THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

FolkRAG

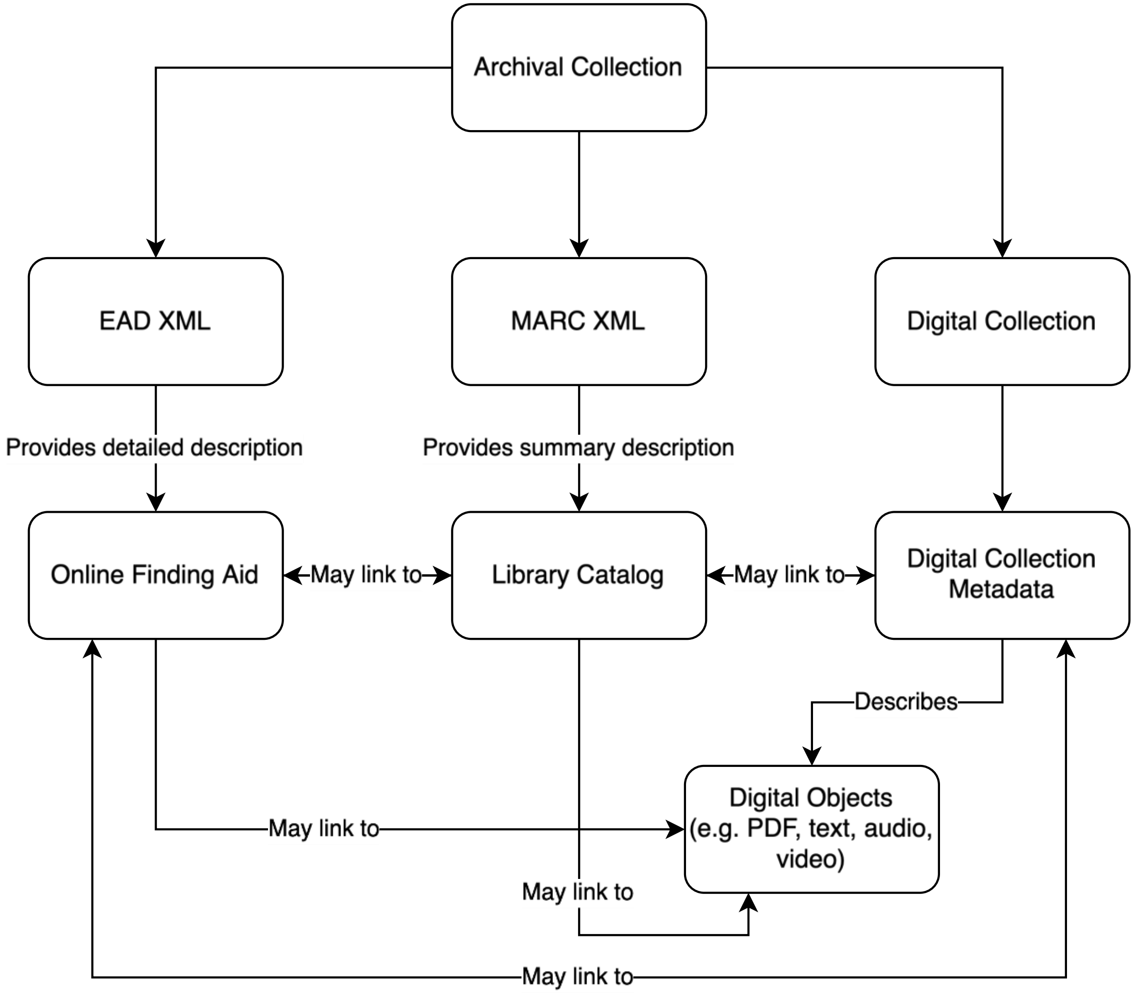
A Retrieval-Augmented Generation System for Cultural Heritage Materials

Jonathan Schild, Paul Kelly

GWU MS DS Capstone, Winter 2024

Archival Data

- Collections described across multiple systems:
  - Catalog records (bibliographic)
  - Finding aids (hierarchical)
  - Digital objects (file-level)
- Current limitations:
  - Multiple disconnected systems
  - Complex navigation
  - Requires expert knowledge



Archival Collection

EAD XML

MARC XML

Digital Collection

Provides detailed description

Provides summary description

Online Finding Aid

May link to

Library Catalog

May link to

Digital Collection Metadata

Describes

Digital Objects (e.g. PDF, text, audio, video)

May link to

May link to

May link to

American Folklife Center

- Division at Library of Congress
- Founded 1976 by American Folklife Preservation Act
- Houses Archive of Folk Culture (est. 1928)
- Materials: Folk songs, oral histories, field notes, manuscripts, photographs, audio/video recordings

FolkRAG

Simple UI that enables researchers and librarians to more efficiently identify their research materials and reference documents.

