

# Mise en place d'un Assistant IA (RAG) pour les Opérateurs de Validation

<b>Mise en place d'un Assistant IA (RAG) pour les Opérateurs de Validation.....</b>	<b>1</b>
1 Contexte et Objectifs.....	1
1-a Problématique initiale.....	1
1-b Solution mise en œuvre.....	2
1-c Objectifs visés.....	2
2 Architecture Technique.....	2
2-a Vue d'ensemble.....	2
2-b Infrastructure de données.....	3
2-c Pipeline d'ingestion documentaire.....	4
2-d Agent conversationnel.....	4
5- Résultats et Performance.....	5
5-a Qualité des réponses.....	5
5-b Impact opérationnel.....	5
6- Aspects Économiques.....	5
Coûts d'exploitation.....	5
7- Limitations Techniques Identifiées.....	6
7-a Synchronisation documentaire.....	6
7-b Extraction de documents complexes.....	6
7-c Autres limitations techniques.....	6
7-d Améliorations techniques possibles.....	7
8- Recommandations.....	7
Court terme.....	7

## 1 Contexte et Objectifs

### 1-a Problématique initiale

Avec l'arrivée de l'été, l'équipe opération accueille un volume important de nouveaux opérateurs nécessitant une montée en compétences rapide.

Idée pour réduire ce 'défi' : Intégrer un outil complémentaire pour faciliter l'apprentissage autonome et réduire les frictions liées à l'acquisition des procédures internes. L'assistant IA

s'inscrit dans cette démarche d'accompagnement, permettant aux nouveaux arrivants d'obtenir des réponses immédiates sans surcharger les tuteurs ou encombrer le canal #opération.

## 1-b Solution mise en œuvre

Développement d'un assistant IA conversationnel intégré directement dans Mattermost, capable de répondre aux questions des opérateurs en s'appuyant sur notre documentation interne. L'assistant, baptisé "Docteur Opérateur", utilise une architecture RAG (Retrieval-Augmented Generation) pour fournir des réponses précises et basées uniquement sur notre documentation.

## 1-c Objectifs visés

- **Autonomisation des opérateurs** : Réduction de la dépendance aux tuteurs pour les questions courantes et une aide disponible 24h/24
- **Centralisation des connaissances** : Point d'accès unique à la documentation interne
- **Traçabilité** : Historique des questions posées et des réponses apportées. Identification des lacunes documentaires et d'éventuelles questions récurrentes
- **Amélioration continue** : Génération de débats constructifs sur les pratiques, détection de la documentation obsolète et uniformisation des méthodes de traitement des dossiers

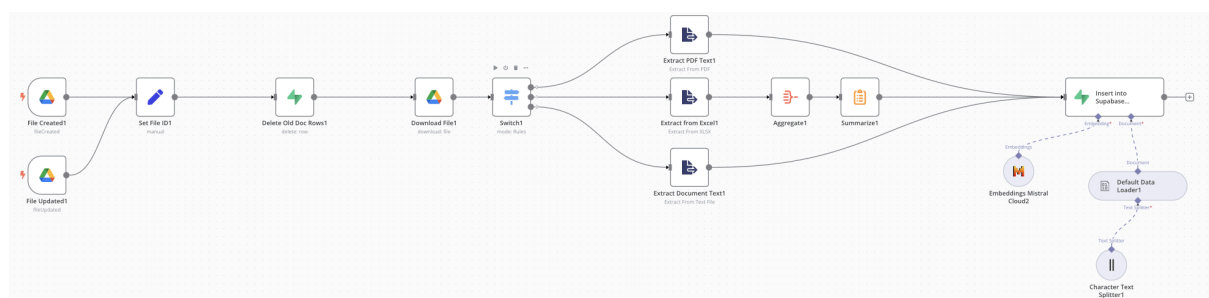
# 2 Architecture Technique

## 2-a Vue d'ensemble

Le système repose sur une architecture n8n déployée en production, composée de deux workflows principaux :

