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

1 INTRODUCTION

The rapid advancements in Natural Language Processing (NLP) have revolutionized information retrieval, enabling more efficient and context-aware systems. Large Language Models (LLMs), such as OpenAI's ChatGPT and Google Gemini, have demonstrated remarkable capabilities in semantic understanding and generation. However, LLMs face inherent limitations in accessing real-time or domain-specific institutional data, which restricts their effectiveness in specialized tasks such as academic literature retrieval. Retrieval-Augmented Generation (RAG) addresses this limitation by integrating LLMs with external knowledge sources, enabling factually grounded responses. Studies by [13] and [14] demonstrate that RAG-augmented LLMs significantly improve accuracy and coherence in knowledge-intensive tasks. Furthermore, [7] highlighted vector databases for semantic indexing, which enhance retrieval efficiency, a capability essential for thesis discovery in academic libraries.

University libraries, including the CSPC Library, maintain extensive collections of undergraduate theses that serve as valuable academic resources. However, these institutions typically rely on traditional, non-digital search systems with restrictive access policies. Physical theses cannot be removed from library premises, and keyword-based catalogs fail to capture semantic intent, resulting in incomplete or irrelevant search results. This limitation hinders student and researcher access to critical academic materials, impeding research progress and academic development.

This study addresses these challenges by developing a Retrieval-Augmented Generation (RAG)-based chatbot integrated with a Large Language Model to enable efficient, semantic-driven thesis discovery. The primary objectives are: (1) to design and implement a RAG pipeline that processes and indexes 290+ thesis PDFs using semantic embeddings and a FAISS vector database; (2) to develop a conversational interface enabling natural language queries for thesis retrieval; (3) to evaluate system performance using the RAGAS framework and user-centered assessment; and (4) to demonstrate that AI-driven retrieval systems significantly improve literature accessibility, search efficiency, and user satisfaction compared to traditional library systems. This research advances information retrieval in academic settings, providing students and researchers with a modern tool for thesis discovery and supporting institutional knowledge management.

2 BACKGROUND

Retrieval-Augmented Generation (RAG) is a technique that combines information retrieval with natural language generation to enhance the capabilities of Large Language Models. RAG systems retrieve relevant external documents or data during inference and use this context to ground the model's responses, thereby reducing hallucinations and improving factual accuracy. This approach is particularly valuable in knowledge-intensive tasks where up-to-date or domain-specific information is required. Large Language Models (LLMs) with integrated RAG techniques have greatly improved knowledge-intensive NLP tasks, overcoming traditional LLM limitations. Studies by [13] and [14] demonstrate how combining RAG with LLMs significantly improves accuracy and coherence in conversations and complex queries. Furthermore, [7] highlighted the use of vector databases for continuous information adaptation integrated with RAG, greatly enhancing retrieval efficiency and relevancy of LLM outputs, which is essential for literature search and thesis retrieval in university libraries. Vector databases such as FAISS (Facebook AI Similarity Search) are specialized data structures designed to store and retrieve high-dimensional vector embeddings efficiently, enabling semantic similarity-based retrieval that

Hypertext '26, June 2025, CSPC, Philippines

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

captures the meaning and intent of queries, resulting in more contextually relevant results compared to traditional keyword-based search.

The application of RAG in various domains has been extensively explored and demonstrates significant potential for transforming information retrieval across diverse fields. For instance, Arzideh et al. 2024 incorporated clinical language embeddings within RAG to improve healthcare information retrieval, while Grigoryan and Madoyan 2024 presented a system that enhances academic retrieval using vector search. Additionally, Aquino et al. 2024 employed RAG for effectively extracting and analyzing Brazilian legal documents, and Ryu et al. 2023 validated RAG's effectiveness in legal question-answering tasks. Evaluation metrics are crucial for assessing RAG system performance, with the RAGAS (RAG Assessment) framework providing automated metrics including Context Precision, Context Recall, Answer Relevancy, and Faithfulness, as emphasized by [11] and [2]. Despite the effectiveness of automated metrics, human evaluation remains important in assessing coherence and user satisfaction [1], providing complementary insights into system usability and effectiveness.

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 Data Sources

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.

3.2 Instruments

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

RAGAS (Retrieval-Augmented Generation Assessment Suite). toolkit was utilized to automatically evaluate the quality of system outputs using metrics such as context precision, faithfulness, and answer relevance [12]. Furthermore, a context recall metric was included, as recommended for evaluating retrieved chunks. These instruments ensured a rigorous and balanced evaluation of the proposed system from both system-level and user perspectives [8].

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 employed librarians and CSPC 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 developed questionnaires 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. The respondents of the study were all from the CSPC including 2 employees of Library, 2 faculty, and 14 students who served as a representative of the whole population.

3.2.1 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.

