Nuclear Chatbot Project

Automated Cognitive Analysis System: Integration of LLMs, Knowledge Graphs, and Vector Storage for PDF Data Utilization

Realized By : **Mohamed Aziz NEBLI**

Sept 2024

TABLE OF CONTENTS

Presentation of the Host Organization

EDF GROUP

- Electricité de France (EDF) was founded in 1946.
- Specialized in the production, transmission, distribution, and supply of electricity.
- Present in more than 22 countries.
- Over 200 000 employees Internationally
- The 5th largest energy producer in the world.

General Context of the Project

Problem Statement

- EDF holds a vast collection of over 40 million PDFs, encompassing innovations and project strategies.

- This large volume of documents makes it challenging for field engineers, technicians, and design engineers to quickly access the necessary information.

Objective

- Designing a hierarchy for a database containing documentation for various projects.

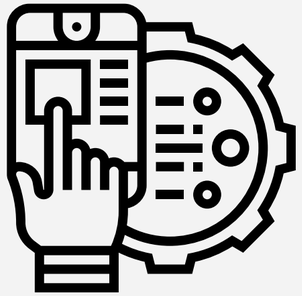
- Creating a CHATBOT for rapid information retrieval within this database, thereby enhancing the productivity of engineers and technicians.

Proposed Solution

An internal database

A chatbot based on internal technologies

Rapid data cleaning and processing of documents



User-friendly interface

Technologies

Data Storage : Graph Data Base

than zero.

- With K_5 negative, the AVR action introduces a positive synchronizing torque component and a negative damping torque component. This effect is more pronounced as the exciter response increases.

  For high values of external system reactance and high generator outputs K_5 is negative. In practice, the situation where K_5 is negative are commonly encountered. For such cases, a high response exciter is beneficial in increasing synchronizing torque. However, in so doing it introduces negative damping. We thus have conflicting requirements with regard to exciter response. One possible recourse is to strike a compromise and set the exciter response so that it results in sufficient synchronizing and damping torque components for the expected range of system-operating conditions. This may not always be possible. It may be necessary to use a high-response exciter to provide the required synchronizing torque and transient stability performance. With a very high external system reactance, even with low exciter response the net damping torque coefficient may be negative.

An effective way to meet the conflicting exciter performance requirements with regard to system stability is to provide a power system stabilizer as described in the following section.

12.5 POWER SYSTEM STABILIZER

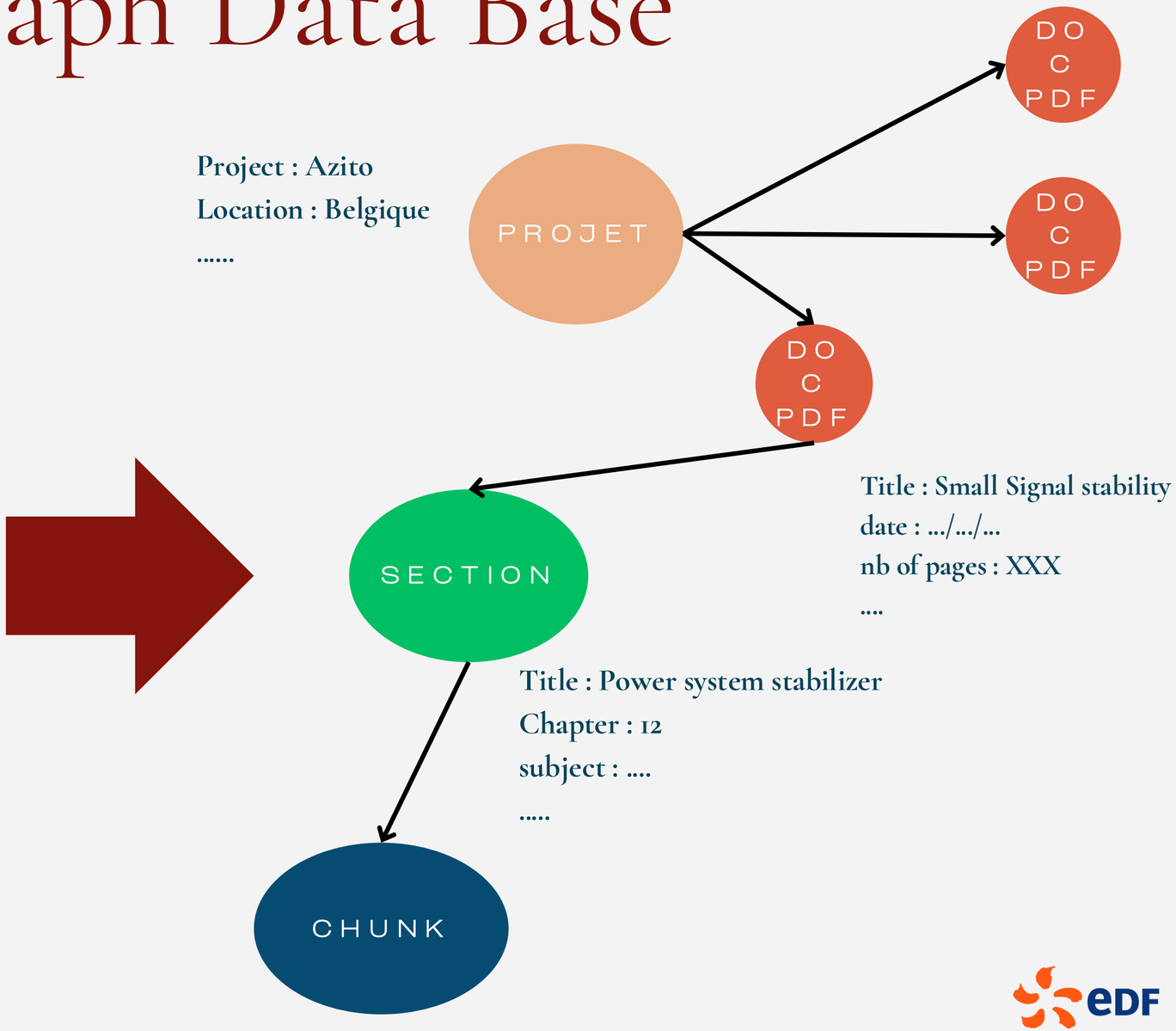
The basic function of a *power system stabilizer* (PSS) is to add damping to the generator rotor oscillations by controlling its excitation using auxiliary stabilizing signal(s). To provide damping, the stabilizer must produce a component of electrical torque in phase with the rotor speed deviations.

