

Towards a RAG based Summarization AI for EIC

[DOI 10.1088/1748-0221/19/07/C07006](https://doi.org/10.1088/1748-0221/19/07/C07006)

Karthik Suresh



College of William and Mary

WILLIAM & MARY

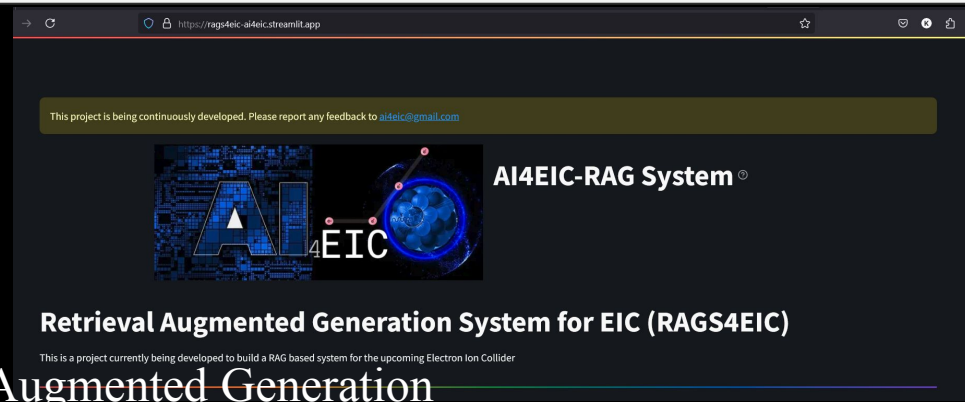
CHARTERED 1693



Image generated by Stable Diffusion XL
From [1]

Outline

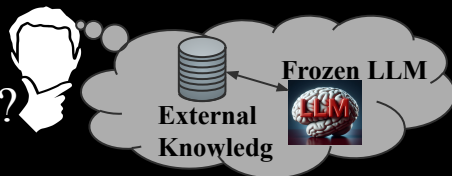
- Need for Summarization
- The envisioned pipeline for Retrieval Augmented Generation using LLM
- Data and embedding methods
- Some example summarizations
- Evaluating the summarizations using quantitative metrics
- Methods of RAG to be explored
- The further works and vision for this project



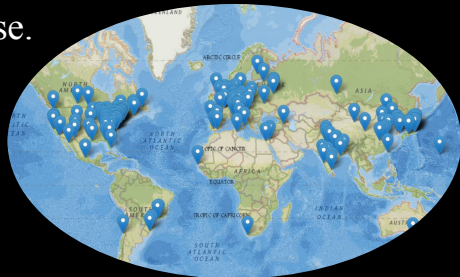
[GitHub Repo](#)

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 under and unfold all contexts using short prompts
- Provide up to date information without “being” constantly trained

An Ongoing HEP effort

chATLAS

An AI Assistant for the
ATLAS Collaboration

DANIEL MURNANE, GABRIEL FACINI,
RUNZE LI & CARY RANDAZZO

AI4EIC WORKSHOP
NOVEMBER 30, 2023



BERKELEY LAB

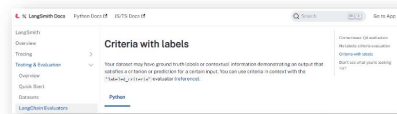


LOUISIANA TECH
UNIVERSITY

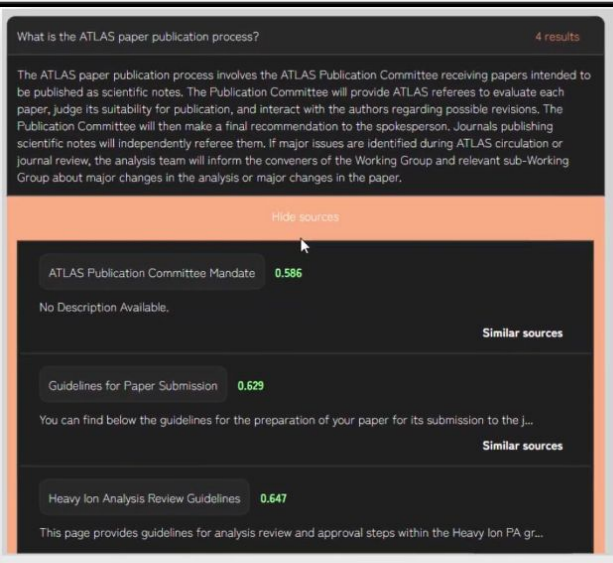
[Talk at AI4EIC-2023](#)

OPEN QUESTIONS & CONCLUSION

- How to avoid hallucinations? Probably “GPT-5 / Q* / AGI” will make this hurdle irrelevant
- How to best “censor” politically incorrect responses (e.g. which analysis team is the best?)
- How to **measure** the quality of responses – LangSmith AI-assisted evaluators?
- What is the best dataset to gather for fine-tuning?
- How to anonymize email threads and discussion forums?



