

Towards a RAG based Summarization AI for EIC

Karthik Suresh, Cristiano Fanelli
College of William and Mary
AI4EIC-2023



WILLIAM & MARY

CHARTERED 1693

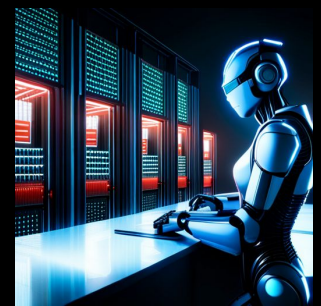


Image generated by Stable Diffusion XL
From [1]

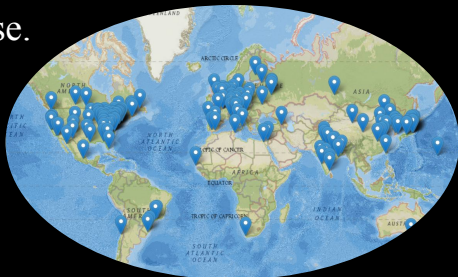
Outline

- Need for Summarization
- The envisioned pipeline for Retrieval Augmented Generation using LLM
- Data and embedding methods
- Some example summarizations
- Evaluating the summarizations using quantitative metrics
- Ongoing and next steps

Why RAG based LLM for Summarization?

Why a summarization tool?

- Ease of retrieving up to date information on topics.
- Extremely useful for new collaborators and especially graduate students.
- Towards building a specialized science for EIC.
- Large Scale Experiments \propto Size of document database.



1400 physicists, 240 institutions and
38 countries....

Why LLM based summary?

- Increased contextual understanding posed by LLM
- Enhanced information retrieval and faster decision making.

The requirement?

- Should be free of hallucinations.
- Should understand and unfold all contexts using short prompts
- Provide up to date information without “being” constantly trained

Retrieval Augmented Generation using LLM pipeline

Ingestion

- Load data (**semantic**) both structured and unstructured from sources
- Split data in small repetitive chunks – **text corpus**
- **Embed** these chunks into a vector space using an embedding model
- Store these vectors in a database for retrieval later.

Inference

- Given a prompt, compute similarity index. Select the most closest vectors
- Choose a response template. Embed the vectors along with input prompt and feed into LLM
- Evaluate the response
 - Model fine-tuning
 - Build metrics for context tuning
- Multi Modal Output

A pilot study – Ingestion

- The Objective – Summarization tool with “relevant citations” for within EIC.
- Framework built has to be scalable and fairly automated

