DeepLearning.AI

Module 1 Introduction

Introduction to RAG

Module Overview

Introduction to RAG

RAG Architecture Deep Dive

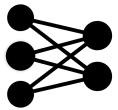
Real-World Applications

Hands-on Projects

Module 1
Basic RAG

Module 2
Retriever

Module 3
Vector Database



Module 4
LLM

Module 5
Monitoring & Evaluation

DeepLearning.AI

Introduction to RAG

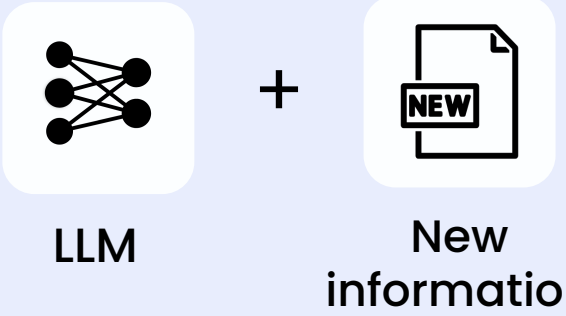
Introduction to RAG

Retrieval Augmented Generation (RAG)

LLMS are already powerful

- Summarize Text
- Generate Code
- Rewrite Content

RAG further improves them



Zain Hasan

Why are hotels expensive on the weekend?

"More people travel on weekends, so there’s more competition for rooms.”

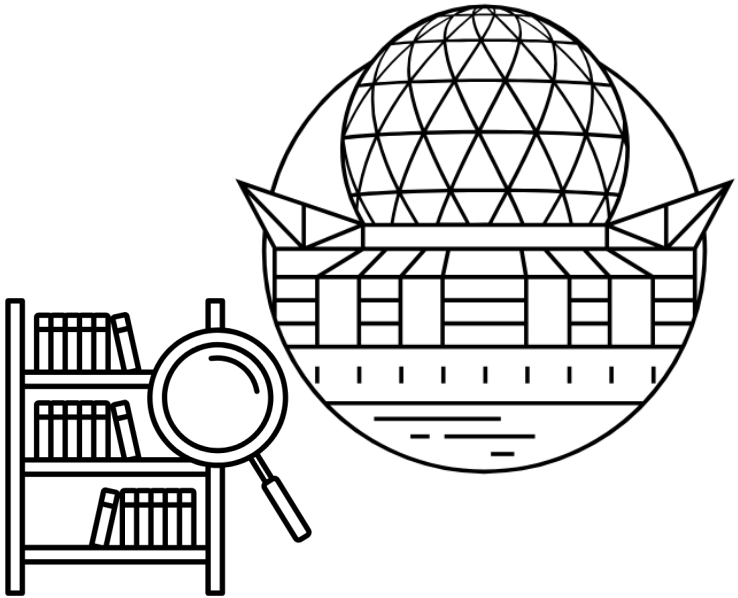
Why are hotels in Vancouver super expensive this coming weekend?

Zain Hasan

Why are hotels in Vancouver super expensive this coming weekend?

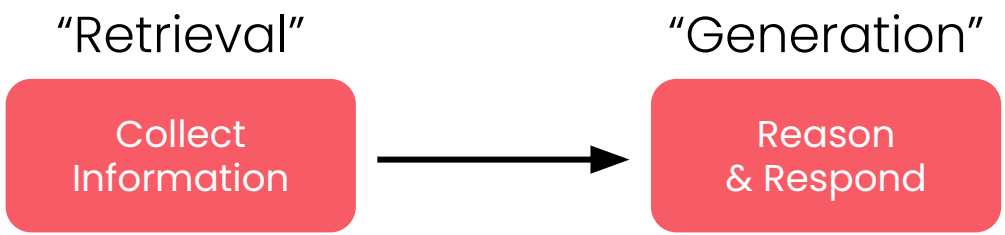
Taylor Swift is in town this weekend

Why doesn’t Vancouver have more hotel capacity close to downtown?



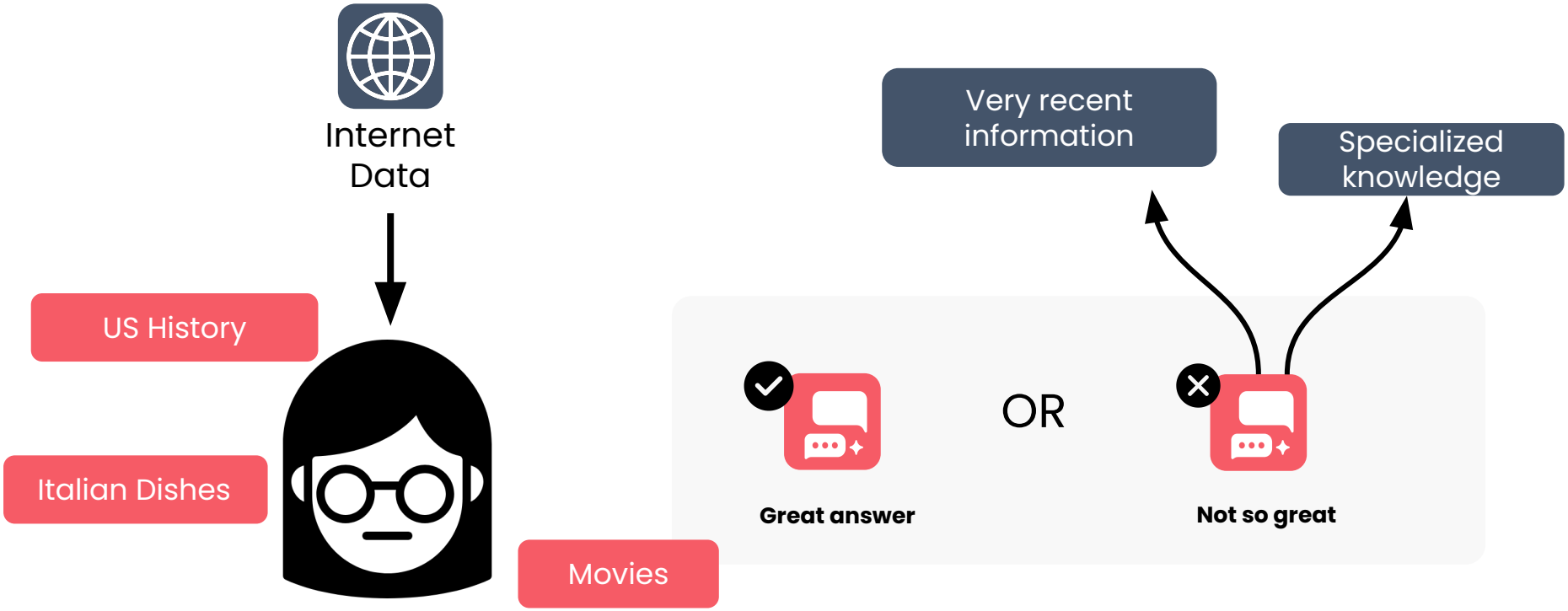
Zain Hasan

Two Steps for Answering Questions

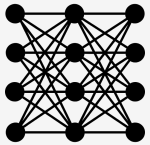


Why are hotels expensive on the weekend?	No need to collect information	Respond based on your knowledge
Why doesn’t Vancouver have more hotel capacity close to downtown?	Extensive Research	Synthesize research

Traditional Language Model Usage



LLM



Mathematical Models

User Query

Why doesn't Vancouver have more hotel capacity close to downtown?

