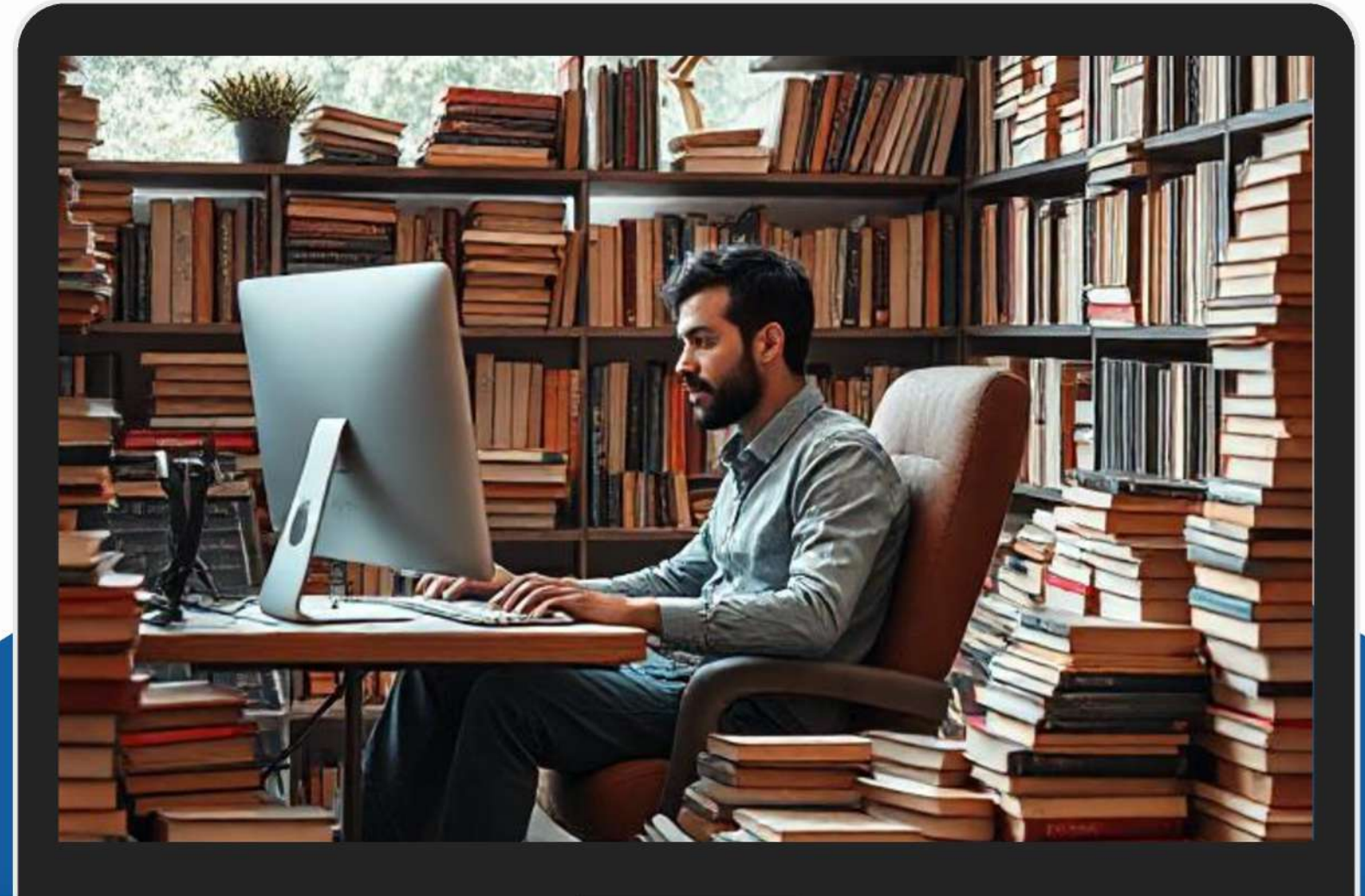


# Smarter Search with RAG: Enhancing Library Digital Repositories

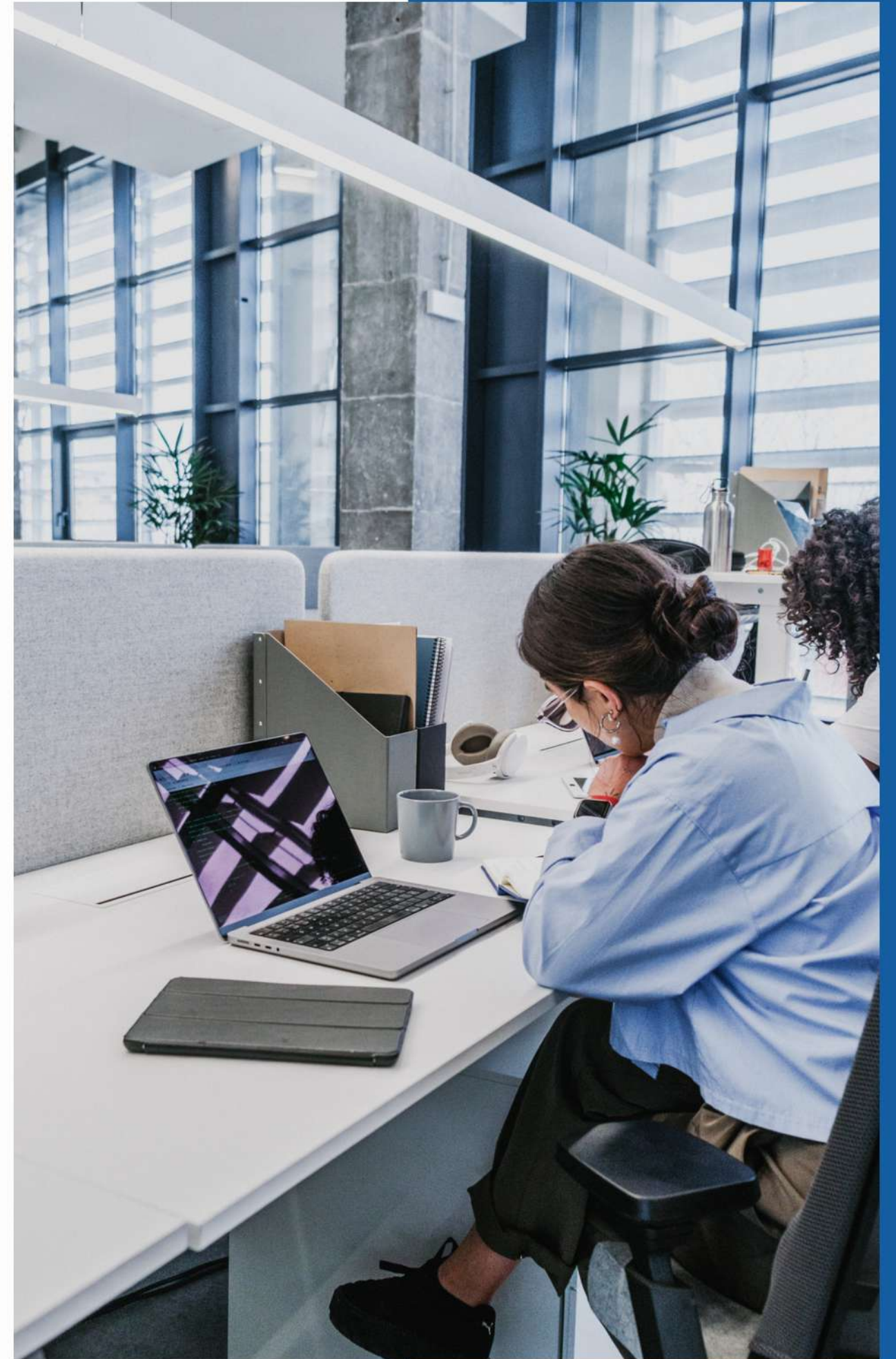
Dr Kevin Wu  
Eric Chu

5th Nov 2025



# Overview

- ▶ Introduction
- ▶ AI and LLMs
- ▶ Challenges of LLMs
- ▶ Introducing RAG
- ▶ Applications & Examples
- ▶ Hands-on Practice
- ▶ Q&A

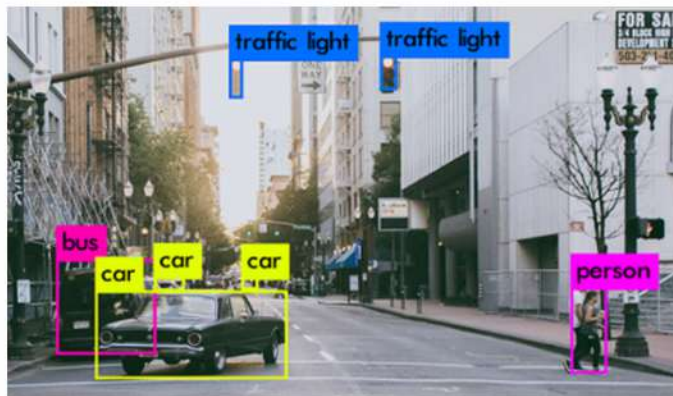


# What is Artificial Intelligence (AI)?

According to IBM, AI is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy.

