Bites of FM4S: [2] LLMs for experiments in fundamental physics

bites of...
Foundation Models
for Science

Themes of discussion:

- Physics-inspired representations
- **LLMs for experiments**
- Applications to Run 3
- Training methods
- Alignment, interpretability
- Datasets and benchmarks

...and more!

Report of Contributions

https://indico.cern.ch/e/1543967

Contribution ID: **1** Type: **not specified**

Multi-Agent Research Validator & Enabler Using LLMs (MARVEL): Experiences from LIGO

Tuesday 3 June 2025 16:30 (20 minutes)

Gravitational wave research at the Advanced LIGO observatories integrates complex, interconnected elements of experimental physics, computational simulations, and theoretical astrophysics. However, decades of valuable knowledge remain scattered across unstructured, multi-modal data and fragmented codebases. Efficient dissemination of this knowledge using large language models (LLMs) can significantly accelerate scientific discovery. In this talk, we share experiences from developing MARVEL, a modular, multi-agent research framework designed to provide scientific assistance in highly technical domains. MARVEL leverages open LLMs to ensure data privacy and is designed to be flexible enough to accommodate a broader range of scientific domains. We highlight challenges, including limitations in fine-tuning, pitfalls of naive Retrieval-Augmented Generation (RAG), model hallucinations, context window constraints, and difficulties in processing scientific documents via optical character recognition. To enhance factual accuracy and reasoning capabilities, MARVEL integrates tool usage and leverages test-time computing at the expense of increased latency. Finally, we emphasize the importance of modular workflows and custom benchmarks to adapt to advances in foundational models rapidly.

Theme of discussion

Presenter: MUKUND, Nikhil (MIT)

Session Classification: Invited talks

Contribution ID: **2** Type: **not specified**

The SpeakYSE: An Agentic LLM for Supernova Science

Tuesday 3 June 2025 16:55 (20 minutes)

Time-domain astronomy is rapidly entering a data-rich era in which wide-field surveys discover millions of transients per year, overwhelming traditional, hand-driven analysis. In this talk we present SpeakYSE, an agentic and open-source language model that turns natural-language requests into end-to-end analyses for the Young Supernova Experiment. The SpeakYSE links literature retrieval, database reasoning, and low-level tool-calling for on-the-fly exploratory data analysis. We describe the SpeakYSE's architecture, early results from its use within the collaboration, and future directions enabled by next-generation reasoning models. Domain-specific LLM agents like SpeakYSE will be essential for exploiting next-generation surveys such as the Rubin LSST.

Theme of discussion

Presenter: GAGLIANO, Alex (IAIFI/MIT/Harvard)

Session Classification: Invited talks

Contribution ID: **4** Type: **not specified**

TBD [LLMs for LHCb]

Session Classification: Invited talks

Contribution ID: **5** Type: **not specified**

chATLAS: An AI Assistant for the ATLAS Collaboration

Tuesday 3 June 2025 15:25 (20 minutes)

The ATLAS Collaboration is composed of around 6,000 scientists, engineers, developers, students and administrators, with decades of institutional documentation spread across wikis, code docs, meeting agendas, recommendations, publications, tutorials, and project management systems. With the advent of retrieval augmented generation (RAG) and sophisticated large language models (LLMs) such as GPT-4, there is now an opportunity to produce a "front door"to this intimidatingly large corpus. ChATLAS is an attempt to provide this entrypoint, as ATLAS'official AI assistant and search system. In this contribution, we review the past year of developments, present the latest updates to the system, and introduce ongoing work to improve back-end performance, agentic information gathering, and science-centric design components.

Theme of discussion

Presenter: MURNANE, Daniel Thomas (Niels Bohr Institute, University of Copenhagen)

Session Classification: Invited talks

Contribution ID: **7** Type: **not specified**

LLM-based physics analysis assistant at BESIII

Tuesday 3 June 2025 15:50 (20 minutes)

The data processing and analyzing is one of the main challenges at HEP experiments. To accelerate the physics analysis and drive new physics discovery, the rapidly developing Large Language Model (LLM) is the most promising approach, it have demonstrated astonishing capabilities in recognition and generation of text while most parts of physics analysis can be benefitted. In this talk we will discuss the construction of a dedicated intelligent agent, an AI assistant names Dr.Sai at BESIII based on LLM, the potential usage to boost hadron spectroscopy study, and the future plan towards a AI scientist.

Theme of discussion

LLMs for operations

Authors: LIU, Beijiang; YUAN, Changzheng; LI, Ke (Chinese Academy of Sciences (CN)); ZHANG, Zhengde (中国科学院高能物理研究所)

Presenters: LIU, Beijiang; YUAN, Changzheng; LI, Ke (Chinese Academy of Sciences (CN)); ZHANG, Zhengde (中国科学院高能物理研究所)

Session Classification: Invited talks

Contribution ID: **8** Type: **not specified**

AccGPT: A Chatbot for CERN Internal Knowledge

Tuesday 3 June 2025 15:00 (20 minutes)

AccGPT is an innovative pilot project utilizing Large Language Models (LLMs) to create a chatbot for interacting with CERN's extensive internal knowledge base. This initiative is primarily led by the CERN Beams and IT departments, while the objective is to make this chatbot available to the entire CERN community. AccGPT is designed to provide quick and straightforward answers to queries, similar to ChatGPT, thereby enhancing productivity and decreasing the time experts spend on support tasks. Looking ahead, there are plans to expand AccGPT's functionalities, for example utilizing enhanced Agent features.

Theme of discussion

LLMs for operations

Author: Dr REHM, Florian (CERN)

Presenter: Dr REHM, Florian (CERN)

Session Classification: Invited talks

Contribution ID: **9** Type: **not specified**

Bridging LLMs and Scientific Infrastructure: A2rchi for Context-Aware Research Support

Tuesday 3 June 2025 17:20 (20 minutes)

We introduce A2rchi (AI-Augmented Research Chat Intelligence), an intelligent, domain-adaptable open source chatbot designed to support research and education workflows.
Beyond its core functionalities, A2rchi also integrates with common communication and workflow tools, including email systems, ticketing platforms, and collaboration platforms like Mattermost, offering seamless assistance across multiple channels.

Leveraging Retrieval-Augmented Generation (RAG), A2rchi combines foundational large language models with custom, project-specific data—such as course materials and documentation—to deliver accurate, context-aware responses.

Originally developed and deployed to support MIT classroom instructions and the Physics Department's analysis facility, A2rchi is now being expanded to serve the CMS experiment at CERN, with applications ranging from Tier-0 operations and data management to end-user physics analysis support.

We present the current implementation, lessons learned from real-world deployments, and the roadmap for scaling A2rchi into a robust, domain-aware assistant for large-scale scientific collaborations.

Theme of discussion

LLMs for operations

Authors: D'ALFONSO, Mariarosaria (Massachusetts Inst. of Technology (US)); LUGATO, Pietro (Massachusetts Inst. of Technology (US))

Presenter: D'ALFONSO, Mariarosaria (Massachusetts Inst. of Technology (US))

Session Classification: Invited talks