

# RETRIEVAL AUGMENTED GENERATION (RAG)

---

Workshop  
January 2025



Jason Haley  
Independent Consultant  
[@haleyjason](#)

## Questions for you

- Do you know what Retrieval Augmented Generation (RAG) is?
- Have you used a RAG application yet?
  - BTW: ChatGPT does a good job with RAG these days
- How about built one?
- How many people have used Azure SQL?

# Phases of LLM Usage

01

General use and awareness

02

Combining LLM with external data and APIs

03

Agents or multi-step workflows with LLMs and other sources

04

Tools or systems that provide actual business value

# Using LLMs in Applications

**Chatbot** – Simple back and forth messaging between system and user

**RAG** – Chat with your own data ← This is what we will explore

**Copilot** – Assistants that work with you to complete tasks

**Agents** – Minimal human intervention to follow instructions and perform tasks autonomously

**Other ways LLMs are used:**

Data generation, summarization, classification, writing drafts, and many more!

# Retrieval Augmented Generation Pattern

## 3 Steps

- Do a search in your retrieval system
- Augment the prompt with search results
- Pass to LLM



# Use Cases for RAG

---

Customer support and help desks

---

Legal Document Analysis

---

Healthcare Information Systems

---

Education and e-Learning

---

Enterprise Knowledge Management

---

Research Assistance

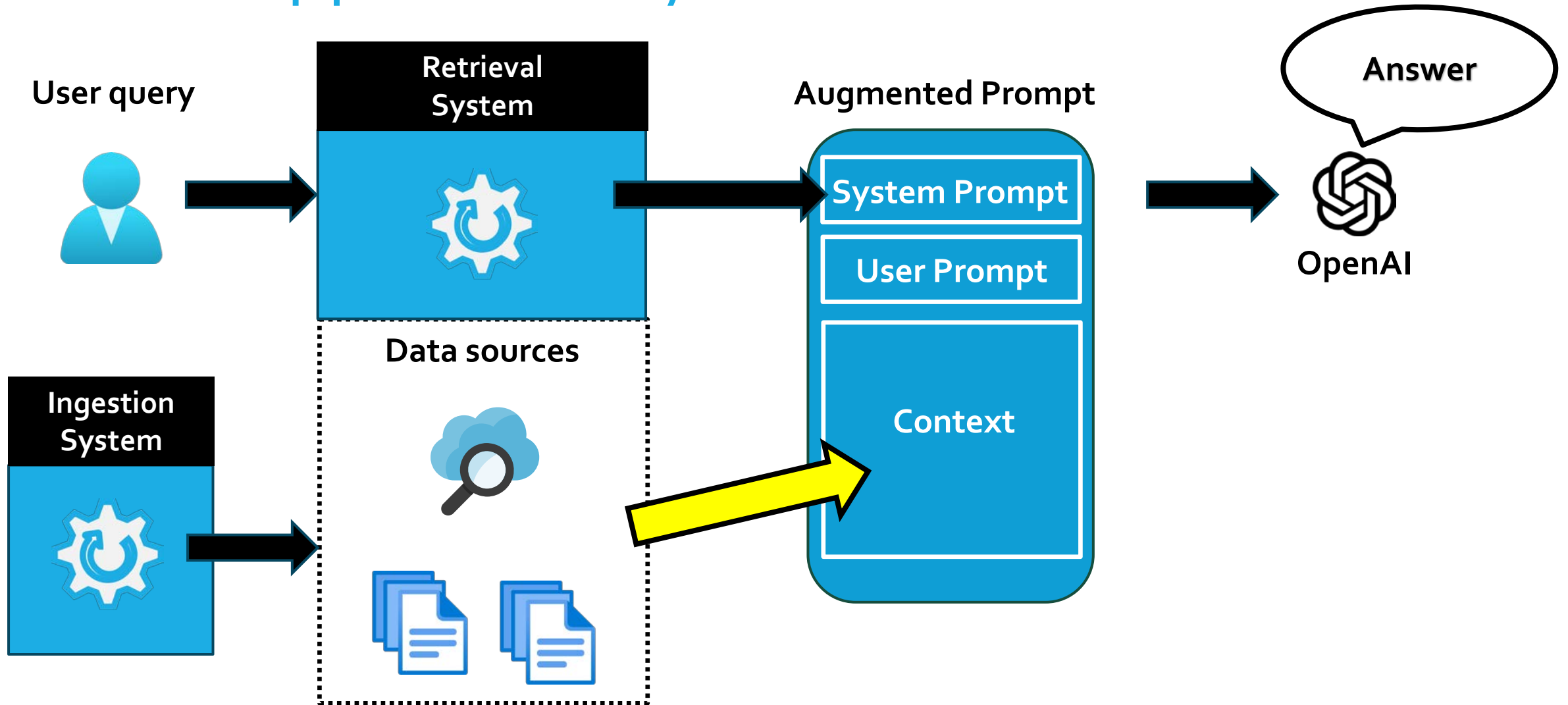
---

Content Generation

---

Many more ...

# RAG Application System Flow



# Retrieval System



Pre-step of searching **your data sources** before calling LLM



Goal is to provide knowledge for LLM to use



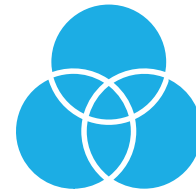
# Searching Types



Keyword



Semantic or Similarity

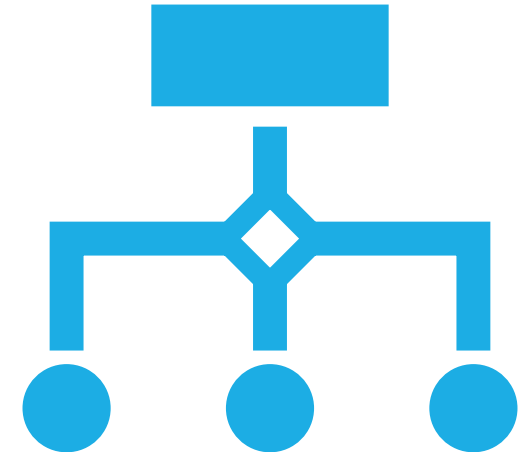


Hybrid (with reranking)

# LAB 3: RAG WITH WEB SEARCH

# Ingestion/Indexing System

- Data quality is directly related to how useful your system will be