The theoretical basis for a PSS may be illustrated with the aid of the block diagram shown in Figure 12.13. This is an extension of the block diagram of Figure 12.12 and includes the effect of a PSS.

Since the purpose of a PSS is to introduce a damping torque component, a logical signal to use for controlling generator excitation is the speed deviation $\Delta\omega_r$.

If the exciter transfer function $G_{ex}(s)$ and the generator transfer function between ΔE_{fd} and ΔT_e were pure gains, a direct feedback of $\Delta\omega_r$ would result in a damping torque component. However, in practice both the generator and the exciter (depending on its type) exhibit frequency dependent gain and phase characteristics. Therefore, the PSS transfer function, $G_{PSS}(s)$, should have appropriate phase compensation circuits to compensate for the phase lag between the exciter input and the electrical torque. In the ideal case, with the phase characteristic of $G_{PSS}(s)$ being an exact inverse of the exciter and generator phase characteristics to be compensated, the PSS would result in a pure damping torque at all oscillating frequencies.

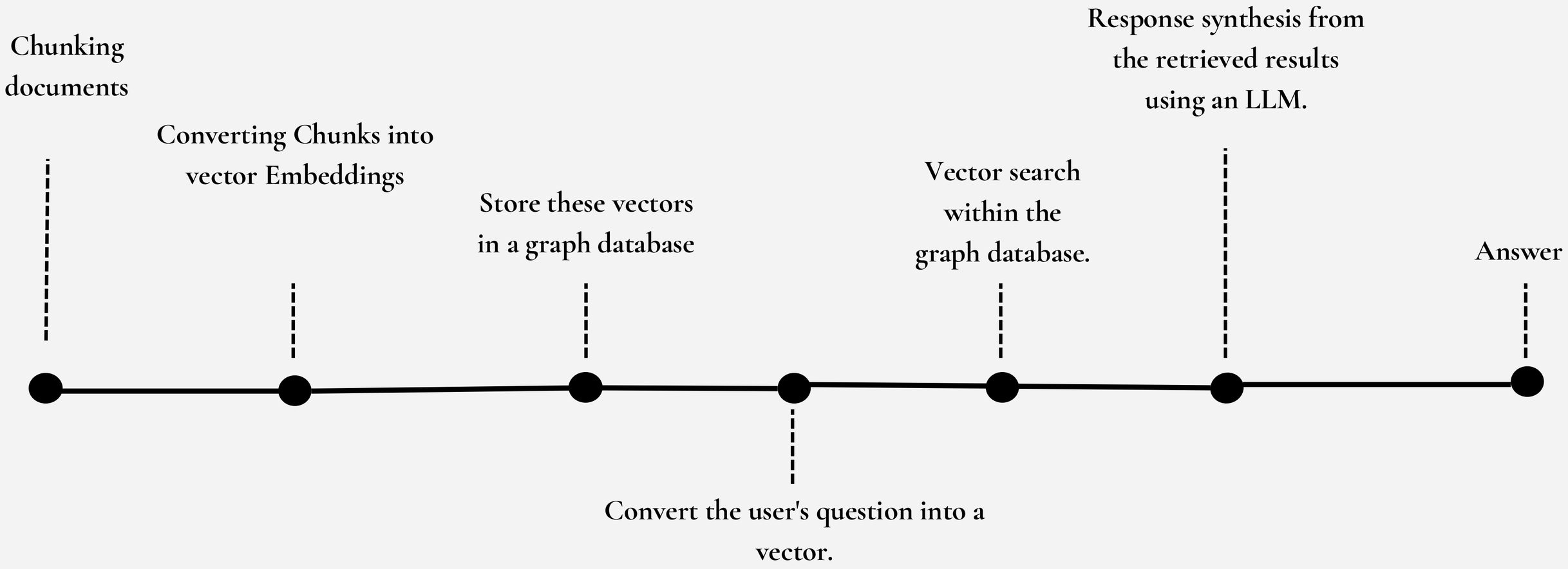
Project : Azito
Location : Belgique
.....



Title : Small Signal stability
date : .../.../...
nb of pages : XXX
....

Title : Power system stabilizer
Chapter : 12
subject : ....
.....

