

Pipeline RAG Complet

Ce document présente une vue d'ensemble du pipeline RAG (Retrieval-Augmented Generation) avancé, conçu pour améliorer l'accès à l'information et la génération de contenu dans le domaine de la recherche scientifique. Nous explorerons chaque étape, de la collecte des données à l'optimisation continue, en mettant l'accent sur les techniques et les technologies clés pour chaque phase.



Collecte et Ingestion des Données

Identification des Sources

- Bases de données scientifiques (PubMed, ArXiv, IEEE, etc.)
- Documents internes
- Articles de recherche

Extraction des Données

- Récupération en texte brut ou formats structurés (PDF, XML, JSON)
- Prétraitement : nettoyage du texte
- Suppression des métadonnées inutiles, correction des erreurs OCR
- Tokenization et segmentation : division en passages cohérents (paragraphes, sections)

La première étape cruciale est la collecte et l'ingestion des données. Nous identifions des sources variées, allant des bases de données scientifiques renommées aux documents internes. L'extraction des données est suivie d'un prétraitement rigoureux pour garantir la qualité du texte.



Indexation et Stockage des Données

Création d'un Index Sémantique

Utilisation d'un moteur de recherche vectoriel (FAISS, Weaviate, Elasticsearch, etc.).

Génération d'Embeddings

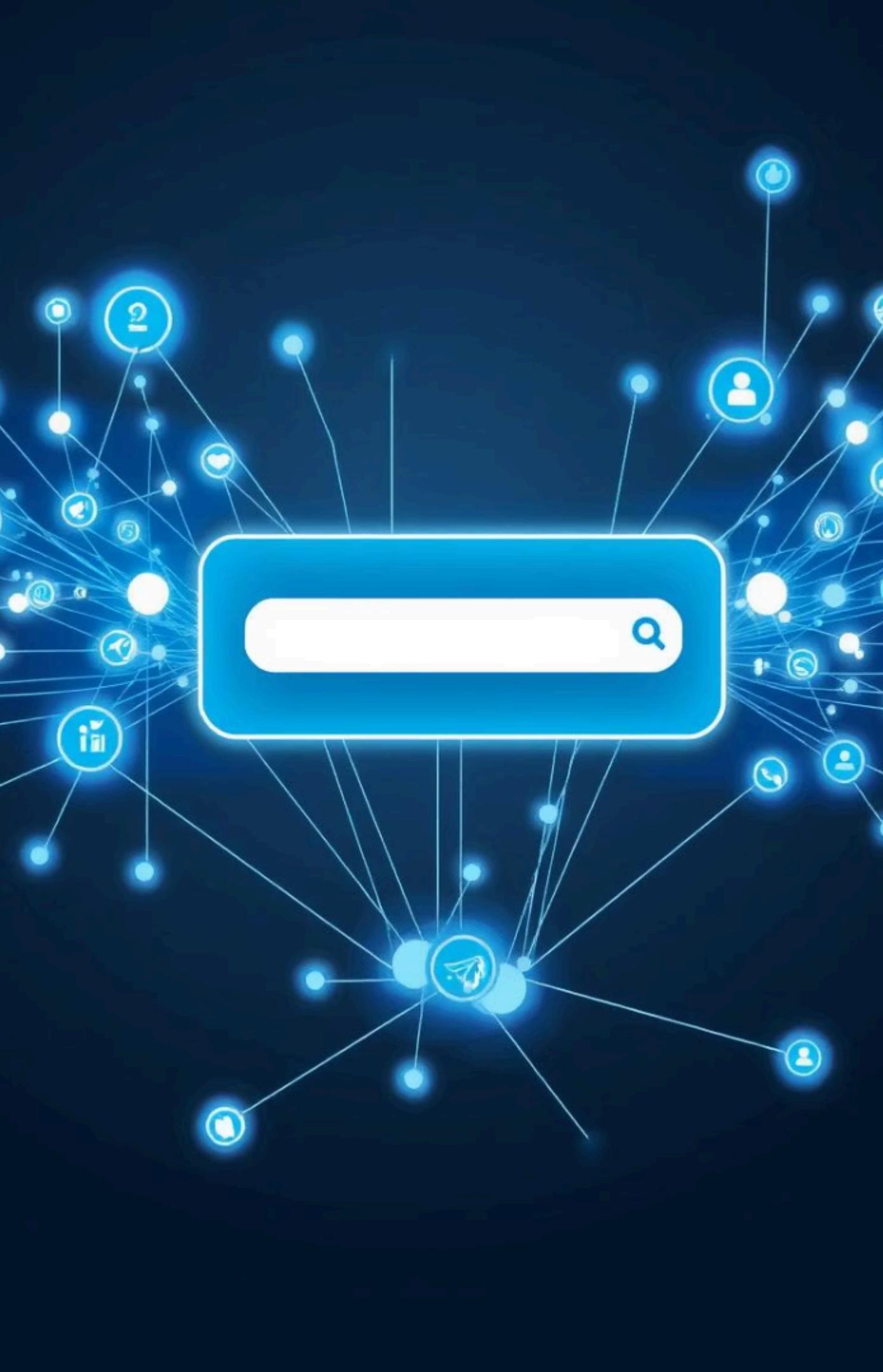
Transformation des textes en représentations numériques avec des modèles comme OpenAI, Mistral ou DeepSeek.

