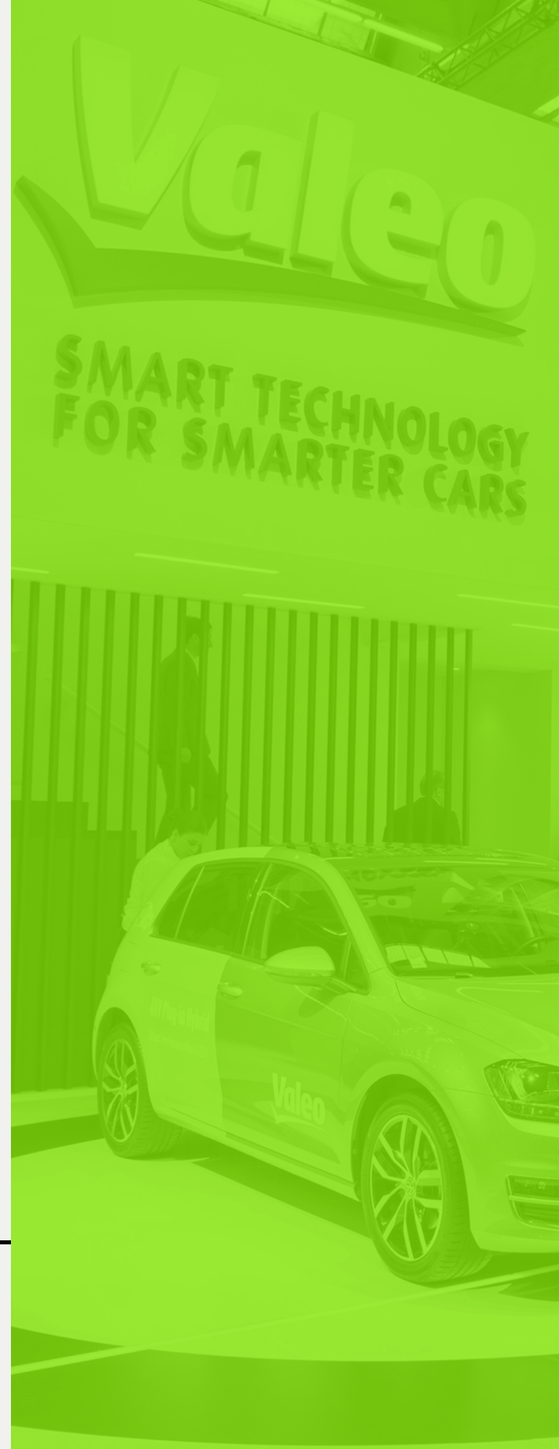
Project : Answering Questions from Documents with AI

"RAG Document Assistant: A Retrieval-Augmented Generation System

Mohamed LAABYDY



Plan :

- Introduction
- Problem Statement
- Solution Overview
- Technical choice
- Challenges and Imporovements

Introduction

The goal of this project is to build a system that answers questions based on document content using retrieval and text generation models.

Why It Matters: Automates extracting insights from large documents.

Goal of the Project and utility

Problem Statement

Extracting specific information from large, unstructured texts.

Need for models that can handle multilingual documents.

Example :

How can we answer the question, “Combien de demandes de brevets Valeo a-t-il fait en 2022 ?”, from a lengthy text (an excerpt from an article in Le Figaro)

Solution Overview

RAG combines **retrieval** (finding relevant parts of the document) with **generation** (producing human-readable answers).

Keys Components for the solution :

- Document preprocessing (chunking).
- Embedding and retrieval with FAISS.
- Generation with Flan T5.

Input Document → **Retrieval** → **Generation** → **Output Answer.**

Technical choice

1. Models and Frameworks

- Flan T5:

⟶ Chosen for its multilingual and contextual understanding capabilities.

⟶ Lightweight and effective for Q&A tasks on structured and semi-structured texts.

- GPT-2:

⟶ Tested for comparison but less effective due to its unidirectional nature and limited contextual understanding.

Technical choice

2. Retrieval Mechanism

- **FAISS (Facebook AI Similarity Search):**

→ Used for efficient similarity-based document retrieval.

→ Handles large-scale document embeddings effectively.

3. Preprocessing

- **LangChain:**

→ Simplifies document chunking and integration with retrieval pipelines.

→ Handles large documents by splitting them into smaller, contextually consistent chunks.

Technical choice

4. Embedding Model

- **sentence-transformers/all-MiniLM-L6-v2:**

→ Compact and efficient for embedding textual data.

→ Balances speed and accuracy in finding relevant information.

Why These Choices?

- Strikes a balance between performance and resource requirements.
- Readily available open-source tools with active community support.
- Prioritized scalability for future enhancements (e.g., more complex models, larger datasets).

