une interface de chat pour toute l'équipe

- Déploiement d'Open WebUI (Docker) : interface type ChatGPT branchée sur le modèle local
- Comptes utilisateurs, historique des conversations, partage de prompts
- Mise à disposition sur le réseau interne de l'entreprise

Intégrer le LLM local dans ses outils

- L'API REST d'Ollama et sa compatibilité avec l'API OpenAI
- Appeler le modèle depuis un script Python (bibliothèque openai), réponses en streaming
- Cas d'usage : résumé de documents, extraction d'informations, rédaction, classification

Interroger ses documents en local (RAG)

- Principe du RAG : découpage, embeddings, recherche sémantique, génération
- Embeddings locaux (nomic-embed-text) et base vectorielle (Chroma)
- Chargement de documents PDF et Word internes, sans aucune donnée envoyée sur Internet

Sécurité, limites et bonnes pratiques

- Exposition réseau, contrôle d'accès, mises à jour des modèles et des outils
- Hallucinations et vérification des réponses ; quand rester sur un LLM cloud
- Feuille de route pour un déploiement en entreprise (poste individuel → serveur partagé)

Travaux pratiques :

- TP3 : déployer Open WebUI et l'utiliser avec le modèle local
- TP4 : construire un assistant qui répond aux questions sur un corpus de documents de l'entreprise, entièrement hors ligne

✓ **Évaluation sommative : QCM final de 15 questions et validation des TP — critère de réussite : 60 %**

5. Méthodes et moyens pédagogiques

Méthodes mobilisées

- Apports théoriques : exposés interactifs avec diaporama et démonstrations en direct
- Mise en pratique : installation guidée sur le poste de chaque participant, TP progressifs
- Pédagogie active : cas d'usage apportés par les participants
- Alternance théorie/pratique : ratio 30 % théorie / 70 % pratique

Moyens techniques

- Salle de formation équipée : vidéoprojecteur, paperboard, connexion Wi-Fi haut débit
- Poste informatique par stagiaire (ou poste personnel répondant aux prérequis)
- En distanciel : Zoom avec partage d'écran, tableau blanc collaboratif
- Supports de cours, scripts Python et fiches d'installation remis aux stagiaires
- Jeu de documents d'exemple fourni pour le TP RAG

6. Modalités d'évaluation

Positionnement à l'entrée

Avant l'entrée en formation, chaque participant complète un test de positionnement évaluant son aisance informatique, ses usages de l'IA générative et la configuration de son poste.

Évaluations en cours de formation

- Évaluation formative : vérification du bon fonctionnement de chaque installation en fin de matinée
- Évaluation sommative : QCM final de 15 questions et réalisation du TP4
- Critère de réussite : obtenir 60 % au QCM final
- Attestation de fin de formation : remise à l'issue, mentionnant objectifs, durée et résultats

Satisfaction

- Questionnaire de satisfaction en fin de formation (formateur, contenu, logistique, atteinte des objectifs)

7. Accessibilité aux personnes en situation de handicap

Pythonia s'engage à rendre ses formations accessibles aux personnes en situation de handicap. Notre référent handicap est disponible pour étudier toute demande d'adaptation :

- Adaptation des supports pédagogiques (taille de police, contraste, format audio)
- Aménagement des conditions de passage (temps majoré, pauses adaptées)
- Orientation vers nos partenaires spécialisés : Agefiph, Cap Emploi, MDPH

Contact référent handicap : contact@pythonia.fr – 06 66 06 38 59

8. Engagement et suivi

- Suivi de l'assiduité : feuille d'émargement signée par demi-journée
- Prévention des abandons : suivi individualisé, adaptation du rythme, relance en cas d'absence
- Réclamations : toute réclamation peut être adressée par email à contact@pythonia.fr. Elle sera traitée sous 48h ouvrées.

PythonIA — Centre de Formation Python & Intelligence Artificielle

255 boulevard Pereire, 75017 Paris | contact@pythonia.fr | 06 66 06 38 59