LiteLLM, PortKey et proxy LLM — Gérer plusieurs modèles en production

Lien : https://innov-systems.com/formation/litellm-portkey-et-proxy-llm-gerer-plusieurs-modeles-en-production

DURÉE
4 jours (28h)

RÉFÉRENCE
BDI115

CATÉGORIE
Intelligence Artificielle Générative (IAG)

Innov Systems

OBJECTIFS DE LA FORMATION

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

- Comprendre les enjeux des architectures multi-LLM
- Maîtriser LiteLLM et PortKey pour la gestion centralisée des modèles
- Concevoir une couche d’abstraction indépendante des fournisseurs
- Réduire les risques de vendor lock-in
- Mettre en place un routage intelligent des requêtes
- Assurer la résilience et la continuité des services IA
- Optimiser les coûts d’exploitation des modèles génératifs
- Superviser les performances et la qualité de service
- Sécuriser les accès et les échanges avec les LLM
- Développer une gouvernance robuste des usages IA
- Déployer des solutions multi-modèles à l’échelle entreprise
- Étudier des cas concrets de production multi-LLM

POUR QUI ?

- Architectes logiciels
- Architectes cloud
- Ingénieurs IA
- Machine Learning Engineers
- MLOps Engineers
- DevOps Engineers
- Data Engineers
- Développeurs backend
- CTO et responsables techniques
- Consultants en transformation digitale
- Product Managers techniques
- Toute personne impliquée dans le déploiement et l’exploitation de LLM en production

Innov Systems

Programme détaillé

1 / Introduction aux architectures multi-LLM

- Évolution des écosystèmes IA générative
- Enjeux du multi-fournisseur et du multi-modèle
- Rôle des proxies LLM dans les architectures modernes

2 / Présentation de LiteLLM et PortKey

- Fonctionnalités clés de LiteLLM
- Fonctionnalités clés de PortKey
- Comparaison des approches et cas d’usage

3 / Concevoir une couche d’abstraction LLM

- Standardisation des appels aux modèles
- Architecture découplée des fournisseurs
- Réduction du vendor lock-in

4 / Routage intelligent des requêtes

- Sélection dynamique du modèle adapté
- Répartition de charge entre fournisseurs
- Gestion des priorités métier

5 / Résilience et continuité de service

- Mécanismes de fallback entre modèles
- Haute disponibilité des services IA
- Gestion des pannes et incidents fournisseurs

Innov Systems

6 / Monitoring, observabilité et gouvernance

- Suivi des performances et de la latence
- Journalisation des interactions
- Tableaux de bord et métriques clés

7 / Optimisation des coûts d’inférence

- Contrôle de la consommation de tokens
- Choix du modèle selon les besoins métier
- Mise en cache et optimisation des appels

8 / Sécurité et conformité des flux LLM

- Gestion sécurisée des clés API
- Protection des données sensibles
- Contrôle d’accès et audit des usages

9 / Mise en œuvre pratique et cas d’usage

- Déploiement d’un proxy LLM en production
- Intégration avec des applications métier
- Bonnes pratiques d’exploitation à grande échelle

Approche pédagogique

- ✔ Support Ecrit et Projection
- ✔ Exposés Interactifs, Podcasts et Vidéos
- ✔ Brainstorming et Jeux de Rôle
- ✔ Cas Pratiques et Labs inclus pour leur impact opérationnel
- ✔ Test de Validation des Acquis des Connaissances

Innov Systems

Prochaines dates programmées

26 au 29 Jan. 2027	Présentiel - Casablanca
23 au 26 Mars 2027	Présentiel - Casablanca
18 au 21 Mai 2027	Présentiel - Casablanca
13 au 16 Juil. 2027	Présentiel - Casablanca
14 au 17 Sep. 2027	Présentiel - Casablanca

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

Réservation & Renseignements

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

Innov Systems — Tous droits réservés

Innov Systems