

Déployer une IA open source en local

Code : EAI-DEPLOY-100

Durée : 1 jour (7 heures)

Public visé

Data Scientists, Data Analysts, ingénieurs Data, MLOps, développeurs, architectes Data, cadres dirigeants, managers, chefs de projets, Product Managers IA.

Prérequis

- Formation IA pour les spécialistes de l'IT ou compétences équivalentes
- Maîtrise du langage Python

Objectifs pédagogiques

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

- Définir les exigences fonctionnelles et tests des modèles
- Comparer les modèles open source selon les métriques principales
- Explorer les alternatives d'entraînement (transfer learning, prompting, LoRA)
- Spécifier les architectures locales pour héberger des modèles
- Déployer et surveiller des modèles en production

Programme

Matin : État de l'art et modèles

Dernières avancées en production IA

- Comparatif closed source vs open source
- Modèles polyvalents : Mistral/Mixtral, Llama, Falcon
- Modèles légers : Phi-3, Gemma, TinyLlama
- Exigences matérielles : mémoire, CPU/GPU/cluster
- Ingénierie des datasets
- Les trois couches du stack IA

Après-midi : Déploiement et optimisation

Modèles fondateurs et alternatives d'entraînement

- From scratch, transfer learning, prompting, distillation

- RLHF, mixture of experts, LoRA, PEFT

Infrastructures de déploiement

- Ollama, LM Studio, llama.cpp, vLLM, Unsloth, Hugging Face

Optimisation

- Quantization, LoRA
- Défis en production pour LLM/LCM

Modalités d'évaluation des acquis

- En cours de formation, par des travaux pratiques
- Et, en fin de formation, par un questionnaire d'auto-évaluation