Data loading



200 recent ArXiv publications
Source files and PDFs

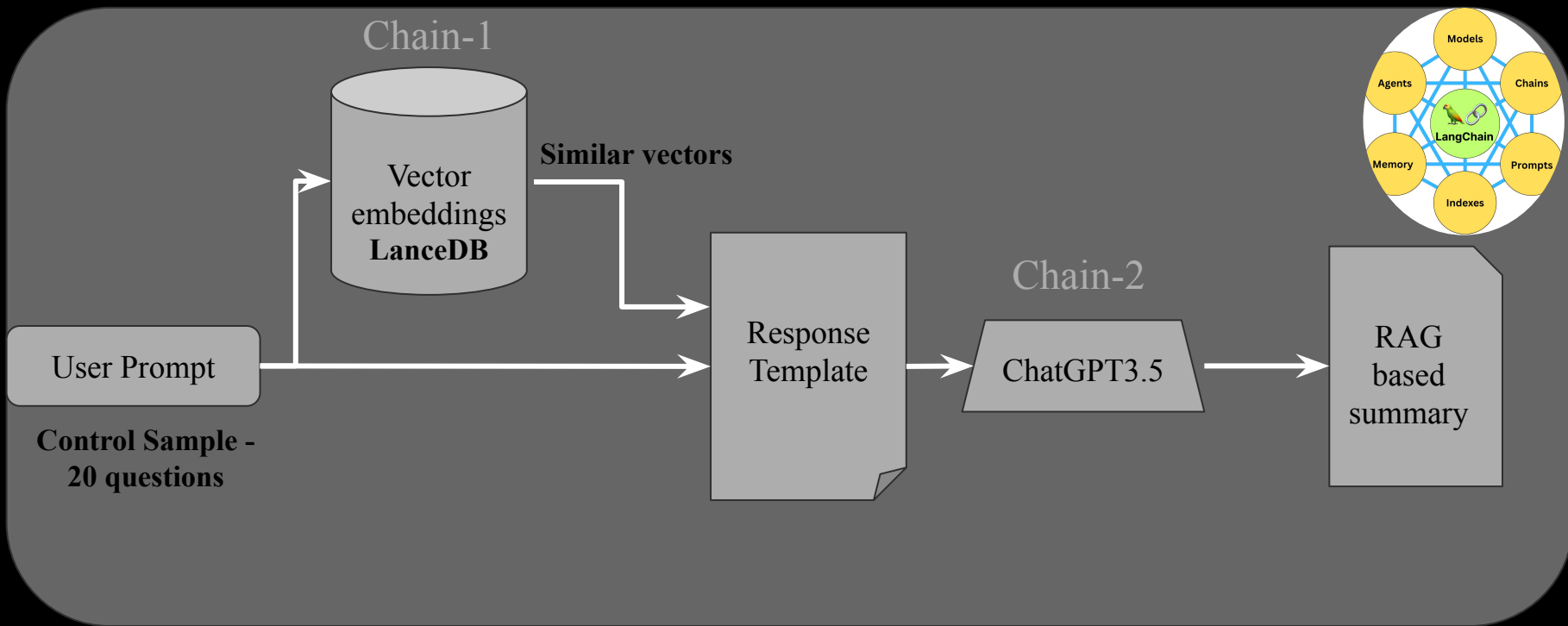
Split in Chunks

- **PyPDF** Reader to read texts from PDF files
- **Latex splitter** – tag based splitting to create chunks. Had to extend and build
- Equations selected as an object in Latex splitter
- Storing Figure location and table location as metadata while captions are split in chunks – can create multi modal output*
- **upto 120 char chunks with 10 char overlaps**

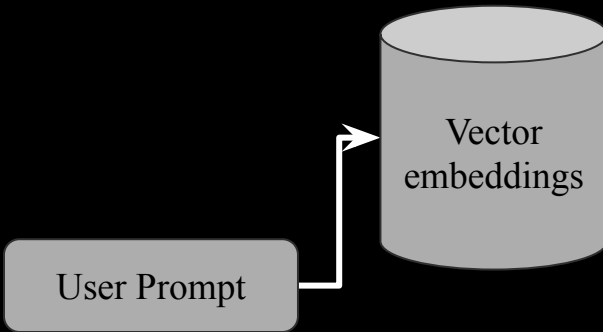
Vector embedding and storage

- text-embedding-ada-002
- 1536 size vector with cosine similarity
- Get first 100 similar vectors
- LanceDB – Lightweight local DB. Ideal for prototyping.
- Scalable solution PineCone

A pilot study – inference



A pilot study – inference – LateX vs PyPDF Splitter



2.1 EIC project detector reference design

The selected EIC project detector reference design consists of a Monolithic Active Pixel Sensor (MAPS) [5] based silicon vertex and tracking detector, a Micro Pattern Gas Detector (MPGD) [6] based tracking detector, an AC coupled Low Gain Avalanche Diode (AC-LGAD) based Time of Flight (ToF) detector, a dual Ring-imaging Cherenkov detector (dRICH), a mirror Ring-imaging Cherenkov detector (mRICH), a Detector of Internally Reflected Cherenkov light (DIRC) PID detector, ElectroMagnetic Calorimeters (EMCal) and Hadronic Calorimeters (HCAL). This proposed detector reference design utilizes the existing Babar magnet with a maximum magnetic field at 1.4 T. It can provide precise primary and displaced vertex determination, tracking reconstruction, particle identification and energy measurements in the pseudorapidity region of $-3.5 < \eta < 3.5$. The layout of the EIC project detector reference design is shown in the left panel of Fig. 1.

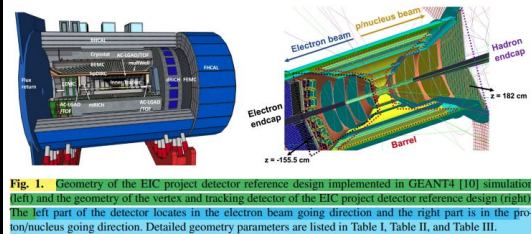


Fig. 1. Geometry of the EIC project detector reference design implemented in GEANT4 [10] simulation (left) and the geometry of the vertex and tracking detector of the EIC project detector reference design (right). The left part of the detector locates in the electron beam going direction and the right part is in the proton/nucleus going direction. Detailed geometry parameters are listed in Table I, Table II, and Table III.

2305.15593v1

Similarity index.

Question 1: Where will EIC experiment be built?
LateX Splitter – 0.37; PyPDF Splitter – 0.32

Question 2: How are dRICH detectors optimized at EIC?
LateX Splitter – 0.26; PyPDF Splitter – 0.21

Question 3: Give me the latest update on EIC tracker
LateX Splitter - 0.35, PyPDF Splitter – 0.29

- PyPDF Splitter splits based on text only, while LateX splitted splits based on “tags”
- Figure caption is tagged separately

Setting context

Setting the appropriate context with LLM.

Crucial in getting desired outcome.