Proof of Concept

- Combines LLMs with information retrieval
- Enables natural language queries
- Maintains archival context, provides citations
- 94% retrieval accuracy

Data Acquisition and Processing

- Web scraping pipeline:
    - EAD XML from LoC finding aids
    - MARC XML from catalog records
    - JSON metadata via LoC API
- Document processing:
    - 10,980 Whisper audio transcriptions
    - 4,292 Tesseract OCR outputs
    - 29,778 library-created text files
- Metadata integration:
    - Pattern matching for identifiers
    - Hierarchical field mapping

Initial Steps

- Started with FAISS:
    - Fast similarity search
    - GPU optimization
    - Metadata limitations
- Switched to DeepLake:
    - Individual tensor per field
    - Preserved metadata hierarchy
    - Efficient filtering capability

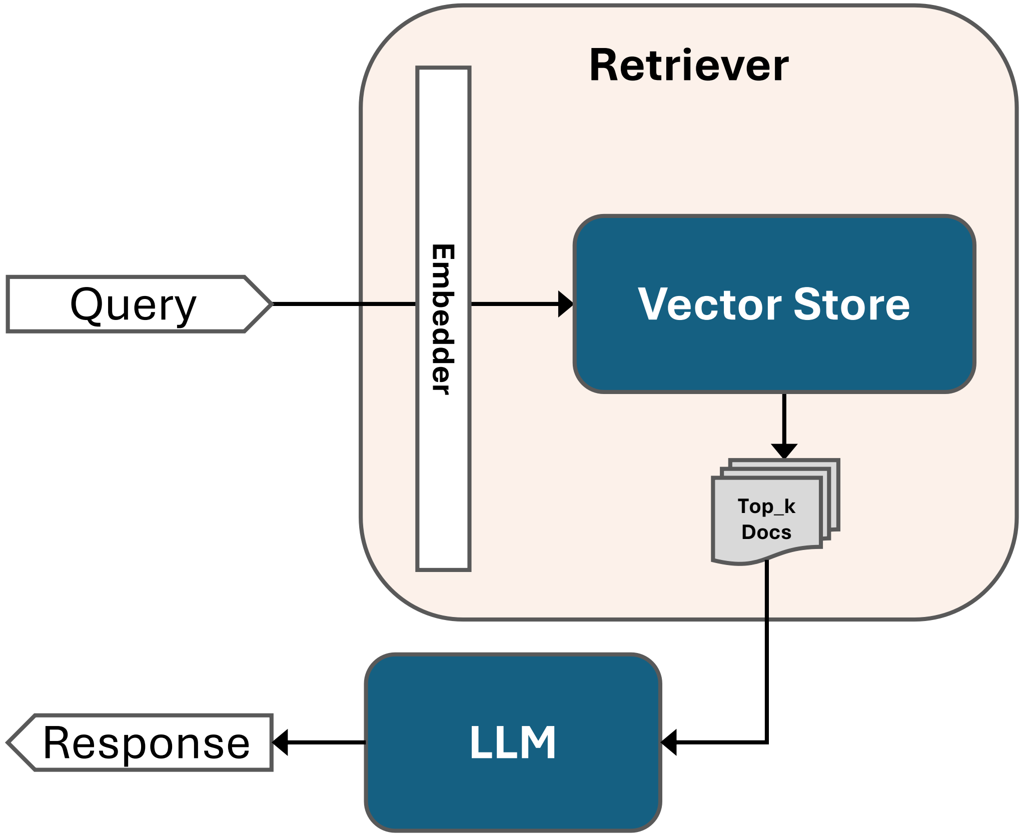
Technical Overview

- DeepLake vector store w/metadata
- Instructor-XL embeddings
- HyDE generator (Claude/Bedrock)
- TF-IDF re-ranking

