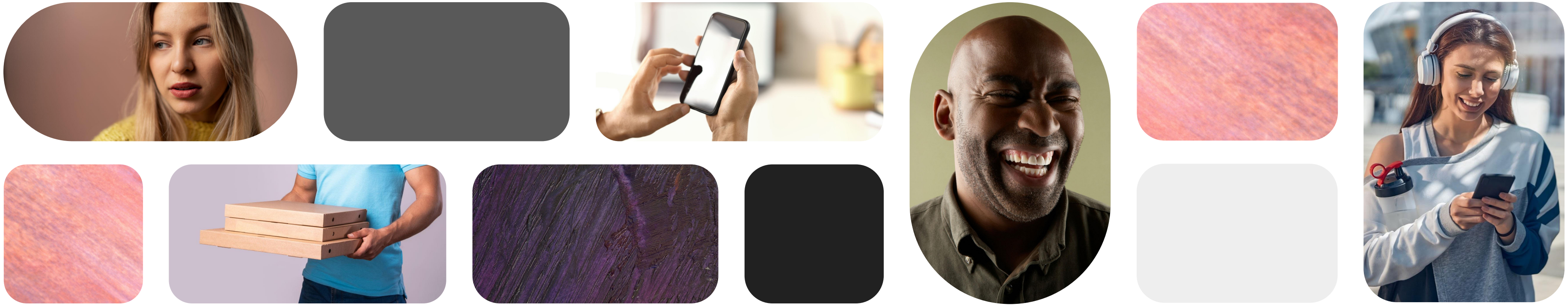
Chat With Your Data

When Your Files Start Talking Back



About me

7+ Years in NLP, ML, AI

- Machine Translation & Grammar Correction
- LLM-based Chatbots & Voicebots
- RAG Pipelines & Semantic Search
- Retrieval, Vector Databases & Multi-Agent Systems

Maryna Bogdan
AI/ML Engineer

Contents

1. Introduction

2. How AI can help

3. AI Limitations

4. Your files, your AI

5. What is RAG?

6. Conclusions

7. Demo

What is the biggest problem we have nowadays that AI can help us with?

TIME

Can AI help us save time?

Can AI help us save time?

Or help us not to waste time?

Can AI help us save time?

Or help us not to waste time?

We see some real cases

✅ KPMG reports **60–80% time savings** in marketing content creation by using AI writing tools with retrieval capabilities (Writer.com/KPMG case study, 2023).

We see some real cases

✅ KPMG reports **60–80% time savings** in marketing content creation by using AI writing tools with retrieval capabilities (Writer.com/KPMG case study, 2023).

✅ A Slack-integrated RAG assistant cut **document search time by 75%** for employees, enabling instant answers from internal knowledge bases (SoftBlues AI Solutions, 2023).

We see some real cases

✅ KPMG reports **60–80% time savings** in marketing content creation by using AI writing tools with retrieval capabilities (Writer.com/KPMG case study, 2023).

✅ A Slack-integrated RAG assistant cut **document search time by 75%** for employees, enabling instant answers from internal knowledge bases (SoftBlues AI Solutions, 2023).

✅ Microsoft's Assembly Software **NeosAI** reduced drafting time for legal documents from **≈40 hours to minutes**, saving **~25 hours per case** (Microsoft Customer Story, 2025).

But there are other problems…

⚠️ No out of the box AI solution.

⚠️ Hallucinations: AI confidently invents false answers.

⚠️ Outdated info: static models don’t know what happened after they were trained.

⚠️Training data biases.

⚠️ Transparency: internal work not fully explainable, “black box”.

⚠️ Risks of misuse or harmful output.

Your files, your AI

💭 Imagine if you could **chat with your own documents** and get accurate answers.

Your files, your AI

💭 Imagine if you could **chat with your own documents** and get accurate answers.

