RETRIEVAL AUGMENTED GENERATION (RAG)

Workshop

December 2024

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 that provide actual business value

Using LLMs in an Application

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 out instructions and perform tasks autonomously

Other ways LLMs are used:
Data generation, summarization, classification, writing drafts, and many more!

Retrieval Augmented Generation Pattern

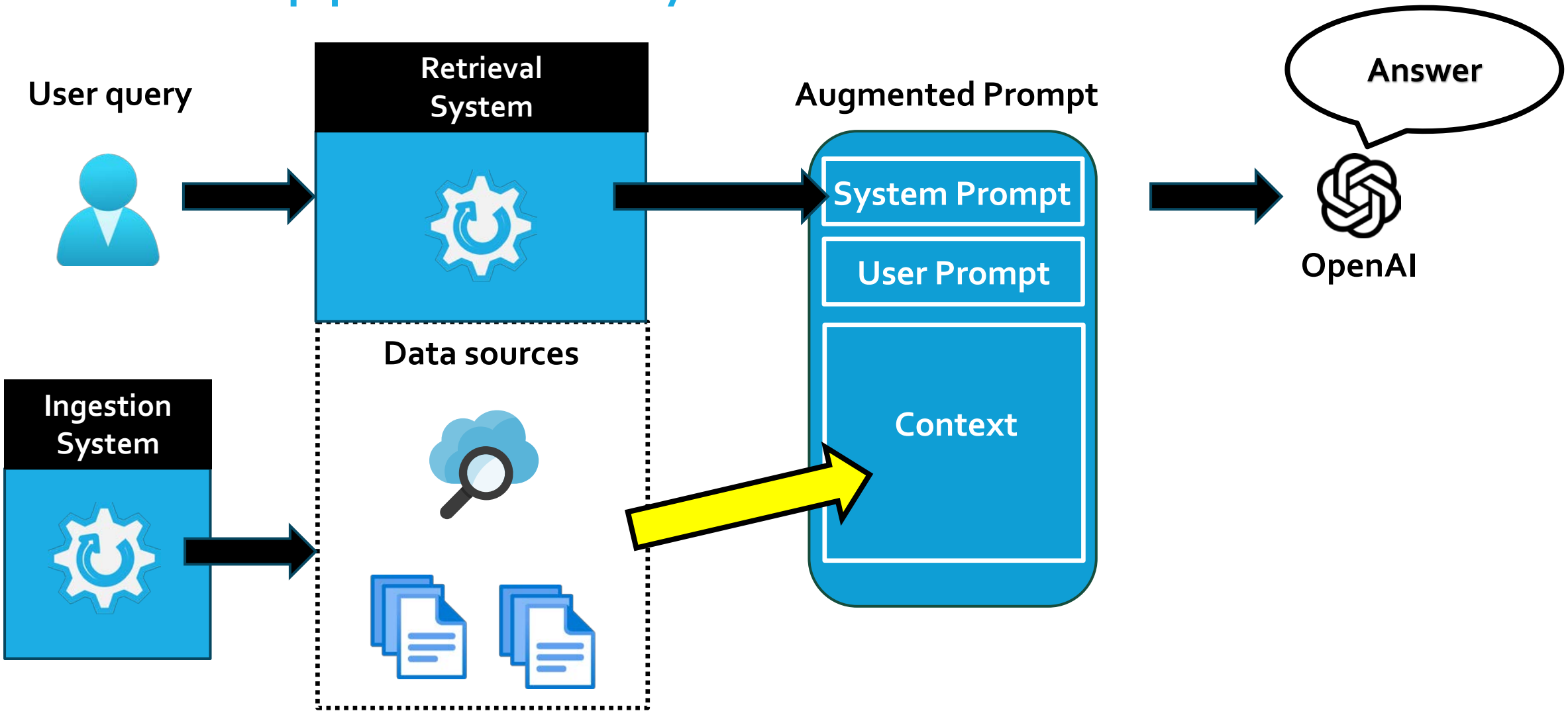
3 Steps

- 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

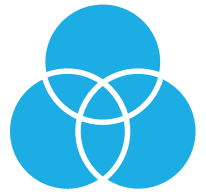
Pre-step of searching your data sources before calling LLM

Goal is to provide knowledge for LLM to use

Searching Types

Keyword

Semantic or Similarity

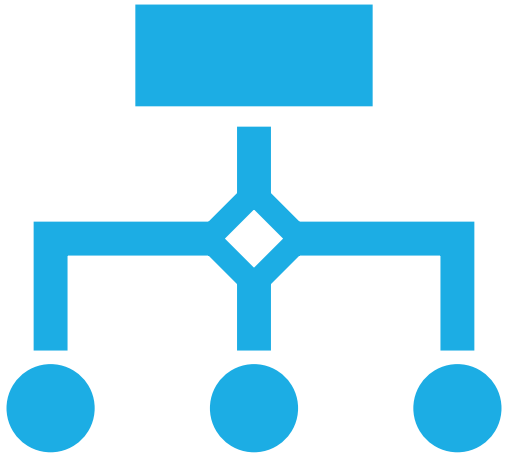


Hybrid (with reranking)

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?
- 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

RAG: from dumb implementation to serious results by Guillaume Laforge **https://bit.ly/3ZHhcPK**

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.

