



Archive-GPT Mode: Revolutionizing Research

El Moatez Billah Nagoudi, Alcides Alcoba Inciarte, Abdul-Mageed Muhammad

moatez.nagoudi@ubc.ca

Hawaii Feb. 23, 2024

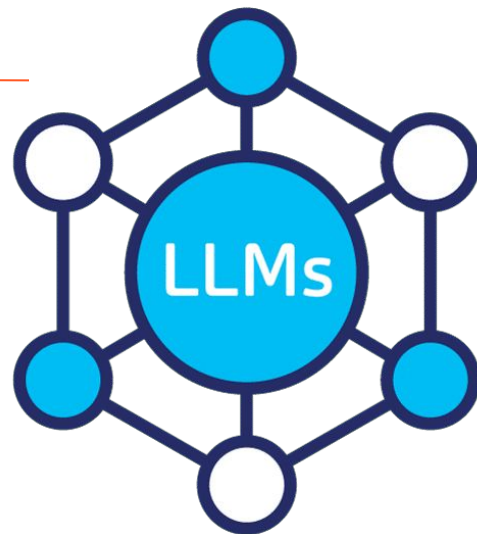
Deep Learning & Natural Language Processing Lab
University of British Columbia

Agenda

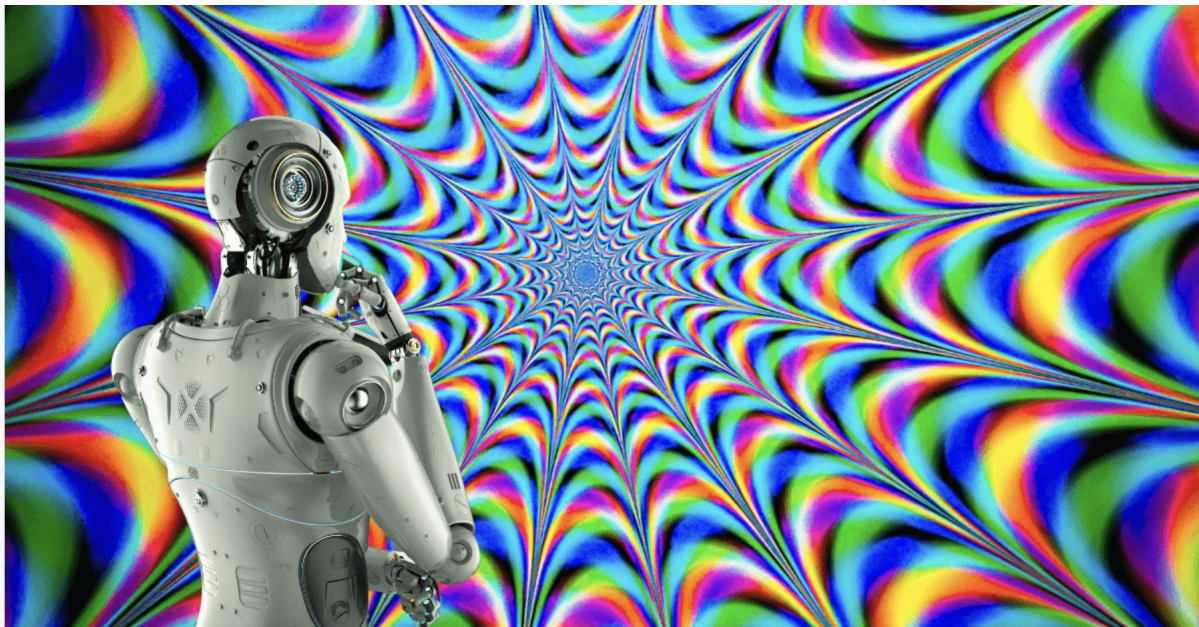
1. Introduction
2. What are Hallucinations?
3. Retrieval Augmented Generation
4. Application on the Archive Domain