🧠 If you had an **intelligent helper** who can:

- Explain and analyze your private data
- Summarize key points
- Draw conclusions tailored to your needs

Your files, your AI

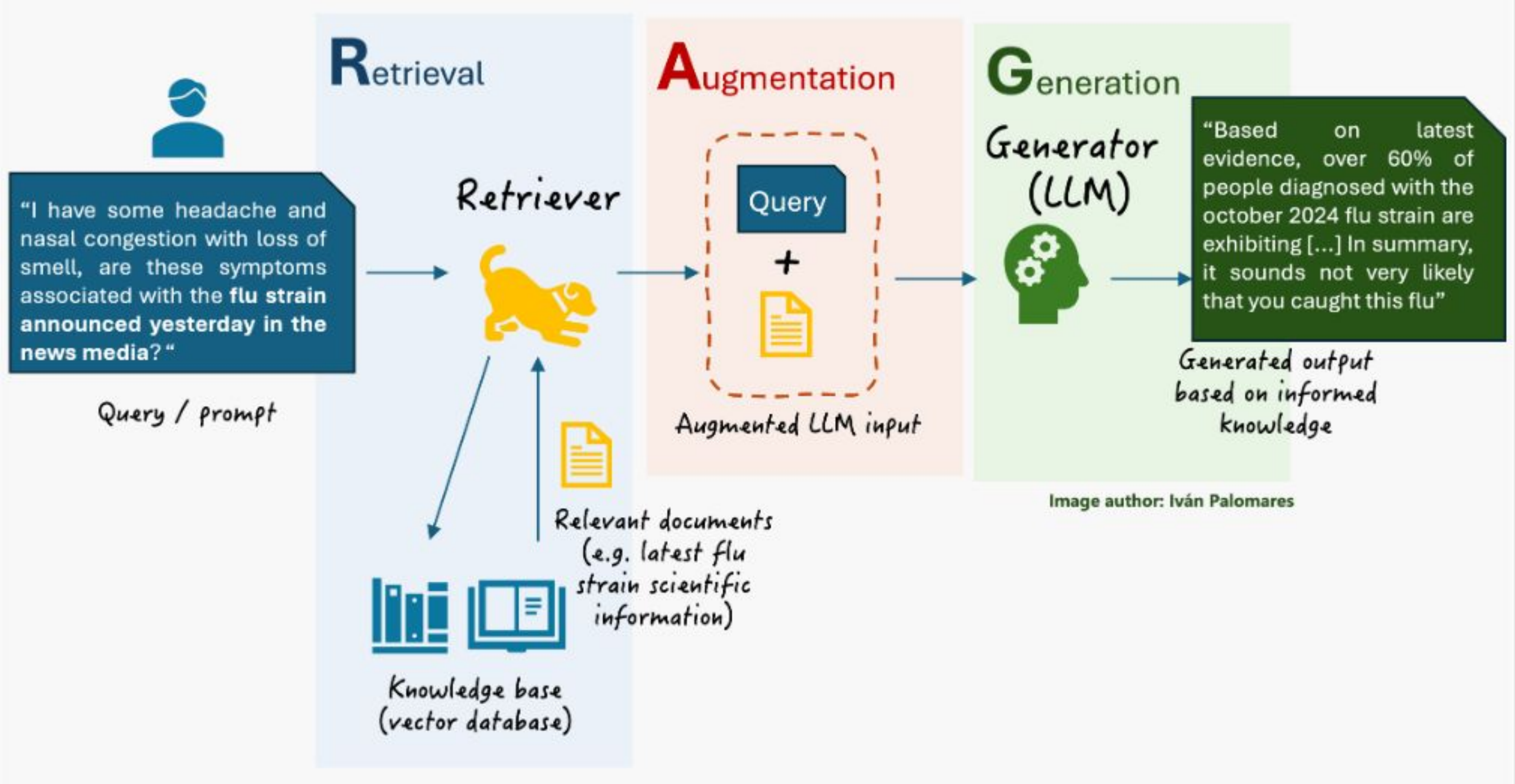
💭 Imagine if you could **chat with your own documents** and get accurate answers.