We are having a lot of fun building this thing from scratch, but if there was an open-source scientific community framework for AI Assistants, it would be even more fun!



Retrieval Augmented Generation using LLM pipeline

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

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

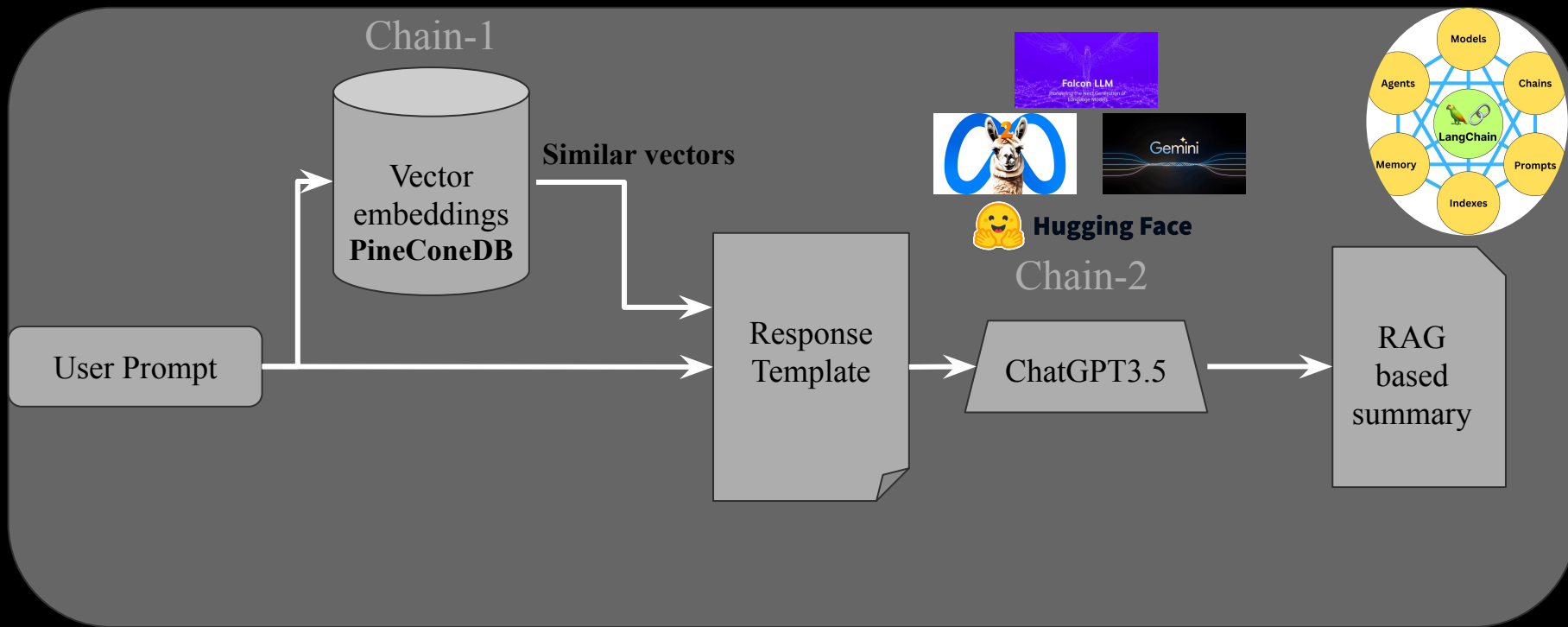
Split in Chunks

- **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**

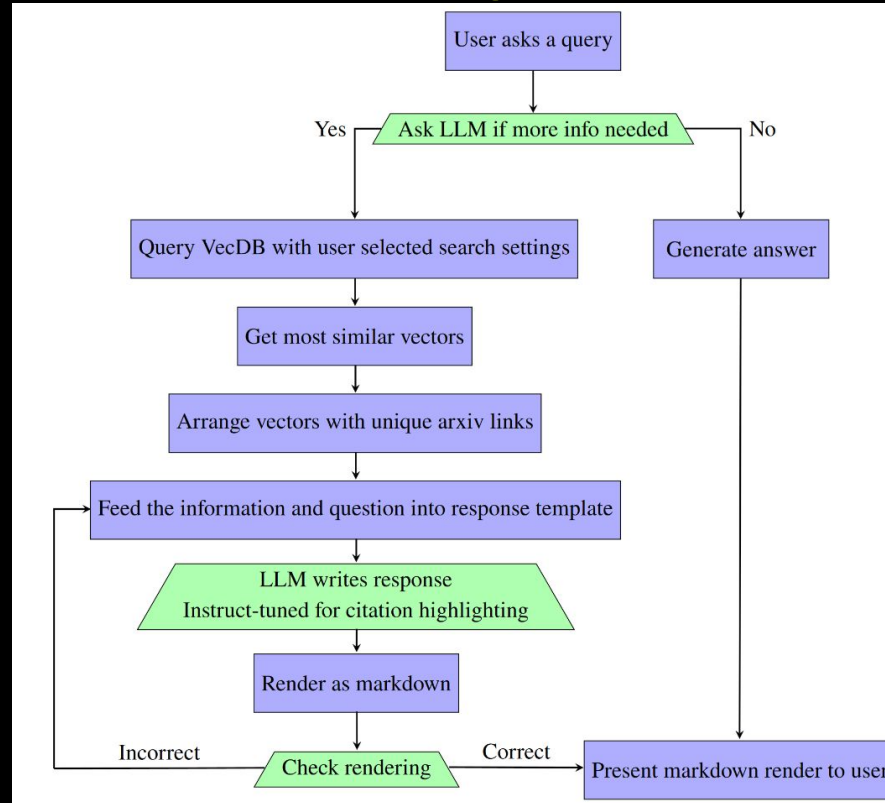
Vector embedding and storage

- 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

The Generic Workflow of building Traditional RAG



The Generic Workflow of building Traditional RAG



The built RAG application

[Check out the full trace of the pipeline in langsmith](https://rags4eic-ai4eic.streamlit.app/RAG-ChatBot)

The screenshot shows a web browser window with the URL `https://rags4eic-ai4eic.streamlit.app/RAG-ChatBot`. The application has a dark theme and a sidebar on the left. The sidebar contains the following elements:

- AI4EIC-RAGAS4EIC
- Login or Signup
- RAG-ChatBot** (highlighted)
- Generate Questions
- View History
- Select VecDB and Properties
- Vector DB: PINECONE
- Retrieval Metric: Cosine similarity
- Max K: A slider set to 10, with values 3 and 120 visible.
- Select Vector DB