RAG Architectures: Naïve

Primary Components

- Vector database
- LLM

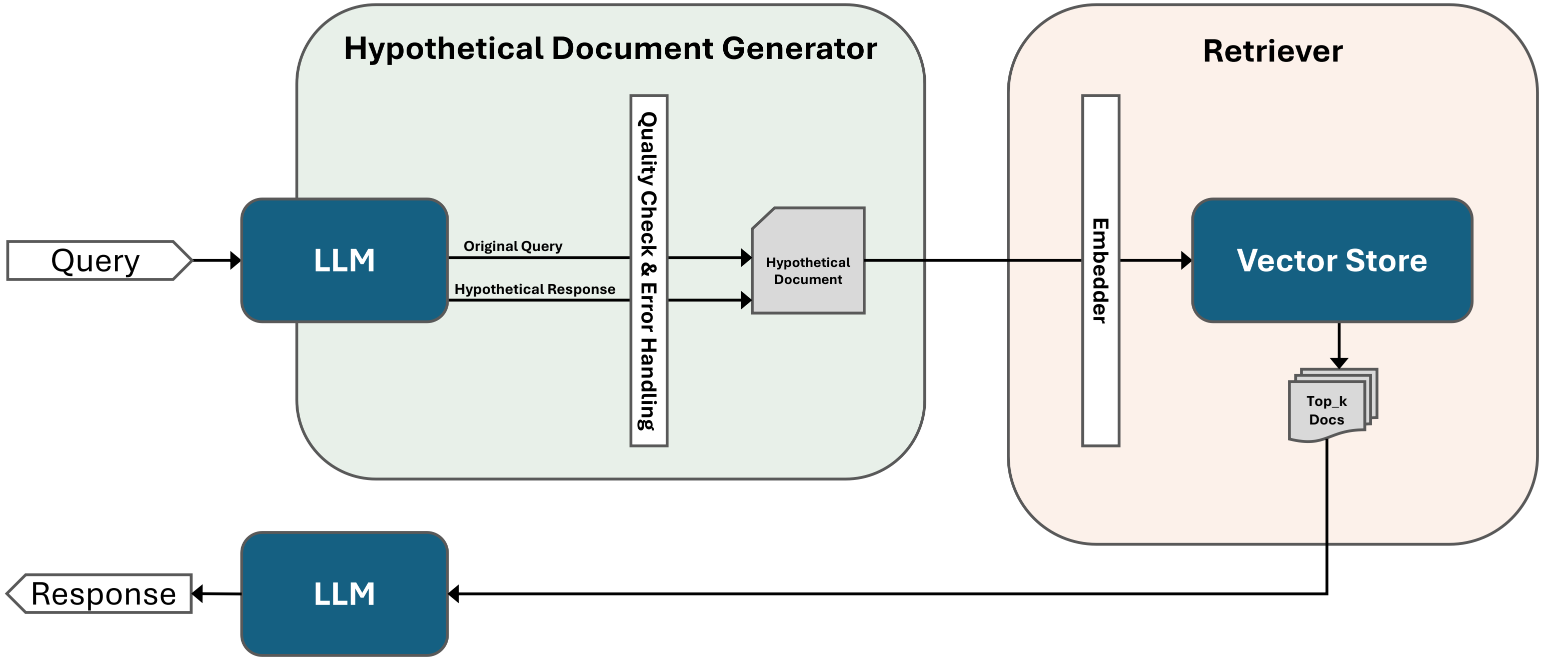


RAG Architectures: HyDE

Naïve + HyDE

Primary Components

- HyDE Generator
- Vector database
- LLM



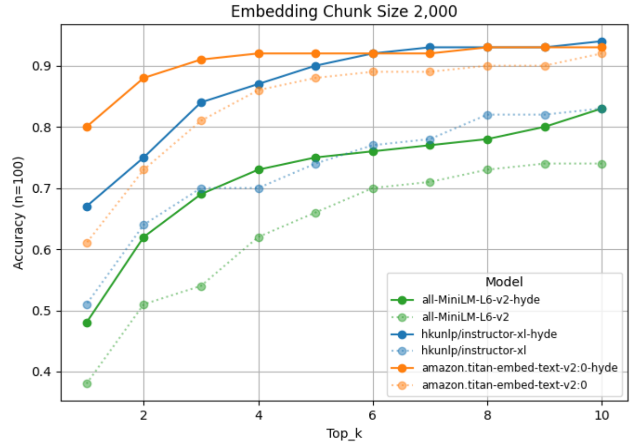
Evaluation: Sample Data

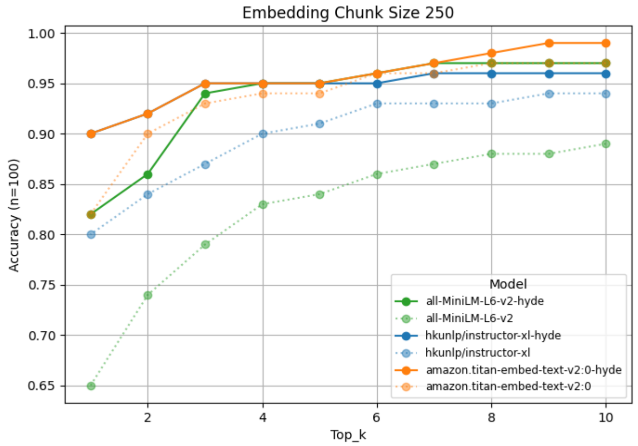
