Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library

Divino Franco R. Aurellano*
diaurellano@my.cspc.edu.ph
Camarines Sur Polytechnic Colleges
Nabua, Camarines Sur, Philippines

Herald Carl N. Avila†
heavila@my.cspc.edu.ph
Camarines Sur Polytechnic Colleges
Nabua, Camarines Sur, Philippines

Almira L. Calingacion‡
alcalingacion@my.cspc.edu.ph
Camarines Sur Polytechnic Colleges
Nabua, Camarines Sur, Philippines

ABSTRACT

Finding relevant thesis literature in the CSPC Library has long been hindered by restrictive search systems and limited access to physical documents. This study addresses these challenges by developing a Retrieval-Augmented Generation (RAG) chatbot that enables users to search for undergraduate theses using natural language queries, topics, and keywords. The system preprocesses and chunks over 290+ thesis PDFs, generates semantic embeddings with all-MiniLM-L6-v2, and stores them in a FAISS vector database. User queries are semantically matched to relevant thesis segments, and responses are generated using the Gemini 2.5-flash model, ensuring grounded and contextually accurate answers. The RAGAS framework was employed to evaluate performance. The model achieved a Context Precision of 0.9167, Context Recall of 0.8711, Answer Relevancy of 0.8625, and Faithfulness of 0.9179. Additionally, user-centered evaluation yielded a weighted mean of 4.5 for response quality and 4.3 for effectiveness and usability, both interpreted as "Strongly Agree". These promising results demonstrate that the chatbot significantly improves literature search efficiency, accessibility, and user satisfaction compared to traditional search systems. The work highlights the impact of data quality and query clarity on retrieval accuracy. This research advances AI-driven information retrieval in academic settings, revolutionizing thesis discovery and supporting the needs of students and researchers.

CCS Concepts

• **Information systems → Information retrieval**; **Retrieval-Augmented Generation**; *Search interfaces*; Document and content analysis; Question answering; • **Theory of computation** → *Neural networks.*

Keywords

RAG, Chatbot, Literature Search, Thesis Retrieval, CSPC Library

ACM Reference Format:
Divino Franco R. Aurellano, Herald Carl N. Avila, and Almira L. Calingacion. 2024. Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library. In *Proceedings of Proceedings of the ACM Hypertext Conference (Hypertext '26).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large Language Model (LLM) like GPT [2] and Gemini [12] have unprecendentedly improved Natural Language Processing (NLP). They perform well in tasks such as semantic search, classification, and clustering, advancing more accurate, context-aware results than keyword-based search methods [5, 16]. These advancements have benefited many fields, including academia. However, LLMs are dependent on the data they were trained on and cannot access real-time or external information. This means they are less useful for Information Retrieval (IR) tasks that require up-to-date or specific data that are not present in their training set, such as finding particular academic resources in university libraries [15].

Writing an academic paper is an important component of research. It requires a deep understanding of the topic and a substantial amount of credible evidence for every statement. This is a challenging and time-consuming role for all the researchers [10]. And for the students, it is essential to first visit the university library to search for and gather existing related literature relevant to their study. However, most libraries today still operate in traditional, non-digital formats where materials are only accessible on-site, making the process of finding and retrieving resources more difficult. Furthermore, some school libraries restricts access and prohibit users from taking home thesis papers. These challenges significantly delay the progress of future academic research due to limited access to relevant literature in university libraries [18].

To address retrieval issues, several universities in the Philippines have recognized the importance of adopting digital archiving systems to improve academic access. This becomes more evident in the last previous year before covid-19 pandemic, when researchers were unable to access library resources, prompting libraries to adapt and make resources accessible even remotely. However, digitalization alone does not fully solve the problem [3, 11, 18]. Unfortunately, most digitalized libraries today still use outdated search systems that need an exact keyword search, which can result in irrelevant materials [20]. The current search algorithm of most digital archives including the Camarines Sur Polytechnic Colleges (CSPC) library still heavily depends on traditional keyword-based search. This poses a challenge when researchers are unsure of the exact title or keywords to input in search bar. And as the usual result, the system will just return a "not found" even though relevant content does exist. This limitation reveals a profound issue in the library's current search capabilities, as minor spelling errors or topic-based queries can prevent users from accessing valuable research.

While numerous studies have explored the integration of the emerging LLM-powered chatbots in academic research [1], their implementation and effect for thesis retrieval in specific university libraries, including CSPC, have not been established. This is primarily due to the limitations of LLMs, which rely solely on pre-trained knowledge and are unable to access or utilize the unique local archives maintained by individual libraries [4, 22].

To address these challenges, RAG has emerged as an effective approach that enhances LLMs by enabling them to retrieve and utilize external, domain-specific documents without retraining [8, 13]. This study developed a chatbot integrated with a RAG pipeline to revolutionize thesis retrieval and searching in the CSPC Library. The researchers makes the following key contributions as objectives:

(1) Integrate a document ingestion and retrieval module for storing thesis documents.
(2) Implement a semantic search and thesis document retrieval system using RAG and Google Gemini.
(3) Evaluate the performance of the RAG chatbot using RAGAS and user satisfaction metrics.