Testing and Results

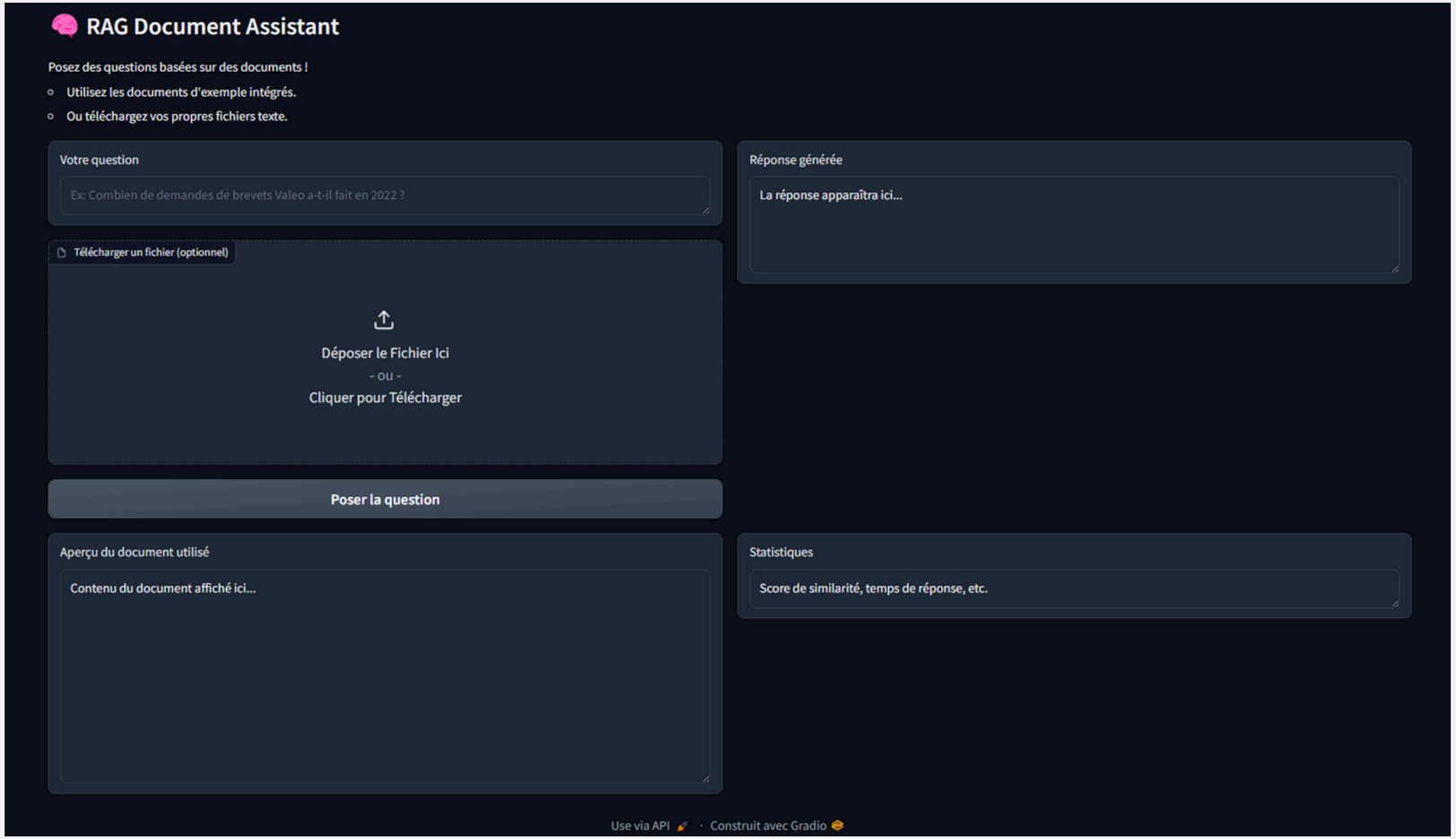
Testing Scenarios:

- Document: Excerpt from Le Figaro article.
- Questions posed: Examples like “Combien de demandes de brevets Valeo a-t-il fait en 2022 ?”

Results:

- Flan T5 performed well for comprehension tasks.
  ⟶ 588

Testing and Results



This interface, built with **Gradio**, allows users to ask questions based on documents. Users can either rely on default example texts or upload their own .txt files for analysis.

Challenges and Imporovements

Challenges :

- Handling multilingual documents.
- Limitations of lightweight models.

Improvements

- Experimenting with robust models like Mistral or Falcon.
- Extending support to more complex documents (legal, technical).
- Optimizing performance for multilingual use cases.
- Scaling for larger datasets.
- Improve response time by investigating alternative models.

Conclusion

Built a functional RAG system for document-based Q&A.

Demonstrated effectiveness with Flan T5 and interactive UI.

Potential for automation in legal, educational, and technical domains.