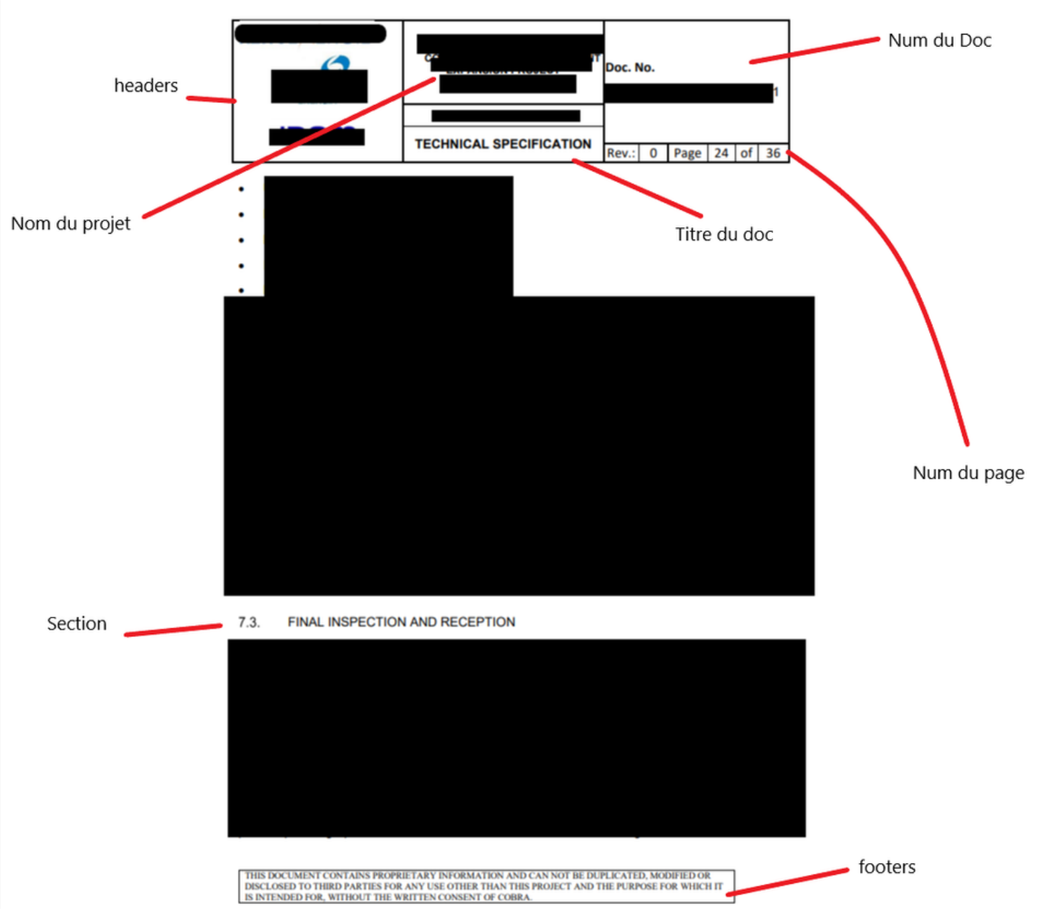
Project Progression



Data Processing

Nature of Data

- PDFs or .txt files with headers and footers, images, tables, often pages containing only numerical codes or symbols, links, etc.
- Database of 500-1000 Pdfs
- Detect the headers and footers on each PDF page, identifying and removing the table of contents from each PDF.



Graph Construction

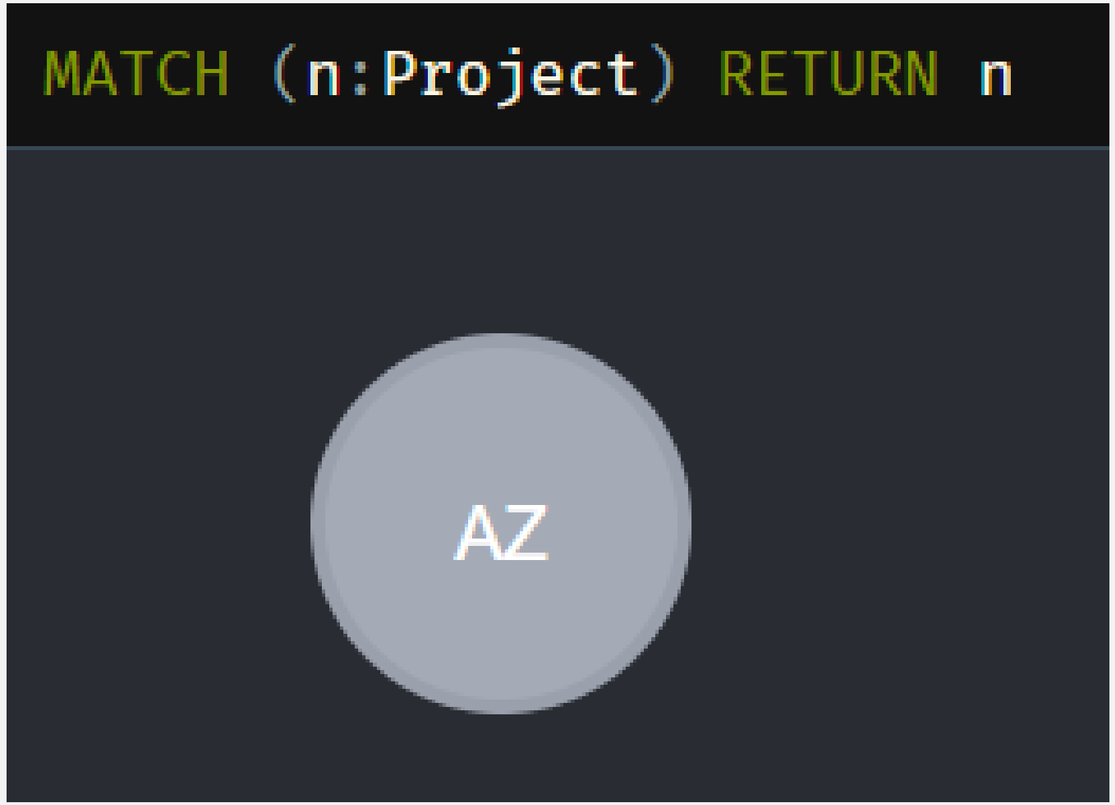
Graph Definition

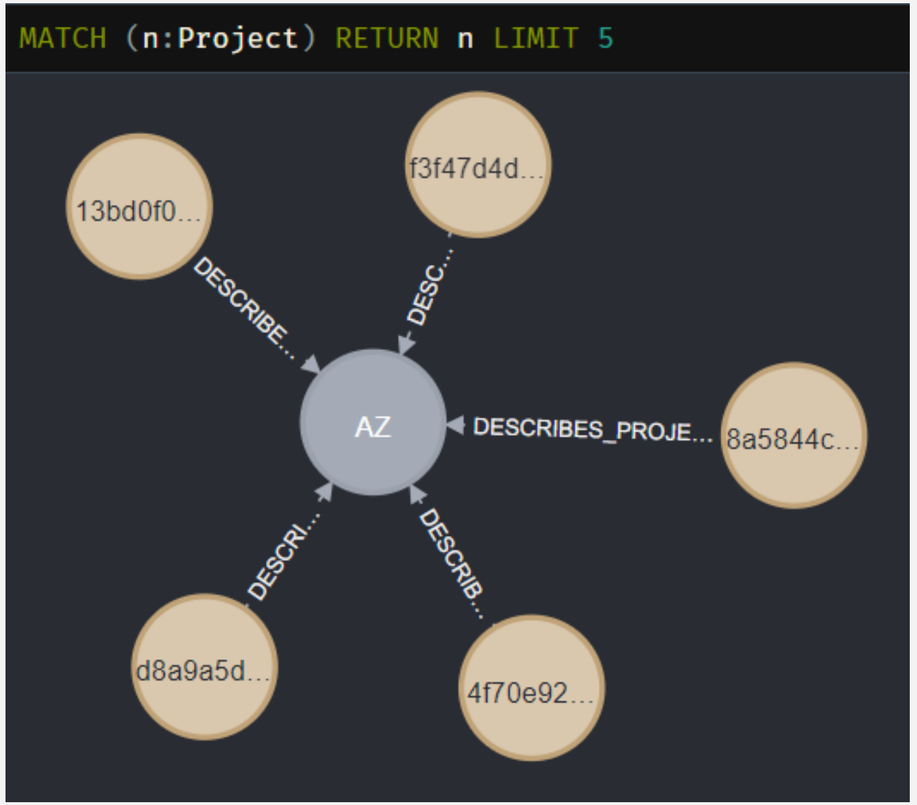
- Create a Graph that :
  - Respects PDFs' internal hiearchy :
    - Title
    - Section
    - Paragraphs
  - Serves as a Vector store :

