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, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large Language Models (LLMs) such as GPT [1] and Gemini [19] have significantly advanced Natural Language Processing by enabling context-aware tasks like semantic search and classification, outperforming traditional keyword-based systems [7, 25]. However, LLMs rely solely on pre-trained data and lack access to real-time or localized information, limiting their effectiveness for Information Retrieval (IR) tasks in university libraries [23]. In the Philippines, many academic libraries, including the Camarines Sur Polytechnic Colleges (CSPC), still rely on traditional or basic digital archives with exact keyword search, making it difficult for students and researchers to retrieve relevant theses, especially when exact titles or keywords are unknown [28, 31]. Although digital archiving has improved access to academic resources, particularly during the COVID-19 pandemic, outdated search mechanisms remain a major limitation [6, 18]. To address these challenges, Retrieval-Augmented Generation (RAG) has emerged as an effective approach that enhances LLMs by enabling them to retrieve and utilize external, domain-specific documents without retraining [14, 20]. This study developed an LLM-powered chatbot integrated with a RAG framework to improve thesis retrieval and information access for the CSPC Library.

The specific objectives of this study are: (1) to design and implement a document ingestion and retrieval module that processes and indexes over 290 thesis PDF documents using semantic embeddings and a FAISS vector database; (2) to develop a semantic search system using Retrieval-Augmented Generation and Google Gemini 2.5-Flash to support natural language queries for thesis retrieval; and (3) to evaluate the performance of the RAG-based chatbot using the RAGAS framework and user-centered assessment metrics to determine improvements in literature accessibility, search efficiency, and user satisfaction compared to traditional library search systems. This study is significant as it demonstrates how Retrieval-Augmented Generation (RAG) can be effectively integrated into university library systems to improve thesis retrieval and information accessibility. By enabling semantic and conversational search over institution-specific academic archives, the proposed system addresses long-standing limitations of keyword-based retrieval and enhances research efficiency. Furthermore, the findings provide practical insights for academic institutions seeking to adopt AI-driven solutions to modernize digital library services.

2 BACKGROUND

Large Language Models (LLMs) have improved academic information retrieval through semantic search, question answering, and document summarization, thereby increasing research efficiency [36, 37]. However, their dependence on pre-trained knowledge and limited access to domain-specific or up-to-date resources constrain their effectiveness in specialized academic contexts [11, 17]. Retrieval-Augmented Generation (RAG) mitigates these limitations by integrating external knowledge sources with LLMs, resulting in improved factual accuracy, contextual relevance, and reduced hallucinations [20, 33].

Recent studies demonstrate that RAG frameworks, when combined with semantic embeddings and vector databases, significantly enhance retrieval accuracy and efficiency across academic, healthcare, and legal domains [4, 12, 15]. In academic settings, RAG-based systems improve scholarly literature retrieval and question-answering performance, while domain-specific implementations in healthcare and legal research highlight gains in information reliability and contextual accuracy [3, 5]. Advances in large language models further reinforce the suitability of RAG architectures for context-sensitive environments such as academic libraries [27].

Hypertext '26, June 2025, CSPC, Philippines

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

Evaluating RAG systems requires specialized frameworks beyond traditional LLM assessment methods. The RAGAS framework is widely adopted to measure context precision, context recall, faithfulness, and response relevance, emphasizing the alignment between retrieved documents and generated outputs [29]. While automated metrics provide valuable insights into system performance, prior studies emphasize the continued importance of human evaluation to capture qualitative factors such as clarity, consistency, and user satisfaction [33, 34]. Despite demonstrated effectiveness across multiple domains, limited research has focused on RAG implementations in academic library environments, particularly for literature search and thesis retrieval. This study addresses this gap by developing a RAG-based chatbot tailored to the CSPC Library.

3 METHODOLOGY

This study employed a constructive research design to develop and evaluate a Retrieval-Augmented Generation (RAG)-based chatbot system integrated with a Large Language Model (LLM). The system aimed to enhance thesis literature retrieval within the CSPC Library by replacing traditional keyword-based search with a vector database and conversational framework. The chatbot was deployed to the cloud for accessibility.

3.1 Theorems, Algorithms, and Mathematical Models

The dataset comprised 290+ undergraduate thesis PDFs sourced from multiple departments within the CSPC library. These documents were provided by library personnel under agreed-upon data handling protocols. The dataset included theses from various academic disciplines, ensuring a diverse and comprehensive knowledge base.

