Boost physics study at High Energy Physics experiments with Dr.Sai

Yipu Liao*

(On behalf of Dr.Sai working group)

Institute of High Energy Physics, CAS, Beijing

University of Chinese Academy of Sciences, Beijing
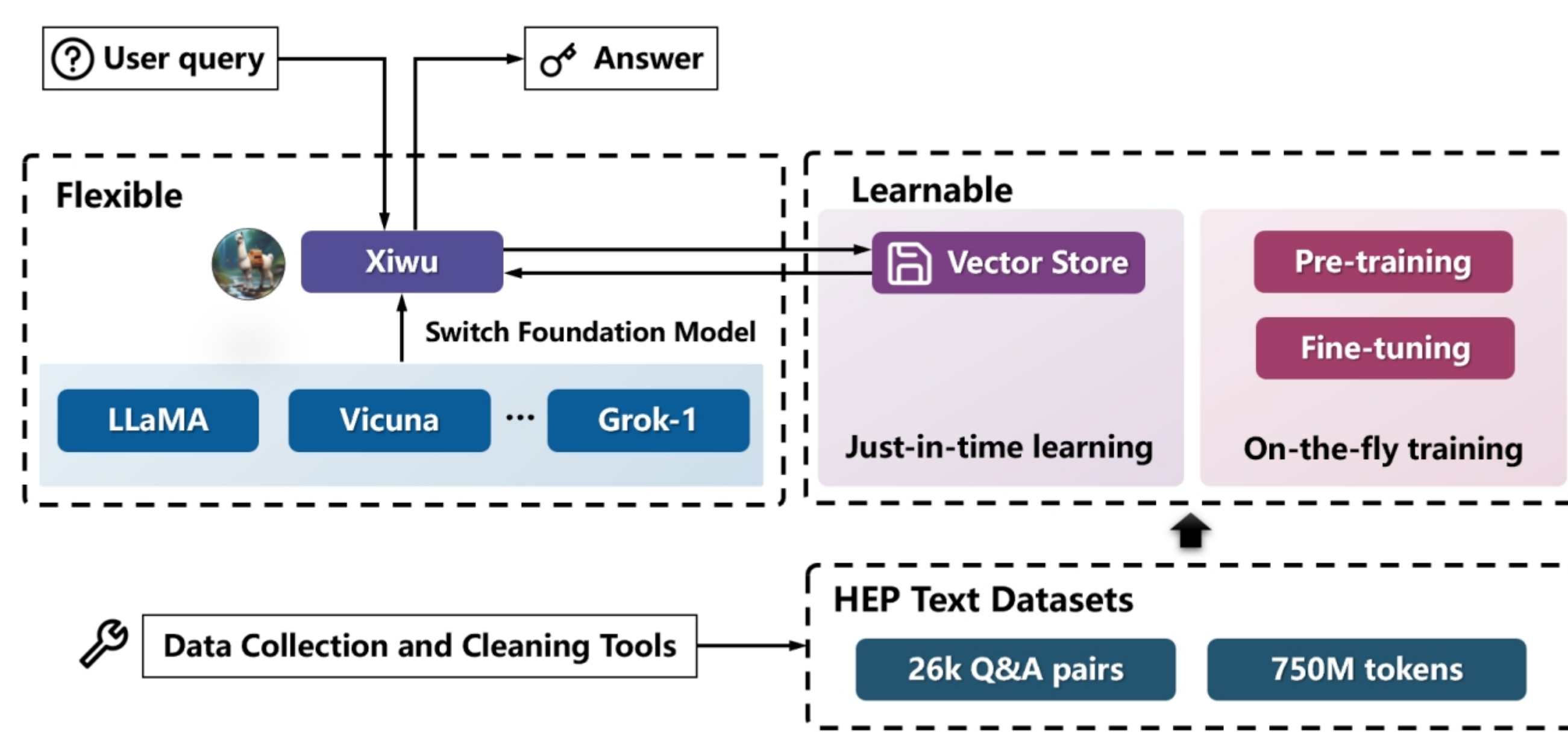
liaoyp@ihep.ac.cn

1. Introduction

The data processing and analyzing is one of the main challenges at High Energy Physics (HEP) experiments, normally one physics result can take more than 3 years to be conducted. 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 benefited. A construction with dedicated intelligent agents, an AI assistant for HEP based on LLM, is proposed to boost data analysis study.

2. Xiwu model and Datasets

Xiwu model is a sophisticated LLM system that facilitates seamless switching between advanced foundation models while rapidly imparting domain knowledge to the model [1].