Urban Development in Vancouver
Zoning laws and city planning since the early 1900s limited hotel growth downtown.

Forum Comment
Probably because land costs are super high near downtown.

Training Data

Massive dataset from the open internet

What LLMs don’t know

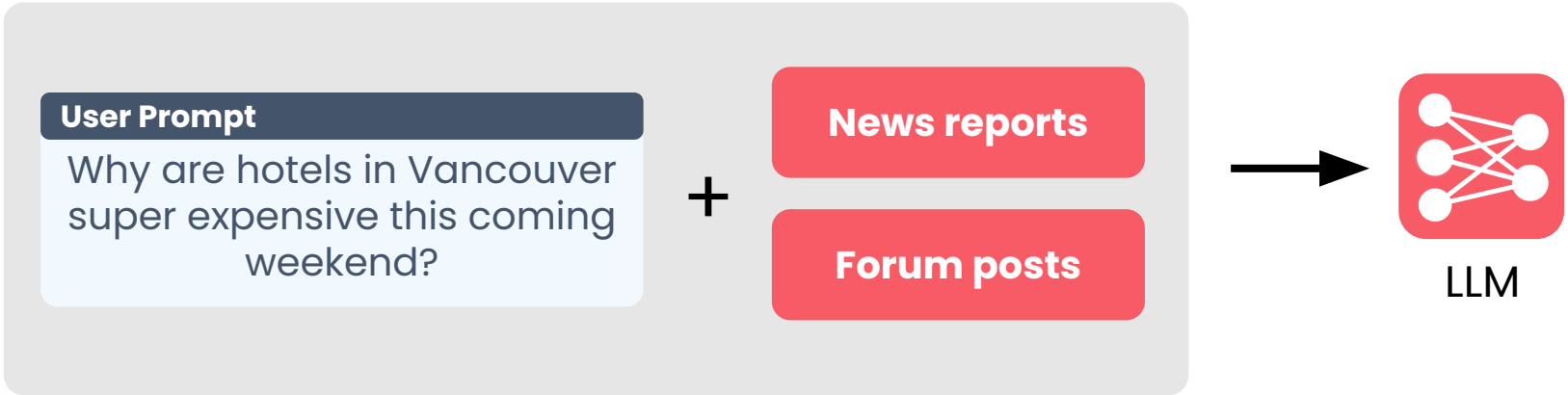
Private Databases
LLMS can’t access confidential information

Hard to access information
Some information isn’t widely available online, making it inaccessible to LLMs.

Real time data
LLMs are trained on past data and don’t update automatically

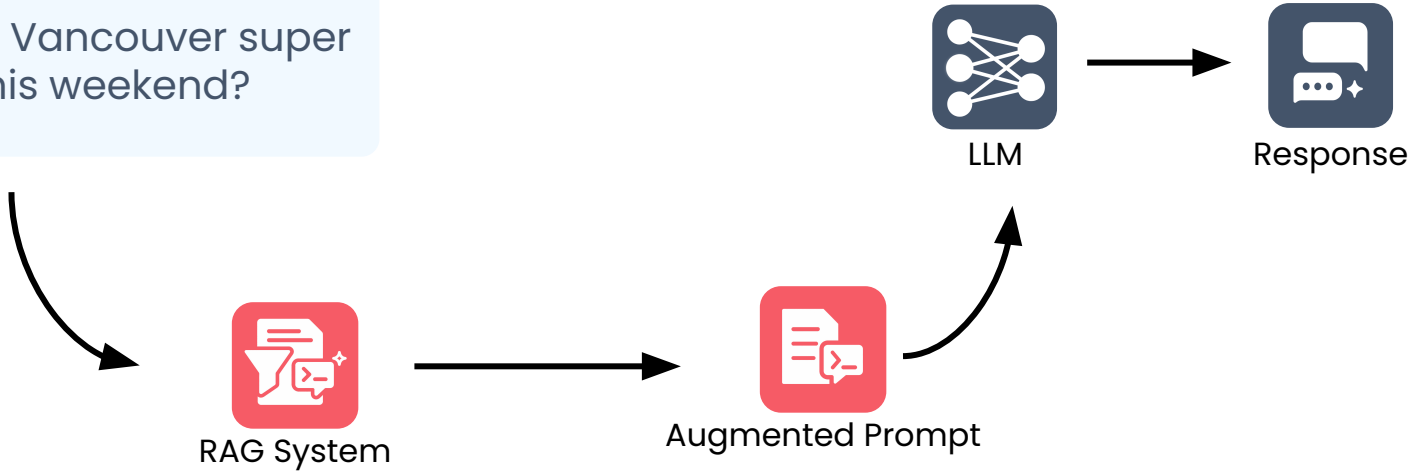
How do you make sure the LLM knows this useful information?

Just put it in the prompt!



User Prompt

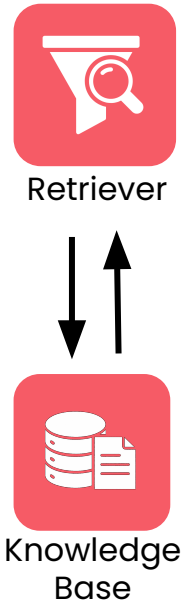
Why are hotels in Vancouver super expensive this weekend?



Taylor Swift is performing her **Eras Tour in Vancouver. This weekend** at BC Place Stadium on **December 6-8, 2024**

Retriever

- Manages knowledge base of trusted information
- Finds the most relevant Information and shares information with the LLM
- Improves generation



Retrieval Augmented Generation

DeepLearning.AI

Zain Hasan

DeepLearning.AI

Applications of RAG

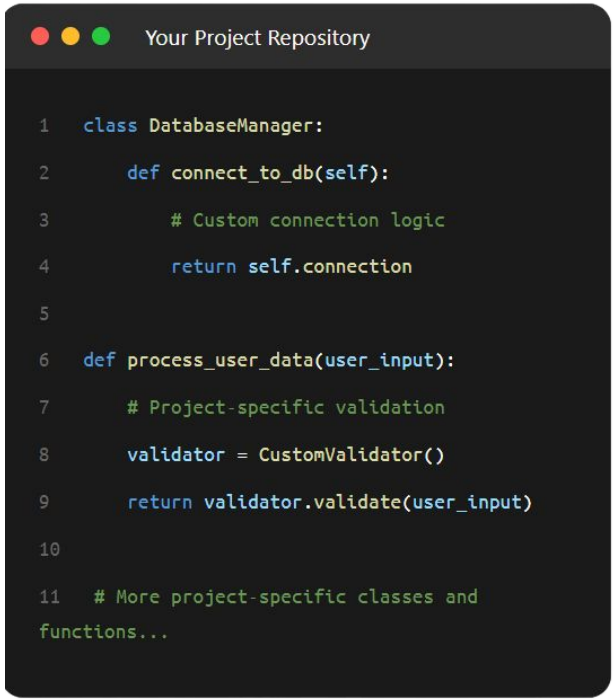
Applications of RAG

Applications of RAG - Code Generation

LLM needs your project's context
Classes, functions, definitions, and coding style

