Agentic AI at the Advanced Light Source

Thorsten Hellert
LBNL
Berkeley, USA
thellert@lbl.gov

Joao Montenegro
LBNL
Berkeley, USA
joaomontenegro@lbl.gov

Antonin Sulc
LBNL
Berkeley, USA
asulc@lbl.gov

Abstract

Language-model-driven agents are transforming operations in safety-critical environments such as scientific facilities, but require architectures that ensure reliability, scalability, and human oversight. We present an application of the Osprey Framework, a production-ready agentic system deployed in the control room of the Advanced Light Source (ALS) particle accelerator. To meet the demands of this high-stakes domain, this work shows a safe, plan-first orchestration model with modular human approval, dynamic tool selection for managing complex capabilities, and a resilient execution environment. In a live deployment, it autonomously executed a multi-stage physics experiment, from historical data analysis to real-time hardware control, based solely on a single high-level prompt from an expert operator. This successful deployment demonstrates that through these architectural principles, agentic systems can move beyond demonstrations to become robust tools for high-stakes domains, providing a blueprint for applications such as automated beamline experiments.

1 Introduction

Particle accelerator environments integrate legacy and modern control systems that generate massive amounts of diverse data. These data sources are distributed. To maximize availability, operators must act on this information under tight time constraints, but retrieving, interpreting, and coordinating across these systems remains cumbersome. Critical expertise is often preserved only as human experience, scattered across institutional memory rather than codified in software. This tension hinders the effective responses to evolving operational demands.

Agentic AI [1], powered by Language Models (LMs), offers a compelling opportunity to address these challenges by translating high-level human intent into complex, automated workflows. However, in highly specialized domains such as particle accelerators, much of the relevant expertise is absent from training corpora and is confined to experts. Leveraging agentic AI in these environments requires architectures capable of operationalizing this internal knowledge while maintaining human-in-the-loop mechanisms to ensure safety and trust.

In this paper, we present the successful deployment of the Osprey Framework [2, 3, 4], a domain-agnostic, production-ready architecture for agentic systems, within the control room of the Advanced Light Source (ALS) user facility. The framework is explicitly designed to address the challenges of high-stakes environments through verifiable safety mechanisms and scalable tool orchestration, as detailed in [3].

We detail its application as a powerful new tool for scientific operations. We demonstrate how a single natural language request from an operator was autonomously translated into a complex, multi-stage physics experiment involving data retrieval, real-time hardware control, and analysis. This serves as a powerful case study for the feasibility and benefit of agentic AI in complex scientific facilities.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

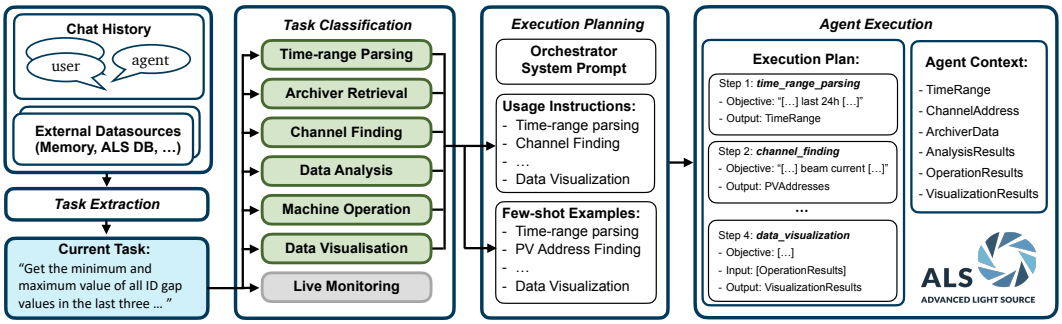

Figure 1: The architecture of the agentic system at ALS, based on the Osprey Framework. The workflow proceeds from Task Extraction to Dynamic Capability Classification (green boxes), where relevant tools are selected. A Plan-First Orchestrator then generates an inspectable plan, which is executed to produce artifacts and update the agent's context.

This work shows that with the right architectural principles, namely plan-first orchestration with human oversight, dynamic and scalable tool integration, and production-ready adaptability an agentic systems can move beyond proof-of-concept demonstrations to become robust tools for day-to-day scientific operations. The successful showcase at the ALS provides an example for applying similar systems to other large-scale facilities.