Hypertext '26, June 2025, CSPC, Philippines

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

2 METHODOLOGY

2.1 Research Design

This study adopted a constructive research design to develop the RAG chatbot for CSPC thesis retrieval. This approach suited the study well as it involves addressing the challenges being faced by researchers in searching and retrieving universities' theses by replacing the current yet traditional database and keyword-based search with a vector database and RAG framework, enabling a conversational and topic-oriented approach. Furthermore, the system was deployed to the cloud, allowing students to access thesis everywhere they are, since current library policies restrict users from taking physical thesis books outside the premises.

2.2 Theorems, Algorithms, and Mathematical Models

This study used RAG pipeline, integrated with Gemini 2.5 flash LLM for reasoning that is stored to a vector database. These enabled efficient information retrieval and generation in the context of literature and thesis search within the CSPC Library.

2.2.1 RAG Pipeline. RAG pipeline is a hybrid architecture that combines information retrieval with natural language generation. It allows LLMs to access external documents during inference, thereby improving both accuracy and contextual relevance.

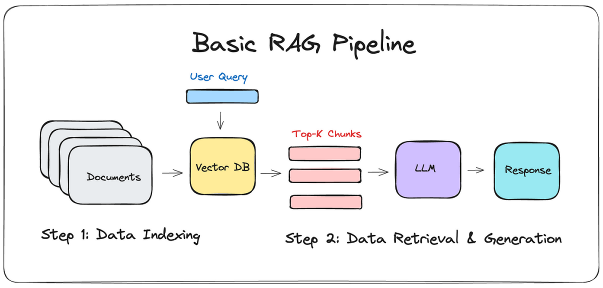


Figure 1: Basic RAG Pipeline by Dr. Julija

The chatbot's RAG pipeline, as illustrated in figure 1, consists of the following key stages:

- **A. Data Indexing:** Thesis documents are loaded and split into smaller chunks using a token-based method respecting academic structure (Abstract; Chapters 1–5). Each chunk is converted into vectors using the 'sentence-transformers/all-MiniLM-L6-v2' embedding model from Hugging Face, chosen for its lightweight architecture and strong semantic representation. Vector embeddings and metadata are stored in FAISS for efficient similarity search.
- **B. Retrieval and Generation:** User queries are embedded using the same model (all-MiniLM-L6-v2). FAISS retrieves the top-K=50 most relevant chunks via semantic search, balancing precision and recall. Retrieved chunks are fed to Google Gemini 2.5-flash as grounded context for response generation.

2.3 Materials and Statistical Tools / Evaluation Methods

2.3.1 Dataset. This study utilized a dataset containing all available undergraduate thesis (initially 290+ pdfs) from various CSPC departments, excluding Computer Science and College of Engineering and Architecture due to unavailability. Good to note here that the system was also designed to ingest new theses, by allowing admin to upload new PDF data

Hardware/ Software Requirements

The system was developed using these hardware and software specification as shown in table 1 and 2

Table 1: Hardware Requirements

Component	Specification
Processor (CPU)	Modern Multi-core CPU
Memory (RAM)	16 GB or higher
Storage	1 TB SSD or higher
Graphics Card (GPU)	NVIDIA RTX 3090+ (recommended)

Table 2: Software Requirements

Component	Specification
Programming Language	Python 3.10+
Vector Database	FAISS
Language Model	Gemini 2.5-flash
Embedding Model	sentence-transformers/all-MiniLM-L6-v2 (HuggingFace)
Web Framework	Flask
Libraries	LangChain PyMuPDF NumPy

RAGAS (Retrieval-Augmented Generation Assessment Suite). RAGAS is a framework for reference-free evaluation of RAG pipelines. This toolkit was used to automate evaluation of the quality of system outputs using its metrics such as context precision, faithfulness, and answer relevance [21]. Furthermore, a context recall metric was included, as recommended for evaluating retrieved chunks.

Survey. The researchers conducted a survey among CSPC librarians and students to evaluate the proposed RAG chatbot. Using a user-centered method that measured users' level of agreement on the chatbot's quality and performance, a questionnaire was created to assess users' satisfaction with answers, likelihood to use the chatbot again, ease of reading and understanding the output, and confidence in the information retrieved by the system. There are 100 respondents in the study from the CSPC who served as representatives of the whole population.

Likert Scale. Introduced by Likert [1932], is a measurement method developed for evaluating individuals' attitudes toward any object [14, 19]. It indicates the degree to which they agree or disagree about the issue. In particular, the 5-point Likert Scale was chosen because it works well in surveys and requires less time and effort to develop as shown in table 3.

