APS-RAG: A DOMAIN-AWARE HYBRID RETRIEVAL AUGMENTED GENERATION SYSTEM FOR ACCELERATOR OPERATIONS AND KNOWLEDGE SYNTHESIS*

R. Sainju†, J. Dariusz, H. Shang, M. Prince, R. Aydelott, M. Cherukara, Y. Sun, M. Borland
Advanced Photon Source, Argonne National Laboratory, Lemont, IL, USA

Abstract

Effective knowledge management is essential for minimizing downtime and maintaining institutional memory in large-scale accelerator facilities. We present APS-RAG, a domain-aware Retrieval-Augmented Generation (RAG) system currently deployed at the Advanced Photon Source (APS), designed to synthesize operational intelligence and facilitate semantic retrieval from dispersed databases. The system consolidates over 10,000 unique documents from four live databases: the Best Electronic Logbook Yet (BELY) scientific electronic logbook, operational Microsoft Teams chat history, the Integrated Content Management System (ICMS), and the Work Request system - under a single retrieval interface. By leveraging the latest frontier LLMs via Argonne's ARGO platform, APS-RAG integrates a specialized query-preprocessing pipeline that performs temporal parsing, domain-acronym resolution, multi-query expansion, and final response generation.

To ensure high precision, a hybrid retrieval architecture is utilized, combining dense vector and keyword searches. Results are aggregated using Reciprocal Rank Fusion (RRF) and refined with cross-encoder reranking to maximize relevance. A 100-question evaluation dataset was built using the InPars methodology, supplemented with qualitative user feedback. Final responses from APS-RAG include inline citations embedded which display the source document chunk and a web-accessible link to the original document. Future developments include multimodal integration and agentic knowledge-graph-aware retrieval.

INTRODUCTION

The Advanced Photon Source (APS), like any large scientific facility, generates an enormous amount of operational data every day. Operating a facility of this scale for stable user operations exercises every part of the operational record: shift entries, logbook entries, trouble shooting and control procedures, equipment replacement and maintenance work orders, and the operations chat messages (e.g., Microsoft Teams). All of these sources contain critical institutional knowledge. Each of these sources is searchable on its own, but no single interface lets APS staff pose a free-form question – e.g., " How to calibrate the new digital cameras in BTS and what's the last study that used those cameras?" -- and receive a synthesized, source-grounded answer.

* Work supported by the U. S. Department of Energy, Office of Science, under Contract No. DE-AC02-06CH11357.
† rsainju@anl.gov

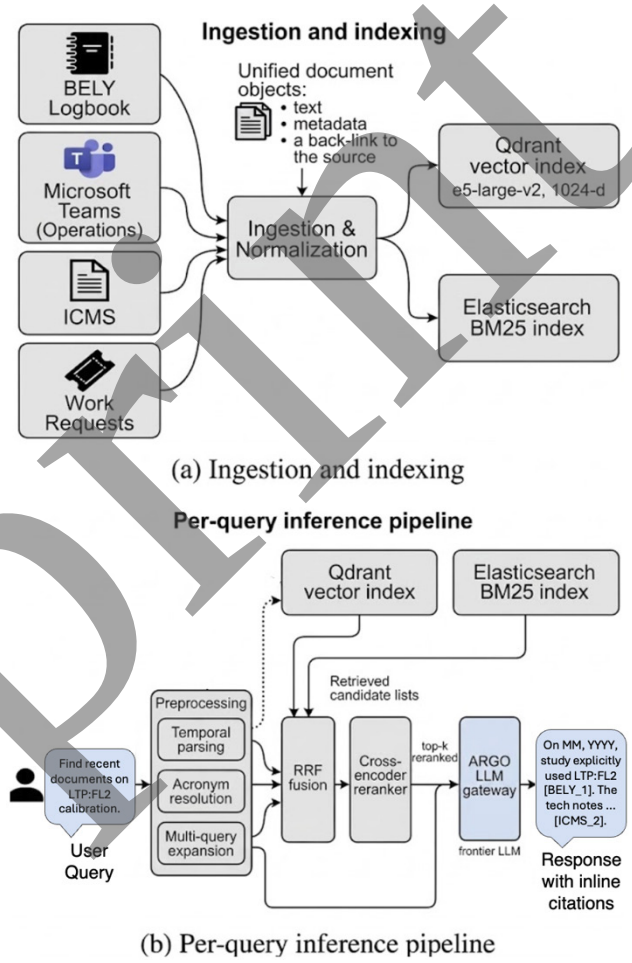

Figure 1: APS-RAG end-to-end architecture. (a) Ingestion and indexing: records from the four operational sources: BELY logbook [1], Microsoft Teams (Operations), ICMS, and the work-request system are normalized into a unified JSON document representation and indexed (b) Per-query inference: the user query passes through temporal parsing, acronym resolution, and multi-query expansion, fans out to the two indices in parallel, and the returned candidate lists are merged by RRF fusion and re-scored by a cross-encoder reranker before the top-k context is passed to a frontier LLM through the ARGO (Argonne National Labs AI platform) gateway, which emits a response with inline, clickable citations.

