AI4EIC

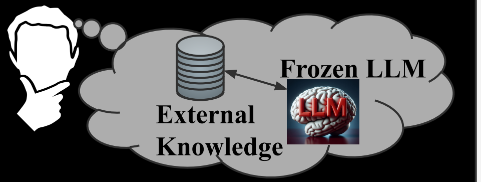
RAG4EIC

July 30th 2025

K. Suresh (William & Mary)

WILLIAM & MARY
CHARTERED 1693

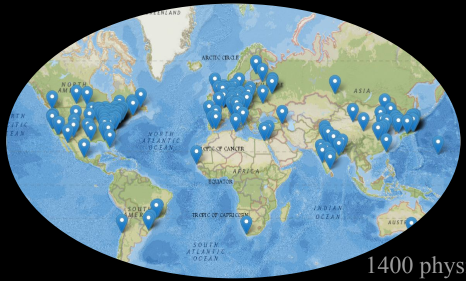
The Project outline



RAG based document retriever for EIC

- A newly reimagined document database where users can search through documents and wiki for the most up to date information.
- A conversational search type where, users' query get answered with LLM assisted responses.
- Real time tools for export such as report building and document creation

Building an AI powered search engine for EIC and its internals.



1400 physicists, 240 institutions and 38 countries….

Five “packages”

1. Advanced RAG System Development

- Expand the database to indico and public wiki links.
- Implement a more advanced RAG system like, Multi-Query Retriever approach.

2. Chain Tracing framework

- Tracing and displaying the “chain”
- Evaluate and compare more open source options with options to deploy on own servers unlike langsmith.
- Open Telemetry, Langfuse are some alternatives

3. Building Retrieval Tools

- Develop specialized tools for WikiReader integration for efficient document retrieval
- Extend PDF reader capabilities with improved metadata extraction and natural language understanding.

4. OpenLLM integration and Scaling

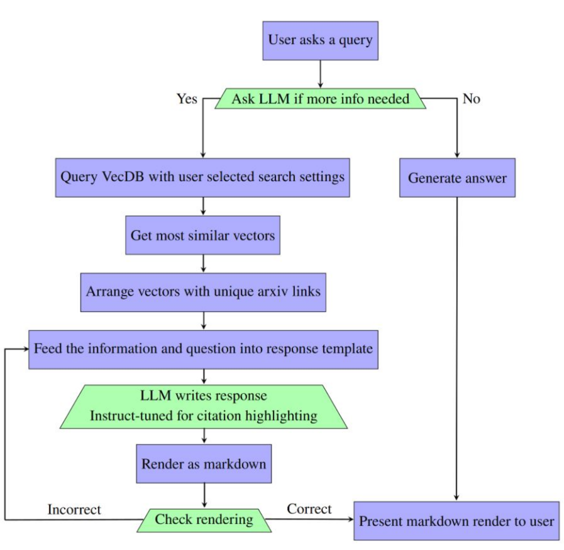
- Integrate OpenLLM models like Llama document retrieval.
- Re-evaluate infrastructure and resources

5. Evaluation Metrics and Model Finetuning

- Build further on RAG based evaluation automated and scalable RAG systems.
- Test out fine tuned models for context retrieval.

What has been done so far

- Advanced routing implementation
  - Check if the query can be answered by LLM itself if not then route to RAG
  - After information extraction, rewrite response using response block
- OpenLLM support (Laama3.2)
  - Now, RAG4EIC supports openLLM models. Can be used to be deployed within an infrastructure.
  - Any models within Ollama is currently supported.
- ChromaDB local support
  - The VectorDB used is Chroma as a local file system. Can be used to deploy within a given infrastructure.



User asks a query

Ask LLM if more info needed

Yes

No

Query VecDB with user selected search settings

Generate answer

Get most similar vectors

Arrange vectors with unique arxiv links

Feed the information and question into response template

LLM writes response
Instruct-tuned for citation highlighting

Render as markdown

Incorrect

Check rendering

Correct

Present markdown render to user

- Langsmith for tracing and feedback is implemented.

