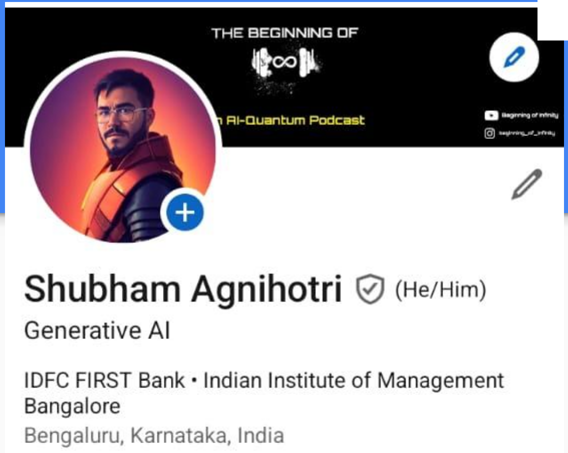
Gemma’s Insights

FROM RETRIEVAL TO GENERATION

Shubham Agnihotri

- 5+ Years of Experience in AI & ML
- Ex Organizer to TFUG Bangalore, hosted 14+ events
- Ex Founder of a Startup and a Coding community group.
- Host at TensorFlow Roadshow, Speaker at Tech Show London, Google DevFest, AWS Community Day, and alot more…
- When I am not coding… I am lost in the Random Forests of Himalayas



Shubham Agnihotri (He/Him)

Generative AI

IDFC FIRST Bank • Indian Institute of Management Bangalore

Bengaluru, Karnataka, India

bit.ly/shubhamagnihotri

RAG

- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

RAG

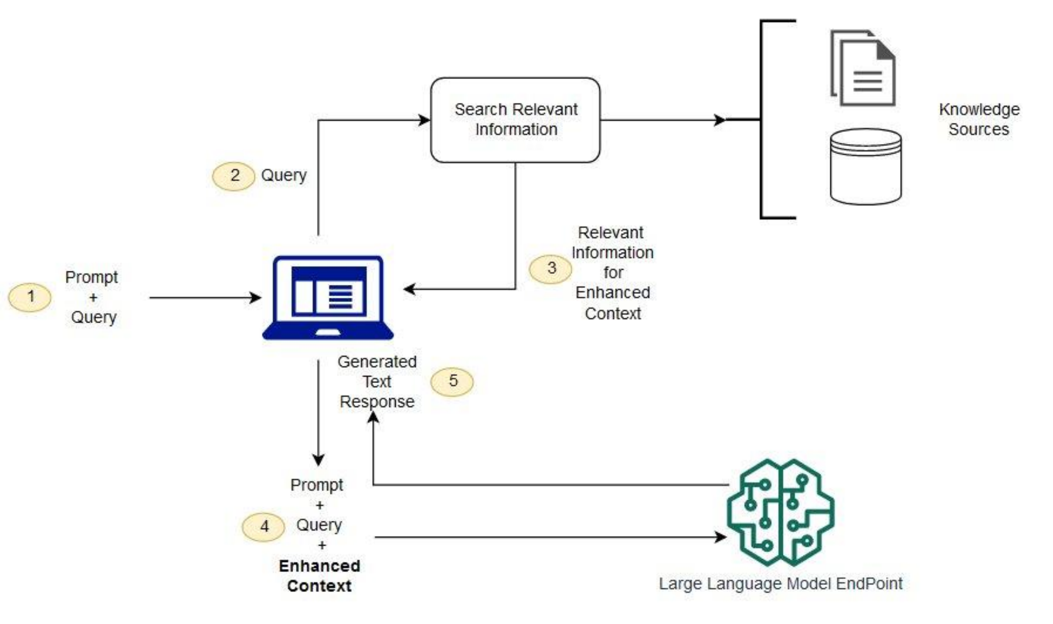
- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

What is RAG?

