

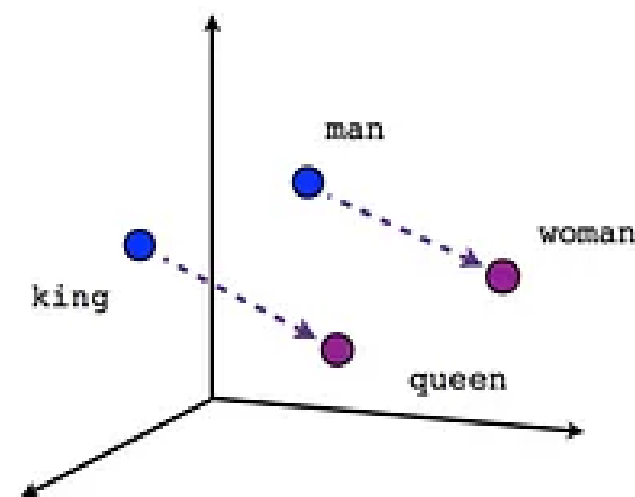
Building Powerful RAG Systems with GCP

Jeet Shah

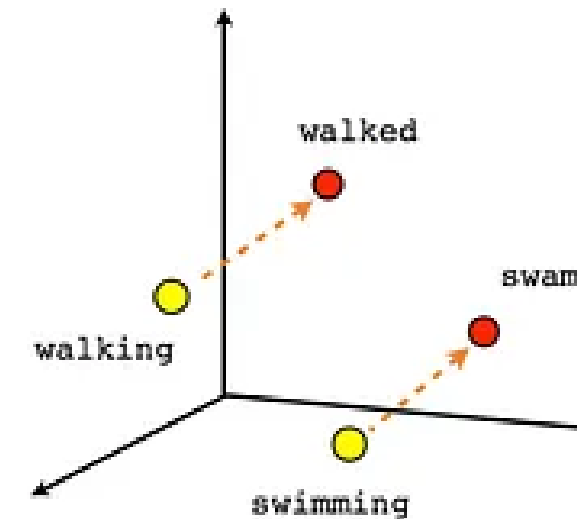
jeet@infocusp.com

Embeddings

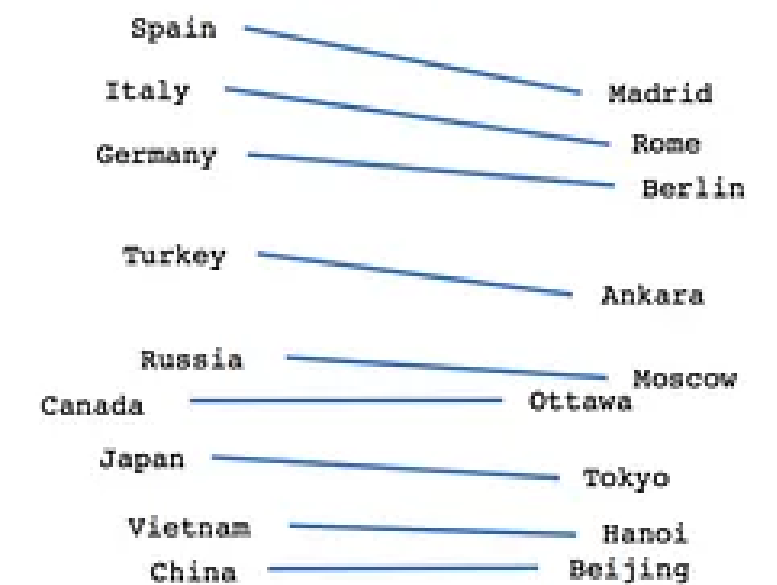
- Encoding sentences as numbers
- Use cases
 - similarity search
 - sentiment analysis
 - clustering
 - classification
 - anomaly detection



Male-Female



Verb tense



Country-Capital

Connections in language captured by embeddings

Image from : [Link](#)

Retrieval Augmented Generation

Generative capabilities of LLMs are limited by their training data, which leads to the following issues:

- ❗ Lack of knowledge
- ❗ Hallucinations
- ❗ Irrelevant answers

What is RAG?

