

# IA et référentiels qualité

*Un RAG pour des réponses précises et rapides*

Clément Le Guyadec - mars 2025



*Quelle est la durée minimale de conservation des dossiers de lot selon les BPF ?*






La réponse se trouve dans un PDF de 400 pages  
ou en allant solliciter le pharmacien responsable

# □ Pourquoi un projet RAG en interne ?

- **Contexte** : Les équipes ont besoin d'un accès rapide et précis aux informations relatives aux BPF.
- **Problème** : Trop d'informations, document complexe avec beaucoup de références croisées.
- **Objectif** : Mettre en place un RAG interne pour améliorer la recherche et l'exploitation du référentiel qualité tout en limitant la sollicitation des experts métier.

# Synthèse des URS retenues par la DSI

La DSI a établi un cadre précis pour garantir un déploiement sécurisé, performant et conforme aux exigences réglementaires. Les critères retenus sont :

-  **Contrôle des accès** : Gestion centralisée des utilisateurs et des permissions pour garantir une utilisation maîtrisée.
-  **Flexibilité et compatibilité** : Intégration fluide avec des modèles IA locaux et des API externes, assurant une adaptation aux besoins métier.
-  **Sécurité et confidentialité** : Traitement des données sous contrôle interne, stockage sécurisé pour éviter toute fuite d'informations sensibles.
-  **Évolutivité et modularité** : Capacité d'ajouter des agents spécialisés (RAG, recherche web...) et d'adapter l'outil aux nouveaux cas d'usage.
-  **Conformité RGPD** : Stockage des données en Europe, mise en place d'un Data Processing Agreement (DPA) et respect strict des réglementations.

# Plan projet

<i>Semaine</i>	1	2	3	4	5	6	7	8	9	10
Enquête métier										
Sélection technologique										
Déploiement du MVP										
Formation des utilisateurs										
Évaluation et amélioration										

Critère	URS	OpenWebUI/Mistral
Gestion des utilisateurs	Multi-accès sécurisé	✓ Oui, gestion granularisée
IA locale et API	Compatibilité hybride	✓ Ollama + API externes
Sécurisation des données	Contrôle interne	✓ Données stockées localement
Extensibilité	Plugins et agents IA	✓ Agents intégrés (RAG, web search)
Conformité RGPD	Données en Europe	✓ Oui, stockage et utilisation contrôlés (selon modèle implémenté)

▫ OpenWebUI



Nouvelle conversation

Espace de travail

Recherche

Conversations

Aujourd'hui

BPF Stockage Matières F ...

BPF Personnel Training

BPF Formation Personnel

BPF Personnel Training

Staff Training Requirements i

BPF Training Requirements

Nouvelle conversation

Formation Personnel BPF

Personnel Training Requirem

Formation Personnel Product

BPF Personnel Training

BPF Personnel Training

BPF Personnel Training

BPF Personnel Training

BPF Personnel Training

Le Guyadec

BPF - mai 2019

+

BPF

BPF - mai 2019

Today at 21:27

1

Parties concernées : Partie I des BPF pour les médicaments à usage humain et vétérinaire.

2

Partie, Chapitre et section :

Partie I, Chapitre 7 : Gestion des matières premières et matériaux de conditionnement.

Partie I, Chapitre 10 : Stockage et distribution.

3

Explication synthétique :

Les matières premières doivent être manipulées et stockées de manière à éviter leur dégradation, leur contamination et la contamination croisée. Elles doivent être stockées dans des conditions appropriées, avec des enregistrements des conditions critiques pour leur conservation. Des zones séparées doivent être prévues pour le stockage des matières en quarantaine, refusées, retournées ou rappelées.

4

Synthèse comparative :

Les matières premières doivent être stockées de manière à préserver leurs caractéristiques qualité et à garantir que les résultats des contrôles demeurent applicables à la matière livrée. Des audits et des contrôles doivent être effectués pour s'assurer de la conformité aux BPF et aux spécifications définies dans le dossier d'autorisation de mise sur le marché.

5

Explication détaillée :

Les matières premières doivent être conservées dans des conditions qui garantissent leur qualité et leur stabilité. Cela inclut le stockage réfrigéré si nécessaire, conformément à l'autorisation de mise sur le marché. Les matières premières doivent être conservées dans des conditions qui garantissent leur qualité et leur stabilité. Cela inclut le stockage réfrigéré si nécessaire, conformément à l'autorisation de mise sur le marché. Les matières premières doivent être conservées dans des conditions qui garantissent leur qualité et leur stabilité. Cela inclut le stockage réfrigéré si nécessaire, conformément à l'autorisation de mise sur le marché.

