

# Overview du concept de Retrieval Augmented Generation (RAG) et cas d'usages – Retour d'Expérience chez IMATECH

Charlotte RIEUX – Jeudi 7 mars 2024

Soirée Global AI community



# Plan

## Overview du concept de RAG et cas d'usages

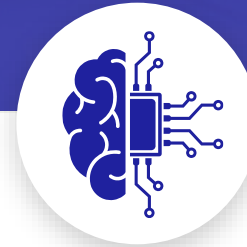
Questionnez vos documents intelligemment : personnalisez votre LLM préféré en le connectant à votre base documentaire (doc technique, CVthèque, base de données) vous permet de chercher les informations pertinentes et générer une synthèse en langage naturel



1.  
**Big picture**



2.  
**Pour quels  
besoins ?**




3.  
**Comment  
ça marche ?**



4.  
**REX IMATECH  
Assistant juridique**

Global **AI** Community



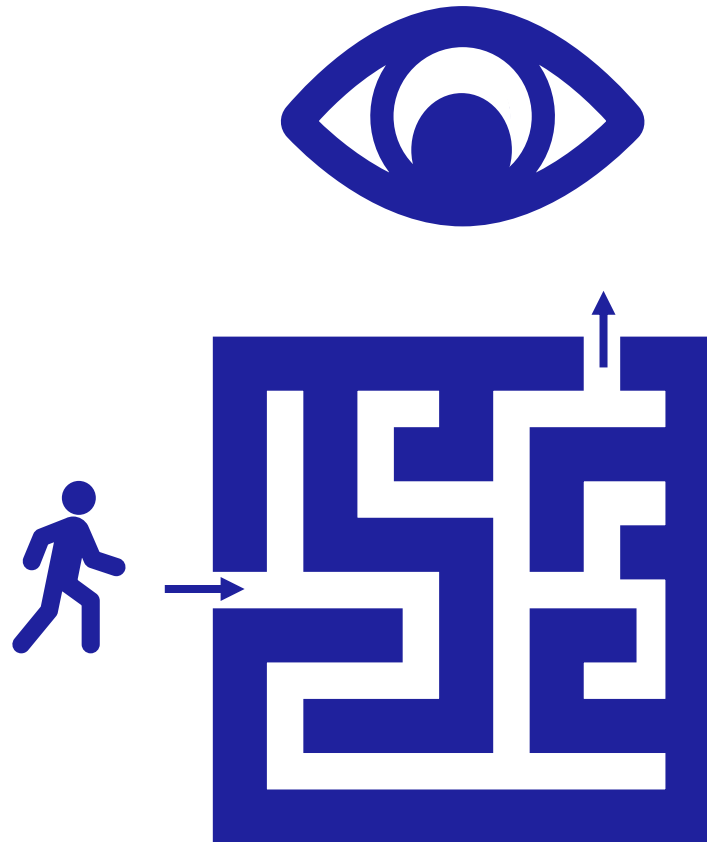
**1**

# Big picture



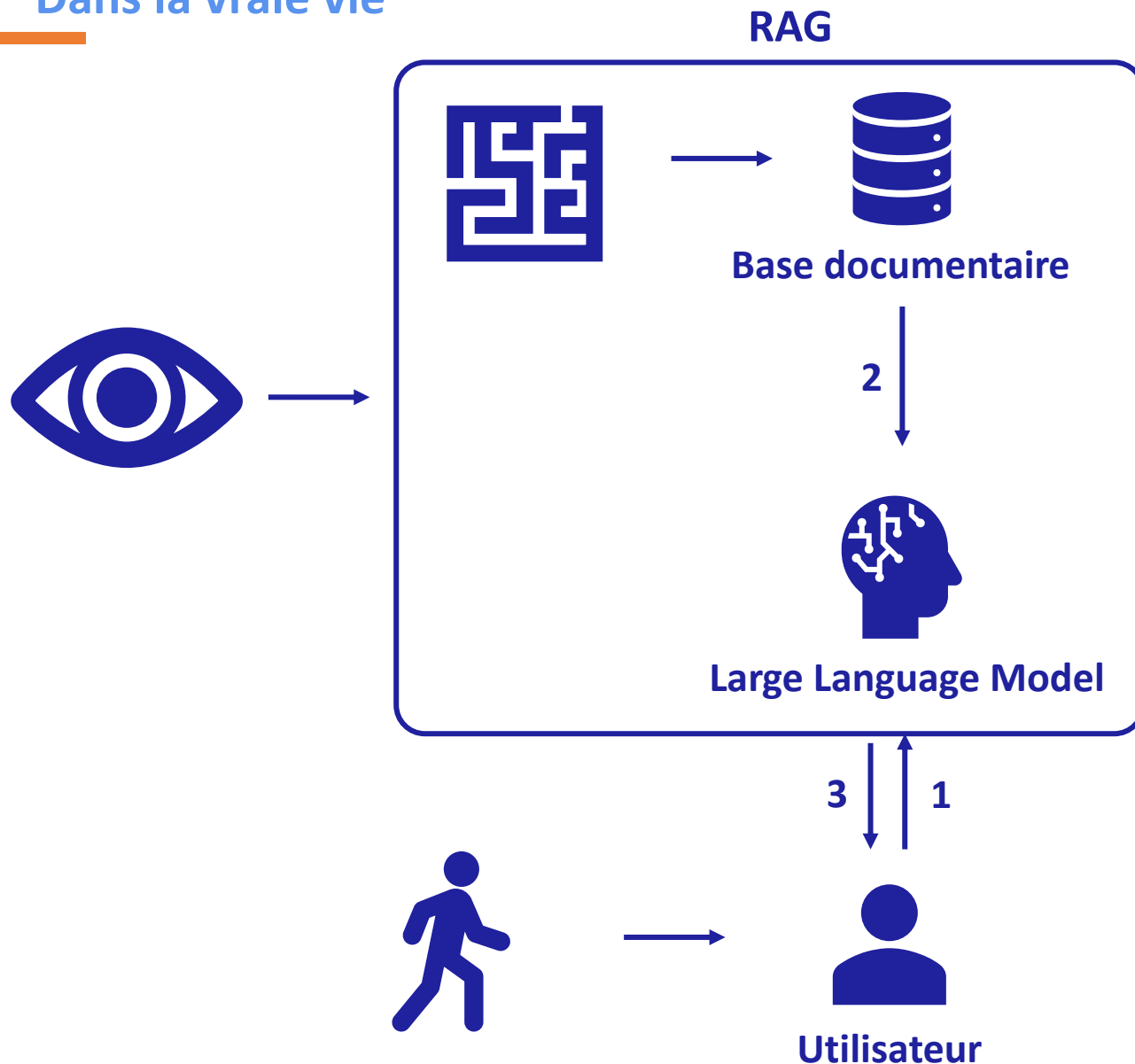
# Big picture (1/2)

L'image du Labyrinthe



# Big picture (2/2)

Dans la vraie vie



1 ) L'utilisateur pose une question

2) Recherche par similarité (dans un vector store) :

- Mots clés
- Sémantique
- Hybride

*et*

Intégration des documents trouvés dans le contexte du LLM (ex. GPT-4)

3) Synthèse de la réponse et envoi à l'utilisateur



**2**

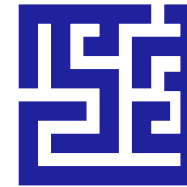
**Pour quels  
besoins ?**



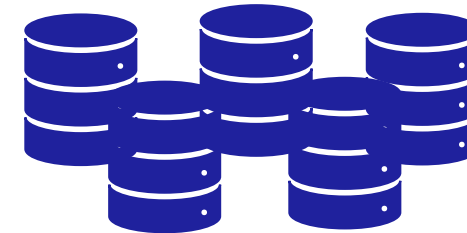
# Pour quels besoins ? (1/4)

Quels sont les problèmes qui peuvent être adressés par un RAG ?

