



BEYOND LLMS: A RAG CHATBOT FOR EFFICIENT LITERATURE SEARCH AND THESIS RETRIEVAL IN CSPC LIBRARY

Divino Franco R. Aurellano,
diaurellano@my.cspc.edu.ph

Herald Carl C. Avila,
heavila@my.cspc.edu.ph

ABSTRACT

This study presents a Retrieval-Augmented Generation (RAG) chatbot designed to enhance literature search and thesis retrieval within the Camarines Sur Polytechnic Colleges (CSPC) Library. Traditional library systems at CSPC rely heavily on exact keyword or title matching, limiting the accessibility and efficiency of research, especially for students unfamiliar with specific titles. With the dramatic rise of Artificial Intelligence (AI) and the growing availability of open-source Large Language Models (LLMs), this project integrates the Deepseek R1 model with a RAG pipeline to enable semantic or conversational search of undergraduate thesis documents. The chatbot retrieves relevant content from thesis PDFs and generates accurate, context-aware responses based on real-world academic texts. The initial prototype was implemented using Python and tested on a dataset of ten undergraduate thesis PDFs. The solution currently operates via a command-line interface (CLI), with future plans to transition into a full-featured web application. This research aims to revolutionize CSPC's library as it can serve as a foundation toward bringing the CSPC library's retrieval method closer to the state-of-the-art standards in the context of modern AI.

KEYWORDS:

Retrieval-Augmented Generation; Large Language Models, Thesis Retrieval, Chatbot, CSPC library



1. RATIONALE

Libraries have been around for centuries, serving as repositories of knowledge, support learning, and connect people through information. In fact, PIA Director-General Jose Torres Jr. highlighted the importance of libraries, stating, “Books not only provide information but also open minds to new ideas, experiences, and perspectives” [1]. However, today we are witnessing a dramatic rise in Artificial Intelligence (AI) and the push for digitalization, alongside the unprecedented development of powerful Large Language Models (LLMs), which hold great potential to revolutionize or transform how libraries retrieve and deliver knowledge [2].

In recent years, the demand for digitalization has increased, especially during and after the COVID-19 pandemic [3]. Camarines Sur Polytechnic Colleges (CSPC) library is one of the many university’s library that has started to go digital and continues to improve. However, the current systems and algorithms being used are still far behind the modern AI standards. Searching is still mostly based on keywords, documents often need to be accessed inside the library, and finding information usually means going through papers manually, page by page.

To address these gaps, we propose the idea of using Retrieval Augmented Generation (RAG) chatbot powered by those LLM. In this way, it can serve as a stepping stone toward bringing the CSPC library’s retrieval method closer to state-of-the-art standards in the context of modern AI. It aims to make research easier and more accessible for students and researchers to engage within institution’s knowledge repository, through a semantic and conversational approach.

The feasibility of this study is supported by the availability of digital thesis PDFs, access to open-source LLMs like Deepseek, and current developments in RAG frameworks [4, 5]. The project’s scope is realistically bounded, focusing only on undergraduate theses and deploying the solution on a local server with access control. Despite computational limitations, the project is implementable within the academic timeline and technical capacity of the research team.



2. SIGNIFICANCE OF THE STUDY

This study aims to revolutionize how the academe community interacts with the CSPC library, making research faster, smarter, and more user-friendly by transitioning to emerging technologies and algorithms. The result of this study will benefit the following:

Students. By integrating semantic search and retrieval capabilities, the chatbot will significantly improve search accuracy and efficiency, reducing the time spent on literature review. This will enable students and researchers to quickly find relevant studies without relying solely on exact keywords or titles.

Faculty Members. The chatbot will serve as a research aid for faculty members by providing easier access to relevant studies. This will enhance their ability to aid students in thesis writing, academic guidance, and collaborative research work, while at the same time reducing the extent of manual effort in literature searching.

CSPC Library Management. The implementation of a RAG-powered chatbot will modernize the library's digital infrastructure, making academic resources more accessible to users. By automating thesis retrieval and search functions, the system will improve library service and optimize resource utilization.

Future Researchers. The study will contribute to the field of AI-driven academic search and retrieval, providing insights into the practical applications of Retrieval-Augmented Generation (RAG). Future researchers can build on this work by exploring ways to further optimize search relevance, retrieval efficiency, and integration with other AI models.