  There are 2 Vector Stores in the graph

Graph : Project Node

- Group the PDFs related to the same project.

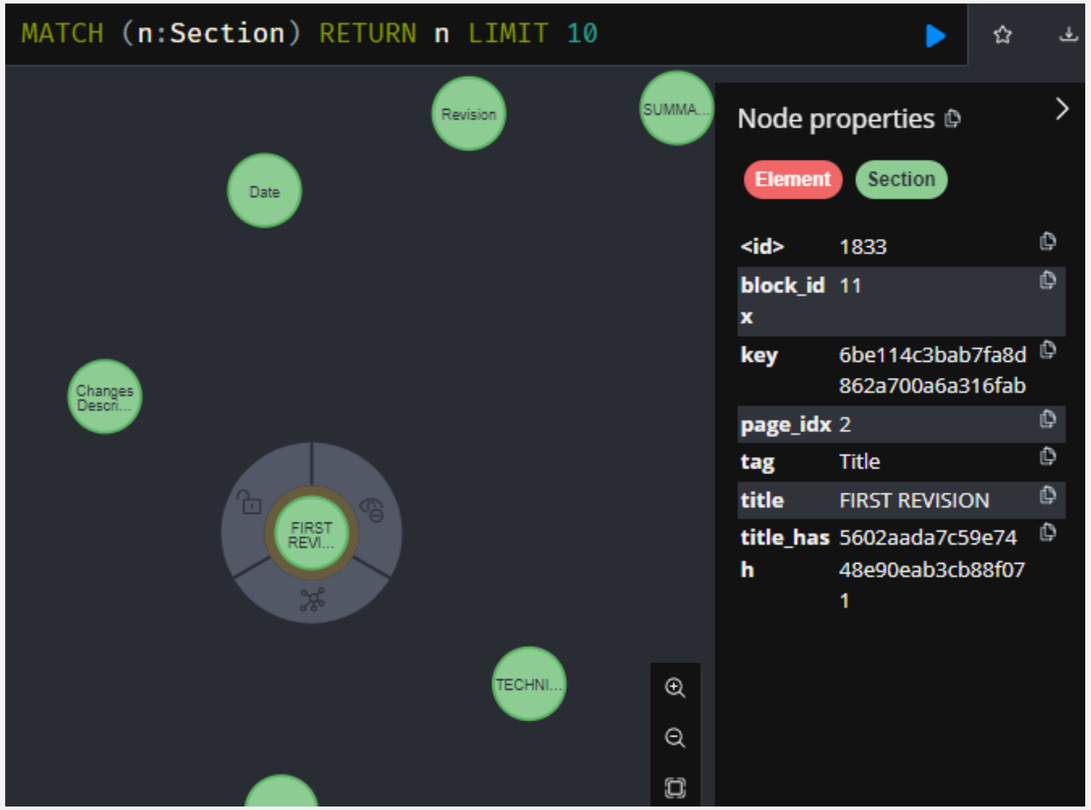




Graph : Section Node

- Creation of nodes grouping the chunks.

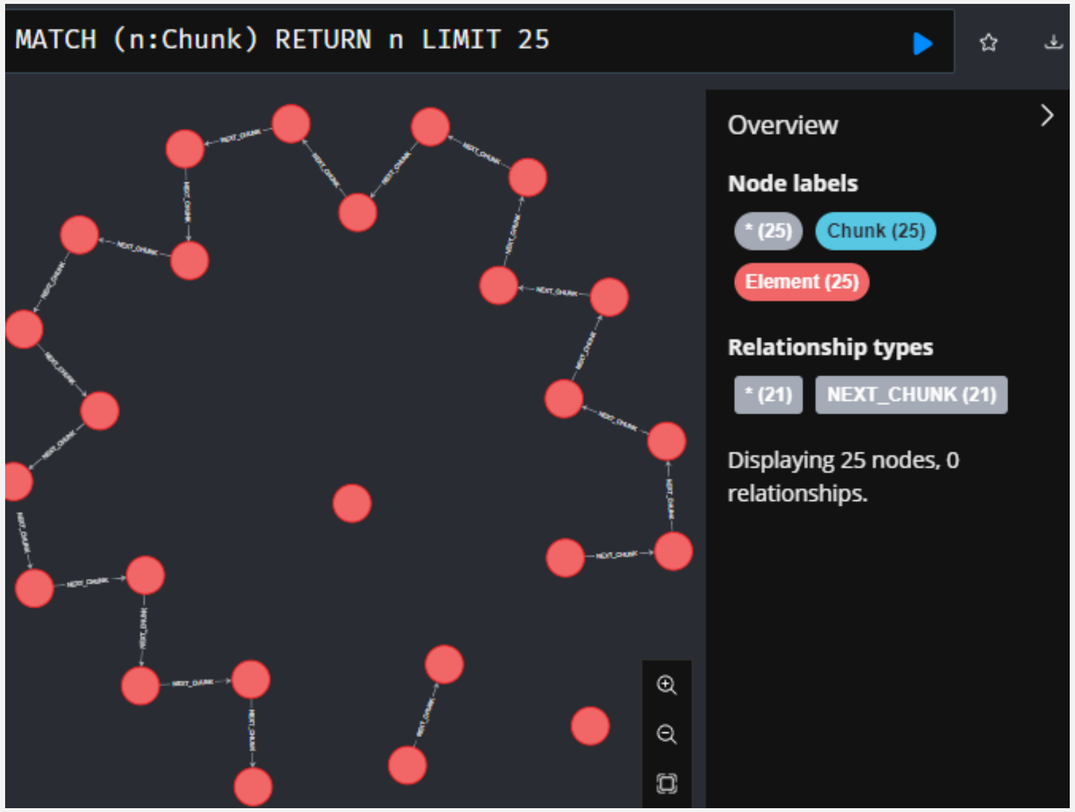
In a single PDF, there can be multiple sections, leading to the creation of the Section node.



