DAWAN

Plan de la formation

www.dawan.fr

Formation Déployer Ollama et des modèles IA open source (Mistral…)

Durée :	2 jours (14 heures)
Tarifs inter-entreprise :	1 875,00 € HT (standard) 1 500,00 € HT (remisé)
Public :	Administrateurs systèmes et DevOps, développeurs back-end, architectes techniques, responsables ou référents IA souhaitant déployer un moteur IA en interne
Pré-requis :	Bonnes bases en environnement Linux (ligne de commande), notions de réseau (ports, firewall, proxy), premières notions sur l'IA générative et les API web
Objectifs :	Comprendre l'architecture d'Ollama et des modèles open source (Mistral…) en contexte entreprise - Savoir installer, configurer et sécuriser Ollama sur un serveur on-premise ou cloud - Être capable d'exposer un ou plusieurs modèles IA aux applications internes (API) en maîtrisant les aspects performance, monitoring et gouvernance
Modalités pédagogiques, techniques et d'encadrement :	• Formation synchrone en présentiel et distanciel. • Méthodologie basée sur l'Active Learning : 75 % de pratique minimum. • Un PC par participant en présentiel, possibilité de mettre à disposition en bureau à distance un PC et l'environnement adéquat. • Un formateur expert.
Modalités d'évaluation :	• Définition des besoins et attentes des apprenants en amont de la formation. • Auto-positionnement à l'entrée et la sortie de la formation. • Suivi continu par les formateurs durant les ateliers pratiques. • Évaluation à chaud de l'adéquation au besoin professionnel des apprenants le dernier jour de formation.

Sanction :	Attestation de fin de formation mentionnant le résultat des acquis
Référence :	INT102822-F
Note de satisfaction des participants:	Pas de données disponibles
Contacts :	commercial@dawan.fr - 09 72 37 73 73
Modalités d'accès :	Possibilité de faire un devis en ligne (www.dawan.fr, moncompteformation.gouv.fr, maformation.fr, etc.) ou en appelant au standard.
Délais d'accès :	Variable selon le type de financement.
Accessibilité :	Si vous êtes en situation de handicap, nous sommes en mesure de vous accueillir, n'hésitez pas à nous contacter à referenthandicap@dawan.fr, nous étudierons ensemble vos besoins

Comprendre Ollama et les modèles IA open source

Panorama des modèles IA open source (Mistral, LLaMA, autres) et cas d'usage en entreprise
Positionnement d'Ollama : moteur local, orchestrateur de modèles, alternative aux API cloud
Principes de base des LLM : tokens, prompts, contextes, limites de taille et coûts indirects
Choisir un modèle en fonction des besoins (génération de texte, chat, résumé, code, etc.)
Atelier pratique : prise en main d'Ollama en local sur poste de travail (installation rapide et premier prompt avec un modèle Mistral)

Préparer l'infrastructure pour un déploiement en entreprise

Préfigurer l'architecture cible : serveur unique, cluster, intégration avec les SI existants
Pré-requis matériels : CPU vs GPU, RAM, stockage, bande passante, dimensionnement de base
Environnement système : choix de la distribution Linux, gestion des utilisateurs, droits sudo
Contraintes réseau : ports, reverse proxy, SSL/TLS, prise en compte des proxies d'entreprise
Atelier pratique : définir une architecture cible et une fiche de dimensionnement pour un premier serveur Ollama d'entreprise

Installer et configurer Ollama sur un serveur

Installation d’Ollama sur un serveur Linux : étapes clés et vérifications indispensables
Gestion du service (systemd) et démarrage automatique, logs et fichiers de configuration
Configuration des options de base : ports d’écoute, répertoires de modèles, accès réseau
Mise en place d’un reverse proxy (nginx ou équivalent) pour exposer Ollama en HTTPS
Atelier pratique : installation complète d’Ollama sur un serveur de test et vérification d’accès via API depuis un poste client

Gérer les modèles Mistral et autres modèles IA

Télécharger, installer et mettre à jour un modèle Mistral avec Ollama (pull, list, remove)
Comprendre l’impact des différentes tailles de modèles (7B, 8B, 12B, etc.) sur la performance
Configurer les paramètres d’inférence : température, top-p, context window, temps de réponse
Stratégies de stockage des modèles et des caches, organisation par environnement (test, prod)
Atelier pratique : déployer plusieurs variantes de modèles Mistral, comparer leurs temps de réponse et ajuster les paramètres d’inférence

Exposer Ollama aux applications internes

Découvrir l’API HTTP d’Ollama : endpoints principaux, formats de requêtes et réponses
Intégrer Ollama dans une application interne (exemples en Python ou Node.js)
Mettre en place des clés d’API ou un proxy d’authentification pour sécuriser l’accès
Bons réflexes pour gérer les files d’attente, les timeouts et la résilience des appels
Atelier pratique : développer un petit service interne (chat ou complétion de texte) connecté à un modèle Mistral déployé sur Ollama

Sécuriser et gouverner l’usage d’Ollama en entreprise

Identifier les risques : fuites de données, prompts sensibles, dérives d’usage
Mettre en place des restrictions d’accès par réseau, authentification et journalisation
Bonnes pratiques pour l’usage de données internes : anonymisation, jeux de test, règles internes
Éléments de conformité (RGPD, confidentialité, localisation des données, journalisation des accès)
Atelier pratique : définir une politique d’usage et un document de bonnes pratiques IA pour les utilisateurs internes

Supervision, optimisation et maintenance du service Ollama

Surveiller l’activité : logs, métriques de performance, consommation CPU/GPU/RAM
Mettre en place un monitoring basique (export de métriques, tableaux de bord simples)
Planifier les mises à jour d’Ollama et des modèles tout en limitant l’indisponibilité
Stratégies de sauvegarde et de restauration : configurations, scripts, documentation technique
Atelier pratique : construire un mini plan d’exploitation (monitoring, sauvegarde, mises à jour) pour le serveur Ollama d’entreprise

Bilan et plan d’actions pour le déploiement en production

Synthèse des choix techniques et organisationnels réalisés pendant la formation
Identifier les évolutions possibles : RAG, connexion à une base documentaire, multi-modèles
Planifier un pilote interne avec un petit groupe d’utilisateurs et des cas d’usage concrets
Formaliser les prochaines étapes : responsabilités, jalons, ressources nécessaires
Atelier pratique : élaborer une feuille de route personnalisée pour passer du serveur de test à un service Ollama opérationnel en production