SUP MTI
École Supérieure de Management, de Télécommunication et d'Informatique

RECONNUE PAR L'ÉTAT

SMART FACTORY

# PRÉSENTATION PROJET PFE

## THÈME: ASSISTANT IA/NLP


PRÉSENTER PAR: ABDOUL FATAOU HAMA ET MAHAMAT YAYA HISSEIN

ENCADRÉ PAR : MME HASNA CHAIBI

# Sommaire

# INTRODUCTION

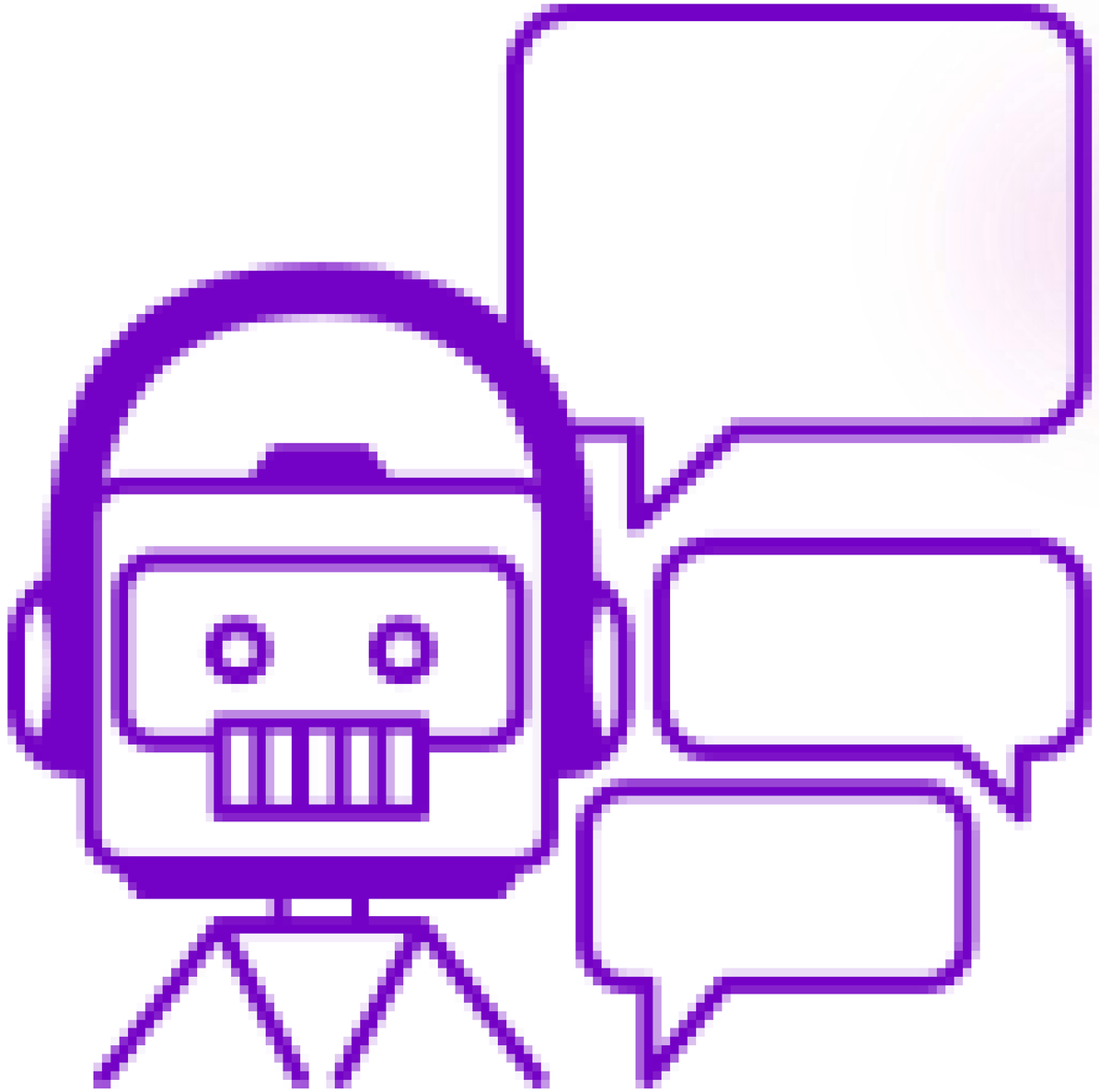
L'essor de l'intelligence artificielle et des grands modèles de langage (LLM) offre des opportunités inédites pour la recherche scientifique et le management d'entreprise. Pourtant, l'accès à l'information pertinente et la gestion efficace des connaissances restent des défis majeurs pour les chercheurs, les doctorants, les ingénieurs et les managers.

