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

Mis à jour le 08/06/2026

S'inscrire

Formation NVIDIA AI Entreprise : déployer et industrialiser des LLM

3 jours (21 heures)

Présentation

NVIDIA AI Enterprise est une plateforme logicielle d'entreprise conçue pour développer, déployer et exploiter des applications d'intelligence artificielle en production.

Elle réunit des frameworks, microservices, SDKs, opérateurs Kubernetes et composants d'infrastructure pour accélérer la mise en œuvre de services IA fiables, sécurisés et scalables.

Notre formation NVIDIA AI Enterprise vous permettra de maîtriser le déploiement et l'industrialisation de LLM en entreprise en vous appuyant sur les briques clés de l'écosystème NVIDIA : NIM, NeMo, TensorRT-LLM, Kubernetes, RAG et observabilité.

Vous apprendrez à déployer des microservices d'inférence, optimiser les performances GPU, concevoir une architecture RAG, sécuriser les usages des modèles et superviser vos endpoints LLM.

À l'issue de la formation, vous serez en mesure de concevoir, déployer et superviser une architecture LLM d'entreprise avec NVIDIA AI Enterprise, tout en maîtrisant les enjeux de sécurité, de coûts, de performance et de gouvernance.

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

Objectifs

- Comprendre les fondamentaux de NVIDIA AI Enterprise
- Déployer des modèles LLM avec NVIDIA NIM

- Optimiser l’inférence avec TensorRT-LLM
- Structurer des workflows d’adaptation et d’évaluation avec NeMo
- Concevoir une architecture RAG d’entreprise
- Industrialiser, sécuriser et superviser des services LLM en production

Public visé

- Ingénieurs IA et ML engineers
- Data scientists souhaitant industrialiser des modèles génératifs
- Architectes cloud et architectes IA
- Ingénieurs DevOps, MLOps et plateformes
- Responsables techniques IA et équipes innovation

Pré-requis

- Connaissances générales en IA générative et modèles LLM
- Notions de Python, API REST et conteneurs
- Compréhension générale de Kubernetes ou des architectures cloud
- Notions de déploiement applicatif et d’observabilité

Pré-requis techniques

- Disposer d’un poste avec Linux, macOS ou Windows avec WSL2
- Avoir accès à un environnement GPU compatible ou à un environnement cloud GPU
- Installer Docker, kubectl, Helm et un éditeur de code
- Disposer d’un compte NVIDIA Developer ou NVIDIA NGC selon les ateliers retenus
- Prévoir une connexion Internet stable

Programme de notre formation NVIDIA AI Enterprise

[Jour 1 - Matin]

Fondamentaux NVIDIA AI Enterprise et IA générative en production

- Comprendre le positionnement de NVIDIA AI Enterprise dans une stratégie IA d’entreprise
- Identifier les composants clés : NIM, NeMo, TensorRT-LLM, drivers, GPU Operator et registry NGC
- Comprendre les enjeux d’industrialisation des LLM : coût, latence, sécurité, scalabilité et gouvernance
- Différencier expérimentation IA, prototype RAG et service LLM prêt pour la production
- Identifier les architectures de déploiement : bare metal, cloud, Kubernetes, OpenShift et environnements hybrides
- Atelier pratique : explorer l’écosystème NVIDIA AI Enterprise, NGC Catalog et les composants utilisés pendant la formation

[Jour 1 - Après-midi]

NVIDIA NIM pour servir des modèles LLM

- Comprendre le rôle de NVIDIA NIM dans le déploiement accéléré de modèles fondation
- Déployer un microservice NIM pour exposer une API d'inférence compatible avec les usages LLM
- Configurer les variables d'environnement, clés d'accès, images conteneurs et paramètres GPU
- Tester l'inférence avec des prompts simples, des requêtes batch et des scénarios conversationnels
- Analyser les notions de débit, temps de réponse, concurrence et consommation GPU
- Atelier pratique : déployer un premier NIM LLM et interroger son endpoint d'inférence