The main chat area on the right displays the title "AI4EIC - RAG QA-ChatBot" with a logo. Below the title, a user query is shown in a grey bubble: "What are the polarization parameters used to study J/ψ polarization in SIDIS at large PT ?". A bot response follows in a blue bubble: "Gathering info from Knowledge Bank for this query...". The response text is: "Hello! Here are the polarization parameters used to study J/ψ polarization in SIDIS at large PT :" followed by a bulleted list:

- The study of J/ψ polarization in large- PT semi-inclusive deep-inelastic scattering (SIDIS) at the Electron Ion Collider (EIC) can help in getting information on the J/ψ formation mechanism, both at large z (high- and low-energy set-ups) and as a function of PT (at large energy)[1[^]].
- Future measurements of transversely polarized $\Lambda/\bar{\Lambda}$ in SIDIS will potentially allow us to gain further insights and to distinguish between different scenarios, contributing to the understanding of the J/ψ polarization in SIDIS at large PT [2[^]].

The response concludes with "I hope this helps to answer your question." and two links:

1. <http://arxiv.org/abs/2301.11987v1>
2. <http://arxiv.org/abs/2301.11987v1>

Below the links is a blue link "View the trace" with a small icon. At the bottom, there is a text input field with the placeholder "What is up? Ask anything about the Electron Ion Collider (EIC)" and a send button.

The built RAG application – Feedback for fine tuning later

Has the output been displayed properly?

Choose an option ▼

Rate the quality of the output, Min = 0, Max = 5

0 – +

What was your expected response?

?

What is an ideal response for this query?

