TP : Création d'un Chatbot basé sur un RAG

Le TP vise à concevoir un système RAG (Retrieval-Augmented Generation) capable de répondre à des questions d'utilisateurs en s'appuyant sur des données de Wikipédia sur la thématique de l'IA générative. Vous allez mettre en œuvre l'ensemble des étapes d'un pipeline RAG :

— Récupération et préparation de documents,

— Indexation et recherche sémantique,

— Génération de réponses contextuelles avec un modèle pré-entraîné.

Étapes du TP

1. Préparation des données

Vous allez récupérer un ensemble de pages Wikipédia portant sur l'IA générative. Ces pages constitueront la base documentaire de votre chatbot.

a) Télécharger les contenus des pages Wikipédia sélectionnées (par exemple : *IA générative*, *GPT*, *Google Gemini*, *Transformers*, etc.).

b) Stocker les textes dans un format exploitable (JSON, CSV,...).

c) Nettoyer les textes : suppression des balises, normalisation des caractères, etc.

2. Indexation des documents et récupération

Indexer les documents pour permettre leur recherche rapide et pertinente.

a) Découper les documents en *chunks* de taille adaptée (par exemple 500–1000 tokens).

b) Créer des **embeddings** pour chaque chunk à l'aide de `OpenAIEmbeddings`.

c) Stocker ces embeddings dans une base vectorielle (`Chroma`).

d) Implémenter une fonction `retrieve_documents(query)` qui :

— génère plusieurs reformulations de la requête utilisateur ;

— récupère les documents les plus pertinents pour chaque reformulation ;

— fusionne et trie les résultats.

— renvoie les 5 documents les plus pertinents.

3. Génération de la réponse

Vous allez maintenant utiliser un modèle de génération (GPT-4-o) pour produire la réponse finale.

— Utilisez les documents renvoyés par `retrieve_documents` comme contexte.

— Générez la réponse en **mode streaming** (flux continu).

— Affichez, à la fin de chaque réponse, les liens vers les pages Wikipédia d'où proviennent les informations.