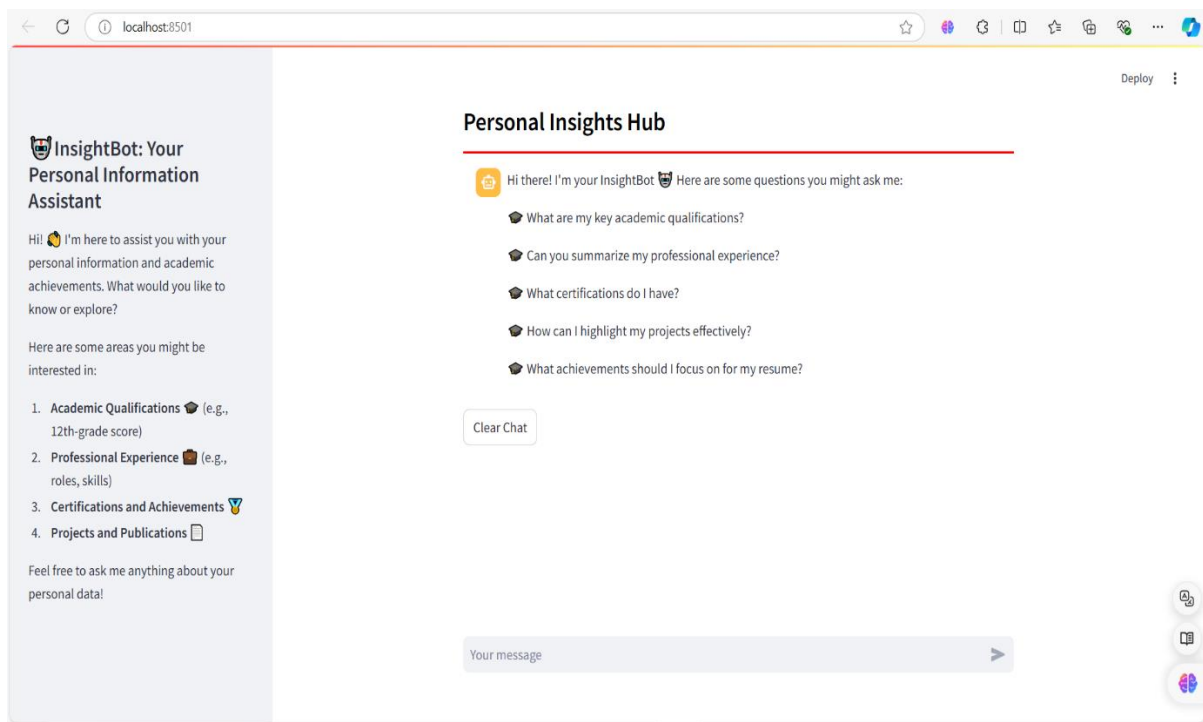


# RAG AI Application (InsightBot)

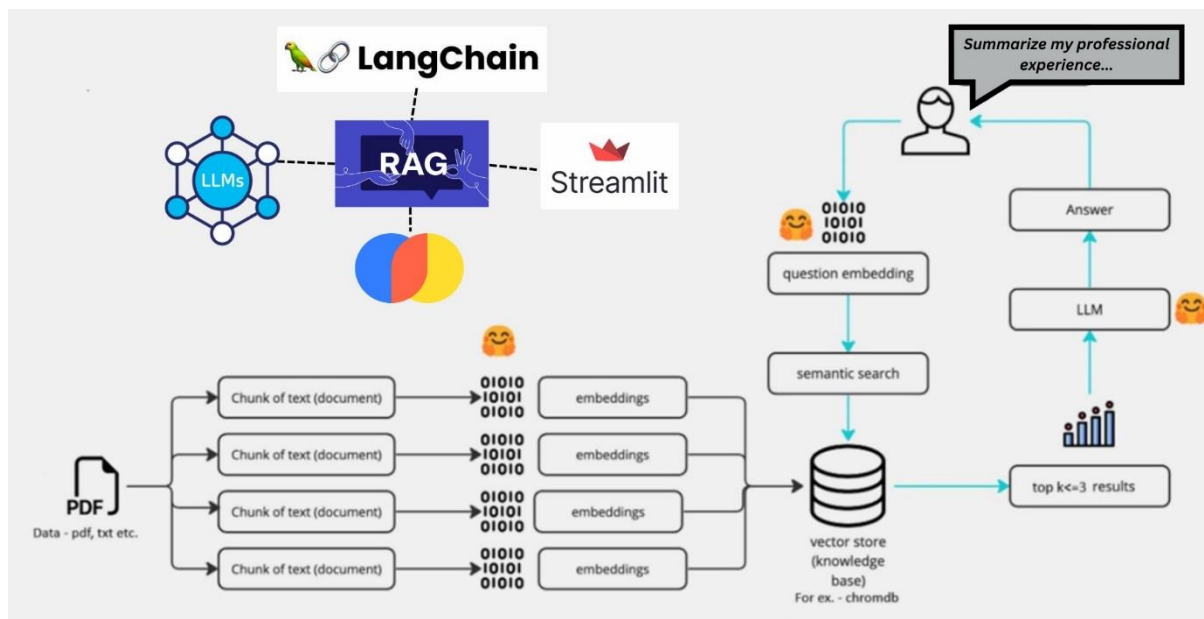
– Chaitanya Madhav

## Introduction

InsightBot is an adaptive RAG system that provides personalized information and insights about your academic and professional achievements through a sophisticated chatbot interface. This system integrates advanced technologies such as LangChain, Hugging Face's Transformers, Retrieval-Augmented Generation (RAG), VectorDB implemented via Chroma, and Streamlit. These technologies collectively enable the system to retrieve and analyze personal data, offering customized responses about your qualifications, experience, and accomplishments efficiently.



# InsightBot Architecture



## Data Ingestion Process in InsightBot

The data ingestion process involves several key steps that ensure raw text data is transformed into a searchable and retrievable format using embeddings and a vector store.

### 1. Environment Setup

The system begins by loading necessary environment variables using `dotenv` to retrieve sensitive information such as API tokens. This ensures secure access to external services like Hugging Face.

### 2. Text Loading

A plain text file/pdf is loaded into the system, containing the raw data that will be processed. This file forms the basis of the content that InsightBot will later retrieve and respond to user queries with.

### 3. Text Splitting

Using LangChain's `RecursiveCharacterTextSplitter`, the loaded text is split into manageable chunks. Each chunk has a maximum size of 2000 characters, with a 300-character overlap to maintain continuity. This splitting is crucial for ensuring that smaller sections of the text can be processed and retrieved efficiently.

### 4. Document Creation

The text chunks are converted into Document objects, which serve as containers for the content. These objects are designed to work seamlessly with LangChain's pipeline, making it easier to manage and retrieve specific pieces of data later on.

### 5. Embedding Generation

The system initializes a pre-trained embedding model (sentence-transformers/all-MiniLM-L6-v2) from Hugging Face. It then uses this model to generate vector embeddings for each document chunk. These embeddings represent the semantic meaning of each chunk, allowing for efficient and accurate similarity-based searches.