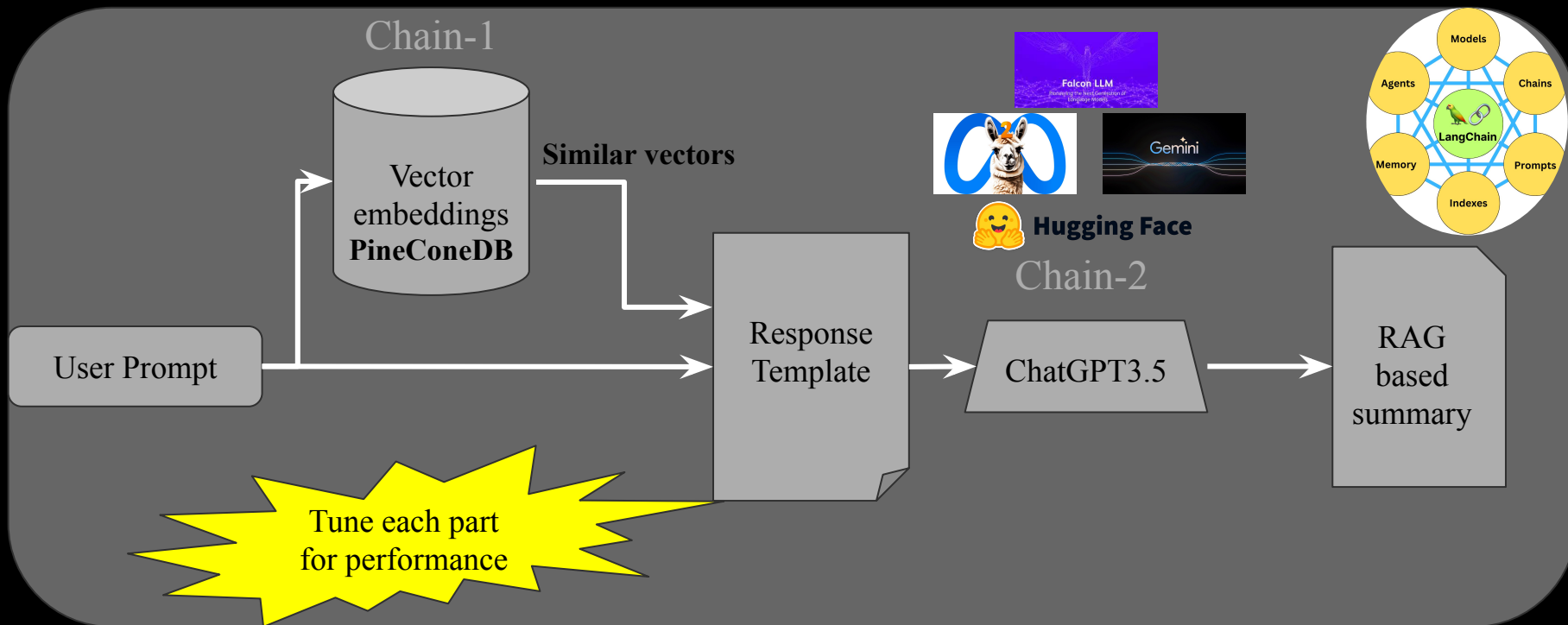
What was the source of the question?

?

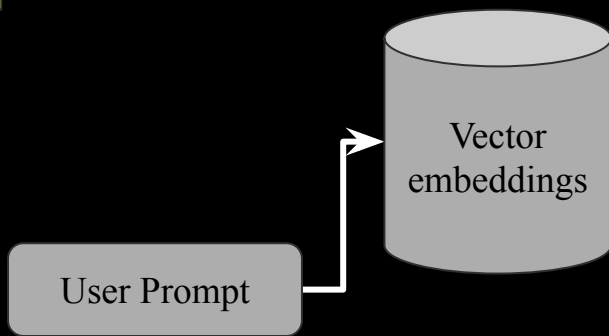
Where did you get this query from, ideally this is the arxiv_id that is used for testing purpose?

Submit Feedback

The Generic Workflow of building Traditional RAG



1. Ingestion tuning – 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.

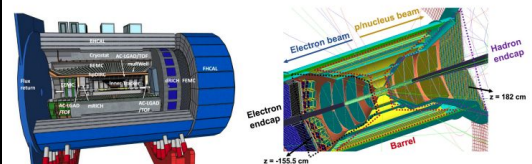


Fig. 1. Geometry of the EIC project detector reference design implemented in GEANT4 [10] simulation (left) and the geometry of the vertex and tracking detector of the EIC project detector reference design (right). The left part of the detector locates in the electron beam going direction and the right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

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”
- Length of each
- Figure caption can be tagged separately