2 Background and Related Work

2.1 Agentic Systems and Tool Use

The foundations of agentic LMs build on the classic definition of agents as entities that perceive their environment and act upon it [5]. Early frameworks such as ReAct unified reasoning and acting in iterative loops for tool-based tasks [6], while AutoGen extended this paradigm toward multi-agent collaboration [7]. Scaling tool use in LMs has since become a central concern. While prior work has advanced the scaling of tool use and memory, these systems often still face challenges of prompt growth and domain grounding at inference time, which are critical hurdles for deployment in specialized scientific environments.

2.2 Applications of Agentic AI in Science

Agentic AI has seen rapid adoption in scientific applications, from autonomous laboratories for materials discovery [8, 9] to chemistry-specific systems like ChemCrow [10] and Co-scientist [11]. Among large-scale scientific facilities, agentic approaches have also been explored for particle accelerators [12]. GAIA demonstrated an early accelerator operations assistant [13], while VISION developed a modular AI assistant for beamline experiments [14]. These systems illustrate the potential of agentic AI, but they often rely on custom implementations that are difficult to scale or generalize. Our work builds on these efforts by demonstrating a successful deployment of a system built on a standardized, scalable, and production-ready architecture.

3 Methods: The Osprey Agentic Framework

The architecture that enabled the ALS deployment is the Osprey Framework [2, 3, 4]. This section provides a concise summary of its core design: a production-ready framework for scalable workflows in complex domains. As illustrated in Figure 1, the framework is built to safely and reliably translate high-level conversational requests into executable, multi-step plans, forming the foundation of the ALS case study presented here. For deployment, we used a containerized environment using Podman for reproducibility and isolation. Model inference is managed through a hybrid architecture with a local Ollama instance on an H100 GPU that provides low-latency responses within the control room network, while a gateway service connects to more powerful external models like GPT-4, Claude 3, and Gemini for complex reasoning tasks. The interface is done through BerkeleyLab's CBorg API.

The workflow begins with **Task Extraction**, where the system transforms multi-turn conversations into a structured, well-defined task. It analyzes the dialogue history and integrates it with external data from user memory, domain-specific knowledge bases (*e.g.*, accelerator physics documentation), and APIs to create a formalized task with clear objectives. This step is crucial for grounding the user's intent in the specific operational context of the facility.

Next, to manage the multitude of potential tools the system employs **Dynamic Capability Classification**. Instead of overwhelming the LM with every possible tool, it performs a per-capability relevance analysis for the specific task at hand. Each tool is independently evaluated in a binary classification step (relevant or irrelevant, see green boxes in Figure 1), ensuring that only the necessary documentation is passed to the planner. This targeted filtering decouples prompt complexity from the total number of tools, which is essential for scalability.

A core principle for ensuring safety is the **Plan-First Orchestration** model. Unlike reactive agents that reason and act in short cycles, this system generates a complete, inspectable execution plan before any hardware is controlled or irreversible action is taken. This plan models the workflow as a graph with explicit input-output dependencies, allowing it to be serialized for inspection, modification, and resumption. This mode allows an operator to review and approve the plan before execution begins, providing a critical human-in-the-loop safety gate. This model is critical for safety and addresses potential LLM unpredictability. Before execution, particularly for any action with write-access to the control system, the plan is presented to an operator for approval. The system includes structured error handling with bounded retries for transient faults and can trigger a re-planning cycle if an unexpected state is encountered, ensuring resilience.

Finally, the execution environment is designed for production-ready reliability. Building on LangGraph's checkpointing for state continuity, it adds structured error classification and recovery strategies, such as bounded retries or re-planning. All operations generate reproducible artifacts, like versioned Jupyter notebooks of the executed code, structured JSON outputs, and logs, creating a complete audit trail for scientific reproducibility and operational review. This addresses the strict need for provenance in scientific facilities. Human-in-the-loop approval is a first-class capability, allowing operators to inspect plans or generated code before execution. For scientific usability, the system supports containerized execution, automatically packages generated code into Jupyter notebooks for inspection, and manages all intermediate and final artifacts.

4 Live Deployment at the Advanced Light Source

The ALS at Lawrence Berkeley National Laboratory is synchrotron facility that provides bright X-ray light to a broad scientific community [15]. We deployed the Osprey framework in the ALS control room, where it was integrated directly with the accelerator control systems to assist operators with complex tasks. This environment poses distinctive challenges that consist of real-time integration, coordination across heterogeneous subsystems, and strict adherence to safety protocols.