Several accelerator laboratories have begun closing this gap with LLMs. GAIA [2], developed at DESY, couples an open-weights LLM with logbook and control-system tools through the ReAct prompting paradigm [3]. A multi-

facility effort has begun to pool logbook-RAG setups across DESY, BESSY, Fermilab, BNL, SLAC, LBNL, and CERN [4]. The Electron-Ion Collider team has prototyped a RAG-based summarization agent over operational notes [5]. Broader agentic frameworks for accelerator control have been outlined [6] and recently realized in production form at the Advanced Light Source [7]. Closely related work has applied LLMs to beamline tuning through natural-language prompts [8]. Each of these efforts builds on the original RAG formulation [9] and a fast-moving body of RAG advances [10].

This paper describes APS-RAG, a domain-aware hybrid RAG system that is now in production at the APS. Relative to prior work, the contribution is the integration choices, not any single algorithm: (i) consolidation of four heterogeneous APS knowledge bases under one retrieval surface; (ii) a query preprocessing pipeline tuned to accelerator and APS terminology and to the abbreviated style of logbooks; (iii) a dense + sparse retrieval pipeline with RRF fusion and cross-encoder reranking; and (iv) verifiable, clickable citations back to the originating documents, which we have found to be an critical for users (i.e., scientist, technicians, engineers, operators) trust. A separate manuscript with quantitative evaluation is in preparation.

SYSTEM ARCHITECTURE

Knowledge Bases

APS-RAG consolidates over 10,000 unique documents drawn from four live operational sources that represent day-to-day APS activity: the BELY scientific logbook, the operational Microsoft Teams chat archive, the ICMS (for technical notes, reports, manuals, and schematics), and the work-request system used for equipment and maintenance tickets. Each source has its own access controls, schema, and update cadence. Figure 1 illustrates APS-RAG's end-to-end architecture. An ingestion step normalizes each record into a unified document object that carries the text content, author, and timestamp metadata, as well as a stable backlink to the originating source so that any later retrieval result can be cited unambiguously (Fig. 1a).

Query Preprocessing

Typical queries are short, time-anchored, and heavy in domain shorthand. Three preprocessing steps mitigate this before retrieval (Fig. 1b). (i) *Temporal parsing* resolves expressions such as "yesterday's beam lifetime study" or "during the May maintenance" into explicit date ranges so that retrieval can apply metadata filters alongside semantic matching. (ii) *Acronym resolution* expands accelerator-specific abbreviations: RF (radio frequency), IOC (input/output controller), DCCT (direct-current current transformer), among many others, before embedding. Such tokens are largely out-of-distribution for general-purpose embedders and silently degrade dense retrieval recall if left as-is. (iii) *Multi-query expansion* uses an LLM to generate N paraphrases (N=5) of the user query, which recovers recall on the terse and telegraphic phrasings that dominate operational notes.