Ongoing efforts to come up with Zero Shot Prompt for the summarization task

```
from langchain.prompts import PromptTemplate
```

```
template2 = """\
```

```
You are an expert in providing up to date information about the Electron Ion Collider (EIC), tasked with answering any question \ about EIC based only on the provided context.
```

```
Generate a comprehensive, and informative answer of 100 words or less for the \ given question based solely on the provided search results (URL and content). You must \ only use information from the provided search results. Use an unbiased and \ journalistic tone. Combine search results together into a coherent answer. Do not \ repeat text. You should use bullet points in your answer for readability. Make sure to break down your answer into bullet points.\ You should not hallucinate nor build up any references, Use only the `context` html block below and its associated `ARXIV_ID` if you find the context relevant. Make sure not to repeat the same context. Be specific to the exact question asked for.\ After each bullet point, cite up to 5 most relevant arxiv_id associated with the `context` html block from which the bullet point was generated. \ The citations should be taken from the context between the tags <ARXIV_ID> and <ARXIV_ID/>. Only quote relevant arxiv_id \
```

```
Only quote the most relevant arxiv_id that you find from `context` block. Note that the contexts are numbered according to the cosine similarity index.\ Place these citations at the end of the sentence or paragraph that reference them. If different results refer to different entities within the same name, write separate \ answers for each entity.
```

```
If there is nothing in the context relevant to the question at hand, just say "Hmm, \ I'm not sure." Don't try to make up an answer. Write the answer in the form of markdown bullet points.\ Make sure to highlight the most important key words in red color. Be precise and concise.\
```

```
Anything between the following `context` html blocks is retrieved from a knowledge \ bank, not part of the conversation with the user. The context are numbered based on its knowledge retrieval and increasing cosine similarity index. \ After each context the arxiv_id is given within the <ARXIV_ID> html block. \ Make sure to consider the order in which they appear context appear. It is an increasing order of cosine similarity index.\ The contents are formatted in latex, you need to remove any special characters and latex formatting before coercing the points to build your answer.\ You will cite no more than 5 citations from the context below.\ Make sure these citations have to be relevant as well as not repetitive in nature.
```

```
<context>  
  {context} <ARXIV_ID> {arxiv_id} <ARXIV_ID/>  
</context/>
```

```
REMEMBER: If there is no relevant information within the context, just say "Hmm, I'm \ not sure." Don't try to make up an answer. Anything between the preceding `context` \ html blocks is retrieved from a knowledge bank, not part of the conversation with the \ user.\
```

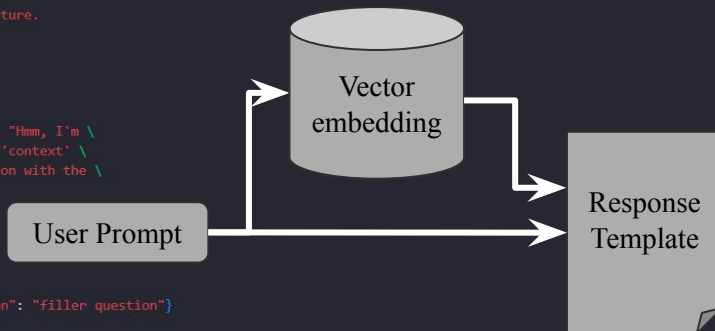
```
Question: {question}
```

```
"""
```

```
rag_prompt_custom = PromptTemplate.from_template(template2)
```

```
print (rag_prompt_custom.invoke(  
  {"context": "filler context", "arxiv_id": "filler arxiv_id", "question": "filler question"}  
).to_string())
```

```
✓ 0.0s
```



A targeted example

How to evaluate this response?

Ragas score

1. Faithfulness –

$$\text{Faithfulness score} = \frac{|\text{Number of claims that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

2. Answer relevance –

$$\text{Answer relevance} = \frac{|\text{\#claims correctly responded}|}{|\text{\#claims in given text}|}$$

3. Context relevance –

$$\text{Context relevance} = \frac{|\text{\# relevant unique refs}|}{|\text{\# refs retrieved}|}$$

4. Answer correctness – Qualitative. Need ground truth.

Question 3: What is the pixel size and expected momentum resolution at 2GeV from a MAPS ITS3 tracking layer

Iteration 1

- The pixel size of the MAPS ITS3 tracking layer is 10 μm
<ARXIV_ID><http://arxiv.org/abs/2108.08314v1></ARXIV_ID> <https://arxiv.org/abs/2108.08314v1>
<ARXIV_ID><http://arxiv.org/abs/2203.07626v2></ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2302.01447v1></ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2305.15593v1></ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2210.08609v2></ARXIV_ID>
- The expected momentum resolution at 2 GeV from a MAPS ITS3 tracking layer is 2.5% to 15% of the initial momentum <ARXIV_ID><http://arxiv.org/abs/2305.15593v1></ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2210.08609v2></ARXIV_ID> <https://arxiv.org/abs/2210.08609v2>

$$\text{Faithfulness} = 2/2 = 1.$$

$$\text{Answer relevance} = 1/2 = 0.5$$

$$\text{Context relevance} = 5/7 \sim 0.7$$

$$\text{Answer correctness} = 1/2 = 0.5$$

Another more complex full example

How to evaluate this response?