**Pourquoi un Assistant IA NLP ?**

**Chercheurs et doctorants** : Difficultés à trouver et analyser rapidement les travaux pertinents.

**Ingénieurs** : Recherche d'informations techniques et support à la rédaction de documentation.

**Entreprises** : Besoin d'un système de veille stratégique et d'aide à la prise de décision.



# Problématiques et solutions proposées

## Défis actuels

- Volume massif de données académiques et techniques.
- Difficultés à extraire des informations précises et pertinentes.
- Temps nécessaire pour analyser et synthétiser des documents.

## Solutions proposées

- Résumer des articles académiques et documents techniques.
- Assister à la rédaction et à la reformulation de textes.
- Automatiser la veille scientifique et technologique.
- Optimiser l'analyse de documents en entreprise.

# Fonctionnalitées clés

1

## Analyse et Recherche Documentaire

Recherche avancée NLP, exploration sémantique, analyse des tendances scientifiques.

2

## Résumé et Extraction d'Informations

Génération de résumés, extraction de citations et références clés, création de fiches de lecture intelligentes.

3

## Assistance à la Rédaction

Génération automatique d'introductions, correction grammaticale et amélioration du style scientifique.

# Solutions existantes

01

**ChatGPT**

Génération de texte, assistance à la rédaction, résumés.

02

**DeepSeek**

Spécialisé dans l'analyse de documents techniques et scientifiques.

03

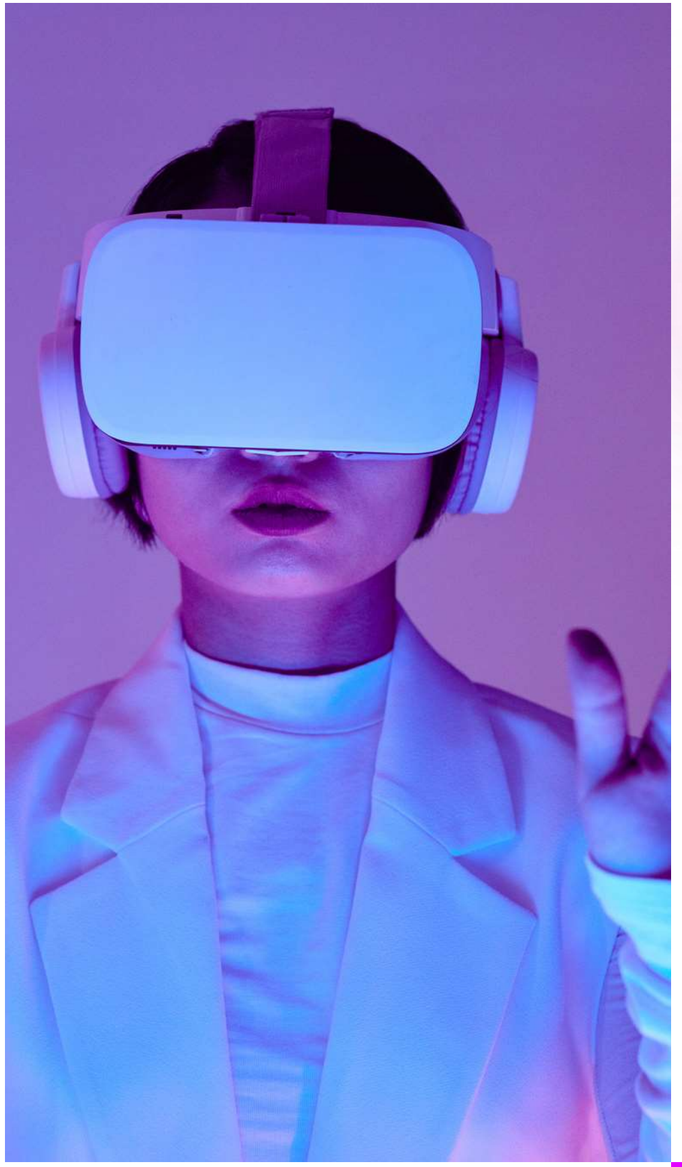
**Google Gemini**

Recherche et traitement avancé d’informations avec intégration multimodale.