Optimisation d'inférence avec TensorRT-LLM

- Comprendre le rôle de TensorRT-LLM dans l'optimisation de l'inférence LLM
- Identifier les leviers de performance : batching, quantization, KV cache, tensor parallelism et paged attention
- Comparer les contraintes de modèles selon taille, contexte, mémoire GPU et latence attendue
- Comprendre les apports de FP8, INT8 et autres stratégies d'optimisation
- Évaluer les compromis entre qualité, coût, débit et temps de réponse

[Jour 2 - Matin]

NeMo pour adapter, évaluer et gouverner les LLM

- Comprendre le rôle de NVIDIA NeMo dans le cycle de vie des modèles génératifs
- Identifier les usages : fine-tuning, adaptation, évaluation, guardrails et préparation de datasets
- Comprendre les concepts de LoRA, PEFT, alignment et personnalisation contrôlée
- Préparer un workflow d'adaptation de modèle selon un besoin métier
- Mettre en place des critères d'évaluation : qualité des réponses, hallucinations, sécurité et pertinence
- Atelier pratique : structurer un jeu de données et définir une stratégie d'adaptation d'un LLM

[Jour 2 - Après-midi]

Construire une architecture RAG d'entreprise

- Comprendre le rôle du RAG pour connecter les LLM aux données internes
- Concevoir une architecture avec ingestion documentaire, embeddings, base vectorielle et moteur de génération
- Définir les stratégies de chunking, métadonnées, filtrage et contrôle d'accès aux sources
- Connecter un endpoint NIM à une chaîne RAG applicative
- Évaluer la pertinence des réponses, la traçabilité des sources et les risques d'hallucination

Sécurité, conformité et gouvernance des LLM

- Identifier les risques liés aux LLM : fuite de données, prompt injection, hallucination et accès non autorisé
- Mettre en place des contrôles de sécurité autour des prompts, réponses et documents internes
- Définir une stratégie de guardrails, filtrage, audit et validation des sorties
- Comprendre les exigences de conformité : confidentialité, logs, traçabilité et conservation des données
- Appliquer les bonnes pratiques de gouvernance pour les modèles, datasets et endpoints
- Atelier pratique : auditer un cas RAG et identifier les points de contrôle sécurité à ajouter

[Jour 3 - Matin]

Déploiement Kubernetes et industrialisation MLOps

- Déployer des workloads IA avec Kubernetes, GPU Operator et NIM Operator
- Comprendre les contraintes GPU : scheduling, ressources, quotas, isolation et multi-tenancy
- Mettre en place un workflow de déploiement versionné pour endpoints LLM
- Automatiser la promotion entre développement, staging et production
- Intégrer NVIDIA AI Enterprise dans une démarche MLOps ou LLMOps

[Jour 3 - Après-midi]

Observabilité, performance et exploitation

- Surveiller les métriques clés : latence, throughput, saturation GPU, erreurs et files d’attente
- Mettre en place des dashboards pour suivre les usages, coûts et performances des endpoints LLM
- Analyser les logs applicatifs, traces, prompts, réponses et événements d’inférence
- Définir des seuils d’alerte pour dégradation de service ou dérive des usages
- Optimiser les ressources selon le modèle, la charge, le contexte et la criticité métier

Cas final : industrialiser un service LLM d’entreprise

- Assembler les composants : NIM, TensorRT-LLM, RAG, sécurité et observabilité
- Définir une architecture cible adaptée à un cas métier d’entreprise
- Mettre en place les bonnes pratiques de déploiement, monitoring, gouvernance et exploitation
- Préparer une checklist de mise en production pour un service LLM
- Identifier les prochaines étapes : montée en charge, adaptation de modèle, coûts et gouvernance multi-projets
- Atelier pratique : concevoir et présenter une architecture LLM industrialisée.

Pour aller plus loin

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.