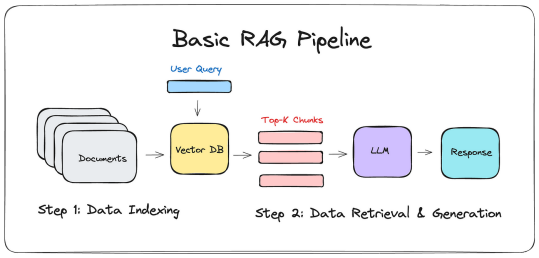

Figure 1: Basic RAG Pipeline by Dr. Julija

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

- **Document Ingestion and Preprocessing:** Thesis PDFs are ingested, and their texts are extracted, enriched with metadata, and segmented into coherent chunks using token-based chunking that respects academic structure (Abstract; Chapters 1–5) to preserve semantic context.
- **Data Indexing:** Text chunks are converted into dense vectors using 'sentence-transformers/all-MiniLM-L6-v2', chosen for its lightweight architecture and strong semantic representation capabilities. These vectors and metadata are indexed in FAISS for efficient similarity search and scalable retrieval over the thesis corpus.
- **Query Processing and Retrieval:** User queries are embedded using the same model employed for indexing. A FAISS-backed retriever performs semantic search to return the top-K most relevant chunks (default $K = 50$), balancing precision and recall based on semantic similarity.
- **Response Generation with LLM:** Retrieved chunks are provided to the Gemini 2.5-flash language model as grounded context. This advanced LLM, part of the Gemini 2.X family, generates factually accurate and relevant responses aligned with source documents, leveraging its optimized architecture for low latency and domain-specific retrieval.
- **User Interface and Deployment:** The system is deployed to the cloud with a web interface for user interaction, displaying responses along with metadata for transparency and traceability.

3.3 Procedures

The study followed a structured approach to develop and deploy the RAG-based chatbot system. First, data preprocessing was conducted by extracting texts from thesis PDFs on a page-by-page basis and enriching them with metadata, such as source and page numbers. Token-based chunking was then applied to segment the texts into coherent pieces aligned with thesis sections (e.g., Abstract, Chapters 1–5), ensuring compatibility with the LLM's context window and improving retrieval fidelity. A sample of the CSPC thesis PDFs used in this process is shown in Figure 2.

Next, indexing and vector database construction involved embedding text chunks into dense vectors using sentence-transformers/all-MiniLM-L6-v2 and indexing them in FAISS for semantic search. During query handling, user queries were encoded and matched against the vector database to retrieve the top-K relevant chunks based on semantic similarity. The Gemini 2.5-flash model then generated responses grounded in the retrieved context, ensuring factual accuracy and relevance. These responses were displayed via a web interface with metadata for transparency. The system's performance was assessed using both automated metrics from the

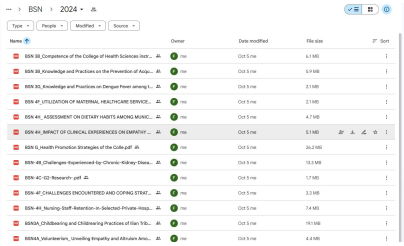

Figure 2: CSPC Thesis PDF Sample

RAGAS framework and a user questionnaire with a 5-point Likert scale to evaluate human-centered performance.

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.

$$\text{Precision@k} = \frac{\text{true positives@k}}{\text{true positives@k} + \text{false positives@k}} \tag{1}$$

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.

$$\text{Context Recall} = \frac{\text{Number of relevant claims supported by retrieved chunks}}{\text{Total number of relevant claims in the reference answer}} \tag{2}$$

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.

$$\text{Response Relevance} = \frac{1}{N}\sum_{i=1}^{N}\cos(E_{g_i}, E_o) \tag{3}$$

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.

$$\text{Faithfulness} = \frac{\text{Number of claims in the response supported by retrieved context}}{\text{Total number of claims in the response}} \tag{4}$$

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 3.

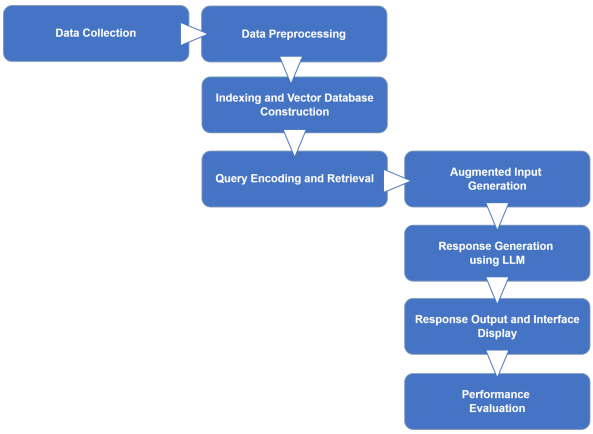

Figure 3: Conceptual Framework

3.5.1 Data Collection. The process began with coordination between the researchers and CSPC Library staff, where the prototype was demonstrated to illustrate how a RAG-powered chatbot could improve thesis discovery beyond exact-keyword search by enabling topic-oriented, semantically grounded retrieval within the library's own repository. In the demonstration, the project's institutional value was emphasized in accelerating literature searches, increasing access to relevant local theses, and supporting academic guidance. Following this formal presentation, the researchers submitted a request to obtain all undergraduate thesis PDFs from various College departments of CSPC to use as the main corpus of the RAG chatbot application.

