LORE: A Deployed AI Agent for Physics Experiment Documentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present **LORE** (Lab Observation and Report Engine), a deployed AI agent that eliminates the documentation bottleneck in experimental physics. A typical measurement session produces notebooks, data files, instrument logs, and lab notes scattered across heterogeneous platforms; synthesizing these into a reproducible record takes hours of researcher time that is routinely deferred or skipped. Given a natural language request via Slack, LORE fetches all experimental artifacts from their native sources, resolves lab-internal software dependencies to recover hardware constants not present in the experiment code, interprets results using domain physics knowledge, and uploads a structured report directly to the laboratory's electronic notebook. Beyond report generation, LORE supports on-demand dilution refrigerator status checks, standalone cryogenic condition reports, and electronic lab-notebook search and retrieval all through natural language. Unlike prior AI-for-science systems that operate as prototypes on curated data, LORE is a running system in an active research laboratory, designed to be experiment-type agnostic. We describe the architecture, two concrete deployment examples, and lessons learned, with superconducting device measurements planned as a near-term extension.

1 Introduction

Experimental physics generates a heavy documentation burden alongside its scientific data. A single measurement session may involve Jupyter notebooks, CSV outputs, instrument logs, configuration files, and hand-written lab notes across multiple platforms. Synthesizing these into a coherent, reproducible record requires hours of researcher time that is often deferred or skipped under the pressure of ongoing work.

LLMs have shown strong capabilities in scientific writing and reasoning Boiko et al. [2023], Lu et al. [2024], and recent work has demonstrated autonomous agents that can *conduct* superconducting qubit experiments Li et al. [2026]. LORE occupies a complementary role: it operates *after* the experiment, assembling and interpreting what a human researcher produced, with no connection to instruments and no risk of hardware interference.

LORE makes two contributions: (1) a deployed multi-source artifact fusion and report generation pipeline, triggered via natural language Slack messages with built-in Slack history search and LabArchives title-based lookup to resolve informal experiment references without requiring the user to supply exact identifiers; and (2) a physics-aware cryogenic capability ranging from quick on-demand DR status checks to full standalone dilution refrigerator condition reports, grounded in superconducting device physics. The system is designed to be experiment-type agnostic; superconducting device measurements are planned as a near-term extension.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

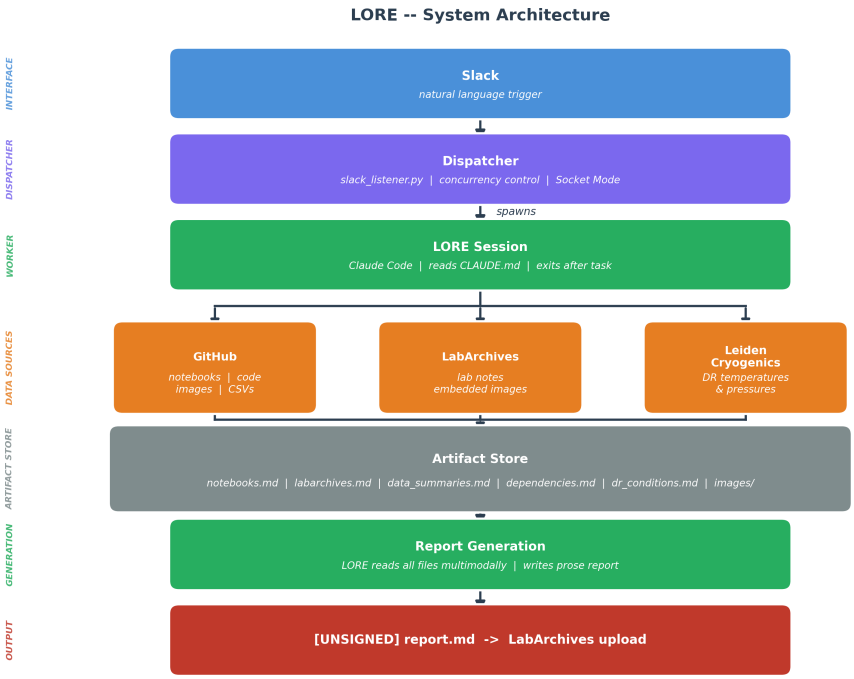

Figure 1: LORE system architecture. A lightweight always-on dispatcher receives natural language Slack messages and spawns on-demand agent sessions. Each session fetches artifacts from multiple sources, assembles a flat artifact store, generates a structured prose report, and uploads it to the laboratory's electronic notebook.

2 System

Figure 1 shows the full architecture. A lightweight always-on dispatcher (`slack_listener.py`) receives Slack messages and spawns a fresh Claude Code session Anthropic [2025] for each request. No persistent LLM state, no concurrency issues. The session reads a structured instruction file (`CLAUDE.md`) encoding all pipeline logic, report schema, and domain physics knowledge, then exits when the task is complete. Although each session carries no persistent LLM state, the Slack channel itself serves as a shared log where prior bot replies and user messages remain in channel history and can be retrieved by a new session when context from previous interactions is relevant. When a request is ambiguous — e.g., "the KID sweep from last Tuesday" the agent proactively searches Slack message history via the bot's `read` scope to locate a relevant GitHub URL or LabArchives page where applicable, before asking the user, a common pattern in fast-moving laboratory environments where experiment references are informal.