04

**Mistral**

Modèles entraînés sur des cas d’usage spécifiques (finance, juridique, santé...).



# Comparaisons des assistants IA existants

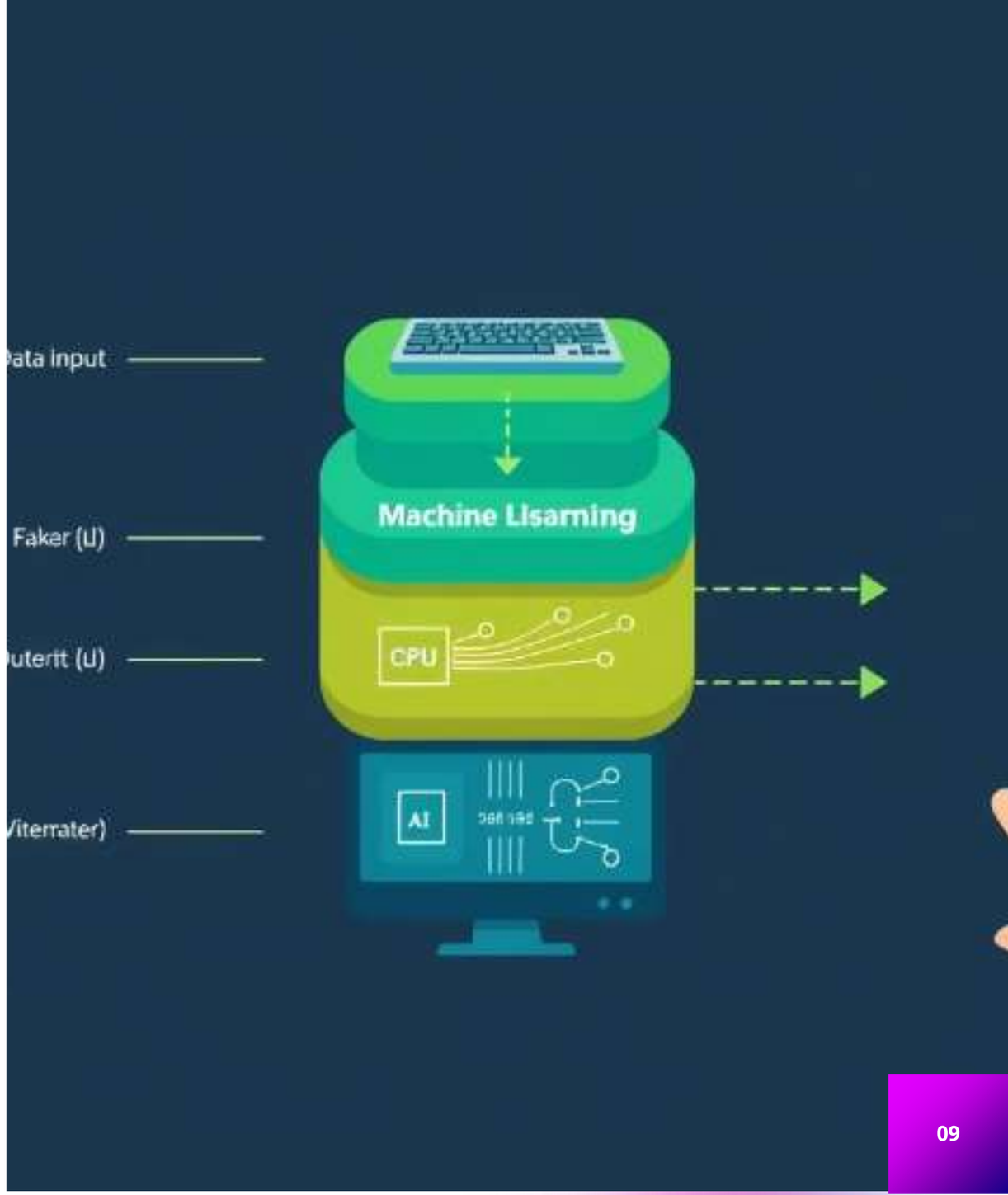
| Critère / Modèle | ChatGPT (OpenAI) | DeepSeek | Mistral (via Le Chat ou Ollama) | Gemini (Google) |
|---|---|---|---|---|
| Compréhension du langage | Très solide (surtout GPT-4) | Bonne, surtout en technique/scientifique | Bonne en français, clair et précis | Très bonne, surtout Gemini 1.5 |
| Génération de texte | Fluide, naturel, multilingue | Précise, un peu plus "technique" | Bonne mais parfois plus brute que GPT | Bonne, contexte long bien géré |
| Capacités en codage | Excellentes avec GPT-4 | Très bon pour la programmation | Correct, mais moins poussé que GPT-4 | Excellentes (surtout Gemini 1.5 Pro) |
| Multimodalité (image, audio) | Oui (image avec GPT-4V, audio avec Voice) | ● Non | ● Non | Oui (image, audio, vidéo pour certains) |
| Context window (mémoire) | GPT-4 classique : 8k à 128k (GPT-4 Turbo) | 32k tokens environ | ● Limité (~8k-16k selon déploiement) | Jusqu'à 1 million de tokens (Gemini 1.5 Pro) |
| Open-source / Local | ● Non, uniquement API/OpenAI | Pas encore local facilement | Oui (Mistral 7B, Mixtral, open source, via Ollama) | ● Non, fermé (usage via Google uniquement) |
| Vitesse de réponse | Très rapide | Rapide | Rapide | Rapide |
| Personnalisation / fine-tuning | Possible avec GPTs + API | Possible localement (poids modifiables) | Facile en local avec Ollama | ● Fermé à la personnalisation pour l'instant |
| Coût d'utilisation | ● Assez élevé (GPT-4) | Gratuit/local ou API abordable | Gratuit en local (via Ollama ou HuggingFace) | Inclus dans certains produits Google |
| Respect des données / confidentialité | Bon, mais dépend du cloud OpenAI | Peut être utilisé en local, plus contrôlable | Usage local = confidentialité maximale | ● Cloud Google = dépendant de leur politique |
| Langues supportées | Très large (dont français, arabe, etc.) | Anglais + chinois bien maîtrisés | Bon en français et langues européennes | Multilingue performant |

