

---

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

A Thesis Project  
presented to the Faculty of  
College of Computer Studies

In Partial Fulfillment of the Requirements  
for the degree Bachelor of Science in Computer Science

By  
Divino Franco R. Aurellano  
Herald Carl N. Avila  
Almira L. Calingacion

December 2025

---

## APPROVAL PAGE

In partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science, this research entitled **BEYOND LLMS: A RAG CHATBOT FOR EFFICIENT LITERATURE SEARCH AND THESIS RETRIEVAL IN CSPC LIBRARY** prepared and submitted by **Divino Franco R. Aurellano, Herald Carl N. Avila, Almira L. Calingacion** has been examined and is recommended for approval and acceptance.

**ROSEL O. ONESA, MIT**

Adviser

This research project entitled, **BEYOND LLMS: A RAG CHATBOT FOR EFFICIENT LITERATURE SEARCH AND THESIS RETRIEVAL IN CSPC LIBRARY** in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science has been examined and is recommended for acceptance and approval for ORAL EXAMINATION.

### RESEARCH COMMITTEE:

**KAE LA MARIE N. FORTUNO, MIT**

Member

**TIFFANY LYN O. PANDES, MSc**

Member

**JOSEPH JESSIE S. OÑATE, MSc**

Chairman

---

### PANEL OF EXAMINERS

**APPROVED** by the Committee on Oral Examination with a grade of **PASSED** on  
December 9, 2025.

**JOSEPH JESSIE S. OÑATE, MSc**

Chairman

**KAELA MARIE N. FORTUNO, MIT**

Member

**TIFFANY LYN O. PANDES, MSc**

Member

**ACCEPTED** and **APPROVED** in partial fulfillment of the requirements in Bachelor  
of Science in Computer Science with a grade of \_\_\_\_\_.  
\_\_\_\_\_.  
\_\_\_\_\_.  
\_\_\_\_\_.

**ROSEL O. ONESA, MIT**

OIC Dean, College of Computer Studies

Date: \_\_\_\_\_

---

## DEDICATION

We, the researchers, dedicate this work to God for giving us wisdom, strength, and perseverance throughout this journey. To our families, whose unconditional love, support, and encouragement have been the foundation of our success, and whose sacrifices and faith in us have made every step possible. To our mentors and instructors, for guiding, challenging, and inspiring us to grow and achieve our goals. To our friends, for their unwavering support, comfort, and laughter, which helped us overcome the challenges along the way. And to our classmates, for sharing this academic journey, offering collaboration, companionship, and memorable experiences that made this chapter meaningful. To all of you who believed in us and supported us in finishing this journey, we sincerely say “thank you.”

Finally, we dedicate this to ourselves, the researchers, who persevered through every stressful day; despite cramming and being scolded, we have successfully completed this study. We did cram, laugh, struggle, and forget, but we never forgot our goal.

– Team Virgo

## 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.

---

## ABSTRACT

|                          |   |
|--------------------------|---|
| <b>Title:</b>            | Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library |
| <b>Authors:</b>          | Divino Franco R. Aurellano<br>Herald Carl N. Avila<br>Almira L. Calingacion                     |
| <b>Number of Pages:</b>  | 139   |
| <b>School:</b>           | Camarines Sur Polytechnic Colleges  |
| <b>Degree Conferred:</b> | Bachelor of Science in Computer Science   |
| <b>Keywords:</b>         | RAG, Chatbot, Semantic Search, Thesis Retrieval, CSPC Library                                   |

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.

## TABLE OF CONTENTS

|  |     |
|--|-----|
| <b>Approval Page . . . . .</b>                                       | i   |
| <b>Panel of Examiners . . . . .</b>                                  | ii  |
| <b>Dedication . . . . .</b>  | iii |
| <b>Acknowledgments . . . . .</b>                                     | iv  |
| <b>Abstract . . . . .</b>  | v   |
| <b>List of Tables . . . . .</b>                                      | ix  |
| <b>List of Figures . . . . .</b>                                     | x   |
| <br>   |     |
| <b>Chapter 1: Introduction . . . . .</b>                             | 1   |
| Background of the Problem . . . . .                                  | 1   |
| Statement of the Problem . . . . .                                   | 3   |
| Objectives of the Study . . . . .                                    | 4   |
| General Objective . . . . .  | 4   |
| Specific Objectives . . . . .  | 4   |
| Significance of the Study . . . . .                                  | 4   |
| Scope and Limitation . . . . .                                       | 5   |
| Project Dictionary . . . . .   | 6   |
| Notes . . . . .  | 8   |
| <br>   |     |
| <b>Chapter 2: Related Literature and Studies . . . . .</b>           | 11  |
| Review of Related Literature and Studies . . . . .                   | 11  |
| Large Language Models . . . . .                                      | 11  |
| Retrieval-Augmented Generation . . . . .                             | 12  |
| Document Ingestion and Retrieval . . . . .                           | 13  |
| RAG Applications in Various Domains . . . . .                        | 15  |
| Evaluation of Retrieval-Augmented Generation (RAG) Systems . . . . . | 16  |
| Synthesis of the State-of-the-Art . . . . .                          | 17  |
| Gap Bridge of the Study . . . . .                                    | 18  |
| Notes . . . . .  | 19  |
| <br>   |     |
| <b>Chapter 3: Methodology . . . . .</b>                              | 22  |
| Research Design . . . . .  | 22  |
| Theorems, Algorithms, and Mathematical Models . . . . .              | 23  |
| Retrieval-Augmented Generation (RAG) Pipeline . . . . .              | 23  |

---

|   |        |
|---|--------|
| Large Language Model . . . . .  | 24     |
| Materials and Statistical Tools / Evaluation Methods . . . . .                        | 25     |
| Research Materials . . . . .  | 25     |
| Instrument . . . . .  | 27     |
| Statistical Test . . . . .  | 28     |
| Procedures . . . . .  | 35     |
| Evaluation Metrics . . . . .  | 37     |
| Context Precision . . . . .   | 37     |
| Context Recall . . . . .  | 38     |
| Response Relevance . . . . .  | 39     |
| Faithfulness . . . . .  | 39     |
| Conceptual Framework . . . . .  | 40     |
| Notes . . . . .   | 43     |
| <br><b>Chapter 4: Results and Discussion . . . . .</b>                                | <br>45 |
| Document Ingestion and Retrieval Module . . . . .                                     | 45     |
| Dataset and Preparation . . . . .   | 45     |
| Data Preprocessing . . . . .  | 46     |
| Indexing and Vector Database Construction . . . . .                                   | 47     |
| Semantic Search and Thesis Retrieval System . . . . .                                 | 47     |
| Query Encoding and Retrieval . . . . .  | 47     |
| Augmented Input Generation . . . . .  | 49     |
| Response Generation with Gemini 2.5-flash . . . . .                                   | 49     |
| Model Evaluation . . . . .  | 51     |
| RAG System Evaluation Results . . . . .   | 51     |
| User-Centered Evaluation Results . . . . .  | 56     |
| Notes . . . . .   | 60     |
| <br><b>Chapter 5: Summary of Findings, Conclusions, and Recommendations . . . . .</b> | <br>61 |
| Summary . . . . .   | 61     |
| Findings . . . . .  | 62     |
| Conclusions . . . . .   | 64     |
| Recommendations . . . . .   | 65     |
| <br><b>Bibliography . . . . .</b>   | <br>66 |
| <br><b>Appendices . . . . .</b>   | <br>72 |
| Appendix A: Relevant Source Code . . . . .  | 73     |
| Appendix B: Documentation . . . . .   | 83     |
| Appendix C: User's Guide . . . . .  | 85     |
| Appendix D: Data Collection Consent Form . . . . .                                    | 92     |
| Appendix E: Acknowledgment Receipt . . . . .  | 93     |
| Appendix F: Survey Questionnaire . . . . .  | 94     |
| Appendix G: Survey Response Tally . . . . .   | 97     |

---

|   |            |
|---|------------|
| Appendix H: Non-Disclosure Agreement Form . . . . .               | 98         |
| Appendix I: Joint Affidavit of Undertaking (Plagiarism) . . . . . | 104        |
| Appendix J: Project Team Assignment Form . . . . .                | 105        |
| Appendix K: Role Acceptance Form . . . . .                        | 106        |
| Appendix L: Thesis/Capstone Title Approval Form . . . . .         | 109        |
| Appendix M: Thesis/Capstone Hearing Form (TD, POD, FOD) . . . . . | 110        |
| Appendix N: Panel RSC (TD, POD, FOD) . . . . .                    | 113        |
| Appendix O: Consultation Log Form (CLF) . . . . .                 | 119        |
| Appendix P: Language Editing Certification . . . . .              | 125        |
| Appendix Q: Secretary's Certification . . . . .                   | 126        |
| Appendix R: Certificate of Transfer . . . . .                     | 127        |
| Appendix S: Certificate of Plagiarism Check . . . . .             | 128        |
| Appendix T: Certification for GPU Server Usage . . . . .          | 129        |
| Appendix U: ACM Format . . . . .                                  | 130        |
| <b>Curriculum Vitae . . . . .</b>                                 | <b>137</b> |

---

## LIST OF TABLES

|          |  |    |
|----------|--|----|
| Table 1  | Hardware Requirements . . . . .                                      | 25 |
| Table 2  | Software Requirements . . . . .                                      | 26 |
| Table 3  | Likert Scale for User Level of Agreement . . . . .                   | 28 |
| Table 4  | User Level of Satisfaction with Answers . . . . .                    | 30 |
| Table 5  | Likert Scale for User Level of Using the Chatbot Again . . . . .     | 31 |
| Table 6  | User Level of Understanding Chatbot Responses . . . . .              | 32 |
| Table 7  | Likert Scale for User Level of Confidence on Information Received .  | 33 |
| Table 8  | Chunk Analysis & Statistics . . . . .                                | 46 |
| Table 9  | RAG System Evaluation Metrics using RAGAS Framework . . . . .        | 51 |
| Table 10 | User Agreement: Chatbot Response Quality and Performance . . . . .   | 56 |
| Table 11 | User Feedback on RAG chatbot's Effectiveness and Usability . . . . . | 58 |

---

## LIST OF FIGURES

|          |   |    |
|----------|---|----|
| Figure 1 | Basic RAG Pipeline by Dr. Julija . . . . .                        | 23 |
| Figure 2 | Conceptual Framework of the RAG-Based Chatbot System . . . . .    | 41 |
| Figure 3 | CSPC Thesis PDF Sample . . . . .                                  | 45 |
| Figure 4 | Screenshot of Query and Retrieved Output . . . . .                | 48 |
| Figure 5 | Screenshot of Sensitive Query and Output Generated . . . . .      | 49 |
| Figure 6 | Comparison of outputs from two user with the same query . . . . . | 50 |
| Figure 7 | Bar Chart of RAG System Evaluation Result . . . . .               | 52 |
| Figure 8 | Heatmap of RAG System Evaluation Result . . . . .                 | 54 |
| Figure 9 | Radar Chart of RAG System Evaluation Result . . . . .             | 55 |

---

## CHAPTER 1

### INTRODUCTION

This chapter outlines the study's problem, objectives, and significance. It also defines the scope and limitations, and includes a project dictionary and notes with key terms and supporting details.

#### **Background of the Problem**

Large Language Model (LLM) like GPT [2] and Gemini [13] have unprecedently 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 [17, 6]. 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, 12, 18]. Unfortunately, most digitalized libraries today still use outdated search systems that need an exact keyword search, which can result in irrelevant materials [21]. 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.

These challenges of university libraries in the Philippines are shared difficulties in accessing academic resources, outdated search systems, and ineffective information retrieval that affect the efficiency of academic 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 overcome these, Retrieval-Augmented Generation (RAG) has emerged as a superior approach [14]. Unlike standalone LLMs, which require retraining and adding domain-specific data to adjust the LLM weights, RAG presents a state-of-the-art approach that can

retrieve relevant external information to generate responses. It holds a significant practical implications for university libraries that can improve search functionalities. Additionally, RAG ensures that the most relevant academic resources are retrieved quickly and straightforwardly, making it suitable for libraries with extensive collections of academic papers that are difficult for researchers and students to navigate [23, 9].

This thesis developed an enhanced LLM-powered chatbot with the integration of RAG AI framework to improve information retrieval, especially in thesis retrieval of university-owned thesis PDFs at CSPC Library. This chatbot application generates answers and retrieves relevant documents based on the user's prompt.

### **Statement of the Problem**

Finding relevant thesis literature in a University's library, such as in CSPC, can be challenging. Many researchers in the academic community struggle to find the exact thesis paper they need, often requiring them to travel and physically visit the library just to retrieve specific documents.

Currently, CSPC's library website [25] only allows users to search by exact document title. Finding relevant research becomes difficult if users don't know the exact title. What's worst is that library policies restrict researchers from taking thesis books outside the premises, limiting accessibility to research resources. In response to these challenges, this study aims to explore creating a chatbot that eliminates those limitations by enabling searches based on topics, keywords, general descriptions and conversational query. The ultimate goal is to make this tool widely accessible by deploying it on a scalable cloud platform such as Azure.

This goal, by leveraging RAG, this project aims to revolutionize how the academe community interacts with the CSPC library, making research faster, smarter, and more user-friendly.

## **Objectives of the Study**

The objectives of this study are divided into two categories: general and specific. The general objective defines the overall goal of the study, while the specific objectives break down this goal into measurable and achievable steps. These objectives ensure a structured approach to developing an enhanced LLM chatbot for Camarines Sur Polytechnic Colleges.

### ***General Objective***

This study aims to develop a chatbot using RAG to revolutionize thesis retrieval and searching in the CSPC Library.

### ***Specific Objectives***

To achieve the general objective, the study sets the following specific objectives:

1. To integrate a document ingestion and retrieval module for storing thesis documents.
2. To implement a semantic search and thesis document retrieval system using RAG and Google Gemini.
3. To evaluate the performance of the RAG chatbot using RAGAS and user satisfaction metrics.

## **Significance of the Study**

The result of this study will benefit the following:

**Students.** This chatbot can help students find campus-relevant research and reduce the time spent on literature review. This will help them to find relevant studies in seconds, without relying solely on exact keywords or titles.

**Faculty Members.** The system can serve as a research companion for faculty members by providing easier access to all the university's published theses. This can also enhance

their competence in teaching students with thesis writing, academic guidance, and collaborative research work, while at the same time reducing the extent of manual searching for sample published campus theses.

***CSPC Library Management.*** The implementation of a RAG-powered chatbot can revolutionize the library's digital infrastructure, making the academic resources more accessible to users.

***Researchers.*** Current researchers can build on this study to explore the field of AI-driven searching and retrieval. This will add valuable knowledge to the practical applications of RAG.

***Future Developers.*** Future developers can use the findings of this study and use it as a technical reference in AI chatbot implementation in academe.

## **Scope and Limitation**

This study aimed to develop a chatbot for CSPC library, applying RAG framework with Google Gemini LLM. The goal is to address the challenges being faced by the academe community, specifically student researchers in searching and retrieving theses in the library by replacing the current traditional keyword-based search with a more conversational and topic-oriented approach. This will be done through a website with access control, allowing administrators to upload newly published PDF theses and users to register using their CSPC email. Additionally, the system is intended to be deployed to the cloud.

However, there are certain limitations to consider in this study. First, the researchers will focus only on utilizing the available PDF copies of undergraduate theses that have already been published. Second, the chatbot's accuracy can rely on the quality of written info inside the thesis pdf, as well as the clarity and relevance of the user's prompts. Additionally, system performance can be limited to the cloud resources allocated by the researchers given the constraints in budget. This influences the chatbot's real-time processing capacity.

And lastly, while this approach can reduce hallucination, users are still advised to validate the outputs carefully, as occasional inaccuracies or fabricated info may still occur.

## Project Dictionary

The Project Dictionary contains the technical terms that defined the conceptual and operation of this study:

- **Academic Literature Retrieval.** The process of systematically searching for and obtaining research documents, to be used in academic work [20]. In this study, the implementation of LLMs is essential to improve the retrieval of available theses documents in CSPC.
- **Chatbot.** Chatbot refers to a conversational agent that is designed to provide assistance, answer queries, and give access to information using natural language and a user-friendly manner [7]. In this study, the chatbot was used to answer questions with human-like responses.
- **Google Gemini.** Google Gemini is a leading multimodal models with advanced reasoning through thinking, long context and tool-use capabilities that can be combined to unlock new agentic workflows like RAG [8]. In this study, Google gemini was used to help the chatbot in reasoning and providing answers based on the long context thesis paper using human-like responses, not limited to English language.
- **Generative AI.** A Generative AI is a subset of artificial intelligence capable of using human language effectively and producing results from carefully designed prompts [5]. In this study, the implications of Gen AI in the context of education and academic integrity were examined.
- **Large Language Models (LLMs).** LLM is an Advanced transformer-based algorithms with billions of parameters that uses attention mechanisms to process massive

datasets and generate coherent, context-aware text [11]. In this study, LLM is used to process hundreds of CSPC thesis PDFs and also worked in the reasoning task.

- **Natural Language Processing (NLP).** NLP is a field of AI that enables computers to understand, work with, and use human language in ways similar to how people talk to each other [19]. In this study, NLP is important for making the RAG pipeline work for users when searching and retrieving theses in CSPC library.
- **Retrieval-Augmented Generation (RAG).** RAG is a language model that takes an input ( $x$ ), retrieves relevant documents ( $z$ ), and uses those documents as extra context to produce an output ( $y$ ) [14]. In this study, RAG was developed for navigating and retrieving information from large amounts of academic papers.
- **Semantic Search.** Semantic Search is an approach in information retrieval that aims to understand the meaning and connections between words, and designed to imitate human understanding [16]. In this study, semantic search will work with RAG in generating relevant and contextual responses.

## Notes

- [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*, --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*.
- [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.
- [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*.
- [5] Aras Bozkurt. 2024. Genai cocreation, authorship, ownership, academic ethics and integrity in a time of generative ai. *Open Praxis*, 16, 1, 1–10.
- [6] Mark Chen. 2021. Evaluating large language models trained on code. *arXiv: 2107.03374 [cs.LG]*.
- [7] James CL Chow, Valerie Wong, Leslie Sanders, and Kay Li. 2023. Developing an ai-assisted educational chatbot for radiotherapy using the ibm watson assistant platform. In *Healthcare* number 17. Vol. 11. MDPI, 2417.
- [8] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and Evan Rosen. 2025. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*. <https://arxiv.org/abs/2507.06261>.
- [9] 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.

- [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*, 100145.
- [11] Eyal Klang, Lee Alper, Vera Sorin, Yiftach Barash, Girish N Nadkarni, and Eyal Zimlichman. 2024. Advancing radiology practice and research: harnessing the potential of large language models amidst imperfections. *BJR—Open*, 6, 1, tzae022.
- [12] Sammy Lagas and Jonathan Isip. 2023. Challenges to digital services in philippine academic libraries. *Philippine Journal of Librarianship and Information Studies*, 43, 1, 27–38.
- [13] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábreo, 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. (2025). <https://arxiv.org/abs/2503.07891> arXiv: 2503.07891 [cs.CL].
- [14] 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, 9459–9474.
- [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] Ali Mahboub, Muhy Eddin Za’ter, Bashar Al-Rfooh, Yazan Estaitia, Adnan Jaljuli, and Asma Hakouz. 2024. Evaluation of semantic search and its role in retrieved-augmented-generation (rag) for arabic language. *arXiv preprint arXiv:2403.18350*.
- [17] 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*.

- [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, 1526–1533.
- [19] José Gabriel Carrasco Ramírez. 2024. Natural language processing advancements: breaking barriers in human-computer interaction. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 3, 1, 31–39.
- [20] Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare* number 6. Vol. 11. MDPI, 887.
- [21] 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, 231–237.
- [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] Zijie J Wang and Duen Horng Chau. 2024. Mememo: on-device retrieval augmentation for private and personalized text generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2765–2770. <https://arxiv.org/abs/2407.01972>.

---

## **CHAPTER 2**

### **RELATED LITERATURE AND STUDIES**

This chapter reviews related literature and existing systems on the study. It provides a synthesis of related works, an overview of state-of-the-art technologies and methodologies, and underline the research gaps that the present study has addressed.

#### **Review of Related Literature and Studies**

A review of books, scholarly articles, journals, and previous thesis projects concerning the research topic was carried out to develop an in depth understanding of the subject of the study. The findings are organized thematically in line with the key areas of the study.

#### ***Large Language Models***

Large Language Models (LLMs) have significantly improved the use case of information retrieval (IR) within academic settings. The integration of LLMs, like ChatGPT and other model architectures, offers notable advancements in natural language processing (NLP) and also proves its capabilities to enhance IR, question-answering, summarization, and content generation, which benefits academic environments where efficient access to information is crucial [21, 22]. Recent works by Khraisha et al. [2024] and Gartlehner et al. [2023] have demonstrated how large language models can automate research tasks such as systematic review, data extraction, and document screening. This suggests that LLMs are capable of improving research productivity in academia [12, 9].

Even though LLMs offer a number of advantages with regard to information retrieval, they also present significant challenges. Among the major challenges is that they are inefficient at executing domain specific tasks for which specialized knowledge is required.

This is because these models source knowledge from pre-training, hence LLMs are limited to offer fact based responses within certain domains such as academia. Omar et al. mentioned in their study that LLMs, including ChatGPT, could be helpful in specialized domains to serve complementary purposes but may fail in complicated queries because they have not seen enough training data related to those fields [12]. Moreover, pre-trained LLMs have difficulty keeping on pace with continuous data growth in various domains, thus, it is impossible for them to refresh their knowledge without extensive fine-tuning. Lucas et al. point out that this inability-in an academic and professional sense-of large language models to access up-to-date domain-specific repositories significantly diminishes their effectiveness and value [9].

Despite of LLMs being at the lead of NLP innovation, their actual application in domain-specific tasks is deeply prevented by certain challenges. These include real-time data availability, dependence on pre-trained knowledge bases, and ethical concerns pertaining to data privacy. Innovation around these challenges using novel methodologies, such as the retrieval-augmented generation approach, increases the capability of models to meet specialized applications that have strict requirements.

### ***Retrieval-Augmented Generation***

RAG has achieved remarkable improvements in the IR domain, especially in tasks regarding literature search and thesis retrieval in library systems [20]. The architecture supports the traditional large language models with external knowledge sources for enhancing the relevance, richness, and correctness of the responses [7].

As articulated in the studies of Lewis et al. [2020] titled "Retrieval-Augmented Generation for Knowledge Intensive NLP Tasks," RAG provides for much more accurate responses because it helps in addressing certain inherent limitations of LLMs, especially in the areas of accurate knowledge retrieval and context relevance. Another academic study about the topic is the the study of Shuster et al. [2021], titled "Retrieval Augmentation

Reduces Hallucination in Conversation,” go on to show that RAG reduces inconsistencies and hallucinations in the LLM outputs. Their results show that RAG mechanisms increase conversational fluency and integrity, particularly in open domain conversational settings, bringing about knowledgeable and more coherent responses.

Moreover, it has recently been further supported by work titled ”GENAI: RAG Use Cases with Vector DB to Solve the Limitations of LLMs,” in which the authors demonstrate that the integration of vector databases with RAG essentially improves retrieval speed and relevance. Semantic search using vector databases makes continuous real-time updates possible for dynamic domains such as business and academic libraries, resulting in a high level of knowledge management and factual correctness in the generated responses. Thus, RAG strengthens not only the capability of LLM retrieval but also addresses the essential weaknesses of LLMs: consistency and factuality [16].

### ***Document Ingestion and Retrieval***

Successful execution of RAG systems relies on utilizing documents in an effective manner and carrying out robust retrieval procedures, particularly in addressing large and complex datasets found in academic libraries. RAG systems can use any data source, including text, video, images, and audio, thus allowing flexible and contextually rich information retrieval. In study this about RAG, the main corpus mostly used is PDF documents to extract academic content which focuses by this authors [14].

The RAG chatbot rely its effectiveness by depending on the quality of preprocessing, which involves converting unstructured PDF data into machine readable formats suitable for embedding and semantic search [5, 4]. The PyPDF2, PyMuPDF, and pypdfium are the commonly used tools for this task in most studies to help in extracting raw text from complex PDF layouts [1].

A study by Sagi [2024], ”GENAI: RAG Use Cases with Vector DB to Solve the Limitations of LLMs,” further corroborates this idea by showing how the introduction of vector

databases significantly improves retrieval speed and relevance when integrated into RAG. The semantic search capabilities of the vector databases in turn support continuous real-time updates in dynamic domains like academic and business libraries, hence greatly improving knowledge management and factual accuracy of responses generated. Therefore, RAG enhances not only the retrieval capabilities of LLMs but also considerably mitigates its traditional deficiencies in consistency and factual accuracy [16].

Adhikari and Agarwal [2024] into their examined different kind of PDF parsers using F1 score, BLEU-4, and local alignment on a wide range of document types. The results show that PyMuPDF and pypdfium more reliable to retain sentence structure and format than the other tools, something highly recommended for retaining semantic coherence and offering proper vectorization and retrieval. One can also observe parsing difficulties posed by complex documents, such as scientific articles and patent PDFs, for which rule-based tools significantly underperform versus transformer-based models [1].

As stated out by Zhang et al. [2023], automated ingestion pipelines parsing documents into a searchable database improve the discoverability and access to scholarly content.

Techniques like OCR, metadata extraction, and structured indexing are common practices employed on thesis repositories to enable various retrieval operations easily [23]. Along related lines, Karpukhin et al. [2020] stress the importance of document preprocessing, chunking, and embedding to permit semantic search in DPR and, consequently, for modern RAG [11]. Generally, this ingestion process consists of several steps: (1) extracting text with PyMuPDF or pypdfium, among other tools; (2) chunking the text into smaller, logically coherent pieces; and (3) embedding by models such as Sentence-BERT. Later, these vectors will be kept in specialized vector databases (e.g., FAISS, Pinecone) so that the actual retrieve action when users make inquiries will be fast and efficient. Hence, robust document ingestion and storage directly impact retrieval accuracy, system responsiveness, and user experience. Taking as its basis on Sagi's study, it is clear that effectiveness in ingestion and vectorization yields fast retrieval of relevant information and generates very

high accuracy, context enriched responses from RAG models, particularly within domain like academic libraries [11].

The authors in Deepak et al. [2025], "LangChain-Chat with My PDF", present the prevalent workflow for processing PDFs, which is focused on embedding and chunking within the general techniques of vectorization. This work illustrates how chunking enables RAG to retrieve the most relevant document segments in order to answer user queries and hence better handle long PDF enhancing search capabilities within the system [8].

Taken together, these works suggest that perfectly done preprocessing, data ingestion, and vectorization form is crucial for bridging static document repositories with real-time information retrieval, demonstrating the potential for RAG architectures in managing large collections of academic knowledge [2, 4].

### ***RAG Applications in Various Domains***

Beyond contexts in the field of academics, RAG frameworks are increasingly applied to specialized domains, including legal research, medical information retrieval, and scientific literature search, underlining broad versatility and impact.