## AI can be employed in different fields:

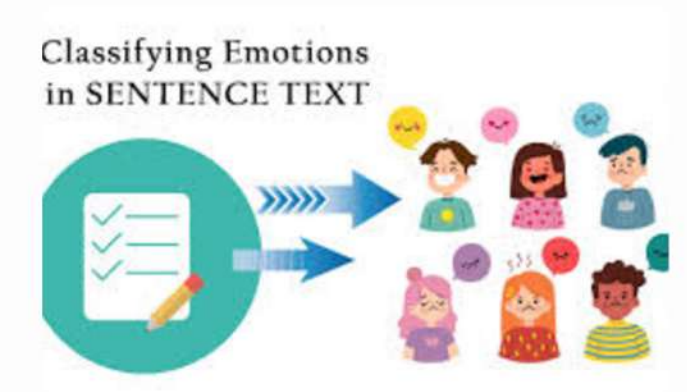
### Visual Perception



### Language Translation



### Sentiment Classification



### Speech Recognition



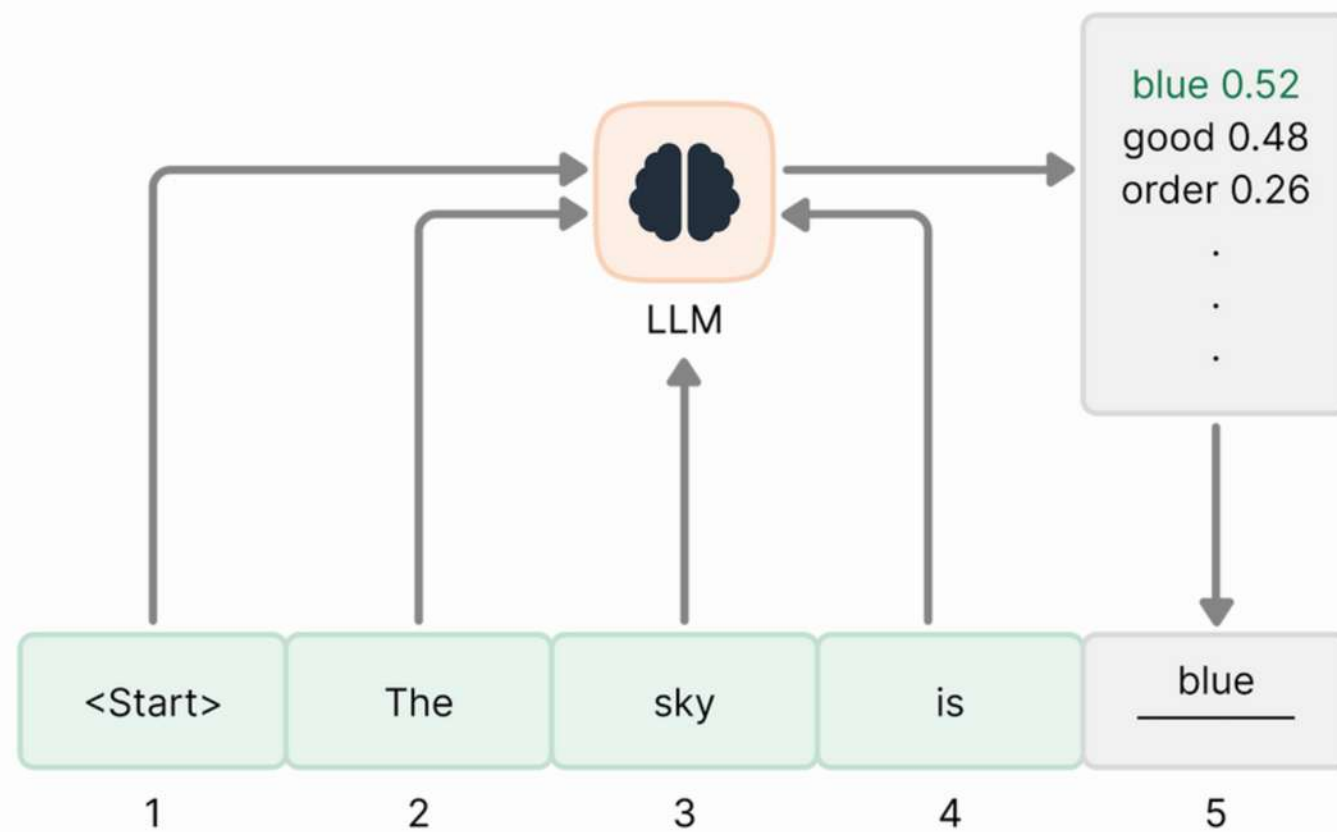
### Trading and Finance



### Fraud Detection



# AI AND LLMs: Next-word prediction



## Large Language Model (LLMs)

- LLM is a type of AI trained to predict the next word based on patterns
- It is trained on millions of data online to understand the relation among words.

## Example

The sky is \_\_\_" → LLM predicts "blue" as highly probable

"I want a cup of \_\_\_" → LLM suggests "tea" or "coffee"

# Example: HKU Main Building



*DALLE2: A young man thinking*

What is the history of  
HKU Main Building?