We tasked the system with the following real-world request from an operator:

> *"Get the minimum and maximum value of all ID gap values in the last three days. Then write a script which moves each ID from maximum to minimum gap and back while measuring the vertical beam size at beamline 3.1. Sample the gap range with 30 points, wait 5s after each new setpoint for the ID to settle and measure the beamsize 5 times at 5Hz. Return a hysteresis plot beam size vs gap."*

This query represents a complex, multi-step physics experiment that requires historical data analysis, hardware control, synchronized diagnostics, and scientific visualization. Manually, such a task would take an expert operator several hours of work to identify control variables, retrieve data, and hand-craft measurement scripts. By contrast, the agentic system automated the entire workflow from this single natural language prompt.

The system first identified the necessary capabilities for parameter discovery, historical data retrieval, data analysis, machine operation, and visualization. The workflow is shown in Figure 1. The orchestrator then produced a structured plan that was presented to the operator for approval. Upon approval, the agent executed the plan:

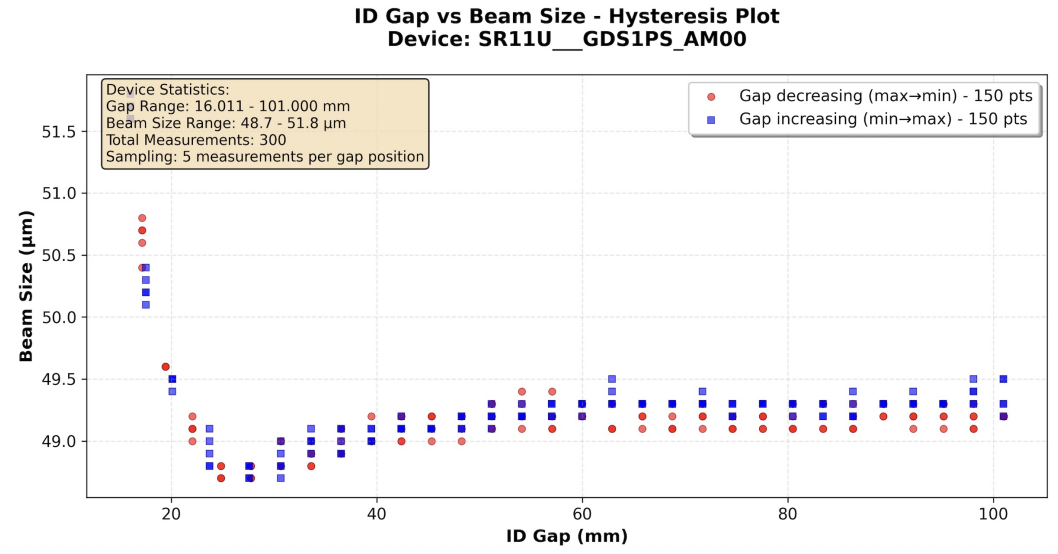

Figure 2: Hysteresis measurement of insertion device gap versus vertical beam size at the ALS. The execution plan generated by the agent combined historical range extraction, automated script generation, and real-time machine control. The agent performed a 30-point bidirectional gap sweep with 5 repeated measurements per point, producing the plot shown here for one device.

1. It navigated over 10,000 control system channels to find the correct variables for "*ID gap values*" and "*vertical beam size at beamline 3.1*."
2. It parsed the human request into a Python native time range
3. It queried the historical data archive and retrieved time series over the last three days.
4. It generated and executed a inspectable Python script to find the min/max gap values.
5. It dynamically generated a measurement script, which contained the logic for the bidirectional hardware scan.
6. Finally, it generated an annotated hysteresis plot from the collected measurement data.

Figure 2 shows the final result, a plot correctly showing no visible hysteresis, validating both the physics measurement and the agent's autonomous execution. This successful deployment, translating a high-level request into a multi-stage experiment with traceable artifacts, demonstrates a significant advance in applying AI to scientific facility operations.

5 Discussion and Future Applications

The successful deployment at the ALS serves as a powerful proof-of-concept with broad implications for other scientific and industrial facilities. The key takeaway is that with a robust, safety-conscious architecture, agentic AI can serve as a powerful co-pilot for human experts and automating tedious tasks, codifying institutional knowledge, and accelerating scientific discovery.

Beamline Experiments: At the ALS, the framework is being extended to automate beamline workflows. By coordinating hardware such as sample changers and detectors with analysis pipelines, it allows scientists to specify complex procedures in natural language and execute them automatically. While our initial focus is on ALS beamlines, the same approach generalizes to other facilities, from neutron sources to telescopes, where complex setups can likewise benefit from natural-language orchestration and automated execution.