3.5.2 Data Pre-processing. Tools such as PyMuPDF were employed to extract plain text from the collected PDFs. The extracted content underwent cleaning and normalization to remove non-informative characters, followed by segmentation into semantically meaningful text chunks that preserved the structural integrity of academic documents.

3.5.3 Indexing and Vector Database Construction. Each text chunk was embedded using the 'sentence-transformers/all-MiniLM-L6-v2' model from Hugging Face, converting semantic meanings into dense vectors that capture both explicit and implicit relationships across CSPC thesis documents. These vectors were indexed in FAISS, enabling fast, context-aware retrieval of relevant content through natural language queries. Metadata was preserved for each vector, maintaining links to source documents and positions, thereby supporting semantic discovery of academic literature and surpassing traditional keyword matching approaches.

3.5.4 Query Encoding and Retrieval. User queries were encoded into dense vectors using the same embedding model as employed for document indexing, ensuring semantic alignment. The system then performed fast similarity searches in FAISS, retrieving the top-K relevant thesis chunks based on conceptual match rather than keyword overlap. This process allowed contextually accurate results even for varied terminology, forming the basis for the chatbot's informed, thesis-grounded responses.

3.5.5 Augmented Input Generation. The augmented input generation phase served as the crucial bridge between retrieved thesis content and intelligent response formulation, where raw document chunks evolved into contextually enriched prompts capable of guiding accurate academic discourse.

3.5.6 Response Generation. The response generation stage represented the culmination of the RAG pipeline, where the Gemini 2.5-flash model transformed augmented academic context into coherent, factually grounded answers that addressed user research inquiries with precision and relevance.

3.5.7 Response Output and Interface Display. Flask was used to create an intuitive web interface that presented the chatbot's responses alongside relevant metadata, such as source thesis titles and sections, ensuring transparency and traceability of retrieved information.

3.5.8 Performance Evaluation. The system's effectiveness was evaluated using two complementary approaches: automated RAGAS metrics (context precision, context recall, answer relevance, and faithfulness) and a user survey using Likert scales. Findings from both evaluation methods guided refinements to retrieval strategies, prompting techniques, and user interface design. This dual-evaluation approach ensured technical robustness, usability, and trustworthy, thesis-grounded answers aligned with CSPC academic needs for library users and researchers.

4 RESULTS AND DISCUSSION

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 coherent chunks, averaging 311 tokens per chunk. 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.

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.

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 generated answers grounded in retrieved context. The system was configured with temperature=0 to ensure deterministic outputs suitable for academic use. Generated content was parsed into clean text for display. While RAG significantly reduced hallucinations, occasional inaccuracies were observed when context was insufficient; users were advised to validate critical findings. The Flask-based web interface facilitated user interaction, providing clear and traceable responses.

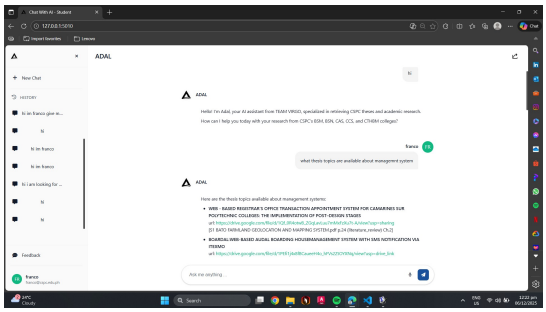

Figure 4: User Interface

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.

4.3 Evaluation Results

4.3.1 Automated Evaluation Metrics. The system was evaluated using the RAGAS framework, focusing on four core metrics: Answer Relevancy, Context Precision, Context Recall, and Faithfulness. Table 1 summarizes the results.

Table 1: RAG System Evaluation Metrics using RAGAS Framework

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

The table presents a performance profile characterized by precise, well-grounded answers. Faithfulness (0.9179) and Context Precision (0.9167) indicate that retrieved evidence is both accurate and tightly focused, yielding citations that trace cleanly to source pages. Context Recall (0.8711) shows broad coverage of relevant thesis passages, while Answer Relevancy (0.8625) confirms that final responses align with user intent in typical literature-search tasks.