🧠 If you had an **intelligent helper** who can:

- Explain and analyze your private data
- Summarize key points
- Draw conclusions tailored to your needs

🚀 **This is now possible—thanks to RAG and retrieval techniques.**

What is RAG?



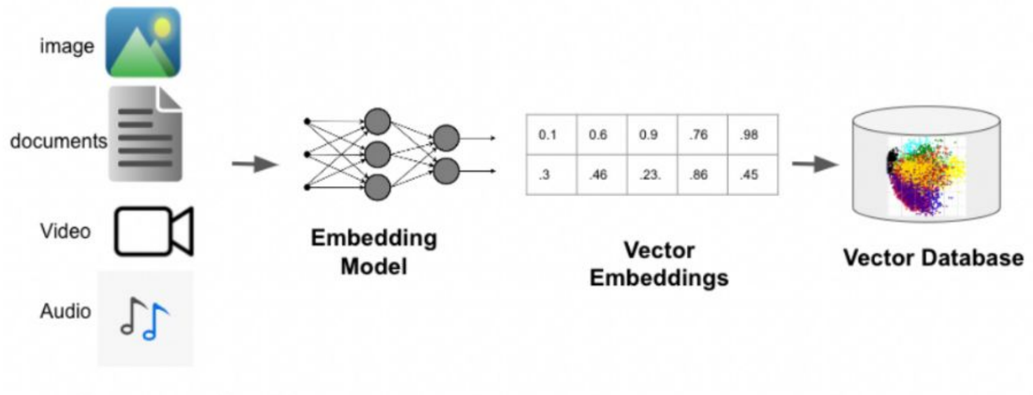
Retrieval
Augmentation
Generation
"I have some headache and nasal congestion with loss of smell, are these symptoms associated with the **flu strain announced yesterday in the news media?**"
Query / prompt
Retriever
Knowledge base
(vector database)
Relevant documents
(e.g. latest flu
strain scientific
information)
Query
+
Augmented LLM input
Generator
(LLM)
"Based on latest evidence, over 60% of people diagnosed with the october 2024 flu strain are exhibiting [...] In summary, it sounds not very likely that you caught this flu"
Generated output
based on informed
knowledge

Image author: Iván Palomares

How does RAG work?

1️⃣**Building the knowledge base:** First, we need to process and store the information that our RAG will use. For this:

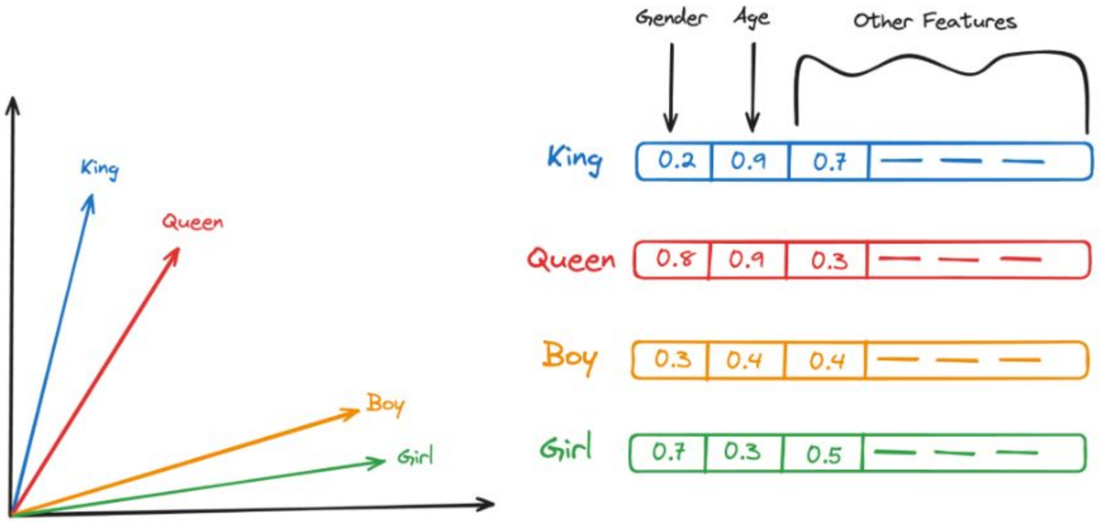
- Collect documents in plain text, PDFs, databases, etc.
- Split them into small fragments (chunks) to improve search.
- Convert them into numerical vectors using an embedding model (e.g., OpenAI *text-embedding-3-small, text-embedding-3-large*).
- Store them in a vector database such as FAISS, Chroma, or Pinecone.