RAG can have several meanings. Here are a few possible interpretations:

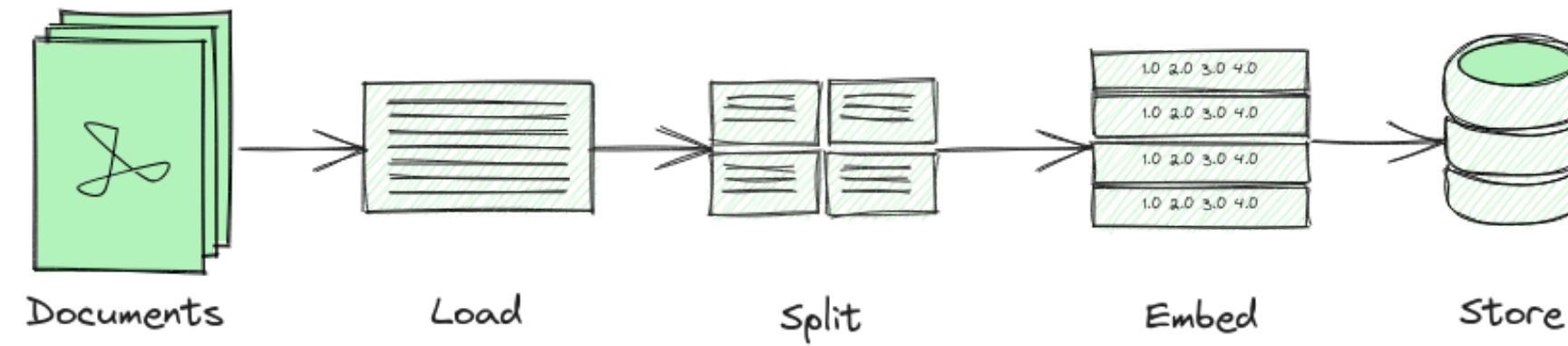
1. **Red, Amber, Green (RAG)**: In project management ...
2. **RAG Rating**: This is a rating system used ...
3. **Recombinase-activating gene (RAG)**: In the context of genetics and immunology ...
4. **RAG Magazine**: There have been various magazines ...

Context: ...
What is RAG?