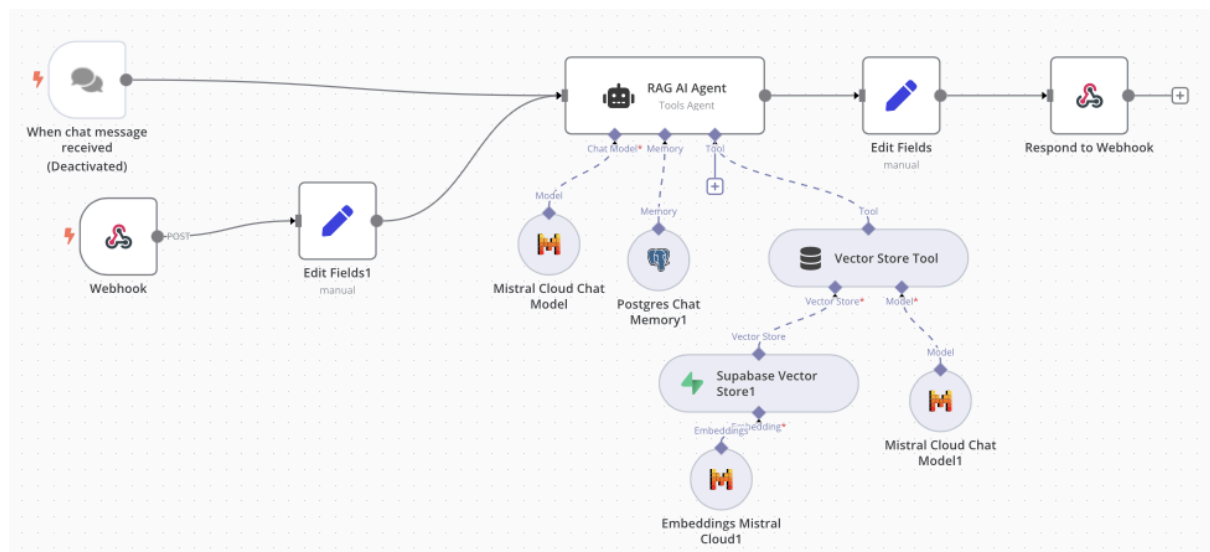
1. **Pipeline d'ingestion documentaire** : Surveillance et traitement automatique des documents

*[Screenshot 1 : Pipeline d'ingestion - de la détection Google Drive au stockage vectoriel]*



## 2. Agent conversationnel : Traitement des requêtes utilisateurs via Mattermost

[Screenshot 2 : Vue d'ensemble du workflow agent conversationnel]



### 2-b Infrastructure de données

#### Stockage vectoriel

- **Base de données** : Supabase (table `documents`) (compte affilié à julien.henry@dossierfacile.fr)
- **Modèle d'embedding** : Mistral Cloud Embeddings
- **Stratégie de chunking** : Segments de 750 caractères avec overlap de 130 caractères pour tenter de garder beaucoup de contexte.
- **Fonction de recherche** : `match_documents` (recherche sémantique)

Les serveurs stockant la base de données sont situés en Europe de l'Ouest mais sous le contrôle d'AWS, une société américaine soumise au Cloud Act.

### Gestion de la mémoire conversationnelle

- **Support** : PostgreSQL via Supabase
- **Persistance** : Par `sessionId` (correspondant à l'ID du message envoyé sur Mattermost)
- **Avantage** : possibilité de retrouver les données facilement par couple (questionHumain/réponseIA)

## 2-c Pipeline d'ingestion documentaire

### Sources de données

- **Répertoire surveillé** : Google Drive (dossier "beta Documentation RAG", situé sur le drive de [julien.henry@dossierfacile.fr](mailto:julien.henry@dossierfacile.fr). Possibilité de migrer le dossier vers "Opérations". Voir point [7-a](#)
- **Types de fichiers supportés** :
  - PDF (extraction via n8n PDF parser)
  - Excel/XLSX (extraction structurée)
  - Google Docs (conversion automatique en texte brut)

### Processus de traitement

1. **Détection** : Triggers Google Drive (création/modification de fichiers)
2. **Nettoyage** : Suppression automatique des anciennes versions vectorisées dans le cas d'un document mis à jour et non créé.
3. **Extraction** : Parsing intelligent selon le type de fichier
4. **Vectorisation** : Génération d'embeddings via Mistral Cloud
5. **Stockage** : Insertion dans Supabase avec métadonnées (file\_id, type)

## 2-d Agent conversationnel

### Modèles LLM

- **Agent principal** : Mistral-Medium-Latest
- **Agent de recherche** : Mistral-Large-Latest (pour les requêtes vectorielles)
- **Justification** : Les modèles plus légers (Mistral 7B, Ministral 3B) ne supportent pas l'utilisation d'outils. Ils ne sont pas compatibles avec le fait d'être des "agents"

### Prompt engineering

- **Identité** : "Docteur Opérateur", assistant dédié aux opérateurs Dossier Facile
- **Comportement** : Usage obligatoire du vector store avant toute réponse. Notifier explicitement s'il ne connaît pas la réponse (d'après les documents mis à sa disposition). L'objectif est de limiter les hallucinations ou les réponses inventées.