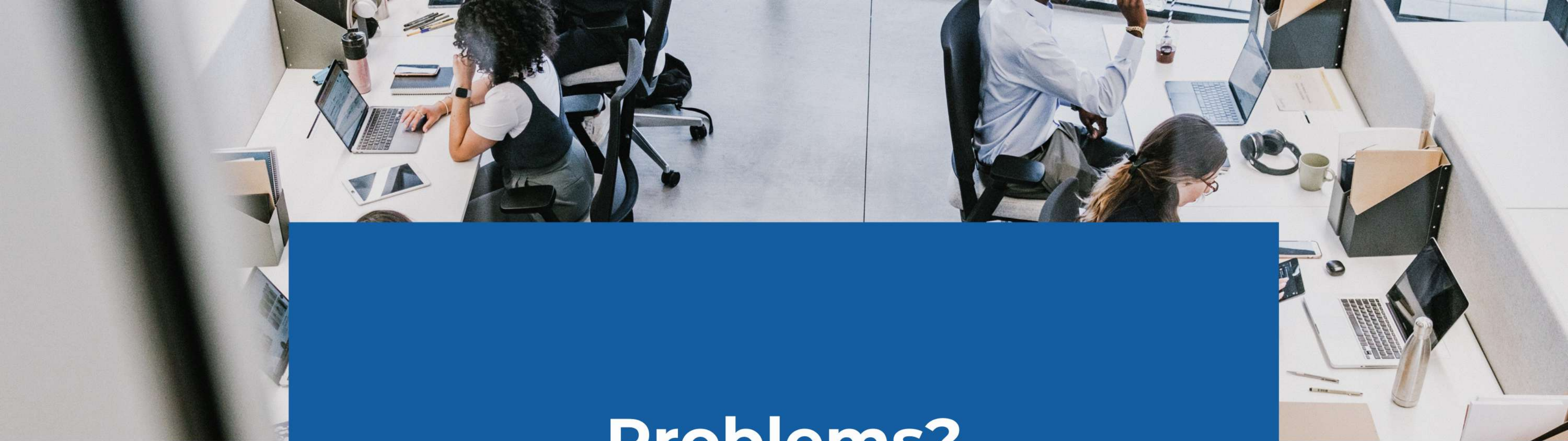


<https://www.alumni.hku.hk/hku-in-films-series-1/>

ChatGPT (Tokens used 209)

The University of Hong Kong (HKU) Main Building is a historic landmark located in Pokfulam, Hong Kong. It was designed by British architect William Henry Bowden and completed in 1912. The building has been the symbolic center of HKU since its construction and has gone through numerous renovations and expansions over the years.





**Problems?**



**ChatGPT** (Tokens used 209)

---

The University of Hong Kong (HKU) Main Building is a historic landmark located in Pokfulam, Hong Kong. It was designed by British architect William Henry Bowden and completed in 1912. The building has been the symbolic center of HKU since its construction and has gone through numerous renovations and expansions over the years.

During World War II, the Japanese Imperial Army occupied the building and used it as a military headquarters. After the war, the building was returned to the University and underwent extensive repairs. In the 1950s and 1960s, additional wings were added to the Main Building to accommodate the University's growing student population.

Today, the Main Building is a Grade 1 historic building in Hong Kong and is a popular tourist attraction. It houses various administrative offices, classrooms, and lecture halls, as well as the University Museum and Art Gallery.

ChatGPT (Tokens used 209)

The University of Hong Kong (HKU) Main Building is a historic landmark located in Pokfulam, Hong Kong. It was designed by British architect **William Henry Bowden** and completed in 1912. The building has been the symbolic center of HKU since its construction and has gone through numerous renovations and expansions over the years.

During World War II, the Japanese Imperial Army occupied the building and used it as a **military headquarters**. After the war, the building was returned to the University and underwent extensive repairs. In the 1950s and 1960s, **additional wings** were added to the Main Building to accommodate the University's growing student population.

Today, the Main Building is a **Grade 1 historic building** in Hong Kong and is a popular tourist attraction. It houses various administrative offices, classrooms, and lecture halls, as well as the **University Museum and Art Gallery**.



## The Main Building

The Main Building is the oldest structure on the HKU Main Campus. This Edwardian Baroque-style building was designed by **Alfred Bryer** of Leigh & Orange. Construction began in 1910 and was completed in 1912. The building was a gift from Sir Hormusjee N. Mody, a distinguished Parsi businessman and land developer in Hong Kong. "Mody Road" in Tsimshatsui bears his name. Without his generosity, the University may not have come into existence.

In December 1941 the Main Building was converted into a **relief hospital** due to the outbreak of war. During the war, parts of it were badly looted; the roof of the Great Hall was removed and the timber used as fuel. Sir Lindsay Ride later recalled that many of the University's buildings, including the Main Building, "... became roofless skeletons, stripped of equipment, furnishings and fittings, floors and stairs".

In the 1950s the Main Building was extended with the addition of two new **courtyards and a further floor** at the rear. In 1956, the old Great Hall was renamed Loke Yew Hall in memory of Dr Loke Yew, an early benefactor of the University.

In 1984, the Main Building was declared a **monument** by the Hong Kong Government. Today, visitors to the Main Building can see the clock tower, four courtyards, and palm trees which are now over 30 feet tall. They will also find a bronze bust of Sir Hormusjee N. Mody on the main staircase and another of Dr Loke Yew on the second floor.

# Limitation 1: Hallucination



**GPT-3.5** (Tokens used 311)

Certainly! Here are some example courses with course codes for a second-year student at the University of Hong Kong (HKU) across various faculties:

## 1. Faculty of Arts and Humanities

- CCCH2001 - Chinese Culture and Society
- ENGL2015 - Shakespeare
- LING2101 - Introduction to Linguistics

## 2. Faculty of Business and Economics

- ECON2101 - Intermediate Microeconomics
- FINA2104 - Corporate Finance
- MGT02101 - Organizational Behavior

## 3. Faculty of Engineering

- COMP2119 - Data Structures and Algorithms
- CIVL2202 - Structural Mechanics
- MECH2401 - Thermodynamics

This is a GPT-3.5 model response for the question:  
"Can you recommend me courses as a year 2 student at HKU?"

**Some of the courses are right,  
but most of the courses don't  
exist!!!**

## Limitation 2: Outdated data

[u3612776@connect.hku.hk](mailto:u3612776@connect.hku.hk)



Who is the latest Physics Nobel Price owner?