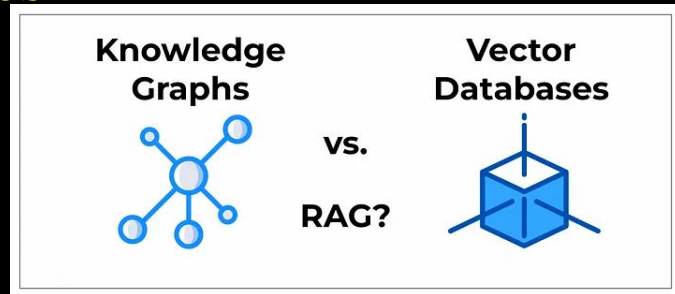
2. Retrieval tuning – Similarity searches

Traditional Vector DB –

- uses similarity indices to retrieve information
- k-Nearest Neighbours
- Maximal Marginal Relevance – reduces redundant phrases

Graph Vector DB -

- Uses Knowledge graph for better contextual retrieval
- Saves information as a node in a graph with connections defining the strength to other “words”/phrases

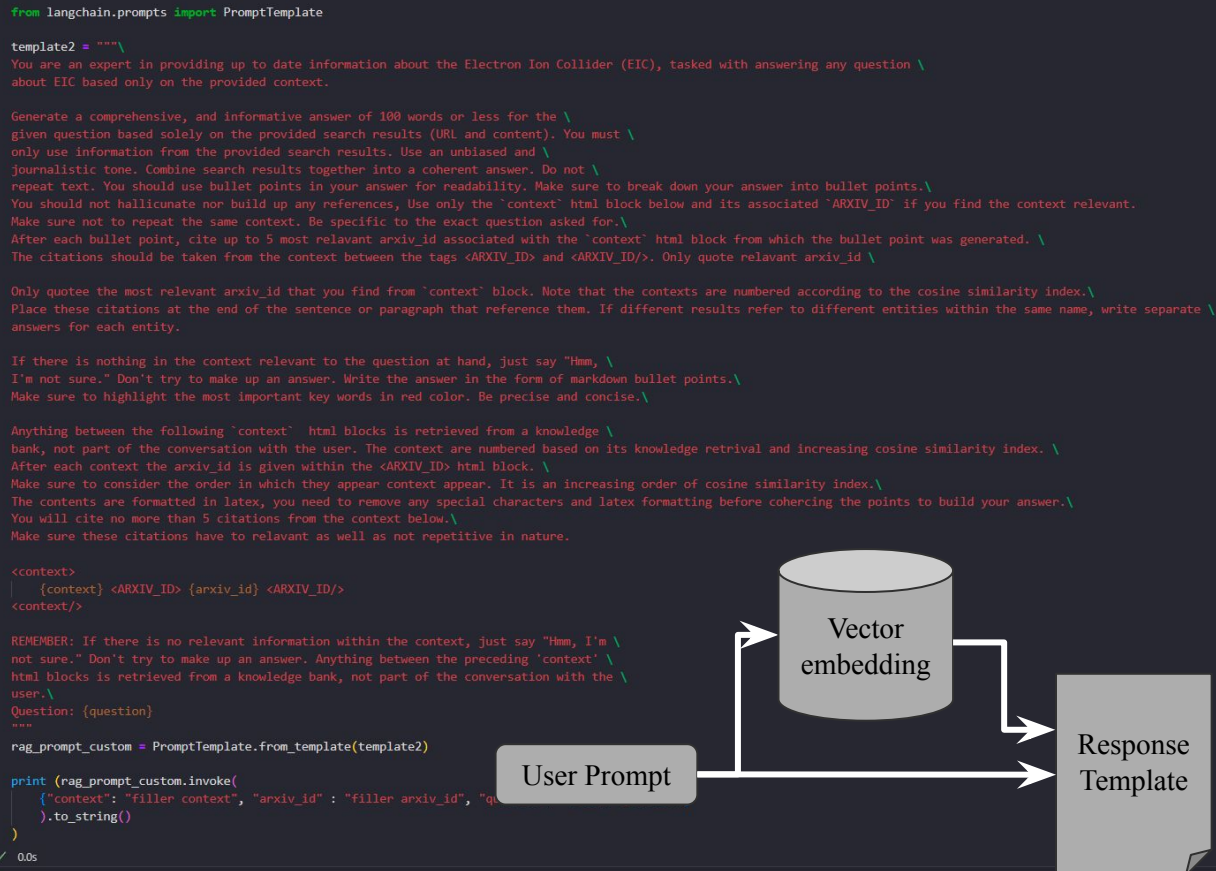


3. Context tuning

Setting the appropriate context with LLM.

Crucial in getting desired outcome.