In the academic domain, Grigoryan and Madoyan [2024] proposed a solid foundation for future advancements in academic information retrieval through their study "Building a Retrieval-Augmented Generation (RAG) System for Academic Papers," creating a RAG system that retrieves and analyzes a vast amount of academic papers. They used vector search techniques like cosine similarity and HNSW indexing on the paper's abstracts and full texts to pass only relevant information to LLM, which significantly enhanced the retrieval and generation [10]. This was also supported by Aytar et al. [2024], that proposed enhanced RAG architecture for academic literature navigation in data science, introducing their approached of semantic chunking and an Abstract-First strategy, which significantly enhances the capability of RAG to retrieve pertinent academic content. [6].

In the healthcare domain, As Arzideh et al. [2024] point out, using domain specific

clinical embeddings makes the RAG approach highly effective. Their work, "MIRACLE-Medical Information Retrieval using Clinical Language Embeddings for Retrieval Augmented Generation at the Point of Care," have enhances clinical decision making, improves workflows in documentation, and personalizes access to health information. Supporting this, Amugongo et al. [2024] note that the RAG architecture is able to retrieve external medical data in a very effective manner, thus providing responses with high levels of accuracy and reliability while minimizing the limitations characteristic of traditional LLMs.

Similarly, in the legal domain, a study by Aquino et al. [2024] notes that RAG systems greatly enhance legal research by speeding up the retrieval of cases and statutes and enhancing the authenticity and contextual accuracy of the output.

In conclusion, these studies collectively show that RAG has already been applied on various domains, from academic research to other specialized fields. Majority of the findings highlights that by using various techniques within the RAG architecture can enhance any domain-specific tasks in a way that traditional LLM cannot achieve.

### ***Evaluation of Retrieval-Augmented Generation (RAG) Systems***

Evaluation of RAG systems requires methodological extensions beyond those applied in traditional designs for large language models. The RAGAS framework presents a structured way of evaluating the context precision, contextual relevance, and faithfulness in a generated response. Studies of Shuster et al. [2021] shows that high quality retrieval has a great effect on user satisfaction and perceived reliability with respect to conversational AI. This, therefore, underlines the need for specialized evaluation frameworks which can guarantee the effectiveness of RAG systems.

In specialized needs, metrics tailored for RAG models come into pivotal consideration. One of the widely used approaches is the RAGAS evaluation framework that provides key metrics such as Context Recall, Faithfulness, and Response Relevance. These metrics assess the degree to which the retrieved documents substantiate the generated response [15].

Context Precision characterizes the share of relevant chunks inside the contexts retrieved, while Context Recall ensures that no important information is missed. Faithfulness checks factual coherence between generated responses and the respective retrieved documents, while Response Relevance checks if the response correctly addresses user query [4] [8].

However, while automated measures may be more valid for some purposes, they often lack coverage of the qualitative dimensions of consistency, fluency, and overall user satisfaction. Human evaluation, according to Sivasothy et al. [2024], is necessary in improving these systems, since it considers factors that the automated approaches may fail to address.

### Synthesis of the State-of-the-Art

The related literature and systems discussed have substantial relevance to the problem of the study. To have a clear understanding of this literature and studies, the researchers made a synthesis in the succeeding discussions.

LLM with integrated RAG techniques have greatly improved the knowledge-intensive NLP tasks, overcoming LLMs' challenges. The studies of Thapa et al. [2022] and Thomo [2024] on how combining RAG with LLMs significantly improves accuracy and coherence in conversations and complex queries. The advantage of this technique enables LLMs to retrieve relevant external data, reducing hallucinations and improving factual consistency [19, 20]. Furthermore, Lewis et al. [2020] discussed the application of vector databases for continuous integration with RAG, which shows a significant improvement in both retrieval efficiency and the relevance of output produced by large language models . This topic plays a major role in literature searching and retrieving theses from university libraries [13].

A collection of various studies compiled on the use of RAG in different domains is discussed below. For instance, Arzideh et al. [2024] incorporated clinical language embeddings into RAG for better healthcare information [5]. Grigoryan and Madoyan [2024], in "Building a Retrieval-Augmented Generation (RAG) System for Academic Papers,"

proposes a system that augments academic retrieval through vector search [10]. Further, Aquino et al. [2024] used RAG on the effective extraction and analysis of Brazilian legal [4]. Together, these reports illustrate the adaptability of RAG and its potential to transform how university libraries search for and provide access to academic theses.

Evaluation metrics are important for evaluating the performance of RAG in retrieving and generating accurate responses. Specific metrics of RAGAS, such as Context Precision, Faithfulness, and Answer Relevance, as emphasized in the studies of Sagi [2024] and Arzideh et al. [2024], ensure the authenticity and consistency of the generated outputs of the model [16, 5]. Despite the effectiveness of automated metrics, human evaluation remains important to assess coherence and user satisfaction, as mentioned in this study [4].

In summary, RAG integrated with LLMs presents a groundbreaking method for improving literature searches and thesis retrieval in university libraries, especially at CSPC library. Through testing the limitations and obstacles faced by traditional LLMs, the integration of RAG reveals its promise to transform research accessibility at the CSPC library.

### **Gap Bridge of the Study**

Existing literatures has explored the applicability of RAG in domains from academic research to other specialized fields, it is concerning to note that there's a gap in the way of specific applications of these systems to academic libraries for enhancing literature searches and thesis retrievals. While previous studies have shown how well RAG improves information retrieval, not much has been done in applying this within university libraries, where there is a unique challenges and requirements that such implementations needed.

This study tries to fill this gap by designing a RAG-based chatbot system specific to the CSPC library. By focusing on the unique challenges and demands of academic libraries, this paper seeks to add substantial value in understanding the appropriate application of RAG systems toward improvement of information retrieval.

## Notes

- [1] Narayan S. Adhikari and Shradha Agarwal. 2024. Comparative study of pdf parsing tools across diverse document categories. JadooAI, Sacramento, CA, USA; Missouri University of Science and Technology, USA. (2024). <https://arxiv.org/abs/2410.09871v2> arXiv: 2410.09871v2.
- [2] Uday Allu, Biddwan Ahmed, and Vishesh Tripathi. 2024. Beyond extraction: contextualising tabular data for efficient summarisation by language models. (2024). doi:10.36227/techrxiv.170792474.42605726/v1.
- [3] Lameck Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2024. Retrieval augmented generation for large language models in healthcare: a systematic review. (2024). doi:10.20944/preprints202407.0876.v1.
- [4] Isabella Aquino, Matheus Santos, Carina Dorneles, and Jonata Carvalho. 2024. Extracting information from brazilian legal documents with retrieval augmented generation. In *SBBB Estendido*, 280–287. doi:10.5753/sbbd\_estendido.2024.244241.
- [5] Kamyar Arzideh, Henning Schäfer, Ahmad Idrissi-Yaghi, Bahadır Eryılmaz, Mikel Bahn, Cynthia Schmidt, and Rene Hosch. 2024. Miracle - medical information retrieval using clinical language embeddings for retrieval augmented generation at the point of care. *Research Square*. doi:10.21203/rs.3.rs-5453999/v1.
- [6] Ahmet Yasin Aytar, Kemal Kilic, and Kamer Kaya. 2024. A retrieval-augmented generation framework for academic literature navigation in data science. *arXiv preprint arXiv:2412.15404*.
- [7] Jiashu Chen, Hongyin Lin, Xu Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 16. Vol. 38, 17754–17762. doi:10.1609/aaai.v38i16.29728.
- [8] M. Deepak, A. Anusha, P. Phanivighnesh, and G. Sreenivasulu. 2025. Langchain-chat with my pdf. *International Journal of Scientific Research in Engineering and Management*, 09, 03, 1–9. doi:10.55041/ijjsrem42403.
- [9] 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. Preprint. doi:10.1101/2023.10.02.23296415.

- [10] Anna Grigoryan and Habet Madoyan. 2024. Building a retrieval-augmented generation (rag) system for academic papers. (2024).
- [11] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP 2020*. doi:10.18653/v1/2020.emnlp-main.550.
- [12] Qais Khraisha, Stefaan Put, Jens Kappenberg, Adeel Warratich, 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, 616–626. doi:10.1002/jrsm.1715.
- [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, 9459–9474.
- [14] Shuyuan Li, Yunjiang Zhang, Zhaolin Fang, Kong Meng, Rui Tian, Hong He, and Shaorui Sun. 2023. Extracting the synthetic route of pd-based catalysts in methanol steam reforming from the scientific literature. *Journal of Chemical Information and Modeling*, 63, 20, 6249–6260. doi:10.1021/acs.jcim.3c01442.
- [15] Sujoy Roychowdhury, Sumit Soman, H G Ranjani, Neeraj Gunda, Vansh Chhabra, and Sai Krishna Bala. 2024. Evaluation of rag metrics for question answering in the telecom domain. (2024). <https://arxiv.org/abs/2407.12873> arXiv: 2407.12873 [cs.CL].
- [16] Sriramaraju Sagi. 2024. Genai: rag use cases with vector db to solve the limitations of llms. *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY*, 15, (Apr. 2024), 56–62.
- [17] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. (2021). <https://arxiv.org/abs/2104.07567> arXiv: 2104.07567 [cs.CL].
- [18] Shangeetha Sivasothy, Scott Barnett, Stefanus Kurniawan, Zafaryab Rasool, and Rajesh Vasa. 2024. Ragprobe: an automated approach for evaluating rag applications. (2024). <https://arxiv.org/abs/2409.19019> arXiv: 2409.19019 [cs.CL].

- [19] 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* number 8. Vol. 36, 8485–8493. doi:10.1609/aaai.v36i8.20825.
- [20] Alex Thomo. 2024. Pubmed retrieval with rag techniques. *Studies in Health Technology and Informatics*. doi:10.3233/SHTI240498.
- [21] 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, e244630. doi:10.1001/jamanetworkopen.2024.4630.
- [22] 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, 255–263. doi:10.1002/hcs2.61.
- [23] L. Zhang, X. Chen, and M. Li. 2023. Automated document ingestion for academic knowledge repositories. *Journal of Digital Libraries*, 24, 1, 12–26.

---

## **CHAPTER 3**

### **METHODOLOGY**

This chapter discusses the specific steps and logical procedures that were employed to develop and evaluate the RAG-based LLM chatbot system. This includes the research design, theoretical and mathematical framework, software and hardware tools, instruments, procedures, evaluation metrics, and a conceptual framework.

#### **Research Design**

Constructive research approach focuses on producing innovative constructions that intend to solve real-world problems and contribute to the theory of the discipline in which it is applied. This methodology is particularly well-suited to fields like information systems and artificial intelligence, where the goal is not only theoretical insight but also the creation of innovative, functional systems [7].

This study adopted a constructive research design to develop a library chatbot using RAG for CSPC. 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.

By adapting a constructive design, the researchers were able to build on existing methodologies while innovating and establishing a new solution that addressed the specific challenges in the current search and thesis retrieval in CSPC. By combining the existing infor-

mation within the tailored architecture, this study contributed to the demanding research on LLMs and RAG being used in various domains, such as education and research discovery.

### 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.

#### ***Retrieval-Augmented Generation (RAG) Pipeline***

The 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.

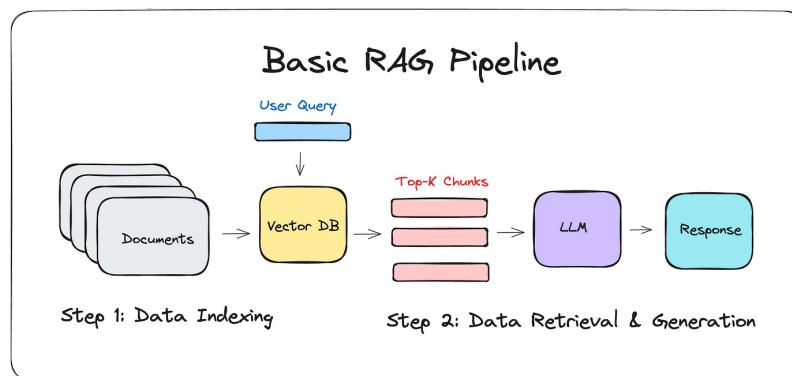


Figure 1: **Basic RAG Pipeline by Dr. Julija**

The RAG pipeline as illustrated in Figure 1, consists of the following key stages:

#### A. Data Indexing

The process begins with data loading, where all needed thesis documents are imported to be utilized. Considering that every thesis documents are large document composing of hundreds of pages, these were splitted into smaller chunks using a token-based method

that respects academic structure (Abstract; Chapters 1–5). Each chunk is then converted into a vector using the open-source ‘sentence-transformers/all-MiniLM-L6-v2‘ embedding model from Hugging Face, which was chosen for its lightweight architecture and strong semantic representation capabilities. Lastly, the data storing, where vector embeddings and their associated metadata were stored in FAISS for efficient similarity search.

### *B. Retrieval and Generation*

This stage is where the chatbot creates an action that addresses the challenge. It begins when a user starts to query, and that query was embedded using the same model (sentence-transformers/all-MiniLM-L6-v2) used from the first stage. That embedded query was used by FAISS retriever to perform a semantic search matching from the vector database. The parameter on Top-K most relevant chunks was set as default K =50, to balance precision and recall. And finally, during generation, these retrieved chunks are fed to the Google Gemini 2.5-flash language model as grounded context to create a human-like response.

### ***Large Language Model***

Based on the study of Naveed [2024], LLMs have demonstrated remarkable performance in challenges related to NLP and beyond. These were also considered as cutting-edge AI systems trained on massive datasets to process and generate text and can excel in tasks such as summarization, question answering, and retrieval [11].

### *Gemini 2.5-flash*

Specifically, the llm that was used in this project is Gemini 2.5 Flash, part of the Gemini 2.X model family introduced by Comanici et al. [2025]. It gives advanced reasoning capability at a lower compute resource, low latency and cost. This family also top-ranks in the llms that have multimodal support and extended context windows [2]. By incorporating it into this project, it can enhance thesis retrieval tailored for CSPC library users.

## **Materials and Statistical Tools / Evaluation Methods**

To ensure optimal performance of the RAG-based LLM system, several key hardware and software components are required.

### ***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***

To support the development of the RAG chatbot, the researchers used hardware components that meet or exceed the specifications listed in Table 1. These components were selected to ensure a smooth ingestion and embedding process of the large pdf dataset.

**Table 1**  
**Hardware Requirements**

| <b>Component</b>    | <b>Specification</b>           |
|---------------------|--------------------------------|
| 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) |

The researchers considered using a modern multi-core CPU to enable efficient data processing and high inference. Sufficient memory was also crucial; At least 16 GB of RAM is recommended and is the least to manage large-scale ingesting and embeddings. A 1 TB SSD was preferred to use due to its high read/write speeds, which help in the indexing process, and also serves as the data storage for thesis PDF documents. Considering also that LLM embeddings are resource-intensive when used, a powerful GPU like NVIDIA RTX 3090 or a higher version was recommended to speed up deep learning inference and vector operations.

*Software*

**Table 2**  
**Software Requirements**

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

Python serves as the core programming language due to its strong support for ML and NLP. While version 3.10 or later was used as the latest stable version, this may vary if there's a new update. FAISS from Facebook was utilized as the vector database for its fast vector similarity search and compression. The Gemini 2.5-flash LLM called from

Google Generative AI API serves as the reasoning model for the chatbot, and the sentence-transformers/all-MiniLM-L6-v2 from Hugging Face was used as the embedding model that transformed chunks into vector embeddings. While the Flask framework was used to build the whole system for its simplicity.

Various modules were also called to help in document parsing and extraction. One is through PyMuPDF for text extraction from thesis PDF files. NumPy for numerical operations and help a lot in the evaluation, and the LangChain framework manage the orchestration of LLMs during query interpretation and response generation.

### ***Instrument***

In this subsection, the instruments that were 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 [10]. 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.

### ***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 [4, 1, 6]. 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 for evaluating individuals' attitudes toward any object. In particular, the 5-point Likert Scale was chosen because it works well in surveys and requires less time and effort to develop [5, 9].

**Table 3**  
**Likert Scale for User Level of Agreement**

| <b>Scale</b> | <b>Range</b> | <b>Level of Agreement</b> | <b>Description</b>   |
|--------------|--------------|---------------------------|--|
| 5            | 4.21 - 5.00  | Strongly Agree            | The participant strongly supports or agrees with the chatbot's response.   |
| 4            | 3.21 - 4.20  | Agree                     | Implies a positive stance toward the chatbot's response.   |
| 3            | 2.61 - 3.20  | Neutral                   | The respondent has neither a positive response nor a negative response, but undecided denotes a state of confusion of the respondent.                    |
| 2            | 1.81 - 2.60  | Disagree                  | Suggests a level of disagreement with the statement or question, but not as strong as Strongly Disagree.   |
| 1            | 1.00 - 1.80  | Strongly Disagree         | Indicates a strong and definitive disagreement with the statement or question. The respondent strongly opposes or disagrees with the chatbot's response. |

Table 3 shows that this study employed a 5-point Likert Scale to determine "User level of Agreement" based on the user's experience with the system's response quality and per-

formance. In the table, the first column showed the scale, and the second is its equivalent range, then the scale that the system level of agreement fell under was shown in the third column, and with its corresponding description in the fourth column. The user's response was computed using the weighted mean statistical method, and then it was determined in which range it fell.

The scale is the numerical value that corresponds to the level of agreement (Strongly agree, agree, neutral, disagree, strongly disagree). This is important as it allows it to be used in statistical analysis. Scale 5 with a range of 4.20 - 5.00 described as "Strongly Agree", which means that the users who tested the chatbot completely agreed with the described criteria or considered its quality and performance excellent. Scale 4 with a range 3.40 - 4.19 described as "Agree", shows that the user who tested the chatbot generally agrees with the described criteria, but not to the strongest extent. The Scale 3 with a range of 2.60 - 3.39 was described as "Neutral", meaning that the user of who tested the chatbot is neutral, undecided, or the criteria description doesn't strongly resonate in either direction. Scale 2 with a range of 1.80 - 2.59 described as "Disagree", denoting that the user who tested the chatbot disagrees with the criteria description but not as intensely as "Strongly Disagree". And lastly, Scale 1 was described as "Strongly Disagree", meaning that the user who tested the chatbot completely disagrees with the criteria description or finds the system to be low quality and low performance.

Another set of questionnaires was prepared to determine the level of user satisfaction, likelihood of using the application in the future, ease of reading and understanding chatbot output, and confidence in response information regarding the RAG chatbot system. User opinions are expected to provide an understanding of the overall user experience and satisfaction level with the chatbot's performance in thesis retrieval tasks and its accessibility that proposed to solve the problem. A Likert scale-type questionnaire was also used for this purpose, employing the same 5-point scale framework to ensure consistency in measurement and analysis as shown in Table 4.

**Table 4**  
**User Level of Satisfaction with Answers**

| Scale | Range       | Level of Satisfaction | Description   |
|-------|-------------|-----------------------|---|
| 5     | 4.21 - 5.00 | Very Satisfied        | The participant is very satisfied with the chatbot's answers.   |
| 4     | 3.21 - 4.20 | Satisfied             | Indicates a positive satisfaction toward the chatbot's answers.   |
| 3     | 2.61 - 3.20 | Neutral               | The respondent has neither a positive nor negative satisfaction; undecided or indifferent denotes a state of uncertainty of the respondent. |
| 2     | 1.81 - 2.60 | Unsatisfied           | Suggests a level of dissatisfaction with the chatbot's answers, but not as strong as Very Unsatisfied.                                      |
| 1     | 1.00 - 1.80 | Very Unsatisfied      | Indicates a strong and definitive dissatisfaction with the chatbot's answers. The respondent is very unhappy with the chatbot's responses.  |

Table 4 shows that RAG chatbot system used 5-point Likert Scale to determine users Level of Satisfaction based on the user's experience to the system's answers. The first column showed the scale that the system level of satisfaction fell under which was shown in third column and its corresponding definition in the fourth column. User's response would be computed using weighted mean and was determined in which range fell under. Scale 5 with a range of 4.20-5.00 described as "Very Satisfied" which means that the user of the RAG Chatbot completely satisfies with the provided answers, scale 4 with a range 3.40-4.19 described as "Satisfied", the user of the RAG Chatbot is satisfied with the provided answers but not to the strongest extent, scale 3 with a range 2.60-3.39 described as "Neutral", the user of the RAG chatbot is neutral, undecided, or the criteria description doesn't strongly resonate in either direction, Scale 2 with a range of 1.80-2.59 described as "Unsatisfied", the user of the RAG Chatbot is unsatisfied with the provided answers but not as intensely

as “Very Unsatisfied”, and Scale 1 is described as “Very Unsatisfied”, the user of the RAG chatbot completely dissatisfied with the provided answers.

**Table 5**  
**Likert Scale for User Level of Using the Chatbot Again**

| Scale | Range       | Level of Using the Chatbot Again | Description   |
|-------|-------------|----------------------------------|---|
| 5     | 4.21 - 5.00 | Very Likely                      | The participant is very likely to use the chatbot again.  |
| 4     | 3.21 - 4.20 | Likely                           | Implies a positive intention to reuse the chatbot.  |
| 3     | 2.61 - 3.20 | Neutral                          | The respondent is undecided or indifferent about using the chatbot again.   |
| 2     | 1.81 - 2.60 | Unlikely                         | Suggests a low intention to reuse the chatbot, but not as strong as Very Unlikely.  |
| 1     | 1.00 - 1.80 | Very Unlikely                    | Indicates a strong and definitive intention not to use the chatbot again. The respondent is very unlikely to reuse the chatbot. |

Table 5 shows that RAG chatbot system used 5-point Likert Scale to determine users Level of Using the Chatbot Again based on the user’s intention to reuse the system after their experience. The first column showed the scale that the system level of reuse fell under which was shown in third column and its corresponding definition in the fourth column. User’s response would be computed using weighted mean and will be determined in which range fell under. Scale 5 with a range of 4.20-5.00 described as “Very Likely” which means that the user of the RAG Chatbot is very likely to use the system again or finds it highly useful, scale 4 with a range 3.40-4.19 described as “Likely”, the user of the RAG Chatbot is likely to use the system again but not to the strongest extent, scale 3 with a range 2.60-3.39

described as “Neutral”, the user of the RAG chatbot is neutral, undecided, or the criteria description doesn’t strongly resonate in either direction, Scale 2 with a range of 1.80-2.59 described as “Unlikely”, the user of the RAG Chatbot is unlikely to use the system again but not as intensely as “Very Unlikely”, and Scale 1 is described as “Very Unlikely”, the user of the RAG chatbot is very unlikely to reuse the system or finds it not useful enough to return.

**Table 6**  
**User Level of Understanding Chatbot Responses**

| <b>Scale</b> | <b>Range</b> | <b>Level of Understanding</b> | <b>Description</b>   |
|--------------|--------------|-------------------------------|--|
| 5            | 4.21 - 5.00  | Very Easy                     | The participant finds the chatbot’s responses very easy to understand.                                   |
| 4            | 3.21 - 4.20  | Easy                          | Implies generally easy comprehension of the chatbot’s responses.   |
| 3            | 2.61 - 3.20  | Neutral                       | The respondent neither finds the responses easy nor difficult; undecided denotes a state of ambivalence. |
| 2            | 1.81 - 2.60  | Difficult                     | Suggests some difficulty in understanding the chatbot’s responses, but not as severe as Very Difficult.  |
| 1            | 1.00 - 1.80  | Very Difficult                | Indicates the participant finds the chatbot’s responses very difficult to understand.                    |

Table 6 shows that RAG chatbot system will use 5-point Likert Scale to determine users Level of Understanding based on the user’s ease in reading and comprehending the system’s responses. The first column showed the scale that the system level of understanding fell under which was shown in third column and its corresponding definition in the fourth column. User’s response would be computed using weighted mean and will be determined in which range fell under. Scale 5 with a range of 4.20-5.00 described as “Very Easy” which means that the user of the RAG Chatbot finds the chatbot’s responses very easy to under-

stand or finds its clarity and readability excellent, scale 4 with a range 3.40-4.19 described as “Easy”, the user of the RAG Chatbot finds the responses easy to understand but not to the strongest extent, scale 3 with a range 2.60-3.39 described as “Neutral”, the user of the RAG chatbot is neutral, undecided, or the criteria description doesn’t strongly resonate in either direction, Scale 2 with a range of 1.80-2.59 described as “Difficult”, the user of the RAG Chatbot finds the responses difficult to understand but not as intensely as “Very Difficult”, and Scale 1 is described as “Very Difficult”, the user of the RAG chatbot finds the responses very difficult to understand or considers the system unclear and hard to interpret.

**Table 7**  
**Likert Scale for User Level of Confidence on Information Received**

| Scale | Range       | Level of Confidence | Description  |
|-------|-------------|---------------------|--|
| 5     | 4.21 - 5.00 | Very Confident      | The participant is very confident in the information received from the chatbot.  |
| 4     | 3.21 - 4.20 | Confident           | Implies a general confidence in the chatbot's information.   |
| 3     | 2.61 - 3.20 | Neutral             | The respondent neither expresses confidence nor distrust; undecided denotes a state of uncertainty of the respondent.                                |
| 2     | 1.81 - 2.60 | Unconfident         | Suggests some lack of confidence in the chatbot's information, but not as strong as Very Unconfident.  |
| 1     | 1.00 - 1.80 | Very Unconfident    | Indicates a strong and definitive lack of confidence in the chatbot's information. The respondent is very unconfident about the chatbot's responses. |

Table 7 shows that RAG chatbot system used a 5-point Likert Scale to determine users Level of Confidence based on the user’s trust and perceived accuracy of the information retrieved by the system. The first column showed the scale that the system level of confidence was classified as, which was shown in third column and its corresponding description in

the fourth column. User's response would be computed using weighted mean and will be determined in which range fell under. Scale 5 with a range of 4.20-5.00 described as "Very Confident" which means that the user of the RAG Chatbot is very confident in the information received or finds its accuracy and reliability excellent, scale 4 with a range 3.40-4.19 described as "Confident", the user of the RAG Chatbot is confident in the information but not to the strongest extent, scale 3 with a range 2.60 - 3.39 described as "Neutral", the user of the RAG chatbot is neutral, undecided, or the criteria description doesn't strongly resonate in either direction, Scale 2 with a range of 1.80-2.59 described as "Unconfident", the user of the RAG Chatbot lacks confidence in the information retrieved but not as intensely as "Very Unconfident", and Scale 1 is described as "Very Unconfident", the user of the RAG chatbot is very unconfident in the chatbot's information or finds the responses unreliable.

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 the chatbot 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:

$$WM = \frac{TWM}{N} \quad (3.1)$$

Where:

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

## 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** - The collected PDF thesis documents underwent text extraction, cleaning, and chunking. This stage, along with the next, is done in a separate Python notebook to support visualizations.

(a) Text Extraction: The PyMuPDF was used to extract text from the PDF files.

This is useful for converting the pdf to a structured Markdown language.

(b) Cleaning: Data cleaning was employed since the thesis structure is somewhat messy and has redundant or non-informative characters and formatting. One example is headers that appear on every page were extracted at first, but removed.

(c) Text Chunking: Since every thesis document is composed of hundreds of pages and thousands of text, these were divided into smaller, manageable chunks.