Criteria

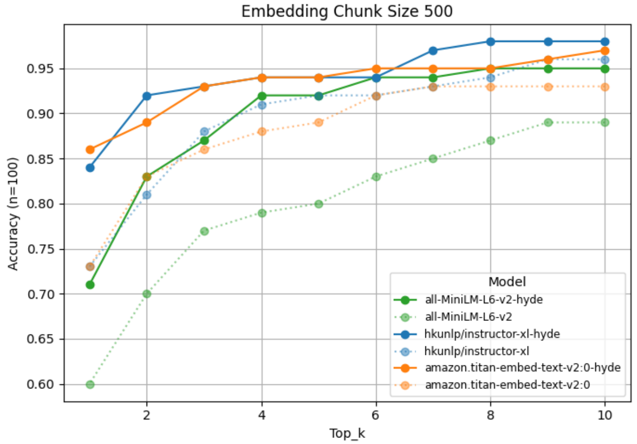
- Accuracy on 100 test questions (Correct document filename among top_k documents)
- RAG System (Naïve, HyDE)

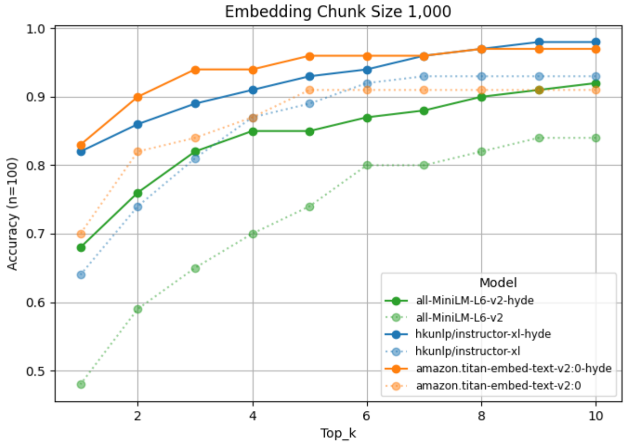
Observations

- Drop MiniLM as an embedding option
- Retain 250, 1,000 sized chunking strategies