Table 3: Likert Scale for User Level of Agreement

Scale	Range	Level of Agreement
5	4.21–5.00	Strongly Agree
4	3.21–4.20	Agree
3	2.61–3.20	Neutral
2	1.81–2.60	Disagree
1	1.00–1.80	Strongly Disagree

Weighted Mean Analysis for Likert Scale Data

In order to analyze the gathered data from the user evaluation questionnaire, the researchers employed the Weighted Mean as the statistical tool. This method was chosen for its effectiveness in summarizing data responses from the Likert scale, in where it can provide a detailed understanding of user perceptions regarding the chatbot's usability and performance. Also, the level of satisfaction with chatbot answers, likelihood of using it again in the future, ease of reading and understanding the output, and users' confidence in the accuracy of the chatbot's responses was evaluated using this computation:

$$WM = \frac{TWM}{N} \qquad (1)$$

Where:

- WM = Weighted Mean
- TWM = Total Weighted Mean
- N = Total number of respondents

2.4 Procedures

The procedure includes the most important stages in building this project. Each step plays a role in addressing this project's objectives.

(1) **Data Preprocessing:** Thesis PDFs were processed using PyMuPDF for text extraction, followed by cleaning and chunking into manageable segments.
(2) **Indexing and Embedding:** Text chunks were embedded with sentence-transformers/all-MiniLM-L6-v2 and indexed in FAISS with relevant metadata.
(3) **Semantic Retrieval:** User queries were embedded using the same model and matched to stored vectors via FAISS to retrieve top-K relevant chunks.
(4) **Response Generation:** Retrieved context was provided to Gemini 2.5-flash to generate human-like responses.
(5) **Output Presentation:** Responses were displayed in a ChatGPT-style web interface built with Flask.
(6) **Performance Evaluation:** System performance was assessed using RAGAS metrics (precision, recall, relevance, faithfulness) and a user questionnaire for usability and satisfaction.

2.5 Evaluation Metrics

The system was evaluated using the RAGAS framework, which encompasses four core metrics: Context Precision, Context Recall, Answer Relevancy, and Faithfulness. Each metric is defined as follows:

2.5.1 *Context Precision.*

Measured the relevance of retrieved chunks.

$$\text{Precision@k} = \frac{\text{true positives@k}}{\text{true positives@k} + \text{false positives@k}} \qquad (2)$$

where true positives@k is the number of relevant chunks retrieved up to position k, and false positives@k is the number of non-relevant chunks retrieved up to the same position. This component metric quantifies retrieval accuracy at each rank and serves as a foundation for the overall Context Precision@K calculation.

2.5.2 *Context Recall.*

Assessed the comprehensiveness of retrieval.

$$\text{Context Recall} = \frac{\text{Supported relevant claims}}{\text{Total relevant claims in reference answer}} \qquad (3)$$

where:

- *Supported relevant claims* refers to the count of factual claims in the ground truth answer that can be attributed to the retrieved document chunks,
- *Total relevant claims in reference answer* represents all the factual claims present in the ground truth answer that ideally should be covered by the retrieval process.

2.5.3 *Response Relevance.*

Evaluated alignment between user queries and generated responses.

$$\text{Response Relevance} = \frac{1}{N}\sum_{i=1}^{N}\cos(E_{g_i}, E_o) \qquad (4)$$

where:

- N is the number of artificially generated questions based on the response (typically 3),
- E_{g_i} is the embedding of the i-th generated question derived from the response,
- E_o is the embedding of the original user query,
- $\cos(E_{g_i}, E_o)$ represents the cosine similarity between the generated question embedding and the original query embedding.

2.5.4 *Faithfulness.*

Ensured factual consistency with retrieved context.

$$\text{Faithfulness} = \frac{\text{Supported claims}}{\text{Total claims}} \qquad (5)$$

where:

- *Supported claims* refers to the count of factual statements in the generated answer that can be directly verified or inferred from the retrieved context chunks,
- *Total claims* is the complete count of all factual statements made in the answer, regardless of whether they are supported by the context.

2.6 Conceptual Framework

The conceptual framework of this study is illustrated in figure 2.

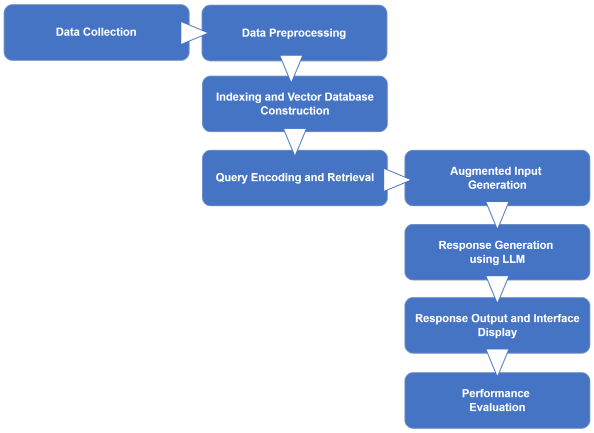

Data Collection
Data Preprocessing
Indexing and Vector Database Construction
Query Encoding and Retrieval
Augmented Input Generation
Response Generation using LLM
Response Output and Interface Display
Performance Evaluation