Graph : Chunk Node

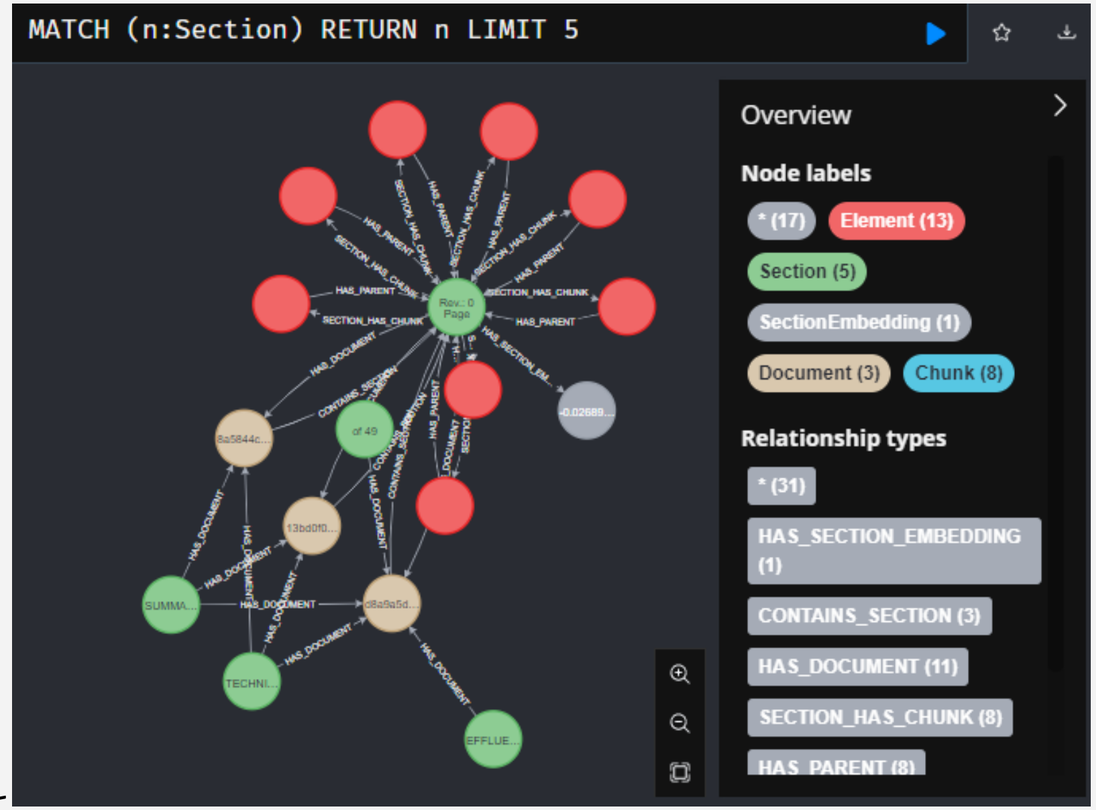
- Creation of nodes representing the chunks.

In a single PDF, there can be multiple Chunks. The 'NEXT CHUNK' relationship is used to identify, for a Chunk N, the previous Chunk (N-1) and the next Chunk (N+1).



