ASKDUNE: CONNECTING DOCUMENTATION, AUTOMATION, AND HPC WORKFLOWS FOR DUNE

Unified Access to DUNE Documentation and Knowledge

A. Rafique – High Energy Physics Division, Argonne National Laboratory
On behalf of the DUNE Collaboration

Contact: aleena@anl.gov

LARGE LANGUAGE MODELS (LLM)

INPUT

Context: You are a helpful assistant....
[potential larger context]
[Chat History: previous interactions]

User: YOUR QUESTION/TASK

send ”out”

PROCESSING

- Ollama (open-weight, local)
- openAI/Antrhopic/Google or: Argo Gateway (closed, remote
- ....
  [model internal reasoning]

OUTPUT

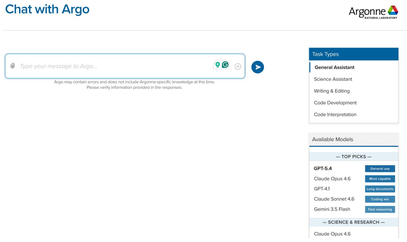
Assistant: GENERATED ANSWER

get ”back”

Impressive: Ingest and “process” 100s of pages within a few seconds
Limitations: Input (and internal reasoning) limited to a few 100 pages of text
Intrinsically only know training data (we only perform inference)
Challenge: Can not provide “all domain knowledge” at inference time
Need to find relevant document/information for user task/user question

LLM INTERFACES

Challenge: can not send internal documentation to any cloud application



Argo Gateway: Provides API interface to “raw” state-of-the-art LLMs: openAI, Anthropic, Google. Cleared for use with internal documents.

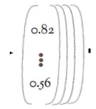
In Addition: Locally hosted models, dedicated Ollama, vLLM server at Fermilab

KNOWLEDGE RETRIEVAL

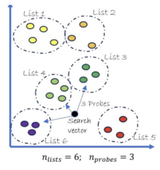
Challenge: Find relevant domain knowledge for a task/user question

Step 1: Raw text extraction from documentation

Step 2: Embed chunk-wise in high-dimensional space



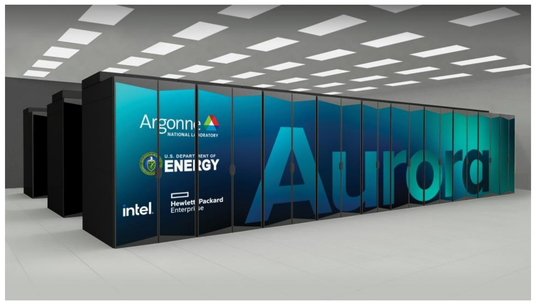
Step 3: Cosine Similarity between embedded chunks and search query



Step 4: Process retrieved knowledge in LLM or agentic system

HPC USE

- Embedding and indexing tasks are distributed and executed on the Aurora supercomputer at Argonne National Laboratory.
- The initial large-scale document ingestion and indexing campaign was performed on Aurora.
- HPC resources enable large-scale parallel processing, with individual jobs handling separate document collections concurrently.
- The Balsam workflow manager is used for job orchestration, monitoring, resource utilization tracking, automated logging, and fault-tolerant execution.
- Following embedding generation and indexing, the resulting vector database and metadata products are transferred to Fermilab for deployment and long-term hosting.



DUNE DOCUMENTATION INTERFACES

Large Language Model Interface for Documentation

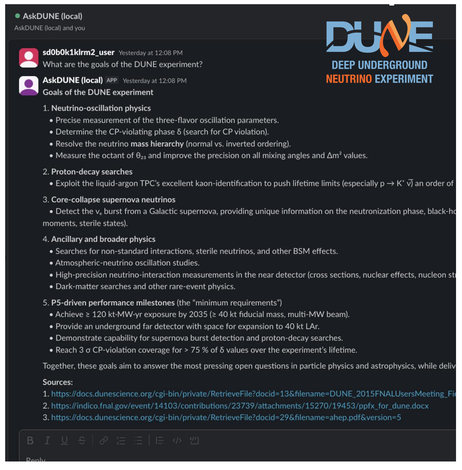
Challenge: i) scattered knowledge: documentation (DocDB, Indico, EDMS, wiki, GitHub) and logbooks/logfiles (ECL)
ii) existing (custom) software often lacks modern (full text) search options

Knowledge Retrieval system + LLMs for processing (RAG or agentic systems)

Example: Web-based unified interface for DUNE [1]

Authentication: Access restricted to authorized DUNE collaborators through Fermilab Single Sign-On (SSO).

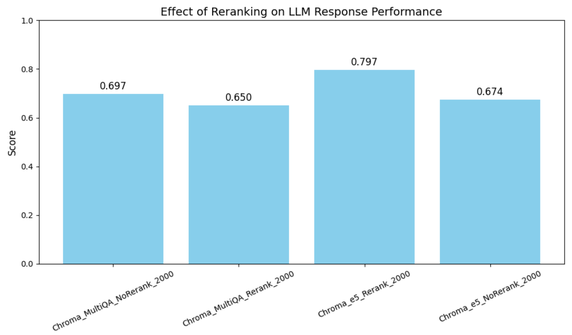
Example: Slack-based interfaces (example from DUNE experiment) with more agentic workflow



BENCHMARKING

Performance Evaluation of the Retrieval Pipeline

A **ChromaDB** vector store combined with intfloat/e5-small-v2 **embedding model** and a document-aware reranking strategy (text vs. slides) provides the best retrieval performance.



Config	[k_docs, k_docs, k_docs]	[k_docs, k_docs, 2k_docs]	[k_docs, k_docs, 3k_docs]
Retrieval Acc.	84.0%	89.17%	90.07%
Latency (s)	1.64	1.82	2.29

We evaluate three retrieval components: **keyword search**, **document-type-aware retrieval** (text vs. slides), and a **reranking layer**. Each component is benchmarked individually and in combination to quantify its impact on retrieval accuracy and latency. The optimized configuration achieves approximately **90% retrieval accuracy** while maintaining low response latency. The retrieval parameter *k* was optimized to a value of **2** [2].

NEXT STEPS:

- Develop and integrate web-based and Slack user interfaces.
- Implement a user feedback and evaluation framework to improve response quality and usability.
- Automate document ingestion, embedding generation, and indexing workflows.
- Expand document coverage and support continuous synchronization with DUNE knowledge repositories.
- Evaluate agentic workflows for enhanced document retrieval and task automation.

[1]: “DUNE Far Detector Technical Design Report, Volumes I, III, and IV”, B. Abi *et al* 2020 *JINST* **15** T08008
[2]: “Large Language Model Integration for Knowledge Retrieval and Interaction for the DUNE Experiment”, https://arxiv.org/abs/2601.05278

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

DEEP UNDERGROUND NEUTRINO EXPERIMENT

Argonne NATIONAL LABORATORY