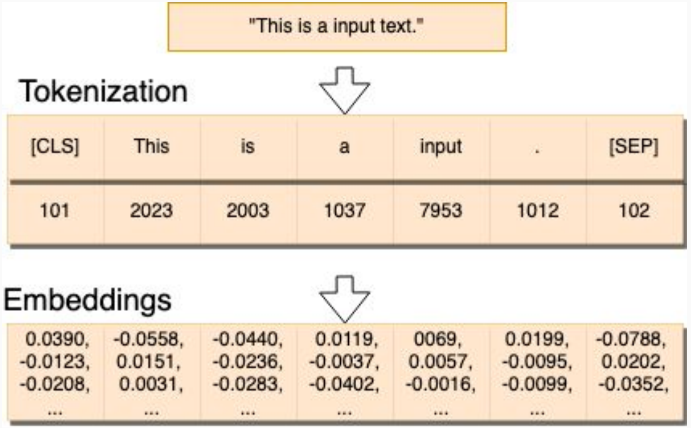
A Process Optimizing outputs of LLMs so it references knowledge base outside training data sources

What is RAG?

A Process Optimizing outputs of LLMs so it references knowledge base outside training data sources


1
Prompt
+
Query
2
Query
Search Relevant Information
Knowledge Sources
Relevant Information for Enhanced Context
3
Generated Text Response
5
Prompt
+
Query
+
Enhanced Context
4
Large Language Model EndPoint

Input Embedding



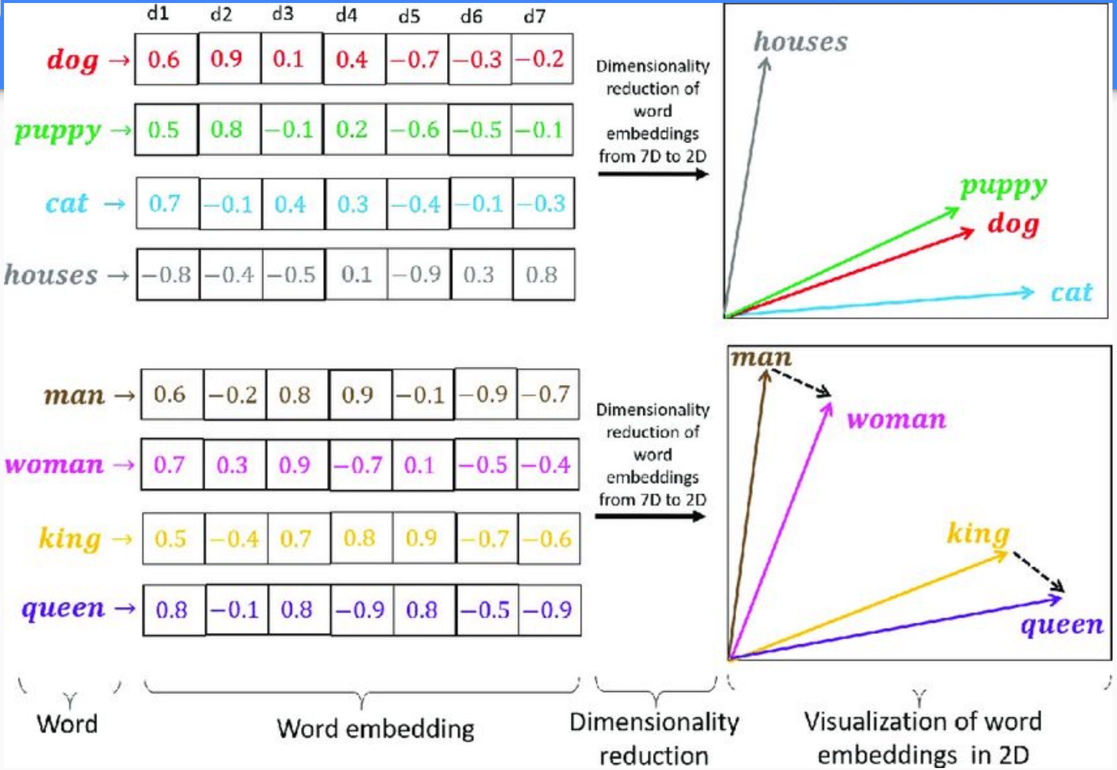
"This is a input text."

Tokenization

[CLS]	This	is	a	input	.	[SEP]
101	2023	2003	1037	7953	1012	102

Embeddings

0.0390, -0.0123, -0.0208, ...	-0.0558, 0.0151, 0.0031, ...	-0.0440, -0.0236, -0.0283, ...	0.0119, -0.0037, -0.0402, ...	0069, 0.0057, -0.0016, ...	0.0199, -0.0095, -0.0099, ...	-0.0788, 0.0202, -0.0352, ...



	d1	d2	d3	d4	d5	d6	d7
dog →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
puppy →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
cat →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D

Dimensionality reduction of word embeddings from 7D to 2D

RAG

- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

RAG

- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

Why RAG?

