

Why RAG Systems?

Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

Published Author



Outline

- Why LLMs?
- What can we do with LLMs?
- Challenges with LLMs
- How to Augment LLMs with private data
- Current Long Context LLMs
- Challenges with Long Context LLMs
- Why RAG Systems?

Why LLMs?

- LLMs are a phenomenal piece of technology for knowledge generation and reasoning.
- They are pre-trained (and fine-tuned) on large amounts of publicly available data.



What can we do with LLMs?



Blog Writing

You
Summarise the following in 50 words only -
Generative AI is an exciting field that has the potential to revolutionize the way we create and consume content. It can generate new art, music, and even realistic human faces that never existed before. One of the most promising aspects of Generative AI is its ability to create unique and customized products for various industries. For example, in the fashion industry, Generative AI can be used to create new and unique clothing designs. In contrast, in interior design, it can help generate new and innovative home decor ideas.

However, Generative AI is not without its challenges. One of the biggest concerns is the ethical implications of using this technology to generate content without proper attribution or consent. Another challenge is ensuring that the generated content is highly relevant to the user.

Despite these challenges, the potential of Generative AI is enormous. As the technology continues to evolve, we can expect to see more innovative applications that will change the way we think about content creation and consumption.

ChatGPT
Generative AI has transformative potential in content creation, producing art, music, and human-like faces. It's valuable for customizing products in various industries like fashion and interior design. Ethical concerns arise around content generation without consent, and ensuring relevance to users is a challenge. Nevertheless, Generative AI promises groundbreaking applications as it evolves, reshaping content creation and consumption.

Summarization

A blurred screenshot of a code editor displaying a snippet of XML or SVG code. The code includes elements like `<linearGradient>`, `<stop>`, and `<feOffset>`, with various attributes such as `x1="100%"`, `y1="0%"`, `color="#06101F"`, and `offset="0%"`. The code is written in a dark-themed code editor.

Code Assistance

What can we do with LLMs?

Generate a list of ideas for writing a blog on RAG Systems (max 5 points)

< 2/2 >

Here's a list of ideas for writing a blog on RAG (Retrieval-Augmented Generation) Systems:

1. **Introduction to RAG Systems: Bridging Retrieval and Generation in AI**
 - Define what RAG systems are and explain how they combine information retrieval with text generation to provide more accurate and context-rich outputs.
2. **Key Components of RAG Systems: How Retrieval and Generation Work Together**
 - Break down the two core elements of a RAG system: the retrieval model (fetching relevant documents) and the generation model (creating responses based on the retrieved information).
3. **Benefits and Use Cases of RAG Systems Across Industries**
 - Explore real-world applications of RAG in healthcare, customer support, legal research, and more, highlighting its ability to enhance decision-making and provide detailed responses.
4. **Challenges and Limitations of RAG Systems: Overcoming Data Silos and Latency**
 - Discuss common challenges such as retrieval latency, context misalignment, and handling vast databases, along with possible solutions or advancements in the field.
5. **Future Trends in RAG Systems: Enhancing AI Responses with Dynamic Knowledge Retrieval**
 - Predict the future of RAG, such as integrating with live databases for up-to-the-minute responses, improving fine-tuning, and advancing model architectures for more efficient retrieval.

Idea Generation

What is a RAG System? Explain briefly.

A RAG system (Retrieval-Augmented Generation system) is a type of AI model that combines information retrieval with generative capabilities. It works by first retrieving relevant information from a large external dataset or knowledge base (the retrieval component) and then using a generative model (like a large language model) to process the retrieved data and generate a coherent response or output.

The advantage of a RAG system is that it improves the accuracy and relevance of the generated text by grounding it in factual, retrieved information. This hybrid approach is particularly useful in cases where the model alone might not have sufficient knowledge or when the information needs to be precise and up-to-date. RAG systems are commonly used in applications like question-answering, customer support, and content generation.

Question Answering

Challenges with LLMs

What was discussed in the Google Q2 Earnings Call 2024? Tell me without searching online only based on your trained knowledge

As of my knowledge cutoff in October 2023, I don't have specific details about Google's Q2 2024 earnings call. However, based on historical trends, Google's earnings calls typically focus on several key areas:

- Financial Performance:** Google (or Alphabet) typically discusses revenue growth, operating income, and net income. Key revenue drivers are often advertising (especially Google Search and YouTube ads), Google Cloud, and other segments like Google Play and hardware. They may also highlight any unusual expenses or one-time items that affected profitability.
- Ad Revenue Trends:** Google places significant emphasis on trends in digital advertising, especially any shifts in advertiser spend across its platforms (Search, YouTube, and Display Network). Discussions often cover how macroeconomic factors, like inflation or changes in consumer behavior, impact ad revenue.

