

Build a RAG system – Case Study

Prepared by:

Abhishek Neeraj
Najeeb Fajandar
Kranthi

Goal:

Build a robust generative search system capable of effectively and accurately answering questions from various Financed Emission related documents. You may use LangChain or LlamaIndex to build the generative search application.

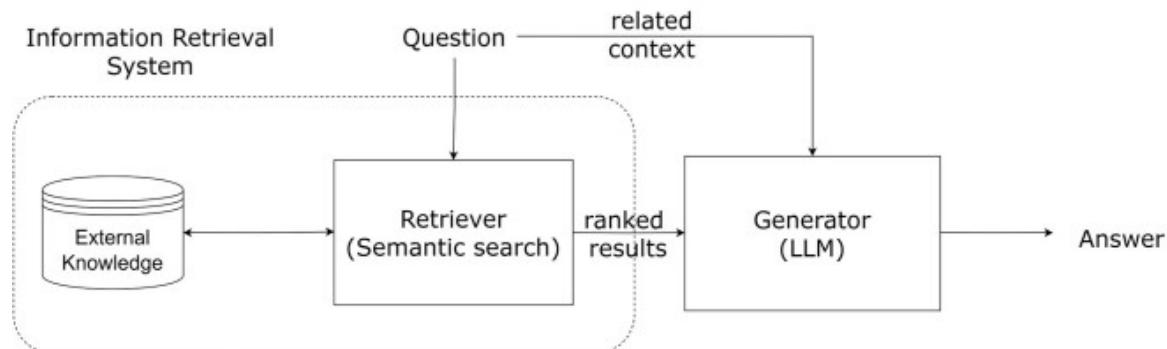
Data Source: Financed Emission related documents in PDF form as below:

<https://github.com/NajeebFajandar/Build-a-RAG-system-case-study>

Notebook: <https://github.com/NajeebFajandar/Build-a-RAG-system-case-study>

System Design:

Retrieval Augmented Generation (RAG)



Retrieval-Augmented Generation (RAG) is the concept to provide LLMs with additional information from an external knowledge source. This allows them to generate more accurate and contextual answers while reducing hallucinations.

Problem

State-of-the-art LLMs are trained on large amounts of data to achieve a broad spectrum of general knowledge stored in the neural network's weights (parametric memory). However, prompting an LLM to generate a completion that requires knowledge that was not included in its training data, such as newer, proprietary, or domain-specific information, can lead to factual inaccuracies (hallucinations), as illustrated in the following screenshot:

ME

What did the president say about Justice Breyer?



I'm sorry, but I don't have access to real-time information, and my knowledge only goes up until January 2022. Therefore, I can't provide you with any specific comments or statements made by the president about Justice Stephen Breyer if they occurred after that date. To find the latest information on this topic, I recommend checking the latest news sources or the official statements from the president's office.

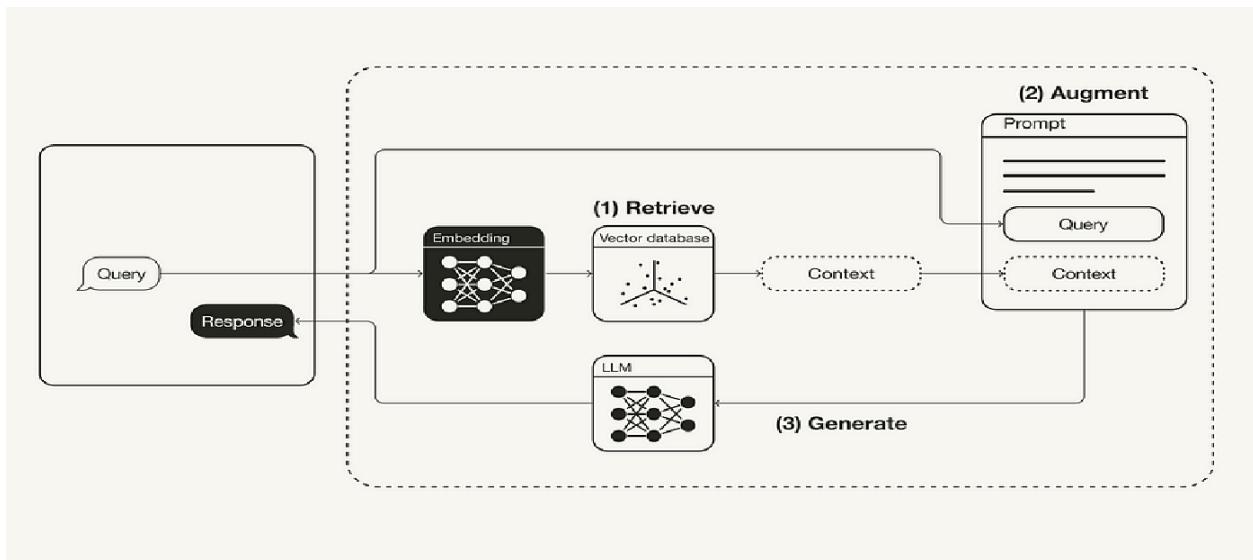
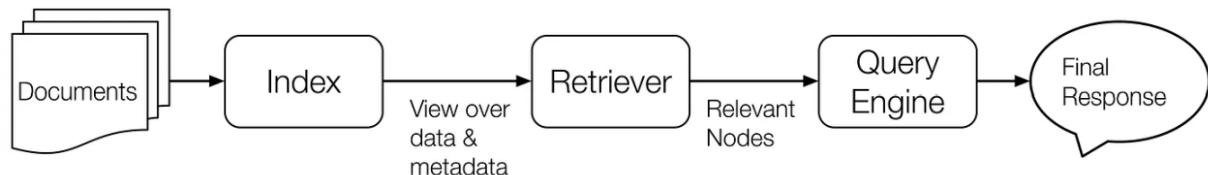
Thus, it is important to bridge the gap between the LLM's general knowledge and any additional context to help the LLM generate more accurate and contextual completions while reducing hallucinations.