Why RAG?

- Cost Effective Implementation
  - Don’t need to train or retrain the model
  - Domain Specific information rich replies

Why RAG?

- Cost Effective Implementation
  - Don’t need to train or retrain the model
  - Domain Specific information rich replies
- Up to date with latest information
  - Any new update can be added in few seconds.

Why RAG?

- Cost Effective Implementation
  - Don’t need to train or retrain the model
  - Domain Specific information rich replies
- Up to date with latest information
  - Any new update can be added in few seconds.
- Enhanced User Trust
  - Get Source of the data

Why RAG?

- Cost Effective Implementation
  - Don’t need to train or retrain the model
  - Domain Specific information rich replies
- Up to date with latest information
  - Any new update can be added in few seconds.
- Enhanced User Trust
  - Get Source of the data
- More Developer Control
  - Easier to Debug
  - Easier to Fix outputs

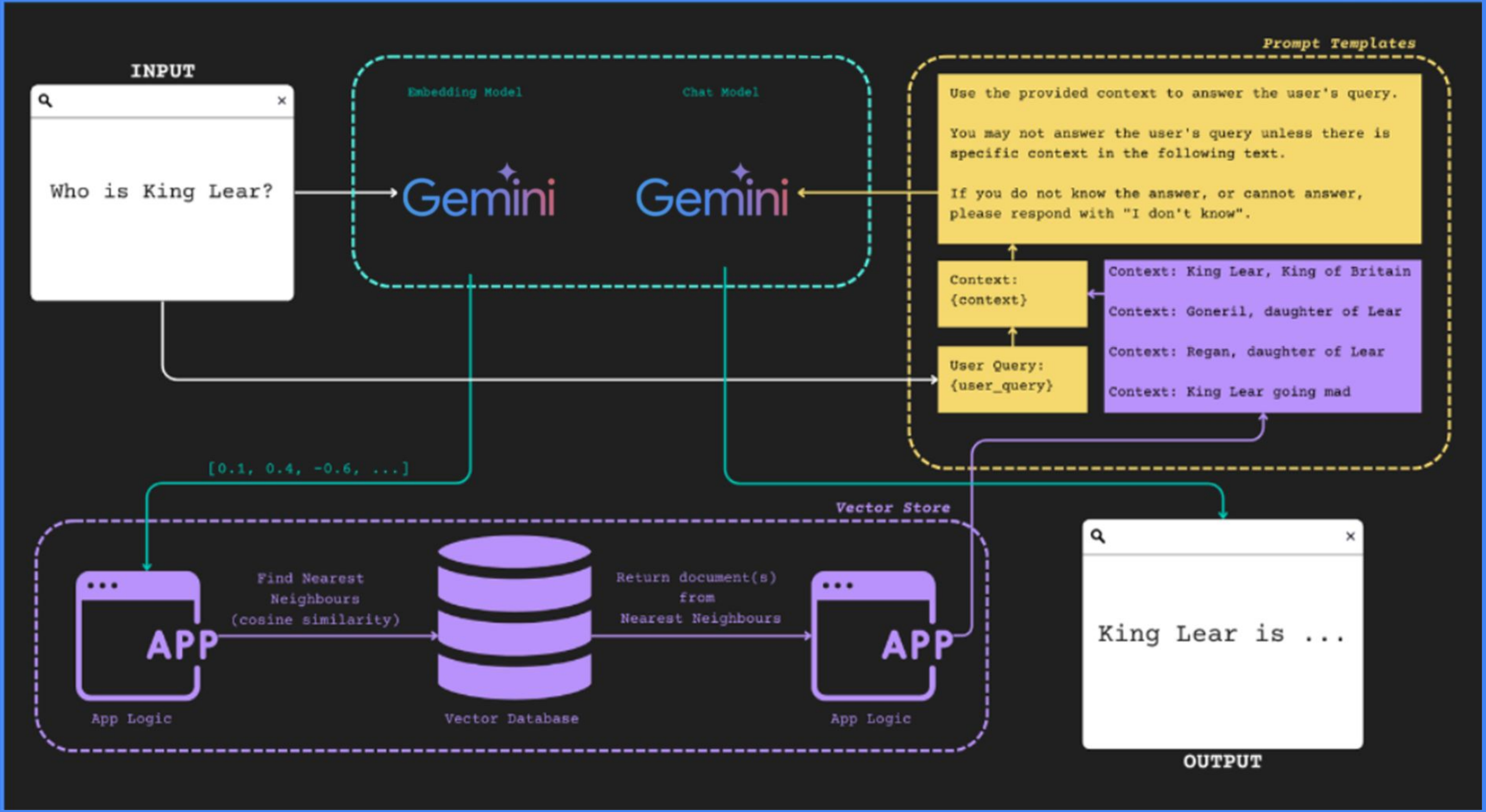
RAG

- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

RAG

- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

How does RAG functions technically?



Use the provided context to answer the user's query.

You may not answer the user's query unless there is specific context in the following text.

If you do not know the answer, or cannot answer, please respond with "I don't know".