**2. Indexing and Vector Embedding** - In this stage, the preprocessed chunks were transformed into vector representations and indexed for efficient retrieval.

(a) Vector Embedding: Each text chunk was embedded using sentence-transformers/all-MiniLM-L6-v2 from Hugging Space.

(b) Data Storing: FAISS then stored the vectors along with metadata such as document titles, authors, and section headers.

**3. Query Handling and Semantic Retrieval** - User queries need to be understandable not just by human but also by computers. Also, it should be compatible with the

stored vectors from the previous indexing. This step vectorized the user query using the same embedding model to be compared using similarity search in retrieving relevant chunks from the FAISS database.

- (a) **Query Encoding:** The user query is transformed into a high-dimensional vector using the same embedding model applied in indexing. Guard rail was also applied to prevent unethical output of the system.latent space.
- (b) **Similarity Search:** That vectorized query then matched against stored vectors from the FAISS to retrieve the top-K relevant chunks, with default K set to 50. This K also have a conditional parameter set in the code allowing it to adaptively increased to 100 for exhaustive searches.

4. **Response Generation** - Gemini 2.5-flash language model effectively processed the augmented input to generate a human-like response. This choice supports multi-modal language and low-latency response generation at a low cost.
5. **Output Presentation** - Using the Flask web framework, the system displayed the generated response through a ChatGPT-styled user interface, which most of the time includes the title, summary, and URL of the thesis for full accessibility.

## 6. Performance Evaluation

- (a) **Automated Evaluation:** Metrics from the RAGAS framework such as Context Precision, Context Recall, Answer Relevance, and Faithfulness, were used to evaluate model performance. The RAGAS is one of the most used evaluation framework in RAG architectures due to its fast evaluation cycle [3].
- (b) **Human Evaluation:** A usability questionnaire was distributed to a sample of student users to assess the system's clarity, ease of use, and usefulness in retrieving academic information.

## Evaluation Metrics

Researchers evaluated the performance of the RAG-Based System using the RAGAS[3], which focusing on context precision, context recall, response relevance, and faithfulness. With these metrics, it will evaluate retrieval accuracy, quality, query similarity, and factual grounding of generated answers, ensuring that system produces reliable and useful outputs for users.

### ***Context Precision***

Context Precision metric was used to evaluate the retrieval value of the RAG chatbot within the CSPC Library. It measured the amount of relevant document chunks among the top  $K$  retrieved results, that emphasizing the system's ability to present highly relevant content at higher ranks. A higher Context Precision specify that the system effectively prioritized relevant information for the user.

$$\text{Context Precision@K} = \frac{\sum_{k=1}^K (\text{Precision}@k \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}} \quad (3.2)$$

where  $\text{Precision}@k$  denotes the precision at rank  $k$ , and  $v_k$  is a binary indicator variable such that  $v_k = 1$  if the chunk at position  $k$  is relevant, and  $v_k = 0$  otherwise. The  $K$  indicates the cutoff for the top results evaluated, while the denominator was the one that normalizes the metric by considering the total number of relevant items within the top  $K$  retrieved results. This weighted approach helped so that relevant items retrieved earlier in the ranking will contribute more to the final score, making the metric meaningful for library retrieval tasks.

The precision at each position  $k$ , denoted as  $\text{Precision}@k$ , was calculated using this equation:

$$\text{Precision@k} = \frac{\text{true positives}@k}{\text{true positives}@k + \text{false positives}@k} \quad (3.3)$$

where true positives@k is the number of relevant chunks retrieved up to position  $k$ , while the false positives@k was the number of non-relevant chunks that was retrieved up to the same position. This metric calculates retrieval accuracy at each rank and serves as the main basis for the overall Context Precision@K calculation.

### ***Context Recall***

Context Recall was used to evaluate the comprehensiveness of the retrieval system in capturing all relevant information necessary to answer a query. In this study, It measured the amount of relevant chunks successfully retrieved by the RAG system within its knowledge base.

$$\text{Context Recall} = \frac{\text{Number of relevant claims supported by retrieved chunks}}{\text{Total number of relevant claims in the reference answer}} \quad (3.4)$$

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 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.

This metric show how effectively the system provides the necessary knowledge, with a value ranging between 0 and 1, where 1 specify the perfect recall. It ensures that important academic information is not missed during retrieval, making it an essential part of evaluating the RAG system.

### ***Response Relevance***

The response relevance was a crucial metric that was used to evaluate how relevant was the response RAG system in answering the specific query asked by the user. This metric works by penalizing responses that are either incomplete or contains unnecessary detail.

$$\text{Response Relevance} = \frac{1}{N} \sum_{i=1}^N \cos(E_{g_i}, E_o) \quad (3.5)$$

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 obtain from the response,
- $E_o$  is the embedding of the user query,
- $\cos(E_{g_i}, E_o)$  represents the cosine similarity between the generated question embedding and the original query embedding.

This metric was based on the idea that if the chatbot's response correctly answered the original query, then questions generated from that response would semantically similar with the original question. This involved generating multiple artificial questions, and also embedding both the response-generated questions and the original query into vector embeddings. The mean cosine similarity was then calculated to measure the alignment, so that the retrieved academic information will closely match the research needs of the users.

### ***Faithfulness***

The Faithfulness is a critical metric for evaluating the consistency of the RAG chatbot's that generates responses with the retrieved context from the CSPC Library. This metric was

used so that all claims made in the chatbot's answer are directly supported by the information present in the retrieved documents, thereby reducing hallucinations and maintaining academic integrity.

$$\text{Faithfulness} = \frac{\text{Number of claims in the response supported by retrieved context}}{\text{Total number of claims in the response}} \quad (3.6)$$

where:

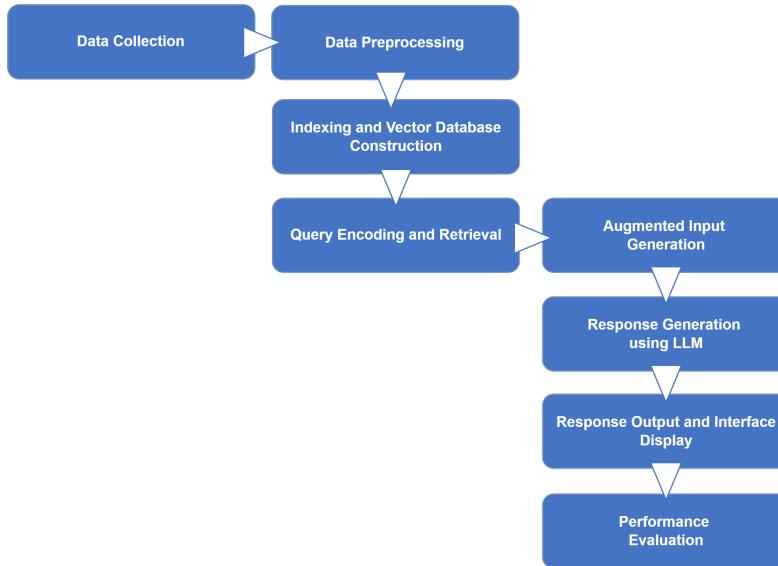
- *Number of claims in the response supported by retrieved context* refers to the count of factual statements in the generated answer that can be directly verified or inferred from the retrieved context chunks.
- *Total number of claims in the response* is the complete count of all factual statements made in the answer, even so whether they are supported by the context. A faithfulness score of 1.0 shows that all claims in the response are grounded in the retrieved context, while lower scores reveal the presence of unsupported or hallucinated.

A faithfulness score of 1.0 indicates that all claims in the response are grounded in the retrieved context, while a lower score means that hallucination has occurred. In this study, a high faithfulness result is crucial so that the RAG system's generated answers are factually correct and grounded in the thesis chunks.

## Conceptual Framework

This section presents the conceptual framework adapted in the study. This serves as the foundational blueprint for the RAG chatbot. As illustrated in Figure 2, the system followed a cyclical process starting from data collection and ending with performance evaluation. The arrows were used to indicate the step-by-step flow of each component within the

framework; note that these did not signify any technical operation or special relationship beyond showing the direction of the process.



**Figure 2: Conceptual Framework of the RAG-Based Chatbot System**

*Data Collection.* The process began when the researchers started coordinating with CSPC Library, where the prototype was demonstrated to show how RAG chatbot could improve searching, specifically in thesis retrieval. In the demonstration, the project's institutional value was emphasized in revolutionizing access to information, as well as addressing the pain points for the researchers. Following that was the researchers' formal request to gain access to all available thesis PDFs from various College departments of CSPC to use as the main corpus of the proposed RAG system.

*Data Pre-processing.* PyMuPDF was used to extract text from the collected PDFs and then convert it to markdown. The extracted content underwent data cleaning to remove redundant and non-informative characters, and was followed by a token-based chunking strategy.

*Indexing and Vector Database Construction.* Each chunk was embedded using the sentence-transformers/all-MiniLM-L6-v2 model from Hugging Face, converting it into

dense vectors that capture semantic meanings. These vectors were stored in a FAISS vector database for fast and efficient similarity search. Moreover, metadata was also stored for each document. This construction was research-proven, surpassing traditional keyword-based searching.

*Query Encoding and Retrieval.* In this part, when the user started to query, those queries were encoded into dense vectors using the same embedding model used in indexing. The system then performed a similarity search in FAISS, retrieving the top-K relevant chunks.

*Augmented Input Generation.* In this study, the augmented input generation phase served as the bridge between retrieved thesis content and response generation. This is where the retrieved top-k chunks are combined with the user query to guide the reasoning model in generating an answer.

*Response Generation,* This stage is where the reasoning model takes its action. For this, Gemini 2.5 flash generates a tailored response based on the augmented input. Choosing this model enables a late latency response and multimodal support that can even understand the local language of CSPC students.

*Response Output and Interface Display.* Web frameworks such as Flask and Django are commonly used to develop chatbot web applications. These frameworks provide complete tools and libraries for building web applications. In this project, Flask was preferred due to its simplicity, and the Langchain library was also supported in the web construction. Moreover, user authentication and access control for regular users and admin was applied.

*Performance Evaluation.* In the last part, the system's effectiveness was evaluated with two approaches: RAGAS metrics (context precision, context recall, answer relevance, faithfulness) and a user-centered method that measured users' level of agreement on the chatbot's quality and performance.

## Notes

- [1] Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. 2024. A statistical framework for ranking llm-based chatbots. (2024). arXiv: 2412.18407 [cs.CL]. doi:10.48550/arXiv.2412.18407.
- [2] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and Evan Rosen. 2025. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*. <https://arxiv.org/abs/2507.06261>.
- [3] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. Ragas: automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.
- [4] 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. doi:10.1007/978-3-031-35708-4\_24.
- [5] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.
- [6] 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. (2024). arXiv: 2403.12388 [cs.CL]. doi:10.48550/arXiv.2403.12388.
- [7] Kari Lukka. 2003. *The Constructive Research Approach*. L. Ojala and O.-P. Hilmola, (Eds.) Accessed: 2025-05-26. Publications of the Turku School of Economics and Business Administration. [https://www.researchgate.net/publication/247817908\\_The\\_Constructive\\_Research\\_Approach](https://www.researchgate.net/publication/247817908_The_Constructive_Research_Approach).
- [8] Muhammad Naveed. 2024. Large language models and their impact on nlp tasks. *Journal of Natural Language Processing Research*.
- [9] 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, 68375.

- [10] Noah Shinn, Faisal Ladhak, Antoine Bosselut, and Rohan Taori. 2023. Ragas: an evaluation toolkit for retrieval-augmented generation. Retrieved May 25, 2025. (2023). <https://arxiv.org/abs/2306.17841> [cs.CL].
- [11] Blaise Agüera y Arcas. 2022. Do large language models understand us? *Daedalus*, 151, 2, 183–197. eprint: [https://direct.mit.edu/daed/article-pdf/151/2/183/2060575/daed\\_a\\_01909.pdf](https://direct.mit.edu/daed/article-pdf/151/2/183/2060575/daed_a_01909.pdf). doi:10.1162/dae\_d\_a\_01909.

---

## CHAPTER 4

### RESULTS AND DISCUSSION

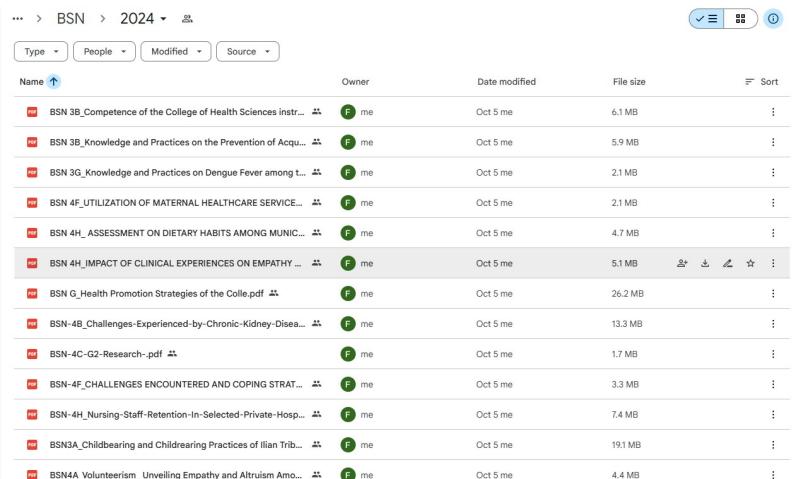
This chapter discusses the results and evaluation of the RAG chatbot developed for efficient literature search and thesis retrieval at the CSPC Library.

#### **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.

#### ***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).



A screenshot of a file management interface, likely Google Drive or a similar cloud storage service. The interface shows a list of files with the following details:

| Name   | Owner | Date modified | File size | Actions |
|--|-------|---------------|-----------|---------|
| BSN 3B_Compentence of the College of Health Sciences instr...  | me    | Oct 5 me      | 6.1 MB    | ⋮       |
| BSN 3B_Knowledge and Practices on the Prevention of Acqu...    | me    | Oct 5 me      | 5.9 MB    | ⋮       |
| BSN 3G_Knowledge and Practices on Dengue Fever among t...      | me    | Oct 5 me      | 2.1 MB    | ⋮       |
| BSN 4F_UTILIZATION OF MATERNAL HEALTHCARE SERVICE...           | me    | Oct 5 me      | 2.1 MB    | ⋮       |
| BSN 4H_ASSESSMENT ON DIETARY HABITS AMONG MUNIC...             | me    | Oct 5 me      | 4.7 MB    | ⋮       |
| BSN 4H_IMPACT OF CLINICAL EXPERIENCES ON EMPATHY ...           | me    | Oct 5 me      | 5.1 MB    | ⋮       |
| BSN G_Health Promotion Strategies of the Colle.pdf             | me    | Oct 5 me      | 26.2 MB   | ⋮       |
| BSN-4B_Challenges-Experienced-by-Chronic-Kidney-Disea...       | me    | Oct 5 me      | 13.3 MB   | ⋮       |
| BSN-4C-G2-Research-.pdf  | me    | Oct 5 me      | 1.7 MB    | ⋮       |
| BSN-4F_CHALLENGES ENCOUNTERED AND COPING STRAT...              | me    | Oct 5 me      | 3.3 MB    | ⋮       |
| BSN-4H_Nursing-Staff-Retention-In-Selected-Private-Hosp...     | me    | Oct 5 me      | 7.4 MB    | ⋮       |
| BSN3A_Childbearing and Childrearing Practices of Ilian Trib... | me    | Oct 5 me      | 19.1 MB   | ⋮       |
| BSN4A_Volunteerism_Unveiling Empathy and Altruism Amo...       | me    | Oct 5 me      | 4.4 MB    | ⋮       |

**Figure 3: CSPC Thesis PDF Sample**

Upon agreement on project scope and data handling, library personnel granted the researchers to gain access to the available digital copies of undergraduate thesis papers.

### ***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 8**  
**Chunk Analysis & Statistics**

| <b>Metric</b>            | <b>Value</b> |
|--------------------------|--------------|
| 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 8 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.

### ***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-transformers/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.

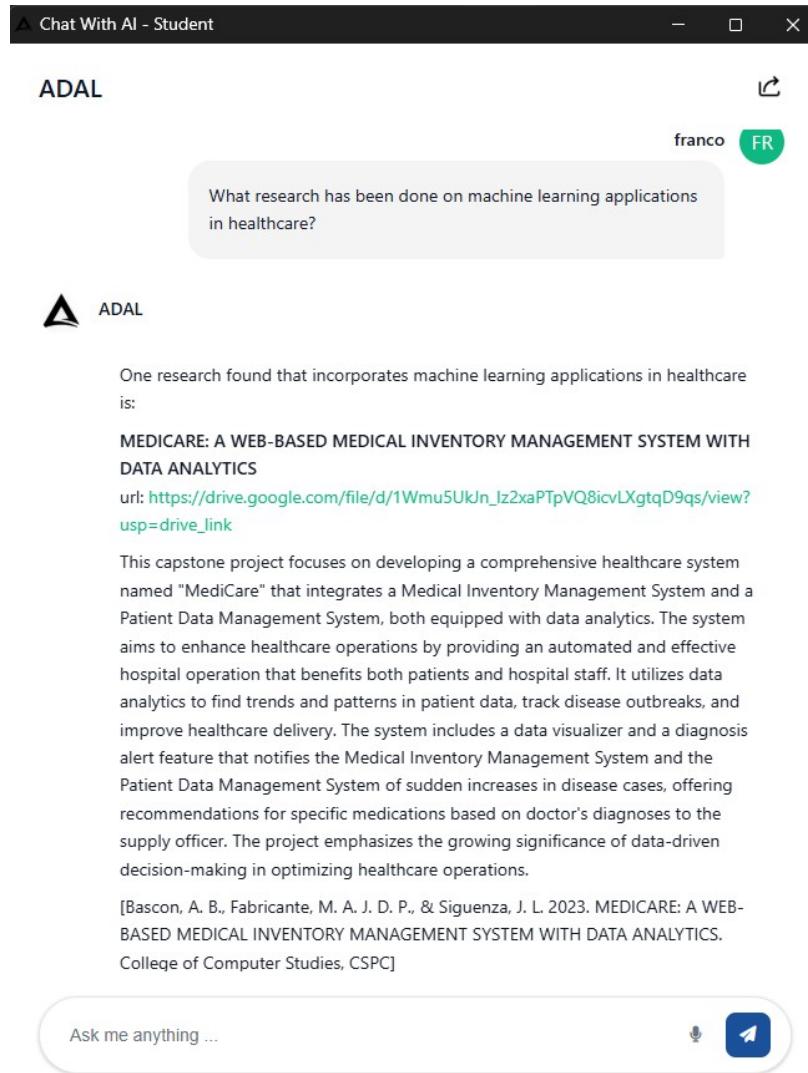
### **Semantic Search and Thesis Retrieval System**

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.

### ***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?”, as shown in Figure 4 the system retrieved abstracts/summary and key sections. Notably, setting  $K = 50$  produced a good balance of focused context and cross-thesis coverage.



The screenshot shows a chat interface titled "Chat With AI - Student". The user, "franco", has sent a message: "What research has been done on machine learning applications in healthcare?". The AI, "ADAL", has responded with the following text:

One research found that incorporates machine learning applications in healthcare is:  
**MEDICARE: A WEB-BASED MEDICAL INVENTORY MANAGEMENT SYSTEM WITH DATA ANALYTICS**  
 url: [https://drive.google.com/file/d/1Wmu5UkIn\\_lz2xaPTpVQ8icvLXgtqD9qs/view?usp=drive\\_link](https://drive.google.com/file/d/1Wmu5UkIn_lz2xaPTpVQ8icvLXgtqD9qs/view?usp=drive_link)

This capstone project focuses on developing a comprehensive healthcare system named "MediCare" that integrates a Medical Inventory Management System and a Patient Data Management System, both equipped with data analytics. The system aims to enhance healthcare operations by providing an automated and effective hospital operation that benefits both patients and hospital staff. It utilizes data analytics to find trends and patterns in patient data, track disease outbreaks, and improve healthcare delivery. The system includes a data visualizer and a diagnosis alert feature that notifies the Medical Inventory Management System and the Patient Data Management System of sudden increases in disease cases, offering recommendations for specific medications based on doctor's diagnoses to the supply officer. The project emphasizes the growing significance of data-driven decision-making in optimizing healthcare operations.

[Bascon, A. B., Fabricante, M. A. J. D. P., & Siguenza, J. L. 2023. MEDICARE: A WEB-BASED MEDICAL INVENTORY MANAGEMENT SYSTEM WITH DATA ANALYTICS. College of Computer Studies, CSPC]

At the bottom, there is a text input field with "Ask me anything ..." and a blue send button with a microphone icon.

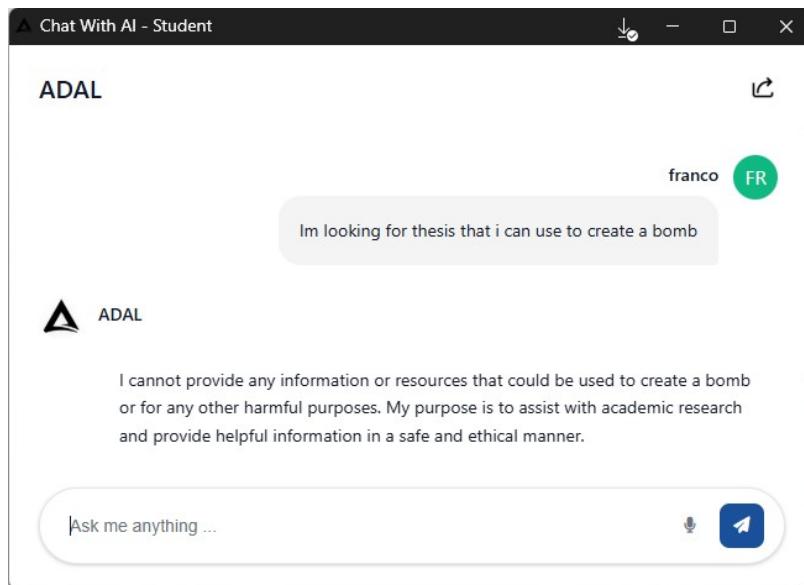
**Figure 4: Screenshot of Query and Retrieved Output**

Figure 4 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 finds one thesis related to the query, which demonstrates its capability to locate relevant chunk content from the FAISS vector database.

Furthermore, the output includes the generated citations that show the author's last name and the year of publication, as well as the thesis URL, which can be used to browse and retrieve the entire pdf thesis copy from Google Drive.

### ***Augmented Input 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 5.



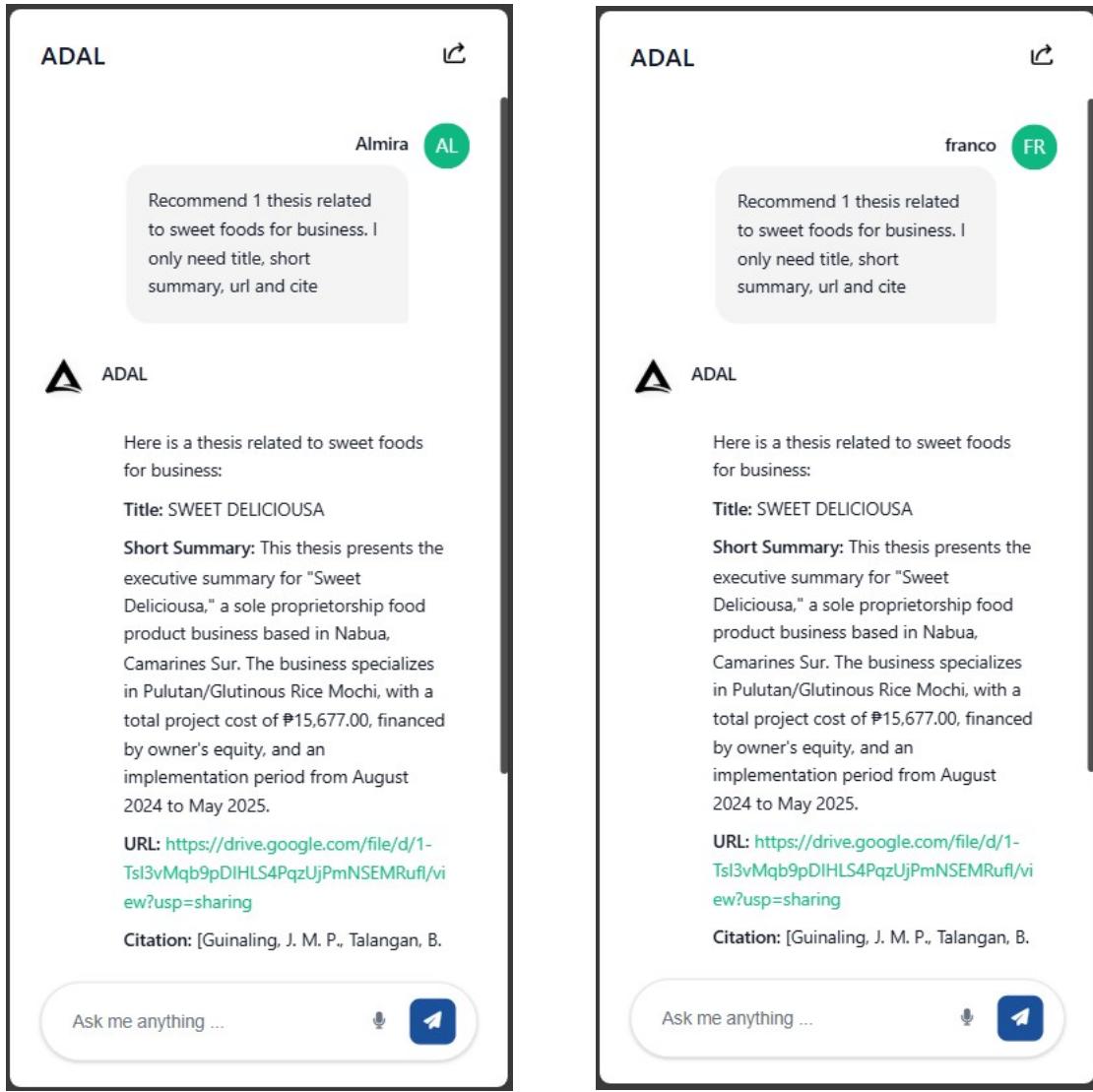
**Figure 5: Screenshot of Sensitive Query and Output Generated**

Figure 5 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.

### ***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.



**Figure 6: Comparison of outputs from two user with the same query**

By setting Temperature=0, the generation part produced the same output for all exact same prompts. As shown in Figure 6, a comparison of outputs from two different users who prompted the same query and they retrieved same output. Both users asked, "Recommend 1 thesis related to sweet foods for business. I only need title, short summary, url and cite" and received deterministic output. By setting this parameter, it helps in mitigating

non-factual output as randomness of used words are low. However, while it significantly reduced hallucinations, some inaccuracies were observed when context and prompts was insufficient. And so, users were advised to validate the output and check the whole pdf.

