

Déployer des LLM en Entreprise — RAG, Fine-tuning et Infrastructure On-Premise

Lien : https://maroc-performance.com/formation/deployer-des-llm-en-entreprise-rag-fine-tuning-et-infrastructure-on-premise

DURÉE
8 jours (56h)

RÉFÉRENCE
IEEI47

CATÉGORIE
IA pour Déploiement LLM, Architecture IA et MLOps

Maroc Performance

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Maîtriser les fondamentaux techniques et architecturaux des Large Language Models en environnement d’entreprise.
- Savoir sélectionner et benchmarker les modèles LLM en fonction des besoins métiers et contraintes techniques.
- Concevoir et dimensionner une infrastructure On-Premise adaptée au déploiement de modèles d’IA générative.
- Déployer des LLM et mettre en place des architectures d’inférence performantes et sécurisées.
- Maîtriser les principales techniques d’optimisation des modèles, notamment quantification, distillation et compression.
- Concevoir des architectures RAG complètes exploitant les connaissances et documents internes de l’entreprise.
- Utiliser les bases vectorielles, embeddings, recherche sémantique et reranking pour améliorer la qualité des réponses.
- Comprendre et mettre en œuvre le fine-tuning et les techniques PEFT, LoRA et QLoRA pour adapter les modèles aux besoins métiers.
- Maîtriser le Prompt Engineering et les mécanismes avancés d’interaction avec les LLM.
- Mettre en place une démarche LLMOps permettant d’industrialiser le cycle de vie des modèles et applications.
- Évaluer, monitorer et optimiser la qualité, la sécurité, la performance et les coûts des solutions LLM.
- Sécuriser les données et les infrastructures LLM tout en garantissant confidentialité, souveraineté et contrôle des accès.
- Intégrer les LLM aux systèmes d’information et applications métiers grâce aux API, outils et agents IA.
- Construire une stratégie de gouvernance et une feuille de route permettant d’industrialiser les LLM à l’échelle de l’entreprise.

Maroc Performance

POUR QUI ?

- Directeurs des Systèmes d’Information (DSI), CIO et responsables de la transformation numérique.
- IT Architects, Enterprise Architects et Cloud/Infrastructure Architects.
- AI Engineers, Machine Learning Engineers et ingénieurs spécialisés en IA générative.
- Data Scientists, Data Engineers et spécialistes des architectures Data & AI.
- DevOps Engineers, MLOps Engineers et LLMOps Engineers.
- Infrastructure Engineers, System Engineers et responsables des infrastructures GPU.
- Responsables Cloud, On-Premise, Edge Computing et architectures hybrides.
- Responsables cybersécurité, sécurité des données et gouvernance des systèmes d’IA.
- Responsables Data, Innovation, R&D et transformation par l’intelligence artificielle.
- Développeurs logiciels et architectes souhaitant intégrer des LLM dans les applications d’entreprise.
- Consultants et experts en IA, Data, Cloud, infrastructure et transformation digitale.
- Responsables métiers souhaitant comprendre les possibilités et contraintes du déploiement de LLM en entreprise.
- Dirigeants et managers souhaitant développer une stratégie d’IA générative souveraine et maîtrisée.
- Professionnels IT souhaitant acquérir une expertise complète en RAG, fine-tuning, LLMOps et infrastructure LLM On-Premise.

Maroc Performance



Programme détaillé

1 / ARCHITECTURE ET FONDAMENTAUX DES LLM EN ENTREPRISE

- Comprendre l’architecture des Large Language Models (LLM), leurs capacités, leurs limites et leurs principaux cas d’usage professionnels.
- Identifier les différences entre modèles propriétaires, open source, Cloud, hybrides et déployés localement.
- Définir une architecture LLM d’entreprise adaptée aux exigences de performance, sécurité, confidentialité et souveraineté des données.

2 / SÉLECTION ET ÉVALUATION DES MODÈLES LLM

- Comparer les modèles selon leurs performances, capacités de raisonnement, contexte, coût, latence et exigences matérielles.
- Évaluer les modèles open source et propriétaires selon les besoins métiers et contraintes techniques de l’entreprise.
- Mettre en place une méthodologie de benchmark et de sélection adaptée aux cas d’usage ciblés.

3 / INFRASTRUCTURE ON-PREMISE POUR L’IA GÉNÉRATIVE

- Concevoir une infrastructure locale adaptée à l’exécution de modèles LLM en environnement d’entreprise.
- Dimensionner les ressources GPU, CPU, mémoire, stockage et réseau selon les modèles et volumes d’inférence.
- Comprendre les architectures de serveurs GPU, clusters IA et infrastructures hybrides Edge-Cloud-On-Premise.

4 / DÉPLOIEMENT ET SERVING DES LLM

- Déployer des modèles LLM dans une infrastructure sécurisée et contrôlée par l’organisation.
- Mettre en place des mécanismes d’inférence, d’API serving, de gestion des requêtes et de montée en charge.