Context:
{context}

User Query:
{user_query}

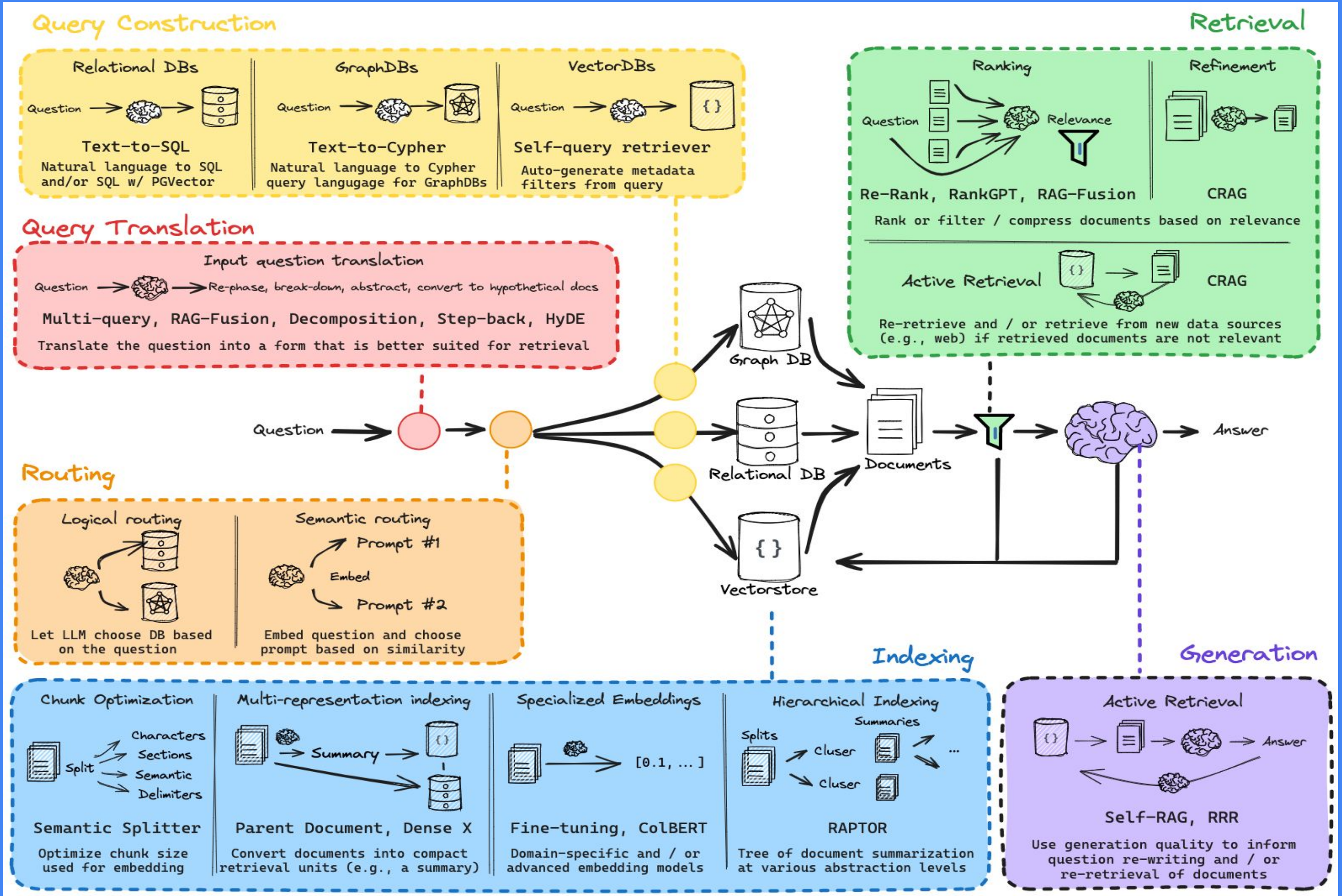
Context: King Lear, King of Britain

Context: Goneril, daughter of Lear

Context: Regan, daughter of Lear

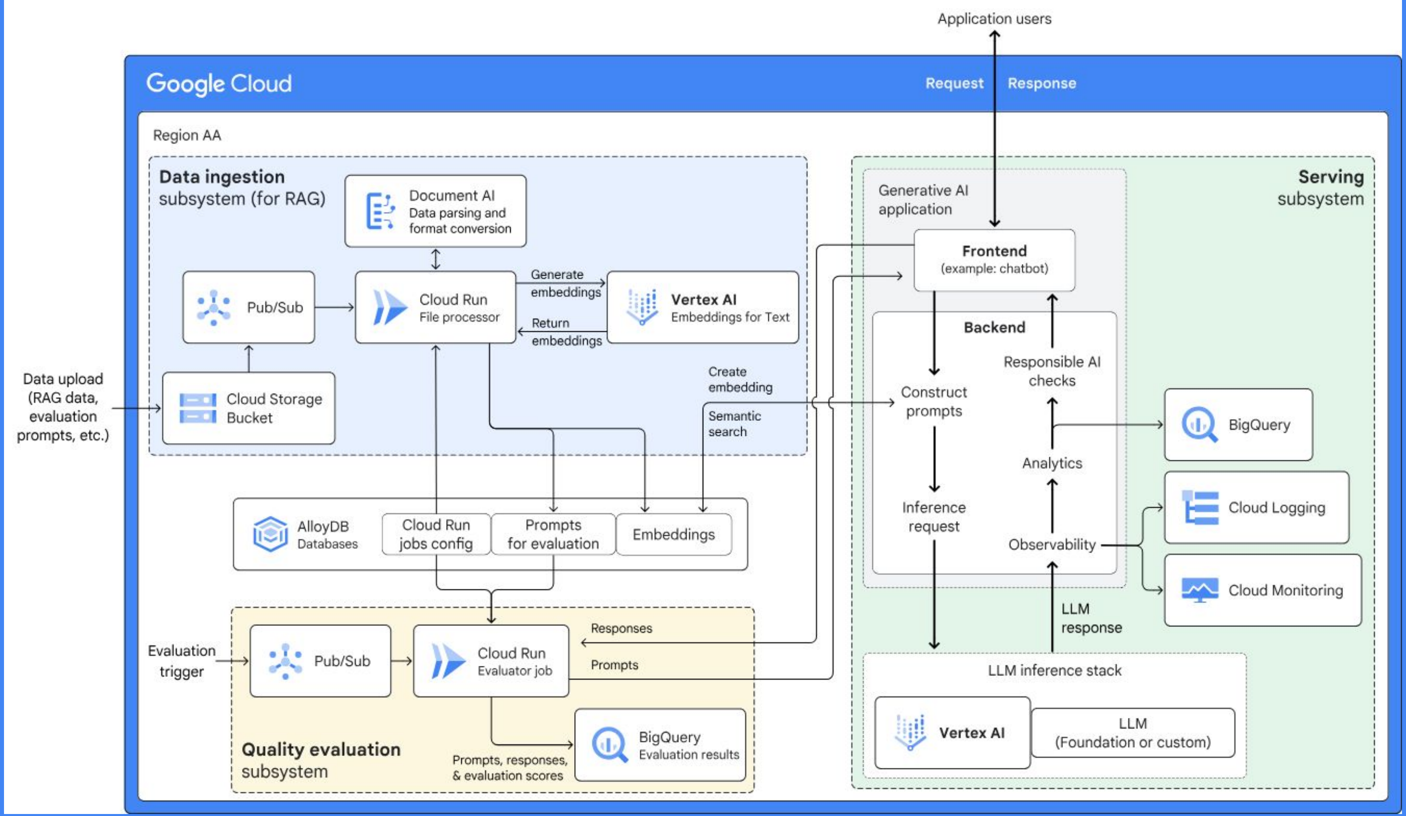
Context: King Lear going mad

Source

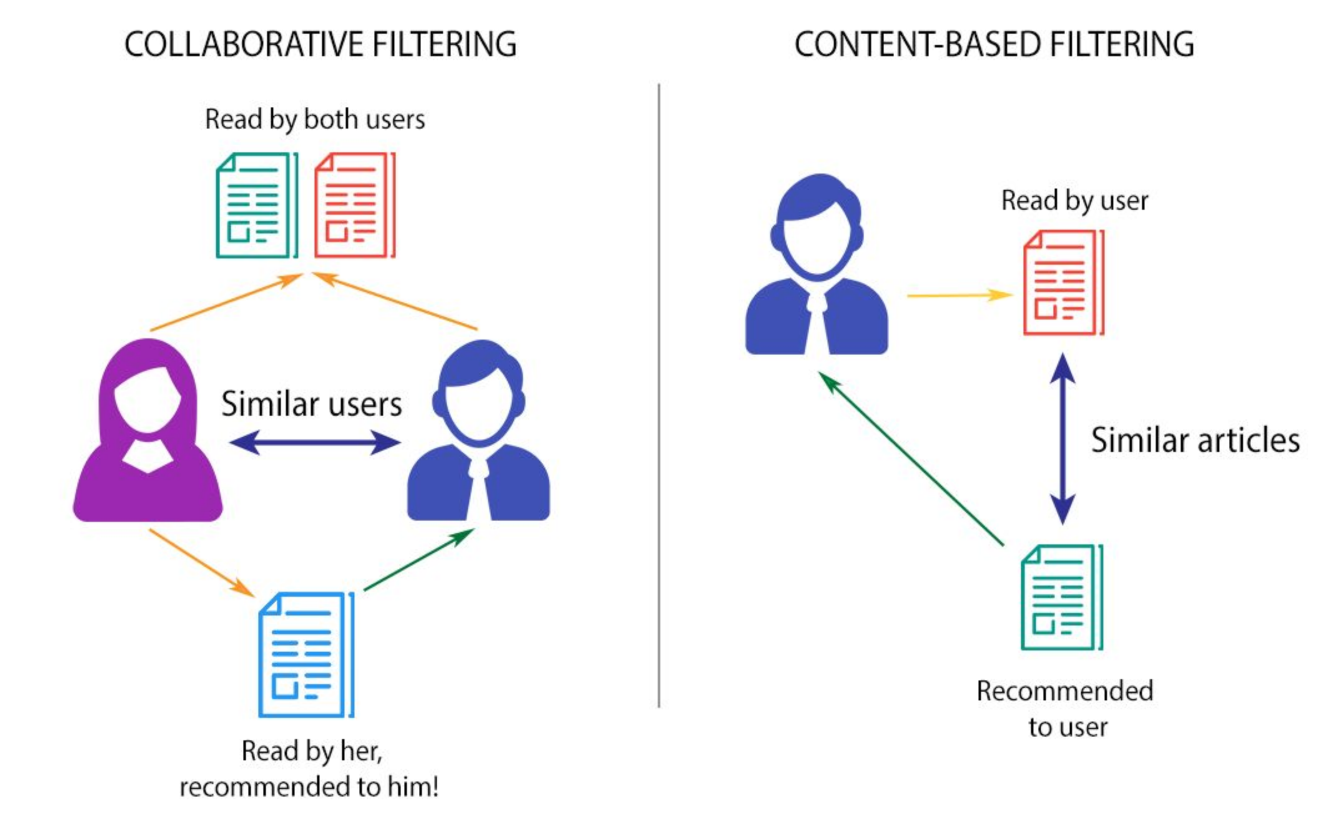

Query Construction
Relational DBs
Question
Text-to-SQL
Natural language to SQL and/or SQL w/ PGVector
GraphDBs
Question
Text-to-Cypher
Natural language to Cypher query langugage for GraphDBs
VectorDBs
Question
{ }
Self-query retriever
