

Sessions garanties disponibles

Formation Ollama : déployer et optimiser des LLM en local

Réf. SF34535

Maîtrisez le déploiement de modèles d'IA en local pour garantir la confidentialité de vos données tout en optimisant les performances !

3 j
Durée2 700 € HT
Inter / participant6 300 € HT
Intra (≤12 pers.)13
Sessions à venir

Objectifs pédagogiques

- ✓ Identifier les avantages du déploiement de LLM en local par rapport aux solutions cloud
- ✓ Installer et configurer Ollama
- ✓ Personnaliser des modèles de langage via les Modelfiles pour répondre à des besoins spécifiques
- ✓ Intégrer Ollama dans des applications via l'API REST et des frameworks comme LangChain
- ✓ Optimiser les performances des modèles par la quantization, le batching et le caching
- ✓ Sécuriser un déploiement Ollama avec Docker, le chiffrement et le contrôle d'accès

Public visé & prérequis

Public visé Data scientists, Développeurs, Ingénieurs DevOps**Prérequis**- Connaissances de base en intelligence artificielle et en modèles de langage - Pratique courante de la ligne de commande (Terminal, PowerShell) - Bases en Python ou langage équivalent - Notions de Docker (images, conteneurs)

Programme détaillé

Introduction à l'IA locale et aux modèles open-source

- Avantages de l'IA locale : confidentialité des données, réduction des coûts et maîtrise de la latence face aux solutions cloud
- Panorama des LLM open-source : positionnement et caractéristiques des modèles Llama, Mistral et Gemma
- Exigences matérielles : impact de la RAM, du GPU et de l'architecture sur les performances des modèles
- Architecture d'Ollama : fonctionnement interne, gestion des modèles et cas d'usage en entreprise

Activités pratiques :

- Comparaison des performances et coûts entre un LLM local et une API cloud sur un même jeu de prompts

Installation et configuration d'Ollama

- Installation multi-plateforme : procédures spécifiques pour macOS, Windows et Linux
- Dépendances et accélération GPU : configuration de Docker et des drivers CUDA pour exploiter le matériel disponible
- Bonnes pratiques de configuration : organisation des modèles, gestion de l'espace disque et variables d'environnement
- Diagnostic et dépannage : résolution des problèmes courants d'installation et de compatibilité

Activités pratiques :

- Installation d'Ollama et lancement d'un premier modèle léger
- Diagnostic et résolution de problèmes d'installation simulés

Utilisation et gestion des modèles

- Interface CLI : commandes essentielles pour télécharger, lancer et gérer les modèles
- Interface WebUI : interaction graphique avec les modèles pour le prototypage rapide
- Gestion du cycle de vie : téléchargement, mise à jour et suppression des modèles

- Techniques de prompting : rédaction de prompts efficaces et tests comparatifs entre modèles

Activités pratiques :

- Lancement de plusieurs modèles et comparaison de leurs réponses sur des prompts variés

Personnalisation des modèles avec les Modelfiles

- Modelfiles : syntaxe et structure pour définir des modèles personnalisés
- Paramétrage avancé : ajustement de la température, du contexte et des instructions système
- Adaptation métier : spécialisation d'un modèle pour un domaine ou une tâche spécifique
- Fine-tuning de base : principes et limites de l'ajustement de modèles en local

Activités pratiques :

- Création d'un Modelfile pour spécialiser un LLM sur la génération de code Python
- Personnalisation d'un modèle avec des instructions système adaptées à un contexte métier

Intégration applicative via l'API REST et LangChain

- API REST Ollama : endpoints disponibles, formats de requêtes et gestion des réponses en streaming
- Intégration Python : appels programmatiques à l'API pour intégrer Ollama dans des applications existantes
- LangChain et Ollama : configuration du connecteur, création de chaînes de traitement et gestion de la mémoire
- Création d'agents conversationnels : architecture d'un chatbot exploitant un modèle local

Activités pratiques :

- Développement d'un chatbot simple intégrant l'API Ollama
- Construction d'une chaîne LangChain connectée à un modèle local

Optimisation des performances

- Quantization : réduction de la taille des modèles et impact sur la qualité des réponses
- Batching : traitement par lots pour améliorer le débit dans les scénarios multi-utilisateurs
- Stratégies de caching : mise en cache des résultats pour accélérer les réponses récurrentes
- Monitoring : suivi des métriques de performance (temps de réponse, consommation mémoire, utilisation GPU)

Activités pratiques :

- Application de la quantization sur un modèle large et mesure de l'impact sur les performances et la qualité
- Mise en place d'un monitoring de base pour suivre l'utilisation des ressources

Sécurisation et déploiement en environnement sensible

- Isolation avec Docker : conteneurisation d'Ollama pour limiter la surface d'attaque
- Chiffrement des données : protection des modèles et des échanges en transit et au repos
- Contrôle d'accès : gestion des permissions et restriction de l'utilisation des modèles
- Cas d'usage en secteurs réglementés : déploiement dans des contextes santé, finance ou défense

Activités pratiques :

- Déploiement d'Ollama dans un conteneur Docker sécurisé avec contrôle d'accès
- Développement d'un agent d'analyse de texte confidentiel exploitant un modèle local isolé

Prochaines sessions

DATE	LIEU	STATUT
30/11/2026	à distance	Sur confirmation
16/03/2027	Grenoble	Sur confirmation
16/03/2027	Paris	Sur confirmation
16/03/2027	Toulouse	Sur confirmation
16/03/2027	à distance	Sur confirmation

DATE	LIEU	STATUT
23/06/2027	Lyon	Sur confirmation

Pourquoi choisir Sparks Formation ?

20 ans d'expertise IT, des formateurs experts et des sessions garanties contre l'annulation.

20 ans
d'expertise

1 068+
formations

11 178
apprenants formés/an

+67
Net Promoter Score

Avis des participants

4,9/5 — 190 avis

Anne-Lise C. — Très bonne adaptation du formateur au public et au niveau d'attente.

Marie G. — La formatrice était très pédagogue et bienveillante ! Elle a su nous partager sa passion et donner envie de pratiquer à la suite Voilà bien longtemps que je...

Votre formateur

Camilo R. — Formateur

Camilo intervient notamment sur l'IA appliquée aux métiers et sa gouvernance, ainsi que sur l'ingénierie des LLM et l'IA générative. Son activité de formation a débuté en 2024. La conception...

Ils nous font confiance

Cette année notre filière Intelligence artificielle & Data Science a accompagné plus de 544 entreprises dont :

SOPRA STERIA THALES GLOBAL SERVICES ARTELIA AXA ASSISTANCE FRANCE GRANT THORNTON FRANCE AKKODIS DIGITAL SAS
MEDIABRANDS CEGEDIM COVAGE INFRA SNCF - E.SNCF SOLUTIONS +534 autres

Financement & certifications

Cette formation est éligible aux financements OPCO et FNE-Formation. Nos équipes vous accompagnent dans le montage de votre dossier.



Réservez votre place dès maintenant

20 ans d'expertise · 1 000+ formations · NPS +67 · réponse sous 24h ouvrées

0 805 950 800

demande@sparks-formation.com