People’s Democratic Republic of Algeria Ministry of Higher Education and
Scientific Research National Polytechnic School of Algiers

National Polytechnic School of Algiers
Department of Industrial Engineering: Data Science & AI

Reliable, fully local RAG agents with LLaMA3

Created by:

- GHRIBI Ouassim Abdelmalek

Supervisors:

- Mr. ARKI Oussama - *National Polytechnic School of Algiers*
- Mr. BETROUNI Hachem - *BIGmama Technology*

ENP 2024
10, Avenue des Fréres Oudek, Hassen Badi, BP. 182, 16200 El Harrach, Alger, Algérie.

Overview

- Introduction
- Background and motivation
- RAG Systems
- Implemented Solution
- Conclusion

Introduction

We are in a transformative era driven by generative AI and large language models (LLMs). Technologies like GPT-4 are revolutionizing various fields, enabling machines to generate human-like content and driving unprecedented efficiency and innovation. This era is reshaping our interaction with technology and addressing significant global challenges.

Background & Motivation

BIGmama Technology

Big Mama is a French-registered startup with an extension in Algeria, specializing in data science & artificial intelligence. With nearly a decade of experience in the field, BIGmama has collaborated with clients worldwide (ooredoo, total energies, manutan...).

Mission

BIGmama’s mission is to democratize the access to AI.
Put people at the heart of technology, i.e. the hybridization of AI.
Make technology a common good, shareable and to which the maximum number of people could participate and contribute.

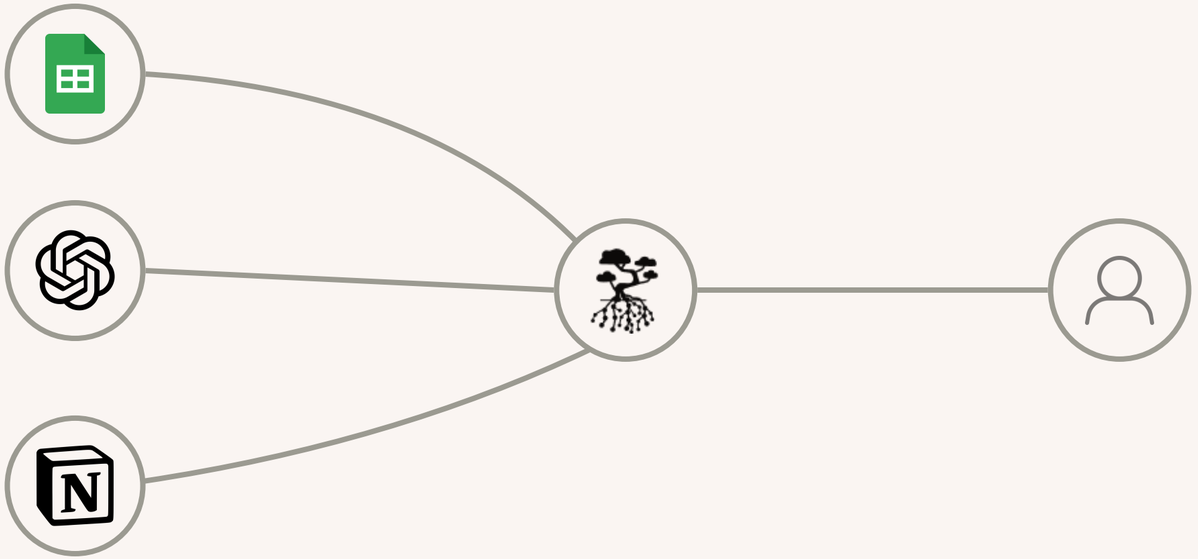
Vision

“Build upon the shoulders of giants”

Artificial intelligence, with tools like ChatGPT, will revolutionize the world like never before. Major companies are mobilizing vast resources to reshape our world. The aim is not to compete with these giants, but to use them as a springboard for technological projects.

Hyko

Hyko is a web app that turns you into an AI app builder. Think of it as using "Lego" bricks from its gigantic toolbox, filled with AI models (>130k), APIs (Stability.ai, Openai, Cohere ..), web scraping tools, utility functions, and more...



Motivation

After numerous meetings with clients and potential investors, we observed a common pattern: the frequent involvement of large language models (LLMs). However, each use case required specific knowledge unique to each client. This insight led us to develop a more effective solution—a node that incorporates the unique knowledge base of each user as input through retrieval-augmented generation (RAG) systems.