Auto-generate metadata filters from query
Retrieval
Ranking
Question
Relevance
Re-Rank, RankGPT, RAG-Fusion
Refinement
CRAG
Rank or filter / compress documents based on relevance
Active Retrieval
{ }
CRAG
Re-retrieve and / or retrieve from new data sources (e.g., web) if retrieved documents are not relevant
Query Translation
Input question translation
Question
Re-phase, break-down, abstract, convert to hypothetical docs
Multi-query, RAG-Fusion, Decomposition, Step-back, HyDE
Translate the question into a form that is better suited for retrieval
Graph DB
Question
Relational DB
Documents
Answer
{ }
Vectorstore
Routing
Logical routing
Let LLM choose DB based on the question
Semantic routing
Prompt #1
Embed
Prompt #2
Embed question and choose prompt based on similarity
Indexing
Chunk Optimization
Split
Characters
Sections
Semantic
Delimiters
Semantic Splitter
Optimize chunk size used for embedding
Multi-representation indexing
Summary
{ }
Parent Document, Dense X
Convert documents into compact retrieval units (e.g., a summary)
Specialized Embeddings
[0.1, ...]
Fine-tuning, ColBERT
Domain-specific and / or advanced embedding models
Hierarchical Indexing
Summaries
Splits
Cluser
Cluser
...
RAPTOR
Tree of document summarization at various abstraction levels
Generation
Active Retrieval
{ }
Answer
Self-RAG, RRR
Use generation quality to inform question re-writing and / or re-retrieval of documents

Source

How does RAG functions technically?



Recommendation system



RAG

- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

RAG

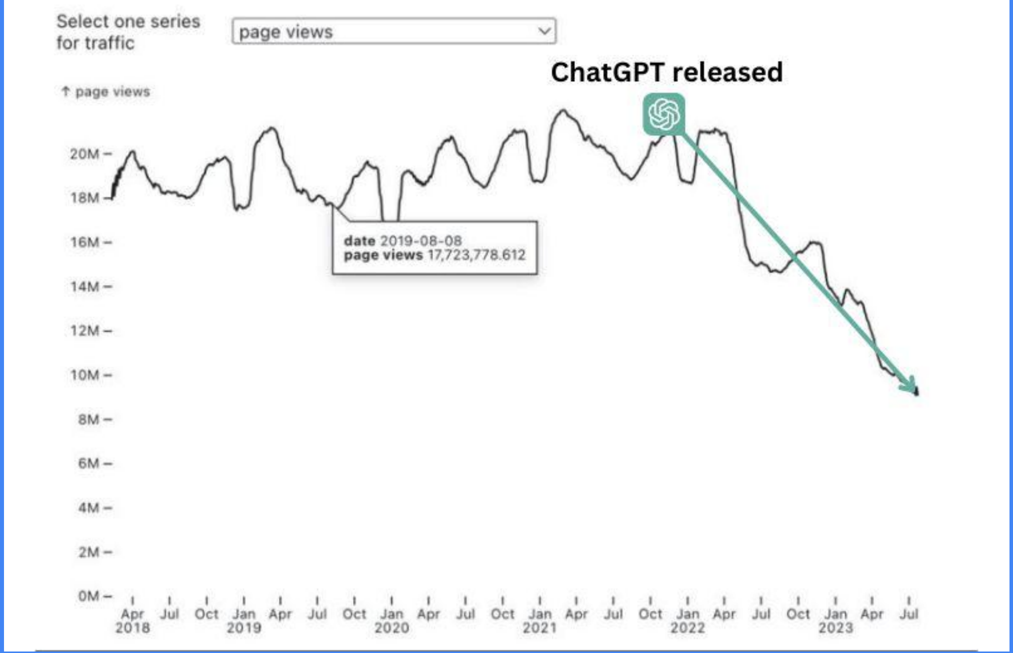
- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

How to Implement RAG?