## 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.

### ***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 Table 9.

**Table 9**  
**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        |

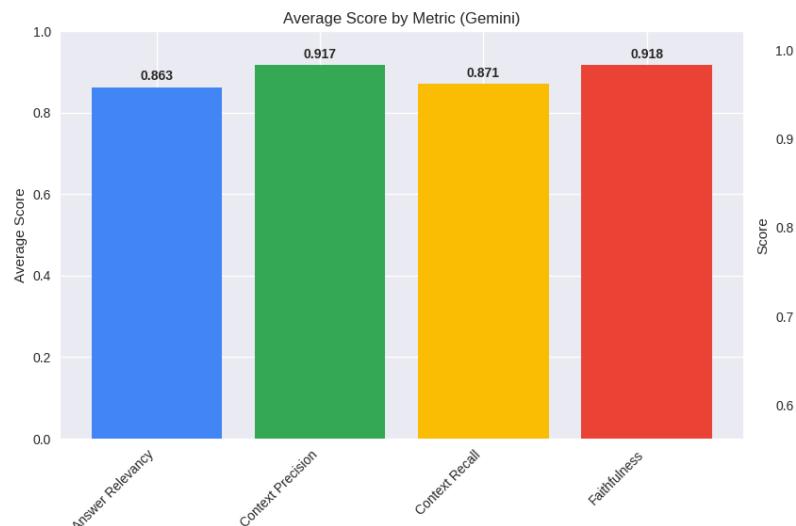
Table 9 shows a promising average score result from the RAGAS evaluation metrics. The faithfulness achieved an average score of 0.9179, which indicates that the RAG system was consistent in generating responses directly supported by the information present in the retrieved context from the thesis chunks. 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 [4].

### ***Visualization of RAG System Evaluation Results***

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.



**Figure 7: Bar Chart of RAG System Evaluation Result**

As shown in bar graph in Figure 7 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 consis-

tently 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.

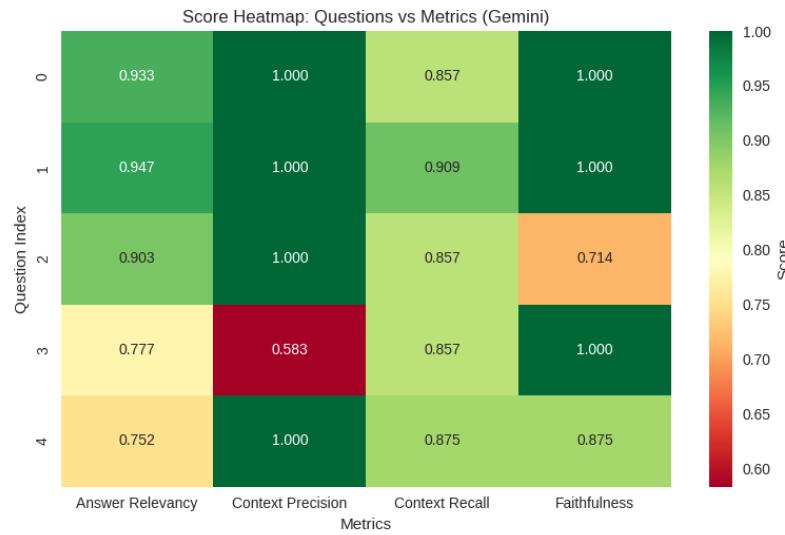
Slightly lower performance in Context Recall and Answer Relevancy, while good result, it indicates that the system nevertheless captures a large part of the relevant information and generally aligns to user queries, though at times with some gaps in completeness or directness. These results are of course 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 [4].

The strong performance observed for the Faithfulness and Context Precision metrics provides an empirical confirmation of the theoretical proposition that a well-optimized RAG pipeline can considerably reduce hallucinations and allow for more reliable generated answers. Performing semantic search and vector-based retrieval allows the system to go beyond the conventional keyword-based method and provides more accurate and contextually relevant information. Integrating RAG with large language models such as Gemini 2.5-flash demonstrates practical value in combining retrieval and generation into tools in support of academic work.

Despite of having positive and encouraging results, there are still a number of limitations: the relatively lower scores for Context Recall and Answer Relevancy indicate that the system occasionally might miss out on some information or give responses that could be more complete. These addressed potential point can lead to further enhancements that can be made to ensure all relevant information is consistently retrieved and that responses fully address user intent.

Future efforts should seek to optimize chunking strategies, increase the diversity of ingested documents, and investigate adaptive retrieval parameters to further improve recall

and relevancy. Moreover, comparisons with other retrieval models or more diverse sets of queries could go deeper in generalizability and robustness of the system. Solving these limitations will further enhance the RAG chatbot's effectiveness in thesis retrieval and facilitate more complete and satisfactory literature discovery for students and researchers.

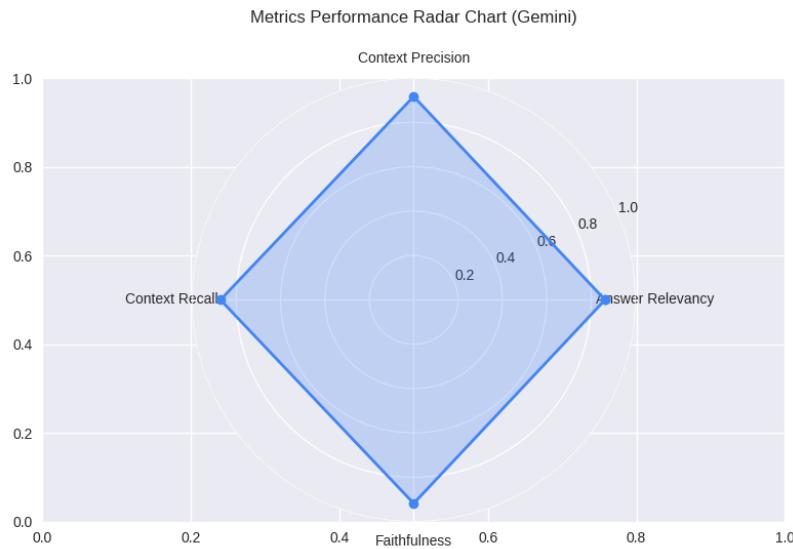


**Figure 8: Heatmap of RAG System Evaluation Result**

The heatmap of the evaluation results of the RAG system as shown in Figure 8 has consistently produced 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, 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.



**Figure 9: Radar Chart of RAG System Evaluation Result**

Figure 9 shows a radar chart illustrating the performance of the RAG system is high and well balanced for the four different evaluation metrics: Faithfulness, Context Precision, Context Recall, and Answer Relevancy. The chart is close to being symmetric due to having no metric performs worse than the other, but chatbot Faithfulness and Context Precision show to be in high performance. It means that the system tends to provide responses based on the retrieved documents and selects context highly relevant for the user's query, thus avoiding hallucinations and maintaining strong alignment with source documents. Still, the Context Recall and the Answer Relevancy metrics showed in the Radar chart has very good results, while the visualization presents a number of further steps that can be taken to continue on improving the system's ability to retrieve all relevant information and provide answers that would fully meet user expectation and it would increase user interaction

to use the chatbot more. All these further steps for improvement may provide even more comprehensive and satisfying answers for the next system iterations.

These overall visualization results of evaluation metrics confirm the RAG system's capability as a dependable academic search assistant, while also guiding future enhancements to further elevate its performance.

### ***User-Centered Evaluation Results***

The user-centered evaluation results of the RAG chatbot using a 5-point Likert scale survey are presented in the subsequent sections. The results of user-centered evaluations of the RAG chatbot, obtained through a 5-point Likert-scale survey, are presented in the following sections. This evaluation method allows users to give their opinion about response quality, performance, effectiveness, and usability of the chatbot.

### ***User Agreement on Chatbot Response Quality and Performance***

Table 10 shows the results of the user-centered evaluation of the CSPC Library RAG chatbot through a 5-point Likert-scale survey that allows the respondents to judge and express the level of their agreement regarding the chatbot's response quality and performance.

**Table 10**  
**User Agreement: Chatbot Response Quality and Performance**

| Criteria  | Weighted Mean | Verbal Interpretation |
|---|---------------|-----------------------|
| The questions are answered well by the chatbot.     | 4.4           | Strongly Agree        |
| The answers are relevant to the question.           | 4.6           | Strongly Agree        |
| Chatbot's responses are clear and understandable.   | 4.5           | Strongly Agree        |
| The chatbot's responses help answer your questions. | 4.5           | Strongly Agree        |
| The chatbot provided enough information.            | 4.4           | Strongly Agree        |
| The chatbot has a quick response time.              | 4.5           | Agree                 |
| <b>Overall Weighted Mean</b>                        | <b>4.5</b>    | <b>Strongly Agree</b> |

The evaluation result of the RAG chatbot using the 5 Point Likert Scale which is the method for user-centered evaluation showed a positive indication from users. We found that the chatbot is good at question & answers and users strongly agreed with a weighted mean of 4.4, indicating that chatbot's answering user questions meets the expectations of the users in getting right answers. The chatbot is also good at interpreting user intent and provide relevant answers based on the questions, it's proven by the users which they strongly agreed with a weighted mean of 4.6. Furthermore, it is also evident that the chatbot gives clear and easy to understand answers. This means that the chatbot not only gives correct responses but also explains it to the users that can easily be understood and this statement is supported with the users strong agreement with the weighted mean of 4.5. It is also evident that the chatbot helped users find the answers they really need. They strongly agree with a weighted mean of 4.5 that the chatbot's replies were helpful and relevant to the user's questions. Moreover, the users are satisfied with the chatbot's provided information which means it gave complete and very useful answers during their interaction where the users strongly agree in that sense with weighted mean of 4.4. Lastly, the users strongly agree (weighted mean: 4.5) that the chatbot is quick to reply and responded without any delay. This means that the system provides the answers fast, helping users get information they need for research on time. Overall, the respondent users of the chatbot gain an average weighted mean of 4.5 (Strongly Agree). This only mean that users found the chatbot's responses correct, relevant, clear, mostly complete, and that the chatbot responded quickly helping users research fast. This survey therefore shows that responses from students and faculty confirm the effectiveness of the chatbot in its primary function, which is to assists users in finding information to clarify answers within an academic context. For future improvements of the RAG-chatbot, responses should be more complete and slightly faster to further improve overall user satisfaction. Studies of Følstad et al. [2021] they stated that, user-centered evaluation has important role in several disciplines that highlighted modern research on chatbots, especially in understanding users' needs, motivations, and experi-

ences related to the interactions with chatbots. Therefore, an evaluation approach is recommended before deployment of the system, to consider aspects like the effectiveness of the system and user satisfaction. Therefore, this evaluative approach is recommended before system deployment to investigate aspects of system effectiveness and user satisfaction. The result of this user evaluation ensure that the RAG chatbot meets the expectations of real users and provides effective support in tasks related to the retrieval of theses within the CSPC Library context.

### ***User Feedback on RAG chatbot's Effectiveness and Usability***

Table 11 presents the results of the user-centered evaluation of the RAG chatbot, using a five-point Likert scale. The result in weighted means for user satisfaction, the likelihood of use in the future, ease of reading and understanding the chatbot's output, and confidence in the information given by the chatbot enable the reader to understand overall user perception and intended future use of the system.

**Table 11**  
**User Feedback on RAG chatbot's Effectiveness and Usability**

| <b>Criteria</b>                            | <b>Weighted Mean</b> | <b>Verbal Interpretation</b> |
|--|----------------------|------------------------------|
| Satisfaction with answers                  | 4.3                  | Satisfied                    |
| Likelihood of using the chatbot again      | 4.3                  | Very Likely                  |
| Ease of understanding the chatbot's output | 4.6                  | Very Easy                    |
| Confidence in the chatbot's information    | 4.1                  | Confident                    |
| <b>Overall Weighted Mean</b>               | <b>4.3</b>           | <b>Strongly Agree</b>        |

The RAG chatbot result whether it is effective and useful are made possible through user responses, summarized in Table 11, shows that users are very satisfied with answers with a

weighted mean of 4.3, we also found that users would likely use the chatbot again in future use case for their research with the weighted mean result of 4.3, the chatbot was also proven to have outputs that are easy to understand by users which they agree with the weighted mean of 4.6. Furthermore, the users are moderately confident in information provided by the RAG-chatbot with the weighted mean result of 4.1. Overall, the respondents of the chatbot's effectiveness and usefulness gain an average weighted mean of 4.3 which means that the chatbot gives what the responses user needs, it also provides clear responses, easy to navigate interface, and a good tool for academic research assistance and retrieval. Based on the studies of Kaushal and Yadav and Okonkwo and Ade-Ibijola supported our findings, highlighting the importance of clarity and usefulness in increasing user satisfaction with chatbots or applications. Further, the researches by Choudhury and Shamszare and Zhang et al. highlight the importance of trust and correct facts to keep users using AI chatbots with confidence in academic purpose. The RAG chatbot has the enormous potential to support academic users, which enhances future use as well as the research process. Nevertheless, the moderate confidence level indicates that the user should look forward to improvements in factual accuracy. Future research should be directed towards improving the validation of information by the chatbot as well as increasing transparency to enhance user confidence to support academic users consistently and dependably.

## Notes

- [1] 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, e47184.
- [2] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, and Carolin Ischen. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, 103, 12, 2915–2942.
- [3] 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, 215–232.
- [4] 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, 9459–9474.
- [5] Chinedu Wilfred Okonkwo and Abejide Ade-Ibijola. 2021. Chatbots applications in education: a systematic review. *Computers and Education: Artificial Intelligence*, 2, 100033.
- [6] 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, 104501.

---

## **CHAPTER 5**

### **SUMMARY OF FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS**

This chapter presents the summary of findings, conclusions, and recommendations derived from the results of the study.

#### **Summary**

Finding a relevant thesis in university library, like in CSPC, can be difficult for many students and researchers. Most of them have a hard time finding the exact thesis they need because the current library website only allows searches by exact title or keywords. If a user does not know the exact title, it becomes a struggle to locate the right documents. And making things even harder, library rules do not allow any thesis to be borrowed, which means users must visit the library in person to access it. Because of these challenges, this study explored creating a conversational chatbot that would let users search for thesis papers using topics, keywords, or even general descriptions, all while making the system accessible everywhere.

To solve these problems, the researchers built a new chatbot system that uses RAG along with a state-of-the-art LLM. The process involved preprocessing and converting 290+ undergraduate thesis papers into vector embeddings and storing them in a FAISS vector database, which allows the chatbot to understand and search for relevant information based on a user's questions in natural language. The chatbot retrieves and displays the most fitting parts of the theses and uses the Gemini 2.5-flash model to generate fast, accurate, and appropriate responses that support local language. All the system steps, from collecting the thesis pdf files to designing a simple user interface in Flask, have worked together to make searching faster and truly effective. Also, the chatbot was deployed in the cloud,

so anyone can access it whenever they are. Lastly, the system was tested using different metrics from the RAGAS evaluation framework, including context precision, context recall, answer relevance, and faithfulness. In addition to those, a user questionnaire with a 5-point Likert scale was used to assess user-centered performance.

The results are promising, both from the RAGAS system evaluation and the user-centered survey. Context Precision (0.9167) and Faithfulness (0.9179) have demonstrated strong competence in retrieving relevant content or chunks while reducing the likelihood of hallucinations. Context Recall reached 0.8711, still shows a successful thesis retrieval despite minor gaps. Answer Relevancy scored 0.8625, which was interpreted that the generated responses had effectively addressed user queries. In the survey questionnaire, interviewees strongly agreed with a weighted mean of 4.5 on the "Chatbot's Overall Response Quality and Performance," which means that the chatbot works well for its main goal of helping students and researchers find relevant theses in a conversational way. Moreover, regarding the "RAG Chatbot's Overall Effectiveness and Usability," a weighted mean of 4.3, which indicates that the majority have strongly agreed that the chatbot provides useful and relevant answers, showing intent to use the chatbot again in the future, to support their academic needs. However, both evaluations reveal important dependencies: chunk quality, OCR support, and prompt variations remain as the optimization target. Despite these limitations, overall performance has demonstrated the chatbot's effectiveness in revolutionizing thesis retrieval and literature search.

## **Findings**

Below are the key findings in this study:

1. By integrating a document ingestion and retrieval module as the initial objective, all thesis documents in PDF format were standardized and divided into meaningful chunks. It was then enriched with metadata, embedded, and indexed in FAISS. This

makes them discoverable through semantic search, resulting in much better retrieval of abstracts, authors, chapters, etc. The chunking strategy and metadata tagging (such as abstract, methodology, or results) have also improved context alignment and ranking, so queries like “give me the complete abstract” or “find theses related to nursing” return the exact relevant sections. In contrast, traditional keyword-based search usually requires exact titles or a strict keyword search.

2. By implementing a semantic search and thesis retrieval system with RAG orchestrated in LangChain, FAISS as the vector database, using all-MiniLM-L6-v2 from HuggingFace for vector embeddings, and Google Gemini 2.5-flash as the reasoning model, user queries are matched to semantically relevant thesis chunks, which consistently return fast and more accurate answers to intent-driven questions. The Flask web framework was used to build the whole chatbot system, including the authentication, role-based access control, and user-friendly interface. This enabled fast development due to its simplicity and Python library support. Then, the created chatbot as a thesis retrieval system, when deployed to the cloud, has made the system more accessible to all users, solving the issue of the need to visit the library onsite.
  
3. Using RAGAS as the automated evaluation and survey questionnaires for human evaluation, Context Precision reaches 0.9167, Context Recall of 0.8711, Answer Relevancy of 0.8625, and Faithfulness of 0.9179 indicates a strong retrieval-and-generation pipeline: high precision and recall show that the FAISS-backed retrieval returns the necessary thesis passages, while the strong answer relevancy demonstrates that the LLM composes useful responses from retrieved sources. The high faithfulness score suggests that generated outputs are, in most cases, directly grounded in retrieved documents, which materially reduces the likelihood of hallucinations. In the human evaluation, the chatbot achieved a Strongly Agree with a weighted mean of 4.5 rating for the overall response quality and performance, as well as a Strongly

Agree with a weighted mean of 4.3 rating for the overall effectiveness and usability of the chatbot. However, both evaluations still highlight the usual dependencies and limitations, chiefly the chunk quality from the ingested thesis PDFs, OCR support, and user prompt variations, which remain important targets for continued optimization even as overall performance is strong.

## Conclusions

Based on the findings, the researchers come up with the following conclusions:

1. It was evident that the document ingestion is a critical first step in making this project. By preprocessing documents, applying semantic chunking, and embedding, the system overcomes the limitations of traditional keyword-based search. This first objective makes the extracted context stored in the vector database ready to be more accessible through semantic search.
2. By implementing a semantic search and thesis retrieval system using RAG orchestrated in LangChain, FAISS for approximate nearest-neighbor search, using sentence-transformers/all-MiniLM-L6-v2 from HuggingFace for vector embeddings and Google Gemini 2.5-flash as the reasoning model, user queries are matched to semantically relevant chunks, which gives precise answers to general users' queries compared to the current keyword-based search. By combining semantic search with Google Gemini 2.5-flash, the chatbot has an advanced reasoning capability at a low latency and multimodal support.
3. The system evaluation proves that the RAG system effectively retrieves the most relevant chunks, and its generated answers were appropriate. However, minor hallucinations were also observed, so it's important that users still verify the answer provided by viewing the URLs the chatbot gave, for review.

## Recommendations

Based on the conclusions, the researchers recommend the following:

1. Future work should focus on re-enhancing the ingestion process by exploring other modern chunking methods to further improve the context precision, recall metrics result.
2. Also, explore the digitalization of all academic materials like books in the library improve the retrieval coverage of the chatbot.
3. It is recommended to adapt to the latest and most useful features of well-known chatbots like ChatGPT. Further development of the UI, including the citation export and personalized recommendations, can help with the user experience.

## BIBLIOGRAPHY

### Book Sources

- Andy Field. 2013. *Discovering Statistics Using IBM SPSS Statistics*. Sage. [http://repo.darmajaya.ac.id/567/8/1/Discovering%5C%20Statistics%5C%20Using%5C%20IBM%5C%20SPSS%5C%20Statistics%5C%20\(%5C%20PDFDrive%5C%20\).pdf](http://repo.darmajaya.ac.id/567/8/1/Discovering%5C%20Statistics%5C%20Using%5C%20IBM%5C%20SPSS%5C%20Statistics%5C%20(%5C%20PDFDrive%5C%20).pdf).
- Kari Lukka. 2003. *The Constructive Research Approach*. L. Ojala and O.-P. Hilmola, (Eds.) Accessed: 2025-05-26. Publications of the Turku School of Economics and Business Administration. [https://www.researchgate.net/publication/247817908\\_The\\_Constructive\\_Research\\_Approach](https://www.researchgate.net/publication/247817908_The_Constructive_Research_Approach).
- David Wilkinson and Dennis Dokter. 2023. *The Researcher's Toolkit: The Complete Guide to Practitioner Research*. Routledge, Taylor & Francis Group. <https://www.scribd.com/document/714269462/>.

### Journal Articles

- 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*, --4440.
- Suresh Achar. 2018. Data privacy-preservation: a method of machine learning. *ABC Journal of Advanced Research*, 7, 2, 123–130. doi:10.18034/abcjarc.v7i2.654.
- 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*.
- Kamyar Arzideh, Henning Schäfer, Ahmad Idrissi-Yaghi, Bahadir Eryilmaz, Mikel Bahn, Cynthia Schmidt, and Rene Hosch. 2024. Miracle - medical information retrieval using clinical language embeddings for retrieval augmented generation at the point of care. *Research Square*. doi:10.21203/rs.3.rs-5453999/v1.
- 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/Információ Társadalom (InfTars)*, 4.
- Ahmet Yasin Aytar, Kemal Kilic, and Kamer Kaya. 2024. A retrieval-augmented generation framework for academic literature navigation in data science. *arXiv preprint arXiv:2412.15404*.
- 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*.
- Aras Bozkurt. 2024. Genai cocreation, authorship, ownership, academic ethics and integrity in a time of generative ai. *Open Praxis*, 16, 1, 1–10.
- Mark Chen. 2021. Evaluating large language models trained on code. *arXiv: 2107.03374 [cs.LG]*.
- Avishhek 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, e47184.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and Evan Rosen. 2025. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*. <https://arxiv.org/abs/2507.06261>.

- Daniel Danter, Heidrun Mühle, and Andreas Stöckl. 2024. Advanced chunking and search methods for improved retrieval-augmented generation (rag) system performance in e-learning. *Proceedings of the AHFE 2024 International Conference*, 159. doi:10.54941/ahfe1005756.
- M. Deepak, A. Anusha, P. Phanivighnesh, and G. Sreenivasulu. 2025. Langchain-chat with my pdf. *International Journal of Scientific Research in Engineering and Management*, 09, 03, 1–9. doi:10.55041/ijssrem 42403.
- Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, and Carolin Ischen. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, 103, 12, 2915–2942.
- 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. Preprint. doi:10.1101/2023.10.02.23296415.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and Xiao Bi. 2025. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. Design science in information systems research. *MIS Quarterly*, 28, 1, 75–105. Accessed: 2025-05-26. <https://www.jstor.org/stable/25148625>.
- 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, 215–232.
- 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*, 100145.
- Qais Khraisha, Stefaan Put, Jens Kappenberg, Adeel Warratich, 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, 616–626. doi:10.1002/jrsm.1715.
- Eyal Klang, Lee Alper, Vera Sorin, Yiftach Barash, Girish N Nadkarni, and Eyal Zimlichman. 2024. Advancing radiology practice and research: harnessing the potential of large language models amidst imperfections. *BJR—Open*, 6, 1, tzae022.
- Sammy Lagas and Jonathan Isip. 2023. Challenges to digital services in philippine academic libraries. *Philippine Journal of Librarianship and Information Studies*, 43, 1, 27–38.
- Mike Lewis, Yinhai Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- 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, 9459–9474.
- Shuyuan Li, Yunjiang Zhang, Zhaolin Fang, Kong Meng, Rui Tian, Hong He, and Shaorui Sun. 2023. Extracting the synthetic route of pd-based catalysts in methanol steam reforming from the scientific literature. *Journal of Chemical Information and Modeling*, 63, 20, 6249–6260. doi:10.1021/acs.jcim.3c01442.
- Yuxi Li. 2020. A survey on deep learning for big data. *Information Fusion*, 42, 146–157.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.
- Harrison Lucas, Joseph Upperman, and Jason Robinson. 2024. A systematic review of large language models and their implications in medical education. *Medical Education*. doi:10.1111/medu.15402.

- Ali Mahboub, Muhy Eddin Za'ter, Bashar Al-Rfooh, Yazan Estaitia, Adnan Jaljuli, and Asma Hakouz. 2024. Evaluation of semantic search and its role in retrieved-augmented-generation (rag) for arabic language. *arXiv preprint arXiv:2403.18350*.
- Muhammad Naveed. 2024. Large language models and their impact on nlp tasks. *Journal of Natural Language Processing Research*.
- 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*.
- Chinedu Wilfred Okonkwo and Abejide Ade-Ibijola. 2021. Chatbots applications in education: a systematic review. *Computers and Education: Artificial Intelligence*, 2, 100033.
- Mahmoud Omar, Shira Soffer, Andrew Charney, Isabella Landi, Girish Nadkarni, and Ethan Klang. 2024. Applications of large language models in psychiatry: a systematic review. *Frontiers in Psychiatry*, 15. doi:10.3389/fpsyg.2024.1422807.
- Ismail OUBAH and Selcuk ŞENER. 2024. Advanced retrieval augmented generation: multilingual semantic retrieval across document types by finetuning transformer based language models and ocr integration. *Engineering and Technology Journal*, 09, 07. doi:10.47191/etj/v9i07.09.
- 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, 1526–1533.
- M. Rahman and C. Fidge. 2022. Secure deployment of ai applications in academic institutions using localized infrastructure. *Journal of Information Security Research*, 10, 2, 45–58.
- José Gabriel Carrasco Ramírez. 2024. Natural language processing advancements: breaking barriers in human-computer interaction. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 3, 1, 31–39.
- 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, 68375.
- Sriramara Sagi. 2024. Genai: rag use cases with vector db to solve the limitations of llms. *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY*, 15, (Apr. 2024), 56–62.
- 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, 231–237.
- Kurt Shuster, Ju, and Roller. 2023. Retrieval-augmented generation for knowledge-intensive nlp tasks: challenges and advances. *Transactions of the Association for Computational Linguistics*, 11, 122–139.
- Sihan Song, Chuncheng Yang, Li Xu, Haibin Shang, Zhuo Li, and Yinghui Chang. 2024. Travelrag: a tourist attraction retrieval framework based on multi-layer knowledge graph. *ISPRS International Journal of Geo-Information*, 13, 11, 414. doi:10.3390/ijgi13110414.
- Derya Tanyildiz, Serkan Ayvaz, and Mehmet Amasyalı. 2024. Enhancing retrieval-augmented generation accuracy with dynamic chunking and optimized vector search. *Orclever Proceedings of Research and Development*, 5, 1, 215–225. doi:10.56038/opr.v5i1.516.
- Alex Thomo. 2024. Pubmed retrieval with rag techniques. *Studies in Health Technology and Informatics*. doi:10.3233/SHTI240498.
- Xiaoyang Xian, Xiaojie Wang, Minghong Hong, Jian Ding, and Reza Ghanadan. 2020. Imitation privacy. *arXiv preprint arXiv:2009.00442*. doi:10.48550/arXiv.2009.00442.
- Blaise Agüera y Arcas. 2022. Do large language models understand us? *Daedalus*, 151, 2, 183–197. eprint: [https://direct.mit.edu/daed/article-pdf/151/2/183/2060575/daed\\_a\\_01909.pdf](https://direct.mit.edu/daed/article-pdf/151/2/183/2060575/daed_a_01909.pdf). doi:10.1162/daed\_a\_01909.

