16. Building a Chatbot

LING-351 Language Technology and LLMs

Instructor: Hakyung Sung

October 23, 2025

Our Goal Today

- Our goal today is to **build a chatbot** that can give factually accurate answers.

Our Goal Today

- Our goal today is to **build a chatbot** that can give factually accurate answers.
- Instead of relying only on what the model remembers, we want it to search for real information.

Our Goal Today

- Our goal today is to **build a chatbot** that can give factually accurate answers.
- Instead of relying only on what the model remembers, we want it to search for real information.
- To do that, we'll use a special type of model called the RAG model.

What is the RAG Model?

- **RAG** stands for Retrieval-Augmented Generation.

What is the RAG Model?

- **RAG** stands for Retrieval-Augmented Generation.
- It combines two parts:

What is the RAG Model?

- **RAG** stands for Retrieval-Augmented Generation.
- It combines two parts:
    - A *retriever* that looks up relevant documents.

What is the RAG Model?

- **RAG** stands for Retrieval-Augmented Generation.
- It combines two parts:
    - A *retriever* that looks up relevant documents.
    - A *generator* that uses those documents to create the final answer.

What is the RAG Model?

- **RAG** stands for Retrieval-Augmented Generation.
- It combines two parts:
    - A *retriever* that looks up relevant documents.
    - A *generator* that uses those documents to create the final answer.
- So the chatbot doesn’t just “guess” — it finds information and then explains it in natural language.

How RAG Works

1. You ask a question (e.g., “What is the capital of France?”)

Retriever + Generator = RAG Model → *A Smarter Chatbot!*

How RAG Works

1. You ask a question (e.g., “What is the capital of France?”)
2. The **retriever** searches a knowledge base for relevant texts.

Retriever + Generator = RAG Model $\rightarrow$ *A Smarter Chatbot!*

How RAG Works

1. You ask a question (e.g., “What is the capital of France?”)
2. The **retriever** searches a knowledge base for relevant texts.
3. The **generator** reads them and produces a fluent, fact-based answer.

Retriever + Generator = RAG Model → *A Smarter Chatbot!*

In the shared code

Section 1

I’ll walk you through the codes.

Section 2: Build Your Own RAG Chatbot

- The chatbot will answer questions based on a small set of documents that you provide.

Section 2: Build Your Own RAG Chatbot

- The chatbot will answer questions based on a small set of documents that you provide.
- You will:

Section 2: Build Your Own RAG Chatbot

- The chatbot will answer questions based on a small set of documents that you provide.
- You will:
  1. Outline what kind of chatbot you want to build: (1) What is its purpose? (2) Who are its users?

Section 2: Build Your Own RAG Chatbot

- The chatbot will answer questions based on a small set of documents that you provide.
- You will:
    1. Outline what kind of chatbot you want to build: (1) What is its purpose? (2) Who are its users?
    2. Prepare a knowledge base for that chatbot: (1) A small collection of short text passages related to your topic (2) The chatbot will retrieve from this ”mini database” before generating an answer