- Are you starting with structured or unstructured data?
  - **Structured** – content is already in a database (relational, document db, graph db)
  - **Unstructured** – data is not organized, such as documents
- How will the data be updated? How often?
- What metadata will be helpful in searching?



# Embeddings

## Represent semantically similar items

- Vectors (arrays) of numbers
- Common distance metrics
  - Cosine Similarity
  - Euclidean Distance
  - Dot Product

## Types

- Word embeddings
- Sentence or document embeddings
- Image embeddings
- Audio and video embeddings
- Knowledge graph embeddings

Hugging Face Leaderboard: <https://huggingface.co/spaces/mteb/leaderboard>

# Vector Databases

- Azure SQL (public preview)
- Azure AI Search
- Cosmos DB
- Pinecone
- Chroma
- Qdrant
- Many others...





# Chunking and Its Challenges

- Process of breaking something up into smaller parts (ie. files or text)
- Keep mind: Embeddings encode a meaning for a given chunk of text
- What does that mean for blocks of text?
  - Character count
    - Overlap (10 – 20 %)
  - Paragraphs
  - Pages
  - Other
- What about tables of text?
- What about images or charts?

## Raw text

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and the sixth-biggest metropolitan region by GDP in the European Union.

Google

Tokens

Characters

155

756

Berlin is the capital and largest city of Germany, both by area and population. With over 3.85 million inhabitants, it has the highest population within its city limits of any city in the European Union. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and the fifth-biggest metropolitan region by GDP in the European Union.

Text

Token IDs

<https://en.wikipedia.org/wiki/Berlin>



# 100 characters split

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and the sixth-biggest metropolitan region by GDP in the European Union.