Solution

Traditionally, neural networks are adapted to domain-specific or proprietary information by fine-tuning the model. Although this technique is effective, it is also compute-intensive, expensive, and requires technical expertise, making it less agile to adapt to evolving information.

In 2020, Lewis et al. proposed a more flexible technique called Retrieval-Augmented Generation (RAG) in the paper [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#). In this paper, the researchers combined a generative model with a retriever module to provide additional information from an external knowledge source that can be updated more easily.

The vanilla RAG workflow is illustrated below:



LangChain vs LlamaIndex: A Basic Overview

LlamaIndex excels in search and retrieval tasks. It's a powerful tool for data indexing and querying and a great choice for projects that require advanced search. LlamaIndex enables the handling of large datasets, resulting in quick and accurate information retrieval.

LangChain is a framework with a modular and flexible set of tools for building a wide range of NLP applications. It offers a standard interface for constructing chains, extensive integrations with various tools, and complete end-to-end chains for common application scenarios.

Here's a comparative table to summarize the key differences:

Feature	LlamaIndex	LangChain
Primary Focus	Search and retrieval	Flexible LLM-powered application development
Data Indexing	Highly efficient	Modular and customizable

Retrieval Algorithms	Advanced and optimized	Integrated with LLMs for context-aware outputs
User Interface	Simple and user-friendly	Comprehensive and adaptable
Integration	Multiple data sources, seamless platform integration	Supports diverse AI technologies and services
Customization	Limited, focused on indexing and retrieval	Extensive, supports complex workflows
Context Retention	Basic	Advanced, crucial for chatbots and long interactions
Use Cases	Internal search, knowledge management, enterprise solutions	Customer support, content generation, code documentation
Performance	Optimized for speed and accuracy	Efficient in handling complex data structures
Lifecycle Management	Integrates with debugging and monitoring tools	Comprehensive evaluation suite (LangSmith)

We would be using Llama Index for implementing RAG, for efficient searching in Financed Emission related documents.

Integrating LangChain and Llama index with RAG can significantly improve the performance and efficiency of AI assistants, making them more capable of understanding and responding to user queries in a more natural and human-like manner.

Implementation:

Sample Query and Response:

Step 5 - Checking responses and response parameters

```
In [41]: response = query_engine.query("What is this document talking about?")  
  
In [42]: #Checking the response  
response.response  
  
Out[42]: 'This document is discussing the TCFD index, governance, strategy, risk management, metrics and targets, and outlook related to climate disclosure.'  
  
In [43]: #Check the source node  
response.source_nodes  
  
Out[43]: [NodeWithScore(node=TextNode(id='da190b41-daf5-48e8-aab8-727f998dcfba', embedding=None, metadata={'page_label': '2', 'file_name': 'climate-disclosure.pdf', 'file_path': 'D:\\Puja\\IIITB_MS\\RAG\\llamaIndex_RAG\\semantic spotter\\Data\\climate-disclosure.pdf', 'file_type': 'application/pdf', 'file_size': 1500685, 'creation_date': '2024-07-21', 'last_modified_date': '2024-07-21'}, excluded_embed_metadata_keys=['file_name', 'file_type', 'file_size', 'creation_date', 'last_modified_date', 'last_accessed_date'], excluded_llm_metadata_keys=['file_name', 'file_type', 'file_size', 'creation_date', 'last_modified_date', 'last_accessed_date'], relationships={<NodeRelationship.SOURCE: '1': RelatedNodeInfo(node_id='7885e6d8-5ee0-41d1-b677-512b3b3ce2cd', node_type=<ObjectType.DOCUMENT: '4', metadata={'page_label': '2', 'file_name': 'climate-disclosure.pdf', 'file_path': 'D:\\Puja\\Settings\\IIITB_MS\\RAG\\llamaIndex_RAG\\semantic spotter\\Data\\climate-disclosure.pdf', 'file_type': 'application/pdf', 'file_size': 1500685, 'creation_date': '2024-07-21', 'last_modified_date': '2024-07-21'}, hash='a0927babb5967675369940223141a01126945b6b66b'})]
```

After Reranking:

```
# Execute a query using the prepared query engine  
response = query_engine.query(  
    "What is Scope 3 emission?", # Query asking about the main characters of a book  
)  
  
In [62]: response.response  
  
Out[62]: "Scope 3 emissions are all other indirect greenhouse gas emissions that occur in the value chain of a reporting company. These emissions are not included in Scope 1 or Scope 2 and can be broken down into upstream emissions that occur in the supply chain and downstream emissions that occur as a consequence of using the organization's products or services."  
  
In [65]: for node in response.source_nodes:  
    print(node.metadata['file_name'])  
    print(node.score)  
    print(node.text)  
  
decommissioning of vehicles after use do not need to be covered because these emissions are difficult to obtain and can be considered rather marginal. However, if a financial institution seeks to account for the production emissions of new vehicles (i.e., embodied emissions), they should report the emissions as follows:  
  
• In the initial financing year, the financial institution shall report the production emissions of the respective vehicle as a lump sum under scope 3 emissions, while the operation emissions in the respective year shall be reported under scope 1 or 2 emissions.  
• In the following financing years, the financial institution shall not report any production emissions of the respective vehicle; they shall only report the operation emissions under scope 1 or 2 emissions.  
This approach on scope 3 emissions only holds for new vehicles, not used vehicles.  
Attribution of emissions  
As a basic attribution principle, the financial institution accounts for a portion of the borrower's annual motor vehicle emissions as determined by the ratio between the outstanding amount (numerator) and the value of the motor vehicle at loan origination (denominator). This ratio is
```

```

response = query_response(user_input)
display(HTML(f'<p style="font-size:20px">{response}</p>'))

```

In [68]: initialize_conv()

Feel free to ask Questions regarding Financed Emissions. Press exit once you are done
What is Financed Emission?

Financed emissions refer to the emissions associated with activities in the real economy that are financed through lending and investment portfolios. These emissions are calculated by multiplying an attribution factor by the emissions of the specific activity being financed, such as a building or a company. Financial institutions use financed emissions to assess climate-related risks, set science-based targets, report to stakeholders, and inform climate strategies and actions. Check further at PCAF-Global-GHG-Standard.pdf document

What is Scope 1 emission?

Scope 1 emissions are direct greenhouse gas emissions that occur from sources owned or controlled by the reporting company, such as emissions from combustion in owned or controlled boilers, furnaces, and vehicles. Check further at PCAF-Global-GHG-Standard.pdf document

who are involved in reporting Financed Emission?

Financial institutions are involved in reporting Financed Emissions. Check further at PCAF-Global-GHG-Standard.pdf document

exit
Exiting the program... bye

Build a Testing Pipeline

Here we feed a series of questions to the Q/A bot and store the responses along with the feedback on whether it's accurate or not from the user

[Activate](#)
[Go to Settings](#)

Testing Pipeline:

In [71]: `import pandas as pd`

In [72]: `testing_pipeline(questions)`

What is Emission reduction target for HSBC?
HSBC's emission reduction targets include being net zero in their own operations and supply chain by 2030, aligning their financial emissions to achieve net zero by 2050, and providing and facilitating \$750bn to \$1tn of sustainable finance and investment by 2030.
Check further at 240311-hsbc-carbon-reduction-plan.pdf document