3.1.1 RAG Pipeline. The Retrieval-Augmented Generation (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.

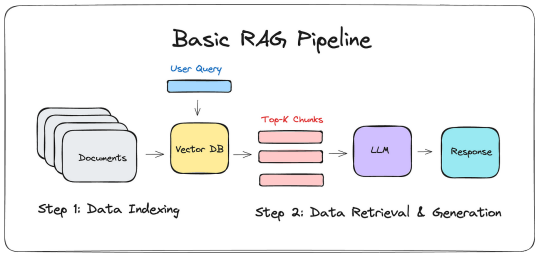

Figure 1: Basic RAG Pipeline by Dr. Julija

The chatbot's RAG pipeline, as illustrated in 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.
- **C. Large Language Model and Gemini 2.5-flash:** LLMs excel in NLP tasks including summarization, question answering, and retrieval [24, 35]. This project integrates Gemini 2.5 Flash [9], which offers advanced reasoning at low latency and cost, with multimodal support and extended context windows, enhancing thesis retrieval for CSPC library users.

3.2 Materials and Statistical Tools / Evaluation Methods

In this subsection, we describe the materials, tools, and evaluation methods used in this study to assess the performance of the RAG-based chatbot system.

3.2.1 Research Materials. This section includes the dataset that was used, as well as the minimum hardware and software needed for the development of the system.

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 and tested on a machine with the following specifications:

- **Processor:** Intel Core i7-10750H CPU @ 2.60GHz
- **RAM:** 16GB DDR4
- **Storage:** 512GB SSD
- **Operating System:** Windows 10 Pro 64-bit
- **Software:** Python 3.8, Flask, PyMuPDF, FAISS, Hugging Face Transformers, Google Gemini 2.5-flash API

3.2.2 Instruments. In this subsection, the instruments that was used by researchers to analyze and evaluate the performance of the RAG chatbot system.

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 [32]. Furthermore, a context recall metric was included, as recommended for evaluating retrieved chunks.

Survey. Instruments served as data collection tools across different areas and provided an effective way to gather information. They were useful when seeking insights into the attributes, preferences, opinions, or beliefs of a specific group. To meet the study objectives, 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, the researchers created a questionnaire 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.

3.2.3 Statistical Test. The RAG system performance was evaluated using the RAGAS framework, focusing on context precision, recall, relevance, and faithfulness to measure how well relevant documents were retrieved and responses generated [2, 13, 22]. Additionally, to assess not only the technical but also the user-centered performance of the system, a user questionnaire was administered to collect feedback regarding usability, accuracy, and overall satisfaction.

The Likert Scale, introduced by Likert [1932], is a measurement method developed for evaluating individuals' attitudes toward any object. 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 [21, 30].

Table 1: 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

User responses from the Likert scale questionnaire were analyzed using the Weighted Mean, a standard statistical method for synthesizing ordinal survey data in human-computer interaction and user experience research. This quantitative approach provides a rigorous measure of aggregate user satisfaction and system performance by computing the arithmetic mean of individual ratings across all respondents. The weighted mean calculation was employed to systematically evaluate user perceptions across four dimensions: satisfaction with response quality, likelihood of continued system adoption, intelligibility of chatbot output, and confidence in information accuracy. This methodology aligns with established conventions in usability research and ensures reproducibility of findings across comparable studies [21].

$$WM = \frac{TWM}{N} \tag{1}$$

Where WM is the Weighted Mean, TWM is the Total Weighted Mean (sum of all individual scores), and N is the total number of respondents. Higher WM values indicate greater user satisfaction and system effectiveness.

3.3 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.

3.4 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:

3.4.1 Context Precision. Measured the relevance of retrieved chunks.

$$\text{Precision@k} = \frac{\text{true positives@k}}{\text{true positives@k} + \text{false positives@k}} \tag{2}$$

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.

3.4.2 Context Recall. Assessed the comprehensiveness of retrieval.

$$\text{Context Recall} = \frac{\text{Number of relevant claims supported by retrieved chunks}}{\text{Total number of relevant claims in the reference answer}} \tag{3}$$

where:

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

3.4.3 Response Relevance. Evaluated alignment between user queries and generated responses.

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

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.