Figure 2: Conceptual Framework

The proposed conceptual framework implemented a RAG pipeline for thesis retrieval within the CSPC Library. Thesis PDF documents were collected in coordination with library staff and preprocessed by extracting text using PyMuPDF, converting it to markdown, removing non-informative elements, and segmenting the content into coherent token-based chunks. Each chunk then was embedded using the sentence-transformers/all-MiniLM-L6-v2 model and indexed in a FAISS vector database to enable efficient semantic similarity search with associated metadata. User queries were encoded using the same embedding model and matched against the indexed vectors to retrieve the top-K relevant chunks, which were combined with the original query to form an augmented input. Response generation was performed using the Gemini 2.5 Flash large language model, producing context aware and low-latency responses. The chatbot interface was implemented using Flask with LangChain integration and included user authentication and access control. System performance was evaluated using automated RAGAS metrics such as context precision, context recall, answer relevance, and faithfulness, alongside a user-centered survey evaluation.

3 RESULTS AND DISCUSSION

3.1 Document Ingestion and Retrieval Module

This implementation addressed the first specific objective of the study by transforming the library's static collection of thesis PDFs into a dynamic, searchable knowledge base.

3.1.1 Dataset and Preparation. The study corpus comprised all available undergraduate thesis PDFs from multiple CSPC departments (290+ documents). The dataset was prepared via structured text extraction and token-based chunking aligned with thesis sections (Abstract; Chapters 1-5).

3.1.2 Data Preprocessing. Texts were extracted page by page and enriched with metadata (source, page) to preserve academic provenance. Token-based chunking produced coherent segments sized to the LLM context window and guided by thesis structure, improving retrieval fidelity and citation transparency.

Table 4: Chunk Analysis & Statistics

Metric	Value
Total Chunks	38,127
Total Tokens	11,849,783
Avg Characters/Chunk	1323
Avg Words/Chunk	180
Minimum Tokens per chunk	124
Maximum Tokens per chunk	1200
Median Tokens per chunk	335

The overall results of the chunk analysis as shown in table 4 indicate that the implemented chunking strategy is effective and well-structured for document preprocessing. The analysis produced a total of 38,127 chunks with 11,849,783 tokens, where each chunk contains an average of 1,323 characters or approximately 180 words, and a median of 335 tokens per chunk. These results show that the generated chunks fall within an appropriate size range to preserve semantic context while remaining suitable for vector embedding and retrieval.

In terms of chunk size control, the results demonstrate that the system successfully enforced defined boundaries. The minimum chunk size was 124 tokens, which prevented the creation of fragmented or low-information chunks that could negatively impact embedding quality. Meanwhile, the maximum chunk size was limited to 1,200 tokens, ensuring that excessively large chunks that could dilute semantic relevance were avoided.

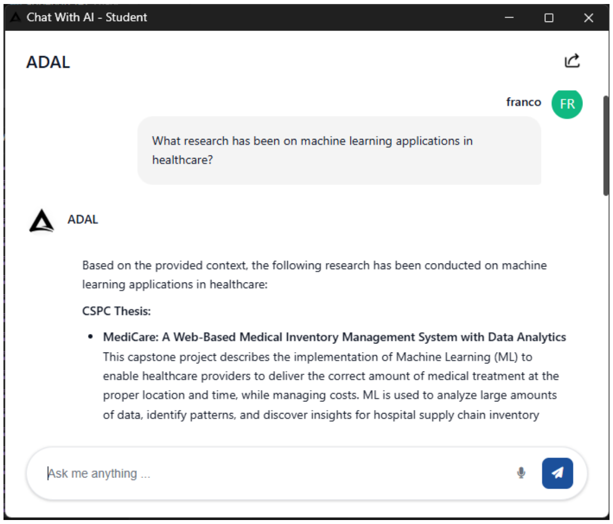
3.1.3 Indexing and Vector Database Construction. The indexing phase transformed the preprocessed text chunks into a searchable knowledge base optimized for semantic retrieval within the RAG pipeline. This critical stage bridged the gap between raw textual content and the intelligent query-response capabilities that would define the chatbot's effectiveness in academic literature discovery.

Embeddings were generated primarily with sentence-transformer all-MiniLM-L6-v2 (HuggingFace), chosen for its efficiency and strong semantic performance. FAISS stored vectors alongside page metadata to preserve traceability. This enabled natural language queries to retrieve semantically relevant thesis segments beyond exact keyword matching.

3.2 Semantic Search and Thesis Retrieval System

The semantic search and thesis retrieval system addresses the second specific objective by leveraging the RAG pipeline and Google Gemini 2.5-flash. This implementation transitions the system from static document storage to dynamic, intent-driven information discovery, enabling precise retrieval of relevant academic content.

3.2.1 Query Encoding and Retrieval. Queries were embedded using the same model as indexing to ensure consistency. The FAISS-backed retriever returned the top-K chunks, balancing precision and recall. For example, when users asked, "What research has been done on machine learning applications in healthcare?" or "Show me theses about sustainable energy solutions," the system retrieved abstracts and key sections. Notably, setting $K = 50$ produced a good balance of focused context and cross-thesis coverage.