# 100 characters split, with 20 of overlap

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and the sixth-biggest metropolitan region by GDP in the European Union.



# Chunking by sentence

Berlin is the capital and largest city of Germany, both by area and by population. Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits. The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany. The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and the sixth-biggest metropolitan region by GDP in the European Union.

Google

Chunking	Size	Overlap	Recall	Precision	Precision <sub>Ω</sub>	IoU
Recursive	800 (~661)	400	85.4 ± 34.9	1.5 ± 1.3	6.7 ± 5.2	1.5 ± 1.3
TokenText	800	400	87.9 ± 31.7	1.4 ± 1.1	4.7 ± 3.1	1.4 ± 1.1
Recursive	400 (~312)	200	88.1 ± 31.6	3.3 ± 2.7	13.9 ± 10.4	3.3 ± 2.7
TokenText	400	200	88.6 ± 29.7	2.7 ± 2.2	8.4 ± 5.1	2.7 ± 2.2
Recursive	400 (~276)	0	89.5 ± 29.7	3.6 ± 3.2	17.7 ± 14.0	3.6 ± 3.2
TokenText	400	0	89.2 ± 29.2	2.7 ± 2.2	12.5 ± 8.1	2.7 ± 2.2
Recursive	200 (~137)	0	88.1 ± 30.1	7.0 ± 5.6	29.9 ± 18.4	6.9 ± 5.6
TokenText	200	0	87.0 ± 30.8	5.2 ± 4.1	21.0 ± 11.9	5.1 ± 4.1
Kamradt	N/A (~660)	0	83.6 ± 36.8	1.5 ± 1.6	7.4 ± 10.2	1.5 ± 1.6
★ KamradtMod	300 (~397)	0	87.1 ± 31.9	2.1 ± 2.0	10.5 ± 12.3	2.1 ± 2.0
★ Cluster	400 (~182)	0	91.3 ± 25.4	4.5 ± 3.4	20.7 ± 14.5	4.5 ± 3.4
★ Cluster	200 (~103)	0	87.3 ± 29.8	8.0 ± 6.0	34.0 ± 19.7	8.0 ± 6.0
★ LLM	N/A (~240)	0	91.9 ± 26.5	3.9 ± 3.2	19.9 ± 16.3	3.9 ± 3.2

<https://research.trychroma.com/evaluating-chunking>

# Challenges – not all questions are easy

Type	Description	
Yes/No	Answer is Yes or No	Has Lady Gaga ever made a song with Ariana Grande?
Comparative	Compare 2 objects on an attribute	Is Mont Blanc taller than Mount Rainier?
Generic	Simple question	Where was Michael Phelps born?
Intersection	Must fulfill 2 or more conditions	Which movie was directed by Denis Villeneuve and stars Timothee Chalamet?
Ordinal	Based on item in ordered list	Who was the last Ptolemaic ruler of Egypt?
Count	Answer required counting	How many astronauts have been elected to Congress?
Difference	Question with a negation	Which Mario Kart game did Yoshi not appear in?
Superlative	Max/Min of an attribute	Who was the youngest tribute in the Hunger Games?
Multi-hop	Requires 2 or more steps to answer	Who was the quarterback of the team that won Super Bowl 50?

<https://github.com/amazon-science/mintaka>

# Resources

## 5 Levels Of Text Splitting

- [https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/tutorials/LevelsOfTextSplitting/5\\_Levels\\_Of\\_Text\\_Splitting.ipynb](https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/tutorials/LevelsOfTextSplitting/5_Levels_Of_Text_Splitting.ipynb)

## Chunking - Optimizing Vector Database Data Preparation

- <https://research.trychroma.com/evaluating-chunking>

## The BEST Way to Chunk Text for RAG

- <https://www.youtube.com/watch?v=Pk2BeaGbcTE&t=2s>

**LAB 4: RAG WITH PDF**

**LAB 5: COMBINE WEB AND PDF**