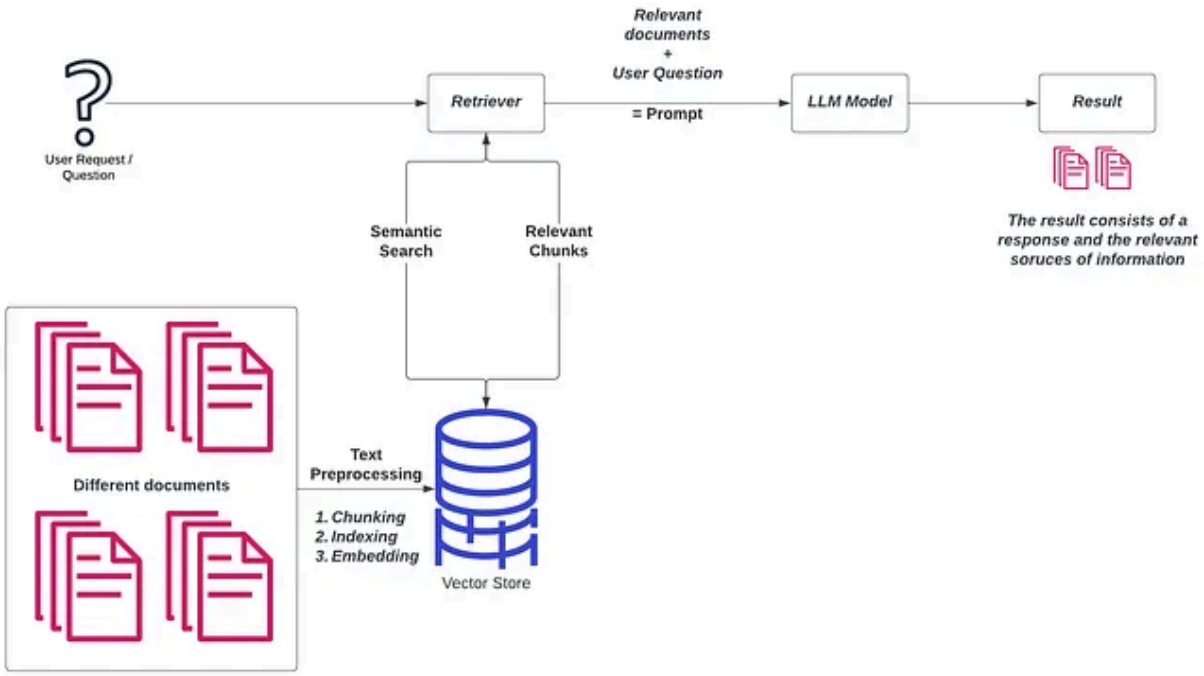
✦ Member-only story

RAG Explained in the Context of LLMs

Revolutionizing AI with Contextual Understanding

Christophe Atten · Follow

Published in AI Advances · 6 min read · Dec 15, 2023



Introducing RAG (Retrieval Augmented Generation) — an innovative AI framework that enhances large language models by retrieving facts from an external knowledge base. By doing so, RAG ensures that LLMs are grounded on the most accurate and up-to-date information.

> *With RAG, users can gain valuable insights into the generative process of LLMs.*

At its core, RAG represents a significant leap in AI's ability to provide contextual information to NLP and GPT processes to provide more accurate responses with references to current LLM, such as OpenAI's original GPT3.5 model, yet was brilliant but had its flaws, such as:

1. **They lack domain-specific knowledge:** LLMs are trained for general tasks and do not have access to your company's private data.
2. **Not up-to-date information:** LLMs have outdated information because it is not feasible to update their vast training datasets.
3. **Black-box:** It's difficult to determine which sources an LLM considered to reach their conclusions.

However, today RAG is already seamlessly integrated into many applications such as ChatGPT Internet search plugins, RAG is present in our lives without knowing it. This makes RAG an intriguing technological advancement and a pivotal tool in the journey towards brilliant AI systems.

I. The Importance of RAG in AI and Generative AI

Large Language Models (LLMs) have gained immense popularity due to their ability to generate responses that are similar to human responses to queries.

However, these models have certain limitations that make them highly unpredictable (hallucination).

The responses generated by LLMs can sometimes be irrelevant to the query, even though they might be accurate at other times. This is because LLMs are based on statistical relationships between words and need help understanding the context and meaning behind them.

To address these limitations, Retrieval-augmented generation (RAG) has been introduced as an Artificial Intelligence system that enhances the LLM's understanding of information by supplementing it with external sources of knowledge. This integration of RAG into LLM-based Q&A systems offers two significant advantages. Firstly, it ensures that consumers can access the model's sources of information, which adds transparency to the process. Secondly, it provides the model with up-to-date and reliable facts, improving the responses' accuracy.

> *RAG is a reliable source of current information from around the world and domain-specific data that can be integrated with your GenAI applications to keep them up-to-date.*

In addition, only some companies, such as Sam Altman, used for ChatGPT, have $ 100 million to train their foundation model.

II. Understanding RAG

We have previously discussed how RAG can provide additional relevant content from your domain-specific database to an LLM at the time of generation. This content is provided through a "context window" accompanying the original prompt or question.

A context window is essentially an LLM's field of vision at a given moment. RAG is like holding up a cue card containing critical points for your LLM to see. This helps it to produce more accurate responses that incorporate essential data.