Large Language Models (LLMs):

Large Language Models (LLMs) like GPT-4 are designed to understand and generate human-like text using vast amounts of data. They excel in tasks such as translation, summarization, and conversational responses, transforming fields like customer service and content creation with their intelligent, context-aware interactions. Their versatility and ability to produce coherent and relevant text make them invaluable in today's technological landscape.

LLMs limitations

- **Outdated Information:** LLMs rely on training data and may not have the latest information.
- **Hallucinations:** LLMs can generate text that appears correct but is factually inaccurate.
- **Lack of Domain-Specific Accuracy:** LLMs often provide generic responses lacking specific details needed for particular domains.
- **Source Citations:** Difficulty in identifying and citing the sources of information generated by LLMs.
- **Long Update Times:** Updating LLMs with new information requires significant time and resources.

Retrieval-Augmented Generation (RAG) Systems:

Retrieval-Augmented Generation (RAG) systems enhance LLMs by integrating them with retrieval mechanisms that pull in specific, relevant information from knowledge bases or external documents. This approach ensures that responses are accurate and context-specific, addressing the limitations of LLMs in accessing specialized or up-to-date information. RAG systems are particularly useful in applications requiring detailed domain knowledge, such as technical support and personalized recommendations.

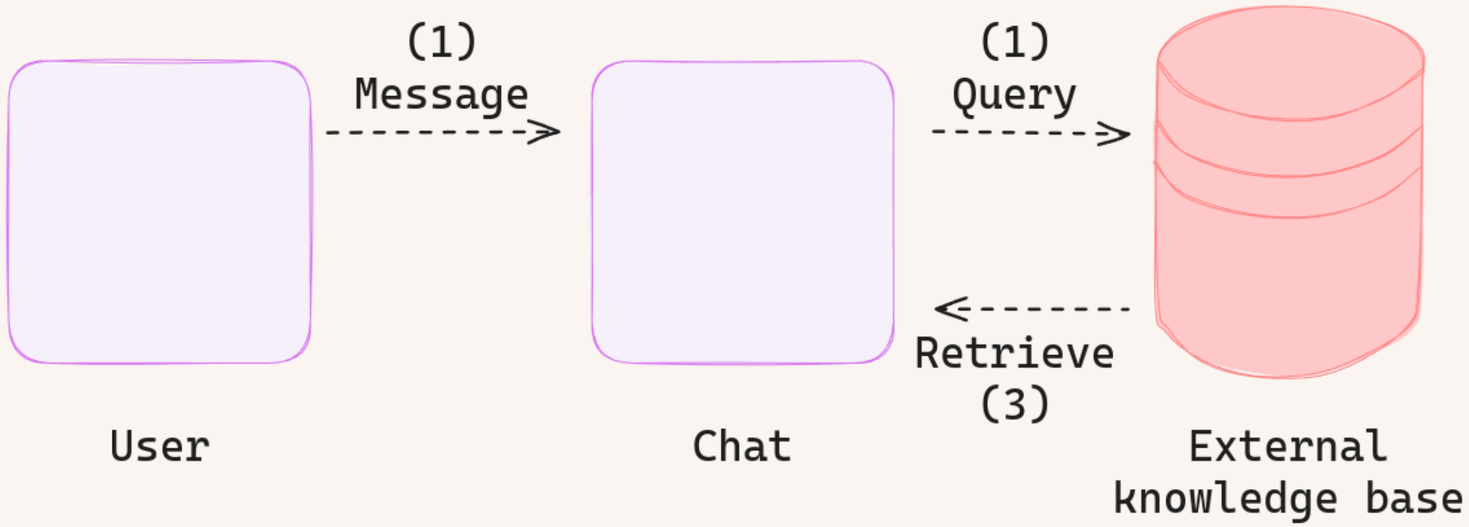
Retrieval-Augmented Generation (RAG) Systems:

RAG systems work in two steps:

- **Retrieval:** The system digs through your data to find useful pieces of information.
- **Generation:** A generative AI model uses the retrieved information to create clear and accurate answers to your questions.

