

A Semantic Retrieval Framework for the CEBAF Logbook

Jefferson Lab

Chris Tennant



U.S. DEPARTMENT
of ENERGY



SURATech

The CEBAF electronic logbook captures years of operational experience and serves as an important record of how the accelerator is operated, diagnosed, and repaired. Finding that information, however, can be difficult because entries are often terse and filled with acronyms, shorthand, device names, and inconsistent terminology. This project explores a more effective way to search that record by combining semantic search – which retrieves information based on meaning – with conventional keyword search, which remains essential for exact technical identifiers. A curated controlled vocabulary connects operator language with canonical terminology and is applied consistently to both historical entries and user queries. The two search methods are then combined into a single ranked result set. The result is a practical hybrid retrieval framework designed around the way accelerator operators actually document and search for information.

1. Why Search Is Hard

- 130K+ log entries per year
- acronyms, shorthand, PV names, device IDs
- same concept often described in different ways
- exact keyword matching misses vocabulary mismatch

2. Clean + Enrich the Corpus

- convert HTML → plain text
- remove formatting noise and normalize content
- preserve technical identifiers and operational language
- apply domain vocabulary before indexing

3. Build the Search Indexes

- **dense index:** semantic embeddings of log entries
- **sparse index:** BM25 keyword representation
- stemming connects forms such as *vent* / *vented* / *venting*
- wildcards preserve searches for structured PV/device names

4. Controlled Vocabulary

- curated with accelerator subject-matter experts
- maps shorthand ↔ canonical terminology
- examples:
 - BPM → Beam Position Monitor
 - quad → quadrupole
 - 1L10 → North Linac Zone 10
- 277 concepts / 336 search triggers

5. Process the User Query

- apply the same vocabulary expansion used for entries
- create a semantic embedding of the expanded query
- preserve the original terms for exact matching
- search both indexes in parallel

6. Hybrid Retrieval

- semantic search: finds related meaning
- keyword search: finds exact technical terms
- Reciprocal Rank Fusion (RRF) combines both rankings
- final result exploits the strengths of both approaches

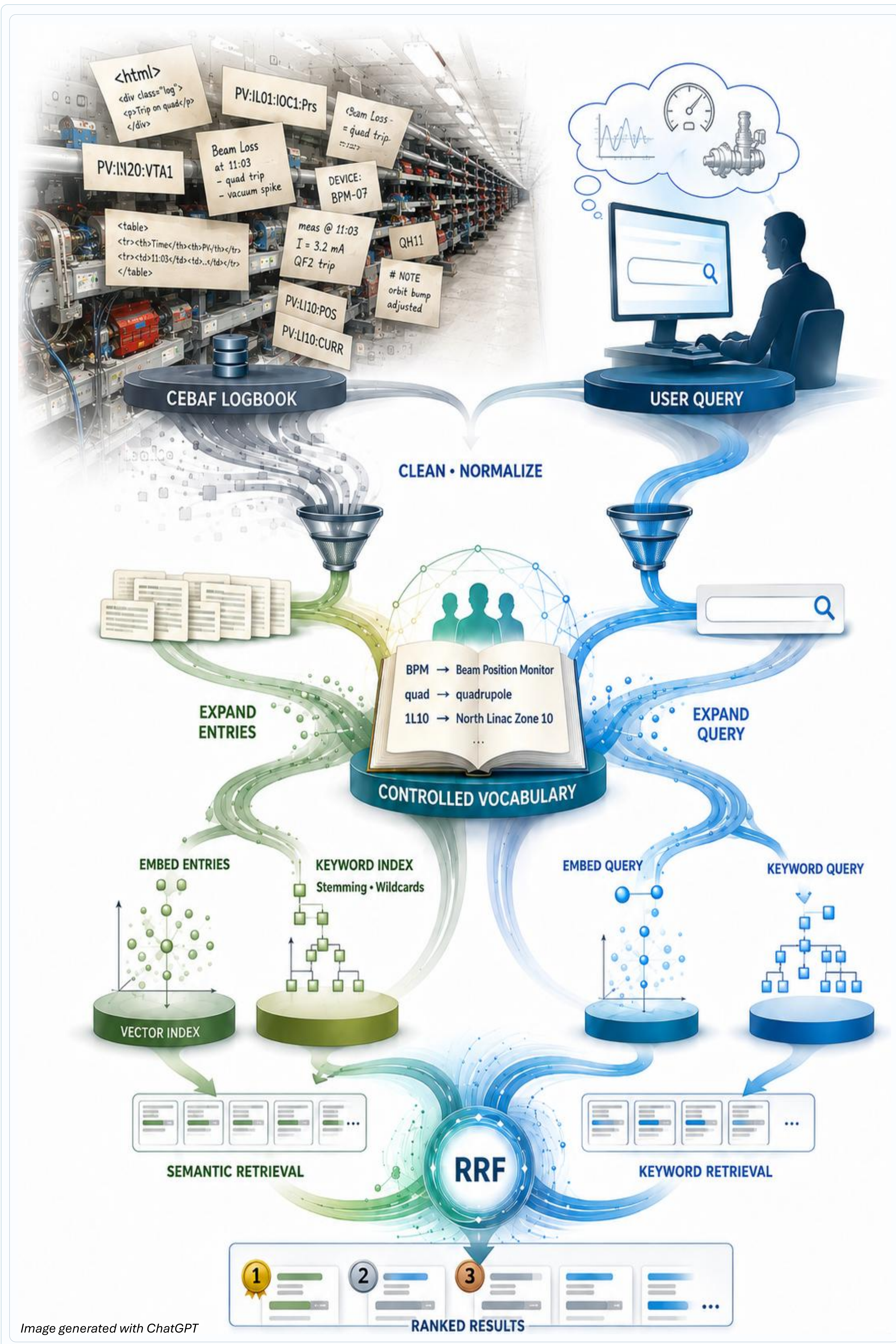


Image generated with ChatGPT

A domain-informed hybrid retrieval pipeline combines semantic understanding with the precision of traditional keyword search

This work was conducted under the Laboratory Directed Research and Development Program at Thomas Jefferson National Accelerator Facility and was supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics, under Contract No. 89243126CSC000213.