Project Statement: Towards an AI agent for Large-Scale Scientific Collaborations

Researchers & Affiliations:

- Abhishikth Mallampalli, under the guidance of Prof. Sridhara Dasu, UW-Madison
- Prof. Balaraman Ravindran, IIT Madras

1. High-Level Science Goals and Need

Experimental physics collaborations today are larger and more productive than ever, leading to an incredible and accelerating pace of scientific discovery. This success, however, brings a new and pressing challenge: managing an ever-growing mountain of complex internal knowledge. For a new PhD student joining a project or a senior researcher trying to synthesize information across multiple analyses, finding precise, context-aware answers within the landscape of internal documents, wikis, and presentations is a major bottleneck. This knowledge retrieval problem slows down the pace of research, complicates the onboarding of new members, and hinders the cross-pollination of ideas between different analysis groups.

To address this challenge, we propose the development of an intelligent AI assistant—a research agent—designed to serve as an expert companion for members of large scientific collaborations. The goal is to create a tool that can understand natural language questions, retrieve the most relevant information from a diverse set of private data sources, and synthesize accurate, trustworthy answers. Such a system would not only dramatically accelerate research but also help democratize access to the collaboration's collective knowledge, making every member more efficient and effective.

2. Current Progress: The RAG Assistant Prototype

Under the guidance of Prof. Dasu at UW Madison, I have developed a working prototype of the assistant using Retrieval-Augmented Generation (RAG). The details of this prototype have been submitted to a major international ML for Physical Sciences workshop and we have also presented it at the Lepton-Photon 2025 conference [1] (with a demo link in slide 12 of [1]). This prototype demonstrates the feasibility and value of our approach. Its key features include:

- **An Automated Ingestion Pipeline:** A system that uses Selenium and advanced Optical Character Recognition (OCR) to automatically retrieve and accurately extract text and its structure from PDF documents stored on internal collaboration websites.
- **A Two-Tiered Database Architecture:** A novel design that first identifies the correct analysis document from a database of abstracts before locking on to the full-text knowledge base for that document. This ensures all answers are grounded in the correct context, a critical requirement for physics research because the same question can have a different, usually conflicting, answer depending on the analysis of interest. It also ensures easy updates to the database.
- **A Fully On-Premise, Privacy-Preserving Framework:** The entire system, from the embedding models to the LLM (a quantized Mistral-7B model), runs on local institutional hardware (GPUs at UW-Madison). This guarantees that no sensitive, proprietary, or unpublished data ever leaves the collaboration's secure network.

- **Validated Performance:** The retrieval component was rigorously benchmarked against a BM25 lexical search baseline. On conceptual queries, the RAG agent was >25x more likely to retrieve the correct document as the top result (Recall@1 of 0.78 vs. 0.03).

While successful, the current prototype is focused on text extraction from a single document type. This 3-month research exchange is designed to build upon this strong foundation, expanding the system's capabilities to create a more comprehensive and powerful research tool.

3. Project Goals and Deliverables (3-Month Timeline)

This collaborative exchange will allow us to learn from the expertise of the IITM team to significantly advance the RAG assistant. The work is divided into two main thrusts: enhancing the system's data extraction capabilities and expanding the breadth of its knowledge base.

3.1. Enhanced Data Extraction: Tables, Equations, and Plots

The first major goal is to enable the assistant to understand and reason about the rich, structured information present in scientific documents.

- Deliverable 1: Table Parsing and Interpretation: We will add the capability to parse and understand information contained within tables. The goal is not just to extract the text, but to understand the structure (rows, columns, headers) so a user can ask questions like, "What is the correlation structure for jet energy scale systematic uncertainty for the W+Jets process?"

- Deliverable 2: Improved OCR for Specialized Content: We will fine-tune or adapt the OCR pipeline to better handle the complex mathematical equations and special symbols common in physics papers.

3.2. Knowledge Base Expansion and Scaling

The second goal is to broaden the assistant's knowledge beyond individual analysis notes and papers to create a more holistic information source.

- Deliverable 3: Integration of Diverse Data Sources: We will expand the ingestion pipeline to process new, unstructured data types, including internal collaboration wikis (e.g., TWiki) and the slide decks from public talks (e.g., from Indico). This will allow the agent to answer a much wider range of questions. We recognize that presentation slides possess highly variable formatting, which presents a significant challenge for automated extraction. However, they also represent a uniquely valuable and timely source of information, making this an interesting development goal.

- Deliverable 4: Scaling the Analysis Set: We will scale up the ingestion process to include a much larger and more diverse set of analysis documents, covering a wider range of physics analyses.

3.3. Initial Assessment of Advanced Visual Reasoning

- Deliverable 5: Feasibility Study for Plot Interpretation: In order to consider answering questions related to plots like “Why is the data-MC bad at a particular mass value?” we will develop a prototype for extracting data points or interpreting the content of stacked 1D histograms.

4. Proposed Timeline

Month 1: Foundational Enhancements

- Develop and integrate a robust table extraction module into the pipeline.
- Begin work on improving the OCR pipeline for equations and special symbols.

Month 2: Knowledge Base Expansion

- Develop and deploy ingestion scripts for collaboration wikis and public talk slides.
- Begin the process of scaling up the ingestion of analysis notes to a larger set.
- Investigate into adding the ability to parse plots.

Month 3: Integration, Evaluation, and Future Planning

- Complete the integration of all new data sources.
- Perform a comprehensive end-to-end evaluation of the enhanced system.
- If applicable, develop and test the initial plot interpretation prototype.
- Analyze results and formulate a detailed future plan

5. Financial Estimates

The estimated budget for the 3-month research visit is requested as a monthly stipend of $1000-$1,100 USD, totaling $3000-$3,300 USD. This figure is a good-faith estimate derived from consultations with the host institution (IIT Madras) regarding the local cost of living for a visiting researcher in Chennai. The stipend is designed to reasonably cover all essential monthly expenses, including furnished accommodation, meals and groceries, local transportation, incidentals and professional expenses. This amount ensures a productive and focused research stay, allowing the researcher to fully immerse in the project without financial constraints. Support for round-trip economy airfare between the researcher's home institution and the host institution is requested. The visiting researcher holds a passport that allows for entry for the proposed duration of the visit. Therefore, no visa sponsorship or support will be required.

6. Summary of Local Support

The researcher will work closely with the IITM team, be provided with lab space and receive guidance on the project's development.

7. References:

[1] https://indico.cern.ch/event/1493037/contributions/6562773/attachments/3125058/5542436/AbhishikthMallampalli_LP2025.pdf