To understand RAG, it is essential first to understand semantic search. Unlike the traditional approach of matching keywords in a user's query, semantic search attempts to find the true meaning behind the query and retrieve relevant information that better fits the user's intent. The goal is to deliver more accurate results that align with the user's needs.

Additional information about the semantic searches and vector databases can be found here:

The Future of AI in Finance: Vector Databases and Beyond

One possible outcome for the future of AI in Finance: What are Vector Databases and how can they be used

levelup.gitconnected.com

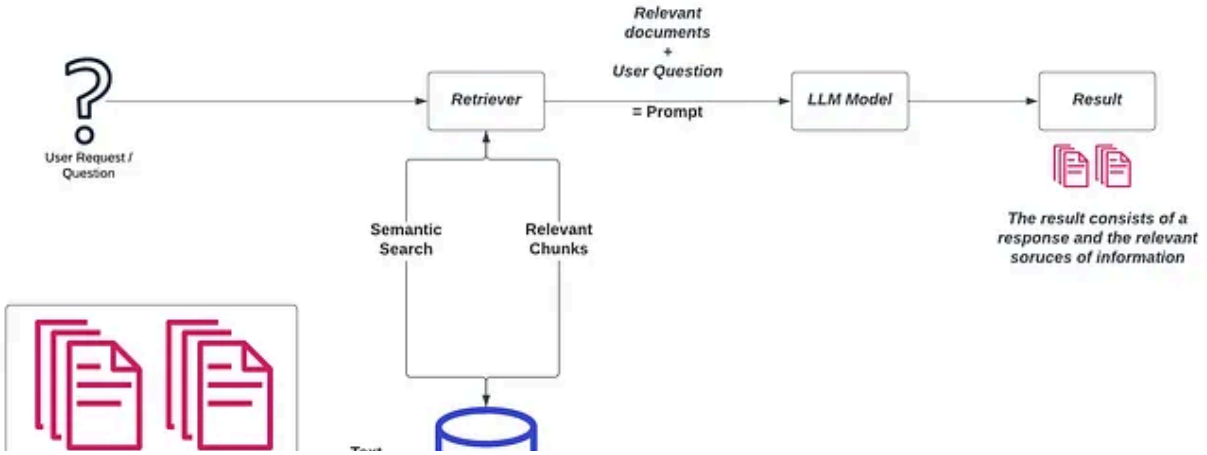
A Symphony of Algorithms: Vector Databases and Generative AI in Finance

Exploring the Possibilities and Challenges Emerging from the Fusion of Vector Databases and GenAI in Finance

medium.datadriveninvestor.com

The main parts of Retrieval Augmented Generation are:

1. Retrieval Phase
2. Augmentation Phase
3. Generation phase



2.1 The Retrieval Phase Explained

RAG has a process that starts with the retrieval phase. When you make a query, RAG searches through large databases (vector databases) to find relevant documents or information.

This process involves more than just finding keywords. RAG aims to understand the context of the query and retrieve the most relevant information. The quality and diversity of the database are essential because the broader and more diverse the database, the more accurate and comprehensive the retrieved information will be.

Information retrieval techniques have evolved significantly, using sophisticated algorithms that go beyond simple keyword matching to understand the semantics and context of the query.

2.2 The Augmentation Phase Explained

After we gather all the necessary information, we move on to the augmentation phase.

We use the retrieved documents during this phase to provide additional context for the generative model. This is where RAG (Retriever-Augmented Generation) truly excels, as it can leverage this context to generate more precise, detailed, and relevant responses.

For example, suppose someone inquires about current global events. RAG can fetch the most recent and relevant articles in that case, resulting in a grammatically correct and factually updated response.

2.3 The Generation Phase Explained

In the final phase of the process, the generative model uses both the original query and the additional information collected to produce a response. This response is not merely a repetition of the gathered data but a coherent and contextually-enriched answer. For instance, in a medical inquiry, RAG can combine medical databases with the user's specific query to provide an accurate and personalized response that fits the query's context.

3. Challenges and Limitations of RAG

RAG, has made significant progress, but it has some challenges and limitations.

One of the primary challenges is the high computational requirements needed for retrieving, augmenting, and generating responses. This process requires powerful hardware and optimized algorithms to function correctly.

Another challenge lies in the quality of the retrieved data. The accuracy and relevance of RAG's responses heavily depend on the quality and recency of the information stored in its database. Therefore, ensuring that the database is regularly updated and free from inaccuracies is crucial.

Moreover, RAG has some limitations in understanding and generating contextually accurate responses. Even though it is an advancement in AI, it is not infallible. Therefore, at times, it may retrieve irrelevant information or

miss the nuances of certain queries, leading to responses that may not fully address the user's intent, even though they may sound fluent.

Conclusion

When generating AI responses, Semantic search and Retrieval Augmented Generation (RAG) are two techniques that provide more relevant and accurate results. Unlike other approaches like building a new model, fine-tuning an existing one, or performing prompt engineering, RAG tackles recency and context-specific issues simultaneously at a lower cost and with less risk. The main objective of RAG is to provide detailed and context-sensitive answers to questions that require access to private data to answer accurately.

Its primary purpose is to provide context-sensitive, detailed answers to questions requiring access to private data to answer correctly. The industry needs more advanced and nuanced systems, like RAG AI, to cater to its evolving needs and challenges.

Looking to the future, RAG AI systems promise even more sophistication, focusing on deeper domain-specific coverage and uncovering hidden insights within vast unstructured data. These advancements will further enhance the precision and relevance of information delivered to various users in the financial sector.