3. DESIGN/PROCESS/OUTPUT OF THE STUDY

Finding relevant thesis literature in a University's library, such as in Camarines Sur Polytechnic Colleges (CSPC), can be challenging. Some in our academic community struggle to find the exact thesis paper they need, often requiring them to travel and physically visit the library just to retrieve specific documents.

Currently, CSPC's library website [19] only allows users to search by exact document title. Finding relevant research becomes difficult if users don't know the exact title. Furthermore, library policies restrict users from taking thesis books outside the premises, limiting accessibility to essential academic resources. In response to these challenges, this study aims to explore creating a chatbot that eliminates those limitations by enabling searches based on topics, keywords, or even vague descriptions or what we called conversational search. With the use of the Retrieval Augmented Generation (RAG) algorithm, will revolutionize how the academe community interacts with the CSPC library, making research faster, smarter, and more user-friendly.

3.1 Thesis PDF Dataset

The dataset used in this study consists of 10 undergraduate thesis PDFs from the CSPC digital archive. Each document contains standard research sections such as the abstract, introduction, and methodology. Using PyPDF2, only the first 10 pages were extracted per file to reduce processing time while retaining key information. Metadata such as title, path, and a 200-character summary were stored for each file. This limited dataset is used to test the prototype and can be expanded in future versions.

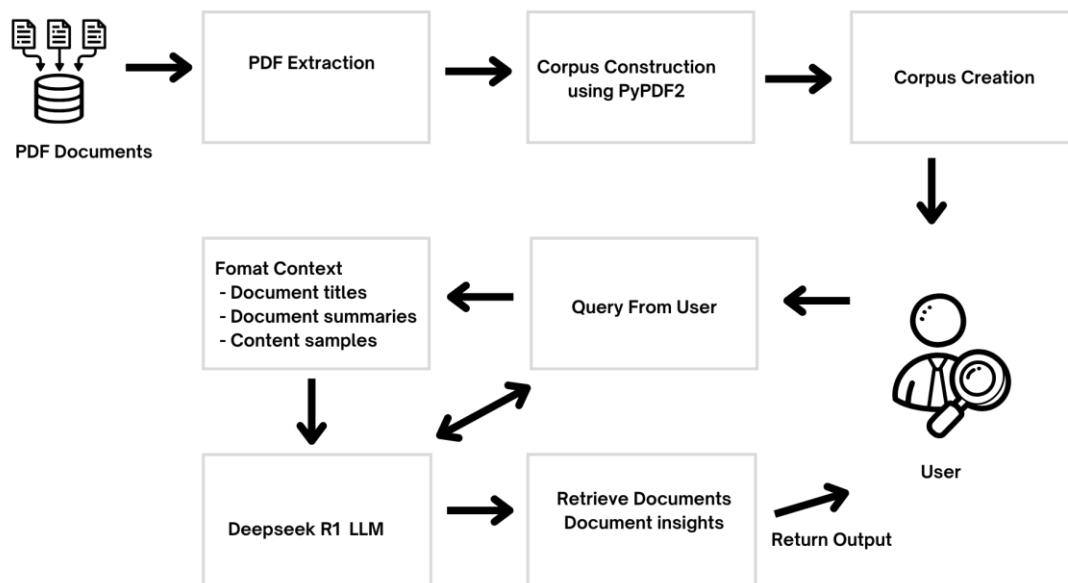
3.2 Methodology

The algorithm implemented in this study is Retrieval-Augmented Generation (RAG), an AI architecture that combines information retrieval with text generation [4]. It enables a Large Language Model (LLM) to provide accurate responses based on real-world documents, in this case, undergraduate thesis papers. Furthermore, Deepseek R1 model was used as the base LLM due to



its open-source availability and strong performance in natural language understanding and generation tasks. The main idea is to retrieve relevant documents and feed them as context into the Deepseek LLM, allowing it to generate fact-based and context-aware answers.

Below is the overall pipeline of our RAG application:



3.3 Implementation

The solution was translated and implemented into a working prototype using the Python programming language. The chatbot runs via a Command Line Interface (CLI) and responds to user queries by analyzing a dataset of undergraduate thesis PDFs from the CSPC Library. The code that was used to run this prototype has been open-sourced and can be found on https://github.com/frankss/Library_Thesis_Retrieval_RAG.git.

Once the program has been launched, the user is prompted to enter any research-related question, keyword, or topic in natural language. The system will then retrieve relevant thesis summaries and provide AI-generated results pointing to the most relevant documents.