Autonomous Laboratories: In the context of self-driving labs for materials science or biology, the framework provides the core capabilities needed to move toward closed-loop operation. While

our current deployments focus on executing single high-level tasks, the same orchestration, error-handling, and re-planning mechanisms can be extended to iterative experiment–analysis–design cycles. This positions the framework as a foundation for future autonomous labs, where agents not only execute experiments but also adapt and propose new ones based on results.

This work represents a shift from viewing AI as a passive analysis tool to an active participant in the operational loop of scientific research.

6 Conclusion

We have presented a case study of a successful deployment of the Osprey Framework in the control room of the Advanced Light Source particle accelerator. By autonomously translating a single, high-level natural language request into a complex, multi-step physics experiment, this work demonstrates that agentic systems can enhance the usability, reliability, and efficiency of scientific facility operations.

The success of this deployment was enabled by key architectural principles, detailed in [2, 3, 4], including a plan-first orchestration model with human-in-the-loop oversight and dynamic capability classification for scalable tool use. Furthermore, result validates the approach for real-world, high-stakes applications and provides a concrete blueprint for the application of agentic AI in other complex scientific and industrial environments. The framework is now being extended to other applications at the ALS, highlighting its readiness for production use and its versatility as a facility-wide tool for accelerating science.

7 Acknowledgment

This research leveraged the CBorg AI platform and resources provided by the IT Division at Lawrence Berkeley National Laboratory. We gratefully acknowledge Andrew Schmeder for his consistent responsiveness and support, which ensured that CBorg served as an invaluable resource for the development of this framework and NLP efforts in general at ALS.

We are grateful to Alex Hexemer, Hiroshi Nishimura, Fernando Sannibale, and Tom Scarvie (LBNL) for stimulating discussions and continued support, and to Frank Mayet (DESY) for sharing insights from his pioneering Gaia prototype, which guided the early trajectory of agentic AI at ALS.

We further acknowledge the use of AI tools during the preparation of this work. Cursor, primarily with Claude 4, was employed extensively during development.

This work was supported by the Director of the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

[1] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. AI agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468*, 2025. arXiv:2505.10468 [cs.AI].

[2] Osprey Developers. The Osprey Framework Repository, 2025. Accessed: 2025-11-04.

[3] Thorsten Hellert, João Montenegro, and Antonin Sulc. Osprey: A scalable framework for the orchestration of agentic systems, 2025.

[4] Thorsten Hellert, Drew Bertwistle, Simon C. Leemann, Antonin Sulc, and Marco Venturini. Agentic ai for multi-stage physics experiments at a large-scale user facility particle accelerator, 2025.

[5] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020.

[6] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023. arXiv:2210.03629 [cs.CL].

[7] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen llm applications via multi-agent conversation. In *CoLM 2024, LLM Agents Workshop at ICLR 2024*, 2023.

[8] Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Doğus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624:86–91, 2023.

[9] Aikaterini Vriza, Michael H Prince, Henry Chan, Tao Zhou, and Mathew J Cherukara. Operating robotic laboratories with large language models and teachable agents. In *ICLR Workshop on AI4MAT – ICLR 2025*, 2025.

[10] A. M. Bran, S. Cox, A. D. White, and P. Schwaller. Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 2024.

[11] D. A. Boiko, R. MacKnight, and G. Gomes. Autonomous chemical research with large language models. *Nature*, 624:486–492, 2023.

[12] Antonin Sulc, Thorsten Hellert, Raimund Kammering, Hayden Houscher, and Jason St. John. Towards Agentic AI on Particle Accelerators. In *Machine Learning and the Physical Sciences Workshop @ NeurIPS 2024*, 2024. arXiv:2409.06336 [physics.acc-ph].

[13] Frank Mayet. GAIA: A general ai assistant for intelligent accelerator operations. *arXiv preprint arXiv:2405.01359*, 2024. arxiv.org:2405.01359 [cs.CL].

[14] Shray Mathur, Noah van der Vleuten, Kevin G Yager, and Esther H R Tsai. VISION: a modular ai assistant for natural human-instrument interaction at scientific user facilities. *Machine Learning: Science and Technology*, 6(2):025051, 2025.

[15] Thorsten Hellert, B. Flugstad, C. Sun, C. Steier, E. Wallén, F. Sannibale, G. Portmann, H. Nishimura, J. Weber, M. Venturini, M. Dach, S. Leemann, S. Omolayo, S. Borra, T. Scarvie, and T. Ford. Status of the advanced light source. In *Proceedings of IPAC'24*, pages 1309–1312, Nashville, TN, USA, May 2024. JACoW Publishing, Geneva, Switzerland. Paper TUPG37.