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

Mis à jour le 12/05/2026

S'inscrire

Formation LLaMA-Factory et Unsloth

3 jours (21 heures)

Présentation

LLaMA-Factory et Unsloth sont aujourd'hui deux outils incontournables pour industrialiser le fine-tuning de modèles de langage open source. Grâce à leurs optimisations avancées, ils permettent d'entraîner rapidement des LLM performants tout en réduisant fortement les besoins matériels et les coûts GPU.

Notre formation LLaMA-Factory et Unsloth vous permettra de maîtriser l'ensemble de la chaîne de fine-tuning LLM : préparation des datasets, entraînement optimisé, quantization, tuning avancé, évaluation et déploiement des modèles.

Vous apprendrez à exploiter efficacement les techniques modernes de QLoRA, PEFT et 4-bit training afin d'optimiser vos infrastructures IA.

À l'issue de cette formation, vous serez en mesure de développer des pipelines complets de fine-tuning IA générative, d'industrialiser des modèles open source et d'optimiser les coûts d'entraînement et d'inférence.

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

Objectifs

- Comprendre les architectures des LLM open source
- Maîtriser les techniques de fine-tuning modernes
- Utiliser efficacement LLaMA-Factory et Unsloth
- Optimiser les entraînements GPU avec QLoRA
- Construire des datasets conversationnels métier
- Déployer des modèles IA générative en production

Public visé

- Data Scientists
- Ingénieurs IA / Machine Learning
- Développeurs Python
- Ingénieurs MLOps
- Architectes IA

Pré-requis

- Bonnes connaissances de Python
- Connaissances générales en Machine Learning
- Notions sur PyTorch
- Expérience des environnements GPU recommandée

Pré-requis techniques :

- Ordinateur portable avec 16 Go de RAM minimum, 32 Go recommandés
- Accès à un environnement GPU NVIDIA compatible CUDA, idéalement avec 16 Go de VRAM minimum
- Espace disque disponible de 50 Go minimum pour les modèles, datasets et checkpoints
- Connexion internet stable pour télécharger les modèles, dépendances et datasets

Programme de notre formation LLaMA-Factory et Unsloth : Fine-tuning LLM

[Jour 1 - Matin]

Environnement LLM open source et stratégies de fine-tuning

- Comprendre les architectures LLM open source
- Panorama des modèles LLaMA, Mistral, Qwen et Gemma
- Différences entre prétraining, instruction tuning et fine-tuning
- Introduction aux techniques LoRA, QLoRA et PEFT
- Préparer un environnement GPU avec CUDA, PyTorch et Transformers
- Atelier pratique : Installation complète de LLaMA-Factory et validation GPU.

[Jour 1 - Après-midi]

Prise en main de LLaMA-Factory

- Architecture et fonctionnement de LLaMA-Factory
- Gestion des datasets conversationnels
- Formats Alpaca, ShareGPT et datasets personnalisés

- Paramétrage des pipelines de fine-tuning
- Monitoring des entraînements et gestion des checkpoints
- Atelier pratique : Fine-tuning d’un premier modèle conversationnel.

[Jour 2 - Matin]

Optimisation mémoire et performances avec Unsloth

- Comprendre les limites GPU et VRAM
- Présentation de Unsloth et des optimisations kernel
- Accélération du fine-tuning avec Flash Attention
- Réduction mémoire avec 4-bit quantization
- Comparaison performances classiques vs optimisées
- Atelier pratique : Optimisation d’un entraînement QLoRA avec Unsloth.

[Jour 2 - Après-midi]

Fine-tuning avancé et jeux de données métier

- Construction de datasets métier spécialisés
- Nettoyage et validation des données d’entraînement
- Techniques d’augmentation et de génération synthétique
- Gestion des biais et hallucinations
- Réglage des hyperparamètres critiques
- Atelier pratique : Création d’un dataset métier et entraînement spécialisé.

[Jour 3 - Matin]

Évaluation, tests et alignment des modèles

- Méthodes d’évaluation qualitative et quantitative
- Benchmarks et métriques LLM
- Introduction au RLHF et à l’alignement
- Détection des dérives et hallucinations
- Optimisation des prompts système
- Atelier pratique : Évaluation comparative de plusieurs modèles fine-tunés.

[Jour 3 - Après-midi]

Industrialisation et déploiement des LLM

- Export des modèles fine-tunés
- Déploiement avec vLLM, Ollama et Text Generation Inference
- Gestion des coûts GPU et stratégie d’inférence

- Introduction aux pipelines MLOps LLM
- Sécurisation et gouvernance des modèles IA
- Atelier pratique : Déploiement d’un modèle fine-tuné en API locale

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.