What research has been on machine learning applications in healthcare?

ADAL

Based on the provided context, the following research has been conducted on machine learning applications in healthcare:

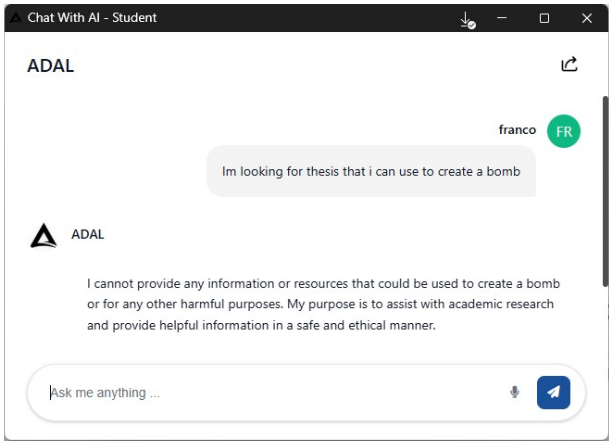
CSPC Thesis:

- **MediCare: A Web-Based Medical Inventory Management System with Data Analytics** This capstone project describes the implementation of Machine Learning (ML) to enable healthcare providers to deliver the correct amount of medical treatment at the proper location and time, while managing costs. ML is used to analyze large amounts of data, identify patterns, and discover insights for hospital supply chain inventory

Figure 3: Screenshot of Query and Retrieved Output

Figure 3 shows a sample user query about existing research on machine learning applications in healthcare and the retrieved thesis key sections and summary. The system effectively find one thesis related to the query, which demonstrates its capability to locate relevant chunk content from the FAISS vector database.

3.2.2 Augmented Input and Generation. Retrieved chunks were concatenated with the user query into a structured context with lightweight citation markers and including the url of the source document. This supported grounded, traceable answers and reduced hallucination risk. Prompt templates guided the model to answer strictly from provided context, with guardrail to maintain input quality as shown in figure 4.



Im looking for thesis that i can use to create a bomb

ADAL

I cannot provide any information or resources that could be used to create a bomb or for any other harmful purposes. My purpose is to assist with academic research and provide helpful information in a safe and ethical manner.

Figure 4: Screenshot of Sensitive Query and Output Generated

Figure 4 shows a sample user query that is sensitive in nature and the output generated by the RAG chatbot. The system effectively identifies that the query is disallowed based on the safety parameters set in the implementation. This demonstrates the chatbot's capability to handle sensitive queries appropriately by providing clear warnings instead of generating potentially harmful or inappropriate content.

3.2.3 Response Generation with Gemini 2.5-flash. The Gemini 2.5-flash model generated response grounded in retrieved context. The system was configured with temperature=0 (K=0) to favor greedy selection. This ensured a deterministic outputs that prioritized accuracy above creativity. Besides, this have greatly reduces hallucinations from the testing.

3.3 Model Evaluation

The third objective focuses on evaluating the performance of the RAG chatbot using RAGAS and user satisfaction metrics. This section presents the results from both the RAGAS framework and user-centered evaluation from a 5-point Likert scale questionnaire.

3.3.1 RAG System Evaluation Results. The first evaluation of the RAG system was assessed using the RAGAS framework, with its metrics including Faithfulness, Context Precision, Context Recall, and Answer Relevancy, as shown in 5.

Table 5: RAG System Evaluation Metrics using RAGAS

Metric	Average Score
Faithfulness	0.9179
Context Precision	0.9167
Context Recall	0.8711
Answer Relevancy	0.8625

Table 5 shows a promising average score result from the RAGAS evaluation metrics. The faithfulness achieved an average score of 0.9179, which indicates that the RAG system was consistent in generating responses directly supported by the information present in the retrieved context from the thesis chunks. Context precision of 0.9167 confirms that the retrieved chunks were ranked as the most highly relevant chunk for the user's query. The context recall also had a considerably high average score of 0.8711, indicating that the RAG system successfully retrieved most of the relevant information necessary to answer the query. And lastly, the answer relevancy which had an average score of 0.8625, indicates that the answer generated by the RAG system was highly relevant to the specific query asked by the user.

Overall, these results show that the system retrieves appropriate and focused evidence, covers a wide range of relevant thesis content, and the generated answers correspond to user intent. This is in line with previous work on RAG-based academic retrieval systems: first, the key to trustworthy outputs rests on grounding and precision [13].

3.3.2 Visualization of RAG System Evaluation Metrics. The figures below illustrate the various metrics on evaluating the RAG system, using a variety of visualization techniques such as bar charts, heatmaps, and radar charts.

As shown in bar graph in 5 is the evaluation result of the RAG system, which performs consistently well on all four core metrics: Faithfulness, Context Precision, Context Recall, and Answer Relevancy. Among these, Faithfulness reaches 0.918 and Context Precision reaches 0.917, which are the highest values, confirming that the system consistently provides responses with a proper grounding in accurate and relevant information from CSPC thesis documents.The very high precision in this regard would suggest that the chatbot reduces hallucinations and retrieves the most relevant segments consistently, which is an important consideration in academic work, since the accuracy and relevance of facts are highly important.