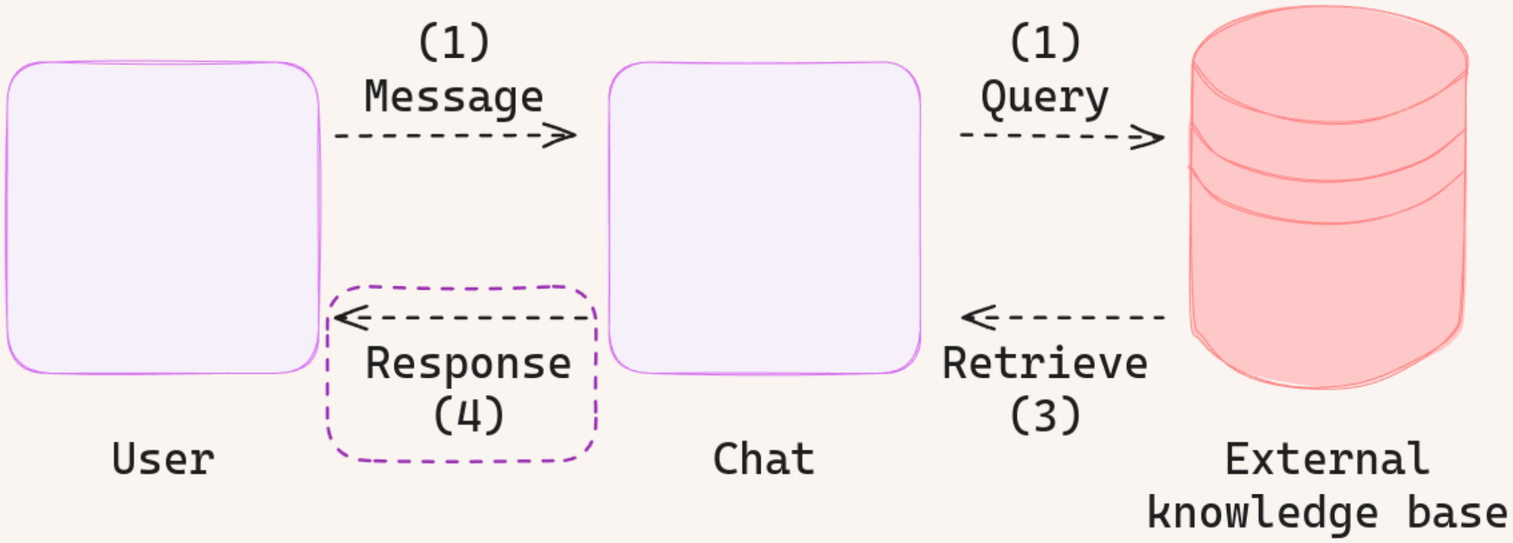
Retrieval

The first part of RAG is to retrieve the right information needed to respond to a user query. Given a user message (1), the Chat endpoint queries an external knowledge base with the relevant queries (2), and finally retrieves the query results (3).



Generation

The second part of RAG is augmented generation. Here, the prompt is augmented with the information retrieved from the retrieval step. The prompt is now grounded with the best information to provide the user with an accurate and helpful response.