Envoyer un message

+

Recherche Web

Code Interpreter

## Exploitation d'un Référentiel Qualité (RAG)

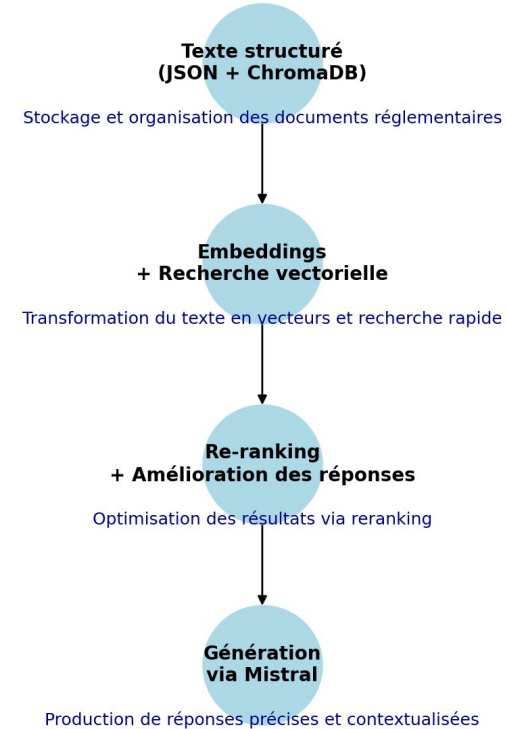
## Technologies choisies

- 🔍 ChromaDB et un model multilangues (intfloat/multilingual-e5-large) pour l'indexation des textes et le stockage local
- 🧠 Modèle Mistral + Re-ranking (bge-reranker-v2-m3) pour améliorer les réponses.
- 📁 Fichiers JSON structurés pour organiser les chapitres et mots-clés et source réglementaire complète au format txt

## Optimisations techniques

- Taille des chunks : 1024 tokens, chevauchement 300.
- Contrôle des prompts pour éviter les réponses hors-sujet.

## Pipeline RAG : De la structuration à la génération



# Architecture et déploiement du RAG



# Défis rencontrés et solutions

⚠️ **Défi 1** : Serveur sans GPU limitant les performances

→ Adoption d'une architecture hybride, combinant des modèles légers en local et l'utilisation d'API externes pour les tâches plus exigeantes.

⚠️ **Défi 2** : Volume et complexité du référentiel

→ Conversion du document en format TXT structuré, segmentation en fichiers JSON par chapitres et mots-clés pour une indexation efficace via ChromaDB.




⚠️ **Défi 3** : Pertinence et clarté des réponses

→ Affinage des prompts pour cadrer la génération de réponses.

→ Ajustement des embeddings et du re-ranking (bge-reranker-v2-m3) pour améliorer la hiérarchisation des résultats

# Validation du modèle

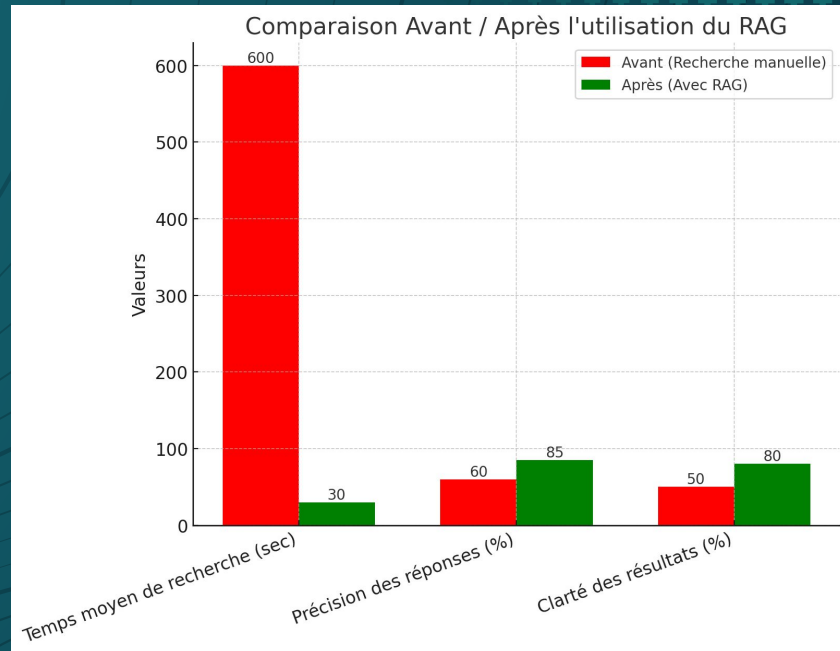
- **Métriques de performance**

-  Exactitude : 50% (Pertinence des réponses).
-  Complétude : 30% (Capacité à récupérer les bonnes informations).
-  Format & clarté : 20% (Facilité de lecture et compréhension).

- **Tests utilisateurs**



- Comparaison avec recherches manuelles
- Réduction des erreurs d'interprétation grâce à l'ajout du re-ranking.

Critère	Recherche manuelle	RAG
Temps moyen de recherche	10 min	30 sec
Précision des réponses	Variable	85% validées
Clarté	Dépend de l'utilisateur	Structurée et synthétique



## Comparaison avant/après

# Conclusion et perspectives

-  Le RAG fonctionne bien et apporte une vraie valeur ajoutée.
-  Améliorations futures :
  - Ajout de nouveaux référentiels,
  - amélioration du modèle,
  - meilleur fine-tuning.
- **Prochaine étape ?**
  - Déploiement à plus grande échelle et retour utilisateur continu.

## ■ Merci pour votre attention

- ➔ <https://github.com/cleguyadec/jedha-Generative-AI>
  - ◆ Prompt système et user prompt pré-configuré dans OpenWebUI
  - ◆ Fichiers alimentant la base de connaissance pour améliorer la qualité des réponses
  - ◆ Grille d'évaluation du modèle et résultat