Use your codebase as a knowledge base
RAG retrieves project-specific content for the LLM

Improves code generation and Q&A
Answers are tailored to your actual repository

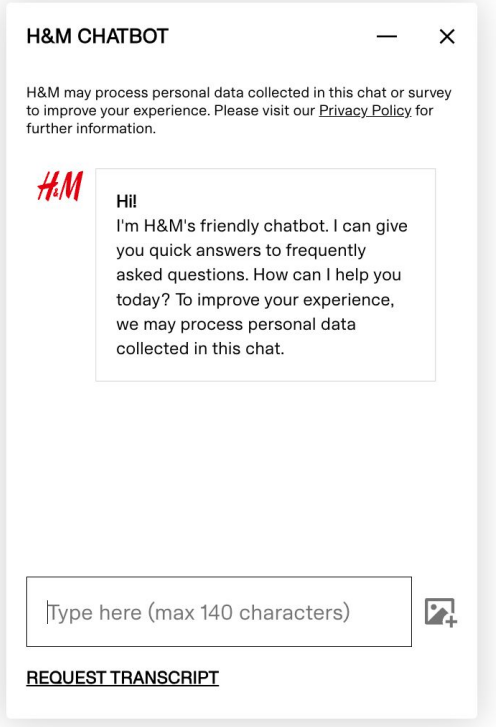


Applications of RAG - Company Chatbots

Tailored to your company
Every business has its own products, policies, and communication style

Uses your internal documents
Manuals, support guides, FAQs

Grounds answers in real context
Reduces generic or incorrect answers



Applications of RAG - Specialized Knowledge

High-impact domains
Legal and medical use cases

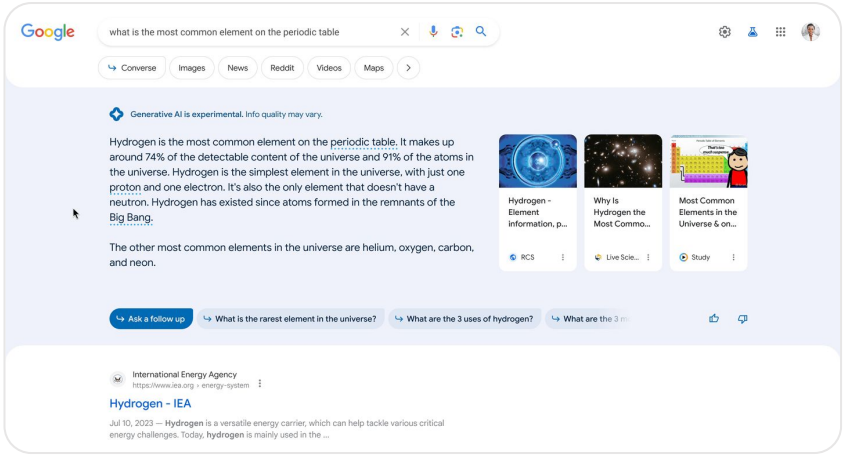
Uses specialized documents
Case files, journals, private data

Enables accurate, secure use
Supports precision and privacy needs

Applications of RAG

Search engines as retrievers
Return websites for a given query

AI summarizes search results
Presents key info in a skimmable format



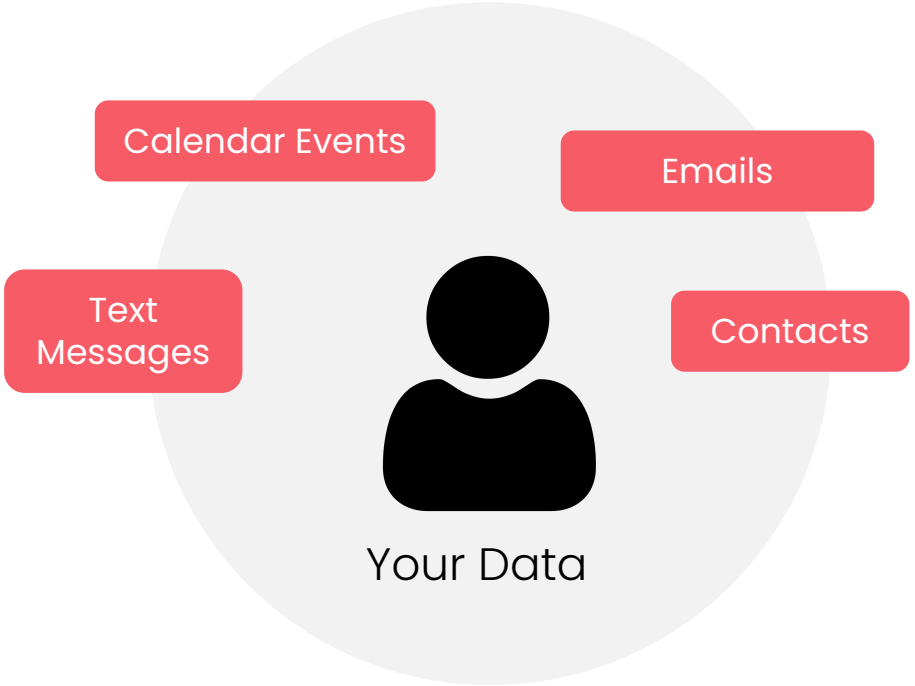
RAG with the internet as a knowledge base
Summaries powered by real-time retrieval

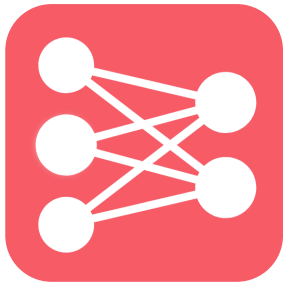
Personalized RAG

More software includes personal assistants
Messaging app, email client, etc.

These tools need context
More context leads to better results

Your data is the knowledge base
Texts, contacts, etc.