Based on your knowledge without any searches, tell me about LangGraph, do not do any guesswork, only tell me the answer if you know about it.

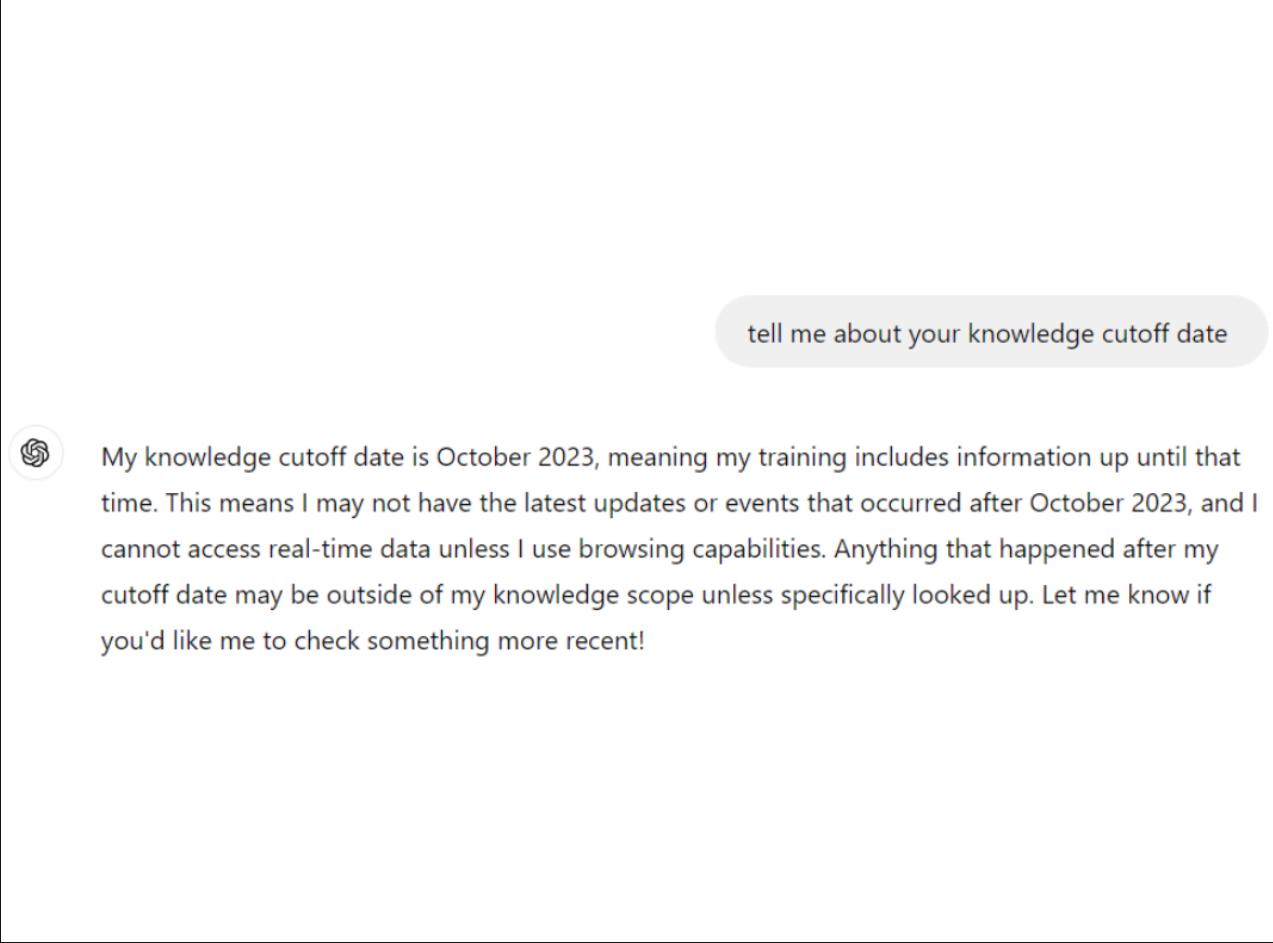
As of my last update, I do not have any specific information about "LangGraph." It does not appear to be a widely recognized tool, framework, or term related to Generative AI, LangChain, or other AI technologies that I have encountered.

If this is a new tool, framework, or concept, it might have emerged after my knowledge cutoff in October 2023. Would you like me to perform a quick search to provide more details?