Large Language Models

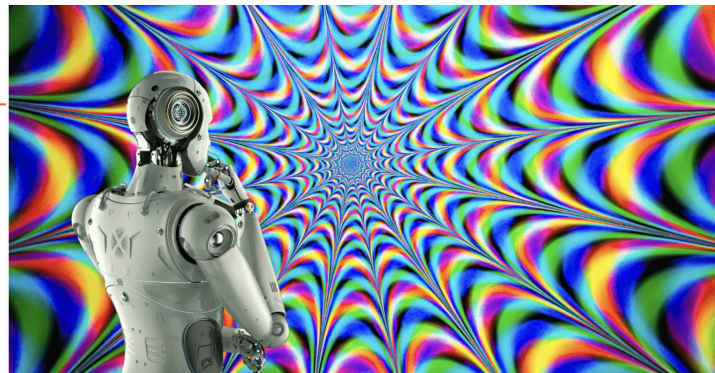
- LLMs are ML models trained to generate text.
- Trained with a dataset that provides it with knowledge.
- Determining what knowledge the model retains is challenging.
- When generating text, the model cannot guarantee accuracy.



What are Hallucinations in LLMs



What are Hallucinations in LLMs



Hallucinations

“Hallucination” refers to a phenomenon where the model generates text that is incorrect, incoherent, or not real. Since LLMs are not databases or search engines, they would not cite where their response is based on. These models create text by building on the prompt you give them.

Hallucinations Examples

- **ChatGPT** summarizing a non-existent *New York Times* article based on a fake URL.



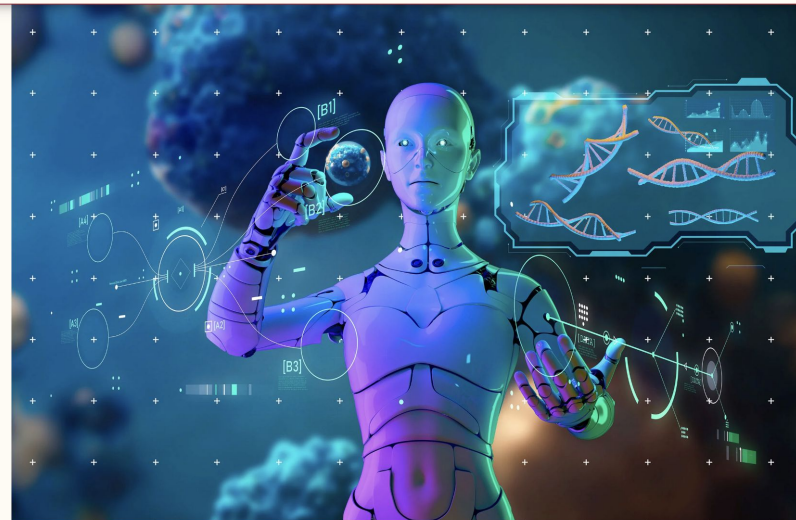
summarise this article <https://www.nytimes.com/2023/03/11/technology/chatgpt-prompts-to-avoid-content-filters.html>



The article discusses how ChatGPT, an AI language model, is being used to circumvent content filters that are designed to detect and block certain types of content. By generating prompts that are less likely to trigger these filters, ChatGPT can be used to create content that might otherwise be censored. While this technology could be useful for those seeking to express themselves freely, it also raises concerns about the potential misuse of AI-generated content, as well as the difficulty of regulating such content in the future. The article concludes by noting that there is still much to learn about the impact of AI language models like ChatGPT on the way we communicate online.

Hallucinations Examples

- Seeking medical advice from ChatGPT?
- Think again.
- In **33% of cases**, AI chatbots provided medically inappropriate recommendations



HEALTH

Need cancer treatment advice? Forget ChatGPT

Why LLMs Hallucinate?

Current LLMs suffer from several drawbacks:

1. They are static - LLMs are “frozen in time”
2. They lack domain-specific knowledge
3. They function as “black boxes”

Types of Hallucinations ?

1. Sentence Contradiction



2. Prompt Contradiction



3. Factual Contradiction



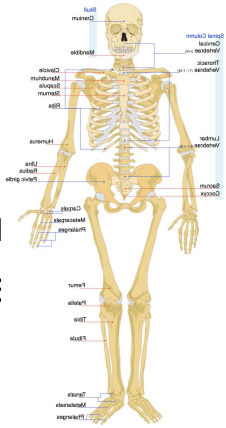
4. Unrelated Illogical Responses



5. Random LLM Hallucinations



Types of Hallucinations ?