- **Format de réponse** : Personnalisation avec prénom utilisateur + réponse la plus proche possible des vecteurs extraits de la base de données.
- **Guardrails** : Recommandation systématique de vérification avec tuteur ou canal #opération

### Intégration Mattermost

- **Point d'entrée** : Webhook sortant n8n configuré pour recevoir les messages
- **Format de réponse** : JSON structuré avec le paramètre "response\_type = comment" pour répondre dans le thread et non pas dans le canal en lui même

## 5- Résultats et Performance

### 5-a Qualité des réponses

- **Précision** : Résultats cohérents et prévisibles
- **Limitation identifiée** : Imprécisions liées davantage à une documentation pas à jour. Un manque de contexte dans la documentation (exemple: dans la documentation parlant des justificatifs d'activité des garants, le terme garant n'apparaît que 2 fois, le premier et le dernier mot du document. Cela engendre le fait que la partie centrale du document, une fois chunkée, ne contient aucune mention de garant ni même du terme "justificatif d'activité")
- **Bénéfice collatéral** : Identification des lacunes documentaires

### 5-b Impact opérationnel

- **Réduction des interruptions** : Diminution des sollicitations directes aux tuteurs
- **Disponibilité** : Service 24/7 sans intervention humaine
- **Traçabilité** : Historique complet des interactions pour analyse des besoins

## 6- Aspects Économiques

### Coûts d'exploitation

- **API Mistral Cloud** : < 0,01€ par requête (coût marginal)
- **Supabase** : Version gratuite utilisée
- **Maintenance** : Système tributaire d'une documentation à jour pour fournir des réponses de qualité.

## 7- Limitations Techniques Identifiées

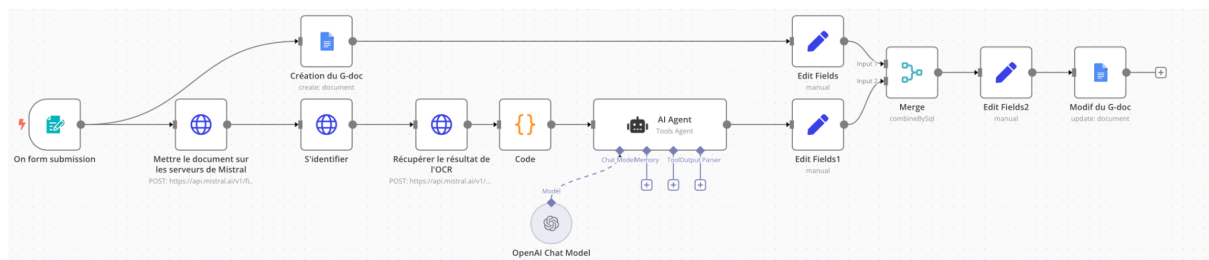
### 7-a Synchronisation documentaire

- **Problème principal** : Le noeud trigger Google Drive "File Created" ne détecte que les créations pures et les imports, pas les événements de copier/coller de documents déjà existant sur le Drive.
- **Impact** : Certaines mises à jour documentaires peuvent ne pas être propagées automatiquement en l'état actuel de la structure de notre drive partagé "Opérations"
- **Contournement actuel** : Monitoring manuel nécessaire pour les copies de fichiers
- **Solution** : Se servir du Dossier monitoré pour stocker directement la documentation dédiée à l'équipe opération et limiter les droits en écriture à ses fichiers à quelques personnes.

### 7-b Extraction de documents complexes

- **Limitation d'extraction** : Les documents PDF contenant beaucoup d'images ou avec des mises en forme complexes ne sont pas traités de manière optimale par les extracteurs natifs n8n. Exemples : guide validation réalisé par Jimmy, présentation Slides sur la détection de faux dossiers, etc
- **Solution développée** : Workflow alternatif utilisant Mistral OCR pour l'extraction avancée de données
- **Note** : Le blueprint JSON du workflow OCR sera fourni séparément pour complément d'information

[Screenshot 3 : extraction de données de PDF complexes vers un format Markdown via Mistral OCR API]



### 7-c Autres limitations techniques

- **Problème de réception de la réponse par Mattermost** : Problème constaté le Jeudi 29 mai à 23h48, suite au message envoyé par un opérateur l'ia n'a pas répondu à son message. Le workflow a pourtant réussi. Néanmoins le workflow a pris 1 mn 17 secondes à s'effectuer contre un temps habituellement situé entre 2 et

20 secondes. Je soupçonne que le “webhook respond” dispose d’un délai au delà duquel Mattermost ne prend pas en considération la requête.

## **7-d Améliorations techniques possibles**

- **Amélioration de la synchronisation** : Développement d'un système de détection des copies
- **Extension des formats** : Support de formats additionnels (images, vidéos explicatives)
- **Amélioration et mise à jour de notre documentation**

## **8- Recommandations**

### **Court terme**

1. **Audit documentaire** : Mettre à jour notre documentation et la développer au besoin.