LLM

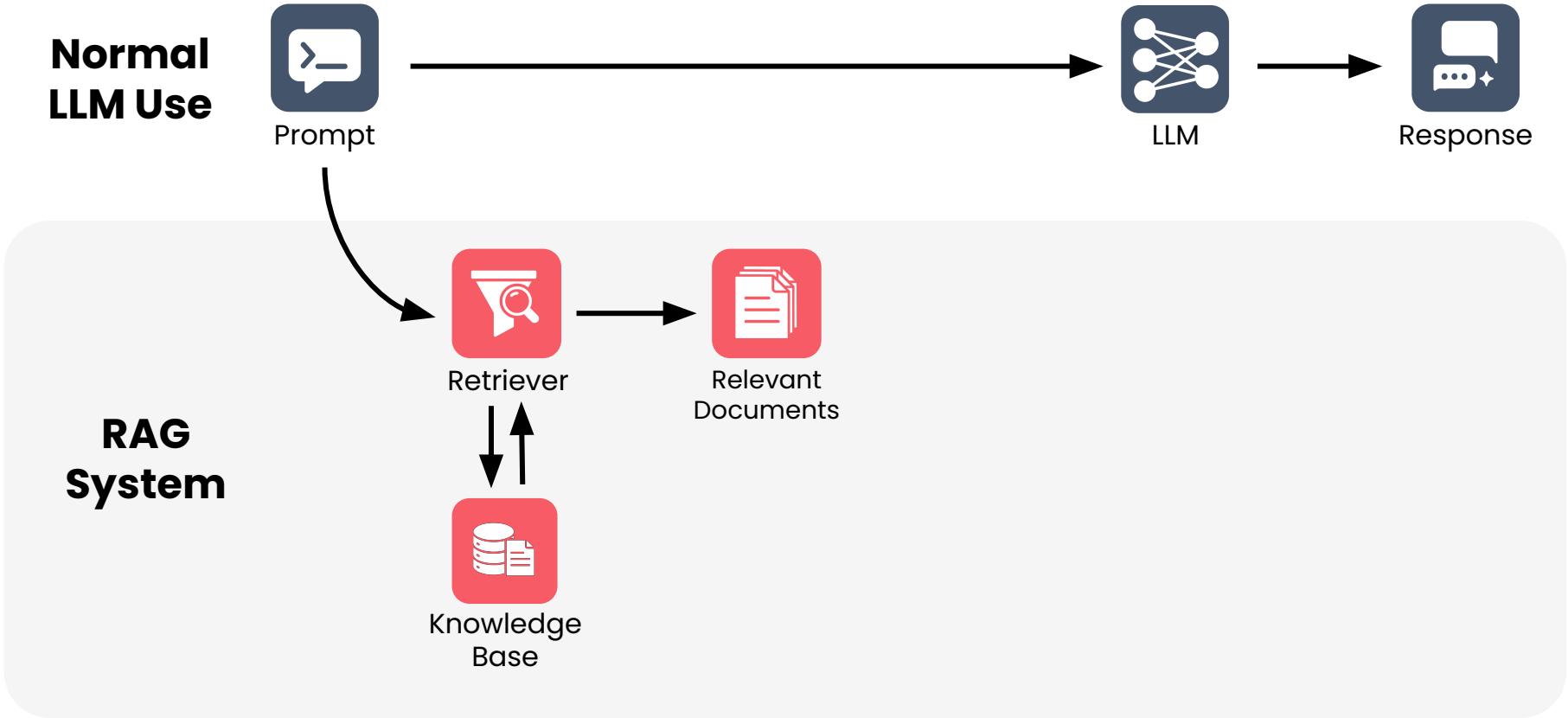
+ your data

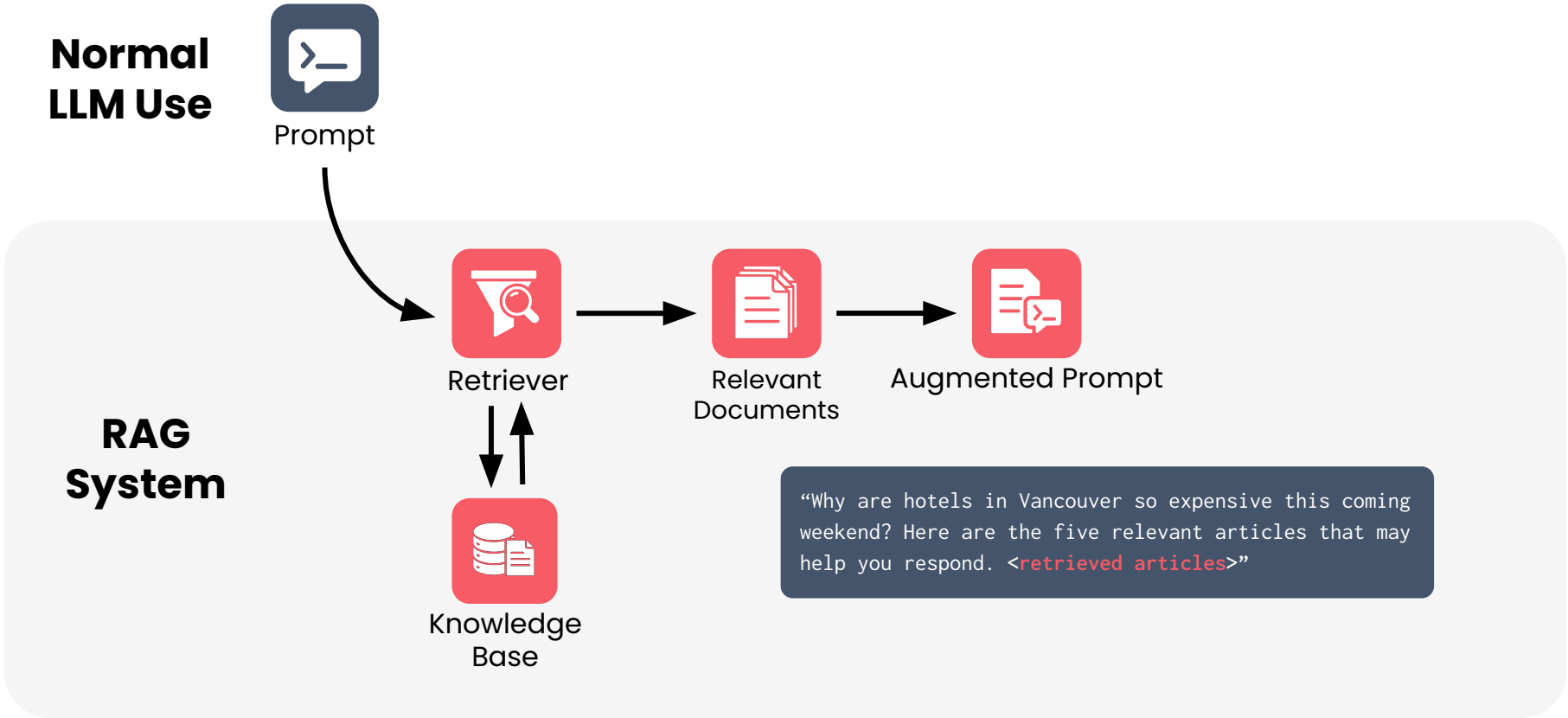
Zain Hasan

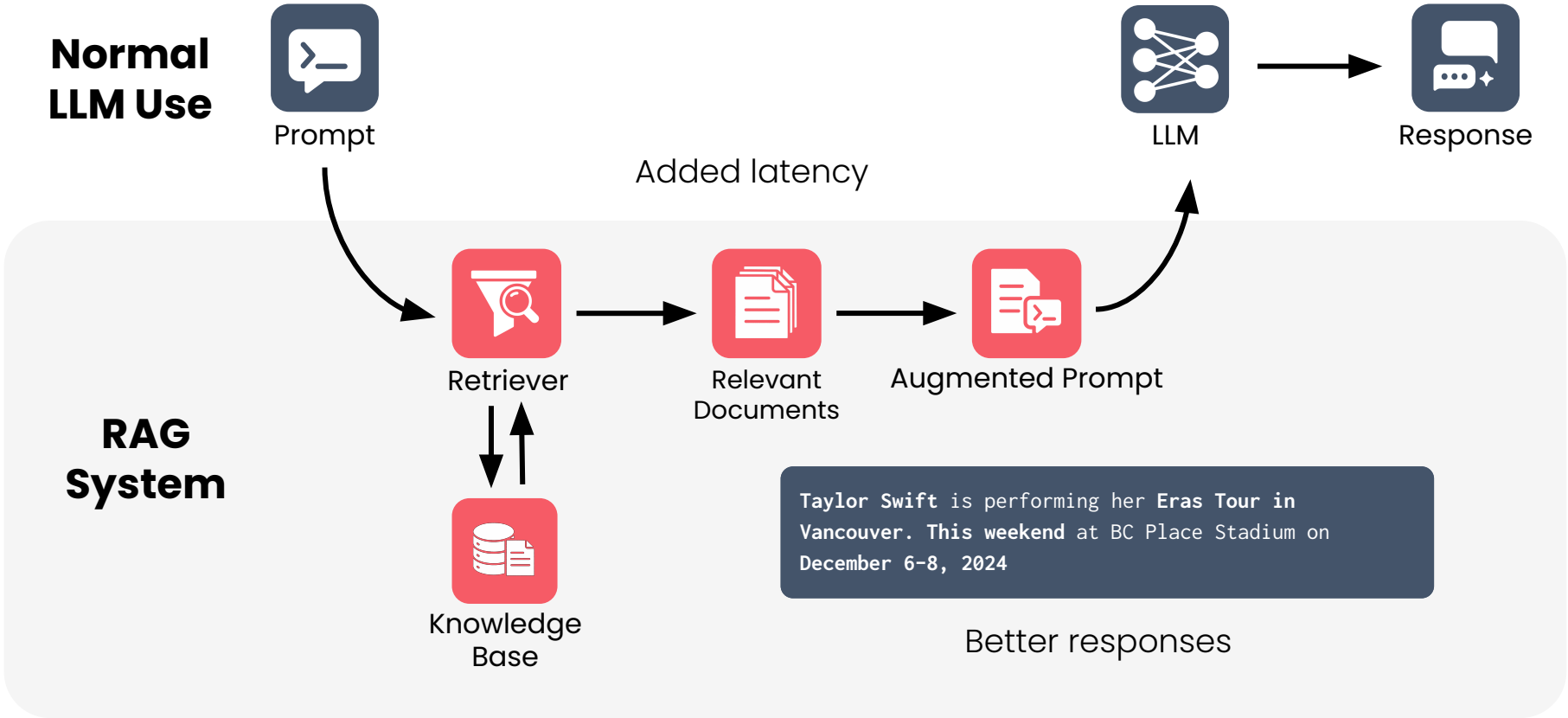
DeepLearning.AI

RAG Architecture Overview

Introduction to RAG


Normal
LLM Use
Prompt
LLM
Response
RAG
System
Retriever
Relevant
Documents
Knowledge
Base


Normal LLM Use
Prompt
RAG System
Retriever
Relevant Documents
Augmented Prompt
Knowledge Base
“Why are hotels in Vancouver so expensive this coming weekend? Here are the five relevant articles that may help you respond. <retrieved articles>”


Normal LLM Use
Prompt
LLM
Response
Added latency
RAG System
Retriever
Relevant Documents
Augmented Prompt
Knowledge Base
Taylor Swift is performing her Eras Tour in Vancouver. This weekend at BC Place Stadium on December 6-8, 2024
Better responses

Advantages of RAG

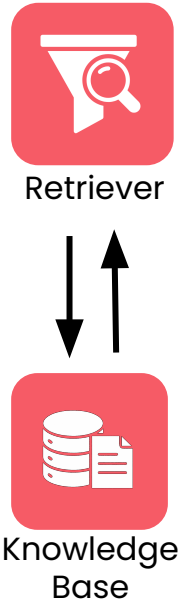
Injects missing knowledge
Adds info not in the training data (e.g. policies, updates)

Reduces hallucinations
Grounds answers with relevant context

Keeps models up to date
Reflects new info by updating the knowledge base

Enables source citation
Includes sources for verifiable answers

