DataBird

Projet final

Construire un RAG (Retrieval-Augmented Generation)

Qu’est ce qu’un RAG ?

Définition

Comment ça fonctionne ?

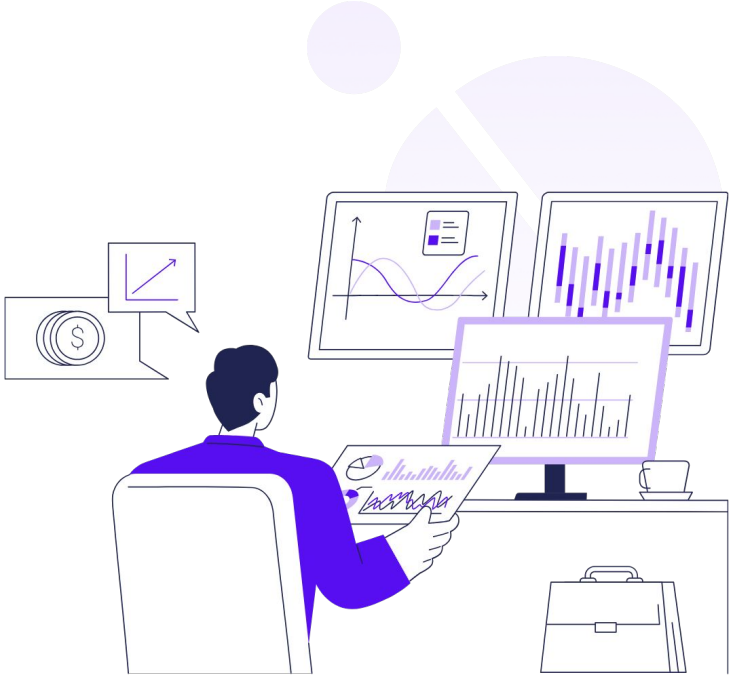
Le RAG (Retrieval-Augmented Generation) est une technique d'IA combinant la récupération d'informations dans un document et la génération de texte via un LLM.

D'abord, le modèle récupère des données pertinentes à partir de grandes bases de données ou documents, puis utilise ces informations pour générer des réponses plus précises et contextuellement appropriées.

Définition

L’utilisation d’un RAG améliore la qualité et la pertinence des réponses fournies par les modèles de langage, notamment dans les tâches de question-réponse et de conversation.

Cela permet d’avoir à disposition un LLM qui s’adapte à des cas d’usage spécifiques.



Les étapes du projet

Quels sont les étapes du projet ?

Notre objectif : avoir à disposition un LLM capable de répondre à nos questions sur une BDD spécifique

- Utiliser **HuggingFace** pour récupérer une **base de données textuelle** (des tweets sur le changement climatique) et un **LLM** (GPT-2)

- Utiliser la librairie python **Langchain** pour faire le lien entre le **LLM** et la **BDD**

- Utiliser **Gradio** pour construire une **interface graphique** simple sur laquelle on peut questionner le modèle

Quelle approche ?

Avoir une approche pas à pas

- Construire le code au fur et à mesure avec la réflexion étape par étape

- Présenter HuggingFace

- Montrer comment choisir le LLM et la BDD, comment créer une clé api

- Comment mettre le projet sur GitHub