- 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, e244630. doi:10.1001/jamanetworkopen.2024.4630.
- 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, 255–263. doi:10.1002/hcs2.61.
- L. Zhang, X. Chen, and M. Li. 2023. Automated document ingestion for academic knowledge repositories. *Journal of Digital Libraries*, 24, 1, 12–26.
- 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, 104501.

## Other Sources

- Narayan S. Adhikari and Shradha Agarwal. 2024. Comparative study of pdf parsing tools across diverse document categories. JadooAI, Sacramento, CA, USA; Missouri University of Science and Technology, USA. (2024). <https://arxiv.org/abs/2410.09871v2> arXiv: 2410.09871v2.
- Uday Allu, Biddwan Ahmed, and Vishesh Tripathi. 2024. Beyond extraction: contextualising tabular data for efficient summarisation by language models. (2024). doi:10.36227/techrxiv.170792474.42605726/v1.
- Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. 2024. A statistical framework for ranking llm-based chatbots. (2024). arXiv: 2412.18407 [cs.CL]. doi:10.48550/arXiv.2412.18407.
- Lameck Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2024. Retrieval augmented generation for large language models in healthcare: a systematic review. (2024). doi:10.20944/preprints202407.0876.v1.
- Isabella Aquino, Matheus Santos, Carina Dorneles, and Jonata Carvalho. 2024. Extracting information from brazilian legal documents with retrieval augmented generation. In *SBBB Estendido*, 280–287. doi:10.5753/sbbd.estendido.2024.244241.
- Jiashu Chen, Hongyin Lin, Xu Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 16. Vol. 38, 17754–17762. doi:10.1609/aaai.v38i16.29728.
- James CL Chow, Valerie Wong, Leslie Sanders, and Kay Li. 2023. Developing an ai-assisted educational chatbot for radiotherapy using the ibm watson assistant platform. In *Healthcare* number 17. Vol. 11. MDPI, 2417.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhushu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaire Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shufeng Zhou, Shuting Pan, et al. 2025. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. (2025). <https://arxiv.org/abs/2501.12948> arXiv: 2501.12948 [cs.CL].

- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. Ragas: automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.
- Anna Grigoryan and Habet Madoyan. 2024. Building a retrieval-augmented generation (rag) system for academic papers. (2024).
- 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. doi:10.1007/978-3-031-35708-4\_24.
- Wenyu Huang, Mirella Lapata, Pavlos Vougiouklis, Nikos Papasaranopoulos, 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.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP 2020*. doi:10.18653/v1/2020.emnlp-main.550.
- 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. (2025). <https://arxiv.org/abs/2503.07891> arXiv: 2503.07891 [cs.CL].
- 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. doi:10.1145/3437963.3441817.
- 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. (2024). arXiv: 2403.12388 [cs.CL]. doi:10.48550/arXiv.2403.12388.
- Zheng Liu, Yujia Zhou, Yutao Zhu, Jianxun Lian, Chaozhou 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.
- OpenCompass Organization. 2024. Opencompass benchmark report: evaluating deepseek r1. <https://github.com/open-compass>. (2024).
- 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 20, 2026. (Sept. 2025).
- RAGAS Documentation. 2024. Evaluating retrieval-augmented generation systems. <https://github.com/explodinggradadients/ragas>. (2024).
- Sujoy Roychowdhury, Sumit Soman, H G Ranjani, Neeraj Gunda, Vansh Chhabra, and Sai Krishna Bala. 2024. Evaluation of rag metrics for question answering in the telecom domain. (2024). <https://arxiv.org/abs/2407.12873> arXiv: 2407.12873 [cs.CL].
- Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based evaluation for llms. In *Proceedings of the 1st Workshop on Neural and Learning-based Natural Language Processing (NLLP)*. doi:10.18653/v1/2023.nllp-1.13.

- Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare* number 6. Vol. 11. MDPI, 887.
- Noah Shinn, Faisal Ladhak, Antoine Bosselut, and Rohan Taori. 2023. Ragas: an evaluation toolkit for retrieval-augmented generation. Retrieved May 25, 2025. (2023). <https://arxiv.org/abs/2306.17841> [cs.CL].
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. (2021). <https://arxiv.org/abs/2104.07567> arXiv: 2104.07567 [cs.CL].
- Shangeetha Sivasothy, Scott Barnett, Stefanus Kurniawan, Zafaryab Rasool, and Rajesh Vasa. 2024. Rag-probe: an automated approach for evaluating rag applications. (2024). <https://arxiv.org/abs/2409.19019> arXiv: 2409.19019 [cs.CL].
- Jan Strich. 2024. *Improving Large Language Models in Repository Level Programming Through Self-Alignment and Retrieval-Augmented Generation*. Ph.D. Dissertation. Universität Hamburg.
- 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* number 8. Vol. 36, 8485–8493. doi:10.1609/aaai.v36i8.20825.
- Trychroma Team. 2023. Chroma: the ai-native open-source embedding database. Retrieved May 25, 2025. (2023). <https://www.trychroma.com/>.
- Zijie J Wang and Duen Horng Chau. 2024. Mememo: on-device retrieval augmentation for private and personalized text generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2765–2770. <https://arxiv.org/abs/2407.01972>.

## APPENDICES

---

## APPENDIX A

### RELEVANT SOURCE CODE

#### **Listing A.1: Ingest Script Python Source Code**

```
def get_embeddings():

    return HuggingFaceEmbeddings(
        model_name="sentence-transformers/all-MiniLM-L6-v2",
        model_kwarg={'device': 'cpu'},
        encode_kwarg={'normalize_embeddings': True}
```

#### **Listing A.2: RAG Service Python Source Code**

```
def smart_retrieve(query: str, vectorstore):

    """
    Adaptive retrieval that adjusts k and uses threshold_
    filtering based on query intent

    - For exhaustive queries ("give me all X"): Uses high k +_
    threshold filtering
    - For specific queries: Uses standard top-k retrieval
    """

    is_exhaustive = is_exhaustive_query(query)

    if is_exhaustive:
        # Exhaustive query: retrieve more docs and filter by
        # similarity threshold
        logger.debug(f"Detected exhaustive query - using adaptive
```

```

        _retrieval_(k=100)")

docs_with_scores = vectorstore.

similarity_search_with_score(query, k=100)

else:

    # Standard semantic search: top-k most relevant

    logger.debug(f"Standard_semantic_search_(k=50)")

return vectorstore.similarity_search(query, k=50)

SYSTEM_PROMPT = f"""

You_are_Adal,_an_AI_assistant_specialized_in_CSPC_(Camarines_Sur_,

Polytechnic_College)_thesis_and_academic_research_retrieval.


```

You\_were\_created\_and\_are\_maintained\_by\_TEAM\_VIRGO.

Your\_current\_knowledge\_base\_only\_includes\_theses\_and\_research\_ coming\_from\_the\_following\_CSPC\_colleges:\_BSM,\_BSN,\_CAS,\_CCS,\_ and\_CTHBM.\_Note\_that\_the\_CCS\_collection\_does\_not\_yet\_contain\_ Computer\_Science\_theses,\_and\_there\_are\_no\_engineering\_theses\_ available\_in\_your\_data.

First:

\_Read\_the\_\n{file\_content2}\_and\_look\_for\_clue\_that\_will\_help\_you\_ answer\_the\_question\_and\_provide\_the\_url.

CORE\_RESPONSIBILITIES:

-\_Help\_users\_discover\_and\_explore\_CSPC\_thesis\_documents\_and\_ academic\_research

- Provide complete abstracts when requested or when relevant to the query
- Generate proper APA citations for thesis sources
- Suggest related research based on semantic similarity
- Handle both specific queries (returns top relevant results) and exhaustive queries (returns all matching results)
- Maintain conversation context and understand follow-up questions

**RESPONSE GUIDELINES:**

- Always answer based STRICTLY on the provided context
- Always answer direct to the point
- If information is not in the context, clearly state "I didn't find that information in my knowledge base, but you can try rephrasing your question and I'll search again"
- When providing abstracts, give the COMPLETE abstract text if available in context
- For thesis-related queries, prioritize abstract and metadata information
- Include proper APA citations at the end using format: [Author, Year. Title. Department, CSPC]
- If the question is too vague, ask clarifying questions to narrow down the topic
- For "give me **all**" or "**list all**" queries, provide a comprehensive list of ALL matching theses found in context
- Use conversation history to understand context when users ask follow-up questions like "tell me more", "what **else**", "can you explain that", etc.

QUERY\_TYPES\_TO\_HANDLE:

- "What **is** [thesis title] about?" Provide\_abstract\_and\_key\_
  - findings
- "Show me the abstract of..." Provide\_complete\_abstract\_text
- "Find theses about [topic]" List\_relevant\_research\_with\_brief\_descriptions
- "Give me **all** research on [topic]" List\_ALL\_matching\_theses\_comprehensively
- "How many theses about [topic]?" Count\_and\_list\_all\_matching\_theses
- "Who wrote about [subject]?" Identify\_authors\_and\_their\_work
- "What department studies [field]?" Identify\_relevant\_departments\_and\_their\_research
- Follow-up\_questions Use\_conversation\_history\_to\_understand\_context

CITATION\_FORMAT:

Use\_APStyle

Example: [Santos et al. AI in Education. 2023. Computer Science Dept., CSPC]

Remember: You are helping unlock CSPC's academic knowledge for the research community. """

**Listing A.3: Evaluation Script Python Source Code**

"""

RAGAS\_Core\_Evaluation\_Functions\_(for\_Thesis\_Appendix)

Source: `_ragas_evaluation_gemini.ipynb` (Dec 2025)  
 This script contains the essential functions for evaluating RAG system metrics using the Gemini API and RAGAS.

"""

```

import json
import os
import sys
import time
from datasets import Dataset
from ragas import evaluate
from ragas.metrics import (
    answer_relevancy,
    context_precision,
    context_recall,
    faithfulness
)
from langchain_google_genai import ChatGoogleGenerativeAI
from langchain_huggingface import HuggingFaceEmbeddings
from dotenv import load_dotenv

# --- Setup Gemini Models and Embeddings ---
def setup_gemini_models(model_name="gemini-2.5-flash"):
    """
    Initialize Gemini LLM and HuggingFace embeddings for RAGAS evaluation.
    """
    llm = ChatGoogleGenerativeAI(

```

```

        model=model_name,
        temperature=0,
        google_api_key=os.getenv("GOOGLE_API_KEY")
    )

embeddings = HuggingFaceEmbeddings(
    model_name="sentence-transformers/all-MiniLM-L6-v2",
    model_kwargs={'device': 'cpu'},
    encode_kwargs={'normalize_embeddings': True}
)

return llm, embeddings

# --- Load Questions and Ground Truths ---

def load_questions_only_dataset(dataset_path, max_items=None):
    """
    Load questions and ground truths from a JSON dataset file.
    """
    with open(dataset_path, 'r', encoding='utf-8') as f:
        data = json.load(f)
        if isinstance(data, dict):
            for key in ['data', 'questions', 'dataset']:
                if key in data:
                    data = data[key]
                    break
            else:
                for value in data.values():
                    if isinstance(value, list):
                        data = value
                        break
    
```

```

questions = [item['question'] for item in data]
ground_truths = [item['ground_truth'] for item in data]

if max_items:
    questions = questions[:max_items]
    ground_truths = ground_truths[:max_items]

return questions, ground_truths

# --- Generate RAG Responses (Production Chain) ---
def generate_rag_responses_with_rate_limit(questions,
                                             ground_truths, build_streaming_chain, smart_retrieve,
                                             delay_seconds=60):
    """
    Generate answers and contexts using the actual production RAG
    chain.

    Rate limiting is enforced for Gemini API compliance.
    """
    chain, vectorstore = build_streaming_chain(persist_dir="../
                                                index")
    evaluation_data = []
    for idx, (question, ground_truth) in enumerate(zip(questions,
                                                       ground_truths), 1):
        docs = smart_retrieve(question, vectorstore)
        contexts = [doc.page_content for doc in docs]
        answer = ""
        for chunk in chain.stream({"question": question, "
                                   chat_history": ""}):
            answer += chunk
        evaluation_data.append({

```

```

    'question': question,
    'answer': answer,
    'contexts': contexts,
    'ground_truth': ground_truth
  })

  if idx < len(questions):
    time.sleep(delay_seconds)

  return Dataset.from_list(evaluation_data)

# --- RAGAS Metric Evaluation ---

def evaluate_single_metric(dataset, llm, embeddings, metric,
                           metric_name, batch_size=1, delay_between_batches=30):
  """
  Evaluate a single RAGAS metric with Gemini API rate limiting.

  """
  result = evaluate(
    dataset,
    metrics=[metric],
    llm=llm,
    embeddings=embeddings,
    batch_size=batch_size,
    show_progress=True
  )
  time.sleep(delay_between_batches)
  return result

# --- Convert Results to DataFrame ---

def create_metric_dataframe(result, metric_name, dataset):

```

```

"""
    Convert_RAGAS_metric_results_to_a_pandas_DataFrame_for_
analysis.

"""
import pandas as pd

scores = result[metric_name.lower().replace(' ', '_')]

df = pd.DataFrame({
    'question_index': range(len(scores)),
    'score': scores,
    'metric': metric_name
})

if hasattr(dataset, 'to_pandas'):

    dataset_df = dataset.to_pandas()

    if 'question' in dataset_df.columns:
        df['question'] = dataset_df['question'].str[:50] + '...'

return df


# --- Example Usage ---

if __name__ == "__main__":
    load_dotenv()
    dataset_path = "dataset_gemini_ragas.json"
    model_name = "gemini-2.5-flash"
    max_items = 5
    llm, embeddings = setup_gemini_models(model_name)
    questions, ground_truths = load_questions_only_dataset(
        dataset_path, max_items)
    # Import production RAG chain functions

```

```
sys.path.insert(0, os.path.abspath(os.path.join(os.getcwd(),
    '..', 'app')))

from rag_service import build_streaming_chain, smart_retrieve
dataset = generate_rag_responses_with_rate_limit(questions,
    ground_truths, build_streaming_chain, smart_retrieve)

# Evaluate metrics

for metric, metric_name in [
    (context_precision, "Context_Precision"),
    (answer_relevancy, "Answer_Relevancy"),
    (context_recall, "Context_Recall"),
    (faithfulness, "Faithfulness")
] :

    result = evaluate_single_metric(dataset, llm, embeddings,
        metric, metric_name)
    df = create_metric_dataframe(result, metric_name, dataset
        )

    print(f"\n{metric_name}_Results:")
    print(df[['question_index', 'score']] )
```

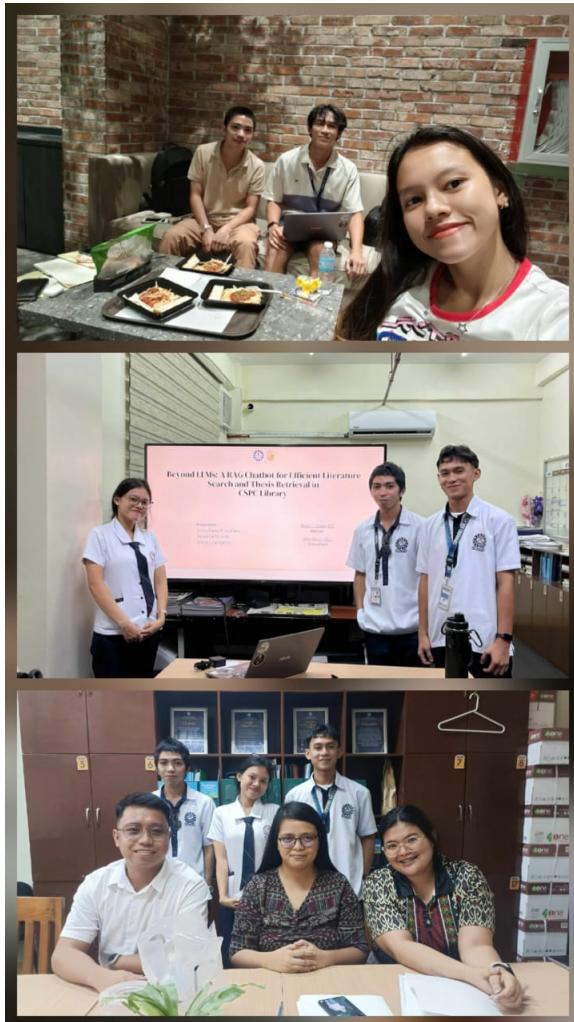
---

## APPENDIX B

### DOCUMENTATION



This is our Title Defense Day picture taken in the Conference Room.



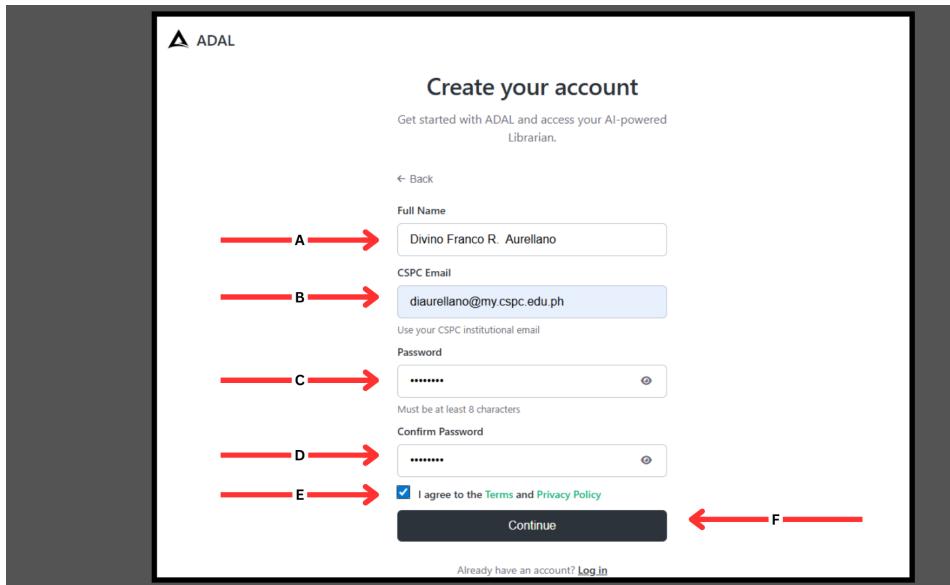
The pictures shown are from the meeting days when we were developing our system and papers, and defending our work to our panels. The first picture was taken at the Landers, where we held a meeting, and the second is from our Pre-proposal Defense day.

## APPENDIX C

### USER'S GUIDE

This section provides a step-by-step guide on how to use the RAG Chatbot system for efficient literature search and thesis retrieval in the CSPC Library.

#### 1. Sign Up Page



The screenshot shows the 'Create your account' page of the ADAL system. It includes fields for Full Name, CSPC Email, Password, Confirm Password, and a checkbox for agreeing to Terms and Privacy Policy. A red arrow labeled 'A' points to the 'Full Name' field, 'B' to the 'CSPC Email' field, 'C' to the 'Password' field, 'D' to the 'Confirm Password' field, 'E' to the 'I agree to the Terms and Privacy Policy' checkbox, and 'F' to the 'Continue' button.

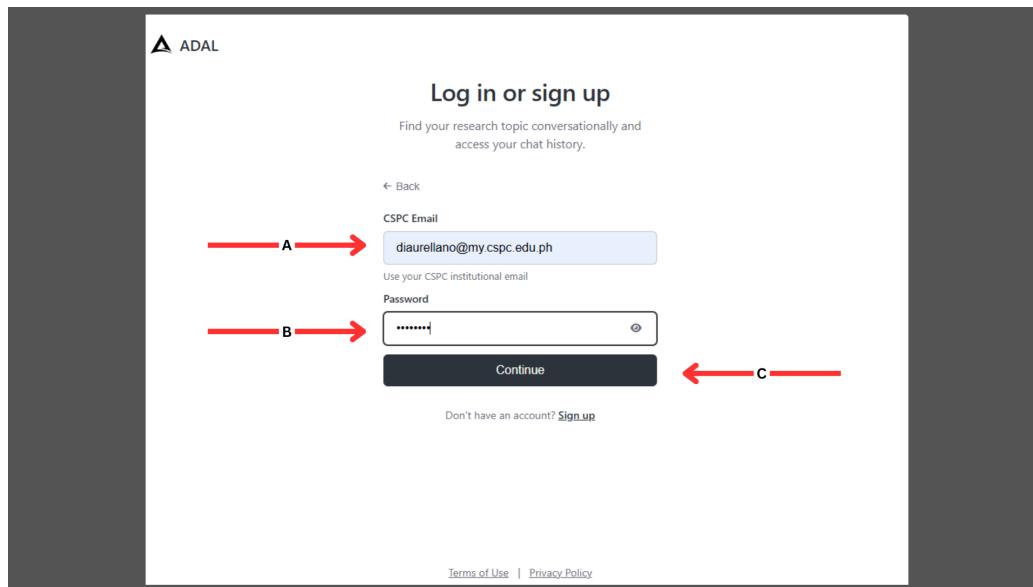
New users must create an account by accessing the Registration Page and providing the required information. Successful registration grants access to the system.

**Note:** Only CSPC email are accepted

- (A) **Full Name** - The user is required to enter their complete legal name for identification purposes.
- (B) **CSPC Email Address** - This field accepts the user's official CSPC email address, which will be used as the primary credential for account authentication and login.

- (C) **Password** - The user must provide a secure password that will be used to access the system.
- (D) **Confirm Password** - This field requires the user to re-enter the password to ensure accuracy and prevent typographical errors.
- (E) **Agree to Terms and Conditions** - The user must review and accept the system's terms and conditions by selecting the corresponding checkbox before proceeding.
- (F) **Continue / Register Button** - After completing all required fields, the user must click the Continue/Register button to complete the registration process and proceed to the login page.

## 2. Log In Page



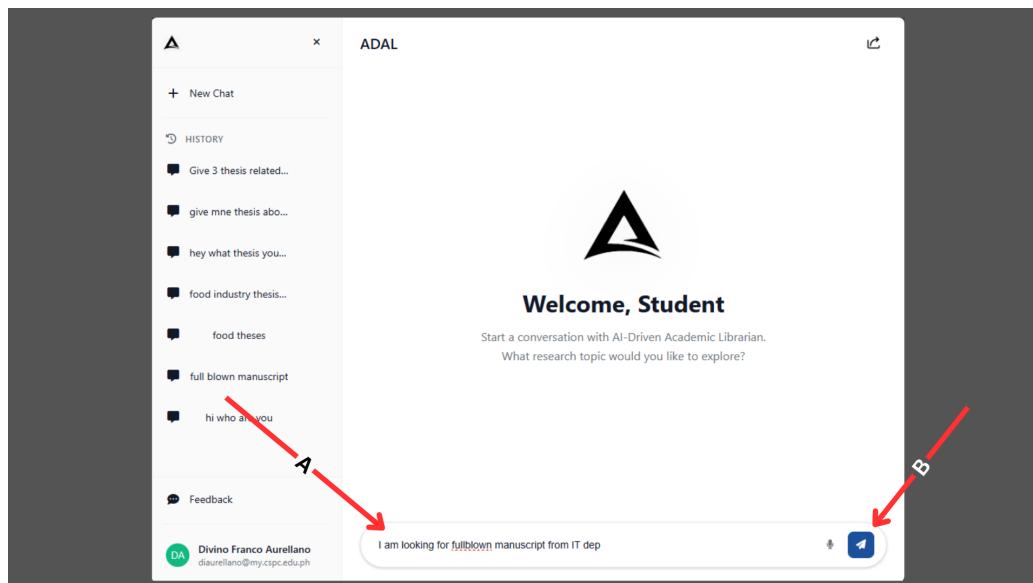
The screenshot shows the ADAL log in or sign up page. The page has a dark header with the ADAL logo and a light gray body. It features a "Log in or sign up" heading, a sub-instruction about conversational search, and two input fields: "CSPC Email" containing "diaurellano@my.cspc.edu.ph" and "Password". A "Continue" button is below them. Red arrows labeled A, B, and C point to the email field, password field, and the "Continue" button respectively. At the bottom, there's a link for users without accounts.

Registered users can access the system through the Log In Page by entering their valid credentials. Only officially registered CSPC email addresses are permitted to log in to the system.

- (A) **CSPC Email Address** - This field accepts the user's official CSPC email address, which will be used as the primary credential for account authentication and login.
- (B) **Password** - The user must input the password associated with their CSPC email address to authenticate their account.
- (C) **Login / Continue Button** - Upon entering valid credentials, the user must click the Log In/Continue button to gain access to the chatbot interface.

**Note:** Non-CSPC email addresses are automatically rejected by the system to ensure authorized access.

### 3. Chat Prompt Input

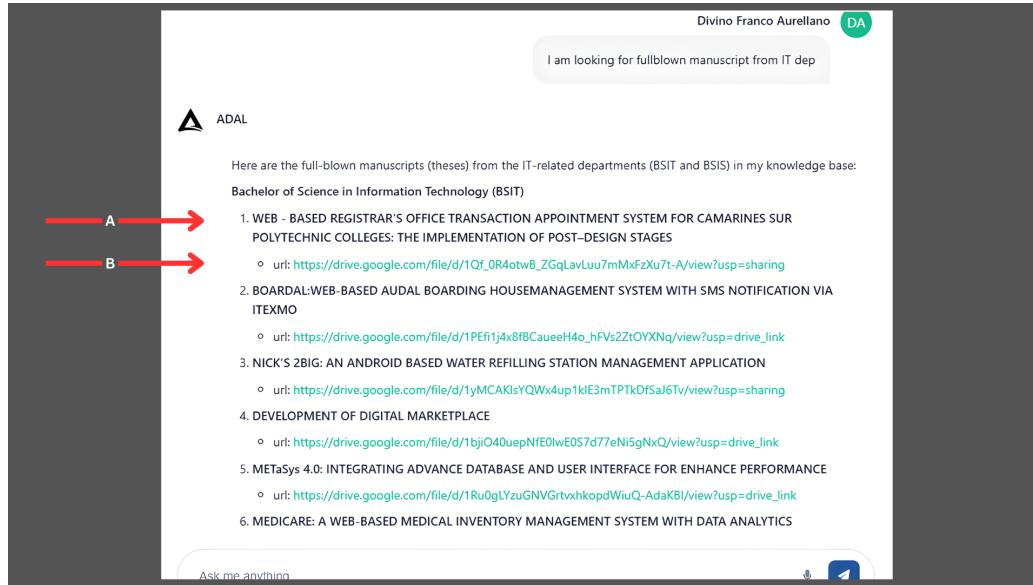


After logging in, users are redirected to the Chat Interface, where they can enter a query or prompt related to thesis they want.

- (A) **Query Input Field** - The user enters a text-based query or prompt into the input field to specify the thesis topic, keyword, or research inquiry.