Focuses model on generation
Retriever finds facts, LLM writes responses



"Why are hotels in Vancouver so expensive this coming weekend? Here are the five relevant articles that may help you respond. <retrieved articles>"

DeepLearning.AI

Introduction to LLMs

Introduction to RAG

LLMs are just fancy autocomplete

Zain Hasan

What a beautiful day, the sun is `out`

Prompt

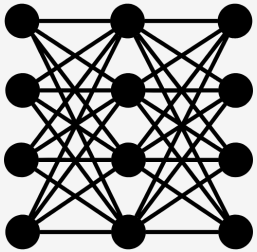
What a beautiful day the sun is shining

What a beautiful day the sun is rising

What a beautiful day the sun is out

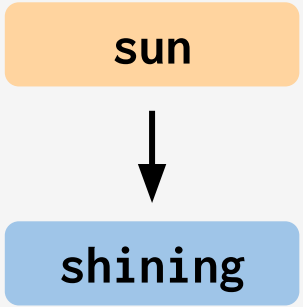
Completions

What a beautiful day, the sun is `exploding`



Neural Network

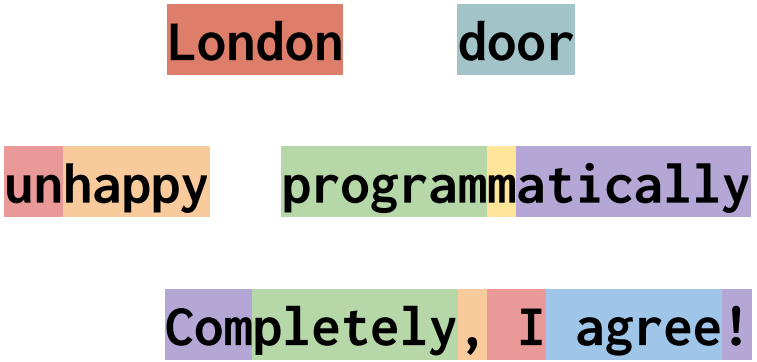
- a complex mathematical model of language
- stores which words frequently appear together, in which order, and contextual meaning
- LLMs use this model to generate text



What a beautiful day, the sun is shining in the sky

Token

- a piece of a word
- some words get single tokens
- compound words use multiple tokens
- punctuation marks
- ~10,000 - 100,000 tokens in LLM's vocabulary, allowing models to represent any possible word with fewer tokens