03/03/2025, 14:21:07



**GPT-4o** (Tokens used 670)

As of my last update in October 2023, the latest Nobel Prize in Physics was awarded to Pierre Agostini, Ferenc Krausz, and Anne L'Huillier for their experimental methods that generate attosecond pulses of light for the study of electron dynamics in matter. Please verify with a current and reliable source for the most up-to-date information.

03/03/2025, 14:21:07



GPT-4o is a large language model trained with data on or before 2023, which means it cannot answer questions that require data from the years after.

## Limitation 3: Limited knowledge on personal and private data

Here are some of the questions that an LLM can't answer

01 **Personal Chat History**  
When is the meeting with my team?

02 **Specific/Specialized Areas**  
What punishments would I get if I graffitied the floor of University Street in HKU?

03 **Private Database tables**  
How do I fetch the username from the table "users"

04 **Personal-related information**  
Where is my current home addresss?

# Introducing RAG: Best of Both Worlds

1

## **RAG**

Retrieval + Generation

2

## **Analogy**

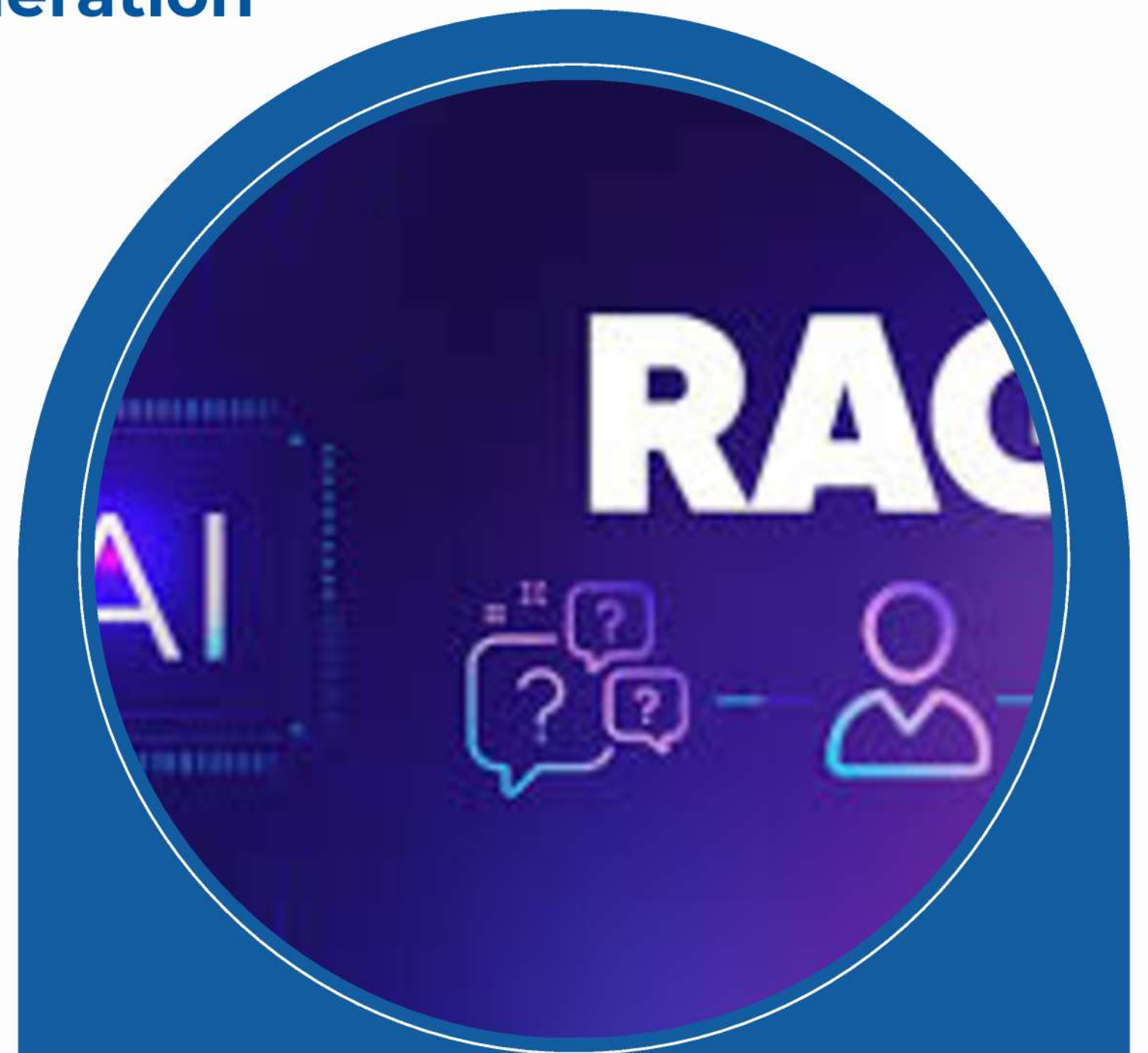
LLM + Search Engine



# Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) is the process of optimizing the output of an **LLM**, so it references an authoritative knowledge base outside of its training **data sources** before generating a response.