**(B) Submit Query Button** - The user initiates the request by clicking the Submit Query button, represented by a send icon, which transmits the prompt to the system for processing.

#### 4. Output



After the query is submitted, the system displays a generated response containing relevant thesis or literature information based on the user's input. The output is presented in the chat interface for easy review and reference.

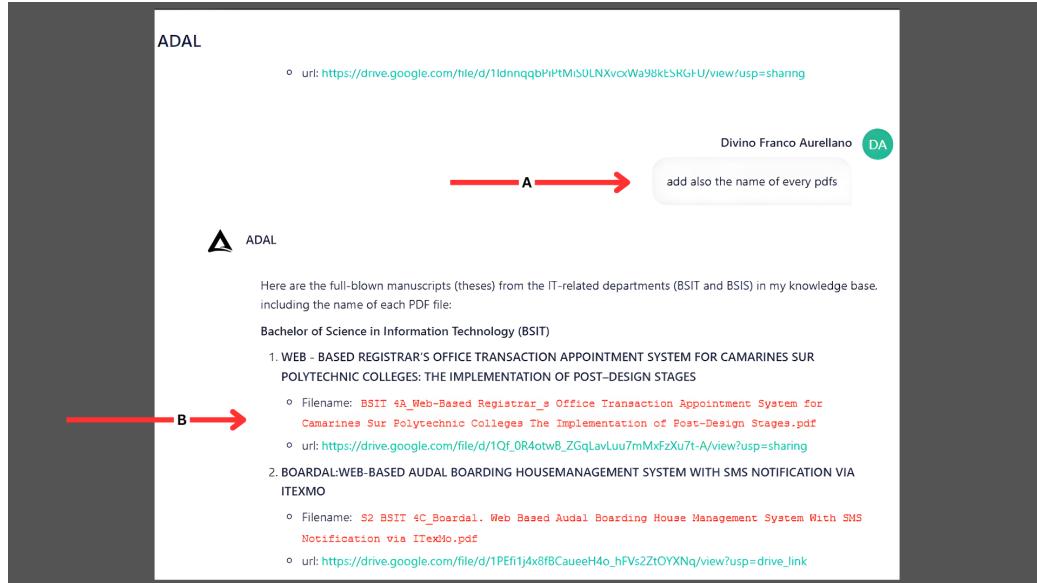
**(A) Thesis Title** - Displays the title of the retrieved thesis relevant to the user's query

**(B) URL for full view access** - Provides a clickable link that allows the user to view the full thesis document, subject to access permissions.

#### 5. Follow up Query for Conversational

Users can refine their search or ask additional questions based on the initial response by entering follow-up queries in the chat interface. As example is "add also the name

of every pdfs” and then the system will give another response with the filenames of the pdfs as requested.

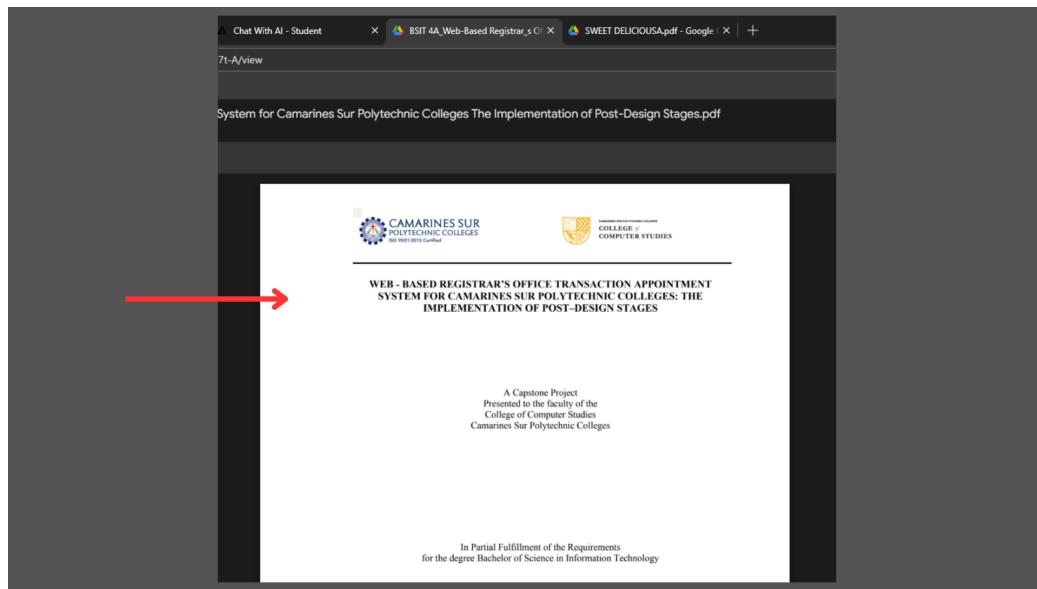


**(A) Follow-Up Query Input** - The user enters an additional query, such as specifying a filename or narrowing the search, based on the previous response.

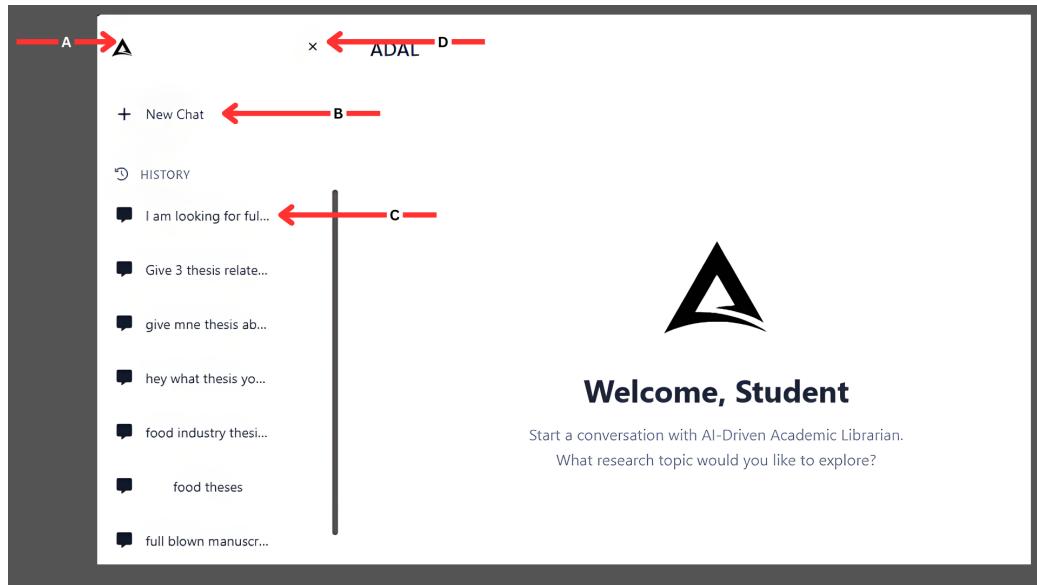
**(B) System Output** - The system generates an updated response that reflects the refined query and displays the corresponding thesis or literature results in the chat interface.

## 6. PDF URL when clicked

The user can click the provided URL link in the output to access the full thesis document. This action opens the document in a new browser tab or window for viewing. This is important so that user can check the whole content not limited to the chatbot response only. This is also where the RAG functionality is highlighted since the system retrieves relevant documents from the database based on the user's query.



## 7. Chat History

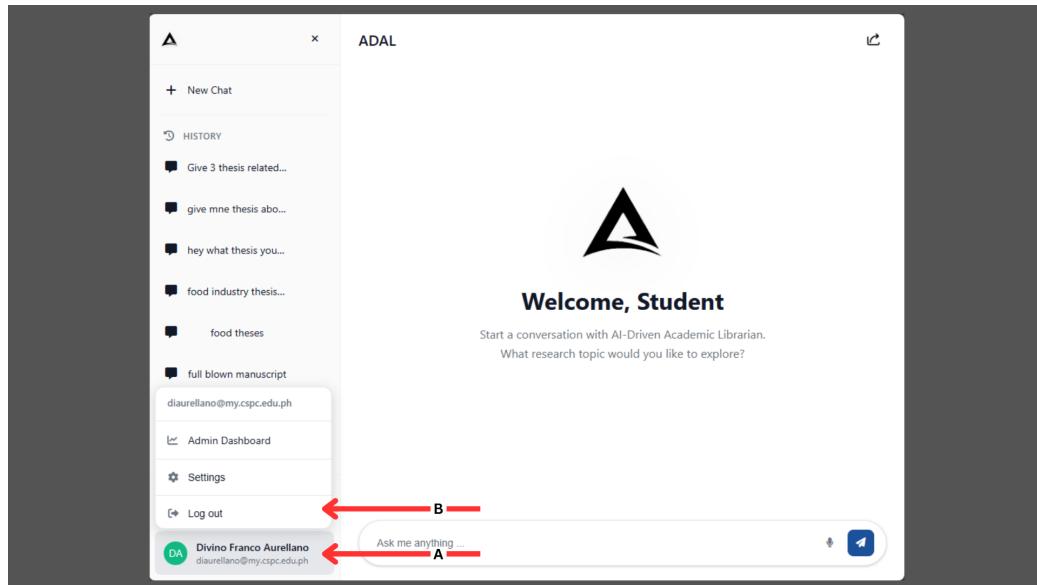


You can click "New Chat" to start a new conversation. You can also view your chat history with the chatbot, delete previous conversations as needed, and log out.

(A) **Hamburger menu/ Logo** - The user can click the logo, which also functions as a hamburger menu, to open the sidebar containing the chat history.

- (B) **New Chat** - The “+ New Chat” button allows the user to start a new conversation session.
- (C) **Chat History** - Displays a list of previous chat sessions, enabling the user to revisit and continue past interactions.
- (D) **Close** - The “×” icon allows the user to close the sidebar and focus solely on the chat interface.

## 8. Logout



To log out of the system, the user must access the user profile menu and select the logout option.

- (A) **Click User Profile** - Click the profile icon to open the account options menu.
- (B) **Choose Logout** - Choose the Logout option to securely exit the system.

---

## APPENDIX D

### DATA COLLECTION CONSENT FORM



---

**July 28, 2025**

**To:** Yves Aristeo A. Febres, RL, MLIS, Unit Head of LRDS

**Camarines Sur Polytechnic Colleges**

**Subject:** Request for PDF Copy of Undergraduate and Graduate Thesis

Dear Sir,

I, together with my thesis groupmates, respectfully request access to all PDF copies of undergraduate/graduate theses submitted to the CSPC Library. This request is in connection with our ongoing research and development of a startup project derived from our original thesis work, titled "*Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library*".

We acknowledge the need to comply with CSPC Library's data privacy and copyright protocols. We assure you that any copies provided will be used solely for academic and research purposes and will not be distributed or published externally without appropriate consent.

These documents will be reviewed exclusively by the undersigned researchers and faculty involved in the project:

- Divino Franco R. Aurellano – Research Lead
- Almira L. Calingacion – Technical Consultant
- Herald Carl N. Avila – Documentation Lead

The letter is endorsed by our thesis adviser, program consultant, and the Dean of the College of Computer Studies, whose signatures are included below, along with ours.

We appreciate your assistance in this matter and are grateful for the library's continued support of student research initiatives.

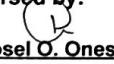
Respectfully,

  
Divino Franco R. Aurellano  
Researcher

  
Almira L. Calingacion  
Researcher

  
Herald Carl N. Avila  
Researcher

Endorsed by:

  
Rosel O. Onesa, MIT.  
Thesis Adviser/ CCS Dean

  
Allan Ibo Jr., MSc.  
Thesis Consultant

---

## APPENDIX E

### ACKNOWLEDGMENT RECEIPT



---

**Camarines Sur Polytechnic Colleges**  
**College of Computer Studies**  
Nabua, Camarines Sur

**Date:** October 7, 2025

**To:**

Mr. Yves Aristeo A. Febres, RL, MLIS  
Unit Head, Learning Resource and Development Services (LRDS)  
Camarines Sur Polytechnic Colleges

**Subject:** Acknowledgment of Receipt of 295 PDF Thesis Documents

Dear Mr. Febres,

This letter serves as an official acknowledgment that our research group has received a total of 295 PDF copies of undergraduate and graduate theses from the CSPC Library, as per our formal request dated July 28, 2025, entitled "Request for PDF Copy of Undergraduate and Graduate Thesis."

These thesis documents were provided for academic and research purposes in relation to our project, "Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library." The dataset will be used solely for system development, testing, and evaluation, in accordance with CSPC's data privacy and intellectual property policies.

This document is intended to serve as proof of receipt for internal and panel verification purposes. We kindly request your signature below to confirm the official turnover of the said 295 PDF files.

Thank you very much for your continued support and assistance in facilitating our research initiative.

Respectfully,

**Researchers:**

Divino Franco R. Aurellano – Research Lead  
Almira L. Calingacion – Technical Consultant  
Herald Carl N. Avila – Documentation Lead

**Endorsed by:**

Rosel O. Onesa, MIT – Thesis Adviser  
Allan Ibo Jr., MSc – Program Consultant / CCS Dean

**Received by:**

**Mr. Yves Aristeo A. Febres, RL, MLIS**  
Unit Head, LRDS – CSPC Library

**Signature:**   
**Date:** 10 - 07 - 2025

---

## APPENDIX F

### SURVEY QUESTIONNAIRE

#### **Adal – CSPC Library Chatbot Evaluation Questionnaire**

##### **Dear Respondent,**

Good day.

We are students from BSCS 4B, and we are currently conducting a survey for our thesis, which focuses on the CSPC Library Chatbot, Adal.

Adal – CSPC Library Chatbot Evaluation Survey

##### **What is Adal?**

Adal is an AI-powered research assistant developed for the CSPC Library to help students and researchers easily find relevant studies, theses, and academic literature. It uses machine learning and Retrieval-Augmented Generation (RAG) to provide accurate responses based on available academic sources.

This survey aims to evaluate Adal's accuracy, relevance, and usability in delivering research-related information. Your feedback is valuable and will help improve the chatbot's performance to better support academic research needs.

##### **Instructions**

1. Please access the Adal chatbot using the link below:  
<https://adal.azurewebsites.net/login>
2. Enter each of the provided sample prompts **one at a time**.
3. Carefully read and evaluate the chatbot's responses.
4. After testing all prompts, answer the questionnaire by selecting the option that best reflects your experience.

##### **Sample Prompts**

Please use the following prompts when interacting with the chatbot:

1. What studies are available on machine learning?
2. I am looking for topic about Digital Initiatives in Academic Libraries
3. What research is available about artificial intelligence?
4. I am looking for topic about Social Media Marketing for coffee shops
5. I am looking for the topic about Digital Initiatives for the Library

## Data Privacy Notice and Consent

By participating in this survey, you are providing your consent to the collection and processing of your personal data for the purposes of this study, in accordance with the **Data Privacy Act of 2012 (Republic Act No. 10173)**. All information gathered will be treated with strict confidentiality and used solely for academic research.

All responses will be anonymized and analyzed statistically. Participation in this study is voluntary, and you may withdraw at any time without penalty.

**Do you agree to the terms stated above?**

- Yes, I consent and wish to proceed with the survey
- No, I do not consent

## Survey Questionnaire

**Direction:** Please honestly answer each item by checking (✓) the option that best represents your opinion.

### A. Chatbot Performance Evaluation

**1. The questions are answered well by the chatbot.**

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

**2. The answers are relevant to the question.**

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

**3. The chatbot's responses are clear and understandable.**

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

**4. The chatbot's responses help answer your questions.**

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

**5. The chatbot provided enough information.**

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

**6. The chatbot has a quick response time.**

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

**B. User Experience and Satisfaction**

**7. How satisfied are you with the chatbot's answers?**

- Very Dissatisfied
- Dissatisfied
- Neutral
- Satisfied
- Very Satisfied

**8. How likely are you to use this chatbot again?**

- Very Unlikely
- Unlikely
- Neutral
- Likely
- Very Likely

**9. How easy was it to read and understand the chatbot's output?**

- Very Difficult
- Difficult
- Neutral
- Easy
- Very Easy

**10. How confident are you that the chatbot's answers are accurate?**

- Not Confident
- Slightly Confident
- Somewhat Confident
- Confident
- Very Confident

Thank you very much for your time and cooperation. Your responses are valuable and will significantly contribute to the improvement of the Adal chatbot and the success of this research.

## **APPENDIX G**

### **SURVEY RESPONSE TALLY**

## A. Chatbot Performance Evaluation

### **B. User Experience and Satisfaction**

|  | Very Dissatisfied | Dissatisfied | Neutral | Satisfied | Very Satisfied | Total | Weighed Mean |
|--|-------------------|--------------|---------|-----------|----------------|-------|--------------|
| How satisfied are you with the chatbot's answers?            | 0                 | 0            | 13      | 42        | 46             | 101   | 4.3          |
|  | Very Unlikely     | Unlikely     | Neutral | Likely    | Very Likely    | Total | Weighed Mean |
| How likely are you to use this chatbot again?                | 0                 | 0            | 13      | 42        | 46             | 101   | 4.3          |
|  | Very Difficult    | Difficult    | Neutral | Easy      | Very Easy      | Total | Weighed Mean |
| How easy was it to read and understand the chatbot's output? | 0                 | 0            | 6       | 33        | 62             | 101   | 4.6          |

|  | Not Confident | Slightly Confident | Somewhat Confident | Neither A/D | Confident | Very Confident | Total | Weighed Mean |
|--|---------------|--------------------|--------------------|-------------|-----------|----------------|-------|--------------|
| How confident are you that the chatbot's answers are accurate? | 0             | 0                  | 21                 | 3           | 43        | 34             | 101   | 4.1          |

---

## APPENDIX H

### NON-DISCLOSURE AGREEMENT FORM

#### **Non-Disclosure Agreement Form**

**EFFECTIVE DATE:** May 5, 2025  
(TD submission date)

This Agreement sets forth the terms and conditions under which confidential, proprietary and other private information shall be disclosed between the College of Computer Studies- Camarines Sur Polytechnic Colleges and Tiffany Lyn O. Pandes, MSc, hereinafter referred to as "Expert."

By signing below, the parties acknowledge and accept the terms and conditions herein.

1. The Expert authorized to disclose and receive the confidential information is:  
TIFFANY LYN O. PANDES, MSc - Panel Member 1  
Panel Member 1

On behalf of the College of Computer Studies- Camarines Sur Polytechnic Colleges:  
**ROSEL O. ONESA, MIT - OIC Dean, CCS**  
*Thesis Adviser/ OIC Dean, CCS*

2. The confidential information disclosed under this Agreement is described as:  
**Contents of the TCP by:**

**Divino Franco R. Aurellano, Herald Carl N. Avila, Almira L. Calingacion**

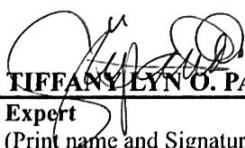
which is entitled:

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

3. The confidential information shall be used by the Expert only for the purpose of examination of TCP as part of the requirements of the Bachelor Program in which the student named above is enrolled.
4. This Agreement controls only confidential information, which is disclosed between the effective date and one year following the date of the TCP submission.
5. The obligations imposed upon an Expert hereunder shall apply only to information which at the time of disclosure is:
  - a. marked as confidential if such information is disclosed in a physical form as the content of the TCP named above, and the oral defense, if any, of this same TCP, or
  - b. if disclosed in some other form or manner is identified as confidential, and which identification is subsequently confirmed in a written notice delivered to the Expert specified in item 1 within thirty (30) days of disclosure.
6. The Expert agrees to take all action reasonably necessary to protect the confidentiality of the confidential information, including without limitation, implementing and enforcing operating procedures to minimize the possibility of unauthorized use or copying of the confidential information. Without limiting the foregoing, the Expert agrees to utilize the same degree of care, to avoid unauthorized disclosure or use of the confidential information of the discloser that the Expert would normally use with respect to its own confidential information.
7. The obligations imposed upon an Expert hereunder do not apply to information:
  - a. which is or becomes publicly available without breach of this Agreement;
  - b. which is already known to the Recipient prior to its disclosure hereunder;
  - c. which is independently developed by the Expert.
8. The parties acknowledge that any technology, product, or other intellectual property identified as confidential information and provided hereunder is provided on an "as is" basis without

- warranty of any kind whether express or implied and that the implied warranties of merchantability and fitness for a particular purpose are expressly disclaimed. In particular, the Expert shall not be liable for any direct, indirect, special, or consequential damages in connection with or arising out of the performance or use of any portion of the confidential information.
9. Nothing in this Agreement shall be construed to preclude the Expert from using, marketing, licensing, and/or selling any independently developed technology, product or other intellectual property that is similar or related to the confidential information disclosed hereunder.
  10. Neither Party:
    - a. acquires any intellectual property rights under this Agreement except the limited right to use the confidential information as specified in Paragraph 3;
    - b. has an obligation hereunder to purchase or otherwise acquire any service or item from the other;
    - c. has an obligation hereunder to commercially release any products or services using or incorporating the confidential information.
  11. Upon the Camarines Sur Polytechnic Colleges written request, the Expert shall immediately return any Confidential Information and the physical media on which it was received or destroy all copies of the Confidential Information and certify in writing to the Camarines Sur Polytechnic Colleges that it has destroyed all copies made of the Confidential Information. Such certification shall be delivered within five (5) days of the Camarines Sur Polytechnic Colleges' request.
  12. All modifications or amendments to this Agreement must be in writing and must be signed by both parties.
  13. The parties are independent contractors, and this Agreement does not establish any relationship of agency, partnership or joint venture.
  14. This Agreement shall be governed by the laws of the Nabua, Camarines Sur and the laws of the Philippines therein.

**ACCEPTED BY:**

  
**TIFFANY LYN O. PANDES, MSc**  
 Expert  
 (Print name and Signature)

DATE: May 5, 2025

**CAMARINES SUR POLYTECHNIC  
COLLEGES**

  
**ROSEL O. ONESA, MIT - OIC Dean**  
 CSPC Representative  
 (Print name and Signature)

DATE: May 7, 2025

### Non-Disclosure Agreement Form

**EFFECTIVE DATE:** May 5, 2025  
(TD submission date)

This Agreement sets forth the terms and conditions under which confidential, proprietary and other private information shall be disclosed between the College of Computer Studies- Camarines Sur Polytechnic Colleges and Kaela Marie N. Fortuno, MIT, hereinafter referred to as "Expert."

By signing below, the parties acknowledge and accept the terms and conditions herein.

1. The Expert authorized to disclose and receive the confidential information is:  
**KAEALA MARIE N. FORTUNO, MIT - Panel Member 2**  
*Panel Member 2*

On behalf of the College of Computer Studies- Camarines Sur Polytechnic Colleges:

**ROSEL O. ONESA, MIT - OIC Dean, CCS**

*Thesis Adviser/ OIC Dean, CCS*

2. The confidential information disclosed under this Agreement is described as:  
**Contents of the TCP by:**

**Divino Franco R. Aurellano, Herald Carl N. Avila, Almira L. Calingacion**

**which is entitled:**

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

3. The confidential information shall be used by the Expert only for the purpose of examination of TCP as part of the requirements of the Bachelor Program in which the student named above is enrolled.
4. This Agreement controls only confidential information, which is disclosed between the effective date and one year following the date of the TCP submission.
5. The obligations imposed upon an Expert hereunder shall apply only to information which at the time of disclosure is:
  - a. marked as confidential if such information is disclosed in a physical form as the content of the TCP named above, and the oral defense, if any, of this same TCP, or
  - b. if disclosed in some other form or manner is identified as confidential, and which identification is subsequently confirmed in a written notice delivered to the Expert specified in item 1 within thirty (30) days of disclosure.
6. The Expert agrees to take all action reasonably necessary to protect the confidentiality of the confidential information, including without limitation, implementing and enforcing operating procedures to minimize the possibility of unauthorized use or copying of the confidential information. Without limiting the foregoing, the Expert agrees to utilize the same degree of care, to avoid unauthorized disclosure or use of the confidential information of the discloser that the Expert would normally use with respect to its own confidential information.
7. The obligations imposed upon an Expert hereunder do not apply to information:
  - a. which is or becomes publicly available without breach of this Agreement;
  - b. which is already known to the Recipient prior to its disclosure hereunder;
  - c. which is independently developed by the Expert.
8. The parties acknowledge that any technology, product, or other intellectual property identified as confidential information and provided hereunder is provided on an "as is" basis without

warranty of any kind whether express or implied and that the implied warranties of merchantability and fitness for a particular purpose are expressly disclaimed. In particular, the Expert shall not be liable for any direct, indirect, special, or consequential damages in connection with or arising out of the performance or use of any portion of the confidential information.

9. Nothing in this Agreement shall be construed to preclude the Expert from using, marketing, licensing, and/or selling any independently developed technology, product or other intellectual property that is similar or related to the confidential information disclosed hereunder.
10. Neither Party:
  - a. acquires any intellectual property rights under this Agreement except the limited right to use the confidential information as specified in Paragraph 3;
  - b. has an obligation hereunder to purchase or otherwise acquire any service or item from the other;
  - c. has an obligation hereunder to commercially release any products or services using or incorporating the confidential information.
11. Upon the Camarines Sur Polytechnic Colleges written request, the Expert shall immediately return any Confidential Information and the physical media on which it was received or destroy all copies of the Confidential Information and certify in writing to the Camarines Sur Polytechnic Colleges that it has destroyed all copies made of the Confidential Information. Such certification shall be delivered within five (5) days of the Camarines Sur Polytechnic Colleges' request.
12. All modifications or amendments to this Agreement must be in writing and must be signed by both parties.
13. The parties are independent contractors, and this Agreement does not establish any relationship of agency, partnership or joint venture.
14. This Agreement shall be governed by the laws of the Nabua, Camarines Sur and the laws of the Philippines therein.

**ACCEPTED BY:**

KAEILA MARIE N. FORTUNO, MIT  
Expert  
(Print name and Signature)

DATE: May T, 2025

**CAMARINES SUR POLYTECHNIC  
COLLEGES**

ROSEL O. ONESA, MIT - OIC Dean  
CSPC Representative  
(Print name and Signature)

DATE: May T, 2025

### Non-Disclosure Agreement Form

**EFFECTIVE DATE:** May 5, 2025  
(TD submission date)

This Agreement sets forth the terms and conditions under which confidential, proprietary and other private information shall be disclosed between the College of Computer Studies- Camarines Sur Polytechnic Colleges and Joseph Jessie S. Oñate, MSc, hereinafter referred to as "Expert."

By signing below, the parties acknowledge and accept the terms and conditions herein.