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



[Checkout the response templates](#) (ai4eic github link)

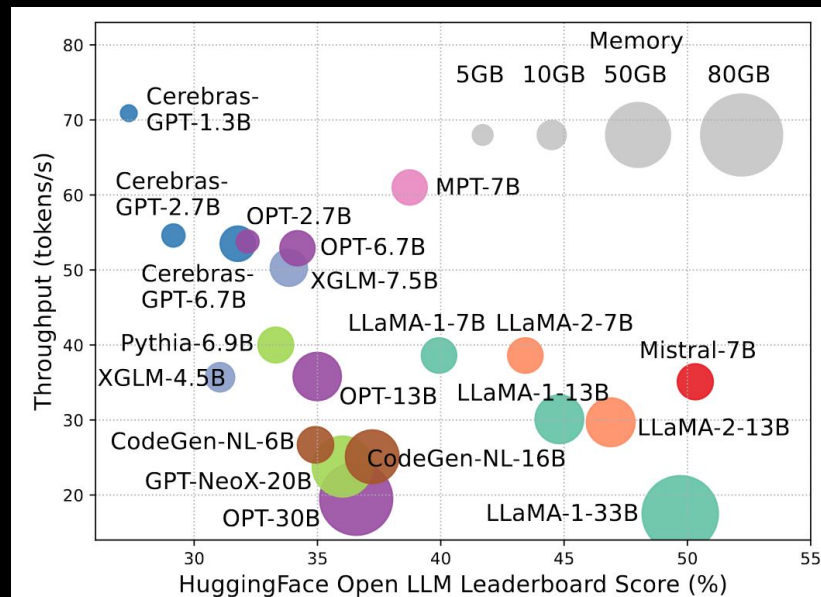
4. The LLM Model

Choice of LLM model as well affects the performance of the RAG system.

Small/Medium Language Models

O(10B) [OpenLLM models](#) are available

Need to consider Model complexity (resource utilization) vs performance.



[arxiv:2312.03863](https://arxiv.org/abs/2312.03863)

3. The RAG Approach

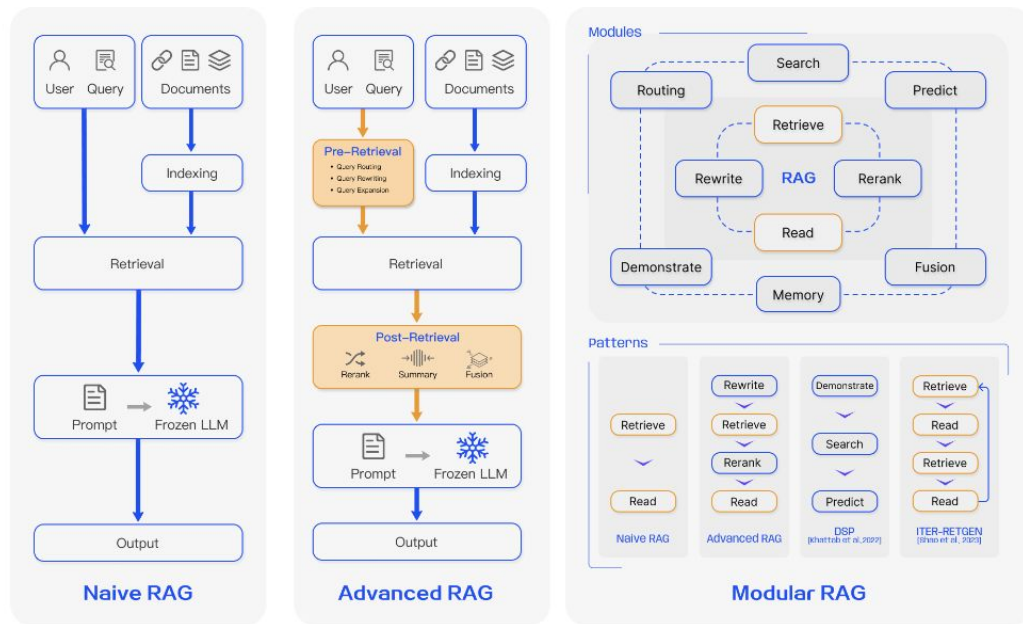


Figure 3: Comparison between the three paradigms of RAG

[arxiv:2312.10997](https://arxiv.org/abs/2312.10997)

Advanced RAG:
Using LLM more than a
orchestrator. Eg. CoT
based RAG

Modular RAG:
Using LLM for routing
logic. Give access to tools
like TWikiReader to
perform actions based on
results from LLM.