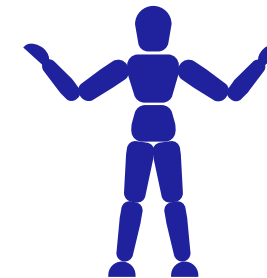
- Base de documents complexes



- Diversité + volumétrie importantes



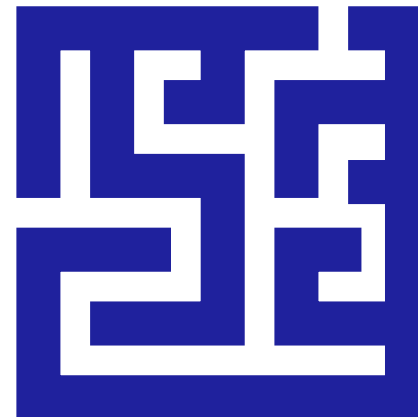
- Contenu peu maîtrisé par les utilisateurs et/ou en constante évolution



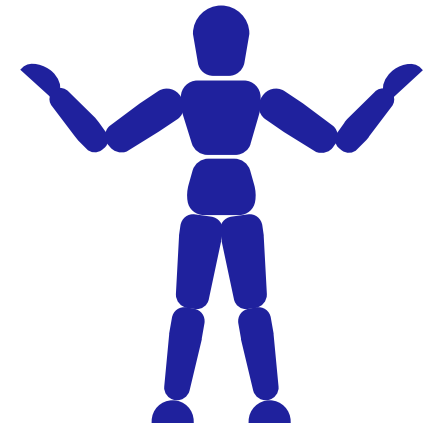
# Pour quels besoins ? (2/4)

## Exemple 1 : Client IMATECH

- Base de documents complexes
- Contenu partiellement maîtrisé par les utilisateurs et/ou en constante évolution



Fiches de synthèse  
juridique



Un turn over important  
dans l'équipe  
des juristes



# Pour quels besoins ? (3/4)

## Exemple 2 : Client dans le domaine de la data

- **Client X : Diversité + volumétrie importantes**

- Articles de presse
- Comptes rendus de conseils municipaux
- Dossiers de consultation pour les entreprises
- Marchés publics

Sur toute la France : des millions de documents



(1) LLM + demande = filtres sur plusieurs champs l'ensemble des documents

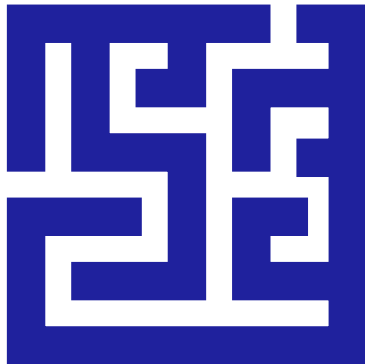
*puis*

(2) Recherche par similarité dans le champ contenu des documents

# Pour quels besoins ? (4/4)

Exemple 3 : client dans le développement de jumeaux numériques

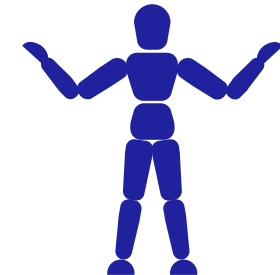
- Client Y : Tout à la fois



Un dashboard complexe  
de 7 pages, 40 graphes et  
tableaux au total  
décrivant un jumeau numérique de  
plusieurs usines



20 tables différentes,  
To de données de  
simulation  
(1) LLM + description des  
tables = Query de la BDD  
*et*  
(2) Recherche dans la base  
de connaissance  
(contexte métier)



Approvisionneur de la  
Supply Chain qui ne  
connait pas le modèle

2

**Comment ça  
marche ?**

# Comment ça marche ? (1/4)

Quelles contraintes cela soulève ?



1) La qualité de la réponse dépend du contexte :

- Une demande précise
- Un prompt efficace
- Des documents pertinents

Le contexte est limité en taille !

2) Recherche de documents dépend de :

- Une indexation intelligente :
  - Choix de la taille des chunks + overlap
  - Nombre k de documents retournés

# Comment ça marche ? (2/4)

## Création d'un vector store

Le skillset est l'ensemble des paramètres d'indexation d'un vector store :



Un document de 350 pages  
(5 000 k tokens)

ne pourra pas passer dans le  
contexte de 16 k tokens

1) Choix du modèle d'embedding

2) Choix des champs (type, contenu) par ex :