Google

RAG: from dumb implementation to serious results by Guillaume Laforge **https://bit.ly/3ZHhcPK**

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.

Google

RAG: from dumb implementation to serious results by Guillaume Laforge **https://bit.ly/3ZHhcPK**

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.

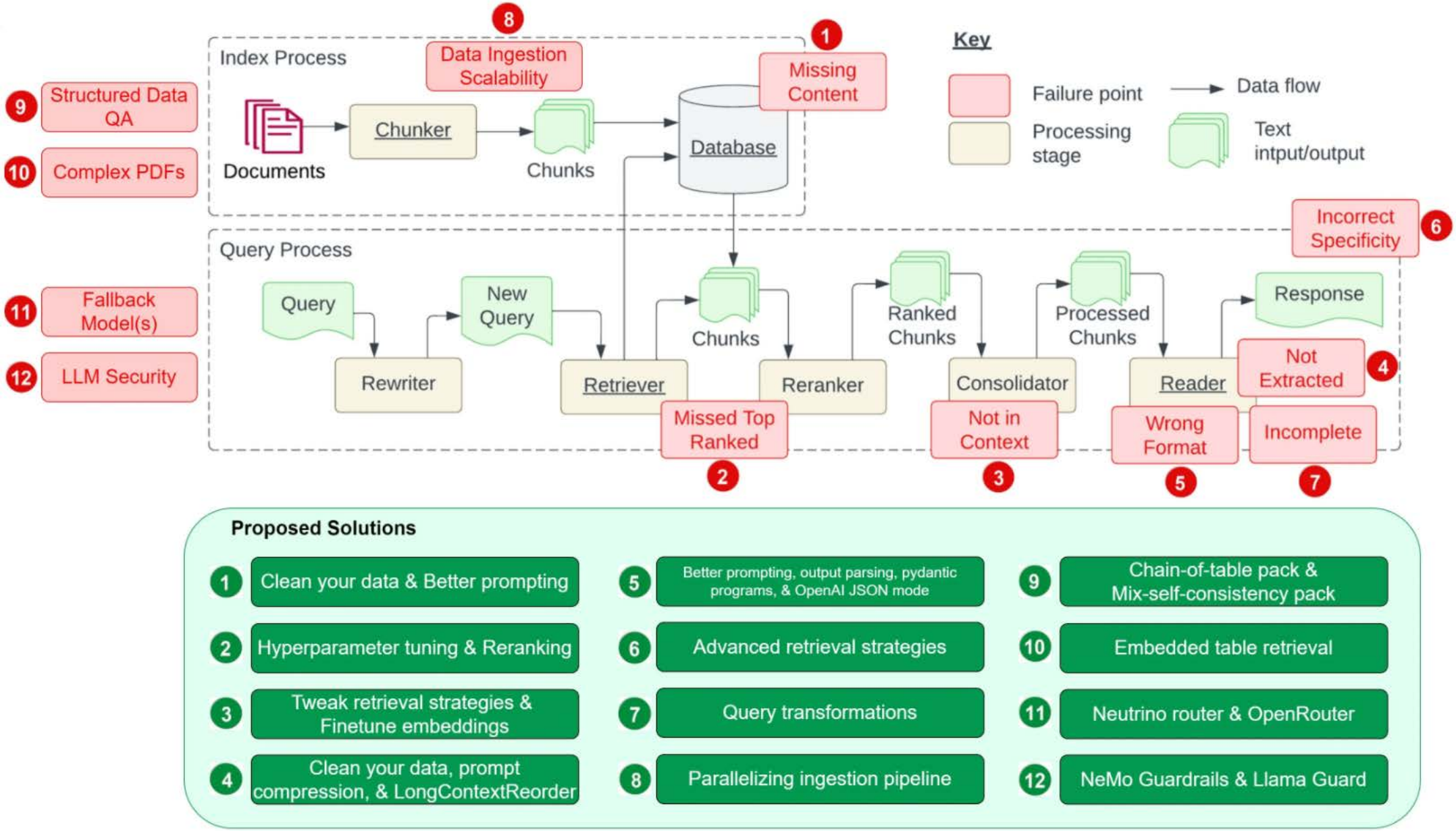
Google

RAG: from dumb implementation to serious results by Guillaume Laforge **https://bit.ly/3ZHhcPK**

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



12 RAG Pain Points and Proposed Solutions, Wenqi Glantz https://bit.ly/3CRASHG

LAB 4: RAG WITH PDF

LAB 5: COMBINE WEB AND PDF