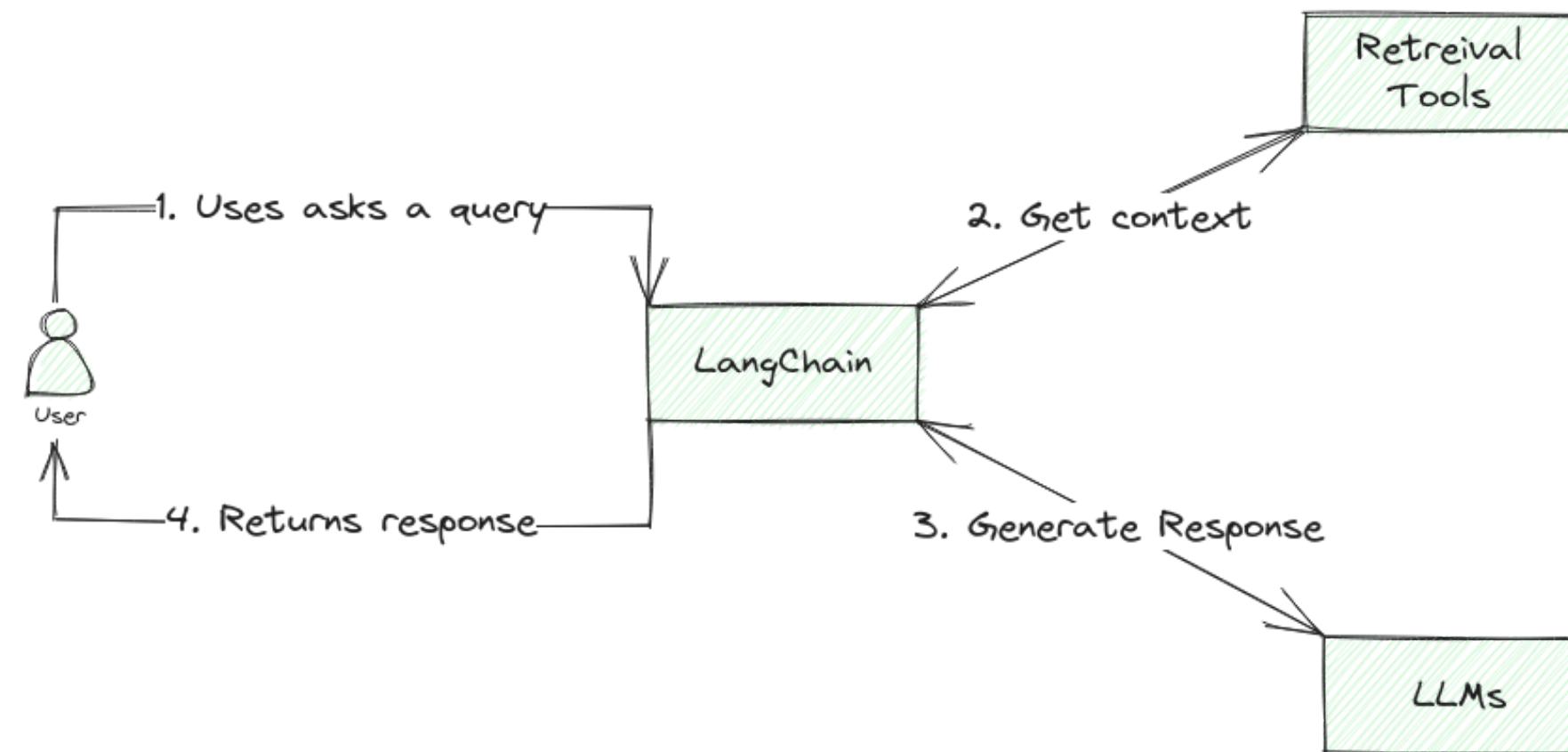
RAG, or Retrieval Augmented Generation, is a method introduced by Meta AI researchers to enhance the performance of language models in knowledge-intensive tasks that require access to external knowledge sources. It is designed to address the limitations of general-purpose language models (LLMs) ...

General RAG Workflow

Indexing Workflow



Generation Workflow



Retrieval Augmented Generation

Query:

How can we load csv data in LangChain?

Embedding:

<1,2,3,4>

Embedding

<1,1,2,4>

Similarity

0.8

CSV

Load CSV data with a single row per document.

```
loader = CSVLoader(file_path='././.csv')
data = loader.load()
```

File Directory

This covers how to load all documents in a directory.

```
loader = DirectoryLoader('../', glob="**/*.md")
docs = loader.load()
```

File Directory

This covers how to load all documents in a directory.

```
loader = UnstructuredHTMLLoader("example.html")
docs = loader.load()
```

Prompt:

How can we load csv data in LangChain?

Context:

CSV

Load CSV data with a single row per document.

```
loader =
CSVLoader(file_path='././.csv')
data = loader.load()
```

Demo

- Search App + Unstructured Data
- Search App + Website Data
- Grounding
- Vector Search

References & Resources

- <https://aman.ai/primers/ai/RAG>
- <https://stackoverflow.blog/2023/10/18/retrieval-augmented-generation-keeping-llms-relevant-and-current/>
- <https://cloud.google.com/vertex-ai/docs/generative-ai>



DELIVERY VIRTUES

→ INNOVATION & RESEARCH

- Building unparalleled product prototypes that are harbingers of innovation in the field
- Patents and publications

→ PRODUCT ENGINEERING













- System architecture
- App development
- UI / UX

→ GROWTH, SCALE, AND SUSTAINABILITY

- Platform scaling and process orchestration
- Maximum efficiency, optimized efficacy with minimum delays
- DevOps

WHAT WE OFFER

DOMAINS WE WORK IN

-  Healthcare
-  FinTech, Banking, and Insurance
-  Wearable Technologies
-  IoT
-  Legal
-  Agriculture and Horticulture
-  Sciences
-  Leisure and Gaming
-  Energy
-  Molecular Chemistry
-  Logistics Supply Chain
-  Geophysics