Evaluating RAG: 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>

Evaluating RAG:

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
- Targeted questions are relatively easier to quantify and evaluate.
 - Hence, Classic Metrics are metric that require human expertise while RAGAS metric can be automated using LLM

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 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>', '<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>

But can we use a LLM for QA Generation?

Use LLM for QA Generation.
[Set context for generating questions.](#)

Select an arxiv article. From the database

Specify the number of claims to generate, along with expected answers.

Using LLM to generate QA benchmarks dataset

Expand to see detailed explanation

Select GPT Version and load an Article from arxiv database to generate questions

☐ Select a Random Article if needed



Select GPT Version



4



ARXIV primary category ?

ARXIV title



physics.i...



The Optimal use of Segmentation for Sampling Calorimeters



Load Article from arxiv....

[Check out the trace here](#)

But can we rely on LLM for QA Generation?

Using a more complex LLM can generate questions along with answers.

LLM can automate the generation of questions to specific format which can be mixed to form complex questions.

[Question Data Set now can be evaluated](#)

Question 1 from 2310.04442v1 at <http://arxiv.org/pdf/2310.04442v1>

Q: What are the sampling fractions for the HCAL and ECAL as computed using a 40 GeV electron? How is the reconstructed energy (EReco) calculated using these sampling fractions? A:

```
{
  "n_claims" : 2,
  "claims": [
    ""sampling fractions for 'HCAL' and 'ECAL'""",
    ""calculation of 'EReco' using sampling fractions""
  ],
  "complete_response": "" The sampling fractions for the HCAL and ECAL are 2.2% and 3.0%, respectively.
                        The reconstructed energy (EReco) is calculated as the sum of the hit energy divided by th
                        expressed as  $EReco = (\text{Sum of } E_i \text{ in HCAL} / SF_{hcal}) + (\text{Sum of } E_i \text{ in ECAL} / SF_{ecal})$ , where S
                        """,
  "answers": [
    ""2.2% for HCAL and 3.0% for ECAL""",
    ""EReco = (Sum of  $E_i$  in HCAL / 2.2%) + (Sum of  $E_i$  in ECAL / 3.0%)""
  ]
}
```

[Check out the trace here](#)

Classic Metrics

Metric Name	Definition	Score
Output Renderer Frequency	Frequency of correctly rendered output response in the markdown format	78.0% \pm 5.8%
Claim Recognition Rate	$\text{CRR} = \frac{ \text{Number of answered claims} }{ \text{Total number of claims in question} }$	96.4% \pm 3.4%
Claim Accuracy Rate	$\text{CAR} = \frac{ \text{Number of correctly answered claims for a question} }{ \text{Total number of recognized claims in question} }$	88.9% \pm 8.3%
Source Citation Frequency	$\text{SCF} = \frac{ \text{Number of time question's source cited} }{ \text{Total queries} }$	85.3% \pm 5.0%
Hallucination Frequency	$\text{HF} = \frac{ \text{Number of hallucinations} }{ \text{Total queries} }$	2% \pm 2%

RAGAS – LLM based evaluation

Metric Name	Definition	Score
Faithfulness	Fraction of correctly rendered output response in the markdown format	87.4% \pm 5.5%
Context Relevance	Relevance of the retrieved context to the question and generated answer	61.4% \pm 4.3%
Context Entity Recall	$\text{CER} = \frac{ \text{claims in context} \cap \text{claims in ground truth} }{ \text{claims in ground truth} }$	98.7% \pm 1.2%
Answer Relevance	Measures relevancy of the generated answer to the retrieved context and the question	77.2% \pm 2.3%
Answer Correctness	Measures correctness of the generated answer to the idea response	72.3% \pm 2.4%

Next steps

- Have a mechanism to evaluate the built RAG pipeline. Each time a new RAG pipeline is created, repeat the Evaluation for its performance.
- Replace OpenLLM models.
- Agent based interactions to better direct output
- Generation of Golden datasets with ideal evaluations
- Building Splitters and Readers for TWiki.
 - Identify Experiments within which this can be integrated.
 - Eg. GlueX, AI4HallC are potential Candidates

Conclusion and final thoughts

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]:

LLM for evaluation is important 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. Potentially replacing traditional Document database for efficient search through information within EIC

Backups

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 dete

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

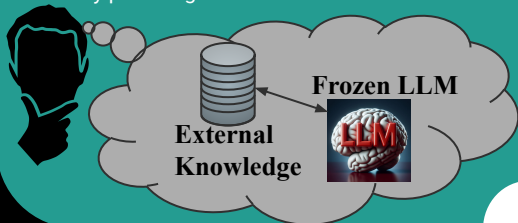
- The MAPS detector used in the EIC has dimensions ranging from 3.67 cm to 43 cm, with a sensor size of 10x10 cm <ARXIV_ID>[<http://arxiv.org/abs/2203.07626v2>], <http://arxiv.org/abs/2203.07626v2>]
- The ePIC MAPS electron endcap detector has a geometry parameter listed, and the ePIC MAPS hadron endcap 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

What is RAG ?