- Titre : string
- Date : datetime
- Contenu : string
- Vecteur : embedding

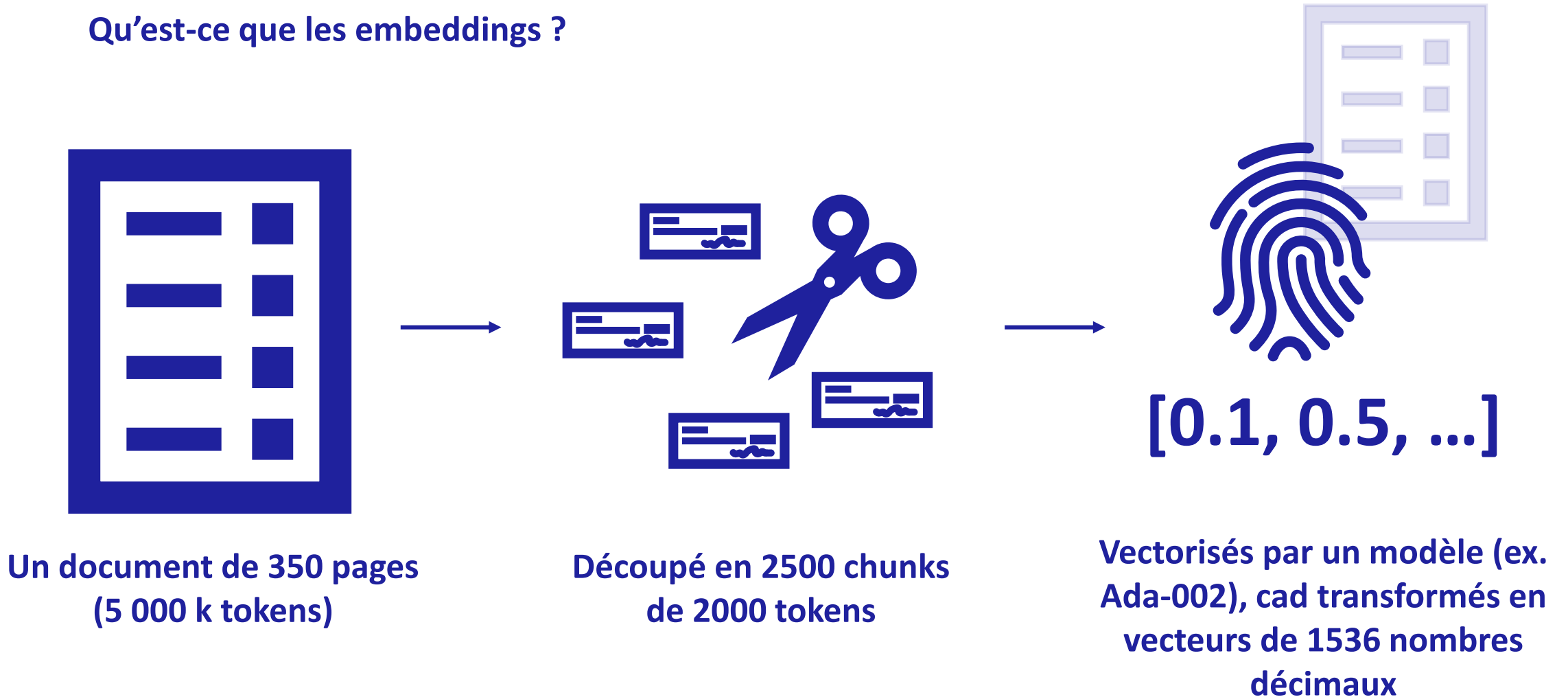
2) Taille des chunks en tokens (1 token = 3-4 caractères)

3) Overlap des chunks

# Comment ça marche ? (3/4)

Indexation des documents dans le vector store

Qu'est-ce que les embeddings ?



# Comment ça marche ? (4/4)

## En résumé

1) Contrainte n° 1 : la taille du contexte du LLM est limitée, comment faire ?

→ Solutions : (1) chunking et (2) recherche des chunks pertinents

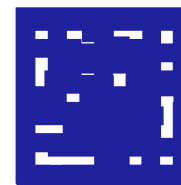
2) Contrainte n° 2 : la machine ne comprend pas le sens du texte, comment trouver les documents pertinents ?

→ Solution : (3) vectorisation, création des embeddings et (4) recherche des embeddings pertinents pour la question

Le choix de la taille des chunks est aussi important pour la recherche par similarité, exemple :



1 information sur  
90 x 90 pixels



3 informations sur  
90 x 90 pixels



4

**REX IMATECH**





# REX IMATECH (1/2)

Assistant juridique



~ 3000 documents sur des thématiques telles que le logement, la famille, la vie socio-professionnelle, etc ...



