Integration of AI Large Language Models in High Energy Physics Research Workflow

Joseph C Wang (joequant@hkphysics.hk) (joequant wechat)

Hong Kong Institute of Physics

24 October 2026

2026 International Workshop on the High Energy Circular Electron Positron Collider

All source code is at

- https://github.com/hkphysics/ai-server
- https://github.com/hkphysics/papers (see cepc2026/examples)
- BSD Licensed

Usage of AI in Research

- Gap in literature. The literature has been focused on using AI models for fundamental research
  and there has been very little literature on how to use AI for research
- Gap between AI software providers and researchers. Software providers are unaware of the needs
  of researchers
- Gap between researchers and general public
- Gap between different domains (experimental / theoretical)

Background

- Compile Key4HEP.
  - Using AI tools for fixing bugs. Very good at diagnosis
  - Problem is integrating AI back into system (SPACK)
- Work as independent researcher
  - Need adminstrative and technical support
  - Reduce friction
  - Manage outside commercial projects - Tokenization of hedge funds

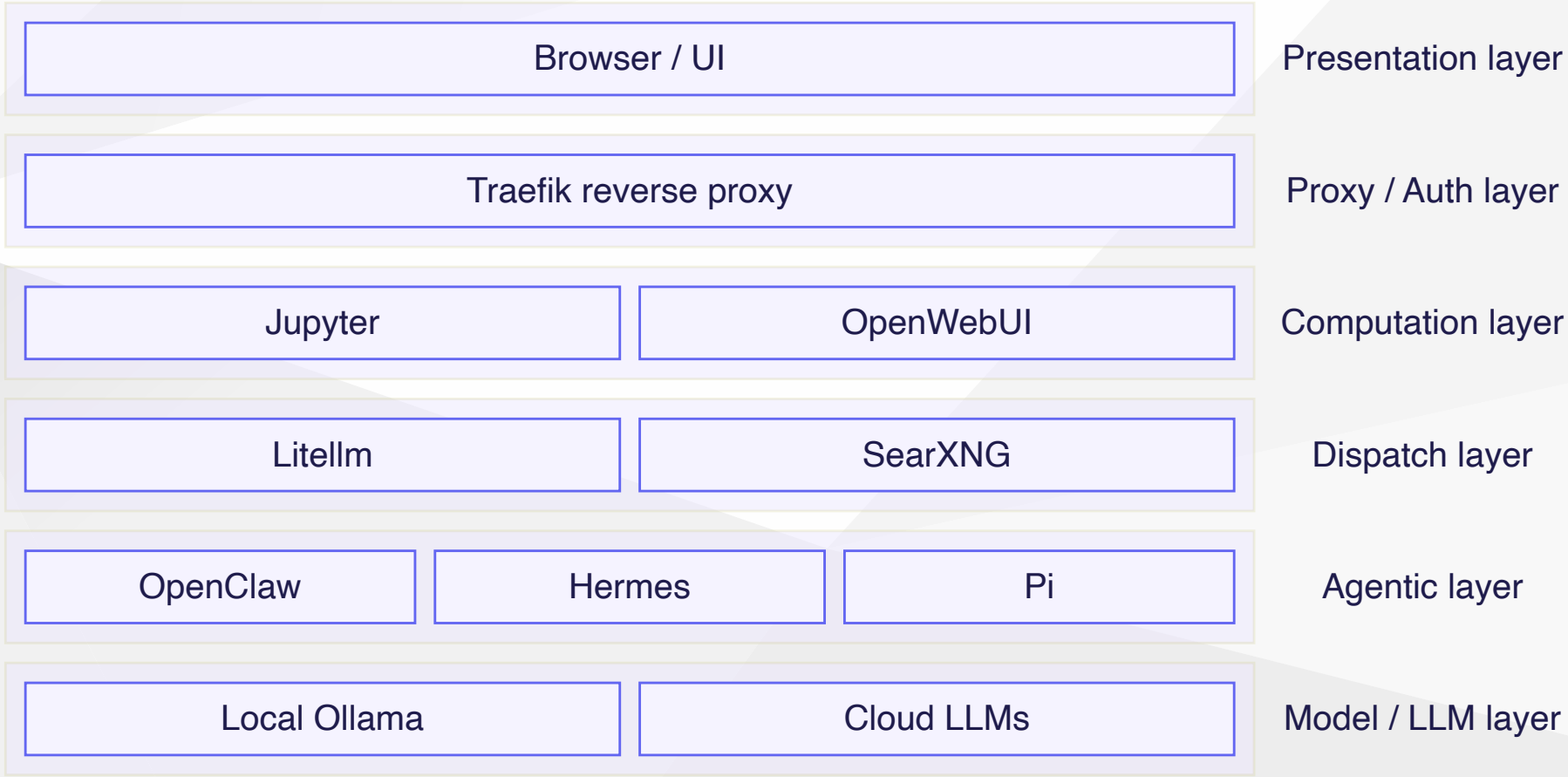
Requirements for HEP Researchers

- Technical users
  - HEP researcher can code but would prefer not to
  - Open source or internal projects
  - No black boxes
  - Very difficult to use SaaS model
- Data security and permissioning
- Obscure packages
  - ROOT, DiRAC, Spack

Platform created

- Github - hkphysics/ai-server
- Intended to be relatively easy to install - docker server
- Jupyter engine and Openwebui with shared internal workspace
- A lot of work with Jupyter AI team (Amazon)
- Three AI instances (Academic, Commercial, Personal)
- Openwebui / Jupyter

System architecture



So what it is good for

- Remote desktop
- Coding tool
  - Integration with Jupyter makes it excellent for semi-vibe coding
  - Data cleaning and utility scripts
  - Extremely good at analyzing large code obscure bases
- Multi-lingual environments

So what is it good for - Part 2

- Great at pedagodical questions and brainstorming
  - Creating problem sets
  - Creating examples
- Hallucinations not always a bad thing (Creative ideas while on AI LSD)
- Barnum/fortune teller effect is useful
- Great at format conversions

Problems

- It halluciates and the answers are unreliable
- Perfect papers that don't exist
- Terrible access control and information boundaries
  - Memory in agents
- It will trash your work
  - I don't trust it with my credit card and calendar
- A lot of plumbing work needs to be done
- Not well integrated into research databases or physics systems

Findings - Open the pod bay doors Hal

- 1960s Star Trek show
- The models are good enough. The problem is the agents
- Thinking tags are important. Voice is terrible. Shared data systems
- Local versus cloud
  - Local models are very good at one shot question
  - Coding requires a lot of context which eats tokens
- AI is extraordinarily useful for multi-language environments
- AI will trash your work data security is a mess
- De-GUIization is good for HEP - The world is discovering LaTeX

Immediate tasks

- More work in integrating AI systems physics tools
  - Integration with math engines
  - Would like to integrate with literature databases
  - Integration with gray repositories (Annas Archive)
- Teach AI about physics tools. ROOT and Spack
- Use as frontend with physics software
  - Generate configuration files and system runs
  - Connect different HPC systems

Future directions

- Reverse engineer physics software for integration into non-physics (ROOT and Spack)
- Controversy has already hit mathematicians and lawyers. Starting to drag in HEP.
- Use of systems for administration
- Questions about education
  - How do we use the technology to better teach physics
  - How do we teach coding?
- How this works with funding and social systems