

DESIGN OF DIGITAL ASSISTANCE ENHANCING DIGITAL LIBRARIES

Milen Kraev

IMI-BAS

INTRODUCTION AND PROBLEM STATEMENT

- Traditional digital academic repositories rely heavily on keyword-based indexing that fails to capture deep semantic meaning.
- This limitation creates a major bottleneck for science corpora.
- User expectations are higher and not limited to receiving raw data.
- To bridge this gap, integrating an Intelligent Conversational Agent offers a semantic solution beyond rigid query systems.

THE CORE OBJECTIVE

- The goal is to transform digital libraries from static storage spaces into dynamic, engaging research assistants.
- By leveraging Large Language Models (LLMs), the assistant aims to understand natural language questions directly from researchers.
- The system synthesizes data and provides clear, concise answers backed by direct quotes from the source material.

RELATED WORK

- Academic case studies demonstrate that conversational AI frameworks can effectively automate user navigation and handle complex FAQs.
- Similarly, real-world deployments like the Ivy chatbot at Lehman College showcase virtual reference assistants that learn from historical user queries.
- The Luigi Einaudi AI chatbot utilizes a large language model trained on over 250,000 words of his original writings, allowing users to have interactive conversations about his life, liberal thought, and complex economic policies.

PROPOSED SYSTEM OVERVIEW AND BULDML

- The developed AI module introduces a retrieval-based system designed to supply context-dependent information dynamically.
- The Bulgarian Digital Mathematical Library (BulDML) serves as the initial testing field to verify the agent's efficiency.
- BulDML's conceptual model seamlessly bridges the user and the repository by combining contextual retrieval with generative response synthesis.

LIGHTWEIGHT RETRIEVAL MECHANISM

- The module scans indexed documents and prioritizes up to three relevant files containing matching keywords.
- To guarantee response grounding under strict computational constraints, the system extracts the predefined amount of content from selected documents.
- If no exact match is found the system return answer to the user asking for more specific question

THREE-LAYER SYSTEM ARCHITECTURE

- The assistant is architected across three interconnected layers:
 - Handling data retrieval
 - LLM processing
 - User interaction.
- The data layer maintains an XML-based document index containing metadata and raw text representations of digitized materials.
- Error handling mechanisms are embedded within this layer.

HANDLING DATA RETRIEVAL

- Lexical enhancements introduce stemming and regular expressions to enable flexible, partial word matching based on grammatical root forms rather than strict exact string matches.
- Weighted retrieval upgrades the document selection logic to statistical ranking algorithms like TF-IDF or BM25 to prioritize files based on keyword frequency and uniqueness.
- Hybrid RAG systems transition the architecture to a vector-based model by breaking texts into overlapping chunks and generating semantic embeddings for deep meaning-based searches.
- Intelligent LLM routing utilizes an upfront AI model to evaluate the user's intent and extract optimized keywords, querying the metadata before passing the selected documents to a downstream synthesis agent.

BACKEND INFERENCE AND LOCAL LLM INTEGRATION

- Once relevant document fragments are retrieved, the system constructs a dense prompt combining the context and the user query.
- This prompt is dispatched to a locally hosted Mistral LLM running on a self-hosted Ollama inference server.
- The local LLM interprets the data to synthesize human-like, factually accurate responses without relying on external cloud APIs.

FUTURE ENHANCEMENTS AND EVALUATION

- Future iterations will focus on software upgrades for query preprocessing and exploring alternative proprietary cloud-hosted models.
- Scaling plans include provisioning additional computational power to support advanced token processing and larger open-source models.
- Ultimately, this implementation marks a successful progression toward evolving digital archives into highly interactive research support systems.

CONCLUSION

- Digital libraries should make their knowledge useable.
- Explainable AI is the road to above goal, integrated in proper ecosystem between the knowledge, agent and the users.