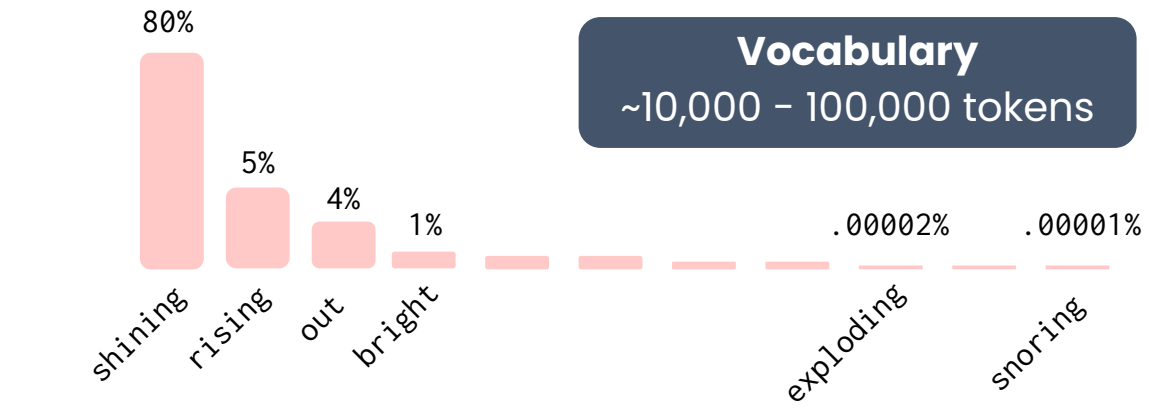


What a beautiful day, the sun is

Process Current State

Calculate Probabilities

Select Next Token



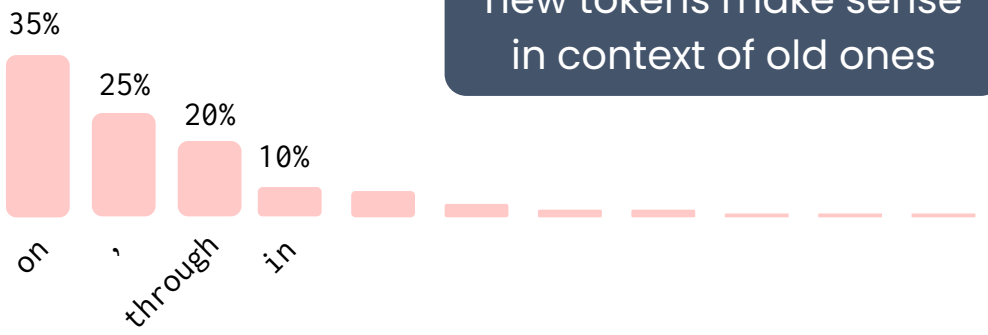
shining

What a beautiful day, the sun is **shining**

Process Current State

Calculate Probabilities

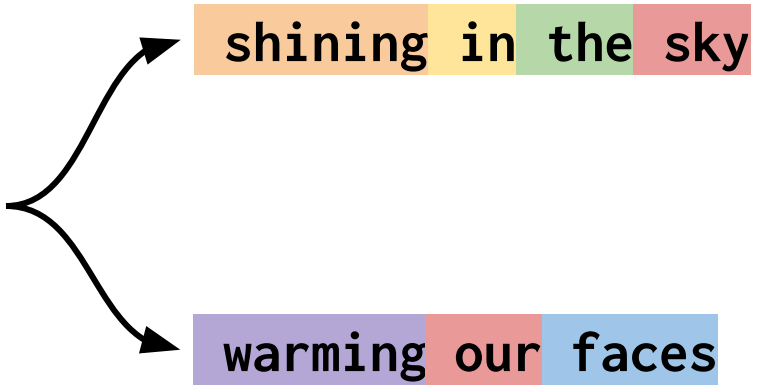
Select Next Token



new tokens make sense in context of old ones

in

What a beautiful day, the sun is



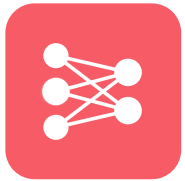
Autoregressive

- “self-influencing”
- new tokens make sense in context of old ones
- running the same prompt leads to different completion

How LLMs Learn

Large text collections

Trains model
usually billions of parameters!

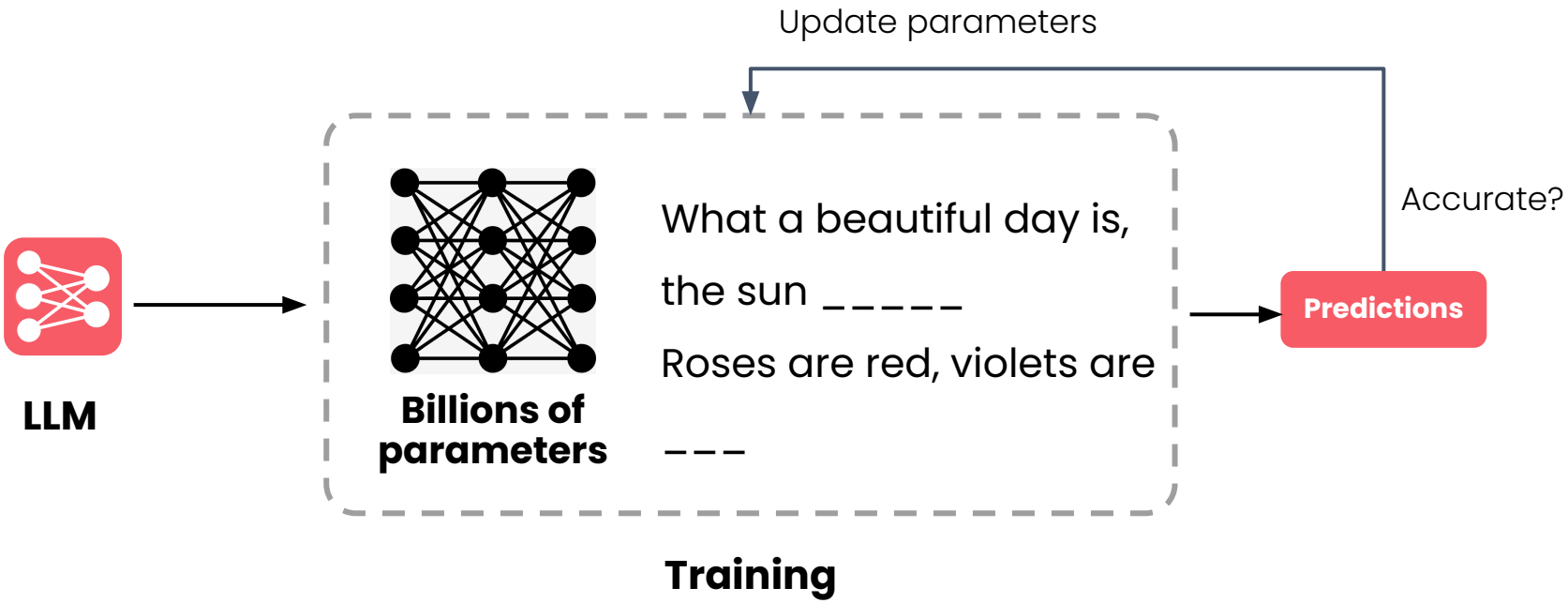


LLM

Before training, LLMs generate gibberish

"Forward to Saturn’s dance floor!" she yowled,
tail transmitting disco beats.