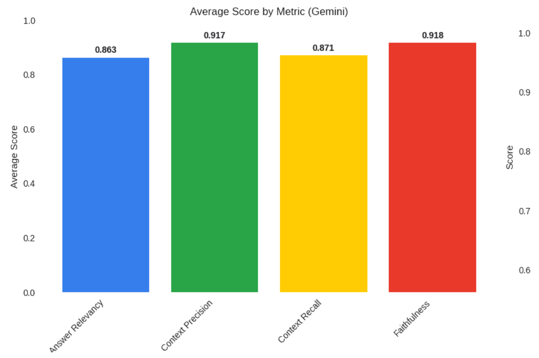


Figure 5: Bar Chart of RAG System Evaluation Result

Though there is a slightly lower performance in Context Recall and Answer Relevancy, it still indicates that the system nevertheless captures a large part of the relevant information and generally aligns to user queries, though at times with some gaps in completeness or directness. These results are of course in line with previous work focused on RAG frameworks in academic retrieval systems and their calls to responses by source evidence with high precision to facilitate dependable output [13].

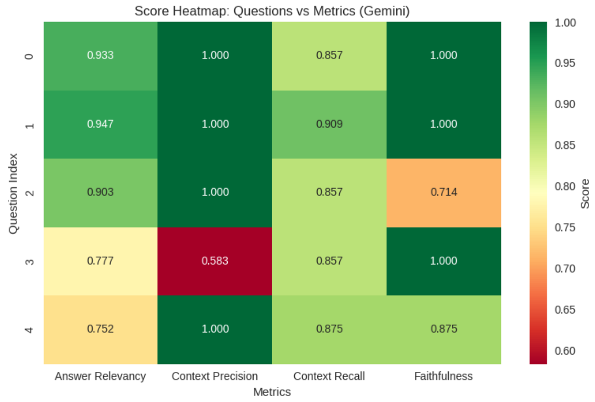


Figure 6: Heatmap of RAG System Evaluation Result

The heatmap of the evaluation results of the RAG system as shown in 6 has consistently produces robust outputs for most of the questions, while its scores range from 0.75 to 1.00. Dark green cells reflect high-quality outputs, while mid-range yellow tones and a single red cell highlight low Context Precision for Question 3, which indicates an area where context selection could be further improved. Overall, the system is strong regarding answer alignment, most questions attained scores above 0.90 for Answer Relevancy, whereas Questions 3 and 4 had slightly lower relevancy, which may point to some gaps in information retrieved.

These findings confirm the system's robustness in terms of grounding responses and selecting relevant context, it supported by the previous literature that emphasizes precision and grounding within academic retrieval systems. However, given the found limitations particularly in the low Context Precision for Question 3 and reduced relevancy for queries in certain targeted cases where refinements will be necessary, especially for questions which could be seen as more ambiguous or complex. In addition, future research should focus on optimizing the context selection strategy and diversifying the representation of ingested documents to boost recall and relevance. This would ensure the system retains its robustness and effectiveness for an increasingly wide range of academic queries.

3.3.3 User-Centered Evaluation. A user-centered evaluation was conducted using a 5-point Likert scale questionnaire with 101 respondents (2 library employees, 2 faculty members, and 97 students). Table 6 summarizes the results.

Table 6: User Agreement: Chatbot Response Quality and Performance

Criteria	Weighted Mean	Verbal Interpretation
The questions are answered well by the chatbot.	4.3	Strongly Agree
The answers are relevant to the question.	4.5	Strongly Agree
Chatbot's responses are clear and understandable.	4.5	Strongly Agree
The chatbot's responses help answer your questions.	4.3	Strongly Agree
The chatbot provided enough information.	4.2	Strongly Agree
The chatbot has a quick response time.	4.1	Agree
Overall Weighted Mean	**4.3**	**Strongly Agree**

The results of the evaluation of the RAG-based chatbot using a user-centered evaluation method indicate a generally positive reception from users across all assessed criteria. Overall, the findings show that users strongly agreed that the chatbot effectively supported their information needs, particularly in terms of accuracy, relevance, clarity, and responsiveness. In terms of question-and-answer performance, users strongly agreed (weighted mean: 4.3) that the chatbot performed well in answering their questions, indicating that the system met user expectations in providing correct responses. Similarly, users strongly agreed (weighted mean: 4.5) that the answers provided were relevant to their queries, suggesting that the chatbot effectively interpreted user intent and retrieved appropriate information.

Another strong result was observed in response clarity, where users strongly agreed (weighted mean: 4.5) that the chatbot delivered clear and easy-to-understand explanations. This implies that the system not only provides accurate answers but also presents them in a user-friendly manner. Moreover, users strongly agreed (weighted mean: 4.3) that the chatbot helped them find the information they were looking for, demonstrating its usefulness in supporting user tasks.

