Build Your Multilingual Support Bot with RAG

Build Your Multilingual

Support Bot with RAG

Multilingual

Support Bot

RAG

[0.12, 0.98, ..., 0.56] [0.80, 0.61, ..., 0.37] [0.74, ..., 0.43]

Search in the Vector DataBase



[9.821, -8.8438, 6.626] [3.1423, -5.5524] [9.681, 0.1253, 6.661] [-7.89, 1.11] [0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33]

Shrijayan Rajendran X - @rshrijayan - ThoughtWorks Al Engineer



Come Inside here

https://shorturl.at/PSWY8



Multilingual Support

Bhashini

What is Bhashimi?

API based, Converting to Indic Languages

Input - Source Language, Target Language, Text

Output - Target Language Text

Bhashini API

Step 1: API Key - https://bhashini.gov.in/ulca/user/login



Sign in to ULCA



New to ULCA ? Create an account



Retrieval-Augmented Generation (RAG)

Terminology to Know

Embedding

Eg: [0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345]

Chunk

Similar to a row in a SQL

[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345]

Articles may include embedded content from other websites, which behaves as if the visitor has visited the other website directly. These sites may collect data about you and use cookies for tracking.

Vector DataBase

n-deminal DataBase



What is AI?

"Artificial intelligence, or AI, is technology that enables computers and machines to **Simulate** human intelligence and problem-solving capabilities."

Traditional AI - Single Task Performer

Generative AI - Multi Task Performer

What is LLM?

A language model is a probabilistic model that assign probabilities to sequence of words.

P ["Shanghai is a city in"] "China"

Prompt

Next Token

We usually train a neural network to predict these probabilities. A neural network trained on a large corpora of text is known as a Large Language Model (LLM)

You are what you eat

A language model can only output text and information that it was trained upon.

This means, that if we train a language model only on English content, very probably it won't be able to output Japanese or French.

To teach new concepts, we need to fine-tune the model.

What is RAG?

Example

What is the return policy of the company named XYZ?

This cannot be answered by a LLM

Solution

Here comes the RAG where we can upload the PDF and answer the Question of the live DATA.

Retrieval Augmented Generation

Why RAG?

- Improves accuracy of large language models (LLMs)
- Leverages external knowledge sources
- Boosts domain-specific performance
- Easier to implement than fine-tuning
- Improves trust and transparency

RAG Architecture



Example

You

Who Won Google Cloud Technology Partner of Year (2023) Award for AI and Machine Learning

ChatGPT

I can't provide real-time information, including award winners beyond my last update in January 2022. To find out who won the Google Cloud Technology Partner of the Year Award for AI and Machine Learning in 2023, you should check the latest announcements from Google or reputable tech news sources. They often cover such awards and recognitions.

007

How to implement RAG?

Part 1 Storing the Data



Load

- Loading the data it will use
- This data serves as the foundation for generating text and can come from various sources:
 - Media File
 - JSON Files
 - URLs

Split / Chunking

- Chunks for efficient processing
- Why not without chunking?
- Then why Chunking?

Embedding

- What embedding captures?
- What will it do?
- Common embedding
 - **TF-IDF**
 - word2vec

Storing

- After generating the embeddings?
- Vector stores are specialized databases designed
- Benefits
 - Fast Retrieval
 - Scalability
 - Search Optimization
- Factors

Simple Visual Implementation

Text from the Document

You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence. Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate! RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.

Chunks

You can think of the LLM as an over-enthusiastic

new employee who refuses to stay informed with

current events but will always answer every

question with absolute confidence.

Unfortunately, such an attitude can negatively

impact user trust and is not something you want

your chatbots to emulate!

RAG is one approach to solving some of these

challenges. It redirects the LLM to retrieve

relevant information from authoritative,

pre-determined knowledge sources.

Chunks

You can think of the LLM as an over-enthusiastic

new employee who refuses to stay informed with

current events but will always answer every

question with absolute confidence.

Unfortunately, such an attitude can negatively

impact user trust and is not something you want

your chatbots to emulate!

RAG is one approach to solving some of these

challenges. It redirects the LLM to retrieve

relevant information from authoritative,

pre-determined knowledge sources.



Embeddings

[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]

[2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33]

[-4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]

Embeddings



Vector DataBase

Part 2 Retrieving the Data



Question

What can I think of the LLM as an over-enthusiastic?

Question Embeddings

[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66]



Vector DataBase

Embedding	Value
[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 1.0, 2.0, 3.0,8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence.
[2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33]	Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate!
[-4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.

Vector DataBase

Embedding	Value
[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 1.0, 2.0, 3.0,8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence.
[2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33]	Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate!
[-4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.

Vector DataBase Results





LLM - LM-Studio

LM-Studio - https://lmstudio.ai/

Reason : Compatible with MAC / LINUX / WINDOWS



Prompting

RRR Prompting

Role: You are a Q&A Chatbot interacting with real-world customers to their questions.

Rule: Only answer based on the content provided. Do not provide any information that is not in the content.

Response: Answer in a interactive way



Fine-Tuning VS RAG



Advance RAG

Techniques

- Text Extraction (Image Understanding, Table Understanding)
- Format Understanding
- Guardrail
- Chunking strategy
- Domain Specific Embedding Model
- High Efficient VectorDB
- Prompting Techniques
- Efficient LLM
- Exception Handling
- RAG Evaluation