3.4.4 Faithfulness. Ensured factual consistency with retrieved context.

$$\text{Faithfulness} = \frac{\text{Number of claims in the response supported by retrieved context}}{\text{Total number of claims in the response}} \tag{5}$$

where:

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

3.5 Conceptual Framework

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

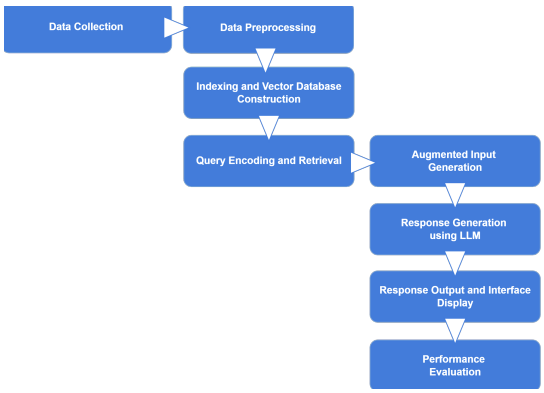

Figure 2: Conceptual Framework

The proposed conceptual framework implemented a Retrieval-Augmented Generation (RAG) pipeline for semantic 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 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 assessing retrieval accuracy, response quality, and overall user satisfaction.

4 RESULTS AND DISCUSSION

This chapter discusses the results and evaluation of the RAG chatbot developed for efficient literature search and thesis retrieval at the CSPC Library.

4.1 Document Ingestion and Retrieval Module

The study processed 290+ undergraduate thesis PDFs from the CSPC Library into a dynamic, searchable knowledge base. Texts were extracted, enriched with metadata, and segmented into 38,127 chunks equating to 11,849,783 tokens, with each chunk containing an average of 1,323 characters or approximately 180 words, and a median of 335 tokens per chunk that show in Table 2. These chunks were embedded using 'sentence-transformers/all-MiniLM-L6-v2' and indexed in FAISS for efficient semantic retrieval. This process ensured compatibility with the RAG pipeline and improved retrieval fidelity. The chunking strategy and semantic indexing played a critical role in ensuring the fidelity and transparency of the retrieved information.

Table 2: 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

4.2 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.

4.2.1 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.

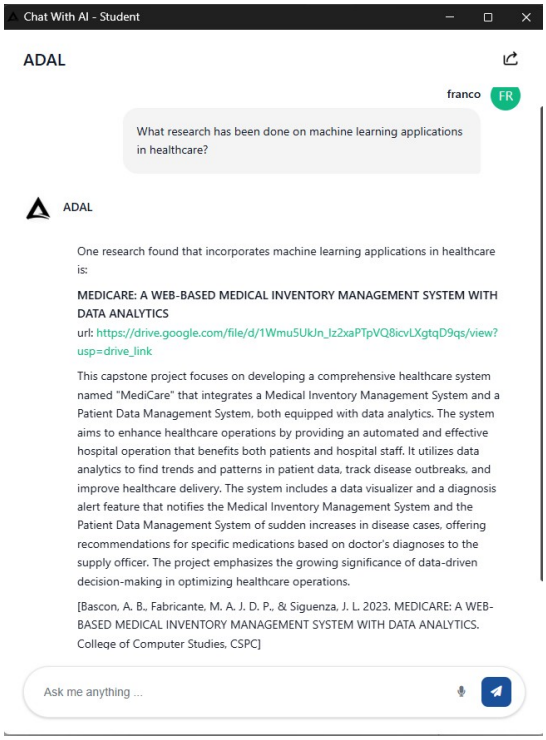

Figure 3: Screenshot of Query and Retrieved Output

In 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.

Moreover, the output also includes the created citations that shows the author's last name and the year of publication in the CSPC. Notably, this retrieved data are all chunks that got from the top-K retrieval process.

4.2.2 Augmented Input and Generation. Retrieved chunks were concatenated with the user query into a structured context with lightweight citation markers. This supported grounded, traceable answers and reduced hallucination risk. Prompt templates guided the model to answer strictly from provided context, with safeguards (token monitoring, truncation) to maintain input quality.

4.2.3 Response Generation with Gemini 2.5-flash. The Gemini 2.5-flash model provided responses based on the retrieved context. Temperature = 0 was used as input to ensure deterministic outputs suitable for scholarly use case. Clean text was generated by parsing the output. RAG substantially reduces hallucinations, but inaccuracies sometimes happened if not enough context was available, therefore, users are encouraged to double-check critical conclusions.

In Figure 4, the system was deployed to the cloud to ensure accessibility for users across various locations. The interface supports conversational exploration with session-based history and safety filters for disallowed queries. Generated responses appear as Markdown with citations and structured text. When queries violate safety parameters, clear warnings are displayed. Deterministic settings improve consistency and build user trust.