Please enter your research topic or question: I am looking for aresearch related to svn that help for brist cancer

Generated response based on PDF documents:
<think>
Okay, let's see. The user is looking for research related to SVM (Support Vector Machine) that helps with breast cancer. They might have a typo with "svn" instead of "SVM". First, I need to check the provided documents to see which ones mention SVM and breast cancer.

Looking at Document 3: The title mentions optimizing breast cancer diagnosis using SVM. That's a direct hit. The content sample talks about using SVM with the Wisconsin dataset, achieving 97% accuracy. Definitely relevant.

Document 4 is about a chatbot using LLM for a help desk. Not related to SVM or cancer. Document 1 and 2 are about chatbots and question-answering systems. Document 1 is about automotive manuals, so not relevant. Document 2 discusses evaluating RAG chatbots, which isn't related to SVM or breast cancer.

So only Document 3 is relevant. The user might be interested in how SVM is applied in cancer diagnosis. The study in Document 3 uses SVM with hyperparameter tuning, which is key. They achieved high accuracy, which is a good finding. No other documents connect here since the others are about chatbots. The most relevant document is clearly Document 3. I should make sure to highlight the methodology and results from that study.

</think>

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS 11
Python + ~ ✕ ...

So only Document 3 is relevant. The user might be interested in how SVM is applied in cancer diagnosis. The study in Document 3 uses SVM with hyperparameter tuning, which is key. They achieved high accuracy, which is a good finding. No other documents connect here since the others are about chatbots. The most relevant document is clearly Document 3. I should make sure to highlight the methodology and results from that study.

</think>

1. Relevant Documents:
  - Document 3 (ML_FINALPROJECT_Aurellano_BSCS2B.pdf) is the most relevant. It explicitly details the use of Support Vector Machine (SVM) algorithms for optimizing breast cancer diagnosis in the Philippines.
  - Document 4 (Aurel_AI_Automating an Institutional Help Desk .pdf) is unrelated, as it focuses on LLM chatbots for institutional help desks.

2. Connection to Query:
  Document 3 directly addresses your interest in SVM for breast cancer, using the Wisconsin Dataset to train an SVM model for early detection. It discusses hyperparameter optimization (e.g., linear kernel, C=0.0001) and evaluates performance metrics like accuracy (97%), precision, and recall.

3. Key Insight:
  The study achieved 99% accuracy in diagnosing breast cancer using SVM, demonstrating its cost-effectiveness and potential for improving early detection in resource-constrained settings like the Philippines.

4. Cross-Document Connection:
  While Documents 1 and 2 focus on chatbots for query resolution, they are unrelated to SVM or healthcare, highlighting that Document 3 is the sole relevant work for your query.

5. Most Relevant Document:
  "Optimizing Breast Cancer Diagnosis in the Philippines Using SVM: A Cost-Effective Method" (Document 3) is the critical resource for your research.

@frankss → /workspaces/BSCS 38 (main) $

```



4. FUTURE PLANS

These plans aim to improve accessibility, accuracy, and scalability for academic research at CSPC.

1. The prototype currently uses 10 thesis PDFs. In the future, the dataset will be expanded to include all undergraduate theses from CSPC and updated automatically with new submissions. And also will be using database to store the chunk texts.
2. The CLI-based system will be transitioned to a web-based interface, making it more user-friendly and accessible for students and faculty.
3. The system will be deployed locally within CSPC's infrastructure, ensuring privacy and secure access for users within the university network.
4. Future iterations will explore using algorithms like RAFT, which combines fine-tuning with LLMs and RAG, to enhance the chatbot's performance and accuracy in retrieving and generating responses.

5. REFERENCES

- [1] Gutierrez, G. L. (2024, March 23). *Public libraries reach over 23M Filipinos; boosts literacy rates*. Philippine Information Agency. <https://pia.gov.ph/public-libraries-reach-over-23m-filipinos-boosts-literacy-rates/>
- [2] Wu, D., Liang, S., Zhao, Y. C., & Zhang, P. (2024, December). Smart Library in the Age of Artificial Intelligence: Vision, Application and Strategy. In Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries (pp. 1-4).
- [3] Ashiq, M., Jabeen, F., & Mahmood, K. (2022). Transformation of libraries during Covid-19 pandemic: A systematic review. *The journal of academic librarianship*, 48(4), 102534.



- [4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
- [5] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.