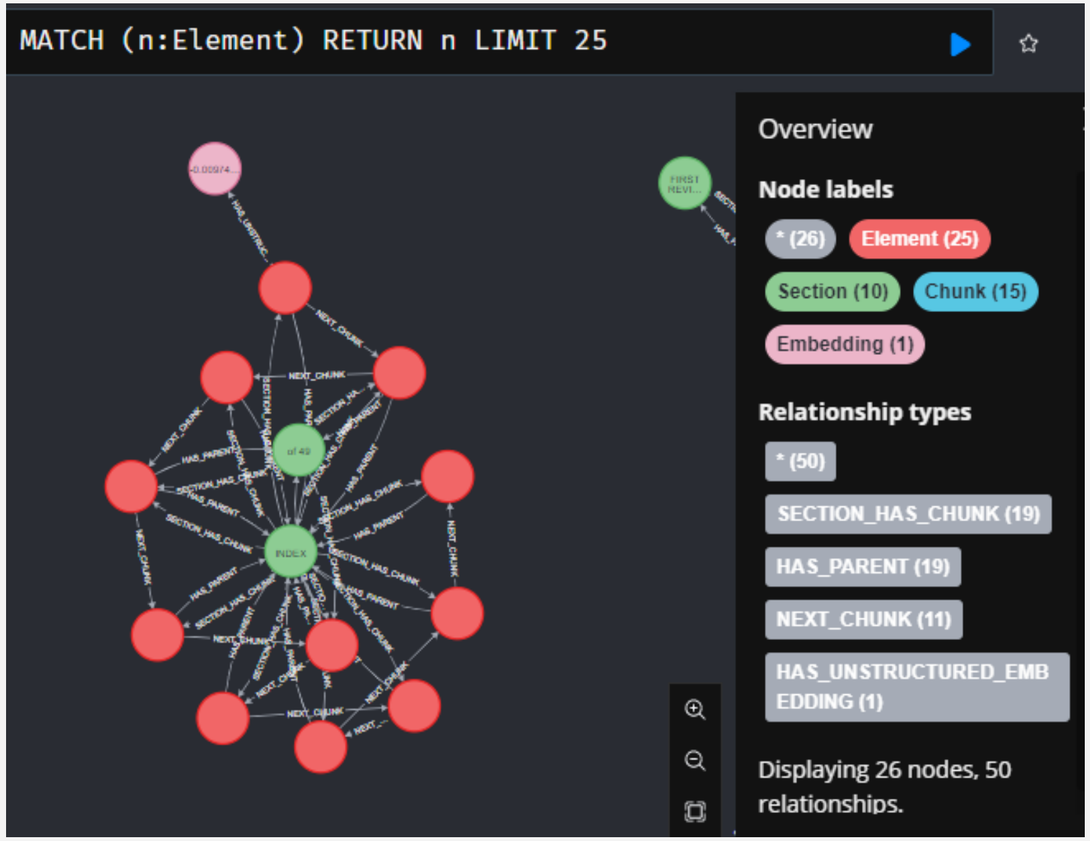
Graph : Hiearchy Cluster

- 4 Nodes:
  - Chunk, Document, Section, Section Embedding
- 5 relationships :
  - Contains section (Doc - section)
  - Has Document (Sec - Doc)
  - Section Has Chunk (Sec - Chunk)
  - Has Parent (Chunk - Section)
  - Has Section Embedding (Section - Embedding)



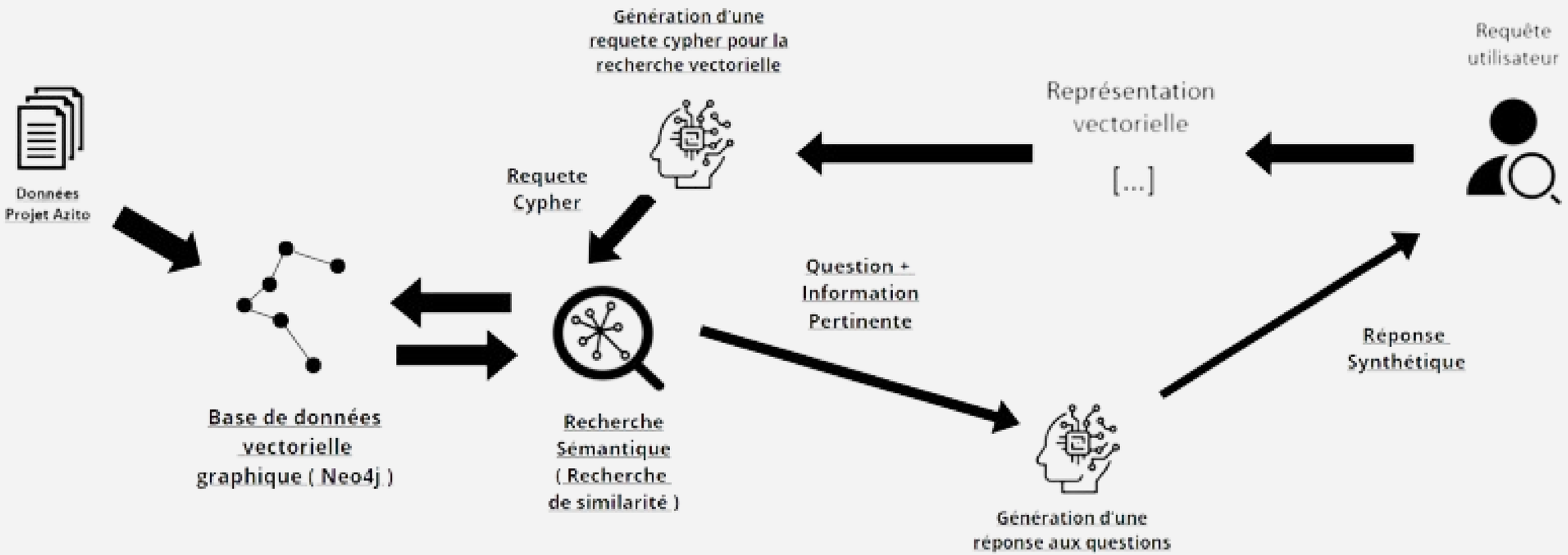
Graphe : Hiearchy Cluster

- 4 Nodes:
  - Chunk, Embedding, Section
- 5 relationships :
  - Section Has Chunk (Sec - Chunk)
  - Has Parent (Chunk - Section)
  - Has Unstructured Embedding (Section - Embedding)
  - Next Chunk (Chunk - Chunk)