**Datasets** are sourced from 8 HEP related domains:

- 8K high-quality Q&A pairs created by proposed Seed Fission technology.
- 10K Q&A pairs derived from Chatbot history within LLM service in HepAI platform [2].
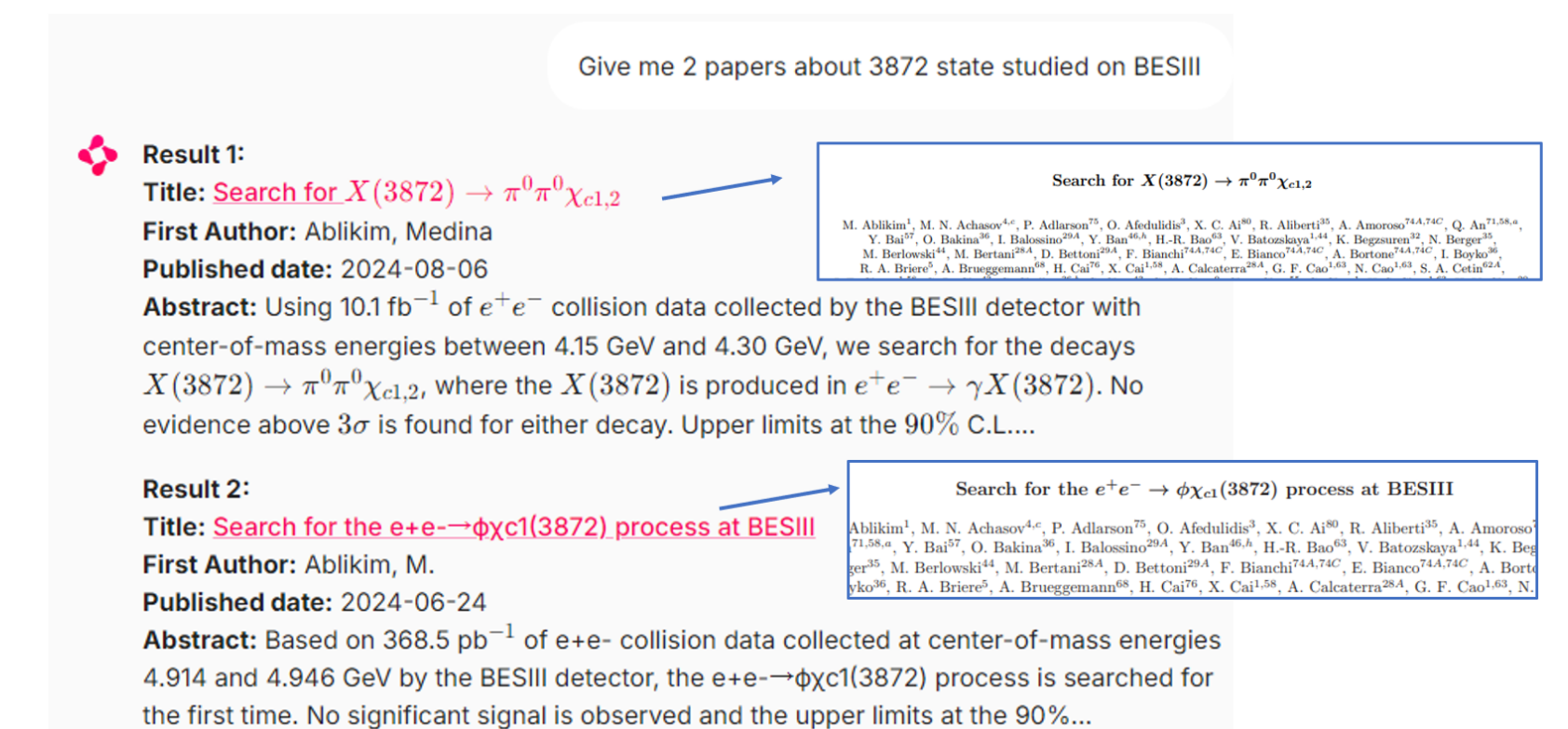- 750M tokens and 8K Q&A pairs cleaned from the abstracts and full texts of research papers [3].

3. Dr.Sai multi-agent collaborative system

An AI agent is a system or software capable of making autonomous decisions or performing actions on behalf of its users. The **Dr.Sai multi-agent collaborative system**[4] is developed based on the AutoGen framework [5] and comprises a variety of functional agents. The trained Xiwu model serves as a crucial component of the agents in the Dr.Sai. A typical workflow involves several steps:

1. Host manages tasks, broadcast information and select other agents.
2. Planner decomposes a user-requested task simpler subtasks.
3. Coder generates codes, then Tester executes codes.
4. A general LLM agent summarizes the results. Human joins GroupChat through HumanProxy.