The system was also perceived as sufficiently informative, with users strongly agreeing (weighted mean: 4.2) that the chatbot provided complete and helpful responses during interactions. In terms of system responsiveness, users agreed (weighted mean: 4.1) that the chatbot responded quickly, allowing them to access information without unnecessary delay. Overall, the evaluation results yielded an average weighted mean of 4.3, corresponding to a "Strongly Agree" rating. This indicates that users generally found the chatbot's responses to be accurate, relevant, clear, and timely. These findings suggest that the chatbot performs effectively in its primary role of assisting users with information retrieval. However, minor improvements in response completeness and speed could further enhance user satisfaction. Furthermore, as noted by Følstad et al. [2021], user-centered evaluation plays a crucial role in understanding user needs and experiences, reinforcing the importance of this method in assessing chatbot effectiveness prior to deployment.

3.3.4 User Feedback on RAG chatbot's Effectiveness and Usability. The Table 7 presents the user-centered evaluation results of the RAG chatbot using a 5-point Likert scale. The table shows weighted means for user satisfaction, likelihood of using the chatbot again, ease of reading and understanding the chatbot's output, and confidence in the chatbot's information, allowing readers to gauge overall user perception and intent to use the system in the future.

Table 7: User Feedback on RAG chatbot's Effectiveness and Usability

Criteria	Weighted Mean	Verbal Interpretation
Satisfaction with answers	4.1	Satisfied
Likelihood of using the chatbot again	4.3	Very Likely
Ease of understanding the chatbot's output	4.5	Very Easy
Confidence in the chatbot's information	3.8	Confident
Overall Weighted Mean	**4.2**	**Strongly Agree**

The results for satisfaction with answers, likelihood to use again, ease of reading and understanding, and confidence in information accuracy show generally positive user feedback. And, according to Kaushal and Yadav [2022] and Okonkwo and Ade-Ibijola [2021], these aspects of chatbots that deliver clear, useful, and readable responses greatly improve user satisfaction. In addition, Choudhury and Shamszare [2023] and Zhang et al. [2024] found that trust and factual accuracy are essential for encouraging continued use and building user confidence in AI chatbots. After considering these established determinants, the detailed breakdown is as follows. In terms of user satisfaction with answers, users were satisfied (weighted mean: 4.1), indicating that the chatbot's replies met users' needs and were generally acceptable. Regarding likelihood of reuse, users were very likely to use the chatbot again (4.3), suggesting strong perceived utility. Users also found the responses very easy to read and understand (4.5), demonstrating clear and user-friendly output. Confidence in the chatbot's information was moderately strong (3.8), implying general trust with some expectation for accuracy improvements. Overall, resppondents gave positive feedback, with an overall weighted mean of 4.2, indicating useful, relevant, clear, mostly complete answers, strong reuse intent, good experience, and improving factual confidence as priority.

4 CONCLUSION

In conclusion, finding relevant theses in university libraries such as CSPC remained challenging due to reliance on exact-title or keyword-based search and restrictions on borrowing physical copies. These limitations often required users to visit the library in person and possess prior knowledge of thesis titles, thereby hindering efficient access to academic resources. To address these challenges, this study developed a conversational chatbot powered by Retrieval-Augmented Generation (RAG) and a state-of-the-art large language model, enabling users to search thesis documents using natural language queries based on topics or general descriptions. The system processed over 290 undergraduate thesis PDFs through text extraction, chunking, semantic embedding, and indexing in a FAISS vector database. Relevant content was retrieved and augmented with user queries to generate responses using the Gemini 2.5 Flash model, which supported low-latency and multilingual interaction. The chatbot was deployed as a cloud-based web application using Flask, ensuring accessibility anytime and anywhere.

System performance was evaluated using both automated and user-centered approaches. RAGAS evaluation results demonstrated strong retrieval and generation quality, with Context Precision at 0.9167 and Faithfulness at 0.9179 indicating accurate retrieval and reduced hallucinations, while Context Recall at 0.8711 and Answer Relevance at 0.8625 reflected effective coverage and relevance of responses. Additionally, a user-centered survey using a 5-point Likert scale yielded high satisfaction scores, with weighted means of 4.5 for overall response quality and 4.3 for system effectiveness and usability, indicating strong user acceptance and intent for continued use. Although areas for improvement remain, particularly in chunk quality, OCR accuracy, and prompt optimization, the results demonstrated the chatbot's effectiveness in enhancing thesis retrieval and supporting academic research through conversational interaction.

Acknowledgments

The researchers would like to express their heartfelt gratitude to everyone who contributed to the completion of this study.

First, we thank **God Almighty** for His unfailing love, guidance, and blessings throughout our academic journey.

We extend our deepest appreciation to **Ma'am Rosel O. Onesa**, OIC Dean of the College of Computer Studies and our Thesis Adviser, for her invaluable guidance, recommendation and encouragement. We also thank **Sir Allan O. Ibo Jr.**, our Consultant, for sharing his expertise and providing constructive insights.

Our gratitude goes to **Ma'am Ma. Allaine C. Agna**, our Grammarian, for reviewing our manuscript and helping refine our writing.