In practice, a query such as "What methodologies are used for detecting academic plagiarism at CSPC?" returns a compact set of segments drawn from Methods and Related Works sections across multiple theses. The system synthesizes these into direct, cited responses; high precision keeps noise low, high recall surfaces cross-department perspectives, and high faithfulness maintains strict grounding in the referenced documents.

These results demonstrate the RAG system's effectiveness in retrieving and generating accurate, relevant, and well-grounded answers based on the indexed thesis documents from the CSPC Library. The high scores across all four evaluation metrics indicate that the system is capable of providing reliable academic assistance, making it a valuable tool for students and researchers seeking information from the library's thesis collection.

4.4 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.

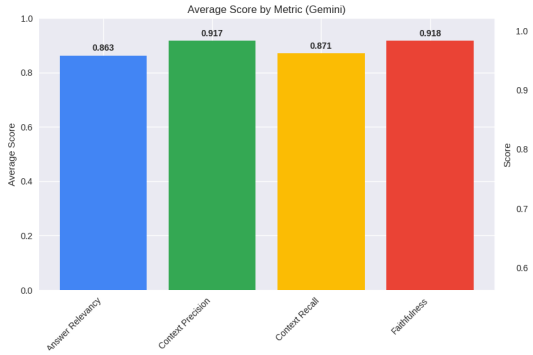

Figure 5: Bar Chart of RAG System Evaluation Result

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.

4.4.1 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 2 summarizes the results.

Table 2: 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.

4.4.2 User Feedback on RAG chatbot's Effectiveness and Usability. 3 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 3: 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

This study successfully demonstrated the feasibility and effectiveness of a Retrieval-Augmented Generation (RAG)-based chatbot system in enhancing thesis literature search and retrieval within the CSPC Library. By integrating semantic embeddings, vector database indexing, and Large Language Models, the system addressed critical limitations of traditional keyword-based search methods.

The findings reveal that the RAG chatbot delivers significant technical and user-centered benefits. Automated evaluation using the RAGAS framework confirmed strong performance, with Faithfulness and Context Precision exceeding 0.91, while Context Recall (0.8711) and Answer Relevancy (0.8625) demonstrated comprehensive and relevant retrieval capabilities. Complementary user-centered evaluation with 18 respondents yielded an overall weighted mean of 4.3 (Strongly Agree), affirming that users found responses accurate, relevant, clear, and appropriately comprehensive, with acceptable response times.

The cloud deployment enhanced accessibility, enabling users to search thesis literature from any location, thereby democratizing access to institutional knowledge resources. The system's innovative conversational approach provides a valuable tool for students and researchers, supporting informed literature discovery and academic guidance.

While promising, the system faces limitations. The dataset of 290+ theses may not fully represent disciplinary diversity, and document preprocessing quality directly impacts retrieval performance. Future research should explore scalability to larger datasets, optimization of chunk sizing and retrieval parameters, and applicability to other academic libraries. Additionally, continuous user feedback incorporation and periodic retraining with updated thesis submissions will maintain system relevance and performance.

In conclusion, this RAG-based chatbot represents a significant advancement in academic information retrieval, offering a modern alternative to traditional library systems and demonstrating the transformative potential of AI-driven semantic search in supporting academic discourse and institutional knowledge management.

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 **Rosel O. Onesa, MIT**, OIC Dean of the College of Computer Studies and our Thesis Adviser, for her invaluable guidance, recommendation and encouragement. We also thank **Allan Ibo Jr., MSc**, our Consultant, for sharing his expertise and providing constructive insights.

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

To **Joseph Jessie S. Oñate, MSc**, 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 **Tiffany Lyn O. Pandes, MSc**, 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 **Kaela Marie N. Fortuno, MIT**, 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] 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

[2] 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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] Jimmy Lin, Ma Ma, and Andrew Yates. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. ACM, New York, NY, 1154–1156. https://doi.org/10.1145/3437963.3441817

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

[10] C. Ryu, S. Lee, S. Pang, C. Choi, H. Choi, M. Min, and J. Sohn. 2023. Retrieval-based Evaluation for LLMs. In *Proceedings of the 1st Workshop on Neural and Learning-based Natural Language Processing (NLLP)*. https://doi.org/10.18653/v1/2023.nllp-1.13

[11] Sriramaraju Sagi. 2024. GENAI: RAG USE CASES WITH VECTOR DB TO SOLVE THE LIMITATIONS OF LLMS. *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY* 15 (04 2024), 56–62.

[12] 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.

[13] Chhagyani Thapa, Mahendran Chamikara, Seyit Camtepe, and Lichao Sun. 2022. Splitfed: When federated learning meets split learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8485–8493. https://doi.org/10.1609/aaai.v36i8.20825

[14] Alex Thomo. 2024. PubMed retrieval with RAG techniques. *Studies in Health Technology and Informatics* (2024). https://doi.org/10.3233/SHTI240498

[15] 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