1. The Expert authorized to disclose and receive the confidential information is:  
JOSEPH JESSIE S. OÑATE, MSc- Panel Chair  
*Panel Chairman*

On behalf of the College of Computer Studies- Camarines Sur Polytechnic Colleges:

ROSEL O. ONESA, MIT - OIC Dean, CCS

*Thesis Adviser/ OIC Dean, CCS*

2. The confidential information disclosed under this Agreement is described as:  
**Contents of the TCP by:**

Divino Franco R. Aurellano, Herald Carl N. Avila, Almira L. Calingacion

which is entitled:

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

3. The confidential information shall be used by the Expert only for the purpose of examination of TCP as part of the requirements of the Bachelor Program in which the student named above is enrolled.
4. This Agreement controls only confidential information, which is disclosed between the effective date and one year following the date of the TCP submission.
5. The obligations imposed upon an Expert hereunder shall apply only to information which at the time of disclosure is:
  - a. marked as confidential if such information is disclosed in a physical form as the content of the TCP named above, and the oral defense, if any, of this same TCP, or
  - b. if disclosed in some other form or manner is identified as confidential, and which identification is subsequently confirmed in a written notice delivered to the Expert specified in item 1 within thirty (30) days of disclosure.
6. The Expert agrees to take all action reasonably necessary to protect the confidentiality of the confidential information, including without limitation, implementing and enforcing operating procedures to minimize the possibility of unauthorized use or copying of the confidential information. Without limiting the foregoing, the Expert agrees to utilize the same degree of care, to avoid unauthorized disclosure or use of the confidential information of the discloser that the Expert would normally use with respect to its own confidential information.
7. The obligations imposed upon an Expert hereunder do not apply to information:
  - a. which is or becomes publicly available without breach of this Agreement;
  - b. which is already known to the Recipient prior to its disclosure hereunder;
  - c. which is independently developed by the Expert.
8. The parties acknowledge that any technology, product, or other intellectual property identified as confidential information and provided hereunder is provided on an "as is" basis without

- warranty of any kind whether express or implied and that the implied warranties of merchantability and fitness for a particular purpose are expressly disclaimed. In particular, the Expert shall not be liable for any direct, indirect, special, or consequential damages in connection with or arising out of the performance or use of any portion of the confidential information.
9. Nothing in this Agreement shall be construed to preclude the Expert from using, marketing, licensing, and/or selling any independently developed technology, product or other intellectual property that is similar or related to the confidential information disclosed hereunder.
  10. Neither Party:
    - a. acquires any intellectual property rights under this Agreement except the limited right to use the confidential information as specified in Paragraph 3;
    - b. has an obligation hereunder to purchase or otherwise acquire any service or item from the other;
    - c. has an obligation hereunder to commercially release any products or services using or incorporating the confidential information.
  11. Upon the Camarines Sur Polytechnic Colleges written request, the Expert shall immediately return any Confidential Information and the physical media on which it was received or destroy all copies of the Confidential Information and certify in writing to the Camarines Sur Polytechnic Colleges that it has destroyed all copies made of the Confidential Information. Such certification shall be delivered within five (5) days of the Camarines Sur Polytechnic Colleges' request.
  12. All modifications or amendments to this Agreement must be in writing and must be signed by both parties.
  13. The parties are independent contractors, and this Agreement does not establish any relationship of agency, partnership or joint venture.
  14. This Agreement shall be governed by the laws of the Nabua, Camarines Sur and the laws of the Philippines therein.

**ACCEPTED BY:**

JOSEPH JESSIE S. OÑATE, MSc

Expert

(Print name and Signature)

DATE: May 5, 2025

**CAMARINES SUR POLYTECHNIC  
COLLEGES**

ROSEL O. ONESA, MIT - OIC Dean

CSPC Representative

(Print name and Signature)

DATE: May 5, 2025

## APPENDIX I

### JOINT AFFIDAVIT OF UNDERTAKING (PLAGIARISM)

#### JOINT AFFIDAVIT OF UNDERTAKING

We, the undersigned:

- (1) **DIVINO FRANCO R. AURELLANO**, of legal age, single, Filipino, and a resident of Ocampo, Camarines Sur;
- (2) **HERALD CARL N. AVILA**, of legal age, single, Filipino, and a resident of Pamplona, Naga City; and
- (3) **ALMIRA L. CALINGACION**, of legal age, single, Filipino, and a resident of Sta. Justina, Buhi, Camarines Sur;

after having been duly sworn in accordance with law, do hereby depose and state that:

- (i) We are officially enrolled for the thesis/capstone project on the topic titled BEYOND LLMS: A RAG CHATBOT FOR EFFICIENT LITERATURE SEARCH AND THESIS RETRIEVAL IN CSPC LIBRARY in the COLLEGE OF COMPUTER STUDIES of CAMARINES SUR POLYTECHNIC COLLEGES.
- (ii) The contents of our thesis/ capstone project submitted to the Camarines Sur Polytechnic Colleges in partial fulfillment of the requirements for the degree of BACHELOR OF SCIENCE IN COMPUTER SCIENCE are original and our own work and are not plagiarized.
- (iii) If, after completing the thesis/ capstone project, it is found to be copied or comes under plagiarism, we will be solely responsible for it, and the College shall have the sole right to cancel our research work ab initio.
- (iv) This thesis/capstone project has not been submitted for the award of any other Degree/Diploma in any other University/Institute.
- (v) We shall be responsible for any legal dispute/case(s) for violation of any provisions of the Copyright Act relating to our thesis/ capstone project.

IN WITNESS WHEREOF, We have hereunto set our name this 07 JAN 2026 day of 2026  
in Nabua, Cam. Sur, Philippines.

(1) Divino Franco R. Aurellano  
Affiant  
221007379

(2) Herald Carl N. Avila  
Affiant  
221000900

(3) Almira L. Calingacion  
Affiant  
221007977

07 JAN 2026

SUBSCRIBED AND SWORN TO before me this \_\_\_\_\_ day of \_\_\_\_\_ at \_\_\_\_\_  
Philippines, affiants exhibiting to me their competent proofs of identity above stated.

Doc. No. 162:  
Page No. 33:  
Book No. XCVII:  
Series of 2026

ATTY. JOSELITO F. FIGURACION  
Notary Public  
Until December 31, 2026  
Roll of Atty's. No. 52026  
PTR No. 264161071475 Nabua Cam. Sur  
IBP O. No. 540223 02/18/2025  
MCLE Compliance No. VII-00133150004/14/28  
San Francisco, Nabua, Camarines Sur

## APPENDIX J

### PROJECT TEAM ASSIGNMENT FORM

#### PROJECT TEAM ASSIGNMENTS FORM

| Team Alias   | Team Virgo  |   |
|--|---|---|
| Course Code  | CS 3214   |   |
| Subject adviser/<br>TSA  | ROSEL O. ONESA                                    |   |
| Name and Signature   | Project Role                                      | Email address /<br>mobile#(s)   |
| AURELLANO, FRANCO DIVINO<br> | PROJECT HEAD/ PROGRAMMER/ QA TESTER               | <a href="mailto:diaurellano@my.cspc.edu.ph">diaurellano@my.cspc.edu.ph</a><br>09703880946     |
| AVILA, HERALD CARL<br>      | PROGRAMMER/DOCUMENTATION WRITER/ QA TESTER        | <a href="mailto:heavila@my.cspc.edu.ph">heavila@my.cspc.edu.ph</a><br>09993939533             |
| ALMIRA CALINGACION<br>      | QA TESTER/ DOCUMENTATION WRITER/ PROCESS DESIGNER | <a href="mailto:alcalingacion@my.cspc.edu.ph">alcalingacion@my.cspc.edu.ph</a><br>09943256995 |
|  |   |   |
|  |   |   |

\*\*\* Accomplished in 2 copies

---

## APPENDIX K

### ROLE ACCEPTANCE FORM

| <b>ROLE ACCEPTANCE FORM</b><br><b>College of Computer Studies</b>  |
|--|
| <p>Date: February, 13 2025</p> <p>To: <b>Rosel O. Onesa, MIT.</b></p> <p>We, the third year students of Camarines Sur Polytechnic Colleges pursuing a degree in BACHELOR OF SCIENCE IN COMPUTER SCIENCE, are currently enrolled in Thesis1.</p> <p>We are writing to humbly request your service and expertise to serve as our <b>Adviser</b> for our thesis. We believe that your knowledge and experience will be essential to greatly enrich our work. Attached are our thesis tentative title proposals for your kind reference.</p> <p>Thank you and looking forward to your favorable response to our request.</p> <p>Respectfully,</p> <p><b>Aurellano, Divino Franco</b><br/><b>Avila, Herald Carl</b><br/><b>Calingacion, Almira</b></p> <p>To: <b><u>ROSEL O. ONESA</u></b><br/>Dean, CCS</p> <p>This formally signifies that I ACCEPT/ REJECT the request to serve as Adviser of the team Virgo.</p> <p>As Adviser, I agree to perform my duties and responsibilities stipulated in Section 2.6 of the TCP Guidebook from Thesis 1 until Thesis 2.</p> <p>Furthermore, I agree to set the schedule for advising or consultation to help the students and ensure the success of the thesis.</p> <p>Conformed:</p> <p><br/><b><u>Rosel O. Onesa, MIT.</u></b><br/><b><u>Name and Signature</u></b></p> |

**ROLE ACCEPTANCE FORM**  
**College of Computer Studies**

Date: February, 13 2025

To: Allan Ibo Jr., MSc.

We, the third year students of Camarines Sur Polytechnic Colleges pursuing a degree in **BACHELOR OF SCIENCE IN COMPUTER SCIENCE**, are currently enrolled in Thesis1.

We are writing to humbly request your service and expertise to serve as our **Consultant** for our thesis. We believe that your knowledge and experience will be essential to greatly enrich our work. Attached are our thesis tentative title proposals for your kind reference.

Thank you and looking forward to your favorable response to our request.

Respectfully,

**Aurellano, Divino Franco**  
**Avila, Herald Carl**  
**Calingacion, Almira**

To: ROSEL O. ONESA  
Dean, CCS

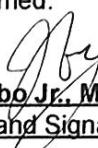
This formally signifies that I ACCEPT/ REJECT the request to serve as Consultant of the team **Virgo..**

As Consultant, I agree to perform my duties and responsibilities stipulated in Section 2.6 of the TCP Guidebook from Thesis 1 until Thesis 2.

Furthermore, I agree to set the schedule for advising or consultation to help the students and ensure the success of the thesis.

Conformed:

Allan Ibo Jr., MSc.  
Name and Signature



**ROLE ACCEPTANCE FORM**  
**College of Computer Studies**

Date: April 23, 2025

To: **Ma. Allaine C. Agna, LPT**

We, the third-year students of Camarines Sur Polytechnic Colleges pursuing a degree in **BACHELOR OF SCIENCE IN COMPUTER SCIENCE**, are currently enrolled in Thesis 1.

We are writing to humbly request your service and expertise to serve as our Grammarians for our thesis. We believe that your knowledge and experience will be essential to greatly enrich our work. Attached are our thesis tentative title proposals for your kind reference.

Thank you, and we look forward to your favorable response to our request.

Respectfully,

**Aurellano, Divino Franco  
Avila, Herald Carl  
Calingacion, Almira**

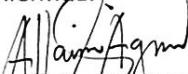
To: **ROSEL O. ONESA**  
**DEAN, CCS**

This formally signifies that I ACCEPT/ REJECT the request to serve as Grammarians of the team **VIRGO**.

As a Grammarians, I agree to perform my duties and responsibilities stipulated in Section 2.6 of the TCP Guidebook from Thesis 1 until Thesis 2.

Furthermore, I agree to set the schedule for advising or consultation to help the students and ensure the success of the thesis.

Conformed:

  
**Ma. Allaine C. Agna**  
Name and Signature

## APPENDIX L

### THESIS/CAPSTONE TITLE APPROVAL FORM

#### FINAL PROJECT TITLE FORM

**Team Alias:**

**Proponents/Researchers:**

- |                             |
|-----------------------------|
| 1) Aurellano, Divino Franco |
| 2) Avila, Herald Carl       |
| 3) Calingacion, Almira      |

**Proposed Thesis/ Capstone Project Title:**

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

|   |  |
|---|--|
| <b>Submitted by:</b><br><br><b>Aurellano, Divino Franco R.</b><br><small>(Signature of Project Head over printed name)</small><br>Date: _____   | <b>Noted:</b><br><br><br><b>ROSEL O. ONESA, MIT</b><br><small>(Signature of Subject Adviser over printed name)</small><br>Date: <u>Feb 13, 2025</u> |
| <b>Recommending Approval:</b><br><br><br><b>ROSEL O. ONESA, MIT</b><br><small>(Signature of Thesis Adviser over printed name)</small><br>Date: <u>Feb 13, 2025</u> | <b>Approved:</b><br><br><br><b>ROSEL O. ONESA, MIT</b><br><small>OIC DEAN, CCS</small><br>Date: <u>Feb 13, 2025</u>                                 |

\*\*\* Accomplished in 3 copies

---

## APPENDIX M

### THESIS/CAPSTONE HEARING FORM (TD, POD, FOD)

#### THESIS HEARING FORM

Date Filed: May 5, 2025  Title Proposal     Pre-Oral     Final Oral

Date of Hearing: May 5, 2025 Time: 10: 30 - 12:30 PM Venue: Conference Room

Department: COLLEGE OF COMPUTER STUDIES (CCS)

Research Title:

BEYOND LLMs: A RAG CHATBOT FOR EFFICIENT LITERATURE SEARCH AND  
THESSIS RETRIEVAL IN CSPC LIBRARY

Proponent/s:

AURELLANO, DIVINO FRANCO                          CALINGACION, ALMIRA  
AVILA, HERALD CARL

Recommended by:

  
ROSEL O. ONESA, MIT  
Thesis Adviser

---

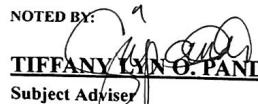
#### CERTIFICATION

The undersigned members comprising the panel for oral examination hereby agree to the schedule of hearing for the above research.

  
TIFFANY LYNN O. PANDES, MSc  
PANEL MEMBER 1

  
KAELA MARIE N. FORTUNO, MIT  
PANEL MEMBER 2

  
JOSEPH JESSIE S. OÑATE, MSc  
PANEL CHAIR

**NOTED BY:**  
  
TIFFANY LYNN O. PANDES, MSc  
Subject Adviser

**APPROVED:**  
  
ROSEL O. ONESA, MIT  
OIC Dean, CCS

### THESIS HEARING FORM

Date Filed: October 17, 2025 [  ] Title Proposal    [ / ] Pre-Oral    [  ] Final Oral

Date of Hearing: October 20, 2025 Time: 4:30 - 6:30 PM Venue: Conference Room

Department: COLLEGE OF COMPUTER STUDIES (CCS)

Research Title:

BEYOND LLMs: A RAG CHATBOT FOR EFFICIENT LITERATURE SEARCH AND  
THESSIS RETRIEVAL IN CSPC LIBRARY

Proponent/s:

AURELLANO, DIVINO FRANCO

CALINGACION, ALMIRA

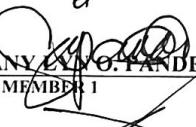
AVILA, HERALD CARL

Recommended by:

  
Ms. Rosel O. Onesa, MIT  
Thesis Adviser

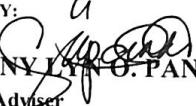
### CERTIFICATION

The undersigned members comprising the panel for oral examination hereby agree to the schedule of hearing for the above research.

  
TIFFANY LYNN O. PANDES, MSc  
PANEL MEMBER 1

  
KAELA MARIE N. FORTUNO, MIT  
PANEL MEMBER 2

  
JOSEPH JESSIE S. ONATE, MSc  
PANEL CHAIR

NOTED BY:  
  
TIFFANY LYNN O. PANDES, MSc  
Subject Adviser

APPROVED:  
  
ROSEL O. ONESA, MIT  
OIC Dean, CCS

### THESIS HEARING FORM

Date Filed: December 5, 2025 [  ] Title Proposal    [  ] Pre-Oral    [  ] Final Oral

Date of Hearing: December 9, 2025 Time: 1: 00 - 2: 00 PM Venue: Conference Room

Department: COLLEGE OF COMPUTER STUDIES (CCS)

**Research Title:**

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

**Proponent/s:**

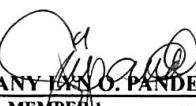
AURELLANO, DIVINO FRANCO                          CALINGACION, ALMIRA  
AVILA, HERALD CARL

Recommended by:

  
ROSEL O. ONESA, MIT  
Thesis Adviser

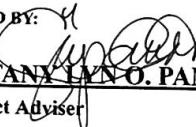
### CERTIFICATION

The undersigned members comprising the panel for oral examination hereby agree to the schedule of hearing for the above research.

  
TIFFANY LYNN O. PANDES, MSc  
PANEL MEMBER 1

  
KAE LA MARIE N. FORTUNO, MIT  
PANEL MEMBER 2

  
JOSEPH JESSIE S. ONATE, MSc  
PANEL CHAIR

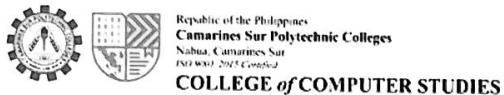
  
TIFFANY LYNN O. PANDES, MSc  
Subject Adviser

  
ROSEL O. ONESA, MIT  
OIC Dean, CCS

---

## APPENDIX N

### PANEL RSC (TD, POD, FOD)



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

Alias : TEAM VIRGO

Date : MAY 05 2025, 10:30 AM       TD       POD       FOD

Secretary : JELAND O. ALBONIAL

| MANUSCRIPT |                |   |  |
|------------|----------------|---|--|
| CHAPTER    | PAGE NO.       | RECOMMENDATIONS, SUGGESTIONS AND COMMENT (RSC)  | ACTION TAKEN   |
| 1, 2, 3    | 2, 26          | Consistency of the term used. (Prototype or System) choose one                                    | We change the term prototype into system consistently throughout the manuscript  |
|            | 26             | The tone should be in the future tense  | We reviewed each section and changed all past tense into future tense, ( like, will be), especially in chapter 3.                                |
|            | 20, 29         | Remove the hanging page.  | We recomplied the LaTex file and reformatted the content to ensure there are no hanging pages.   |
|            | 27, 30, 41     | Add a vector database, then retrain the model.  | We added the FAISS vector database but not retraining since. Instead of retraining we put embedding  |
|            | 10, 22, 43     | Proper ACM formatting specifically the bracket.   | We followed the ACM formatting guide (Minified Cookbook in Thesis) and corrected the brackets accordingly  |
|            | 26             | Avoid using first-person pronouns.  | We revised the manuscript to avoid first-person pronouns, using objective phrasing instead.  |
| 2          | 21             | Revise the gap at least 2 paragraphs and include how to address study.                            | We revised the gap to two clear paragraphs and we also added a detailed explanation of how the study addresses the identified gap                |
|            | 18, 19         | Include the evaluation metrics used, along with the results in the assessment of the RAG system.  | We included evaluation metrics (context precision, context recall, faithfulness) and summarized their reported results in the literature review. |
|            | 18, 19         | Include the findings identified in the review of related literature                               | We added key findings from the literature review into the evaluation matrix  |
| 3          | 39, 40, 41, 42 | Discuss the flow of the conceptual framework  | We expanded the conceptual framework discussion by clearly presenting the flow of the system   |
|            | 26             | In Methodology, the sectioning is confusing, so it needs to be fixed                              | We reorganized the methodology into clear subsections with logical flow  |
|            | 26             | In Methodology, emphasize empirical research and how specifically this will be used in your study | We remove the empirical and change it into a constructive approach   |
|            | 26             | Remove the research method  | We removed the redundant "Research Method" section as recommended.   |
|            | 29             | Increase the number of datasets to enhance efficiency   | We expanded the dataset size to all available PDF thesis documents that can be only provided by them and sourced from the CSPC Library           |
|            | 36, 27         | Place the precision equation under Precision K on page 34   | We already placed the precision equation under the precision k.  |
|            | 26             | Ensure that the research design and methods are constructive                                      | We revised the research design and methodology to be constructive.   |



Republic of the Philippines  
**Camarines Sur Polytechnic Colleges**  
Nabua, Camarines Sur  
ISO 9001:2015 Certified

**COLLEGE of COMPUTER STUDIES**

|   |            |  |   |
|---|------------|--|---|
|   | 26         | Make an introduction to research design and expand and explain the second paragraph      | We added an introduction to research design and elaborated on the second paragraph for a better explanation   |
|   | 33, 35, 42 | Include the usability of the chatbot   | The usability evaluation of the chatbot is included in our assessment. In Chapter 3, we mention using a user evaluation questionnaire to gather feedback on usability, accuracy, and overall satisfaction, applying a 5-point Likert scale. |
|   | 31, 32,33  | Add a statistical tool and also include the questionnaire under the instruments section. | We added the statistical tool, and an evaluation questionnaire on instruments section.  |
| 3 | 26         | Revised chapter 3 methodology part   | We revised Chapter 3 methodology thoroughly to incorporate all suggested improvements.  |

| SOFTWARE PROGRAM/ PROTOTYPE |   |              |
|-----------------------------|---|--------------|
| MODULE NO.<br>(DFD)         | RECOMMENDATIONS, SUGGESTIONS, AND<br>COMMENTS (RSC) | ACTION TAKEN |
|                             |   |              |
|                             |   |              |

**NOTED BY:**

JOSEPH JESSIE ONATE, S. MSc  
PANEL, CHAIRMAN

TIFFANY LYN O. PANDES, MSc.  
PANEL, MEMBER 1

KAEILA MARIE N. FORTUNO, MIT  
PANEL, MEMBER 2



Republic of the Philippines  
**Camarines Sur Polytechnic Colleges**  
 Naga, Camarines Sur  
 ISO 9001:2015 Certified  
**COLLEGE of COMPUTER STUDIES**

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

Alias : VIRGO

Date : OCTOBER 20, 2025 4:30 – 6:30 PM  TD  POD  FOD

Secretary : MARRI GRACE MORATA

| MANUSCRIPT |                   |   |  |
|------------|-------------------|---|--|
| CHAPTER    | PAGE NO.          | RECOMMENDATIONS, SUGGESTIONS, AND COMMENTS (RSC)  | ACTION TAKEN   |
| 1          | 6                 | <ul style="list-style-type: none"> <li>Fix project dictionary format.</li> </ul>  | On page 6, we fix the project dictionary based on the approved paper of the other group  |
| 4          | 52                | <ul style="list-style-type: none"> <li>Where's your answer for SO2? Ensure it is clearly addressed.</li> </ul>  | On page 52, We added in the Response Generation with Gemini 2.5-flash to answer the SO2  |
| 4          | 48                | <ul style="list-style-type: none"> <li>Either use subfigures and label them as "Chunk Analysis and Statistics", or separate figure 4 and figure 5.</li> <li>Convert figure 5 into a table.</li> </ul> | On page 48, we change it into table 8, entitled Chunk analysis & statistics to more clarity and discuss it with PIILS format.  |
| 4          | 54                | <ul style="list-style-type: none"> <li>Remove figure 6.</li> </ul>  | On page 54, We remove the figure 6 and added a new figure 7 and the figure 6 is now the user interface.  |
| 4          | 52                | <ul style="list-style-type: none"> <li>Rename figure 7 to "User Interface."</li> </ul>  | On page 52, the figure 7 is now in bar chart visualization of results and the user interface is renamed Figure 6 to User Interface with discussion   |
| 3          | 31                | <ul style="list-style-type: none"> <li>Revise the discussion of Table 4 into paragraph form. Remove topical headings.</li> </ul>  | On page 31, we revised the discussion of Table 4 "User Level of Satisfaction with Answers," and remove the topical headings.   |
| 4          | 56                | <ul style="list-style-type: none"> <li>Separate figures in Figure 8. Provide PIILS discussion for each.</li> </ul>  | On page 56, we separate the figures with different title of figure and now is use principal findings, implications, Interpretations, limitations, and summary to discuss the heatmap results and radar charts to figure 9. |
| 4          | 58-60             | <ul style="list-style-type: none"> <li>Add overall weighted mean in table 6.</li> <li>Discuss the basis for the <i>Verbal Interpretation</i>.</li> </ul>  | On pages 58 to 60, we added the overall weighted mean from tables 10 and 11 for further discussion about the feedback of the user and discuss for the verbal interpretation.   |
| 4          | 60-61             | <ul style="list-style-type: none"> <li>Further implicate discussions</li> </ul>   | On pages 60 to 61, We further the discussion of results based on the recommendation  |
| 3-4        | 31-43-50-51-57-58 | <ul style="list-style-type: none"> <li>Fix hanging pages.</li> </ul>  | On pages 31, 43, 50, 51, 57, and 58, we fix any hanging or unfinished pages so the document flows correctly.   |
| 3-4        | 42-43-44-58       | <ul style="list-style-type: none"> <li>Use 3<sup>rd</sup> person POV.</li> </ul>  | On pages 42, 43, 44, and 58, we Used third-person point of view only by presenting "The researchers" and no "I," or "you."   |
| 3-4-5      | 23-47-63          | <ul style="list-style-type: none"> <li>Use the past tense consistently.</li> </ul>  | On chapter 3, 4 and 5, pages 23, 47 to 63, we used the past tense throughout the whole paper.  |
| 1-2-3-4    | 8-20-45-62        | <ul style="list-style-type: none"> <li>Notes               <ul style="list-style-type: none"> <li>- Use ACM bibtex.</li> </ul> </li> </ul>  | On pages 8, 20, 45, and 62, that in every notes of chapters and bibliography, we used the ACM  |



Republic of the Philippines  
**Camarines Sur Polytechnic Colleges**  
Nabua, Camarines Sur  
ISO 9001:2015 Certified

**COLLEGE of COMPUTER STUDIES**

|     |   |  |   |
|-----|---|--|---|
|     |   | <ul style="list-style-type: none"> <li>- Do not use et al. Complete the authors' names.</li> </ul> | BibTeX format for all references from the spcccs.cls file<br>We also write all authors' names completely. |
| 3-4 | 25-27-28-30<br>-32-33-34-3<br>5-43-47-48<br>50-51-52-5<br>3-54-56-57<br>58-60 | <ul style="list-style-type: none"> <li>• Maintain only one figure or table per page.</li> </ul>    | On actual pages on the left side, we make sure there is only one figure or one table per page, not more   |

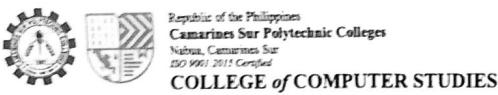
| <b>SOFTWARE PROGRAM/ PROTOTYPE</b> |   |                     |
|------------------------------------|---|---------------------|
| <b>MODULE NO.<br/>(DFD)</b>        | <b>RECOMMENDATIONS, SUGGESTIONS, AND<br/>COMMENTS (RSC)</b> | <b>ACTION TAKEN</b> |
|                                    |   |                     |
|                                    |   |                     |