How LLMs Learn



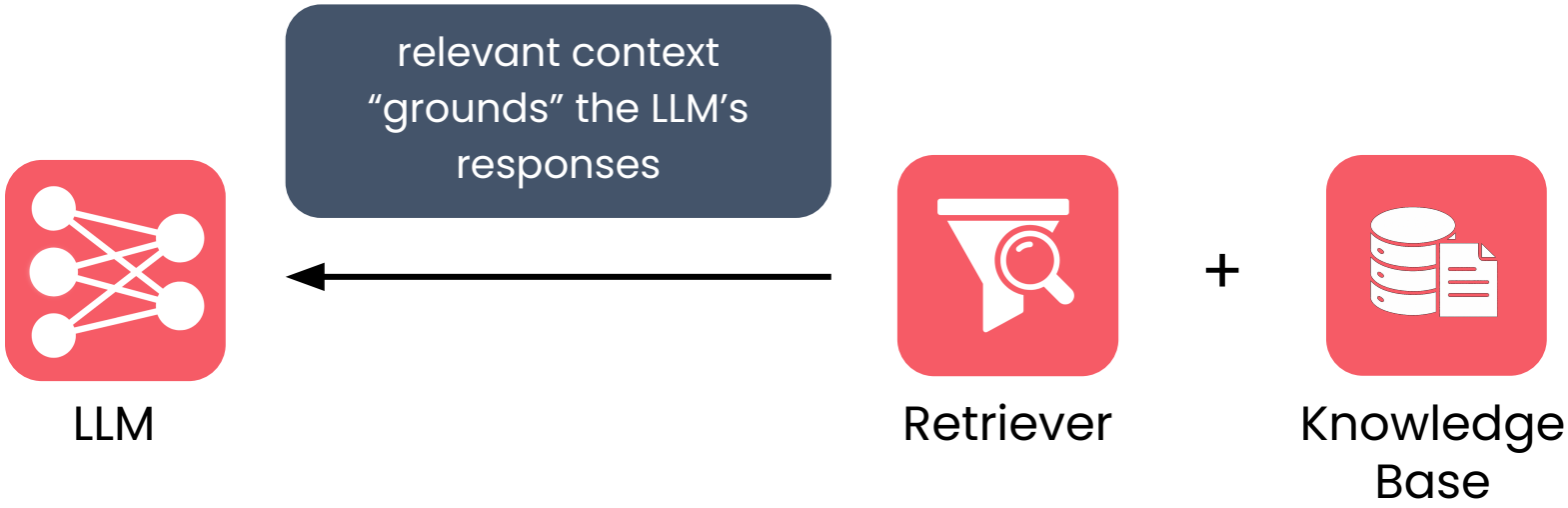
Why LLMs Hallucinate

- **LLMs generate probable word sequences**
  LLMs just reproduce statistical patterns from their training data

- **Knowledge gaps cause inaccurate responses**
  Responses can “sound right” but aren't true.

- **Truthful ≠ probable**
  LLMs are designed to generate "probable" text, not truthful text

How RAG solves the problem



Why not add everything?

Higher Computational Cost

- Longer prompts take more computation to run
- Model performs computationally complex scan of every token
- Scan happens before generating each new token

Context Window Limit

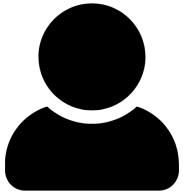
- Eventually you hit the limit of LLM's context window
- Smaller models: only a few thousand tokens
- Largest models: millions

Zain Hasan

DeepLearning.AI

Introduction to information retrieval

Introduction to RAG



Your Question

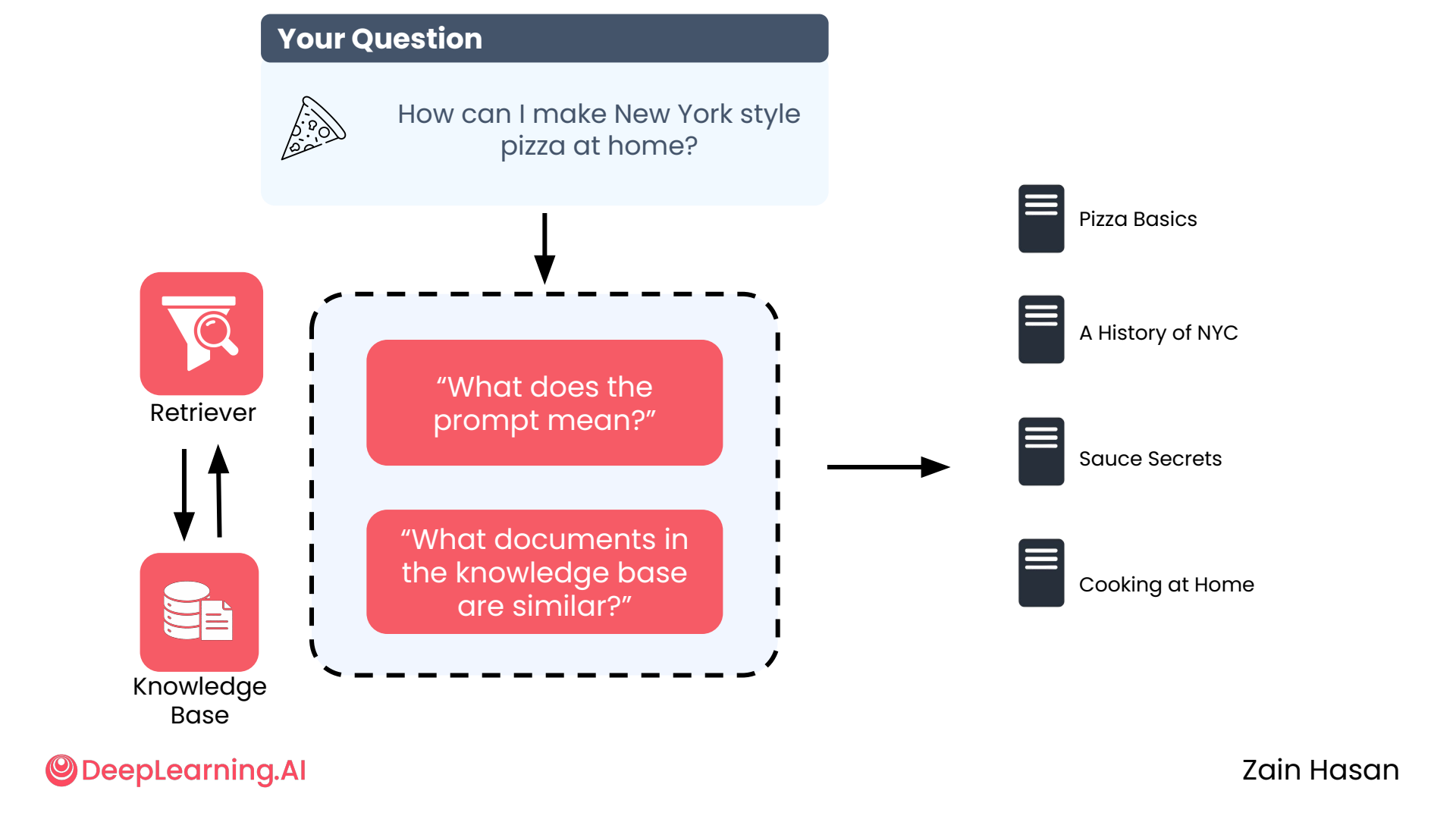
How can I make New York style pizza at home?

