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

Mis à jour le 26/02/2026

S'inscrire

Formation LLMOps : Inférence vLLM & Accélération Unsloth

3 jours (21 heures)

Présentation

LLM Ops : Inférence vLLM & Accélération Unsloth vous apprend à servir des modèles de langage en production avec une latence réduite et un coût GPU maîtrisé. Vous verrez comment déployer une API d'inférence, optimiser le débit (throughput) et accélérer le fine-tuning pour des cas d'usage RAG, chat interne ou assistants métiers.

La finalité est de passer d'un notebook à un service robuste : choix du runtime, configuration GPU, gestion du batching, streaming, quantization et observabilité. Nous comparons les compromis qualité/performance et les patterns d'architecture (API, worker, file d'attente) adaptés aux charges réelles.

L'approche est 100% pratique : ateliers guidés, démos reproductibles et scripts prêts à réutiliser. Livrables : un serveur vLLM opérationnel, un pipeline Unsloth pour fine-tuning rapide, et une checklist de mise en production (tests, métriques, limites, sécurité).

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

Objectifs

- Déployer un serveur d'inférence vLLM avec API et streaming.
- Optimiser le throughput via batching, KV cache et paramètres GPU.
- Mettre en place la quantization et évaluer l'impact qualité/latence.
- Accélérer un fine-tuning avec Unsloth (LoRA/QLoRA) et valider les gains.
- Industrialiser le run : logs, métriques, tests de charge et limites d'usage.

Public visé

- ML Engineers / LLM Engineers
- Data Scientists souhaitant passer en production
- DevOps / SRE impliqués dans le déploiement GPU
- Développeurs backend intégrant des LLM via API

Pré-requis

- Python (environnements, packages, scripts)
- Notions de Transformers, tokenisation, embeddings
- Base Linux/CLI et gestion de processus
- Compréhension API HTTP/JSON et logs applicatifs

Pré-requis techniques

- Machine Linux ou Windows avec WSL2, ou macOS (GPU NVIDIA recommandé)
- 16 Go RAM minimum, 32 Go conseillé
- GPU NVIDIA conseillé (CUDA) avec 12 Go VRAM minimum, 24 Go recommandé
- Python 3.10+ et outils : Git, terminal, éditeur de code
- Accès à un environnement GPU (local ou distant) pour les ateliers

Programme de formation LLMOps : Inférence vLLM & Accélération Unsloth

[Jour 1 - Matin]

Fondamentaux LLM Ops et architecture d’inférence

- Clarifier les objectifs latence, débit, coût et qualité (SLO/SLA)
- Comprendre le pipeline d’inférence : tokenisation, prefill, decode, KV cache
- Choisir un format de modèle : HF Transformers, GGUF, AWQ/GPTQ (impacts perf/qualité)
- Préparer l’environnement GPU : drivers, CUDA, PyTorch, vérifications et diagnostics
- Atelier pratique : Mesurer une baseline d’inférence (latence p50/p95, tokens/s) sur un modèle HF.

[Jour 1 - Après-midi]

Mise en place de vLLM pour servir un LLM en production

- Installer et configurer vLLM (versions, compatibilités GPU, paramètres clés)
- Démarrer un serveur OpenAI-compatible : endpoints, modèles, limites et timeouts
- Optimiser le throughput avec PagedAttention, batching et gestion du KV cache
- Gérer la concurrence : files d’attente, quotas, backpressure et stratégies de rejet
- Atelier pratique : Déployer un serveur vLLM et lancer un test de charge (concurrence + tokens/s).

[Jour 2 - Matin]

Accélération vLLM : quantification, parallélisme et tuning

- Choisir une stratégie de quantification : FP16/BF16 vs INT8/INT4 (qualité, VRAM, perf)
- Configurer le tensor parallelism et la répartition multi-GPU (contraintes et gains)
- Ajuster les paramètres vLLM : max model len, max num seqs, block size, swap et limites mémoire
- Mettre en place des benchmarks reproductibles : prompts, tailles de batch, métriques et comparaisons
- Atelier pratique : Comparer 2 configurations vLLM (quantifiée vs non quantifiée) et documenter les gains.

[Jour 2 - Après-midi]

Unsloth : fine-tuning LoRA rapide et préparation au serving

- Comprendre Unsloth : objectifs, accélérations, limites et modèles compatibles
- Préparer un dataset d'instruction : formats, nettoyage, split, contrôles qualité
- Lancer un fine-tuning LoRA/QLoRA : hyperparamètres, VRAM, stabilité et checkpoints
- Exporter et packager : merge LoRA, sauvegarde HF, versioning et traçabilité des artefacts
- Atelier pratique : Fine-tuner un modèle avec Unsloth (QLoRA) puis exporter un artefact prêt à servir.

[Jour 3 - Matin]

Intégration : servir un modèle fine-tuné avec vLLM

- Charger un modèle fine-tuné : chemins, configs, tokenizer et compatibilité vLLM
- Paramétrer la génération : temperature, top_p, max_tokens, stop sequences et garde-fous
- Mettre en place une stratégie de prompts : templates, system prompts, et tests de non-régression
- Valider la qualité : jeux de tests, scoring simple, comparaisons avant/après fine-tuning
- Atelier pratique : Déployer le modèle Unsloth dans vLLM et exécuter une suite de tests de régression.

[Jour 3 - Après-midi]

Exploitation LLM Ops : observabilité, sécurité et industrialisation

- Mettre en place l'observabilité : métriques (latence, tokens/s, erreurs), logs structurés et traces
- Gérer les coûts : limites par client, cache, politiques de timeouts et dimensionnement GPU

- Sécuriser l’API : authentification, rate limiting, filtrage d’entrées et protection contre abus
- Industrialiser : versioning modèles, rollback, canary, et runbooks d’exploitation
- Atelier pratique : Construire un mini-runbook (alertes + actions) et une checklist de mise en production.

Introduction au Deep Learning

Formation Pytorch

Formation Tensorflow

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.