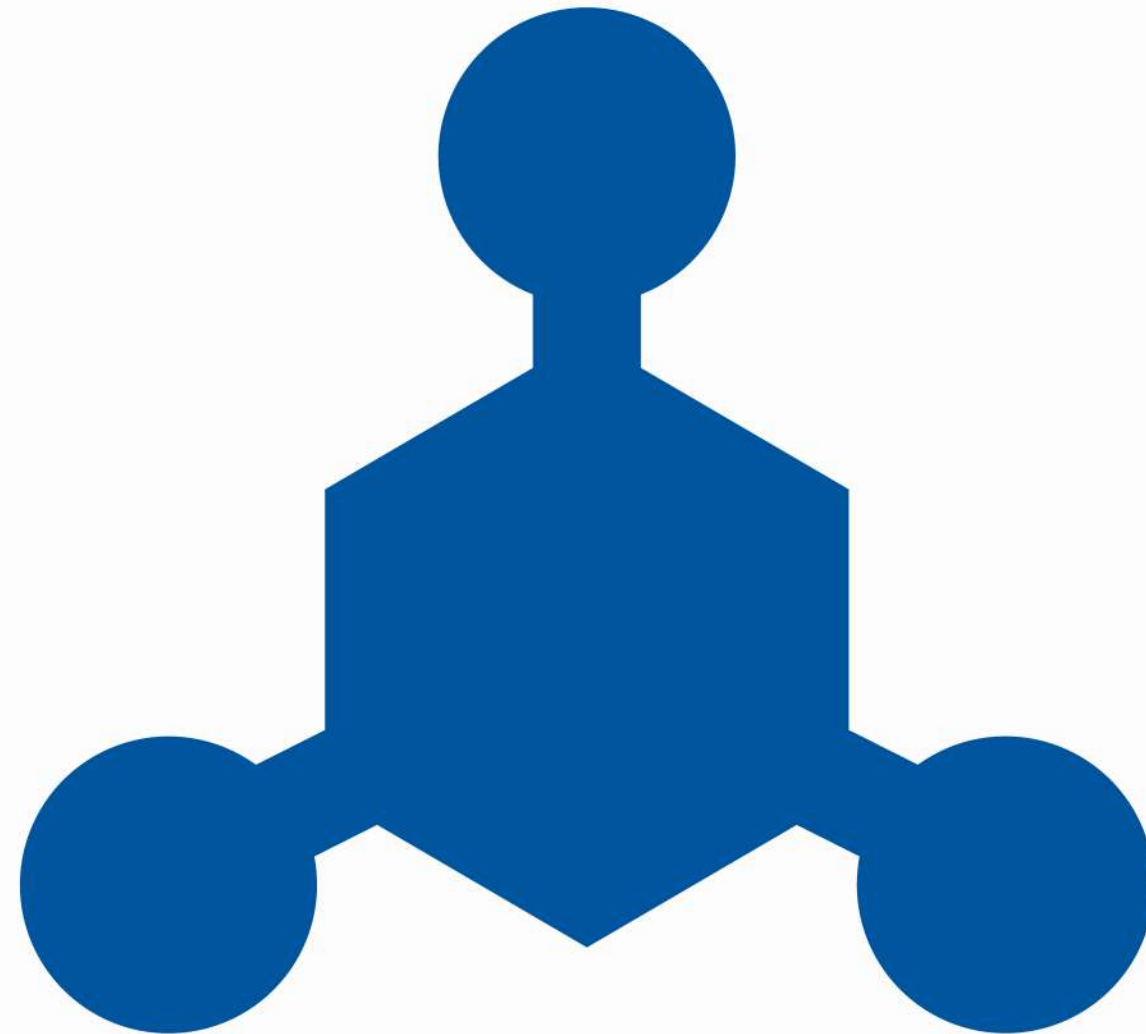
RAG extends the already powerful capabilities of LLMs to specific domains or an organization's internal knowledge base, all without the need to retrain the model. It is a cost-effective approach to improving LLM output so it remains **relevant, accurate**, and useful in **specific contexts**.