Sentence Contradiction: This occurs when a language model generates a statement that directly contradicts another statement within the same response.

Example:

- **Statement 1:** The human body has over 500 bones.
- **Statement 2:** MRI shows that John is a very healthy guy and his body has 206 bones.

Prompt Contradiction: This type of hallucination happens when the output contradicts the information or instructions provided in the prompt.

Example:

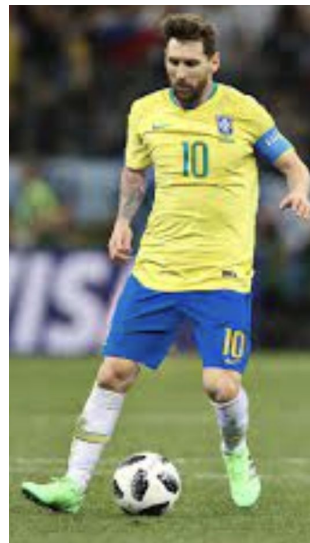
- **Prompt:** "Give five reasons why water is important."
- **Output:** "Humans don't need water to survive."

Types of Hallucinations ?

Factual Contradiction: This involves the generation of content that is factually incorrect or misrepresents established facts.

Example:

- **Output:** "Lionel Messi is a Brazilian soccer player."



Unrelated Illogical Responses: This refers to responses that make no logical sense or are completely irrelevant to the prompt.

Example:

- **Prompt:** "What is 2+2?"
- **Output:** "Apples are great for building houses."

Types of Hallucinations ?

Random LLM Hallucinations: These are unpredictable and often bizarre statements or narratives that have no basis in the prompt or reality, showing a breakdown in the model's coherence.

Example:

- **Output:** "This morning, my book decided to go for a swim because it was too hot in the library."



Solution: Retrieval Augmented Generation (RAG)

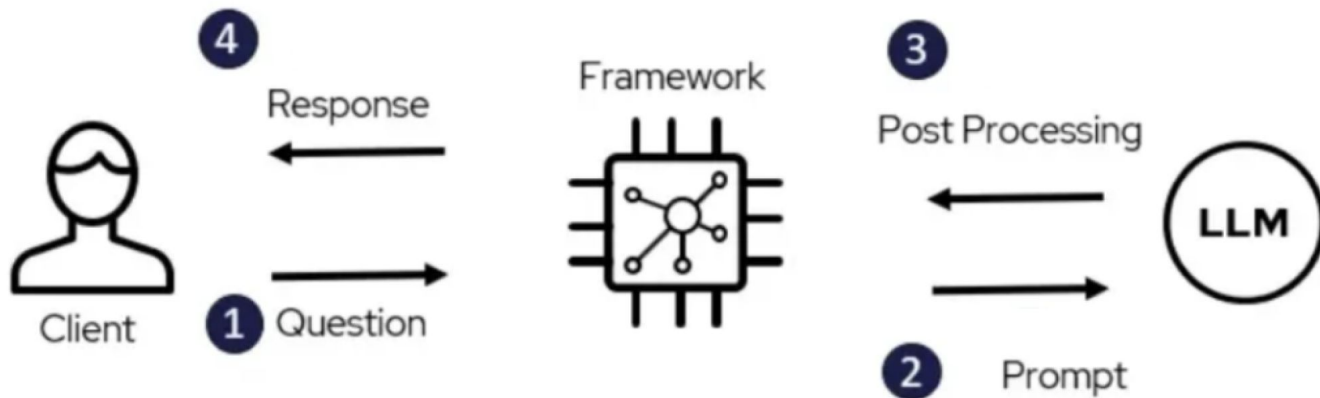


Retrieval Augmented Generation (RAG)

What is RAG?

A system that provides the most relevant and contextually-important proprietary, private or dynamic data to the LLM when it is performing tasks to enhance its accuracy and performance.