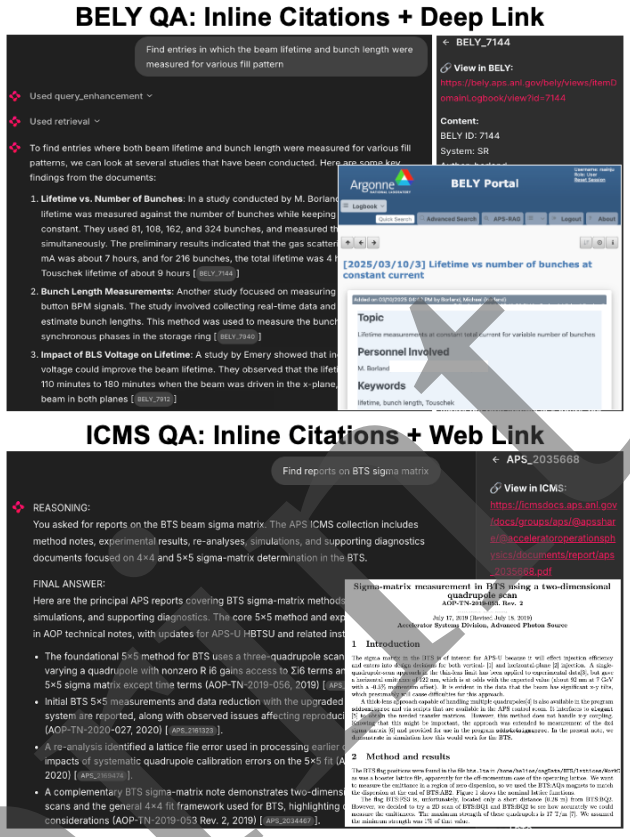

Figure 2: Production user interface and citation surface of APS-RAG. A representative user query and the synthesized response in the chat-style interface, with citation markers rendered inline; sensitive operational details have been redacted. Citation drill-down: clicking a citation marker opens a side panel containing the underlying source chunk together with a deep-link back to the original record in BELY, ICMS, Teams, or the work-request system, so that every claim in the response is verifiable at the source.

Hybrid Retrieval and Reranking

The preprocessed query is dispatched in parallel to two retrievers as shown in Fig 1b. The dense path uses Qdrant as a vector store over 1024-dimensional `e5-large-v2` embeddings; the sparse path uses Elasticsearch with BM25. The two ranked lists are merged using Reciprocal Rank Fusion [11], which is robust to mismatched score scales and avoids brittle linear weight tuning. Hybrid retrieval is critical in this domain. Dense search recovers paraphrased or semantically related entries - for example, a logbook entry mentioning "DCCT noise" in response to a query about "beam current readback drift" - while BM25 reliably anchors on exact identifiers such as sector numbers, process-variable (PV) names, and BELY IDs that operators routinely include verbatim. The top-k fused candidates are then rescored by a cross-encoder reranker [12], which jointly encodes the query and each candidate document, yielding more accurate relevance scores than the bi-encoder used in the initial retrieval stage.

Response Generation and Citation

The reranked context is passed to a frontier LLM (e.g., Claude Opus, GPT, Gemini) accessed through Argonne's ARGO platform, which provides an authenticated gateway to multiple model backbones. Figure 2 shows the production user interface and citation surface of APS-RAG. The synthesis prompt instructs the model to ground each substantive claim in a numbered citation that maps to a specific retrieved chunk. Citations are rendered inline in the response and are clickable in the user interface: a tap opens the source chunk together with a deep link to the originating document in BELY and ICMS. Chats in Teams and the Work request system are opened as HTML links. This citation surface is the primary trust mechanism. In production we consistently observe that users treat the response as a navigation aid into the underlying records, not as an authoritative answer, and the design supports that posture by construction.

EVALUATION AND DEPLOYMENT

Expert and Synthetic-Question Benchmark

Hand-labeling a high-quality evaluation set for a domain as specialized as accelerator operations is expensive and needs to be updated as algorithm improves. In addition to human expert benchmark dataset, we apply the InPars strategy [13] for each document chunk, a strong LLM is prompted with a small number of few-shot examples to generate a plausible operator query whose answer would be found in that chunk. We sample to balance across a taxonomy of question types, including factual lookup, troubleshooting, summarization, and expert routing. The resulting benchmark contains 100 synthetic query-document pairs and is used to measure recall at multiple cutoffs, the relative contribution of each retriever, and the marginal value of each preprocessing step.

Production Deployment

APS-RAG is deployed in production at APS and used by accelerator operations staff and physicists in their normal workflow. The system is exposed via a chat-style interface co-located with existing operator tools, featuring streaming response generation and inline citation rendering. Internal logs of question patterns, follow-up rates, and citation click-throughs provide a second, qualitative evaluation channel alongside the synthetic benchmark.

Observations

Three patterns from early operational use are worth recording because we believe they generalize beyond APS.

First, acronym resolution improves retrieval. Disabling it materially degrades dense-retrieval recall on real operator queries: shift entries and chat messages are written in dense shorthand, and a frozen general-purpose embedder has no prior for tokens like the dozens of three- and four-letter shorthand strings that appear in the knowledgebases. The preprocessing step pays for itself well before the LLM is ever invoked.

Second, the spare retrieval (BM25) matters more than intuition suggests. A nontrivial fraction of real queries from users includes explicit identifiers: PV names, work-order numbers, sector and equipment labels. On exactly these queries dense retrievers are at their least reliable. Hybrid retrieval is not just a quality vs. latency trade-off: it is a precondition for the system being trustworthy for identifier-heavy questions.