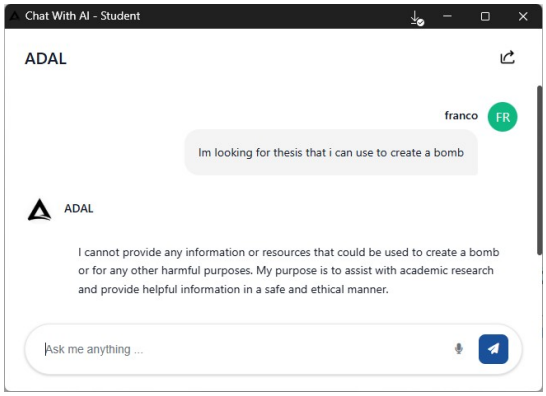

Figure 4: User Interface

4.3 Model Evaluation

This section reveals the evaluation result of the RAG chatbot using the four metrics from RAGAS including Answer Relevancy, Context Precision, Context Recall, and Faithfulness. Each metrics captures the system accuracy, coverage, and grounding of responses. Another method applied for evaluation is the user-centered evaluation specifically using a 5-point Likert scale questionnaire.

4.3.1 RAG System Evaluation Results. The RAG system's effectiveness at retrieving and generating accurate, relevant and well grounded answers grounded in the indexed thesis documents from the CSPC Library is shown by the evaluation outcomes provided through the four core metrics of the RAGAS framework summarized in the following 3.

Table 3: RAG System Evaluation Metrics using RAGAS Framework

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

The 3 shows how the high result of the RAG system performance in terms of Faithfulness, Context Precision (0.9167), Context Recall (0.8711), and Answer Relevancy (0.8625), where all these result point to a reliable and well-grounded responses.

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 [20].

The results indicated that the system could be used as a useful academic support tool, which would improve the literature search to find theses for students and researchers. For some reason, the current evaluation result is not considered the absolute reflection of the system performance, which might affect generalizability. Future work will further compare the performance on a broader class of queries, including diverse or ambiguous ones, against alternative retrieval models in order to validate and enhance this system.

4.3.2 Visualization of RAG System Evaluation Metrics. The bar chart in Figure 5 illustrates the performance of the RAG system across all four evaluation metrics. Faithfulness achieved the highest score at 0.9179, indicating that the system's responses are strongly grounded in retrieved context, with minimal hallucinations or unsupported claims. Context Precision followed closely at 0.9167, demonstrating that retrieved chunks are highly relevant to user queries, minimizing noise in the retrieval results. Context Recall scored 0.8711, reflecting comprehensive coverage of relevant thesis segments necessary to answer user questions. Answer Relevancy, at 0.8625, shows strong alignment between generated responses and original user queries, confirming that the system effectively interprets user intent and provides pertinent information.

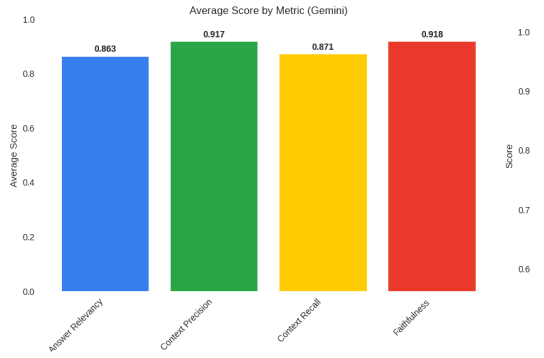

Figure 5: Bar Chart of RAG System Evaluation Result

The consistency of high scores across all metrics validates the RAG system's robustness in delivering accurate, relevant, and well-grounded responses. This balanced performance profile is particularly important for academic applications, where factual accuracy and comprehensive coverage are critical. The slightly lower Context Recall score suggests minor gaps in retrieval completeness, which could be addressed through optimization of chunk size, retrieval parameters, or refinement of the embedding model. Nevertheless, these results demonstrate that the RAG-augmented chatbot is reliable for thesis discovery and academic information retrieval within the CSPC Library context.

4.3.3 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 4 summarizes the results.

Table 4: 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.

4.3.4 User Feedback on RAG chatbot's Effectiveness and Usability. The Table 5 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 5: 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.

5 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] 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).