0.1	0.6	0.9	.76	.98
.3	.46	.23.	.86	.45

What is embedding?

An **embedding** is a mathematical representation of text in a multidimensional space. It converts words or phrases into vectors that capture their meaning and semantic relationships. In this space, terms with similar meanings are located closer to each other. Embedding models learn these representations by analyzing the context in which words appear, allowing a better understanding of their relationships.

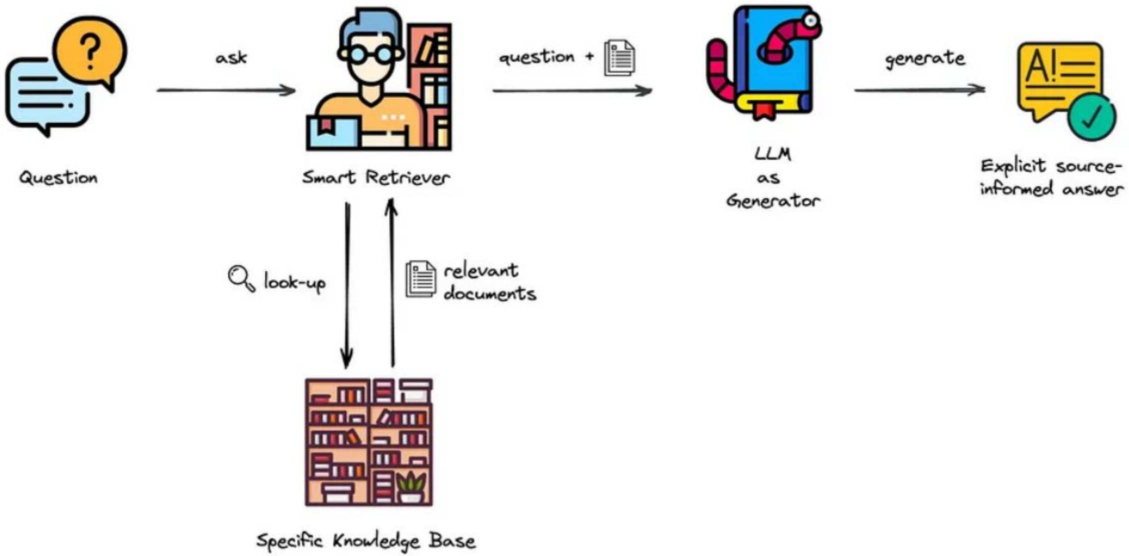


	Gender	Age	Other Features			
King	0.2	0.9	0.7	—	—	—
Queen	0.8	0.9	0.3	—	—	—
Boy	0.3	0.4	0.4	—	—	—
Girl	0.7	0.3	0.5	—	—	—

2️⃣ Efficient Information Retrieval

🔍 When a user asks a question:

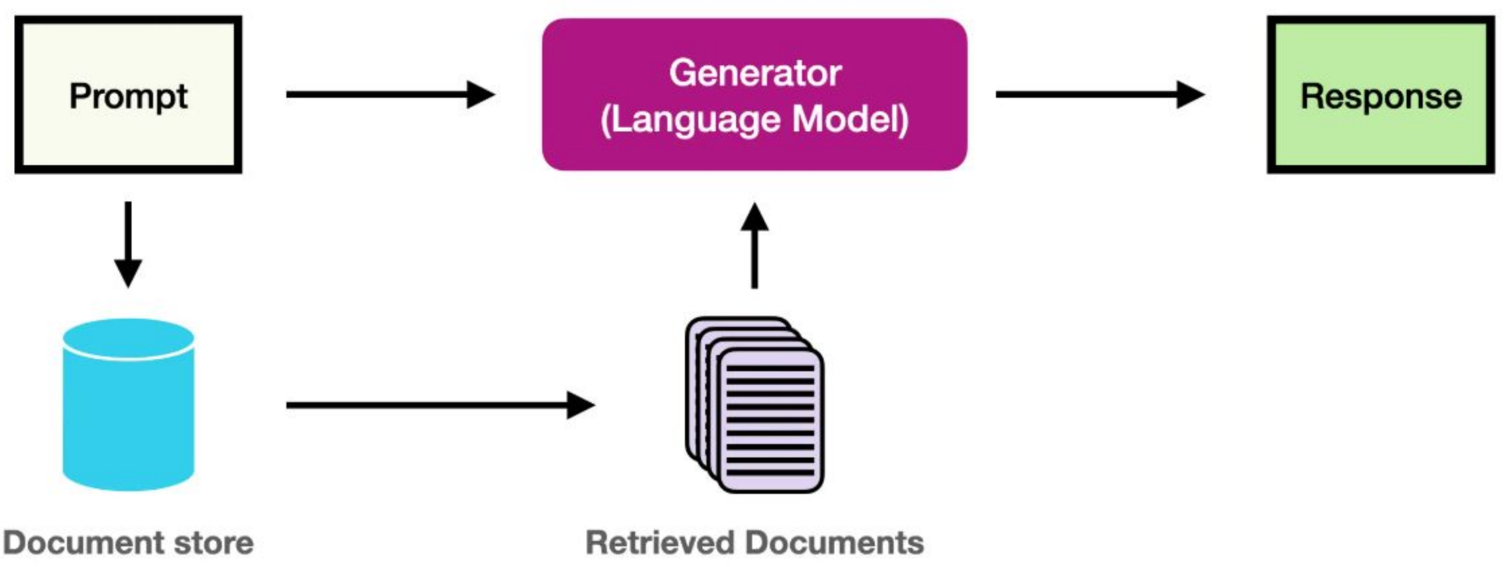
- The question is converted into a numerical vector using the same embedding model.
- The vector database is searched for the most relevant fragments.
- A set of documents containing useful information is retrieved.



3️⃣ Response Generation Using an LLM (Generation)

📝 With the retrieved documents:

- A prompt is built for a language model (e.g., GPT-4, Mixtral, or Llama).
- The question is provided along with the retrieved context.
- The model generates a response based on the supplied information.



Conclusions

🎯 **Benefits of RAG**

✅ **More accurate answers:** Avoids the “hallucination” problem of generative models.

✅ **Up-to-date information:** You can add new data without retraining a model from scratch.

✅ **Lower costs:** You don’t need a huge model with all the knowledge—just an efficient database.

A RAG application combines the best of two worlds:

- **Efficient retrieval** → Finds the best information fragments.
- **Text generation** → Uses an LLM to create natural, well-structured responses.

It is a fundamental technique for intelligent chatbots, recommendation systems, and virtual assistants.

Demo

Demo app: https://demo-rag-chat.streamlit.app/

Public repository: https://github.com/maryna-b/rag_chat_app

Demo questions:
https://www.notion.so/instaboost-ai/Demo-Questions-271dea8a61a780ebb90af406f410d106

Thank you!