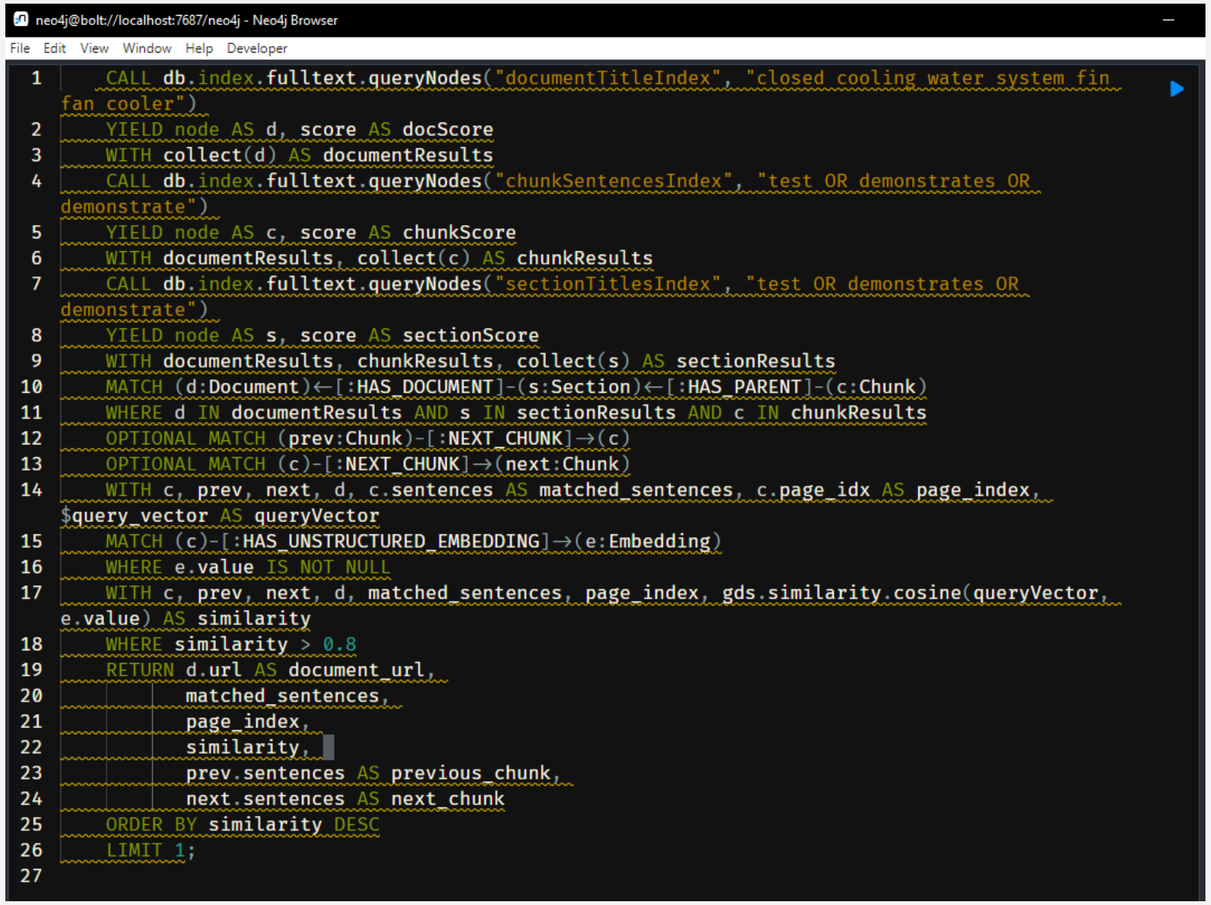


RAG System

Project Design



Query Generation



We extract the keywords from the question and filter the corresponding documents and sections. With the relevant results, we perform a semantic search to identify the chunks containing the answer.

Retrieval Result

```
MATCH (c)-[:HAS_UNSTRUCTURED_EMBEDDING]→(e:Embedding)
WHERE e.value IS NOT NULL
WITH c, prev, next, allChunks, d, matched_sentences, page_index, gds.similarity.cosine(queryVector, e.value) AS similarity
WHERE similarity > 0.8
RETURN
  d.url AS document_url,
  matched_sentences,
  page_index,
  similarity,
  prev.sentences AS previous_chunk,
  next.sentences AS next_chunk,
  collect(allChunks.sentences) AS all_section_chunks
```

document_url	matched_sentences	page_index	similarity	previous_chunk	next_chunk	all_section_chunks
"C:\Users\G22084\Desktop\Input_Doc\inputs_docs\experiment\AZ-20-YMV-PI-SPE-IDM-0002_00_Valves Technical Specification (RNC).pdf"	"Vendor shall provide an equipment of a well-proven design. The equipment shall be designed, manufactured and inspected in accordance with the applicable standards and comply with the terms and conditions of this document as regards to the requested performance, design criteria, construction requirements, guarantees and time schedule."	8	0.8994791194811631	"The equipment shall also be designed so that all provisions for easy access for inspection, cleaning, maintenance and repair are accounted for. The inspection facilities and openings required for this are to be included as well."	"The recommended materials indicated in document “AZ-20-YMV-PI-SPE-IDM-0003 Valve Class” are considered as being of the minimum quality desirable in the design and manufacture of the equipment. However, the final selection of the equipment materials will be the responsibility of the"	["The recommended materials indicated in document “AZ-20-YMV-PI-SPE-IDM-0003 Valve Class” are considered as being of the minimum quality desirable in the design and manufacture of the equipment. However, the final selection of the equipment materials will be the responsibility of the", "The scope of supply must include all the components required to ensure co

Using this method, we retrieve the document's URL containing the result, the page index of the result, the chunk containing the result, as well as the chunks before and after, and finally, all the chunks of the section where the result is located.

Response Generation

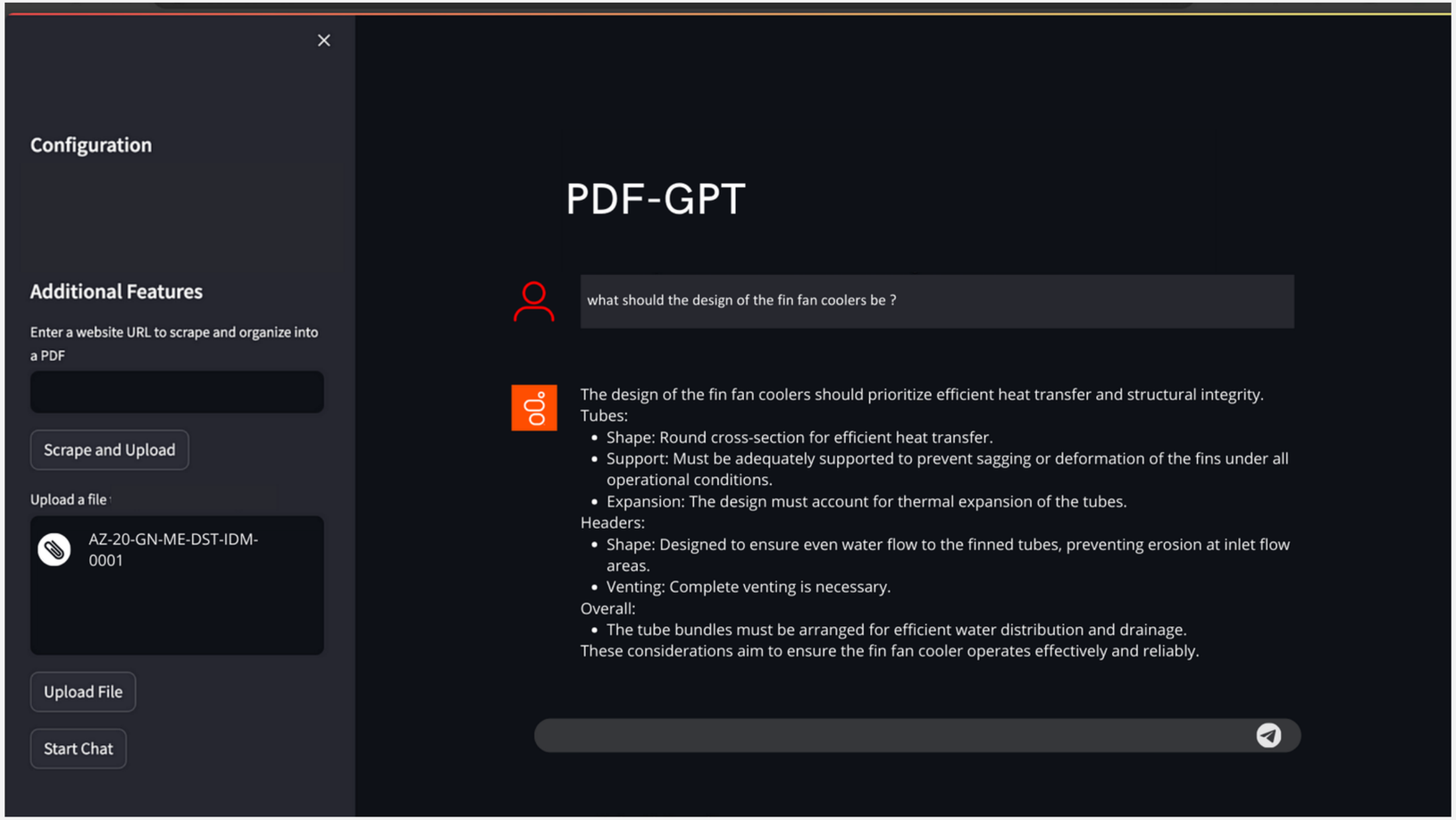
```
#the conversation :
the_narrow_context,the_broader_context,page_index,document=respond_to_question(user_question)
text1= f"""
i will give you a question, a narrow context and a broader context, formulate a comprehensive response on the question only based on
the contexts i provided, don't add any additional information :
the question :  {user_question}
the narrow context : {the_narrow_context}
the broader context : {the_broader_context}
"""
response_query = generate_response()
```