Please provide your feedback on the response provided by the bot. ("GOOD"/"BAD")
GOOD

What are the Disclosure checklist for reporting?
The Disclosure checklist for reporting includes criteria related to general disclosure, coverage, absolute emissions, avoided emissions and emission removals, and recalculation and significance threshold. Each criterion must be answered with either "Yes" or "No", with a justification required for any deviations from the standard. Additionally, recommendations for reporting can be excluded but it is advised to provide justification. If a disclosure point is not applicable, it should be marked as "No" with "non-applicable" noted in the comment box. The checklist emphasizes the importance of meeting the minimum requirements outlined by the Standard and encourages institutions to exceed these requirements where possible.
Check further at PCAF-Disclosure-Checklist-Part-A-Financed-Emissions-July-2024.pdf document

Please provide your feedback on the response provided by the bot. ("GOOD"/"BAD")
GOOD

What is the role of the financial sector in climate risk?
The financial sector plays a crucial role in addressing climate risk by directing capital towards decarbonization efforts. Financial institutions need to understand and measure the greenhouse gas emissions associated with their loans and investments to manage risks, set emission reduction goals, and disclose progress. By measuring financed emissions, financial institutions can identify actions to lower their climate impact and engage stakeholders in discussions to support lower financed emissions.
Check further at PCAF-Global-GHG-Standard.pdf document

Please provide your feedback on the response provided by the bot. ("GOOD"/"BAD")
GOOD

[Activate](#)
[Go to Settings](#)

	Question	Response	Page	Good or Bad
0	What is Emission reduction target for HSBC?	HSBC's emission reduction targets include being...	document	GOOD
1	What are the Disclosure checklist for reporting?	The Disclosure checklist for reporting include...	document	GOOD
2	What is the role of the financial sector in cl...	The financial sector plays a crucial role in a...	document	GOOD

Building a custom prompt template

We saw that in one of the questions the model was not able to answer properly, or rather give a very detailed answer. Let's see how we can use a custom prompt template to mitigate the issue

```
In [73]: response = query_engine.query("How to calculate Financed emission for any sector?")
```

```
In [74]: response.response
```

```
Out[74]: 'To calculate financed emissions for any sector, the process involves multiplying the attribution factor by the emissions of the respective entity in that sector. This attribution factor represents the proportional share of the entity, which is determined by factors such as outstanding amount, total equity, debt, and revenue. The total financed emissions are then calculated based on this attribution factor and the emissions data specific to the entity in the sector.'
```

Activate

Using Prompt:

```
In [77]: retrieved = response.source_nodes[0].node.text + response.source_nodes[1].node.text
```

```
In [78]: messages = [
    {"role": "system", "content": "You are an AI assistant to Financial emission related queries."},
    {"role": "user", "content": f"How to calculate Financed emission for any sector? Check in '{retrieved}' """},]
```

```
In [79]: response2 = openai.chat.completions.create(
    model="gpt-3.5-turbo",
    messages=messages)
response2.choices[0].message.content
```

```
Out[79]: 'To calculate the financed emissions for any sector, you can use the following equations:\n\n**For a loan or investment in a company:**\n\\text{Financed Emissions} = \\text{Attribution Factor} \\times \\text{Emissions of the Borrower/Investee Company}\n\n**For a listed equity and corporate bonds portfolio:**\n\\text{Financed Emissions} = \\frac{@00000000}{FFF}\\underline{ } \\times CCCCCC\n\n**For mortgages:**\n\\text{Financed Emissions} = \\text{Attribution Factor} \\times \\text{Emissions of the Building}\n\nThese equations will allow you to calculate the financed emissions based on the attribution factor and the emissions associated with the specific sector or investment.'
```

```
In [80]: print(response2.choices[0].message.content)
```

To calculate the financed emissions for any sector, you can use the following equations:

For a loan or investment in a company:

$$\text{Financed Emissions} = \text{Attribution Factor} \times \text{Emissions of the Borrower/Investee Company}$$

For a listed equity and corporate bonds portfolio:

$$\text{Financed Emissions} = \frac{@00000000}{FFF}\underline{ } \times CCCCCC$$

For mortgages:

Activate
Go to Settings

Further Improvement:

This model can be further improved by using better/cleaner dataset or utilizing better data pre-processing techniques.

We can also use custom node and LLMs for better results.

We can use Caching for faster retrieval of the responses, and for frequent queries.