Accelerating materials discovery with ADAM: An AI co-pilot for physical experimentation summary

Anderson Borch, *Member, ACM,* Steven R. Spurgeon

***Abstract*—The rapid discovery of new materials for energy storage, advanced computing, and other technologies requires us to remake scientific experimentation beyond the limits of human cognition and manual interaction. Most existing autonomous experimentation relies on deterministic scripts or narrowly-scoped post-hoc machine learning. Large Language Models offer a fundamentally different capability, in the form of generalized reasoning from natural language goals, but we must still address safe LLM reasoning with real-time instrument feedback in a closed-loop workflow. Here we present ADAM, a multi-agent large language model framework integrating a retrieval augmented generation pipeline grounded in internal documentation, a validated instrument tool registry, and human-in-the-loop oversight positioned at semantic reversibility boundaries. By inferring these boundaries from natural language rather than hard-coded rules, ADAM achieves high autonomy in routine experiments and applies precise human oversight only when needed. We demonstrate this system in two materials characterization and processings examples.**

***Index Terms*—Artificial Intelligence, electron microscopy, materials science, autonomous experimentation**

I. Summary

The use of Large Language Models (LLM) in physical experimentation is in its infancy, especially within materials science. Materials experiments are diverse, across various types of instrumentation, making it difficult to design a generalizable model. Many instruments require lots of operational expertise and are designed around human-in-the-loop (HITL) workflows, demanding mastery of intricate controls, safety protocols and multi-step procedures, with subjective, experience-based judgments required throughout. Consequently, quality data and efficiency depend on expert operators, leading to severe bottlenecks in throughput and reproducibility. There is an urgent need to abstract low-level instrument complexities while ensuring safe, reliable experimentation, allowing researchers to focus on high level goals as the demand for faster materials design grows.

Autonomous characterization and processing experiments present both an opportunity and challenge for autonomous reasoning models. These experiments generate vast amounts of multi-modal data, whose interpretation turns on subtle knowledge of materials features, acquisition parameters, and potential artifacts. In addition, the reasoning needed to analyze a material is different from the reasoning needed to process a material. Artificial intelligence (AI) and machine learning (ML) are well suited to these problems, allowing for reproducible, rich statistical analysis of data. Combining agentic LLMs with existing autonomous workflows is the logical next step. However, most autonomous materials science currently focuses on synthesis, leaving a critical bottleneck in characterization and processing. Self-driving laboratories have demonstrated impressive throughput for autonomous compound synthesis, but they presuppose a mature characterization pipeline to interpret their output—a bottleneck this work directly addresses. In electron microscopy, the platform we discuss here, ML methods have been widely adopted for post-hoc tasks like image classification and defect analysis rather than real time operational support and control. Previous autonomous experimentation routines rely heavily on deterministic automation scripts or rigid Bayesian optimization loops. These systems can accelerate calibration tasks but lack generalized reasoning and natural language safeguards on autonomy. Thus, a significant gap remains between a researcher's high-level experimental intent and the low-level physical operations required to execute it.

Frontier models display impressive out-of-the-box reasoning ability for science, interpreting experimental data, proposing mechanistic explanations, and designing follow-on experiments across scientific disciplines without fine-tuning. Still, deploying standard models directly to physical hardware poses significant risks. They tend to hallucinate group specific domain knowledge, controls, and parameters, making them a liability for mission-critical instrumentation. Agentic systems can "sleepwalk," taking arbitrary, potentially disastrous actions; as a result, the field has tended to deploy either rigid deterministic scripts or burdensome approval checkpoints that preclude true autonomy. For autonomous experimentation to advance beyond toy examples to general problems, we must reconcile these approaches. A key part of the solution is grounding in internal, group specific data, protocols, and objectives. Along these lines, research has demonstrated how in-context learning combined with meticulous guidance prompts can enable expert-level autonomous single-crystal X-ray diffraction. This is a clear, important step toward safe autonomy, but it reveals several gaps. First, there is a need for grounding that moves from hand-tuned prompting to a living knowledge base. Second, HITL oversight at every checkpoint undermines the very nature and value of autonomous experimentation. Third, an architecture that can generalize to different instruments is needed to progress from specialized to useful autonomy. Retrieval-Augmented Generation (RAG) directly addresses the first of these gaps: by bounding an LLM response to a retrieved context of verified documents, it significantly improves factual accuracy and safety in scientific domains. By applying oversight only where irreversibility demands it and a modular hardware abstraction layer, this approach moves beyond brittle prompting, blanket approval

checkpoints, and single-instrument specialization.

In this work, we present ADAM, a deployment-ready, multi-agent platform grounded in living scientific knowledge and targeted HITL oversight, validated on diverse materials experimentation. Model grounding is implemented through a custom RAG database built on domain- and group-specific data like publications, procedures, and discussions; this approach ensures a living research corpus, auditability, and group-specific aims, not just general knowledge. Human oversight is integrated with the same philosophy: ADAM infers the boundary between reversible and irreversible actions from natural language, not hard-coded rules, allowing for selective HITL intervention. Inference sits atop structural engineering controls, including separation between conversational and hardware-control agents and a validated tool registry. The registry ties agent actions to instrument-specific APIs, decoupling the reasoning layer from low-level control code and imparting native safety, since hallucinated commands fail validation before ever reaching hardware. The ADAM architecture is designed for generality, and we demonstrate this capacity on two very different instruments and experiments: automated ion processing on a plasma focused ion beam (PFIB) microscope and open-ended discovery on a scanning transmission electron microscope (STEM). Following developments at synchrotron facilities, we employ an emulator-to-live-instrument approach that allows us to deeply probe the planning and execution of the architecture. Collectively, RAG-grounded reasoning, registry-bounded actions, and selective HITL gating provide a blueprint for safe, scalable physical experimentation.