

Librarian, Query My Library! [1]

Question-Answering with Open-Source LLMs and Local Books

Razvan Florian Vasile

razvanflorian.vasile@studio.unibo.it



Department of Computer Science
University of Bologna

October 20, 2024



- 1 Introduction & Motivation
 - The Need for a Study Companion
 - Overcoming Limitations
 - Agent characteristics
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- 1 Introduction & Motivation
 - The Need for a Study Companion
 - Overcoming Limitations
 - Agent characteristics
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- ***AI Study Companion:*** friendly and useful.
- ***Key Attributes:*** user-education, explainability, rationality.



- 1 Introduction & Motivation
 - The Need for a Study Companion
 - **Overcoming Limitations**
 - Agent characteristics
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- ***Specialized Domain Questions:*** specialized knowledge.
- ***Fact Verification:*** truth of a statement.
- ***Jeopardy Style Questions:*** questions with cues, then guess.



- 1 Introduction & Motivation
 - The Need for a Study Companion
 - Overcoming Limitations
 - **Agent characteristics**
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- **TAO pipeline.** Break down large questions into smaller subtasks.
- **Citations.** Verify sources and the credibility of the response.



1 Introduction & Motivation

2 Retrieval Augmented Generation

- Key characteristics
- Components

3 Librarian Architecture

4 Visualization Tools

5 Ethical Considerations

6 Future Directions and Conclusion



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
 - Key characteristics
 - Components
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- Embedding *external knowledge*.
- Combines *parametric* and *non-parametric* memory.
- Result: *accurate*, *factual* and *context-aware* responses.



1 Introduction & Motivation

2 Retrieval Augmented Generation

- Key characteristics

- Components

3 Librarian Architecture

4 Visualization Tools

5 Ethical Considerations

6 Future Directions and Conclusion



- ① **Retriever:** Fetches relevant documents from a knowledge base.
- ② **Query Encoder:** Transforms the user's query into a vector representation.
- ③ **Document Index:** Stores the vector representations of documents for efficient retrieval.
- ④ **Generator:** A pre-trained LLM that generates the final answer based on retrieved information and the user's query.



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture**
 - Learning paradigms
 - Key Features
 - TAO Idiom
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture**
 - **Learning paradigms**
 - Key Features
 - TAO Idiom
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- ***Unsupervised Learning:***

- Semantic embeddings (knowledge representation)
- Vector store creation (no labeled data)
- Two distinct embedding models (dimensionality reduction)

- ***Self-supervised Learning:***

- Pre-trained LLM (e.g., Llama 3)
- Answer generation

- ***Reasoning and Planning:***

- TAO (Thought-Action-Observation) cycle
- Dynamic task decomposition



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture**
 - Learning paradigms
 - Key Features**
 - TAO Idiom
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- ***One vector store per book:***

- Separate contextual information.
- Prevents mixing information from different books.

- ***Semantic Chunking:***

- Groups related sentences.
- Improves retrieval accuracy.
- Provides relevant information to LLM.

- ***ReAct Framework:***

- Thought-Action-Observation (TAO) idiom.
- Iterative answer refinement.



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture**
 - Learning paradigms
 - Key Features
 - **TAO Idiom**
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- **Thought:**

- *Assess information* sufficiency.
- *Select* appropriate *tools*.
- Decide to *respond*.

- **Action:**

- *Retrieve additional* information.
- Utilize *tools*.



- **Observation:**

- *Synthesize* fetched information.
- *Reason* about the retrieved information.

- **Final Response:**

- Generated from *observation sequence*.
- Comprehensive *answer formulation*.



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools**
- 5 Ethical Considerations
- 6 Future Directions and Conclusion



- **2D Cluster Visualization:**

- Captures *thematic similarities* between books.
- Highlights *distinct clusters* for different topics.

- **Activation Heatmap:**

- Demonstrates *successful clustering* of information.
- Reveals *thematic similarities* as well.

- **Key Findings:**

- *Effective organization*
- *Clear differentiation* between book themes
- Supports *accurate retrieval* and context understanding



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations**
 - Potential Concerns
 - Mitigation Strategies
- 6 Future Directions and Conclusion



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations**
 - Potential Concerns
 - Mitigation Strategies
- 6 Future Directions and Conclusion



- ***Environmental Impact:***

- High *electricity consumption* of LLMs

- ***Information Integrity:***

- Potential for generating *misleading* or *false information*
- "*Hallucinations*" in AI-generated content

- ***Security Risks:***

- *Malicious actors* embedding undesirable goals
- Potential for system *misuse* or *manipulation*



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations**
 - Potential Concerns
 - Mitigation Strategies**
- 6 Future Directions and Conclusion



- **Energy Efficiency:**

- *Power capping* and *model compression*
- Use of *smaller models* when appropriate

- **User Education:**

- Inform about AI *limitations*
- Emphasize importance of *fact-checking*

- **Governance:**

- Establish *regulations* and frameworks
- Prevent *misuse* and *bias* in AI systems



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion**
 - Future Work
 - Conclusion



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion
 - Future Work
 - Conclusion



- ***External Knowledge Integration:***
 - Wikipedia, *arXiv*, and other sources
 - *Expand* knowledge base
- ***Multi-Agent Workflows:***
 - Enable *interactive conversations*
 - Users interacting with *multiple agents*
- ***Continuous Improvement:***
 - *Scalability* with new open-source LLMs
 - Ongoing *performance enhancements*



- 1 Introduction & Motivation
- 2 Retrieval Augmented Generation
- 3 Librarian Architecture
- 4 Visualization Tools
- 5 Ethical Considerations
- 6 Future Directions and Conclusion**
 - Future Work
 - **Conclusion**



- **Key Technologies:**

- Leverages *RAG*.
- Utilizes *agent planning* using *ReAct*.

- **Contributed Value:**

- Valuable tool for *exploring complex texts*.
- Enhances *analysis*.
- *Facilitates learning* and understanding.

Thank you for the attention



Thank you for the attention!

Any Questions?



- [1] Razvan F. Vasile.

Librarian's repository.

<https://github.com/atomwalk12/librarian>, 2024.