## **6. Vector Store Initialization**

A Chroma vector store is initialized to store the generated embeddings. This vector database is persistent, meaning that the stored embeddings can be used for future queries without having to reprocess the raw text every time.

## **7. Document Ingestion into Vector Store**

The processed document chunks (now represented as embeddings) are added to the Chroma vector store. This step ensures that the documents are indexed and can be retrieved based on user queries.

## **8. Validation and Testing**

To validate the setup, a test query is executed against the vector store using a similarity search. The results are deduplicated using an OrderedDict to ensure only unique responses are considered. The system returns the top 3 most relevant results, confirming that the data has been ingested correctly and is ready for retrieval.

# **Integration and Implementation of RAG in InsightBot**

The integration of Retrieval-Augmented Generation (RAG) in InsightBot combines a retrieval mechanism with a generative language model to deliver highly relevant, context-specific responses. This hybrid approach enhances the system's ability to address user queries related to personal data—such as academic qualifications and professional achievements—by efficiently retrieving information and generating coherent, precise answers. The following is a detailed explanation of the integration process:

## **Data Loading and Vectorization**

**Data Extraction:** LangChain's PyPDFLoader or textfileloader is utilized to extract text from personal documents (e.g., resumes, scorecards), forming the foundation of a searchable database.

**Text Chunking:** Using LangChain's RecursiveCharacterTextSplitter, the extracted data is divided into manageable chunks, ensuring that meaningful context is preserved for accurate retrieval.

## **Embedding Generation**

**Embedding Model:** The HuggingFaceEmbeddings class, with the model sentence-transformers/all-MiniLM-L6-v2, generates vector embeddings of the text chunks. These embeddings capture the semantic meaning of the documents, enabling efficient matching of user queries to relevant data.