2.1 Artifact Pipeline and Report Generation

Given a GitHub URL and one or more experiment references, the pipeline: (1) discovers and ingests all repository files (notebooks, CSVs, images, configuration files), extracting notebook cells and computing data statistics; (2) resolves dependencies by matching notebook imports against the organization's GitHub repositories, fetching lab-internal package source to give the agent access to hardware constants and measurement class implementations not present in the experiment code itself; (3) retrieves LabArchives [2026] page content via authenticated API calls, pages are matched by approximate title so users need not supply internal identifiers, including embedded images; and (4) writes flat markdown files for the agent to read. No pre-summarization is applied and the agent receives all artifacts and applies its own judgment about relevance.

The agent then reads all output files including images interpreted multimodally and writes a structured prose report, scaling depth to the complexity of the experiment: from a concise summary for a simple single-sweep to a full ten-section document for multi-session campaigns. Reports are physics-specific: the Methods section integrates hardware constants from fetched dependency source code,

and results are interpreted against physical expectations rather than simply summarized. Reports are uploaded via Markdown-to-HTML, with images inlined as JPEG data URIs (under 1 MB), and marked `[UNSIGNED]` until reviewed by a human researcher.

2.2 Cryogenic Monitoring and Reporting

Leiden Cryogenics dilution refrigerators produce continuous plain-text `.dat` log files. A dedicated parser reads these directly, no database required, extracting readings from all thermometry channels. Gap detection flags thermal excursions and characterizes their severity.

The cryogenic capability operates at two levels. For *quick status checks* ("how's the DR right now?"), the agent parses the last two hours of logs and replies in Slack within $\sim$30 s — no report generated, no upload required. For *standalone condition reports* ("log the DR conditions for Feb 18, 14:00–22:00"), the agent produces a full structured report covering temperature stability, pump-line pressure, and a physics-grounded verdict on measurement suitability, then uploads it to LabArchives. When a condition report accompanies an experiment report, the agent assesses whether conditions were suitable for the measurements taken. For example, flagging that the refrigerant pump was inactive during part of the measurement window, meaning the system was warming rather than actively cooling.

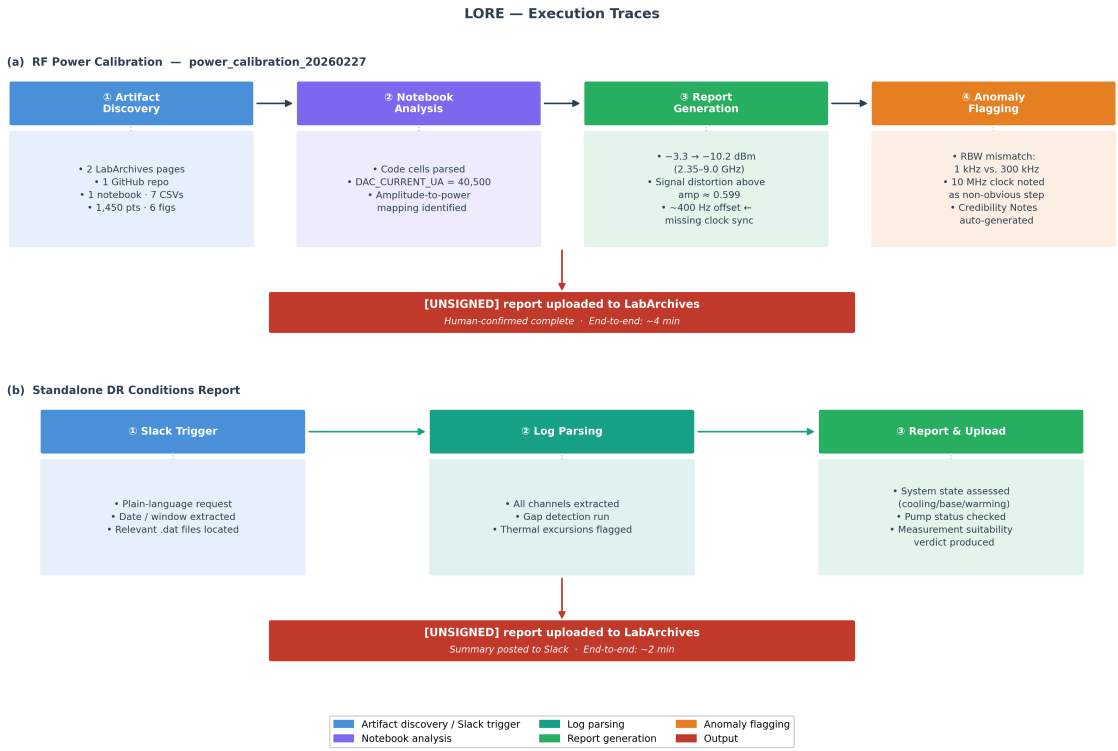

Figure 2: LORE execution traces for both deployment examples. (a) RF power calibration: four stages — artifact discovery, notebook analysis, report generation, and anomaly flagging. (b) Standalone DR conditions report: three stages — Slack trigger, log parsing, and report upload.