# Key Components of a RAG system

**Generator**

LLM



**Retriever**

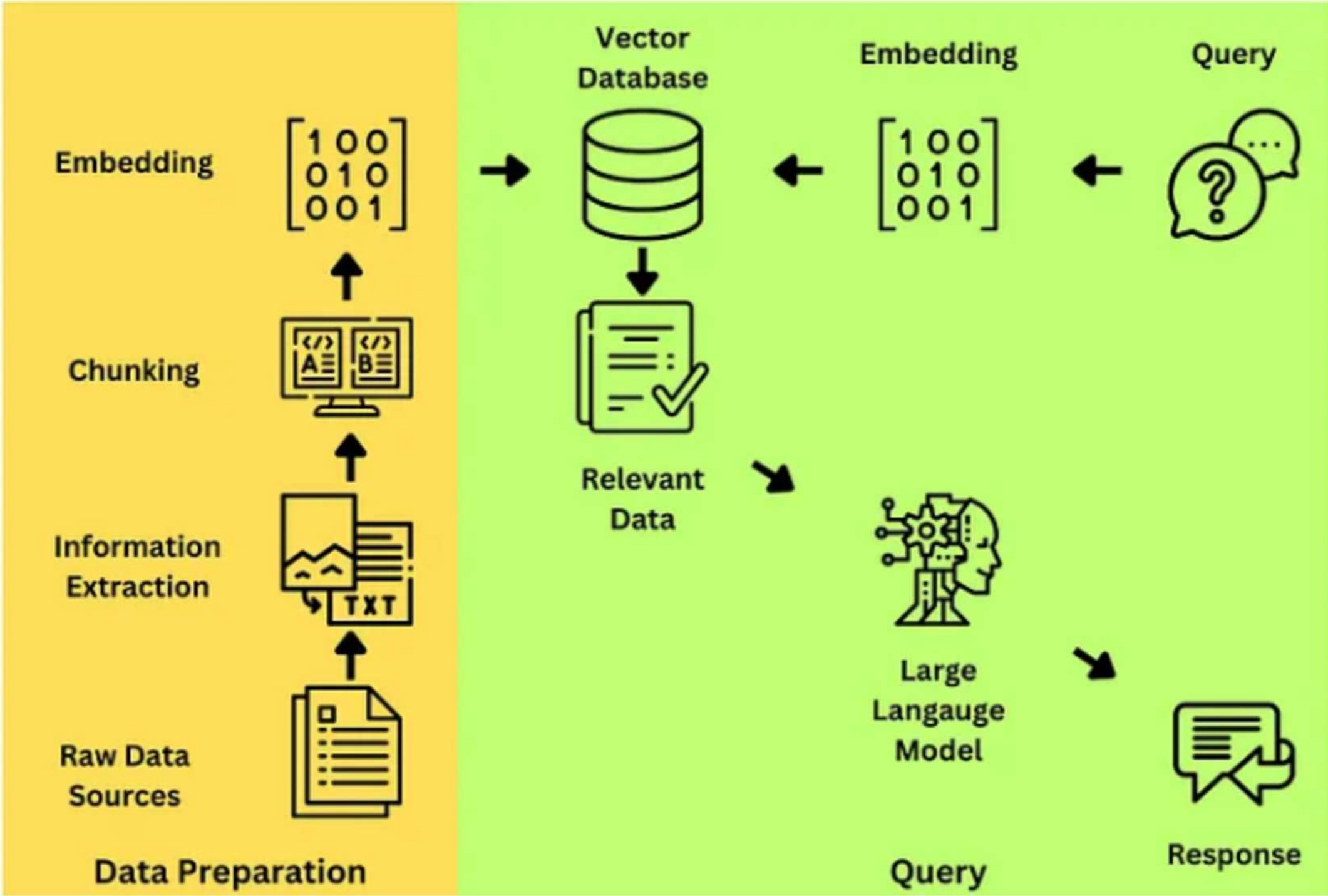
Search Engine

**Knowledge base**

Data Source

*How does this improve  
LLM's answer?*

# RAG Architecture



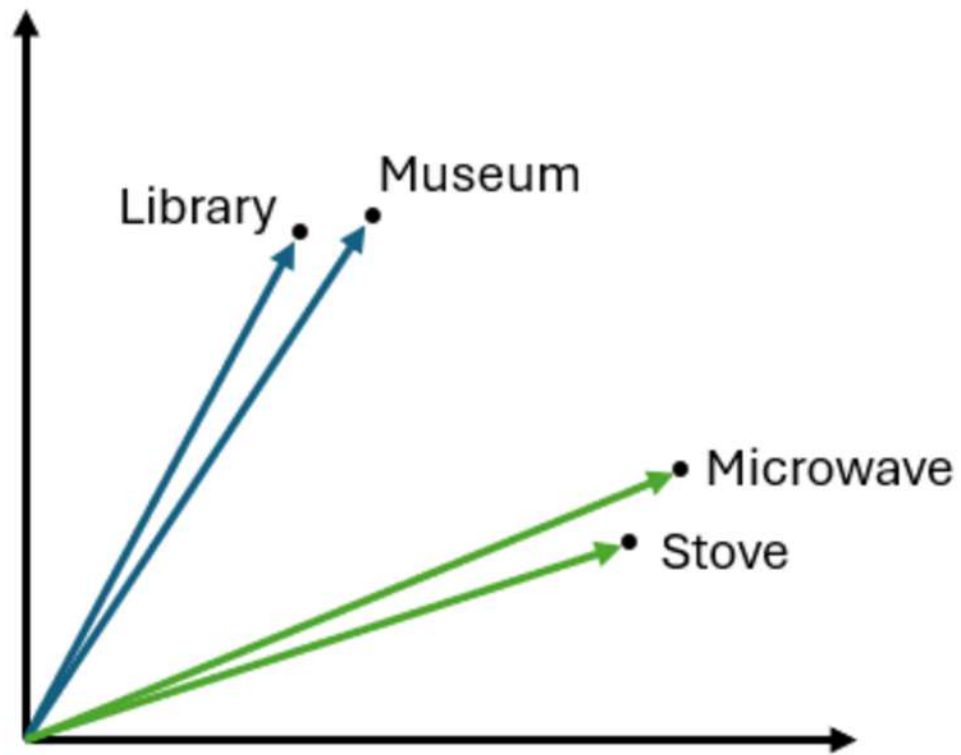
The RAG pipeline combines document retrieval with language model generation.

During **indexing**, documents are split into chunks, converted into embeddings, and stored in a database.

For **retrieval and generation**, a user's question is embedded to find relevant document chunks from the database, which are then fed into a LLM alongside the question.