The question : User's Question
The narrow context : previous chunk + result chunk + Next Chunk
The broader context : All the chunks of the section where the result is found.

Results

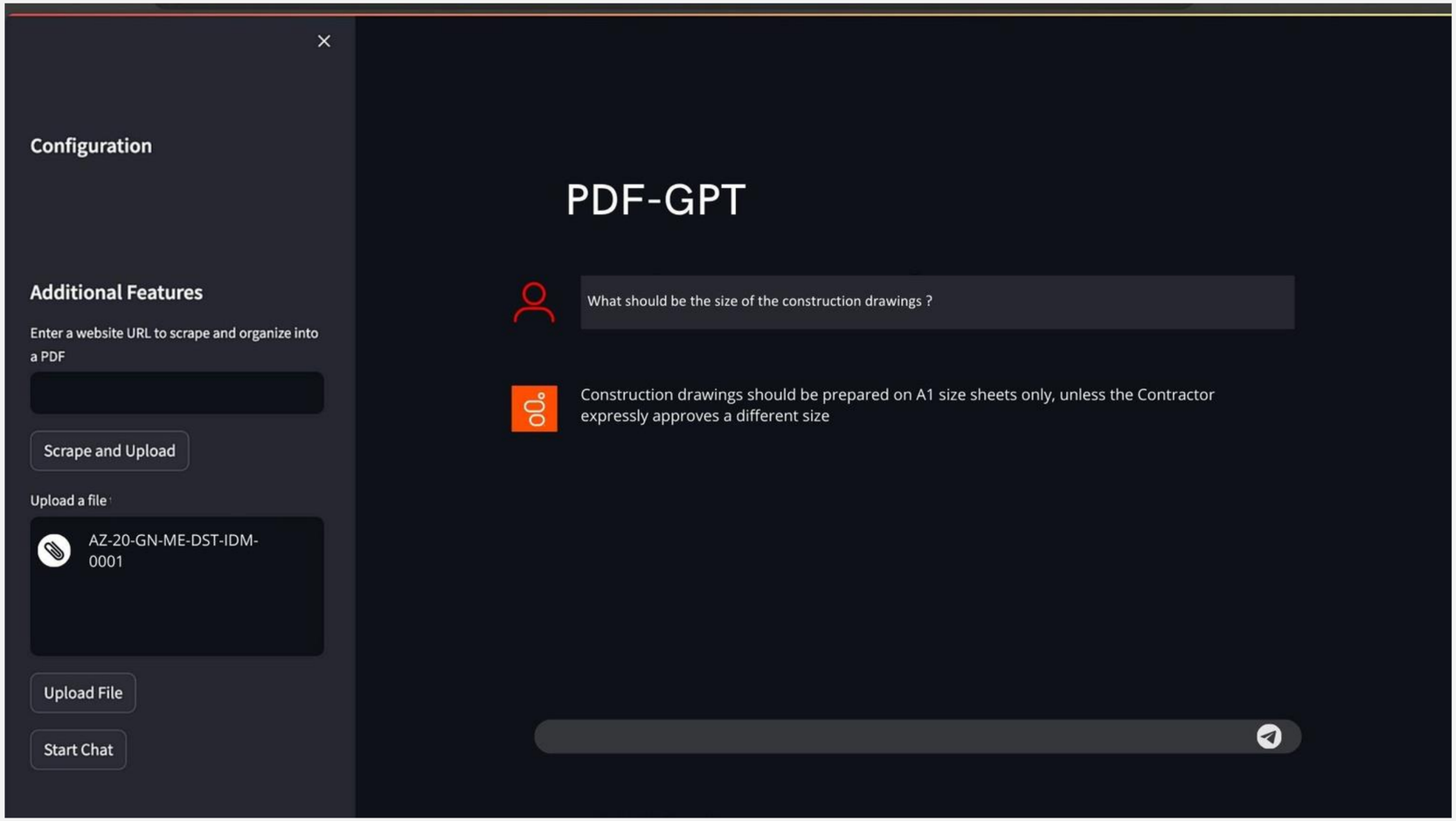
General Questions

- General Question



Specific Questions

- Specific Question



Thank you

Questions ?