# Architecture et Pipelines d’assistant basé sur RAG

Cette solution innovante combine les forces de l'IA et du NLP pour offrir une assistance intelligente dans divers domaines, de la recherche scientifique à l'analyse commerciale. Nous allons explorer les différentes couches de l'architecture, les outils et technologies utilisés, ainsi que les modules fonctionnels qui composent cet assistant. Préparez-vous à plonger au cœur de l'IA et à découvrir comment elle peut transformer votre façon de travailler.

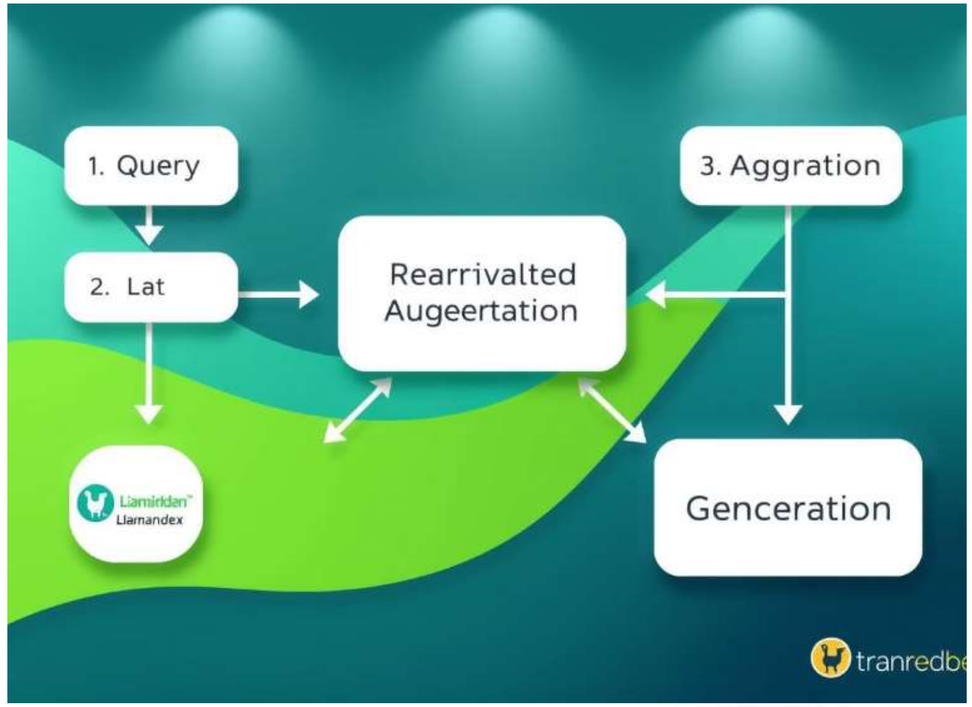
# Aperçu de l'Architecture

**Notre architecture est structurée en plusieurs couches, chacune ayant un rôle spécifique dans le pipeline NLP-IA. La couche Frontend utilise React et Next.js avec TailwindCSS pour une interface utilisateur moderne et réactive. L'Orchestration est gérée par LangChain, qui coordonne les différents composants. Le RAG Pipeline s'appuie sur LlamaIndex et une base de données vectorielle (FAISS/ElasticSearch) pour la recherche et la génération de contenu. Les LLMs (DeepSeek, Mistral, LLaMA, GPT, Claude) fournissent les capacités de traitement du langage. Enfin, le Data Storage utilise Azure CosmosDB, PostgreSQL, S3 et Data Lake pour stocker les données.**



# Pipeline d'Assistant Basé sur RAG (LlamaIndex + LangChain)

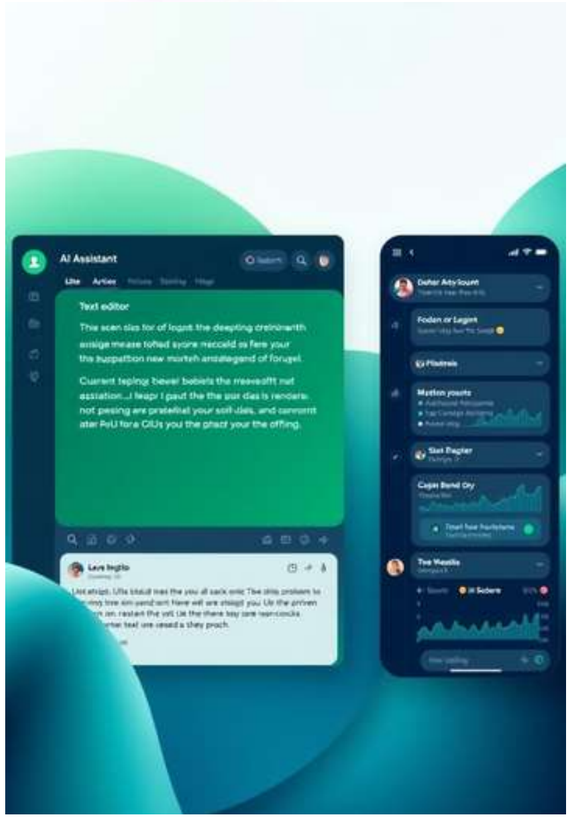