The **Retrieval-augmented generation (RAG)** is also used to enhance the HEP-related accuracy and reliability of LLM. The Dr.Sai developed a vector store based on LlamaIndex [6], for storing exampled HEP codes and documents. The RAG could be helpful to user's searching request.

4. Application on HEP

Now, the Dr.Sai is capable of handling simple tasks and decomposing complex tasks into subtasks. For example:

1. **Code generation and execution**: Generated code can be automatically sent to remote worker for execution, based on the distributed deployed framework. It supports languages such as Python, CERN ROOT, and Shell, among others.
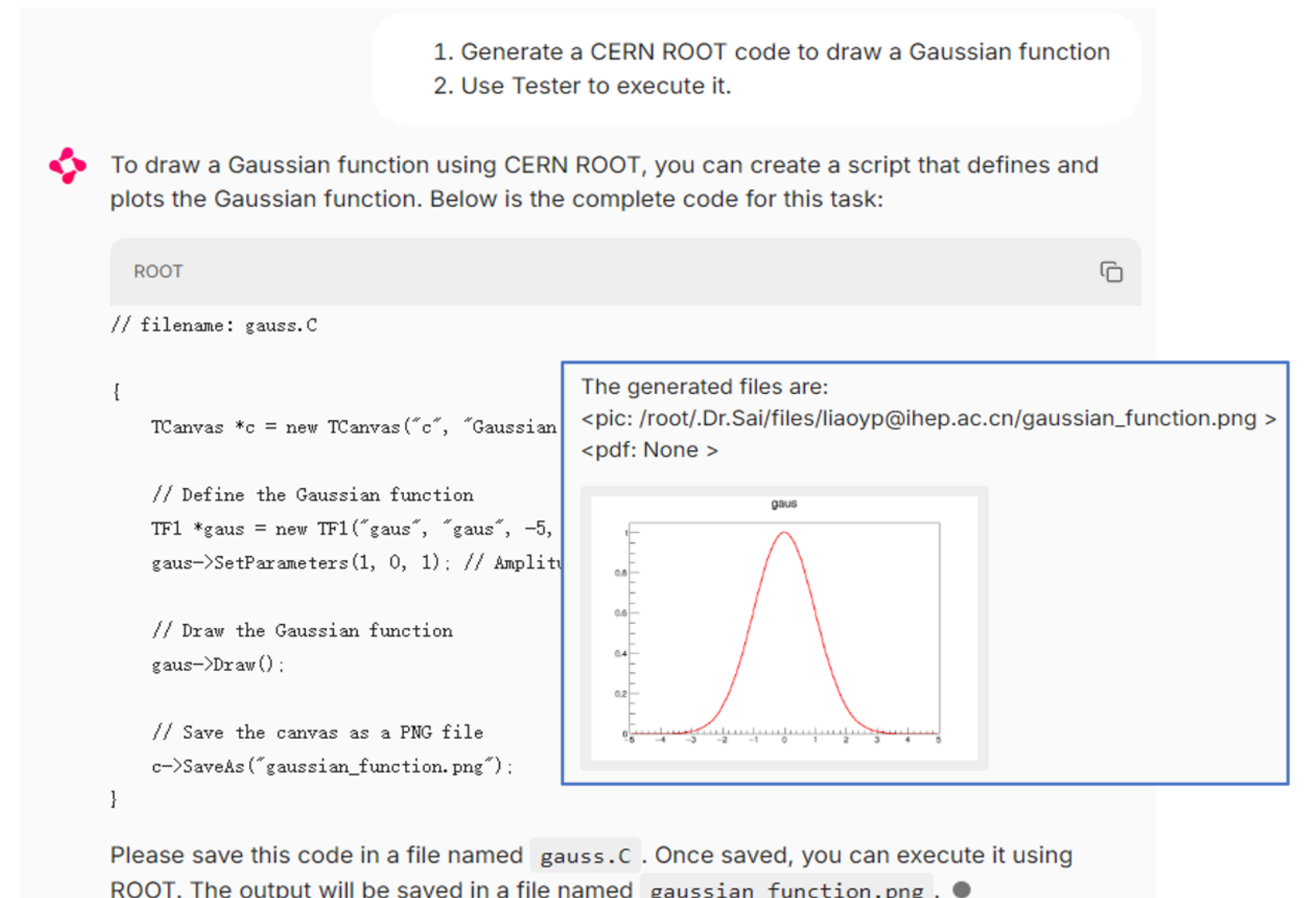
2. **HEP paper searching**: LLMs extract keywords and construct query URLs for searching. It currently supports ArXiv and INSPIRE.
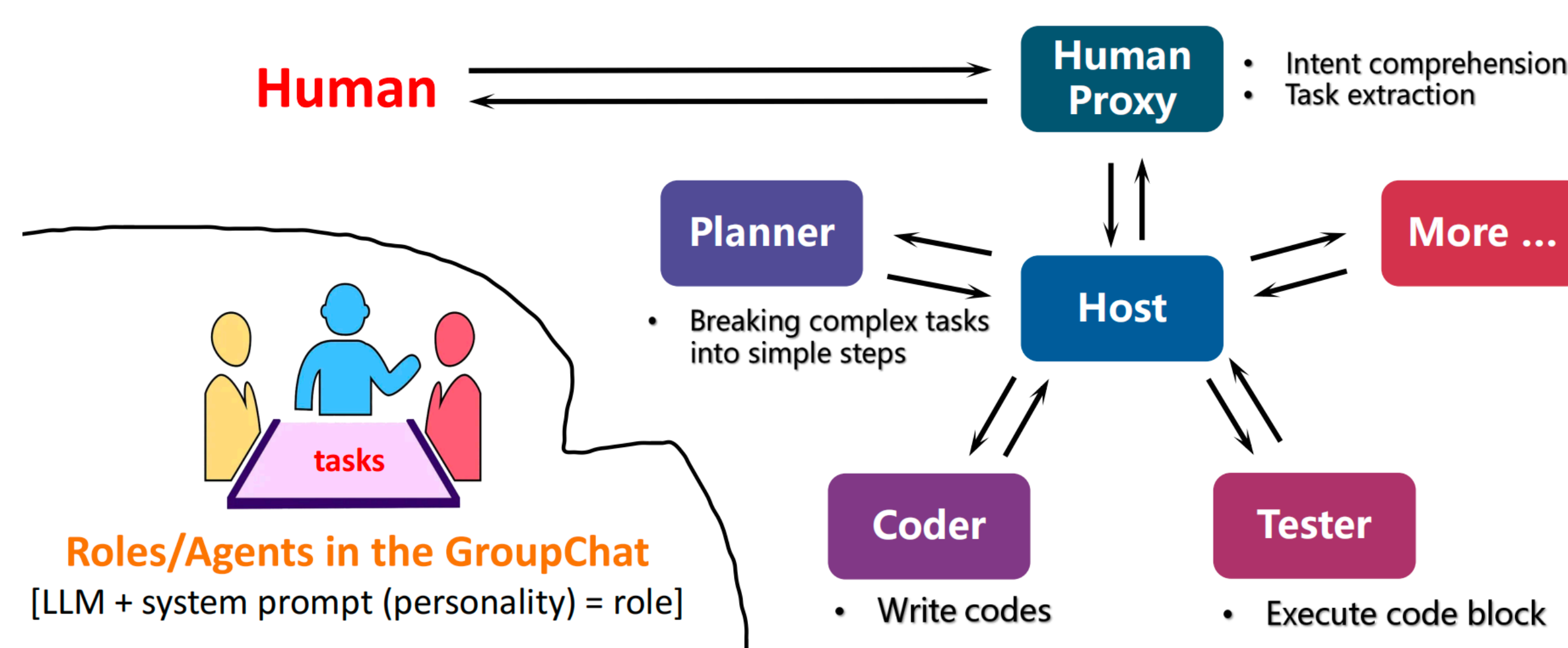

