LHCbFinder: Semantic Search and Knowledge Discovery Framework

Mohamed Elashri Conor Henderson Michael David Sokoloff

University of Cincinnati & LHCb Collaboration

Vision and Motivation

Current Knowledge Management Challenges at LHCb:

- **Fragmented knowledge** across multiple platforms (TWiki, Indico, arXiv, internal notes)
- **Valuable institutional knowledge** often undocumented or difficult to discover
- **Steep learning curve** for newcomers joining the collaboration

LHCbFinder Solution:

- Centralize scattered documentation in a semantic framework
- Enable intuitive natural language search across all resources
- Preserve and share institutional knowledge
- Reduce entry barriers for new members

Semantic Search Foundation

LHCbFinder employs a powerful semantic search pipeline that forms the foundation of our knowledge platform:

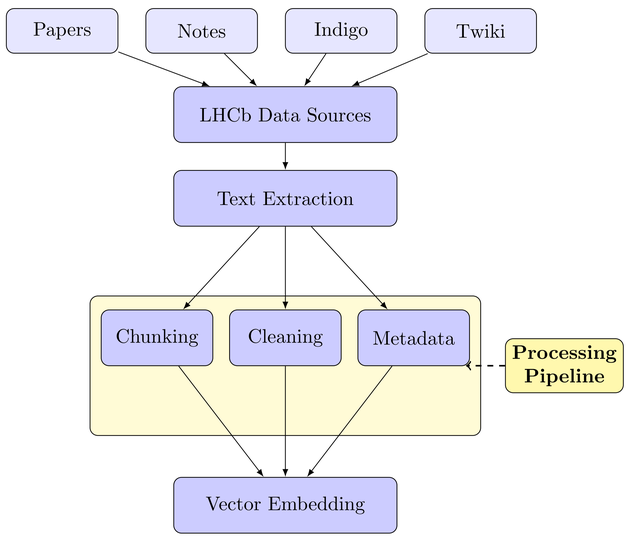

Figure 1. Text processing pipeline for creating vector embeddings

Strategic Implementation:

- **Phase 1:** Abstracts for rapid validation
- **Phase 2:** Full paper introductions for deeper context
- **Phase 3:** Comprehensive scaling to all document types

Key Features

Our semantic search system offers significant advantages over traditional keyword search:

- Understands meaning and context rather than just matching keywords
- Finds conceptually related documents using vector similarity
- Supports natural language queries for intuitive discovery

Understanding Embeddings and Vector Search

The heart of LHCbFinder is our embedding system that converts scientific text into semantic vectors:

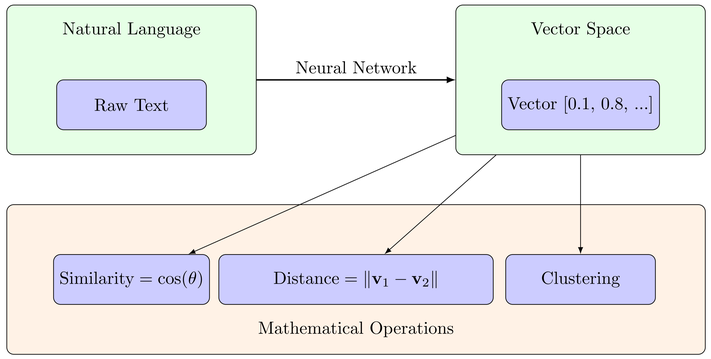

Figure 2. Converting text to vectors for semantic search

Embedding Model: BAAI/bge-large-en-v1.5

- 1024-dimensional vector embeddings
- Optimized for scientific literature
- Excellent technical term handling

Vector Database:

- Efficient similarity search with ChromaDB (local) and Pinecone (cloud)
- Scales to millions of vectors
- Supports hybrid deployment for flexibility

RAG: Retrieval-Augmented Generation Framework

The next phase of LHCbFinder integrates LLMs with our vector knowledge base:

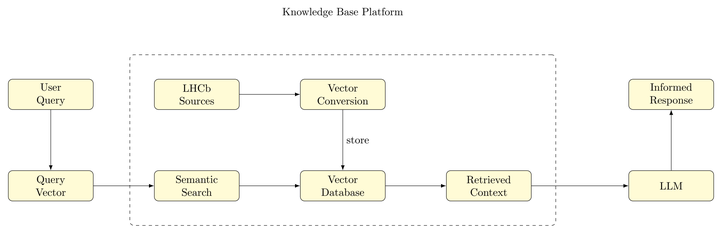

Figure 3. Retrieval-Augmented Generation architecture

Live Demo and Current Status

lhcbfinder.net offers semantic search across LHCb papers with natural language queries. Try searches like *"CP violation in B decays"* or *"Machine learning for tracking"* to explore related papers ranked by semantic similarity.

Development Roadmap

Three-Phase Strategy

Phase 1: Foundation *Current*	Full semantic search for LHCb papers. Optimized embeddings and refined user experience.
Phase 2: Integration *Next*	Developing document scraping/knowledge grabbing pipeline for diverse LHCb resources and integrating with LLM models.
Phase 3: Expansion *Future*	Framework-agnostic development to extend beyond LHCb, enabling knowledge discovery across different HEP experiments.

Current Priorities: Content acquisition, framework modularization, and LLM integration.

Technical Implementation

System Architecture:

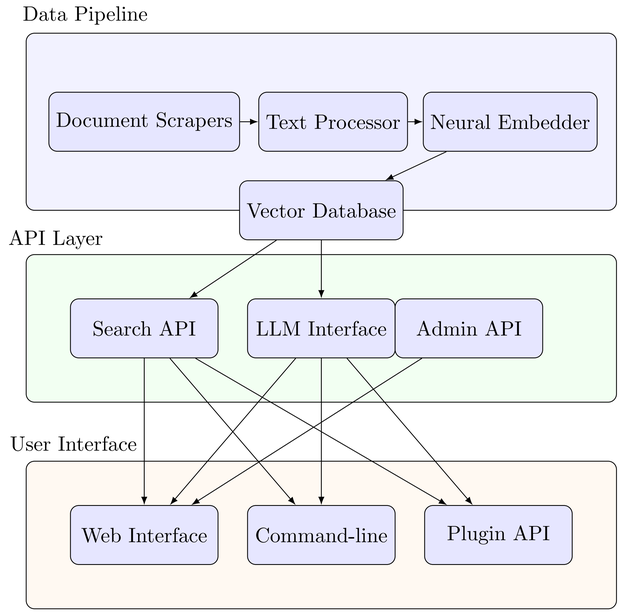

Figure 4. LHCbFinder system architecture

Future Research Directions

LHC-Specific Embedding Models: Training specialized vector embeddings on physics domain terminology to improve semantic understanding of technical concepts in HEP documentation.

LLM Integration Framework: Evaluating various open-source LLMs for compatibility with scientific knowledge retrieval, including benchmarking domain-specific response quality.

Experiment-Agnostic Architecture: Developing a modular design to extend beyond LHCb, enabling knowledge discovery across all LHC experiments with minimal adaptation.

mohamed.elashri@cern.ch