**Le cœur de notre assistant NLP-IA est le pipeline RAG, qui combine LlamaIndex et LangChain pour une génération de contenu augmentée par la recherche. La première étape est l'Ingestion et le Prétraitement des Documents avec LlamaIndex,, les Embeddings sont créés et stockés dans une base de données vectorielle (FAISS). L'Intégration LLM via LangChain utilise des LLMs comme DeepSeek ou Mistral pour la complétion des prompts, avec le Retriever de LlamaIndex fournissant le contexte. Enfin, le LangChain Tooling + Agents implémente des outils comme le chargeur de documents, le surligneur de mots-clés et le générateur de résumés.**



# Modules Fonctionnels (Codés avec LlamaIndex + LangChain)

**Notre assistant NLP-IA est composé de plusieurs modules fonctionnels, chacun conçu pour répondre à des besoins spécifiques. Le module Recherche & Analyse gère les requêtes sémantiques avec VectorStoreIndex et analyse les tendances en utilisant des embeddings de documents filtrés par le temps. Le module Résumés & Extraction utilise DocumentSummaryIndex et RAKEKeywordTableIndex pour extraire les termes clés et générer des résumés. L'Assistance à la Rédaction propose des modèles avec des prompts LangChain pour la rédaction d'introductions et de revues de littérature. Enfin, le module R&D / Business Insights génère des rapports de marché à partir de clusters de documents et analyse les tendances de l'industrie.**