RAG vs. Fine-tuning

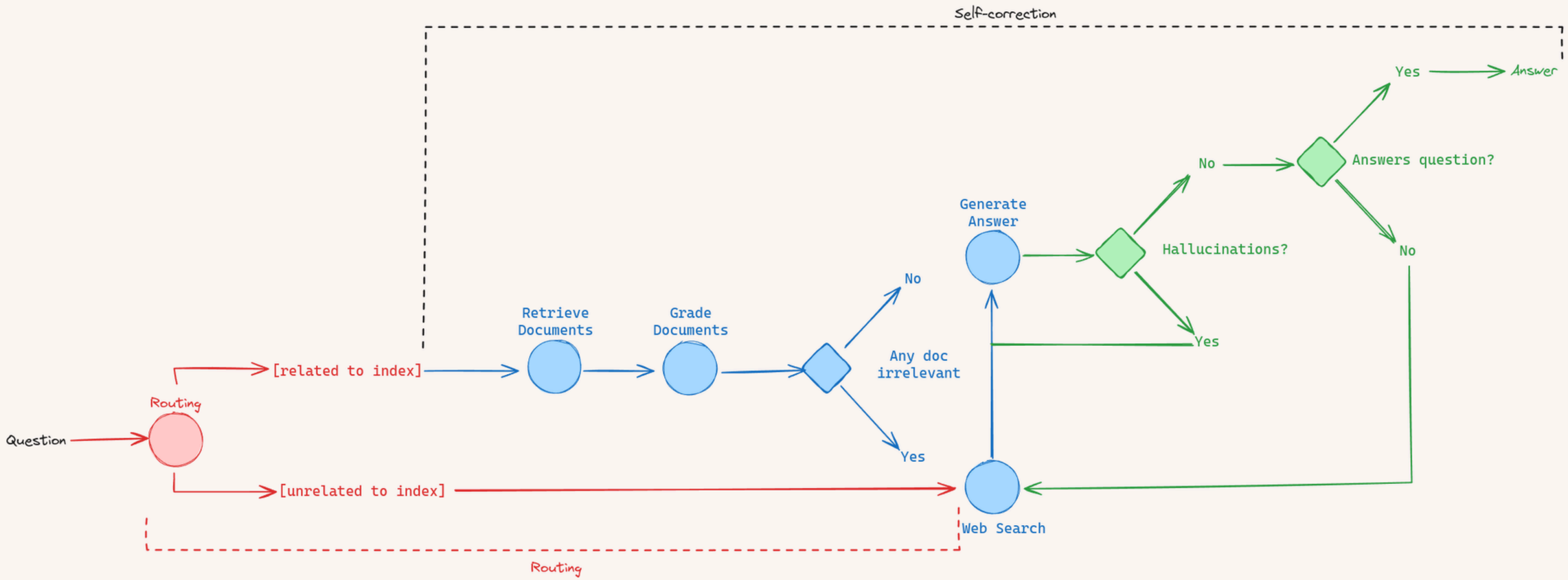
There is ongoing debate about when to use RAG versus fine-tuning. RAG integrates new knowledge, while fine-tuning enhances model performance and efficiency. These methods are complementary and can be combined to improve LLMs for complex, knowledge-intensive applications.

Solution Overview

In our solution we were able to integrate ideas from different RAG papers into a RAG agent:

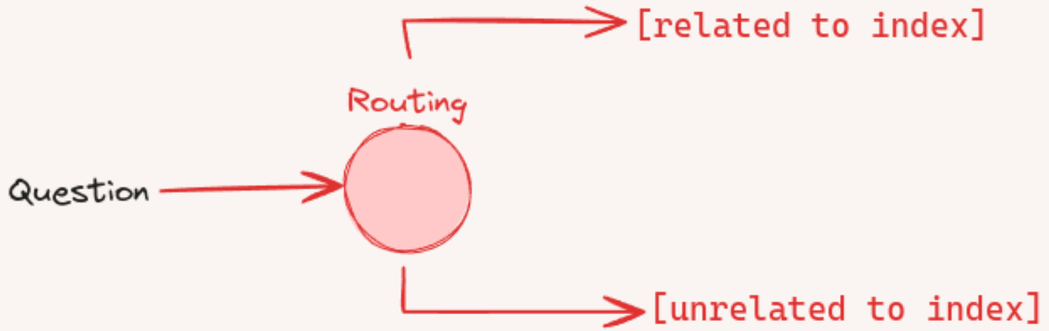
- **Routing:** Route questions to different retrieval approaches.
- **Fallback:** Fallback to web search if docs are not relevant to query.
- **Self-correction:** Fix answers w/ hallucinations or don’t address question