3. **Navigation**: Internal websites and documents searching by the RAG.
4. **Text polishing and Image Parsing**: A fine-tuned LLM has been implemented.

5. Conclusions and Outlook

To accelerate the analysis in HEP, the development of an AI assistant based on LLMs can be both exciting and beneficial. The following related works have been explored:

- A high-level LLM, the Xiwu model, is trained and fine-tuned for handling HEP tasks.
- A multi-agent collaborative system, Dr.Sai, is proposed and developed.
- The RAG method is employed to enhance HEP-related capabilities.

The development of a knowledge graph-based RAG, the next iteration of Xiwu with more parameters, and test sets specifically designed for HEP-LLMs are currently underway. Efforts are also being made to process collision data directly using a Large Science Model.

6. Acknowledge and References

[1] Zhengde Zhang et al. Xiwu: A Basis Flexible and Learnable LLM for High Energy Physics. ArXiv: 2404.08001.

[2] HepAI platform: ai.ihep.ac.cn

[3] Jianwen Luo et al. Hainougat: An Academic Document Parser that Preserves Formulas and Tables for High-Energy Physics. SSRN: dx.doi.org/10.2139/ssrn.4982380

[4] Dr.Sai: drsai.ihep.ac.cn

[5] AutoGen: github.com/microsoft/autogen

[6] LlamaIndex: www.llamaindex.ai