https://latribunelibre.com/emploi/llmops-f-h-3

LLMOps F/H

Description

Dans le cadre de notre stratégie de recrutement et de croissance, nous recherchons des **profils spécialisés en LLMOps** afin d'accompagner nos clients dans leurs projets autour des grands modèles de langage (LLM).

Missions principales

1. Mise en production et exploitation

- Déployer, superviser et maintenir des modèles LLM (open source ou propriétaires) dans des environnements cloud, on-premise ou hybrides.
- Gérer l'ensemble du cycle de vie des modèles : entraînement, fine-tuning, déploiement, mises à jour.
- Garantir la haute disponibilité, la performance et la résilience des applications LLM.
- Assurer le déploiement de modèles à grande échelle et la gestion du versioning (avec rollback si nécessaire).
- Mettre en œuvre des pratiques avancées : A/B testing, shadow deployment, suivi de la dérive des données et des performances.
- Développer des mécanismes de feedback loop et de supervision humaine (RLHF / RLAIF).
- Optimiser le temps de réponse et contrôler les coûts liés à l'inférence.
- Participer à la conception de produits "LLM-friendly" (chatbots, copilotes, assistants intelligents, etc.).

2. Optimisation et intégration

- Concevoir et déployer des pipelines CI/CD adaptés aux modèles d'IA générative.
- Optimiser l'usage des ressources GPU/CPU afin de maîtriser les coûts et améliorer l'efficacité.
- Intégrer les modèles dans les produits et services métiers via des API, microservices ou agents intelligents

Organisme employeur
ETIX WAY

Type de poste
Temps plein

Secteur
CONSEIL EN SYSTÈMES ET LOGICIELS INFORMATIQUES

Lieu du poste
93051, NOISY LE GRAND, NOISY LE GRAND, France

Salaire de base
40000 € - **Salaire de base**
45000 €

Date de publication
28 septembre 2025 à 15:06

Valide jusqu'au
28.10.2025

Qualifications

- **Formation** : Bac+5 en informatique, data science, intelligence artificielle ou équivalent (école d'ingénieurs, université).
- **Expérience** : 2 ans minimum dans des fonctions liées au **MLOps, DevOps, Data Engineering ou AI Engineering**, avec une expertise récente ou un fort intérêt pour les **LLM et l'IA générative**.

Compétences techniques

- Maîtrise des concepts et outils de **MLOps**/**DevOps** (Kubernetes, Docker, CI/CD, GitOps).
- Bonne connaissance des frameworks IA et NLP : **HuggingFace, PyTorch, TensorFlow, LangChain**.
- Expérience avec les **bases vectorielles** (FAISS, Weaviate, Pinecone, Milvus).
- Compétences en **observabilité et monitoring** (Prometheus, Grafana, ELK, MLflow).
- Connaissances en **optimisation d'inférence** : quantization, distillation, optimisation GPU/TPU.
- Maîtrise de l'intégration via API, microservices et architectures cloud.
- MLOps avancé
- NLP et LLM
- Infrastructure et DevOps
- Optimisation des modèles
- RAG (Retrieval-Augmented Generation

Soft skills

- Forte appétence pour l'**innovation et l'IA générative**.
- Rigueur, autonomie et sens de l'organisation.
- Capacité à travailler dans un environnement agile, collaboratif et en forte évolution.
- Bonnes compétences en communication, pour interagir avec les équipes techniques comme avec les métiers.
- Esprit analytique, sens du service et orientation résultats.