

CAHIER DES CHARGES

Projet

Conception et mise en œuvre d'un système de Retrieval-Augmented Generation (RAG) pour l'exploitation intelligente des données internes de l'entreprise

1. Contexte

Les entreprises produisent quotidiennement une quantité importante de données sous différentes formes :

- Documents PDF
- Procédures internes
- Notes de service
- Manuels techniques
- Contrats
- Rapports
- Comptes rendus de réunions
- Courriers
- Bases documentaires
- Fichiers Word, Excel et PowerPoint

Ces informations sont généralement dispersées dans plusieurs systèmes (serveurs de fichiers, SharePoint, NAS, GED, Cloud, etc.), ce qui rend leur consultation difficile.

Les collaborateurs perdent un temps considérable à rechercher une information pertinente.

L'émergence des modèles de langage (LLM) offre aujourd'hui la possibilité d'interroger ces données en langage naturel.

Cependant, les LLM ne connaissent pas les données internes d'une entreprise.

Le Retrieval-Augmented Generation (RAG) permet de résoudre cette problématique en combinant :

- la recherche sémantique,
- les bases vectorielles,
- les modèles d'embeddings,
- et un modèle de langage génératif.

Le présent projet consiste à concevoir puis développer un assistant intelligent capable de répondre aux questions des collaborateurs à partir des documents internes de l'entreprise.

2. Problématique

Comment permettre aux collaborateurs d'accéder rapidement aux connaissances internes de l'entreprise à travers une interface conversationnelle alimentée par un système RAG garantissant la pertinence, la traçabilité et la sécurité des réponses ?

3. Objectif général

Concevoir et implémenter une plateforme RAG permettant l'exploitation intelligente des connaissances de l'entreprise.

4. Objectifs spécifiques

Le système devra permettre de :

- centraliser les documents internes ;
- indexer automatiquement les documents ;
- générer les embeddings ;
- stocker les vecteurs dans une base vectorielle ;
- rechercher les passages les plus pertinents ;
- générer des réponses contextualisées ;
- citer les documents sources ;
- gérer les droits d'accès aux informations ;
- administrer la base documentaire.

5. Périmètre du projet

Inclus

- Import de documents PDF
- Import de documents Word
- Import de fichiers texte
- Import de documents Markdown
- Import de fichiers Excel
- Découpage automatique (Chunking)
- Nettoyage des données
- Génération des embeddings
- Base vectorielle

- Recherche sémantique
- Chat IA
- Historique des conversations
- Gestion des utilisateurs
- Gestion des rôles
- Interface d'administration
- Réindexation des documents
- Journalisation des traitements

Hors périmètre

- OCR de documents papier
- Traitement vidéo
- Traitement audio
- Traduction automatique
- Fine-tuning d'un LLM

6. Fonctionnalités attendues

Module Administration

- Authentification
- Gestion des utilisateurs
- Gestion des rôles
- Tableau de bord
- Gestion des documents
- Paramétrage des modèles IA
- Gestion des bases vectorielles
- Statistiques

Module Gestion documentaire

- Dépôt de documents
- Suppression
- Versionning
- Catégorisation

- Tags
- Métadonnées

Pipeline RAG

Le pipeline devra suivre les étapes suivantes :

1. Import du document
2. Extraction du texte
3. Nettoyage
4. Découpage en chunks
5. Génération des embeddings
6. Indexation
7. Stockage dans la base vectorielle

Moteur de recherche

Le système devra permettre :

- recherche sémantique
- recherche hybride
- filtrage par métadonnées
- filtrage par catégorie
- recherche par utilisateur

Assistant IA

L'utilisateur pourra :

- poser une question en langage naturel ;
- recevoir une réponse argumentée ;
- consulter les sources utilisées ;
- afficher les extraits ayant servi à la réponse ;
- poursuivre une conversation.

7. Exigences fonctionnelles

Le système devra :

- répondre en moins de 5 secondes pour une question classique ;
- prendre en charge plusieurs milliers de documents ;

- permettre la mise à jour des index ;
- assurer la confidentialité des données ;
- empêcher l'accès aux documents non autorisés ;
- enregistrer les conversations.

8. Exigences non fonctionnelles

Sécurité

- Authentification
- RBAC
- Chiffrement
- Journalisation
- HTTPS

Disponibilité

- Sauvegarde
- Haute disponibilité

Évolutivité

Le système devra permettre l'ajout futur :

- d'autres LLM ;
- d'autres bases vectorielles ;
- d'autres formats de documents.

9. Résultats attendus

À l'issue du projet, l'entreprise devra disposer d'un assistant conversationnel intelligent capable d'interroger efficacement la base documentaire interne, de fournir des réponses fiables, contextualisées et traçables, tout en garantissant la confidentialité des informations et une expérience utilisateur fluide.