Recent Updates

Enterprise Data

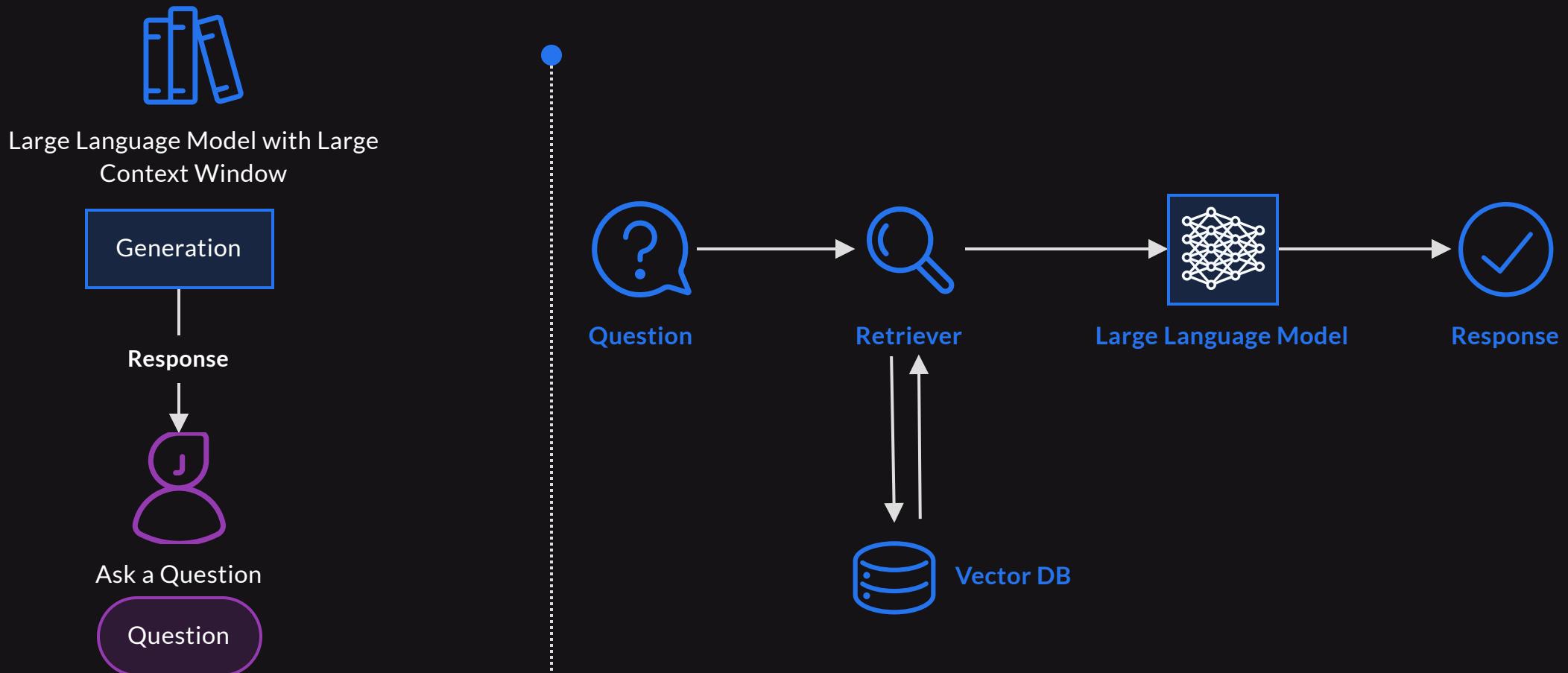
GPT-4o Knowledge Cutoff Date

A screenshot of a GPT-4o AI interface. At the top right is a blue button labeled "tell me about your knowledge cutoff date". Below it is a circular profile icon with a white swirl symbol. The main text area contains the following message:

My knowledge cutoff date is October 2023, meaning my training includes information up until that time. This means I may not have the latest updates or events that occurred after October 2023, and I cannot access real-time data unless I use browsing capabilities. Anything that happened after my cutoff date may be outside of my knowledge scope unless specifically looked up. Let me know if you'd like me to check something more recent!

LLMs usually have a knowledge cut-off date (data until this date was used in its training) and are not updated frequently with new information

How To Best Augment LLMs With Our Own Private Data



Long Context LLMs

Retrieval Augmented Generation (RAG)

Current Long Context LLMs

LLM	Organization	Context Limit (Tokens)
GPT-4o	OpenAI	128,000
Claude 3.5	Anthropic	200,000
Gemini 1.5	Google DeepMind	1,000,000
Llama 3.2	Meta	128,000

Challenges with Long Context LLMs

- Latency and Processing Time
- More documents increase the time taken to find the context for queries
- More documents also increase cost as more LLMs charge per token
- Lost-in-the-middle issue when finding information

Why RAG Systems?

- Separate retrieval stage which can scale to millions of documents
- Cheap, easy to implement and scalable
- Can deal with long context data and many documents
- Easily integrate custom data to augment LLM's knowledge
- No need for fine-tuning LLMs (most of the time)
- Tries to prevent hallucinations

Thank You