Third, citation surfaces are as important as answer fluency. The most common follow-up action after a response is a citation click, not a rephrased question. APS staff read the response, locate the source it pointed to, and continue their investigation from there. This argues for engineering effort spent on the citation surface: chunk granularity, clarity of back-links, and latency from click to source, rather than on chasing marginal generation quality.

ONGOING WORK

Two threads of work extend APS-RAG. The first integrates multimodal content from logbook attachments, screenshots, beam-position images, and plots using vision-language encoders, enabling an operator asking, "show me the orbit feedback behavior during the October trip," to recover both the text entry and the figures it referenced. The second introduces an agentic, knowledge-graph–aware retrieval layer that complements vector and keyword retrieval with structured traversal of operator, system, and event entities extracted from the same corpus. This is aimed at multi-hop questions - "which work requests preceded the most recent camera replacement?" - that are inherently difficult for purely chunk-based retrieval. Both extensions are designed to slot in behind the existing preprocessing and citation surfaces so that the operator-facing experience remains stable.

CONCLUSION

APS-RAG demonstrates that careful, domain-tuned engineering of an otherwise standard hybrid RAG stack, including temporal parsing, accelerator-acronym resolution, multi-query expansion, dense + sparse fusion via RRF, cross-encoder reranking, and verifiable citation, produces a system that operators, scientists, and technicians use in their work at a major synchrotron light source. The system is in production at the APS and provides a foundation for ongoing multimodal and agentic, knowledge-graph–aware extensions. The same architecture should port directly to other accelerator facilities whose operational knowledge is split across logbooks, content-management systems, ticket queues, and group chats.

ACKNOWLEDGEMENTS

The authors express their gratitude to all members of the AOP and Controls group for insightful discussions on the development of APS-RAG. We gratefully acknowledge the computing resources provided on Improv and Swing, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory.

REFERENCES

[1] D. Jarosz, E. Chandler, G. Shen, L. Xiao, N. Arnold, and S. Veseli, "Logging a new era at the APS using BELY," in *Proc. ICALEPCS'25*, Chicago, IL, USA, Sep. 2025, pp. 1482-1487. doi:10.18429/JACoW-ICALEPCS2025-THMG007

[2] F. Mayet, "GAIA: A General AI Assistant for Intelligent Accelerator Operations," May 2024, arXiv:2405.01359. doi:10.48550/arXiv.2405.01359

[3] S. Yao *et al.*, "ReAct: Synergizing Reasoning and Acting in Language Models," Mar. 2023, arXiv:2210.03629. doi:10.48550/arXiv.2210.03629

[4] A. Sulc *et al.*, "Towards Unlocking Insights from Logbooks Using AI," May 2024, arXiv:2406.12881. doi: 10.48550/arXiv.2406.12881

[5] K. Suresh, N. Kackar, L. Schleck, and C. Fanelli, "Towards a RAG-based Summarization Agent for the Electron-Ion Collider," Jun. 08, 2024, arXiv:2403.15729. doi:10.48550/arXiv.2403.15729

[6] A. Sulc, T. Hellert, R. Kammering, H. Hoschouer, and J. S. John, "Towards Agentic AI on Particle Accelerators," Sep. 2025, arXiv:2409.06336. doi:10.48550/arXiv.2409.06336

[7] T. Hellert, D. Bertwistle, S. C. Leemann, A. Sulc, and M. Venturini, "Agentic artificial intelligence for multistage physics experiments at a large-scale user facility particle accelerator," *Phys. Rev. Res.*, vol. 8, no. 1, p. L012017, Jan. 2026. doi:10.1103/jtqy-9jz1

[8] J. Kaiser, A. Lauscher, and A. Eichler, "Large language models for human-machine collaborative particle accelerator tuning through natural language," *Sci. Adv.*, vol. 11, no. 1, p. eadr4173, Jan. 2025. doi:10.1126/sciadv.adr4173

[9] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 9459–9474. [Online], https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[10] S. Minaee *et al.*, "Large Language Models: A Survey," Mar. 2025, arXiv:2402.06196. doi:10.48550/arXiv.2402.06196

[11] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proc. the 32nd international ACM SIGIR conference on Research and development in information retrieval*, Boston, MA, USA: Jul. 2009, pp. 758–759. doi:10.1145/1571941.1572114

[12] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," Apr. 2020, arXiv:1901.04085. doi:10.48550/arXiv.1901.04085

[13] L. Bonifacio, H. Abonizio, M. Fadaee, and R. Nogueira, "InPars: Unsupervised Dataset Generation for Information Retrieval," in *Proc. the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, Jul. 2022, pp. 2387–2392. doi:10.1145/3477495.3531863