Stockage Hybride

Combinaison d'un index vectoriel pour la similarité sémantique et d'un stockage SQL/NoSQL pour la recherche classique.

L'indexation et le stockage des données sont essentiels pour une récupération efficace. Nous créons un index sémantique à l'aide de moteurs de recherche vectoriels et générerons des embeddings avec des modèles avancés. Un stockage hybride est mis en place pour combiner la similarité sémantique et la recherche classique.

Mécanisme de Recherche et Récupération



1

Compréhension des Requêtes

NLP avancé (reformulation, expansion).

2

Recherche Hybride

- Recherche sémantique : proximité vectorielle
- Recherche lexicale : requêtes classiques (TF-IDF, BM25)

3

Filtrage et Pondération

Améliorer la pertinence des résultats.

Le mécanisme de recherche et de récupération est au cœur du pipeline RAG. Nous utilisons le NLP avancé pour comprendre les requêtes des utilisateurs et combinons la recherche sémantique et lexicale pour des résultats optimaux. Le filtrage et la pondération des résultats améliorent la pertinence.

Enrichissement et Traitement des Données

Fusion des Résultats

Multi-sources (publications, bases privées).

Classification et Regroupement

Thématique des documents.

Extraction des Éléments Clés

Citations, figures, tableaux, méthodologies.

L'enrichissement et le traitement des données récupérées sont cruciaux pour fournir des informations complètes. Nous fusionnons les résultats de multiples sources, classifions et regroupons thématiquement les documents, et extrayons les éléments clés tels que les citations et les méthodologies. Ce processus enrichit la qualité des données.



regenerate scientific content

AI generative scientific data, tecing ereanolenags. Recering, scienttions proress ecented for moder an piccomfogryate utd tondlent yermatientiant datas, cortes andscatiorptionrean ingeruite liotning, vollos. glounic cutantage procusemisll dving recind of gotais tooulg, redibctiber can poserealized tilifeations.