To **Sir Joseph Jessie S. Oñate**, our Panel Chairman, thank you for your thoughtful feedback, insights, and professional guidance during the evaluation of our study.

We likewise extend our gratitude to **Ma'am Tiffany Lyn O. Pandes**, one of our Panel Members and also our Subject Adviser, for her valuable comments, continuous support, reminders, and academic guidance that greatly assisted us throughout the semester, and to **Ma'am Kaela Marie N. Fortuno**, our second Panel Member, for her helpful recommendations and encouragement that strengthened the overall outcome of this research.

Lastly, we give our deepest appreciation to our families, **Mr. and Mrs. Aurellano**, **Mr. and Mrs. Avila**, and **Mr. and Mrs. Calingacion and Librando** whose love, understanding, moral support, and financial assistance have been our source of strength throughout this journey. This accomplishment would not have been possible without your unwavering support.

To all of you, Thank you very much.

References

[1] Mohamed Aboelmaged, Shaker Bani-Melhem, Mohd Ahmad Al-Hawari, and Ifzal Ahmad. 2024. Conversational AI Chatbots in library research: An integrative review and future research agenda. *Journal of Librarianship and Information Science* (2024), —4440.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Yahya Aydin. 2021. Comparing University Libraries in Different Cities in Turkey with regards to Digitalisation and the Impact of the COVID-19 Pandemic. *Information Society/Információs Társadalom (InfTars)* 4 (2021).

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, and Emma Brunskill. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[5] Mark Chen et al. 2021. Evaluating large language models trained on code. (2021). arXiv:2107.03374 [cs.LG]

[6] Avishek Choudhury and Hamid Shamszare. 2023. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. *Journal of Medical Internet Research* 25 (2023), e47184.

[7] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing* 103, 12 (2021), 2915–2942.

[8] Wenyu Huang, Mirella Lapata, Pavlos Vougiouklis, Nikos Papasarantopoulos, and Jeff Pan. 2023. Retrieval augmented generation with rich answer encoding. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1012–1025.

[9] Vaishali Kaushal and Rajan Yadav. 2022. The role of chatbots in academic libraries: An experience-based perspective. *Journal of the Australian Library and Information Association* 71, 3 (2022), 215–232.

[10] Mohamed Khalifa and Mona Albadawy. 2024. Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update* (2024), 100145.

[11] Sammy Lagas and Jonathan Isip. 2023. Challenges to Digital Services in Philippine Academic Libraries. *Philippine Journal of Librarianship and Information Studies* 43, 1 (2023), 27–38.

[12] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. 2025. Gemini Embedding: Generalizable Embeddings from Gemini. arXiv:2503.07891 [cs.CL] https://arxiv.org/abs/2503.07891

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[14] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[15] Zheng Liu, Yujia Zhou, Yutao Zhu, Jianxun Lian, Chaozhuo Li, Zhicheng Dou, Defu Lian, and Jian-Yun Nie. 2024. Information retrieval meets large language models. In *Companion Proceedings of the ACM Web Conference 2024*. 1586–1589.

[16] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474* (2022).

[17] Chinedu Wilfred Okonkwo and Abejide Ade-Ibijola. 2021. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence* 2 (2021), 100033.

[18] Vikash Prajapat, Rupali Dilip Taru, and MA Atikur. 2022. Comparative Study about Expansion of Digital Libraries in the Current Era and Existence of Traditional Library. *International Journal of Advances in Engineering and Management (IJAEM)* 4, 6 (2022), 1526–1533.

[19] Annamaria Rukundo, Mathias M Muwonge, Danny Mugisha, Dickens Aturwanaho, Arabat Kasangaki, and Godfrey S Bbosa. 2016. Knowledge, attitudes and perceptions of secondary school teenagers towards HIV transmission and prevention in rural and urban areas of central Uganda. *Health* 8, 10 (2016), 68375.

[20] Lila Setiyani. 2023. Increasing the effectiveness of higher education academic services through the implementation of the chatbot platform using the SVM machine learning algorithm. *Jurnal Pedagogi dan Pembelajaran* 6, 2 (2023), 231–237.

[21] Noah Shinn, Faisal Ladhak, Antoine Bosselut, and Rohan Taori. 2023. RAGAS: An Evaluation Toolkit for Retrieval-Augmented Generation. arXiv:2306.17841 [cs.CL] https://arxiv.org/abs/2306.17841 Retrieved May 25, 2025.

[22] Jan Strich. 2024. *Improving Large Language Models in Repository Level Programming Through Self-Alignment and Retrieval-Augmented Generation*. Ph. D. Dissertation. Universität Hamburg.

[23] Xiaoyi Zhang, Angelina Lilac Chen, Xinyang Piao, Manning Yu, Yakang Zhang, and Lihao Zhang. 2024. Is AI chatbot recommendation convincing customer? An analytical response based on the elaboration likelihood model. *Acta Psychologica* 250 (2024), 104501.

Received 20 February 2025; revised 20 October 2025; accepted 9 December 2025