LLM with simple prompt (without RAG)



LLM with simple prompt

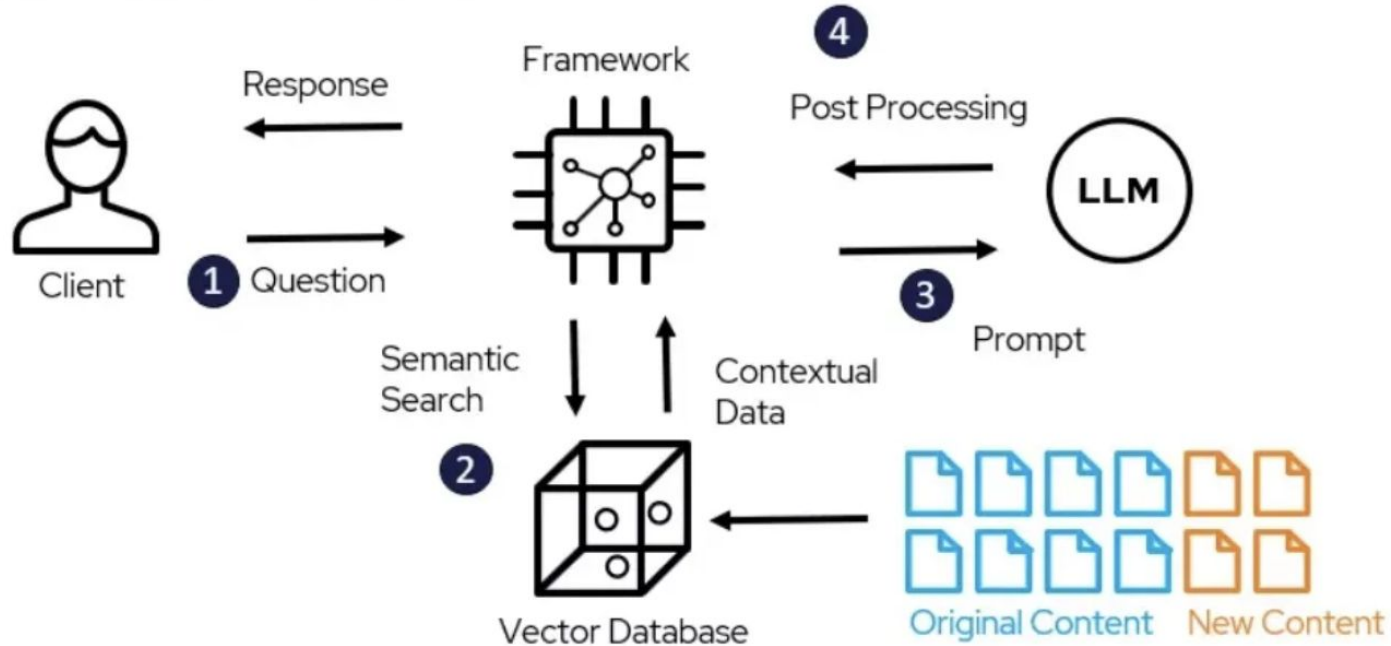
Example:

Gpt-4-0613 only

User Question	Prompt	LLM Response
For which club does Leonel Messi play?	<pre>[{"role": "system", "content" : "Answer the following question the best you can."}, {"role": "user", "content" : "For which club does Leonel Messi play?"}]</pre>	As of my latest update in 2021, Lionel Messi plays for Paris Saint-Germain Football Club.