- Retrieval Augmented Generation (RAG).
- Access up to date information without explicitly training of LLM.
- Reduce “Hallucination” of LLM.
- Grounding LLM to truth to increase reliability by providing citations.



Why need RAG for Large Scale Physics Experiments?

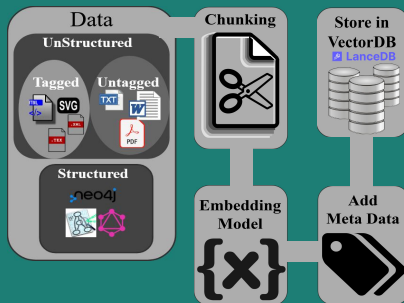
- Electron Ion Collider (EIC) is a large scale experiment.
- Regular updates to documents, Run Wiki
- Newbies may take 6 months to get to know the full experimental details.
- Document size \propto Scale of experiment



1400 physicists, 240 institutions and 38 countries

Ingestion

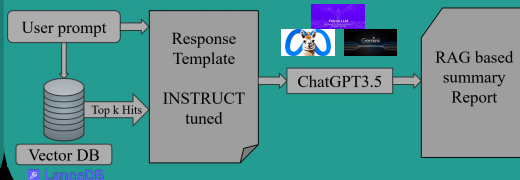
- Creation of the vectorized knowledge base.
- Every node below influence RAG performance
- 200 recent arxiv papers on EIC (since 2021)



The inference*



- Given a prompt compute similarity index to most similar vectors in VectorDB
- Use LLM to further narrow down and summarize the finding



*Naive RAG pipeline

The research approach

What we have built ?

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 MAMTIS tracking layer is 10 $\mu\text{m}/\text{cm}$
 $\langle \text{AROV}_i \rangle = \sqrt{\langle \text{AROV}_i^2 \rangle} = \sqrt{2100.00316} \text{ e}^{-1} \langle \text{AROV}_i \rangle$
 $\langle \text{AROV}_i \rangle = \sqrt{\langle \text{AROV}_i^2 \rangle} = \sqrt{2203.07626} \text{ e}^{-1} \langle \text{AROV}_i \rangle$
 $\langle \text{AROV}_i \rangle = \sqrt{\langle \text{AROV}_i^2 \rangle} = \sqrt{2302.01441} \text{ e}^{-1} \langle \text{AROV}_i \rangle$
 $\langle \text{AROV}_i \rangle = \sqrt{\langle \text{AROV}_i^2 \rangle} = \sqrt{2305.15591} \text{ e}^{-1} \langle \text{AROV}_i \rangle$
 $\langle \text{AROV}_i \rangle = \sqrt{\langle \text{AROV}_i^2 \rangle} = \sqrt{2210.08609} \text{ e}^{-1} \langle \text{AROV}_i \rangle$
- The expected momentum resolution at 2 GeV from a MAMTIS tracking layer is 2.5% to 15% of the initial momentum $\langle \text{AROV}_i \rangle = \sqrt{\langle \text{AROV}_i^2 \rangle} = \sqrt{2305.15591} \text{ e}^{-1} \langle \text{AROV}_i \rangle$

- Creation of benchmark evaluation dataset.
- Evaluate performance before scaling.

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 4.3 cm, with a sensor size of 10x10 cm. <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>
- The ePIC MAPS electron endcap detector has a geometry parameter listed, and the ePIC MAPS hadron endcap detector also has geometry parameters listed. <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>
- The ePIC MAPS vertex detector design can reach as low as 0.05, and the imaging layers of MAPS sensors provide excellent position resolution. <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>
- The MAPS sensor used in the EIC is based on low mass, low power consumption MAPS in 65 nm technology. <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>
- The MAPS sensor used in the EIC is based on low mass, low power consumption MAPS in 65 nm technology. <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>, <https://arxiv.org/abs/2207.01447v1>

Methods to evaluate RAG's performance

- ### 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 —

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

Methods to improve RAG architecture

- Better chunking strategies. LaTeX Splitter, TWikiSplitter
- Metadata based filtering.
- Response Template fine-tuning. INSTRUCT tuning
- Model fine-tuning. Computationally costly.