RAG (Retrieval-Augmented Generation) — Architecture et implémentation

Lien : https://innov-systems.com/formation/rag-retrieval-augmented-generation-architecture-et-implementation

DURÉE
5 jours (35h)

RÉFÉRENCE
BDI187

CATÉGORIE
Intelligence Artificielle Générative (IAG)

Innov Systems

OBJECTIFS DE LA FORMATION

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

- Comprendre l’architecture RAG et ses composants
- Concevoir des systèmes de génération augmentée par récupération
- Maîtriser les embeddings et bases vectorielles
- Implémenter des pipelines RAG complets
- Intégrer des LLM dans des systèmes d’information
- Améliorer la pertinence des réponses IA
- Optimiser les performances des systèmes RAG
- Utiliser des frameworks comme LangChain et LlamaIndex
- Gérer la sécurité et la gouvernance des données
- Déployer des solutions RAG en production
- Réduire les hallucinations des modèles génératifs
- Étudier des cas concrets d’applications IA avancées

POUR QUI ?

- Data scientists
- Machine learning engineers
- AI engineers
- Data engineers
- Backend developers
- Software engineers
- Architectes IA
- DevOps / MLOps engineers
- Consultants IA
- Product engineers
- Toute personne impliquée dans les systèmes IA et LLM

Innov Systems

Programme détaillé

1 / Introduction au RAG

- Définition du Retrieval-Augmented Generation
- Limites des LLM sans accès à des données externes
- Cas d’usage en entreprise (chatbots, recherche, support)

2 / Architecture globale d’un système RAG

- Pipeline retrieval + generation
- Composants clés : LLM, retriever, vector store
- Flux de données et orchestration

3 / Préparation et ingestion des données

- Collecte et nettoyage des documents
- Chunking et segmentation du contenu
- Embeddings et vectorisation des données

4 / Bases vectorielles et indexation

- Concepts de vector databases
- Similarité sémantique (cosine, dot product)
- Outils : FAISS, Pinecone, ChromaDB

5 / Moteurs de recherche sémantique

- Recherche hybride (keyword + vectorielle)
- Ranking et re-ranking des résultats
- Optimisation de la pertinence

Innov Systems

6 / Intégration des LLM dans le pipeline RAG

- Prompt engineering pour RAG
- Injection de contexte dans les modèles
- Gestion de la fenêtre de contexte

7 / Frameworks et outils RAG

- LangChain, LlamaIndex
- Orchestration de pipelines IA
- Connecteurs vers sources de données

8 / Optimisation des performances

- Réduction de latence
- Cache des embeddings
- Amélioration de la qualité des réponses

9 / Sécurité et gouvernance des systèmes RAG

- Protection des données sensibles
- Filtrage et contrôle des sources
- Prévention des hallucinations

10 / Cas pratiques et déploiement

- Construction d’un chatbot RAG
- Système de recherche documentaire intelligent
- Déploiement en production cloud

Innov Systems

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

Prochaines dates programmées

23 au 27 Nov. 2026	Distanciel
25 au 29 Jan. 2027	Présentiel - Casablanca
22 au 26 Mars 2027	Présentiel - Casablanca
17 au 21 Mai 2027	Présentiel - Casablanca
12 au 16 Juil. 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