Maroc Performance

- Optimiser la latence, le throughput, la disponibilité et l’utilisation des ressources de calcul.

5 / OPTIMISATION DES MODÈLES : QUANTIZATION, DISTILLATION ET COMPRESSION

- Comprendre les techniques de quantification permettant de réduire la consommation mémoire et les besoins en calcul.
- Exploiter la distillation, pruning et autres méthodes de compression pour adapter les modèles aux infrastructures disponibles.
- Évaluer les compromis entre taille du modèle, performance, précision, coût et latence.

6 / RAG — RETRIEVAL-AUGMENTED GENERATION

- Comprendre l’architecture RAG et son intérêt pour exploiter les connaissances internes de l’entreprise.
- Construire une chaîne complète intégrant ingestion documentaire, chunking, embeddings, recherche et génération augmentée.
- Réduire les hallucinations et améliorer la pertinence des réponses grâce à des sources documentaires contextualisées.

7 / VECTOR DATABASES ET SYSTÈMES DE RECHERCHE SÉMANTIQUE

- Comprendre le rôle des embeddings et des bases vectorielles dans les architectures RAG.
- Concevoir des mécanismes de recherche sémantique, hybride et de reranking adaptés aux données d’entreprise.
- Optimiser la pertinence des documents récupérés et la qualité du contexte transmis au LLM.

8 / FINE-TUNING ET ADAPTATION DES LLM AUX BESOINS MÉTIERS

- Comprendre les différences entre prompting, RAG, fine-tuning et entraînement spécialisé.
- Mettre en œuvre des techniques d’adaptation telles que PEFT, LoRA et QLoRA selon les contraintes du projet.
- Constituer, nettoyer et structurer des jeux de données permettant d’adapter un modèle à un domaine métier spécifique.

9 / PROMPT ENGINEERING ET STRUCTURATION DES INTERACTIONS LLM

- Concevoir des prompts robustes pour améliorer la précision, la cohérence et la contrôlabilité des réponses.

Maroc Performance

- Exploiter les techniques de few-shot prompting, structured output, chain-of-thought alternatives et tool calling.
- Construire des templates de prompts réutilisables adaptés aux différents processus métiers.

10 / LLMOPS ET INDUSTRIALISATION DES APPLICATIONS IA

- Mettre en place des pratiques LLMOps pour gérer les modèles, prompts, datasets, versions et environnements.
- Automatiser les processus de déploiement, tests, monitoring, évaluation et mise à jour des applications LLM.
- Assurer la reproductibilité, la traçabilité et la gouvernance du cycle de vie des solutions d'IA générative.

11 / ÉVALUATION, OBSERVABILITÉ ET PERFORMANCE DES LLM

- Définir des métriques permettant d'évaluer la pertinence, factualité, robustesse, sécurité et qualité des réponses.
- Mettre en place des mécanismes de monitoring des modèles, prompts, coûts, latence et consommation des ressources.
- Construire des jeux de tests et benchmarks permettant de détecter les régressions et dérives des performances.

12 / SÉCURITÉ, CONFIDENTIALITÉ ET SOUVERAINETÉ DES DONNÉES

- Sécuriser les données sensibles utilisées dans les systèmes RAG, fine-tuning et applications LLM.
- Mettre en place des mécanismes d'authentification, contrôle d'accès, chiffrement, isolation et traçabilité.
- Maîtriser les risques liés aux fuites de données, prompt injection, data poisoning, exfiltration et accès non autorisés.

13 / INTÉGRATION DES LLM AUX SYSTÈMES D'ENTREPRISE

- Connecter les LLM aux ERP, CRM, bases de données, GED, API, outils métiers et systèmes d'information existants.
- Mettre en œuvre le tool calling et les agents IA capables d'interagir avec les applications de l'entreprise.
- Concevoir des architectures permettant aux LLM d'exécuter des tâches tout en conservant des mécanismes de contrôle et de validation.

Maroc Performance

14 / GOUVERNANCE, COÛTS ET STRATÉGIE DE DÉPLOIEMENT À GRANDE ÉCHELLE

- Définir une gouvernance des LLM couvrant sécurité, conformité, responsabilité, modèles, données et usages.
- Piloter les coûts d'infrastructure, GPU, stockage, énergie, inférence et maintenance des modèles.
- Construire une feuille de route permettant de passer du Proof of Concept à une plateforme LLM d'entreprise industrialisée.

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Mises en Situation pour faciliter l'assimilation
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

04 au 13 Jan. 2027	Présentiel - Casablanca
01 au 10 Mars 2027	Présentiel - Casablanca
26 Avr. au 05 Mai 2027	Présentiel - Casablanca
21 au 30 Juin 2027	Présentiel - Casablanca
16 au 25 Août 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Maroc Performance

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@maroc-performance.com
Web : https//www.maroc-performance.com

Maroc Performance — Tous droits réservés