## **Vector Database (Chroma)**

**VectorDB Integration:** Chroma is used to store the embeddings, allowing fast and scalable similarity searches. The system persistently saves these embeddings, enabling swift retrieval of relevant information when needed.

## **Semantic Similarity Search**

**Retriever Mechanism:** Upon receiving a query, the system conducts a semantic similarity search using the Chroma vector store. The method `vector_store.as_retriever(top_k=3)` ensures that the top three most relevant text chunks are selected, enhancing response accuracy.

### Language Model for Response Generation

**Generative Model (LLM):** The retrieved information is processed by HuggingFaceHub's language model (meta-llama/Meta-Llama-3-8B-Instruct), which generates human-like responses based on the context and user query. The model is optimized for coherence and relevance through specific settings (e.g., temperature, token limits).

### Custom Prompt Template

**Custom Prompting:** A tailored PromptTemplate guides the language model, focusing on specific topics such as academic achievements and professional experience. This ensures that the generated responses are accurate, avoiding irrelevant information.

### Streamlit Interface

**Interactive User Interface:** The Streamlit-powered interface provides a user-friendly experience, allowing users to query personal data and receive real-time responses. Features like chat history maintenance and reset options enhance user interaction.

## Evaluation and Testing of InsightBot

To ensure that **InsightBot** operates effectively and provides accurate responses to personal information queries, a comprehensive evaluation and testing phase was conducted. This process involved submitting a range of queries, from basic to complex, to test the system's ability to accurately retrieve and generate responses related to academic qualifications, professional experience, and other personal data.

**Note:** I have ingested my 12<sup>th</sup> Std details and resume in .txt/pdf file to the system.

### Testing Methodology

InsightBot was subjected to a series of tests covering various query types:

#### 1. Simple Queries

These straightforward queries assess the bot's ability to retrieve direct information from the database. Examples include:

- "What are my key academic qualifications?"



What are my academic qualifications?



You have completed your 12th standard from Mahesh PU College with a total score of 555/600 and a distinction. You have also completed your B.Tech in Computer Science from PES University with a CGPA of 7.43/10.

Please refine your question if you need more specific information.

o "Can you summarize my professional experience?"



Can you summarize my professional experience?"



Based on the provided information, I can summarize your professional experience as follows:

You have worked as a Graduate Engineering Trainee at Microland Ltd since June 2024, where you are undergoing comprehensive training in various technologies. Prior to this, you worked as an AI Technical Intern at Qualitas Technologies Pvt Ltd from January 2024 to June 2024. During this period, you developed image processing pipelines, automated visual inspection processes, and innovated AI email automation and case study automation using "make.com". You also led research on next-generation AI technologies like Microsoft Autogen and Crew AI.

Please note that this summary is based on the information provided and may not be exhaustive. If you would like to add or clarify any details, please feel free to do so.

## 2. Complex/Nuanced Queries

These queries test the system's ability to handle detailed inquiries and generate contextually relevant responses. Examples include:

o "How can I highlight my research publications and certifications for my resume?"



How can I highlight my research publications and certifications for my resume?



To highlight your research publications and certifications on your resume, you can create a separate section for "Publications" and "Certifications" under the "Education" or "Skills" section. Here's an example:

**Publications:**

- Chaitanya Madhav, et al. "Real-Time Word-Level Recognition of Indian Sign Language." ICICV-2024 Conference, Jaipur, 2024.
- [Insert other research publications]

**Certifications:**

- AWS Educate - Introduction to Cloud 101
- Make.com - Basics and Foundation Certificate (Low-Code Automation Course)
- ChatGPT and AI Hacks in MS Office – SKILL NATION
- Linux Foundation – Beginner's Guide to Linux Kernel Development

By highlighting your research publications and certifications, you can demonstrate your expertise and commitment to your field, which can be attractive to potential employers. Make sure to tailor your resume to the specific job you're applying for and highlight the most relevant information.

o "What are my key achievements in AI/ML to mention in a cover letter?"



What are my key achievements in AI/ML to mention in a cover letter?"



As a skilled AI/ML enthusiast, your key achievements include:

- Spearheading the development of image processing pipelines from data preprocessing to model deployment, leveraging deep learning techniques to automate visual inspection processes and achieving notable results in image classification and segmentation.
- Innovating AI Email Automation and Case Study Automation using "make.com" for streamlining Sales processes in the company.
- Researching on Next Generation AI Technologies like Microsoft Autogen and Crew AI for solving complex processes in the company.
- Developing a real-time dynamic recognition system for Indian Sign Language using machine learning and deep learning models, addressing key challenges like signer dependency, skin color variations, and high processing power requirements.