Pathway towards realizing the RAG4EIC Discussion

Current proposal – Serve EIC community

- Data Ingestion
  - Ingest a large corpus of EIC documents into the current RAG pipeline
  - Use existing cloud-based vector store (e.g., Pinecone)
  - Integrate updated GPT-based models to serve initial responses
- Beta User Rollout
  - Launch access to ~100 beta users from the EIC community
  - Users are expected to:
    - Test the app regularly and give feedback on UI rendering
    - Evaluate retrieval quality and grounding of responses
    - Rate model answers using RAGAS-style LLM-as-judge scoring
  - Update to app Every 4 weeks:
    - App updates with improved capabilities and bug fixes
    - New content ingestion or UX enhancements

- Goal:
  - Evaluate various RAG strategies
  - Develop various ingestion strategies
  - Build a high-quality, supervised dataset for fine-tuning
- Cloud resource support:
  - Provided by W&M for AI4EIC

*Upto 8 months

Serving the EIC community – Beta users

- Cloud-based deployment planned
  - To enable scalable access for beta testers and ensure smooth delivery of RAG4EIC.
- Corpus size and system load increasing
  - As we ingest more EIC related documentation, self hosting becomes less feasible* without dedicated infrastructure.
- Need for robust model performance
  - Open-source small language models, without fine-tuning, struggle to meet quality benchmarks like **RAGAS**
- Dedicated compute required
  - Hosting large models or experimentation with fine-tuned LLMs (e.g., LLaMA, Mistral, Phi) demands consistent cloud-backed resources.
- Supports iterative dataset creation
  - Beta user queries and usage logs will help build a real-world EIC-focused dataset to improve RAG accuracy and utility.
- Laying the groundwork for model fine-tuning
  - Early user interactions help us gather a high-quality dataset for future domain-specific training.

*currently with traditional RAG

Current proposal – Modularized implementation

- MCP Integration
  - Implement data sources as MCP servers (same corpus as cloud)
  - Mainly to alleviate privacy concerns
  - Develop batteries to run model inference + retrieval via MCP
  - Focus: Internal/private hosting of RAG workflows (e.g., BNL, JLab)
  - Role based authentication for VectorDB

- LangGraph Agentic Pipeline
  - Replace LangChain with LangGraph for modular graph-based control.
  - Introduce:
    - Source-aware routing
    - Multi-hop retrieval
    - Agentic scoring or fallback logic
  - Improved orchestration and auditability of queries

- Unified Deployment
  - Combine all in one stack to be deployed in a site

*currently with traditional RAG

List of tasks / areas of involvement

- Towards serving the first version to beta users
  - Ingestion
    - Arxiv papers
    - Indico meetings page
  - Inference
    - Add conversational memory
    - Add support for multiple vector base calls
  - Web interface
    - Authentication using GitHub OAuth
    - Improve feedback mechanism

- Modular implementation
  - Agentic workflow
    - Replace LangChain with LangGraph
    - Supervisor Agent
    - Network agent implementation
  - Algorithm
    - Advanced RAG
    - Graph RAG
    - Evaluation of RAG pipeline
  - MCP server implementation
    - Wiki sources
    - Zenodo sources
  - Web application development
    - Implementation of OpenWebUI interface

Ofcourse, new issues are welcome

How to get involved

- Email to support@eic.ai
- Subject line: Involvement in RAG4EIC as developer (or beta user)
- If developer (highly recommend including)
  - Current institution and a brief experience with LLM development (Just to get to know)
  - A brief description on which of the areas you would like to work on
- If beta user (highly recommend including)
  - Current institution and a brief description on how you are currently using Language model in your research
  - Your area of expertise (Eg. Theory, experimental, hadron spectroscopy)
  - Would you be interested in curating a golden data set in your area of expertise
- Regular RAG4EIC working group meetings
  - Once a month, Next meeting By August 26 2025 (Tuesday) anticipated. A reminder will follow.