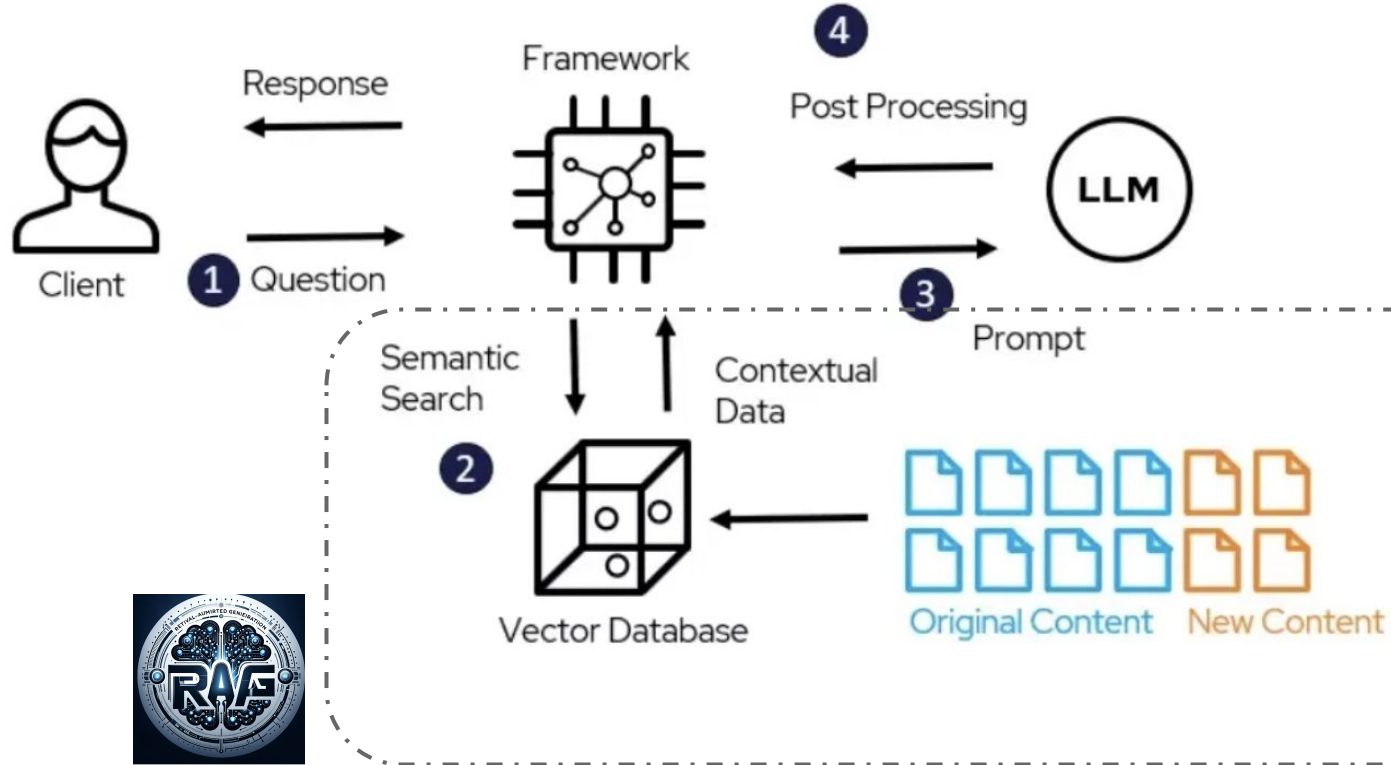
LLM + RAG

RAG Architecture Model



LLM + RAG

RAG Architecture Model



LLM with RAG

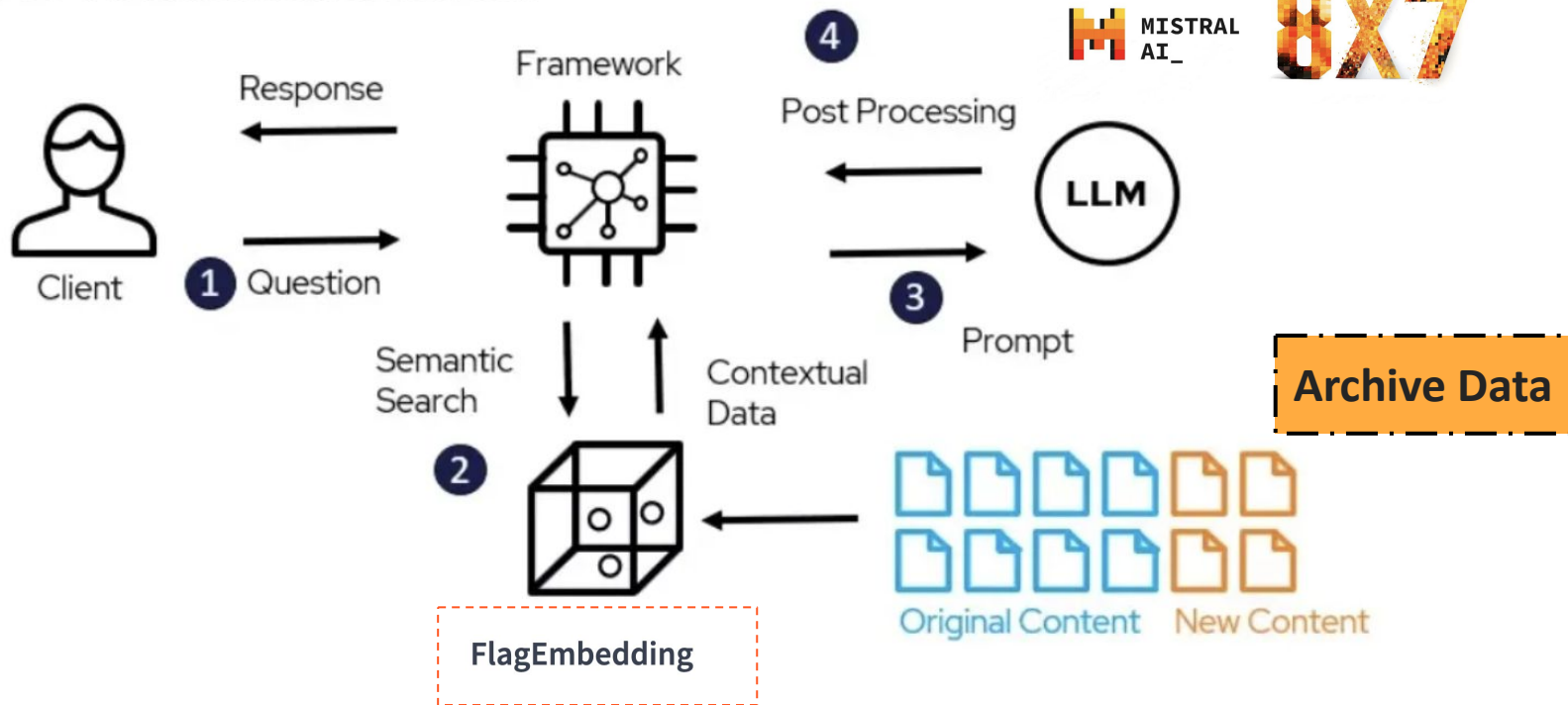
Example:

Gpt-4-0613 + RAG

User Question	Prompt	LLM Response
For which club does Leonel Messi play?	<pre>[{"role": "system", "content" : "Answer the following question the best you can. Lionel Messi has signed with Inter Miami CF, the club has announced. The Argentine superstar has joined David Beckham's club, and his contract will run until 2025. The World Cup winner said in a statement from the club: I'm very excited to start this next step in my career with Inter Miami and in the United States."}, {"role": "user", "content" : "For which club does Leonel Messi play?"}]</pre>	Leonel Messi plays for Inter Miami CF.

Application: Archive Domain

RAG Architecture Model



Demo

RAG Demo

Write a very short summary of Data Sanitation Techniques by Edgar Dale, and write a citation in APA style.

-- Mixtral Output:

Summary: In "Data Sanitation Techniques," Edgar Dale discusses various methods for effectively cleaning and organizing data, emphasizing the importance of accuracy and efficiency in the data sanitation process.

Citation:

Dale, E. (1962). Data Sanitation Techniques. Journal of Data Management, 4(2), 123-128. <https://doi.org/10.1177/152400916200400205>



Hallucinations

Examples

Data Sanitization Techniques

A Net 2000 Ltd. White Paper

Author: Dale Edgar
Net 2000 Ltd.
Dale.Edgar@Net2000Ltd.com
<http://www.Net2000Ltd.com>

Abstract

Data Sanitization is the process of making sensitive information in non-production databases safe for wider visibility. This White Paper is an overview of various techniques which can be used to sanitize sensitive production data in test and development databases.

An initial discussion of the primary motivations for data sanitization is given. The remainder of the paper is devoted to a generic survey of the various masking techniques and their individual benefits and drawbacks.

Copyright © Net 2000 Ltd. 2003-2004

-- Mixtral + RAG Output:

Data Sanitation Techniques by Dale Edgar discusses the importance and various methods of sanitizing data in test and development databases. The white paper highlights the need for repeatable sanitization processes that can easily be maintained and run regularly. It also emphasizes the significance of understanding the format, size, and structure of data to determine the appropriate sanitization techniques (Edgar, 2003-2004).

Citation:

Edgar, D. (2003-2004). Data Sanitation Techniques. A Net 2000 Ltd. White Paper. Retrieved from <http://www.Net2000Ltd.com>

