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

Mis à jour le 09/01/2026

S'inscrire

Formation vLLM : Déploiement et Optimisation

3 jours (21 heures)

Présentation

vLLM est un moteur de serving de modèles de langage haute performance, pensé pour les environnements de production. Grâce à des innovations comme PagedAttention et le continuous batching, il permet d'exécuter des modèles tels que Llama ou Mistral avec une efficacité optimale et une consommation mémoire maîtrisée.

Cette formation vLLM : Déploiement & Optimisation vous guide pas à pas dans la mise en place d'une infrastructure de serving fiable, scalable et observée. Vous apprendrez à déployer vLLM sur des environnements Kubernetes et Cloud, à en surveiller les performances, et à automatiser vos déploiements avec des outils CI/CD modernes.

L'approche pédagogique, orientée terrain, alterne apports techniques et ateliers pratiques pour vous permettre de comprendre, paramétrer et exploiter efficacement un service de génération basé sur LLM.

Vous découvrirez comment intégrer vLLM à vos applications existantes, optimiser les performances, réduire les coûts d'exploitation et garantir la sécurité ainsi que la conformité de votre environnement de production.

À l'issue de la formation, vous serez capable de concevoir, déployer et superviser un environnement complet de serving vLLM, tout en adoptant une démarche professionnelle d'industrialisation et de pilotage de la performance.

Comme toutes nos formations, celle ci s'appuie sur la dernière version à jour de vLLM.

Objectifs

- Déployer un service vLLM robuste et prêt pour la production.
- Optimiser la latence, le débit et la consommation GPU.
- Superviser le service grâce à Prometheus et Grafana.

- Automatiser les déploiements avec CI/CD et GitOps.
- Appliquer les meilleures pratiques de sécurité et de conformité.

Public visé

- Ingénieurs DevOps et MLOps
- Architectes Cloud
- Ingénieurs IA/ML
- SRE

Pré-requis

- Bonnes notions de Docker et de Kubernetes
- Connaissances de base en Python et en administration Linux
- Accès à un environnement Cloud ou à une machine GPU pour les travaux pratiques

Programme de formation vLLM : Déploiement et Optimisation

[Jour 1 - Matin]

Découvrir vLLM et son écosystème

- Présentation du rôle de vLLM dans l’architecture de serving de modèles de langage
- Principes de fonctionnement : PagedAttention, continuous batching et gestion du KV cache
- Compatibilités et intégrations : API OpenAI-compatible, modèles Llama et Mistral
- Usages professionnels : assistants internes, chatbots, moteurs conversationnels et analyse de texte
- Contraintes techniques : gestion du GPU, drivers, runtime (CUDA ou équivalent) et dépendances Python
- Atelier pratique : Installer vLLM et exécuter une première requête d’inférence.

[Jour 1 - Après-midi]

Installer et configurer un environnement fiable

- Création d’un environnement Docker ou virtuel stable et maintenable
- Configuration des paramètres clés
- Gestion des modèles : téléchargement, licences et stockage efficace
- Sécurisation de la configuration réseau et du déploiement
- Tests de validation et vérification des performances initiales

Maîtriser l’architecture et les modes d’exécution

- Compréhension des modes mono et multi-GPU
- Répartition du calcul et parallélisation des requêtes
- Réduction de la latence avec le prefix caching et le préchargement des modèles
- Mesure des performances : temps de réponse, taux d'erreur et consommation mémoire
- Bonnes pratiques d'exploitation continue et gestion des mises à jour
- Atelier pratique : Observer et analyser les métriques d'exécution avec un jeu de requêtes contrôlé.

[Jour 2 - Matin]

Déployer vLLM dans Kubernetes

- Création et configuration des manifests Deployment, Service et HorizontalPodAutoscaler
- Paramétrage des nœuds GPU, gestion des sélecteurs et des tolérances
- Stockage des modèles dans des volumes persistants (S3, GCS, PVC)
- Stratégies de déploiement : Rolling Update, Blue/Green et Canary
- Gestion des redémarrages et supervision des pods critiques
- Atelier pratique : Déployer vLLM sur un cluster Kubernetes et vérifier sa montée en charge.

[Jour 2 - Après-midi]

Surveiller, fiabiliser et maîtriser les coûts

- Collecte et visualisation des métriques avec Prometheus et Grafana
- Configuration d'alertes sur la latence, les erreurs et l'utilisation GPU
- Suivi des SLO et SLI pour évaluer la fiabilité du service
- Optimisation financière : ajustement du dimensionnement et des stratégies de batching
- Construction de tableaux de bord de pilotage

Optimiser la performance et la scalabilité

- Équilibrer le débit (throughput) et la latence pour un service réactif
- Optimiser les paramètres d'exécution : batching, cache et nombre de workers
- Utiliser la quantization pour réduire les coûts sans perte significative
- Réaliser des benchmarks comparatifs et interpréter les résultats
- Atelier pratique : Mettre en œuvre une session de tests de charge et ajuster les paramètres en conséquence.

[Jour 3 - Matin]

Connecter vLLM à des applications existantes

- Création d'une API façade pour la consommation interne ou externe
- Développement de clients Python et JavaScript robustes
- Gestion du contexte de requêtes et de la mémoire conversationnelle
- Streaming des réponses pour un affichage progressif des tokens

- Surveillance des appels via le traçage distribué

Automatiser les déploiements avec CI/CD et GitOps

- Conception d’une chaîne CI/CD complète pour vLLM
- Utilisation de Terraform et Helm pour l’infrastructure
- Mise en place d’une approche GitOps avec ArgoCD ou Flux
- Intégration des audits de sécurité et vérification automatique des dépendances
- Atelier pratique : Créer un pipeline de déploiement automatisé et valider un déploiement en production simulée.

[Jour 3 - Après-midi]

Assurer la sécurité et la conformité

- Gestion des secrets et des droits d’accès par RBAC
- Application du principe du moindre privilège et séparation des environnements
- Traitement et protection des données sensibles (PII)
- Contrôle des contenus générés et politiques de filtrage
- Validation de la checklist de mise en production sécurisée

Exploiter et améliorer en continu

- Suivi des performances, tendances d’usage et consommation des ressources
- Maintenance préventive et mise à jour des modèles et dépendances
- Gestion des incidents et mise en œuvre des actions correctives
- Communication des résultats et indicateurs de disponibilité
- Atelier pratique : Réviser un runbook opérationnel et ajuster les seuils de 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.