RAG

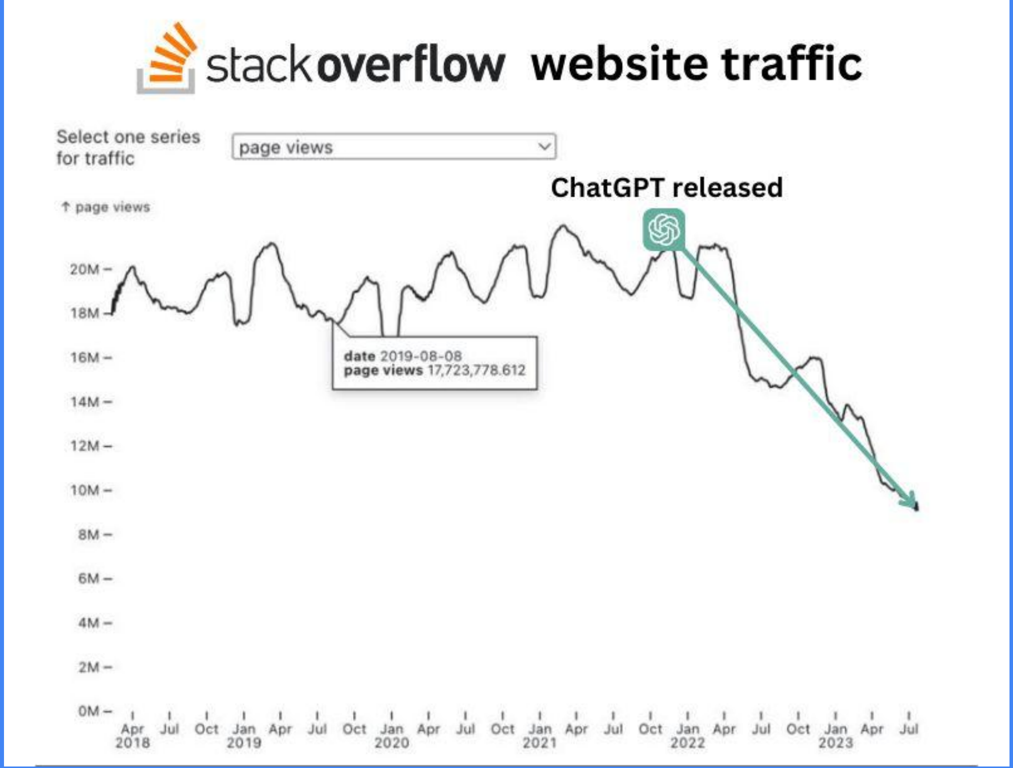
- What is RAG?
- Why RAG?
- How does RAG functions technically?
- How to Implement RAG?
- Usecase

Use Case



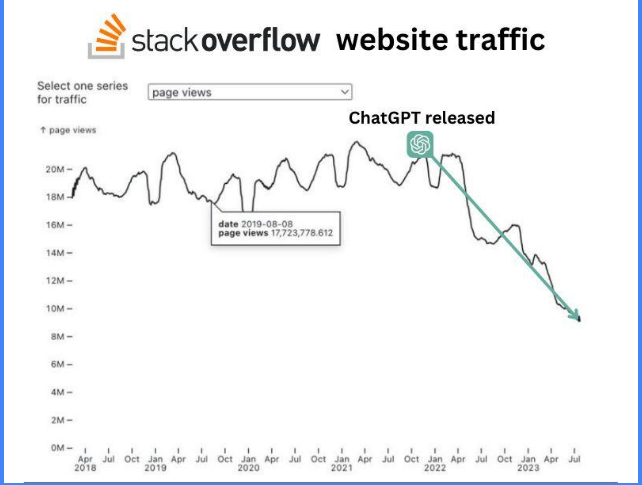
date 2019-08-08
page views 17,723,778.612

Use Case



stackoverflow website traffic
Select one series for traffic
page views
ChatGPT released
↑ page views
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
date 2019-08-08
page views 17,723,778.612
Apr 2018
Jul
Oct
Jan 2019
Apr
Jul
Oct
Jan 2020
Apr
Jul
Oct
Jan 2021
Apr
Jul
Oct
Jan 2022
Apr
Jul
Oct
Jan 2023
Apr
Jul

Usecase


stackoverflow website traffic
Select one series for traffic
page views
↑ page views
ChatGPT released
date 2019-08-08
page views 17,723,778.612
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
Apr 2018
Jul
Oct
Jan 2019
Apr
Jul
Oct
Jan 2020
Apr
Jul
Oct
Jan 2021
Apr
Jul
Oct
Jan 2022
Apr
Jul
Oct
Jan 2023
Apr
Jul

overflow AI

Thanks!

Shubham Agnihotri

Senior Manager - Generative AI

Shubham Agnihotri (He/Him)

Generative AI

IDFC FIRST Bank • Indian Institute of Management Bangalore

Bengaluru, Karnataka, India

bit.ly/shubhamagnihotri