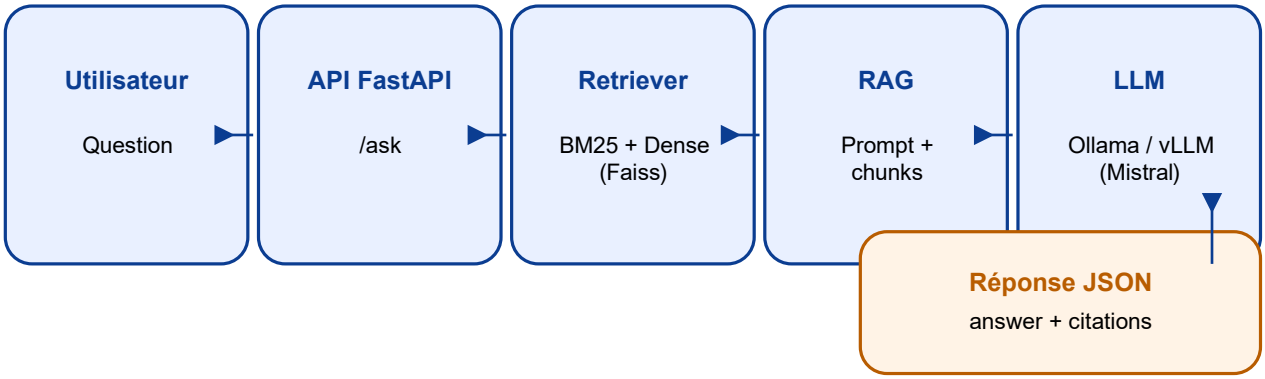
Livrable — Assistant juridique RH (RAG) avec LLM open-source

Objectif : fournir un assistant RH en français capable de répondre à des questions juridiques fréquentes en s'appuyant sur un corpus documentaire fiable et en citant ses sources.

1) Pipeline (vue d'ensemble)

Le système combine un **retriever hybride** (BM25 + embeddings) et un **LLM** pour générer une réponse structurée et sourcée (JSON : *answer* + *citations*).



Ingestion	Docs Markdown (.txt) → nettoyage, découpage (chunks), métadonnées (Doc_ID, pays, dates, tags).
Indexation	Embeddings (E5) → index FAISS + index BM25 (lexical) sur les mêmes chunks.
Recherche	Hybrid scoring (alpha) + top-k final → sélection des passages les plus pertinents.
Génération	Prompt RAG : instructions + passages → réponse concise + citations (doc_id / source / URL).
Évaluation	Script eval_full : taux de réponse, taux de citations, latence, erreurs/timeouts.

2) Outils, modèles et configuration

Back-end & API : Python, **FastAPI** (endpoint /ask), Uvicorn, CORS.
LLM : serveur local (ex. **Ollama**) ou endpoint compatible OpenAI (ex. vLLM) ; modèle Apache-2.0 (ex. Mistral/Mixtral).
Retrieval : index lexical (BM25) + index vectoriel (**FAISS**) ; **sentence-transformers** avec *intfloat/multilingual-e5-small*.
Évaluation : script *eval_full.py* (requests) + export CSV (logs).

3) Stratégie d'adaptation (points clés)

- **Aligner les questions sur le corpus** : reformulation orientée 'mots-clés' (termes juridiques exacts, unités, seuils).
- **Normaliser le corpus**: titres stables, Doc_ID, pays, dates, suppression du bruit, liens sources conservés.
- **Chunking adapté au juridique** : découpe par sections (titres/puces), taille contrôlée pour limiter les citations hors sujet.
- **Forcer la traçabilité** : prompt demandant explicitement une liste *citations* + fallback côté code à partir des chunks retenus.
- **Réglage retrieval** : top_k_dense/BM25/final et alpha calibrés sur un petit jeu d'éval (itérations rapides).
- **Garde-fous** : option *refuse_if_no_context* (refus si pas de passages) + filtrage par métadonnées (country=FR).

4) Risques identifiés et mitigations

Risque	Mitigation
Hallucinations / réponses non sourcées	Imposer citations, refuser si pas de contexte, limiter température, logs d'audit.
Droit qui évolue (obsolescence)	Mettre à jour les docs, ajouter date de version, avertissement 'à vérifier' si contexte ancien.
Citations incorrectes (mauvais passage)	Chunking plus fin, reranking, tests unitaires sur citations, dédoublonnage.
Attaques prompt-injection via documents	Sanitiser les instructions dans le prompt (priorité aux règles système), filtrer contenus suspects.
Confidentialité / RGPD	Minimisation des données, redaction, contrôle d'accès, journalisation, politique de conservation.
Latence / coût	Cache embeddings/index, limiter top-k, modèle plus léger, streaming si besoin, monitoring P95.