Ragas score

1. Faithfulness ~ 4/25
 2. Answer relevance – 1
 3. Context relevance – 1
 4. Answer correctness – Qualitative
- Repeated this exercise for 50 questions mainly on the EIC detector topics
 - Targeted questions are relatively easier to quantify and evaluate.

# of Hallucination	Faithfulness	Answer relevance	Context relevance	Answer correctness
3/50	26%	62%	100%	Qualitative

Question 3: What are the dimensions of MAPS detector used in EIC?

Response

- The MAPS detector used in the EIC has dimensions ranging from 3.67 cm to 43 cm, with a sensor size of 10x10 cm <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>, <http://arxiv.org/abs/2302.01447v1>, <http://arxiv.org/abs/2210.04879v2>, <http://arxiv.org/abs/2305.02079v2>, <http://arxiv.org/abs/2302.01447v1>]
- The ePIC MAPS electron endcap detector has a geometry parameter listed, and the ePIC MAPS hadron endcap detector also has geometry parameters listed <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>, <http://arxiv.org/abs/2305.15593v1>, <http://arxiv.org/abs/2305.15593v1>, <http://arxiv.org/abs/2305.15593v1>]
- The ePIC MAPS vortex detector design can reach as low as 0.05, and the imaging layers of MAPS sensors provide excellent position resolution <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>, <http://arxiv.org/abs/2305.15593v1>]
- The MAPS sensor used in the EIC is based on low-mass, low-power consumption MAPS in 65 nm technology <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>]
- The MAPS sensor used in the EIC is based on low-mass, low-power consumption MAPS in 65 nm technology <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>]

<https://arxiv.org/abs/2203.07626v2>

<https://arxiv.org/abs/2302.01447v1>

<https://arxiv.org/abs/2210.04879v2>

<https://arxiv.org/abs/2305.02079v2>

<https://arxiv.org/abs/2305.15593v1>

Conclusion and next steps

Methodology Development:

- Focused on developing a robust evaluation methodology for the EIC dataset and its corresponding answers.
- Developing a frontend using lang-serve and lang-smith to better improve call tracing and user experience
- Working on coercing figures and table metadata to produce a multi-modal summary report.

LLM as Judge for Evaluation^[1]:

Currently pursuing the LLM as Judge solution, recognizing the importance of base truth in specific cases. The idea is to reduce the total number of evaluations using base truth.

Summarization Assistant Proof of Concept:

Successfully presented the initial proof of concept for a potential Summarization assistant.

Back ups

A full example

```
ques = r"What are the dimensions of MAPS detector used in EIC?"
print ("# Question 3: " + ques)
for i in range(5):
    print (f"## Iteration {i+1}")
    for chunk in rag_chain_with_source.stream(ques):
        if(chunk.get("answer")):
            print(chunk["answer"], end="", flush=True)
    print ("\n")
```

✓ 28.6s

Python

Question 3: What are the dimensions of MAPS detector used in EIC?

Iteration 1

- The MAPS detector used in the EIC has dimensions ranging from 3.67 cm to 43 cm, with a sensor size of 10x10 cm, and a length of 2.4 m. The detector consists of 5 layers and utilizes (

Iteration 2

- The MAPS detector used in the Electron Ion Collider (EIC) has dimensions ranging from 3.67 cm to 43 cm, with the electron endcap detector having a geometry of 23 cm to 43 cm. The det

Iteration 3

- The MAPS detector used in the EIC has dimensions ranging from 3.67 cm to 43 cm, with various inner and outer radii for different components of the detector. The ePIC MAPS electron en

Iteration 4

- The MAPS detector used in the EIC has dimensions ranging from 3.67 cm to 43 cm, with a sensor size of 10x10 cm <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>], <http://arxiv.org/abs/2203.07626v2>]
- The ePIC MAPS electron endcap detector has a geometry parameter listed, and the ePIC MAPS hadron endcap detector also has geometry parameters listed <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>]
- The ePIC MAPS vertex detector design can reach as low as 0.05, and the imaging layers of MAPS sensors provide excellent position resolution <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>]
- The MAPS sensor used in the EIC is based on low-mass, low-power consumption MAPS in 65 nm technology <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>]
- The MAPS sensor used in the EIC is based on low-mass, low-power consumption MAPS in 65 nm technology <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>]

Iteration 5

- The MAPS detector used in the EIC has dimensions ranging from 3.67 cm to 43 cm, with a sensor size of 10x10 cm, and a length of 2.4 m. The detector also features a 65 nm Monolithic A