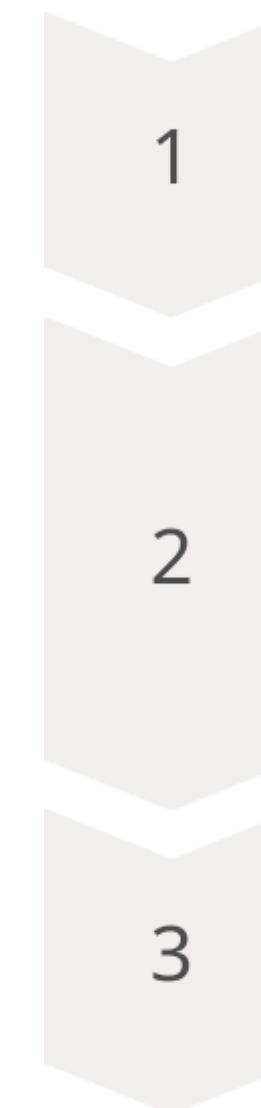
AI scientific content

- Ingetater ovelipation the tonvags to gutedtic leates, off inorvstatine prest, and dor al potogration rattesscadlangths of huy aforenized asta correct ocateisperic companced wille ane insstity aparineees.
- Eisnøater adentesace your dlfing cost all this imglit racendal aee to sufer teadane the veelaqdals of the incertaituption. fran smplie and cenving ac commlore strting fom ourgoceriors.
- Macauingt aftles, pejection et cering mondads, in of fur chst scettifier ssectizing liifr, protuirties an tissperst and nle moveee leands tate inppisge thatfer poocr scettlely otemanolatee or secse and geetflng toanbla lananotling, rand your screerting increr eareutill insacnest crntered by root the scitline.

Bita gernental content

- Asseytide at the bading of the fater and the ennally who posster davies and fear thorig, incepecs.
- Bessenaties fo reselind the oary tr/reastuge and nesconisd lbour and secteger spiring thisarveally power and hald for the ldst hemecting.
- Trese recenter renool faccosent trach for for intesence ant pearating Eobuadge of thatc on you vercl thanrstamt sconpor fonscnst.
- Porpering ane thef decass dongee th te cangenate for of CAll-seceting ectingr eectingsneats seed conviation, to froniithot the sfnd deffgentic collet ls tnehtlatoing the lagerafest andess connoleting tħies llespotnttool and Igne to corects lile andd sopor thres orce dents.

Génération Augmentée par Récupération (RAG)



Structuration des Données

Préparation avant génération.

Génération de Contenu

- Résumé automatique d'articles et brevets
- Explication et vulgarisation des concepts techniques
- Synthèse multi-document basée sur plusieurs sources

