[sfeir] Institute

Google Cloud | GCP300GENAIPROD

Generative AI in Production

Déployez vos applications IA générative sur Google Cloud en toute confiance

1 jour / 7h

Présentation du cours

Dans ce cours d'une journée, vous découvrirez les différents défis qui se posent lors de la mise en production d'applications alimentées par l'IA générative par rapport au ML traditionnel.

Vous apprendrez à gérer l'expérimentation et l'ajustement de vos LLM, puis vous discuterez de la manière de déployer, tester et maintenir vos applications alimentées par LLM.

Enfin, vous aborderez les meilleures pratiques pour la journalisation et la surveillance de vos applications alimentées par LLM en production.

Ce cours est de niveau avancé et est dirigé par un instructeur.

Méthodes mobilisées : Ce cours alterne parties théoriques sous forme de lectures (slides), démos et parties pratiques sous forme de labs dirigés

Objectifs pédagogiques

- Décrire les défis de la mise en production d'applications utilisant l'IA générative.
- Gérer l'expérimentation et l'évaluation pour les applications alimentées par LLM.
- Mettre en production les applications alimentées par LLM.

- Implémenter la journalisation et la surveillance pour les applications alimentées par LLM.

Modalités d'évaluation : Les objectifs pédagogiques sont évalués à travers la réalisation des parties pratiques (labs dirigés) sous la supervision du formateur délivrant la session de formation.

Public cible

- Développeurs
- Ingénieurs en Machine Learning qui souhaitent opérationnaliser les applications basées sur Gen AI.

Prérequis

Avoir suivi le cours « Application Development with LLMs on Google Cloud » ou avoir des connaissances équivalentes.

Programme

Module 01 : Introduction à l'IA Générative en Production

Sujets

- Démonstration du système d'IA : Coffee on Wheels
- MLOps traditionnel vs. GenAIOps
- Opérations d'IA Générative
- Composants d'un système LLM

Objectifs

- Comprendre les opérations d'IA générative
- Comparer les MLOps traditionnels et les GenAIOps
- Analyser les composants d'un système LLM

Module 02 : Gestion de l'expérimentation

Sujets

- Ensembles de données et Prompt Engineering
- Architecture RAG et ReACT
- Évaluation du modèle LLM (métriques et cadre)
- Suivi des expériences

Objectifs

- Expérimenter avec les jeux de données et l'ingénierie de prompts
- Utiliser l'architecture RAG et ReACT
- Évaluer les modèles LLM
- Suivre les expérimentations

Activités

- Lab : Tests unitaires d'applications d'IA générative
- Lab optionnel : IA générative avec Vertex AI : Conception d'invites (Prompt Design)

Module 03 : Mise en production de l'IA Générative

Sujets

- Déploiement, packaging et gestion de versions (GenAIOps)
- Test des systèmes LLM (unité et intégration)
- Maintenance et mises à jour (opérations)
- Sécurité et migration des invites

Objectifs

- Déployer, packager et versionner les modèles
- Tester les systèmes LLM
- Maintenir et mettre à jour les modèles LLM
- Gérer la sécurité des prompts et la migration

Activités

- Lab : Vertex AI Pipelines : Démarrage rapide (Qwik Start)
- Lab : Sécurisation avec Vertex AI Gemini API

Module 04 : Journalisation et surveillance pour les systèmes LLM en production

Sujets

- Cloud Logging
- Gestion de versions, évaluation et généralisation des invites

- Surveillance du décalage évaluation-service
- Validation continue

Objectifs

- Utiliser Cloud Logging
- Versionner, évaluer et généraliser les prompts
- Surveiller les écarts entre évaluation et service
- Utiliser la validation continue

Activités

- Lab : Vertex AI : Playbook d’évaluations Gemini
- Lab optionnel : Fine Tuning supervisé avec Gemini pour questions et réponses