

Retrieval-Augmented Generation

Campagne d'évaluation TREC RAG - cas d'usage BNF

Encadrement : Laure Soulier (laure.soulier@isir.upmc.fr)

Mots-clés : IA, modèles de langue (ChatGPT), moteur de recherche

1 Contexte

La campagne d'évaluation TREC¹ propose tous les ans de nombreuses tâches en recherche d'information ou traitement automatique du langage naturel.

Dans le cadre de ce projet de recherche, nous nous intéressons au RAG (Retrieval-Augmented Generation) dont le but est de répondre à des questions en langage naturel en s'appuyant sur les capacités d'un moteur de recherche qui accède à des bases de données documentaires. Lorsqu'une question est soumise à un système de dialogue, les modèles d'IA fonctionnent en deux étapes :

- L'étape de recherche (retrieval) : une base de données est interrogée pour identifier les documents en lien avec la question
- L'étape de génération : les documents et la requête sont fournis en entrée d'un modèle de langue (type ChatGPT) pour générer une réponse en langage naturel.

Les modèles de langue calculent alors des probabilités de mots pour générer la réponse étant donné le contexte (la question + les documents). Un enjeu supplémentaire consiste à lier les parties de la réponse générée avec les documents supports, on appelle ce processus l'attribution de sources.

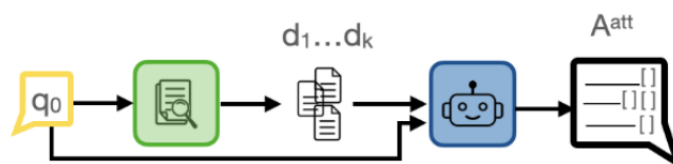


FIGURE 1 – Illustration d'un modèle de type RAG avec attribution des sources

2 Objectifs du projet

L'objectif de ce projet de recherche est multiple :

- Réalisation d'un état de l'art sur le RAG
- Implémenter des modèles existants et éventuellement proposer un modèle innovant
- Plusieurs cas d'usages/cadres expérimentaux pourront être abordés : la campagne d'évaluation TREC RAG, mais aussi un cas d'usage réel avec la Bibliothèque Nationale de France.
- Réaliser une campagne d'évaluation comparant les performances de différents modèles
- Analyser qualitativement les résultats et identifier les échecs des modèles évalués.

1. <https://trec.nist.gov/pubs/call2024.html>