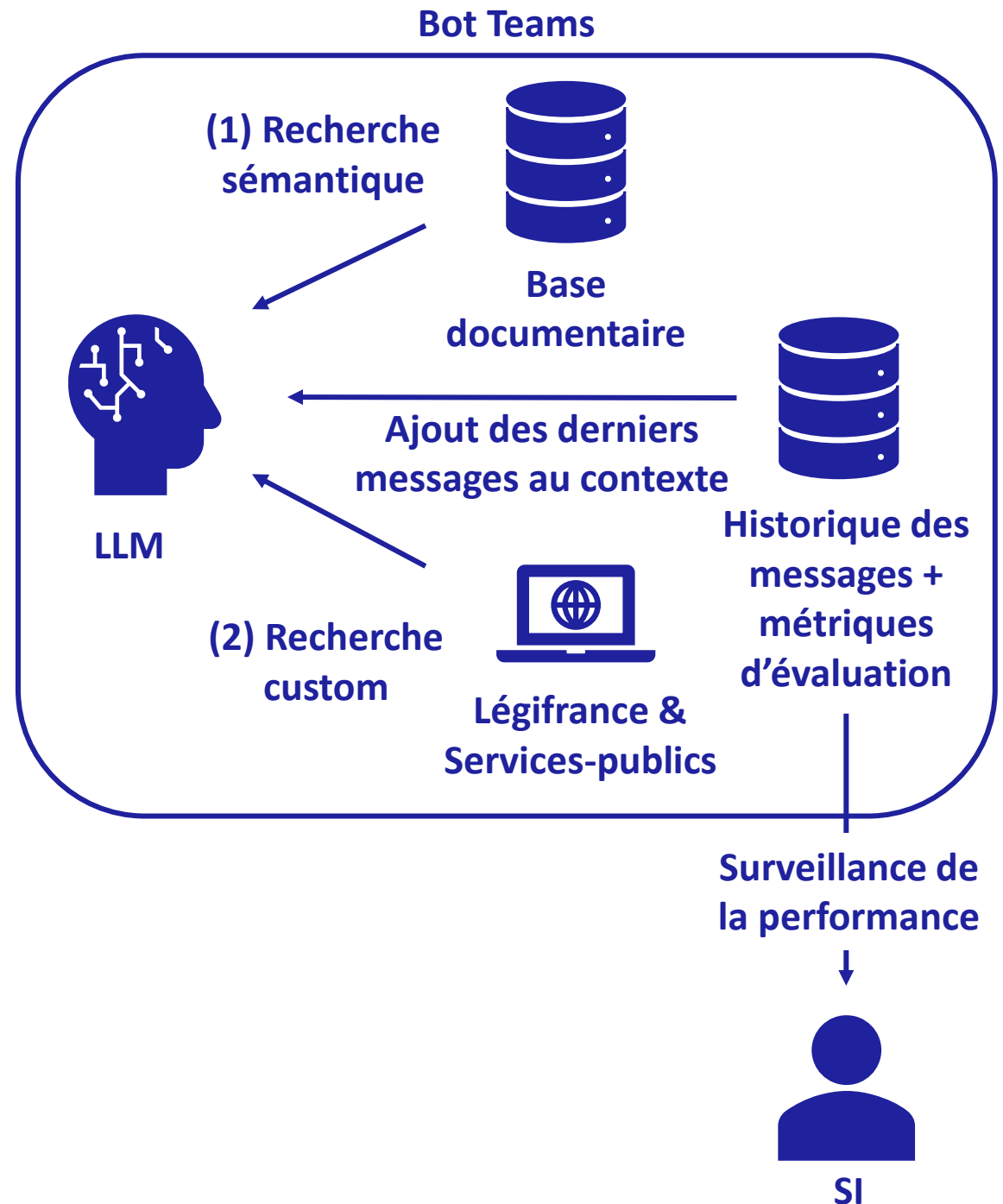
centre d'appel de  
~ 150 juristes,  
veille nécessaire pour  
mettre à jour les  
documents



~ 15 appels par jour/  
juriste  
~ 2250 appels par jour au  
total

# REX IMATECH (2/2)

## Assistant juridique





**Merci !**