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

Mis à jour le 12/06/2026

S'inscrire

Formation Small Language Models avec Hugging Face et Ollama

2 jours (14 heures)

Présentation

Small Language Models désigne des modèles de langage compacts capables de traiter des tâches ciblées avec moins de ressources qu'un grand LLM généraliste. Associés à Hugging Face et Ollama, ils permettent de créer des assistants IA plus rapides, plus privés, plus sobres et plus économiques.

Notre formation Small Language Models avec Hugging Face et Ollama vous permettra de sélectionner, exécuter, personnaliser et intégrer des modèles de langage compacts dans des cas d'usage concrets d'entreprise.

Vous apprendrez à explorer le Hugging Face Hub, lire les model cards, comparer plusieurs modèles, analyser les licences, les résultats d'évaluation et les contraintes de déploiement local ou privé.

À l'issue de la formation, vous serez en mesure d'exécuter des modèles avec Ollama, de créer un assistant spécialisé avec un Modelfile, de connecter un SLM à une application et de mettre en place un premier RAG frugal sur des documents internes.

Cette formation aborde également l'évaluation des réponses, la sécurité, la gestion des données sensibles, la documentation du choix de modèle et les bonnes pratiques d'industrialisation pour un usage professionnel des SLM.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Comprendre les usages, bénéfices et limites des Small Language Models
- Sélectionner et comparer des SLM sur Hugging Face Hub
- Exécuter et personnaliser des modèles localement avec Ollama
- Créer une application connectée à un modèle local via l'API Ollama
- Construire un RAG frugal avec documents internes, embeddings et SLM
- Évaluer, sécuriser et industrialiser l'usage des SLM en entreprise

Public visé

- Développeurs et tech leads
- Ingénieurs IA et ML engineers
- Data scientists et MLOps engineers
- Architectes IA et architectes cloud
- Équipes innovation souhaitant déployer des modèles IA locaux ou privés

Pré-requis

- Connaissances générales en IA générative ou modèles de langage
- Notions de Python, JavaScript ou TypeScript appréciées
- Compréhension de base des APIs, fichiers JSON et environnements de développement

Pré-requis techniques

- Disposer d'un ordinateur avec Linux, macOS ou Windows avec WSL2
- Créer un compte Hugging Face pour accéder aux modèles, datasets et espaces privés
- Prévoir une connexion Internet stable pour télécharger les modèles et dépendances

Notre programme de formation Small Language Models avec Hugging Face et Ollama

[Jour 1 - Matin]

Comprendre les Small Language Models et leurs cas d'usage

- Comprendre ce qu'est un Small Language Model et le différencier d'un LLM généraliste
- Identifier les bénéfices des SLM : latence réduite, coûts d'inférence maîtrisés, confidentialité et déploiement local
- Comprendre les cas d'usage adaptés : classification, extraction, résumé, assistant métier, chatbot interne et automatisation documentaire
- Identifier les limites des modèles compacts : raisonnement complexe, contexte long, hallucinations, robustesse et spécialisation métier
- Positionner les SLM dans une architecture d'entreprise : poste local, serveur interne, edge, cloud privé ou API applicative

- Atelier pratique : identifier les cas d’usage d’entreprise pertinents pour un SLM et définir les critères de choix du modèle

[Jour 1 - Après-midi]

Sélectionner un modèle avec Hugging Face

- Utiliser Hugging Face Hub pour rechercher des modèles de langage compacts adaptés à un besoin métier
- Lire une model card : usages prévus, limites, licence, données d’entraînement, évaluation et contraintes de déploiement
- Comparer les familles de modèles : Phi, Gemma, Qwen, Llama, Mistral, SmolLM et modèles instruct
- Analyser les critères de sélection : taille, licence, langue, contexte, format, performance, ressources et compatibilité locale
- Exploiter les benchmarks, leaderboards et métriques publiées pour comparer plusieurs modèles candidats
- Atelier pratique : sélectionner plusieurs modèles sur Hugging Face et construire une grille comparative pour un cas métier

Exécuter et personnaliser un SLM avec Ollama

- Installer et configurer Ollama sur un poste local ou un serveur de test
- Télécharger, lancer et tester un modèle compatible avec la CLI Ollama
- Utiliser les principales commandes : pull, run, list, show, remove et serve
- Comprendre les paramètres d’inférence : température, top_p, contexte, seed, répétition et nombre de tokens
- Créer un assistant spécialisé avec un Modelfile, un prompt système et des paramètres adaptés
- Atelier pratique : lancer un SLM avec Ollama, créer un assistant métier local avec Modelfile et tester plusieurs stratégies de prompt

[Jour 2 - Matin]

Intégrer un SLM dans une application d’entreprise

- Utiliser l’API locale d’Ollama pour connecter un SLM à une application Python, JavaScript ou TypeScript
- Créer une API interne pour exposer un modèle local à une application métier
- Gérer les sessions, prompts, paramètres, erreurs, timeouts et réponses structurées
- Mettre en place des formats de sortie exploitables : JSON, résumé structuré, classification ou extraction d’informations
- Ajouter des garde-fous simples : validation d’entrée, filtrage de sortie, logs, restrictions d’usage et gestion des données sensibles
- Atelier pratique : créer un mini-service applicatif qui interroge un SLM local via l’API Ollama

[Jour 2 - Après-midi]

RAG frugal avec Hugging Face, Ollama et documents internes

- Comprendre le rôle du RAG pour connecter un SLM à des documents d’entreprise
- Concevoir une architecture RAG frugal : ingestion, découpage, embeddings, recherche vectorielle et génération
- Choisir entre modèle local, embeddings locaux, embeddings hébergés et base vectorielle selon les contraintes de confidentialité
- Réduire les coûts et la latence : taille des chunks, nombre de documents récupérés, contexte utile et prompts courts
- Limiter les hallucinations avec grounding, citations, sources et consignes de refus
- Atelier pratique : construire une mini-application RAG locale avec documents internes, embeddings et SLM exécuté via Ollama

Évaluation, sécurité et industrialisation des SLM

- Construire un jeu d’évaluation avec prompts, réponses attendues, cas limites et exemples métier
- Comparer plusieurs SLM selon qualité, latence, consommation mémoire, taille du modèle et coût d’exploitation
- Identifier les risques : hallucinations, biais, données sensibles, réponses non conformes et dépendance aux modèles
- Documenter le choix du modèle : licence, limites, cas d’usage, benchmarks, paramètres et règles d’utilisation
- Définir une architecture cible : poste local, serveur interne, container, API privée, edge ou environnement cloud
- Atelier pratique : concevoir une architecture SLM d’entreprise avec Hugging Face, Ollama, RAG, évaluation, sécurité et supervision

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.