Practical RAG: Building Specialized Chatbots Step-by-Step

Hocine Abdellatif Houari, @hahouari

LLMs?

LLMs (or Large Language Models) are artificial intelligence models that are trained on large amounts of text data to generate human-like text.

LLMs?

They are considered as general purpose AIs, unlike specialized AI models that are designed to perform specific tasks (e.g. Face Recognition, OCR, etc.).

Quiz 1

What is the primary purpose of Large Language Models (LLMs)?

- a) To perform face recognition tasks
- b) To generate human-like text based on large amounts of data
- c) To scrape data from websites
- d) To analyze and process audio data

Quiz 1 (Answers)

What is the primary purpose of Large Language Models (LLMs)?

- ❌ To perform face recognition tasks
- ✅ To generate human-like text based on large amounts of data
- ❌ To scrape data from websites
- ❌ To analyze and process audio data

LLMs and Their Limitations

While LLMs are powerful, they have limitations:

- **Static Knowledge**: They are only trained on data available up to a certain point.
- **Large Size**: They require significant resources for fine-tuning or training.
- **Context Limitation**: They struggle to retrieve specific information efficiently (or hallucinate).

Bridging the Gap: Why RAG?

To address these limitations, **Retrieval-Augmented Generation (RAG)** combines 2 steps:

1. **Retrieval Systems**: For fetching up-to-date, task-specific, or large-scale information on demand.
2. **LLMs**: For generating coherent and contextually relevant text.

This synergy enhances the effectiveness of AI systems in dynamic and specialized use cases.

Why not Fine-Tuning?

Fine-tuning is a process of taking a pre-trained model to train it and tweak its parameters to perform better on a specific task.

Pros & Cons over Fine-Tuning

RAG			Fine-Tuning
Up-to-Date	✅		Can still be outdated
No Training	✅		Training required
Easy to Switch Model	✅		Hard to Switch
Retrieval Quality	👌	👌	Training Quality
Extra Retrieval Step		✅	Real-time

Quiz 2

Which of the following statements are true?

- a) Fine-tuning can give outdated answer unlike RAG.
- b) Fine-tuning allows easy switching between models, while RAG requires retraining for each new model.
- c) RAG can be updated with new data more easily, while fine-tuning requires retraining the model.
- d) Fine-tuning is faster to implement than RAG since it doesn’t require retrieval systems.

Quiz 2 (Answers)

Which of the following statements are true?

- ✅ Fine-tuning can give outdated answer unlike RAG.
- ❌ Fine-tuning allows easy switching between models, while RAG requires retraining for each new model.
- ✅ RAG can be updated with new data more easily, while fine-tuning requires retraining the model.
- ❌ Fine-tuning is faster to implement than RAG since it doesn't require retrieval systems.

Components of RAG System

- Large Language Model (e.g. OpenAI GPT-3, Anthropic Sonnet, or Google Gemini)
- Structured Data
- Embedding Model for Semantic Search
- Database with vector storage & search capabilities

What is scraping & structuring data?

- Scraping: Extracting useful data from websites, PDFs, or APIs.
- Structuring Data: Converting scraped data into a structured format (e.g. JSON, CSV, XML, Class Objects, etc).

Components of RAG System

- Large Language Model ✅
- Structured Data ✅
- Embedding Model for Semantic Search 🤔❔
- Database with vector storage & search capabilities ⏳

Embedding Models? 🤔

They are specialized ML models that convert data (like text, images, or audio) into vectors (embeddings). These vectors allow us to perform semantic search.

Databases with vector storage & search capabilities

Examples:

- MySQL (9.1 or later), MariaDB (11.7 or later)
- PostgreSQL (using pgVector)
- SurrealDB
- Pinecone
- Milvus

Quiz 3

Which of the following best describes the process of structuring data?

- a) Extracting raw data from websites or APIs
- b) Converting extracted data into a usable format like JSON or CSV
- c) Searching through data using embedding models
- d) Storing data into database

Quiz 3 (Answers)

Which of the following best describes the process of structuring data?

- ❌ Extracting raw data from websites or APIs
- ✅ Converting extracted data into a usable format like JSON or CSV
- ❌ Searching through data using embedding models
- ❌ Storing data into database

Quiz 4

What is the purpose of embedding models in a RAG system?

- a) To generate human-like text
- b) To convert data into vectors for semantic search
- c) To store large datasets efficiently
- d) To perform real-time data retrieval

Quiz 4 (Answers)

What is the purpose of embedding models in a RAG system?

- ❌ To generate human-like text
- ✅ To convert data into vectors for semantic search
- ❌ To store large datasets efficiently
- ❌ To perform real-time data retrieval

Thank You!

Thank you all for your attendance and active participation.
Your interactions and engagement made this session insightful and enjoyable!