# Intégration & UX/UI



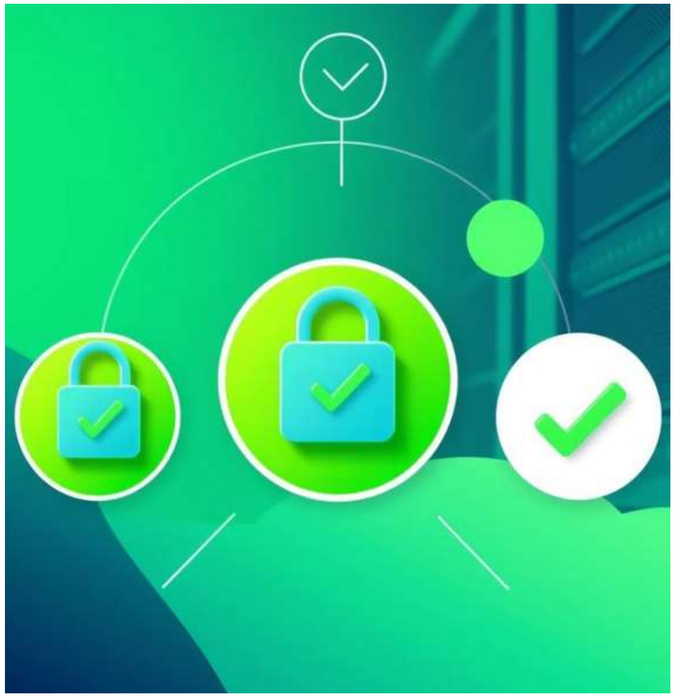

**L'API Backend est construite avec FastAPI pour fournir des points de terminaison pour le téléchargement de documents, l'interrogation de l'assistant NLP, la récupération de résumés/analyses et la sauvegarde des sessions utilisateur. L'UX Frontend offre une interface utilisateur intuitive avec un éditeur de texte, un chat AI dans la barre latérale et des informations générées automatiquement (résumés, cartes de citations). Un sélecteur de langue est également inclus. L'objectif est de fournir une expérience utilisateur fluide et efficace pour interagir avec l'assistant NLP-IA.**

# Stratégie d'Évaluation

| Composant | Métrique |
| --- | --- |
| Récupération | Recall@K, MRR, Précision |
| Génération | ROUGE, BLEU, Évaluation humaine |
| Utilisabilité | Taux de réussite des tâches, Tests UX |

Pour garantir la qualité et l'efficacité de notre assistant NLP-IA, nous avons mis en place une stratégie d'évaluation rigoureuse. Nous évaluons la Récupération en utilisant des métriques telles que Recall@K, MRR et la Précision. La Génération est évaluée avec ROUGE, BLEU et des évaluations humaines. Enfin, l'Utilisabilité est mesurée par le taux de réussite des tâches et des tests UX. Cette approche nous permet didentifier les points forts et les points faibles de notre système et de l'améliorer continuellement.

# Sécurité, Conformité & Déploiement



La sécurité des données est une priorité absolue. Nous chiffrons les téléchargements et anonymisons les documents pour protéger la vie privée des utilisateurs. Nous respectons également la conformité GDPR en permettant la suppression des données utilisateur sur demande. Un contrôle d'accès basé sur les rôles (étudiant, professionnel, entreprise) est mis en place pour gérer les autorisations. Pour le déploiement, nous dockerisons le backend et déployons sur Azure (Data Lake + Cosmos DB + API Gateway). Nous utilisons les API HuggingFace ou Replicate pour DeepSeek/Mistral.

# Conclusion

L'avenir de notre assistant NLP-IA est prometteur. Nous prévoyons d'ajouter des fonctionnalités telles que la collaboration multi-utilisateurs en temps réel, ainsi que d’offrir à l’utilisateur le choix du model donc d’utiliser pour notre assistant plusieurs model d’IA via un menu déroulant.

# THANK YOU