A Lab-Personalized AI Agent for Reproducible Digital Pathology Research

Varun Kalidindi, Avilash Angirekula, Zarif Azher, Nehan Mohammed, Joshua Levy
Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center

Dartmouth Health

ABSTRACT

Background

- Research projects accumulate across HPC directories, notebooks, scripts, and undocumented institutional knowledge. As projects grow, recovering the data and code behind a prior result becomes difficult. Clinical data governance like HIPAA restricts the use of commercial AI assistants to search them.

Methods

- We deployed an fully local 30.5-billion-parameter language model on Dartmouth's HPC cluster and built a retrieval-augmented generation pipeline over a laboratory project's own files. The retrieval accuracy was evaluated on 30 questions against a real cytology lab project. We're able to make this accessible for lab groups through a Slack interface.

Results

- The system indexed a 49-file laboratory project into 637 retrievable passages and delivered answers from behind the cluster firewall. Profiling identified directory traversal as the dominant indexing cost, reducing index time from 340 seconds to under 1 second. On 30 questions, the correct source file appeared in the top three results for 92% of answerable questions (23/25).

Conclusion

- A locally hosted, retrieval-grounded AI agent can accurately locate prior work within a laboratory's own unpublished files while satisfying institutional data-governance constraints, offering a path to faster onboarding and more reproducible research.

INTRODUCTION

Modern digital pathology projects combine whole-slide images, spatial transcriptomics, clinical metadata, and analysis code. As a project concludes, that work disperses across cluster directories, notebooks, and the undocumented knowledge of whoever ran it. New researchers spend their first weeks reading old folders rather than doing science.

Why an AI agent?

Large language models can answer questions about a codebase, but only when grounded in the actual files rather than training data. Retrieval-augmented generation constrains the model to answer from retrieved source material and cite it, making the output verifiable rather than something a researcher must trust.

Why not a commercial service (like ChatGPT or Claude)?

Using a commercial API with protected data requires a Business Associate Agreement and institutional review, and data-use agreements often prohibit third-party transfer regardless. Cloud endpoints are also versioned and retired, so a result produced against one may not reproduce a year later. A local model keeps weights frozen and citable.

Aims

Ⅰ. Deploy an open-weight model on institutional hardware, connected to laboratory files

Ⅱ. Enable recovery of the data, code, and decisions behind completed results

Ⅲ. Evaluate retrieval accuracy quantitatively against a real project

METHODS

Infrastructure

Qwen3-Coder (30.5B parameters, quantized to 18.6 GB) served locally on one NVIDIA V100 via SLURM on Dartmouth's Discovery cluster. No outbound network requests at any stage

Corpus

A laboratory cytology project (panCyto, A. Panchal), used with permission: 49 files of scripts, configuration, and notebooks, indexed as 637 retrievable passages. Image data and slide manifests excluded by allowlist.

Retrieval Pipeline

Files are segmented into overlapping passages, embedded, and stored in a vector index. A query retrieves the five nearest passages, which become the model's sole context. Responses cite the files they came from.

Access and Safety

Delivered through a Slack interface using an outbound connection, since the cluster cannot accept inbound traffic. Read access is restricted to an approved directory allowlist; write and delete operations are unavailable to the serving process.

Evaluation

30 questions written by reading the corpus, each paired with the file containing its answer. Stratified as locate (10), comprehend (8), reproduce (7), and unanswerable controls (5), the last verified absent from the index before testing.

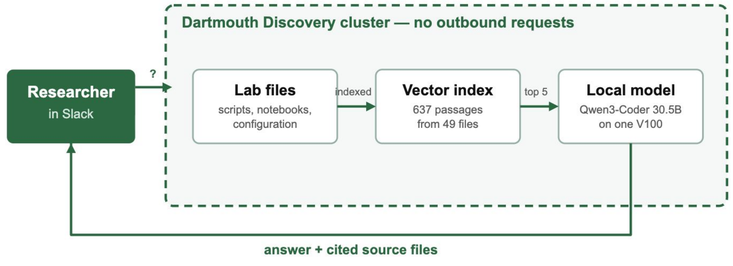

Figure 1. System Architecture

RESULTS

Metric	Result
Correct file in top 3	92.0% (23/25)
Correct file ranked first	52.0% (13/25)
Correct abstention, unanswerable	80.0% (4/5)
False abstention, answerable	0.0% (0/25)
Median response latency	14.0 s

Table 1: Retrieval & Abstention Performance

Retrieval Accuracy

- Questions were written against a project none of the authors contributed to, so performance reflects retrieval over unfamiliar code rather than recall of material the system's builders already knew.
- On five questions whose answers were absent from the corpus, the system correctly declined four times, indicating it does not reliably fabricate when grounding is unavailable.

Results

Indexing Latency

- Profiling attributed indexing cost to directory traversal, not embedding: each run enumerated ~200,000 image files solely to check extensions. Excluding dataset directories reduced index construction from 340 seconds to under 1 second.
- Two interventions were tested and rejected: restricting file types changed nothing, as filtering occurs after enumeration; relocating embedding to CPU produced a fivefold regression and was reverted.

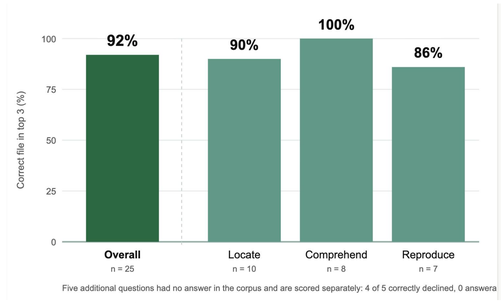

Figure 2: Retrieval accuracy by task type

Failure Analysis

- Two of three failures share one cause. A cervical extraction query returned the thyroid equivalent, and an RGB statistics query missed the file named for those terms. Dense-vector retrieval matches meaning, not literal tokens. Hybrid lexical-plus-dense search should resolve both and is testable against the same questions.
- The third was over-interpretation, not fabrication: a compute-allocation identifier reported as a funding source. The citation verifies, making it harder to detect than invention.

CONCLUSION

A 30.5-billion-parameter model was deployed on institutional hardware and grounded in a laboratory's own unpublished files with no external network request, locating the correct source file for 92% of questions about a project authored by another investigator and declining when information was absent. Indexing a separate finding: file paths and slide manifests carry patient identifiers that content-level filtering does not catch, which generalizes to any retrieval system indexing clinical directories. Limitations include five control questions, a single project, and ground truth pending author confirmation. Continued work will deploy to the laboratory workspace with usage instrumentation and implement hybrid retrieval.

References

edit.