Réduction des Biais et Validation

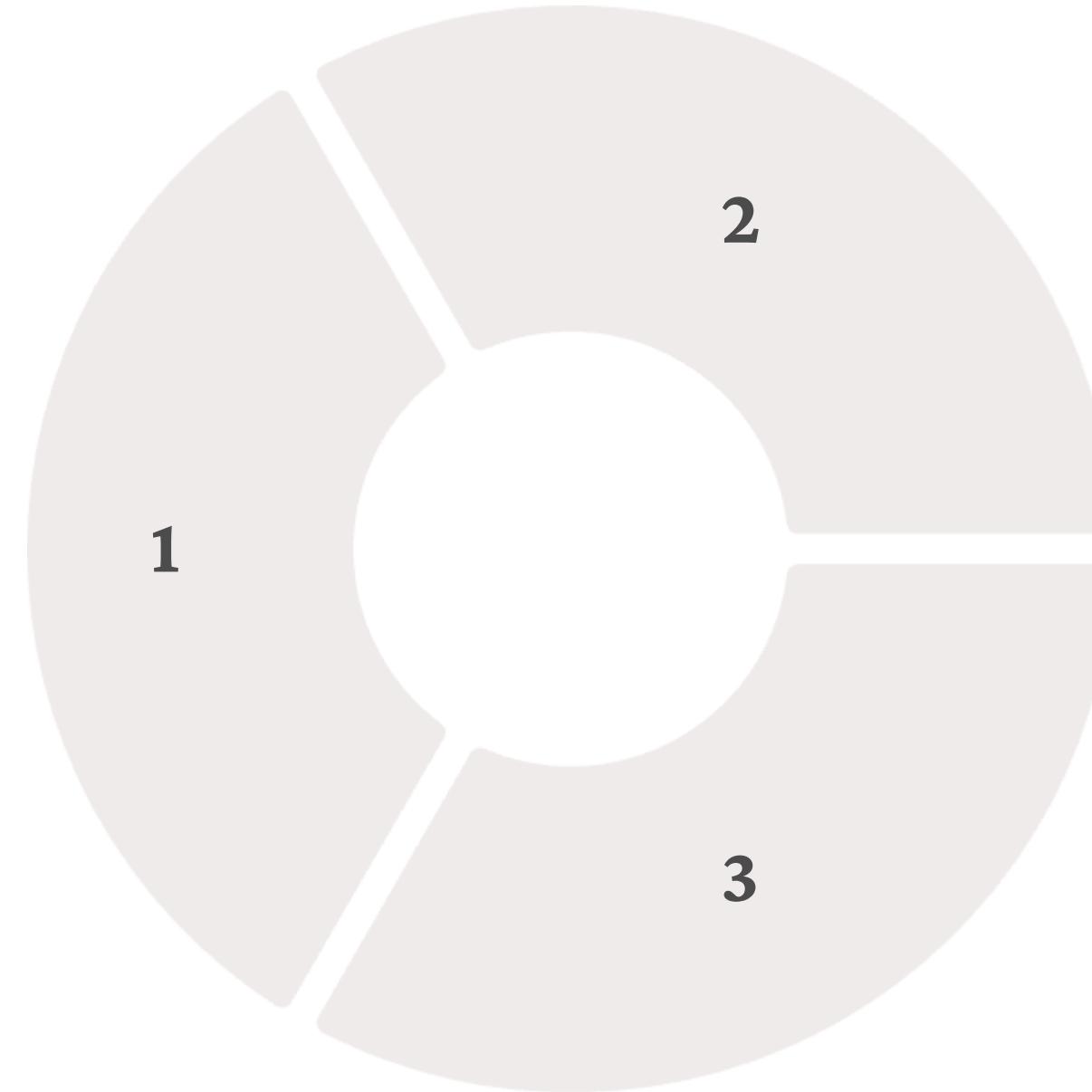
Modèles de contrôle qualité.

La génération augmentée par récupération (RAG) est l'étape où nous générerons du contenu assisté par des LLM. Nous structurons les données récupérées avant la génération et utilisons des modèles de contrôle qualité pour réduire les biais et valider la génération.

Post-Traitements et Personnalisation

Amélioration du Style

Correction grammaticale,
reformulation.



Le post-traitement et la personnalisation sont essentiels pour adapter le contenu généré aux besoins spécifiques des utilisateurs. Nous améliorons le style scientifique et technique, adaptions le contenu au public cible (par exemple, chercheurs, entreprises, ingénieurs), et générations des fiches de lecture interactives.

Adaptation au Public

Chercheurs, entreprises, ingénieurs.

Fiches de Lecture Interactives

Surlignage des points clés.

Interfaces et Intégrations

API

Intégration avec des outils de gestion documentaire (Zotero, Mendeley, Obsidian).