Embedding Chunk Size 2,000


Embedding Chunk Size 250


Embedding Chunk Size 500

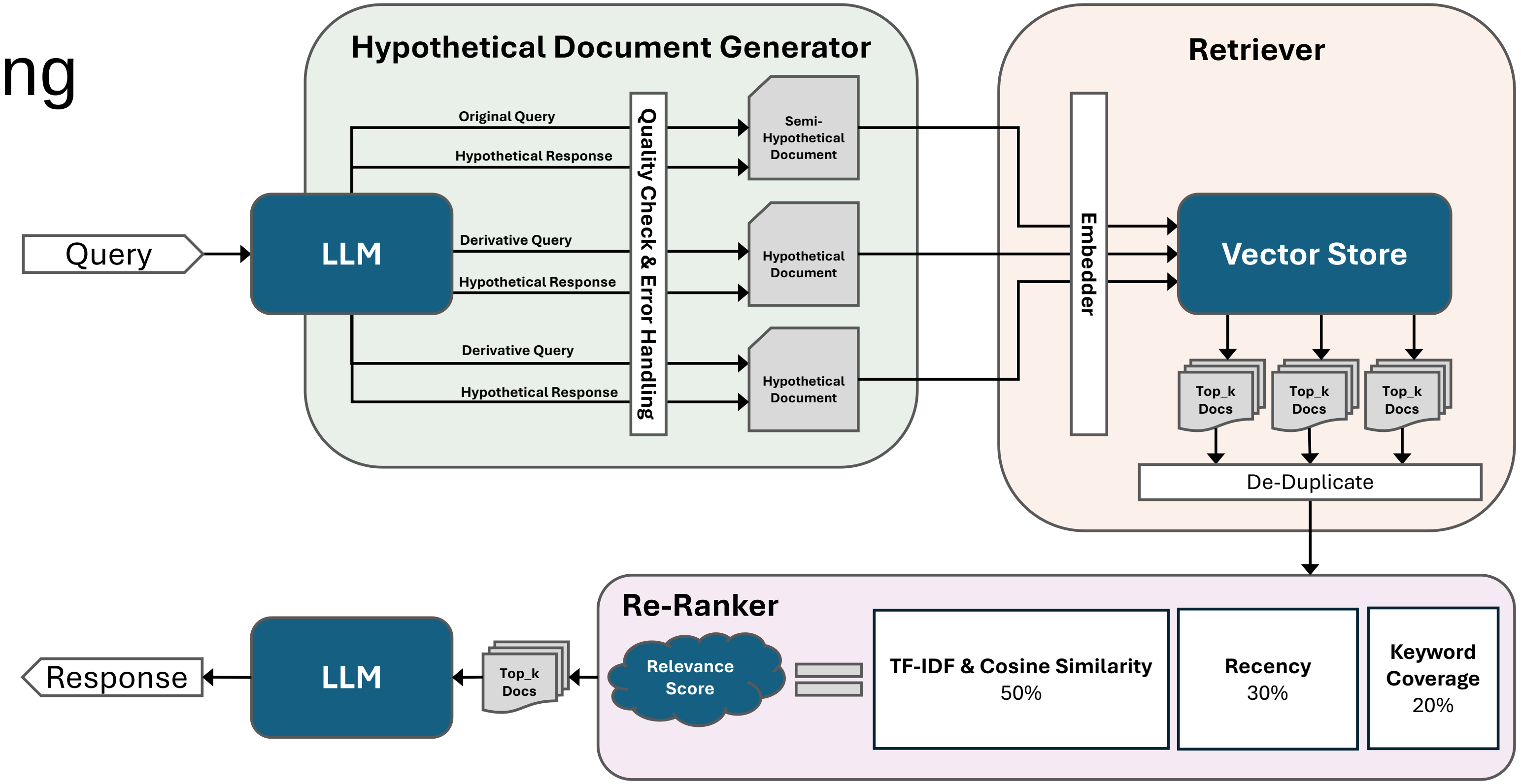

Embedding Chunk Size 1,000

RAG Architectures: HyDER

Naïve + HyDE + Re-ranking

Primary Components

- Enhanced HyDE Generator
- Vector database
- Custom Re-ranker
- LLM



Query

LLM

Hypothetical Document Generator

Original Query

Hypothetical Response

Derivative Query

Hypothetical Response

Derivative Query

Hypothetical Response

Quality Check & Error Handling

Semi-Hypothetical Document

Hypothetical Document

Hypothetical Document

Retriever

Embedder

Vector Store

Top_k Docs

Top_k Docs

Top_k Docs

De-Duplicate

Re-Ranker

Relevance Score

TF-IDF & Cosine Similarity 50%

Recency 30%

Keyword Coverage 20%

Top_k Docs

LLM

Response

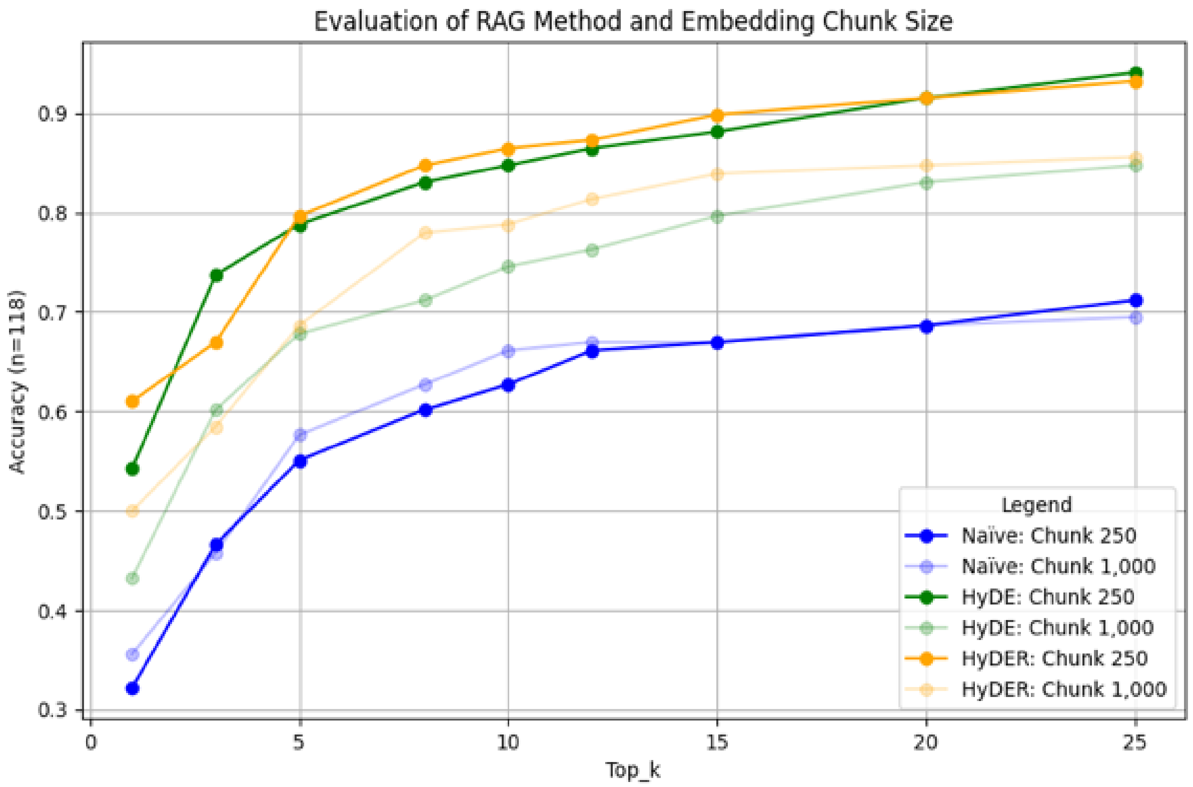
Evaluation: All Data

Criteria

- Accuracy on 18 test questions (Correct document filename among top_k documents)
- RAG System (Naïve, HyDE, HyDER)

Best Model Config

- Instructor-xl
- Smaller chunking strategy (250)
- HyDER (slight preference)



Evaluation: Generation

Query: *What is Captain Pearl R. Nye's favorite ballad?*

HyDER (250 Chunked Vector DB)

Captain Pearl R. Nye's favorite ballad is Barbara Allen. This information can be found in the context, where Captain Nye states, "I guess Barbara Allen is my favorite of them all." The recording was made on June 27, 1937 in Akron, Ohio, as part of a field interview conducted by John A. Lomax. This information is part of a sound recording titled "Two Sailors: Sea Shanties and Canal Boat Ballads," which is one of 10 radio programs in "The Ballad Hunter" radio series. The original filename for this information is ***sr28a_en.txt****.*

HyDER (1,000 Chunked Vector DB)

Captain Pearl R. Nye's favorite ballad is Barbara Allen. This information can be found in the file "~~***afc1933001_ms403_002.txt***~~*". The context explicitly states: "I guess Barbara Allen is my favorite of them all." This quote is attributed to Captain Nye in the passage. Additionally, the context mentions that the Library of Congress has eighty-three versions of Barbara Allen, further emphasizing the popularity and significance of this ballad.*

The metadata indicates that this information comes from the John A. Lomax and Alan Lomax papers, specifically from a manuscript titled "Adventures of a Ballad Hunter." The document was created between 1907 and 1969 in the United States, though Captain Nye's exact recording date is not specified in the given information.

Results

- Naive RAG baseline: ~45% accuracy
- Basic HyDE: ~75% accuracy
- HyDER: ~90% accuracy at top-k=15
- Chunk size comparison:
    - 250 chars: 94% max accuracy
    - 1000 chars: 87% max accuracy
    - 2000 chars: 82% max accuracy

Conclusions and Future Work

- Future Work:
  - Multimodal Integration:
    - Image understanding
    - Cross-modal relationships
    - Visual context preservation
  - Conversational Interface:
    - Memory of past interactions
    - Context accumulation
    - Thread-based exploration