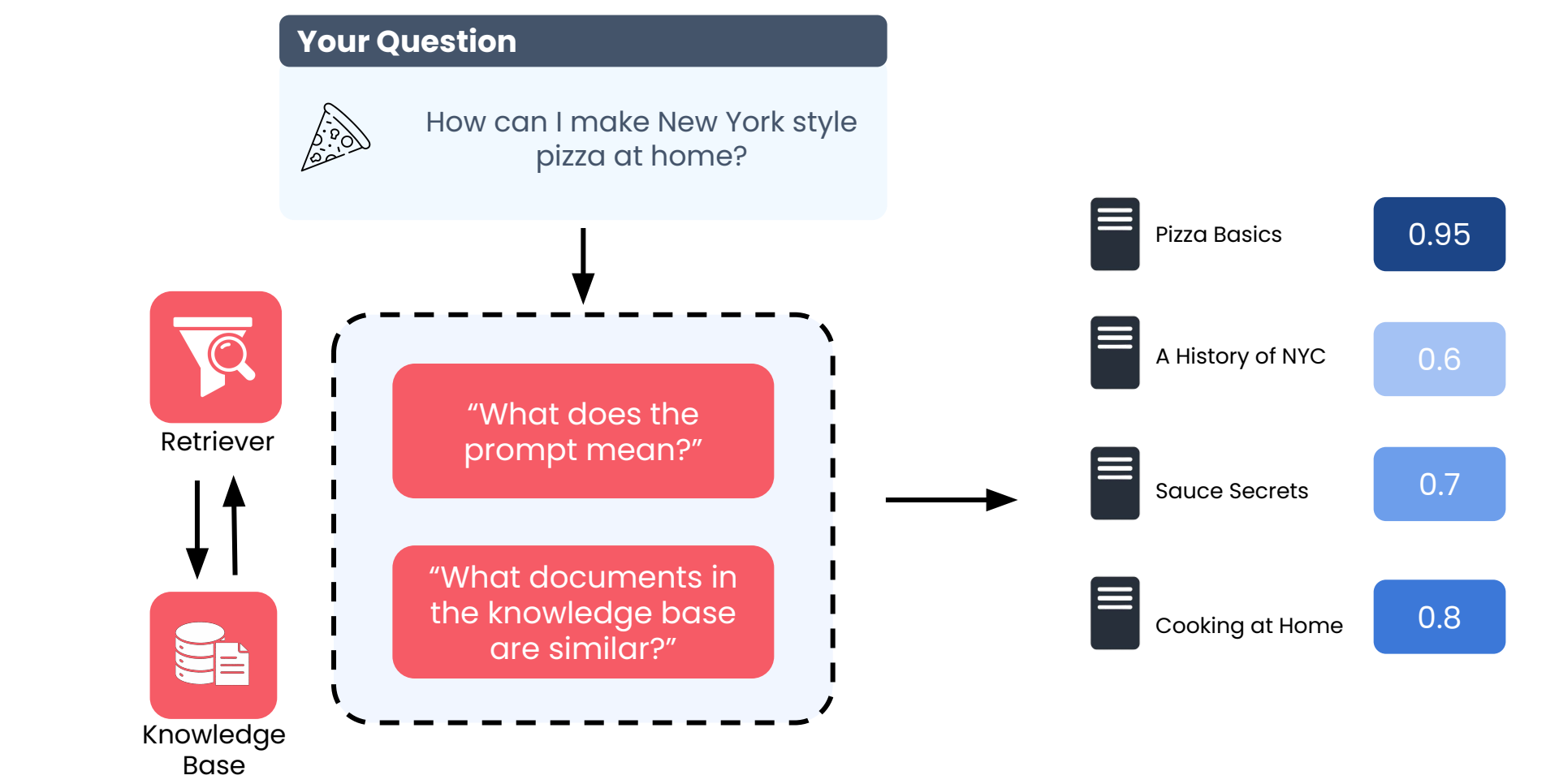
	Collection	Organization	Search
 **Library**	Books on many topics	Different shelves and sections	Librarian helps you find best sections or books
 **Retriever**	Documents in a database	“index” for search	Retriever searches the index

Search with a librarian

Librarian

- Understands the meaning of your question
- Identifies the right shelves to search
- Eventually finds relevant books


Your Question
How can I make New York style pizza at home?
Retriever
Knowledge Base
"What does the prompt mean?"
"What documents in the knowledge base are similar?"
Pizza Basics
A History of NYC
Sauce Secrets
Cooking at Home


Your Question
How can I make New York style pizza at home?
Retriever
Knowledge Base
"What does the prompt mean?"
"What documents in the knowledge base are similar?"
Pizza Basics
0.95
A History of NYC
0.6
Sauce Secrets
0.7
Cooking at Home
0.8

Retriever Tradeoffs

- **Relevance vs irrelevance**
  Need to return relevant documents and withhold irrelevant ones
- **Return every document?**
  Mountains of irrelevant docs. Wastes context window
- **Return the single highest ranked document?**
  Miss valuable information
- **No perfect solution**
  Retriever usually doesn’t perfectly rank documents
- **Monitor and experiment**
  Change settings to find what works

Search Engine

Retrieves relevant webpages

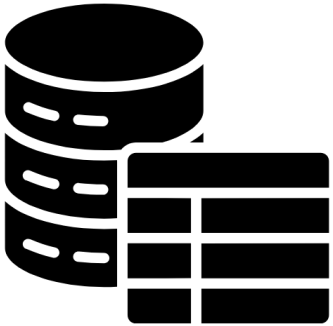
Database

Retrieves relevant tables and rows

Historical Context

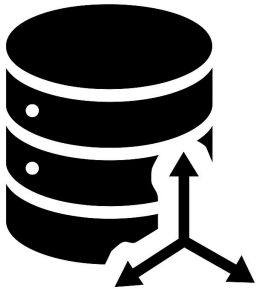
Information Retrieval was already mature when LLMs were first developed

Practical Implementation



Relational Database

Already widely adopted



Vector Database

Specialized for retrieval in a RAG system

DeepLearning.AI

Module 1 Conclusion

Introduction to RAG

Key Concepts

- RAG pairs an LLM with a knowledge base
- Data is private, recent, or highly specific and so missing from the LLM's training data
- Retriever finds relevant documents and adds them to an augmented prompt
- LLMs ground their responses in the retrieved information