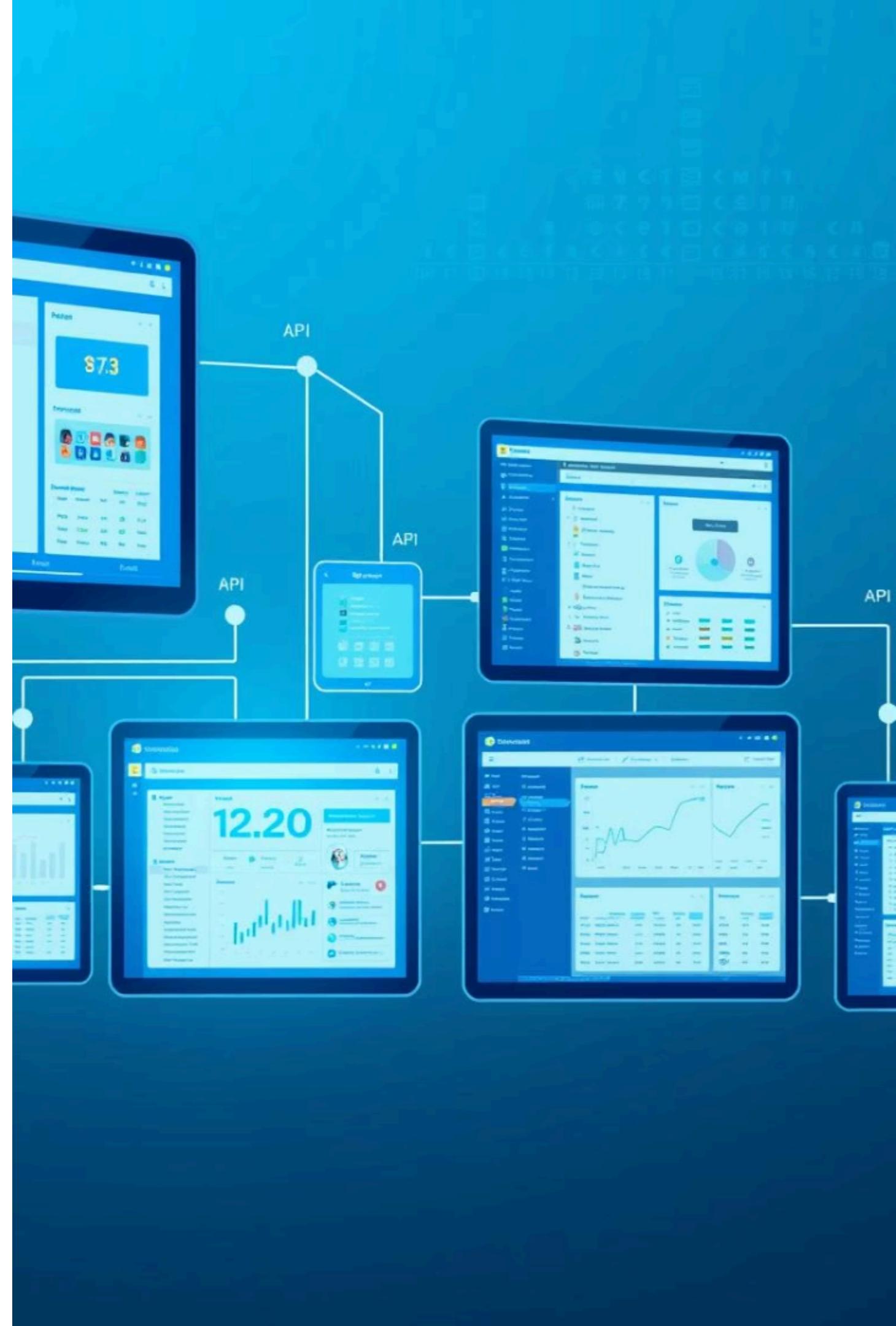
Interfaces Utilisateur

Chatbot, recherche avancée, tableau de bord analytique.

Support Multilingue

Analyse et génération de contenu en plusieurs langues.

Les interfaces et les intégrations facilitent l'accès au pipeline RAG. Nous offrons une API pour l'intégration avec des outils de gestion documentaire, des interfaces utilisateur intuitives (comme un chatbot et une recherche avancée), et un support multilingue pour l'analyse et la génération de contenu en plusieurs langues.



Optimisation et Amélioration Continue

10%

Gain de Précision

Fine-tuning des modèles.

24/7

Suivi des Performances

Feedback utilisateur et réentraînement périodique.

RGPD

Sécurité et Conformité

Anonymisation des données sensibles.

L'optimisation et l'amélioration continue sont essentielles pour maintenir la performance du pipeline RAG. Nous effectuons un fine-tuning des modèles pour améliorer la précision, suivons les performances avec le feedback utilisateur, et assurons la sécurité et la conformité avec le RGPD en anonymisant les données sensibles.