These achievements demonstrate your expertise in AI/ML, problem-solving skills, and ability to innovate and adapt to new technologies. Highlighting these achievements in your cover letter will showcase your value as a candidate and increase your chances of standing out in a competitive job market.

### 3. Scenario-Based Queries

Practical, contextual queries require InsightBot to integrate multiple factors and provide personalized advice. Examples include:

- "I'm preparing for a job interview. What aspects of my academic background and professional experience should I emphasize?"



○ I'm preparing for a job interview. What aspects of my academic background and professional experience should I emphasize?



As you prepare for your job interview, it's essential to highlight your academic achievements and professional experience that align with the job requirements. In your case, you can emphasize your:

- Strong academic record, particularly your 2nd PUC scorecard, which showcases your academic excellence.
- Relevant projects, such as the Microservice Communication of Student Database, IoT Cloud-based Color Sorter, Speech Emotion Recognition, and Crime Report Management System, which demonstrate your technical skills and problem-solving abilities.
- Professional experience, including your roles as a Graduate Engineering Trainee and AI Technical Intern, which highlight your ability to work in a team, lead projects, and adapt to new technologies.
- Certifications, such as AWS Educate - Introduction to Cloud 101, Make.com - Basics and Foundation Certificate (Low-Code Automation Course), ChatGPT and AI Hacks in MS Office – SKILL NATION, and Linux Foundation – Beginner's Guide to Linux Kernel Development, which demonstrate your commitment to continuous learning and professional development.

Remember to tailor your responses to the specific job requirements and emphasize your unique strengths and qualifications. Good luck with your interview!

- "How should I present my project experience in AI/ML during a technical assessment?"



○ How should I present my project experience in AI/ML during a technical assessment?



When presenting your project experience in AI/ML during a technical assessment, focus on the following key aspects:

1. **Problem Statement:** Clearly articulate the problem you aimed to solve, and how it relates to the AI/ML domain.
2. **Approach:** Describe the approach you took to tackle the problem, including the algorithms, techniques, and tools used.
3. **Implementation:** Highlight the implementation details, such as the programming languages, frameworks, and libraries employed.
4. **Results:** Share the results, including any metrics, visualizations, or insights gained from the project.
5. **Lessons Learned:** Reflect on the challenges faced, lessons learned, and how you would improve the project if given another opportunity.

Remember to be concise, clear, and specific in your presentation, and be prepared to answer follow-up questions from the assessors. Practice your presentation to ensure you can effectively communicate your project experience and achievements.

#### 4. Irrelevant/Out-of-Scope Queries

These queries ensure that the bot can appropriately redirect or respond to inquiries unrelated to its domain. Examples include:

- o "Do you know the weather forecast for today?"



o Do you know the weather forecast for today?



I'm happy to help you with your query! However, I don't have information about the weather forecast for today as it's not related to my stored personal data. Could you please refine your question or ask something related to my academic achievements or professional experience? I'm here to assist you with any questions you may have!

#### 5. Misspelled Queries

These tests evaluate the system's robustness in handling queries with incorrect spelling, ensuring it still provides accurate responses. Examples include:

- o "Whatt are my acadmic qualfications?"



Whatt are my acadmic qualfications?"



You have completed your 12th standard from Mahesh PU College with a total score of 555/600 and a distinction. You have also completed your B.Tech in Computer Science from PES University with a CGPA of 7.43/10.

- o "Proffesional experince highlights?"



Proffesional experince highlights?



As a Graduate Engineering Trainee at Microland Ltd, I am undergoing comprehensive training in various technologies including Networking, Cyber-Security, Cloud, Linux, and Windows Server. Prior to this, I worked as an AI Technical Intern at Qualitas Technologies Pvt Ltd, where I spearheaded the development of image processing pipelines, leveraging deep learning techniques to automate visual inspection processes. I also innovated AI Email Automation and Case Study Automation using "make.com" for streamlining Sales processes in the company. Additionally, I led research on Next Generation AI Technologies like Microsoft Autogen and Crew AI for solving complex processes in the company.

By conducting these tests, InsightBot's performance was validated for its accuracy, contextual relevance, and resilience in handling a variety of user interactions. This ensures the bot delivers reliable, personalized responses to all valid queries.