High-Level Architecture



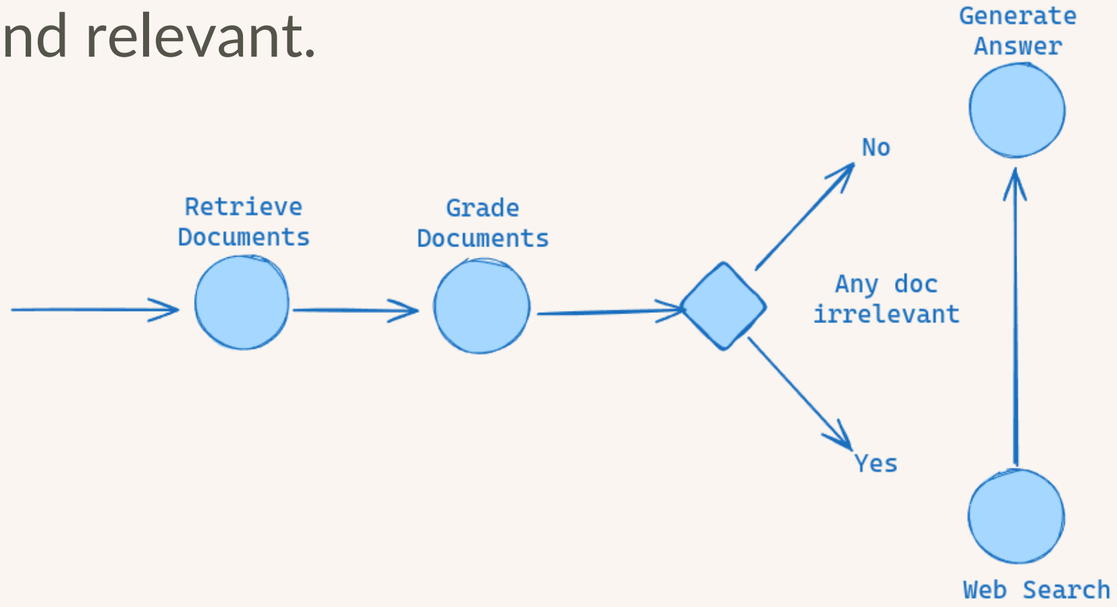
Question Routing

The workflow starts with a user submitting a query, the system then evaluates the query's relevance to indexed documents. If relevant, the query is routed to the local retrieval system; if not, it is redirected to a fallback mechanism using web search. This ensures users receive comprehensive answers, regardless of whether the local index contains the necessary information.



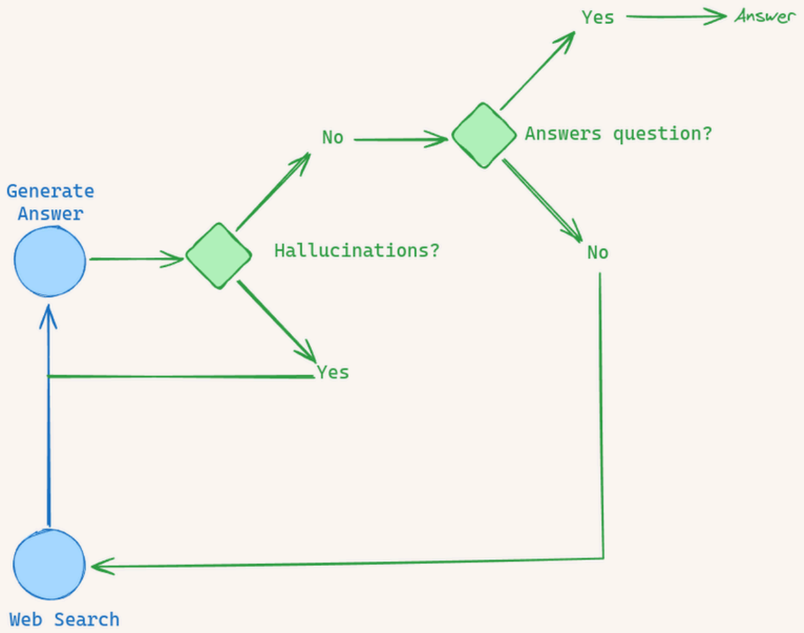
Document Retrieval and Grading

For queries routed to the local retrieval system, ChromaDB fetches relevant documents using vector search techniques. The LLaMA3 model then grades these documents for relevance and quality. If any document is deemed irrelevant, it is flagged, and the system switches to a fallback web search to ensure the information provided is accurate and relevant.



Answer Validation and Correction

Generated answers are initially validated for relevance and accuracy, flagging and correcting issues like hallucinations. If needed, a fallback mechanism performs a web search to refine answers, ensuring final accuracy and system reliability.



Conclusion

In this project, we developed and deployed a Local Retrieval-Augmented Generation (RAG) system, integrating advanced technologies like LLaMA3, FastAPI, Langchain, and Docker. Our goal was to create a versatile platform for precise document retrieval and accurate answer generation. Despite challenges, our modular design and strategic use of LLaMA3's prompting techniques ensured efficient implementation and improved system performance. By iterating on validation and refinement processes, we enhanced response accuracy and reliability. This project exemplifies an integrated approach to leveraging technology for effective information retrieval and generation, ensuring scalable and reliable performance over time.