3 Deployment Examples

3.1 RF Power Calibration

We trace LORE's execution on a DAC power calibration experiment that characterized the Presto platform's output power across 2.35–9.0 GHz as a function of amplitude and bias current, a prerequisite for converting software amplitude settings to calibrated power levels at the quantum device. Figure 2(a) shows the execution trace. Execution proceeded in four stages:

1. **Artifact discovery.** LORE fetched two LabArchives pages (resolved by title search) and one GitHub repository, classifying the contents into one notebook, seven CSV calibration files (1,450 data points), and six figures.
2. **Notebook analysis.** The agent extracted the hardware constant `DAC_CURRENT_UA = 40,500` directly from the notebook code cells, a value essential for interpreting the amplitude-to-power mapping.
3. **Report generation.** The agent interpreted all six figures multimodally and identified: output power rolls off from -3.3 dBm at 2.35 GHz to -10.2 dBm at 9.0 GHz; signal distortion appears above amplitude ≈ 0.599; and a $\sim$400 Hz frequency offset traced to missing reference clock synchronization.
4. **Anomaly flagging.** The agent independently flagged a resolution bandwidth mismatch between sessions not noted by the experimenter, and identified the shared 10 MHz clock connection as a non-obvious reproducibility step absent from the notebook.

The report was confirmed by a human researcher as a suitable permanent record. End-to-end time from Slack trigger to uploaded report: $\sim$4 min.

3.2 Standalone Dilution Refrigerator Conditions Report

LORE can also generate a DR conditions report with no experiment input at all. A plain Slack message — e.g., "give me a DR report for Feb 18, 14:00–22:00" — triggers the parser for the specified window. The agent extracts temperature and pressure readings; identifies any thermal excursions via gap detection; and writes a structured report covering: system state (cooling, base, or warming); mixing-chamber temperature and spread (min/median/max); pump-line pressure confirming active refrigerant circulation; and a physics-grounded verdict on measurement suitability. The report is uploaded to LabArchives and a brief summary posted back to Slack. This workflow is entirely decoupled from any specific experiment and serves equally as a cooldown log, a retrospective conditions check, or an infrastructure health record. End-to-end time: $\sim$2 min.

4 Discussion and Limitations

Generalizability. The dispatcher-worker pattern applies to any laboratory using GitHub for code and LabArchives as an electronic notebook. Only the domain physics knowledge in `CLAUDE.md` needs adaptation for a different field; data fetching, Slack search, LabArchives lookup, report structure, and instrument log parsing are field-agnostic. Within superconducting hardware, the next planned extension is superconducting device measurements — qubit coherence and resonator spectroscopy — where the DR conditions capability is directly relevant to assessing the reliability of measured results.

On-demand over always-on. Maintaining a persistent agent session proved fragile in practice: long-running LLM sessions accumulate context, degrade in responsiveness, and cannot handle concurrent requests. The spawned-session model trades ~10–15 s startup overhead for reliability and full context on every task.

Limitations. LORE does not take experimental actions and relies on human sign-off before reports are considered final. Evaluation spans two distinct capability modes (experiment reporting and cryogenic monitoring); broader assessment across additional experiment types and laboratories is an important direction for future work.

Broader impacts. LORE can reduce documentation overhead and improve reproducibility in busy experimental settings, but it also introduces risks if automatically generated reports are over-trusted, if retrieved laboratory context is incomplete, or if sensitive internal records are surfaced too broadly. In the current deployment, these risks are mitigated by keeping the agent out of the control loop and requiring human review before any report is treated as final.

LORE demonstrates that a deployed AI agent grounded in domain physics knowledge can meaningfully reduce the documentation burden in an experimental physics laboratory. The combination of Slack-mediated natural language access, multi-source artifact fusion, notebook code analysis, and physics-interpretive cryogenic monitoring represents a qualitatively different capability from generic scientific writing assistants. The system architecture and pipeline code are available at `https://anonymous.4open.science/r/Lab-Agent-6F11/`.

References

Anthropic. Claude code. Software documentation, 2025.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 2023.

LabArchives. Labarchives electronic lab notebook. Product website, 2026.

Xinyu Li et al. Hal: Closed-loop learning and control for quantum experiments, 2026. Workshop manuscript / preprint.

Chris Lu et al. The ai scientist: Towards fully automated open-ended scientific discovery, 2024.