[2] Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. 2024. A Statistical Framework for Ranking LLM-Based Chatbots. https://doi.org/10.48550/arXiv.2412.18407 arXiv:2412.18407 [cs.CL]

[3] L. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel. 2024. Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review. https://doi.org/10.20944/preprints202407.0876.v1

[4] I. Aquino, M. Santos, C. Dorneles, and J. Carvalho. 2024. Extracting Information from Brazilian Legal Documents with Retrieval Augmented Generation. In *SBBD Estendido*. 280–287. https://doi.org/10.5753/sbbd_estendido.2024.244241

[5] K. Arzideh, H. Schäfer, A. Idrissi-Yaghi, B. Eryılmaz, M. Bahn, C. Schmidt, and R. Hosch. 2024. MIRACLE - Medical Information Retrieval Using Clinical Language Embeddings for Retrieval Augmented Generation at the Point of Care. *Research Square* (2024). https://doi.org/10.21203/rs.3.rs-5453999/v1

[6] 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).

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

[8] 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.

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025). https://arxiv.org/abs/2507.06261

[10] 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.

[11] Gerald Gartlehner, Laura Kahwati, Roxanne Hilscher, Ivan Thomas, Susan Kugley, Kristina Crotty, and Rebecca Chew. 2023. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. (2023). https://doi.org/10.1101/2023.10.02.23296415 Preprint.

[12] A. Grigoryan and H. Madoyan. 2024. Building a Retrieval-Augmented Generation (RAG) System for Academic Papers.

[13] Samuel Holmes, Raymond R. Bond, Anne Moorhead, Vivien Coates, and Michael F. McTear. 2023. Towards Validating a Chatbot Usability Scale. In *Human Interface and the Management of Information*. 321–339. https://doi.org/10.1007/978-3-031-35708-4_24

[14] 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.

[15] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, and W. Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of EMNLP 2020*. https://doi.org/10.18653/v1/2020.emnlp-main.550

[16] 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.

[17] Qais Khraisha, Stefaan Put, Jens Kappenberg, Adeel Warraitch, and Kaitlyn Hadfield. 2024. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods* 15, 4 (2024), 616–626. https://doi.org/10.1002/jrsm.1715

[18] 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.

[19] 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

[20] 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.

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

[22] Ying-Chun Lin, Jennifer Neville, Jack W. Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and J. Teevan. 2024. Interpretable User Satisfaction Estimation for Conversational Systems with Large Language Models. https://doi.org/10.48550/arXiv.2403.12388 arXiv:2403.12388 [cs.CL]

[23] 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.

[24] Muhammad Naveed et al. 2024. Large Language Models and Their Impact on NLP Tasks. *Journal of Natural Language Processing Research* (2024).

[25] 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).

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

[27] Arjun Prabhulal. 2025. Build a RAG Pipeline with Gemini Embeddings and Vector Search – A Deep Dive (Full Code). https://medium.com/google-cloud/build-a-rag-pipeline-with-gemini-embeddings-and-vector-search-a-deep-dive-full-code-bcd521ad9e1c. Accessed: January 16, 2026.

[28] 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.

[29] S. Roychowdhury, S. Soman, H.G. Ranjani, N. Gunda, V. Chhabra, and S.K. Bala. 2024. Evaluation of RAG Metrics for Question Answering in the Telecom Domain. arXiv:2407.12873 [cs.CL] https://arxiv.org/abs/2407.12873

[30] 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.

[31] 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.

[32] 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.

[33] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. arXiv:2104.07567 [cs.CL] https://arxiv.org/abs/2104.07567

[34] S. Sivasothy, S. Barnett, S. Kurniawan, Z. Rasool, and R. Vasa. 2024. RAGProbe: An Automated Approach for Evaluating RAG Applications. arXiv:2409.19019 [cs.CL] https://arxiv.org/abs/2409.19019

[35] Blaise Agüera y Arcas. 2022. Do Large Language Models Understand Us? *Daedalus* 151, 2 (2022), 183–197. https://doi.org/10.1162/daed_a_01909

[36] Anirudh Yalamanchili, Bhavya Sengupta, Ji Song, Stephanie Lim, Trevor Thomas, Bhavesh Mittal, and Peter Teo. 2024. Quality of large language model responses to radiation oncology patient care questions. *JAMA Network Open* 7, 4 (2024), e244630. https://doi.org/10.1001/jamanetworkopen.2024.4630

[37] Ruicong Yang, Tianyu Tan, Wenhao Lu, Arun Thirunavukarasu, Daniel Ting, and Nan Liu. 2023. Large language models in health care: development, applications, and challenges. *Health Care Science* 2, 4 (2023), 255–263. https://doi.org/10.1002/hcs2.61

[38] 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