The **LLM synthesizes** the retrieved information to generate a precise, context-aware answer, enhancing accuracy by grounding responses in external knowledge.

# Word Vectors



**Word Vectors** creates a representation of each word present in our vocabulary into a vector.

**An embedding model** converts chunks of text from a knowledge base (like documents) into dense numerical vectors. Similar words have similar embeddings.

# Coding Session

Let's build a basic RAG chatbot

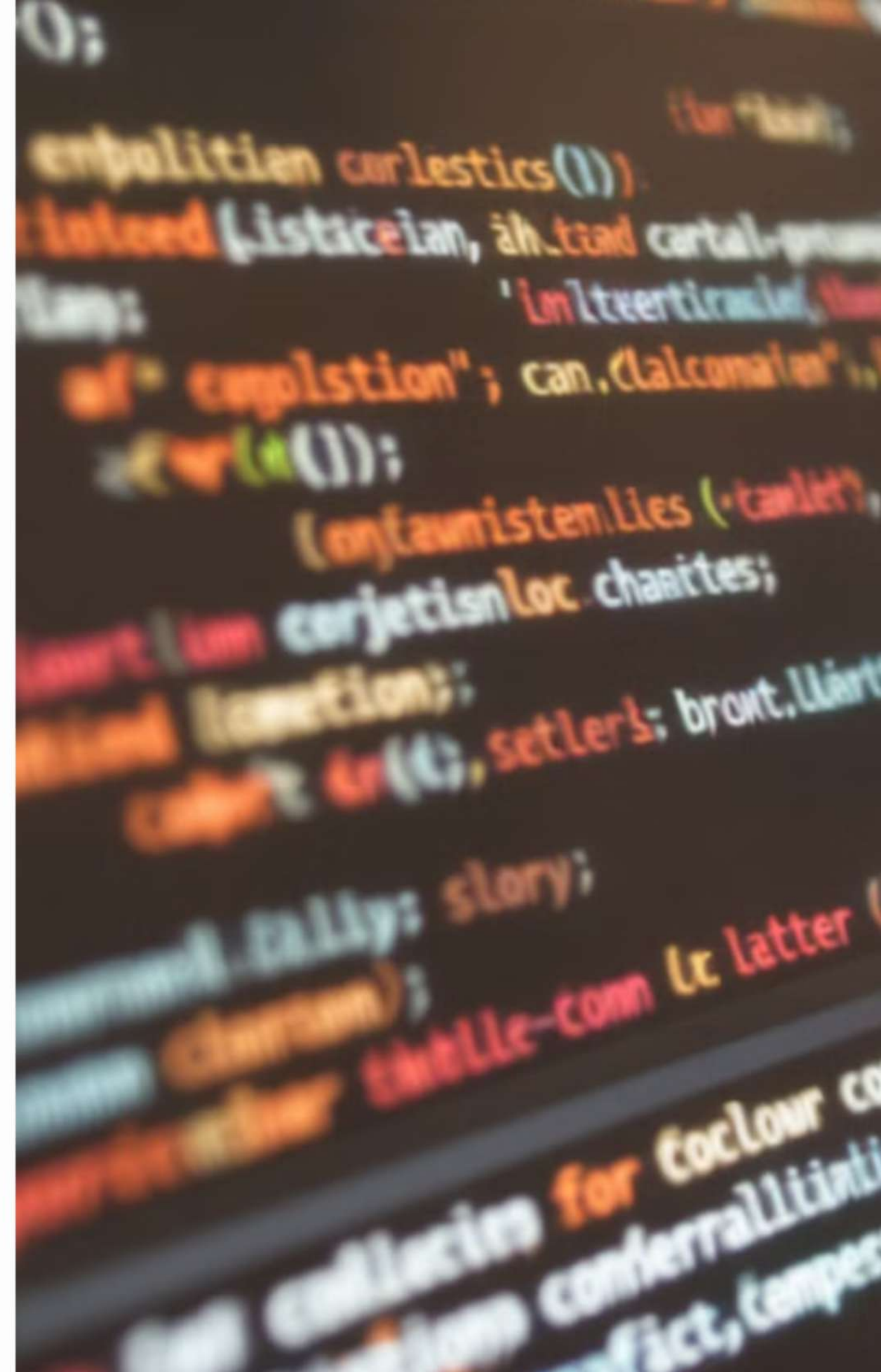
To get started use this link to access coding materials

[https://colab.research.google.com/drive/1hAxI3eJIfuZ2OtGpsdapYDotFQ35\\_-qc](https://colab.research.google.com/drive/1hAxI3eJIfuZ2OtGpsdapYDotFQ35_-qc)

Download data files from here:

[https://hkuhk-my.sharepoint.com/:f:/g/personal/wukevin\\_hku\\_hk/EraHOYXvPLJAK-0L4FSeaAUBRI1uTHFyDsvG4CYRODaI\\_w?e=Zi9WFs](https://hkuhk-my.sharepoint.com/:f:/g/personal/wukevin_hku_hk/EraHOYXvPLJAK-0L4FSeaAUBRI1uTHFyDsvG4CYRODaI_w?e=Zi9WFs)

Openrouter: <https://openrouter.ai>





**Q&A**