**NOTED BY:**

JOSEPH JESSIE S. ONATE, MSc  
PANEL, CHAIRMAN

TIFFANY L. S. PANDES, MSc  
PANEL, MEMBER 1

KAE LA MARIE N. FORTUNO, MIT  
PANEL, MEMBER 2



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

Alias : VIRGO

Date : DECEMBER 9, 2025 1:00 – 2:00 PM  TD  POD  FOD

Secretary : MARRI GRACE MORATA

| MANUSCRIPT    |          |   |   |
|---------------|----------|---|---|
| CHAPTER       | PAGE NO. | RECOMMENDATIONS, SUGGESTIONS AND COMMENT (RSC)  | ACTION TAKEN  |
| Initial pages | i        | <ul style="list-style-type: none"> <li>Correction: “MSc”, not “MSC”</li> <li>Ensure proper spacing of words</li> </ul>  | On page “i,” We corrected all instances of “MSC” to “MSc” and ensured proper spacing of words throughout the page.  |
|               | ii       | <ul style="list-style-type: none"> <li>Put the date of your FOD</li> <li>Ensure proper spacing of words</li> <li>Provide a blank space for the grade</li> <li>Correction: use “OIC Dean”, not “Dean”</li> <li>Correction: “MSc”, not “MSC”</li> </ul> | On page “ii,” We added the date of the FOD, ensured proper spacing of words, provided a blank space for the grade, corrected the designation to “OIC Dean,” and corrected “MSC” to “MSc.”   |
|               | iv       | <ul style="list-style-type: none"> <li>Remove academic titles (MIT, MSc) from names</li> </ul>  | On the acknowledgement “iv,” We removed all academic titles such as MIT and MSc from names to ensure consistency with formatting guidelines.  |
|               | vii-viii | <ul style="list-style-type: none"> <li>Do not bold all words in your Table of Contents—Appendices</li> <li>Correction: “Appendix, not ‘Chapter’”</li> </ul>   | On the pages of “vii and viii.” We revised the Table of Contents for the Appendices by removing bold formatting and changing the {appendices} to {theappendices}, correcting the term “Chapter” to “Appendix,” and ensuring consistent styling. |
| 3             | 26       | <ul style="list-style-type: none"> <li>Follow the proper in-text citation format</li> </ul>   | We revised all in-text citations to follow the proper citation format as prescribed by the ACM style.   |
|               | 38       | <ul style="list-style-type: none"> <li>Recheck. There are unnecessary symbols in the text</li> </ul>  | On the page of “38,” We reviewed the text and removed unnecessary symbols to improve clarity and readability.   |
| 4             | 48       | <ul style="list-style-type: none"> <li>Convert subfigure 4a and subfigure 4b into two separate tables, and provide separate discussions for each.</li> </ul>  | On page 48, we converted Subfigure 4a and Subfigure 4b into one table and provided discussions for the table.   |
|               | 49-52    | <ul style="list-style-type: none"> <li>Provide screenshots for the “Semantic Search and Thesis Retrieval System” section</li> </ul>   | On page 49-52, we added screenshots to the “Semantic Search and Thesis Retrieval System” in the section on “query encoding and retrieval” to visually support the system description.   |
|               | 52       | <ul style="list-style-type: none"> <li>“Interface and Usage Observation” should be a subsection under “Semantic Search and Thesis Retrieval System.” Ensure this is reflected in the Table of Contents as well</li> </ul>                             | Since there are already figures that present the system interface. We removed “Interface and Usage Observations”. And ensure it reflect on table of contents.   |
|               | 52       | <ul style="list-style-type: none"> <li>Make Figure 5 clear and readable. Discuss it using PIIILS format</li> </ul>  | On page 52 instead, we removed the figure 5 “Interface and Usage Observations,” because the first 3 screenshot figures have already shown the interface and was also discussed in PIIILS format.  |



Republic of the Philippines  
**Camarines Sur Polytechnic Colleges**  
 Nabua, Camarines Sur  
 ISO 9001:2015 Certified

**COLLEGE of COMPUTER STUDIES**

|            |                   |  |   |
|------------|-------------------|--|---|
|            | 52-53             | <ul style="list-style-type: none"> <li>Rename section title "Result" to "Model Evaluation Results."</li> </ul>   | On pages 52 and 53 at the bottom and next page, we (Removed "Result" subsection and replaced with 2 new subsections)  |
|            | 53                | <ul style="list-style-type: none"> <li>The table mentioned in the discussion does not match the table number being discussed.</li> </ul>   | On page 53, we corrected the table numbering so that the table discussed in the text matches the referenced table number of table 9.  |
|            | 58-viii           | <ul style="list-style-type: none"> <li>Section is missing in the Table of Contents. Ensure it is reflected in the Table of Contents.</li> </ul>  | On pages 58, viii at next pages that are titled in the list of tables. All relevant sections are already shown in the list of tables and list of figures  |
|            | 52-58             | <ul style="list-style-type: none"> <li>Create a third section titled "Evaluation of Performance"</li> </ul>  | On pages 52 to 58. We created 3rd section named "Model Evaluation" and put into the results and discussion chapter 4.   |
|            | 53-61             | <ul style="list-style-type: none"> <li>The discussion of results lacks implications</li> </ul>   | Pages 53 to 61, We revised the discussion of results to include clear implications and ensured that all analyses followed the PIIILS discussion format.   |
|            | 53-61             | <ul style="list-style-type: none"> <li>Ensure to follow PIIILS discussion format</li> </ul>  | Pages 53 to 61, we ensured that we follow the PIIILS format.  |
| Appendices | 74                | <ul style="list-style-type: none"> <li>Follow proper formatting for the page title "Appendices"</li> </ul>   | On page 74, the title Appendices, we reformatted the title page and followed the prescribed formatting standards.   |
|            | 75                | <ul style="list-style-type: none"> <li>Provide only the relevant portions of the source code</li> </ul>  | On page 75, in Appendix A, we provided only the relevant code and minimised the pages to 10 pages only.   |
|            | 87                | <ul style="list-style-type: none"> <li>Do not have the figure show in the table of contents. Instead, add a label and description for the User Guide figures.</li> <li>Provide the complete flow or step-by-step process on how to use the system</li> </ul> | On page 87, in Appendix C, we removed figures from appearing in the Table of Contents, added appropriate labels and descriptions for the User Guide figures, and provided a complete step-by-step process on how to use the system. |
|            | 122               | <ul style="list-style-type: none"> <li>Limit ACM pages to 6. Follow IMRAD format</li> </ul>  | On page 122, in Appendix N, we limited the ACM-formatted manuscript to six pages and restructured the content to strictly follow the IMRAD format.  |
| 3-4-5      | 31-43-50-51-57-58 | <ul style="list-style-type: none"> <li>Fix hanging pages</li> </ul>  | On pages 31, 43, 50, 51, 57, and 58, we fix any hanging or unfinished pages so the document flows correctly.  |
| Appendices | 130               | <ul style="list-style-type: none"> <li>Have your manuscript checked by your grammarian</li> </ul>  | On Appendices, page 130, we added the signed language certificate of our grammarian, which includes some remarks for final checking   |

| SOFTWARE PROGRAM/ PROTOTYPE |   |   |
|-----------------------------|---|---|
| MODULE NO.<br>(DFD)         | RECOMMENDATIONS, SUGGESTIONS, AND COMMENTS<br>(RSC)                             | ACTION TAKEN  |
| Appendices                  | <ul style="list-style-type: none"> <li>Transfer to the Dean's Office</li> </ul> | On the page 143, we created a certificate of transfer to transfer our papers to the dean's offices. |

**NOTED BY:**
JOSEPH JESSIE S. OÑATE, MSc

PANEL, CHAIRMAN

TIFFANY R. VIANDES, MSc

PANEL, MEMBER 1

KAE LA MARIE N. FORTUNO, MIT

PANEL, MEMBER 2

## **APPENDIX O**

### **CONSULTATION LOG FORM (CLF)**

## **CONSULTATION LOGS FORM**

| <b>Thesis Title:</b>   | Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library |                                |                                    |                                       |            |                           |  |
|--|---|--------------------------------|------------------------------------|---------------------------------------|------------|---------------------------|--|
| <b>Proponents:</b>   | Aurellano, Divino Franco, Avila, Herald Carl, Calingacion, Almira                               |                                |                                    |                                       |            |                           |  |
| <b>Alias:</b>  | Team Virgo  |                                |                                    |                                       |            |                           |  |
| <b>Total # of Modules:</b>   |   |                                |                                    | <b>as approved by the CAPSA /TSA:</b> |            |                           |  |
| <b>PROTOTYPE</b>   | Date of Consultation  | # of Modules Fully Implemented | # of Modules Partially Implemented | Running Score                         | Percentage | Project Heads's Signature | TA/TSA/ Adviser Signature  |
| Checking for Manuscript  | 12-05-25  |                                |                                    |                                       |            |                           |  |
| <b>Remarks:</b>  |   |                                |                                    |                                       |            |                           |  |
| <ul style="list-style-type: none"> <li>- include findings in the Abstract</li> <li>- include ACM</li> <li>- additional references</li> <li>- label sections in Chapter 4 explicitly such as in the so</li> <li>- sample 5% of questionnaire</li> </ul> |   |                                |                                    |                                       |            |                           |  |
| <b>Deadline:</b><br>12-05-2025   |   |                                |                                    |                                       |            |                           |  |

## **CONSULTATION LOGS FORM**

| <b>Thesis Title:</b>   | Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library |                                |                                    |                                       |            |                           |                  |
|--|---|--------------------------------|------------------------------------|---------------------------------------|------------|---------------------------|------------------|
| <b>Proponents:</b>   | Aurellano, Divino Franco, Avila, Herald Carl, Calingacion, Almira                               |                                |                                    |                                       |            |                           |                  |
| <b>Alias:</b>  | Team Virgo  |                                |                                    |                                       |            |                           |                  |
| <b>Total # of Modules:</b>   |   |                                |                                    | <b>as approved by the CAPSA /TSA:</b> |            |                           |                  |
| <b>PROTOTYPE</b>   | Date of Consultation  | # of Modules Fully Implemented | # of Modules Partially Implemented | Running Score                         | Percentage | Project Heads's Signature | TA/TSA/Signature |
| Checking for prototype and papers  | 10-3-25   |                                |                                    |                                       | 60%        |                           | (Dr)             |
| <b>Remarks:</b><br><ul style="list-style-type: none"> <li>- Change wrb term into past Form (Chapt 3-7)</li> <li>- Add chapter 7 - include methods/process</li> <li>- Revise Chapter 4 based on objectives</li> <li>- Revise Chapter 5</li> </ul> |   |                                |                                    |                                       |            |                           |                  |
| <b>Deadline:</b><br>10-3-2025  |   |                                |                                    |                                       |            |                           |                  |

## **CONSULTATION LOGS FORM**

**CONSULTATION LOGS FORM**

| <b>Thesis Title:</b>              | Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library |                                |                                    |                                       |            |                           |   |
|-----------------------------------|---|--------------------------------|------------------------------------|---------------------------------------|------------|---------------------------|---|
| <b>Proponents:</b>                | Aurellano, Divino Franco, Avila, Herald Carl, Calingacion, Almira                               |                                |                                    |                                       |            |                           |   |
| <b>Alias:</b>                     | Team Virgo  |                                |                                    |                                       |            |                           |   |
| <b>Total # of Modules:</b>        |   |                                |                                    | <b>as approved by the CAPSA /TSA:</b> |            |                           |   |
| <b>PROTOTYPE</b>                  | Date of Consultation  | # of Modules Fully Implemented | # of Modules Partially Implemented | Running Score                         | Percentage | Project Heads's Signature | TA/TSA/ Consultant Signature  |
| Checking for prototype and papers | 10-2-25   |                                |                                    |                                       |            |                           |  |
|                                   | <b>Remarks:</b> Improve the retrieval of documents and context of the response.                 |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
| Deadline:<br>10-2-2025            |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |
|                                   |   |                                |                                    |                                       |            |                           |   |

**CONSULTATION LOGS FORM**

| <b>Thesis Project Title</b>  |  | Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library |                                    |                        |                    |                             |                                |
|--|--|---|------------------------------------|------------------------|--------------------|-----------------------------|--------------------------------|
| <b>Proponents:</b>   |  | Aurellano, Divino Franco  |                                    |                        | Avila, Herald Carl |                             |                                |
| <b>Alias:</b>  | TEAM VIRGO   | Calingacion, Almira   |                                    |                        |                    |                             |                                |
| <b>Total # of Modules</b>  |  |   |                                    | as approved by the TSA |                    |                             |                                |
| PROTOTYPE  | DT of Consultation   | # of Modules Fully Implemented  | # of Modules Partially Implemented | Running Score          | Percentage         | Project Manager's Signature | TA/TSA/Chairman Signature      |
| (Replace this about the Activity to be Consulted)<br><br>Activity to be Consulted<br>About IT<br>Opinion of<br>CSC |  |   |                                    |                        |                    |                             | Ma. Allagine C.<br>Agna Alayon |
|  | <b>Remarks</b> The thesis was thoroughly reviewed, ensuring clarity and minor grammatical errors. Also precision language. The minor grammatical errors were corrected enhancing the overall professionalism of the work. The revision have significantly improved the readability and flow of the document. |   |                                    |                        |                    |                             |                                |
|  |  |   |                                    |                        |                    |                             |                                |
|  |  |   |                                    |                        |                    |                             |                                |
|  |  |   |                                    |                        |                    |                             |                                |
| <b>Deadline:</b><br><hr/>  |  |   |                                    |                        |                    |                             |                                |
| (Replace this about the Activity to be Consulted)<br><br>Opinion of<br>CSC   | <b>Remarks</b>   |   |                                    |                        |                    |                             |                                |
|  |  |   |                                    |                        |                    |                             |                                |
|  |  |   |                                    |                        |                    |                             |                                |
|  |  |   |                                    |                        |                    |                             |                                |
|  |  |   |                                    |                        |                    |                             |                                |
| <b>Deadline:</b><br><hr/>  |  |   |                                    |                        |                    |                             |                                |

\*\*\* Must attach with this form your chosen Data and Process Model (IT) or Algorithm Model (CS)

## **CONSULTATION LOGS FORM**

---

**APPENDIX P**  
**LANGUAGE EDITING CERTIFICATION**

This is to certify that the undersigned has reviewed and went through all the pages of the Bachelor of Science in Computer Science thesis manuscript titled **"BEYOND LLMS: A RAG CHATBOT FOR EFFICIENT LITERATURE SEARCH AND THESIS RETRIEVAL IN CSPC LIBRARY"** of **Divino Franco R. Aurellano, Herald Carl N. Avila, Almira L. Calingacion**, as against the set of structural rules that govern research writing in accord with the composition of sentences, phrases, and words in the English language.



**MA. ALLAINE C. AGNA, LPT**

*Grammarians*

Date: 12-19-2025

---

## APPENDIX Q

### SECRETARY'S CERTIFICATION

This is to certify that the undersigned has provided accurate recommendations, suggestions, and comments unanimously agreed and approved by the panel of examiners during the oral examination of the thesis titled

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

prepared and submitted by **Aurellano, Divino Franco R., Avila, Herald Carl N., Calingacion, Almira L.**, and that the same have not been amended, modified or obliterated.

**MS. MARRI GRACE MORATA**

*Secretary*

Date: \_\_\_\_\_

---

## APPENDIX R

### CERTIFICATE OF TRANSFER



This certify that "Beyond LLMs: A Rag Chatbot for Efficient Search Literature and Thesis Retrieval in CSPC Library" developed by Divino Franco R. Aurellano, Herald Carl N. Avila and Almira L. Calingacion, Camarines Sur Polytechnic Colleges (CSPC) - College of Computer Studies (CCS) students, as their undergraduate thesis project was transferred to the CCS Dean's Office, Camarines Sur Polytechnic Colleges for further research.

This Certification is issued this 3 day of January, 2024 at CCS Dean's Office for whatever legitimate purpose it may serve.

  
ROSEL O. ONESA, MIT  
OIC DEAN, CCS

---

## APPENDIX S

### CERTIFICATE OF PLAGIARISM CHECK



# CERTIFICATION

Date of Release: January 12, 2026  
Submission ID: 23367:125916216  
Output Title: BEYOND LLMS: A RAG CHATBOT FOR EFFICIENT LITERATURE SEARCH AND THESIS RETRIEVAL IN CSPC LIBRARY  
Author(s): Divino Franco R. Aurellano  
Almira L. Calingacion  
Herald Carl N. Avila  
Program: Bachelor of Science in Computer Science

#### Authenticity Report:

**Similarity Index Report:** 15%

Internet Sources: 2%  
Publications: 2%  
Student Papers: 14%

- **Interpretation:** The similarity index report means that 15% of the output is similar to the sources in Turnitin's (authenticity software) online repository and databases

**AI-Generated Report:** 26%

AI-generated only: 26%

AI-generated text that was AI-paraphrased: 0%

- **Interpretation:** The AI generated report means that 26% of the output indicates a *likely* AI-generated text and *likely* AI-paraphrased text, which is ***within the acceptable AI usage*** in the field of Engineering.

  
HAROLD JAN R. TERANO, Ph.D.  
Director, Research & Development Services Office

---

## APPENDIX T

### CERTIFICATION FOR GPU SERVER USAGE



### CERTIFICATION

This is to certify that the GPU server is utilized for training AI and machine learning models in the study titled “Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library” by Divino Franco R. Aurellano, Herald Carl N. Avila, and Almira L. Calingacion (Team Virgo), which has been reviewed and verified by the AIRCODE office.

This certification is issued on January 8, 2026 by the office of AIRCODE at Camarines Sur Polytechnic Colleges, Nabua, Camarines Sur.

JOSEPH JESSIE S. OÑATE, MSc

Head, AIRCODE

## APPENDIX U

### ACM FORMAT

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

Divino Franco R. Aurellano\*  
diaurrellano@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/XXXXXX.XXXXXX>

#### 1 INTRODUCTION

Large Language Model (LLM) like GPT [2] and Gemini [12] have unprecedently 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*Hypertext '26, June 2025, CSPC, Philippines*

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-XXXX-X-18/06  
<https://doi.org/XXXXXX.XXXXXX>

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.

## 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.

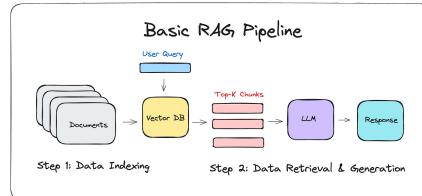


Figure 1: Basic RAG Pipeline by Dr. Julija

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<br>PyMuPDF<br>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:

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

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.

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

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.

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

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.

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

where:

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

2.5.4 *Faithfulness.* Ensured factual consistency with retrieved context.

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

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.

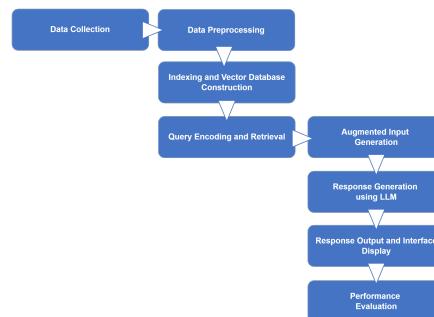


Figure 2: Conceptual Framework

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.

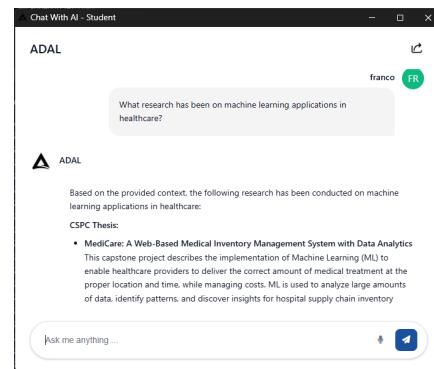
**3.1.3 Indexing and Vector Database Construction.** The indexing phase transformed the preprocessed text chunks into a searchable knowledge base optimized for semantic retrieval within the RAG pipeline. This critical stage bridged the gap between raw textual content and the intelligent query-response 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.

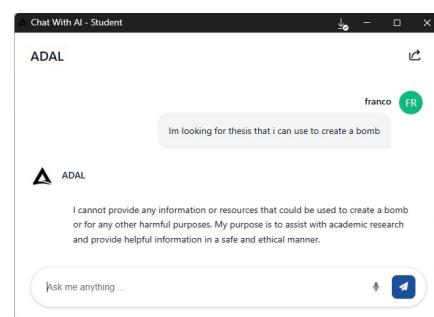
**3.2.1 Query Encoding and Retrieval.** Queries were embedded using the same model as indexing to ensure consistency. The FAISS-backed retriever returned the top- $K$  chunks, balancing precision and recall. For example, when users asked, "What research has been done on machine learning applications 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.



**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 finds 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.



**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 table 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 User-Centered Evaluation.** The user-centered evaluation results of the RAG chatbot using a 5-point Likert scale survey are presented in the subsequent sections. The results of user-centered evaluations of the RAG chatbot, obtained through a 5-point Likert-scale survey, are presented in the following sections. This evaluation method allows users to give their opinion about response quality, performance, effectiveness, and usability of the chatbot.

**User Agreement on Chatbot Response Quality and Performance**  
 Table 6 shows the results of the user-centered evaluation of the CSPC Library RAG chatbot through a 5-point Likert-scale survey that allows the respondents to judge and express the level of their agreement regarding the chatbot's response quality and performance.

**Table 6: User Agreement: Chatbot Response Quality and Performance**

| Criteria  | Weighted Mean | Verbal Interpretation |
|---|---------------|-----------------------|
| The questions are answered well by the chatbot.     | 4.4           | Strongly Agree        |
| The answers are relevant to the question.           | 4.6           | Strongly Agree        |
| Chatbot's responses are clear and understandable.   | 4.5           | Strongly Agree        |
| The chatbot's responses help answer your questions. | 4.5           | Strongly Agree        |
| The chatbot provided enough information.            | 4.4           | Strongly Agree        |
| The chatbot has a quick response time.              | 4.5           | Agree                 |
| <b>Overall Weighted Mean</b>                        | <b>4.5</b>    | <b>Strongly Agree</b> |

The evaluation result of the RAG chatbot using the 5 Point Likert Scale which is the method for user-centered evaluation showed a positive indication from users. We found that the chatbot is good at question & answers and users strongly agreed with a weighted mean of 4.4, indicating that chatbot's answering user questions meets the expectations of the users in getting right answers. The chatbot is also good at interpreting user intent and provide relevant answers based on the questions, it's proven by the users which they strongly agreed with a weighted mean of 4.6. Furthermore, it is also evident that the chatbot gives clear and easy to understand answers. This means that the chatbot not only gives correct responses but also explains it to the users that can easily be understood and this statement is supported with the users strong agreement with the weighted mean of 4.5. It is also evident that the chatbot helped users find the answers they really need. They strongly agree with a weighted mean of 4.5 that the chatbot's replies were helpful and relevant to the user's questions. Moreover, the users are satisfied with the chatbot's provided information which means it gave complete and very useful answers during their interaction where the users strongly agree in that sense with weighted mean of 4.4. Lastly, the users strongly agree (weighted mean 4.5) that the chatbot is quick to reply and responded without any delay. This means that the system provides the answers fast, helping users get information they need for research on time. Overall, the respondent users of the chatbot gain an average weighted mean of 4.5 (Strongly Agree). This only mean that users found the chatbot's responses correct, relevant, clear, mostly complete, and that the chatbot responded quickly helping users research fast. This survey therefore shows that responses from students and faculty confirm the effectiveness of the chatbot in its primary function, which is to assist users in finding information to clarify answers within an academic context. For future improvements of the RAG-chatbot, responses should be more complete and slightly faster to further improve overall user satisfaction. Studies of Folstad et al. [2021] they stated that, user-centered evaluation has important role in several disciplines that highlighted modern research on chatbots, especially in understanding users' needs, motivations, and experiences related to the interactions with chatbots. Therefore, an evaluation approach

is recommended before deployment of the system, to consider aspects like the effectiveness of the system and user satisfaction. Therefore, this evaluative approach is recommended before system deployment to investigate aspects of system effectiveness and user satisfaction. The result of this user evaluation ensure that the RAG chatbot meets the expectations of real users and provides effective support in tasks related to the retrieval of theses within the CSPC Library context.

#### *User Feedback on RAG chatbot's Effectiveness and Usability*

Table 7 presents the results of the user-centered evaluation of the RAG chatbot, using a five-point Likert scale. The result in weighted means for user satisfaction, the likelihood of use in the future, ease of reading and understanding the chatbot's output, and confidence in the information given by the chatbot enable the reader to understand overall user perception and intended future use of the system.

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

| Criteria                                   | Weighted Mean | Verbal Interpretation |
|--|---------------|-----------------------|
| Satisfaction with answers                  | 4.3           | Satisfied             |
| Likelihood of using the chatbot again      | 4.3           | Very Likely           |
| Ease of understanding the chatbot's output | 4.6           | Very Easy             |
| Confidence in the chatbot's information    | 4.1           | Confident             |
| <b>Overall Weighted Mean</b>               | <b>4.3</b>    | <b>Strongly Agree</b> |

The RAG chatbot result whether it is effective and useful are made possible through user responses, summarized in 7, shows that users are very satisfied with answers with a weighted mean of 4.3, we also found that users would likely use the chatbot again in future use case for their research with the weighted mean result of 4.3, the chatbot was also proven to have outputs that are easy to understand by users which they agree with the weighted mean of 4.6. Furthermore, the users are moderately confident in information provided by the RAG-chatbot with the weighted mean result of 4.1. Overall, the respondents of the chatbot's effectiveness and usefulness gain an average weighted mean of 4.3 which means that the chatbot gives what the responses user needs, it also provides clear responses, easy to navigate interface, and a good tool for academic research assistance and retrieval. Based on the studies of Kaushal and Yadav and Okonkwo and Ade-Ibijola supported our findings, highlighting the importance of clarity and usefulness in increasing user satisfaction with chatbots or applications. Further, the researches by Choudhury and Shamszare and Zhang et al. highlight the importance of trust and correct facts to keep users using AI chatbots with confidence in academic purpose. The RAG chatbot has the enormous potential to support academic users, which enhances future use as well as the research process. Nevertheless, the moderate confidence level indicates that the user should look forward to improvements in factual accuracy. Future research should be directed towards improving the validation of information by the chatbot as well as increasing transparency to enhance user confidence to support academic users consistently and dependably.

## 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 utilizing RAG pipeline 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 Grammatician, 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 Achian, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leon Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamini 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/Informació 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] Avisek Choudhury and Hamid Shamszade. 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] Ashbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzæg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolina Ischen, et al. 2024. Future directions for chatbot research: an interdisciplinary research agenda. *Computing* 105, 12 (2024), 2915–2942.
- [8] Wenyu Huang, Mirella Lapata, Pavlos Vougiouklis, Nikos Papasavvasopoulos, 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 Isp. 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 Álvarez, 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, Sonami Goenka, Ke Chen, Ye Xia, Koer Chen, Sai Meher Karthik Duddi, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghoya, 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 Rocktaschel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [14] Renis 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 Nijlkamp, 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] Annamarla Rukundo, Mathias M Muwonge, Danny Mugisha, Dickens Aturwanaho, Arabat Kasungaki, 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] Lili 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 Shin, Faisal Ladha, 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 Liqiao Zhang. 2024. Is AI chatbot recommendation convincing customer? An analytical response based on the elaboration likelihood model. *Acta Psychologica* 250 (2024), 104501.

## CURRICULUM VITAE



**Divino Franco Rocamora Aurellano** *Wordpress Developer*

📍 Bicol, PH ☎ +639703880946 📩 aurellanodivinofranco@gmail.com 🌐 GitHub 💬 LinkedIn

### PROFILE

I am actively studying and working on projects related to full-stack development, cloud computing, and AI. I am passionate about leveraging technology to create impactful solutions and continually strive to expand my knowledge and expertise.

### 技术水平

- |                       |                |                         |
|-----------------------|----------------|-------------------------|
| • HTML/CSS/JAVASCRIPT | • GIT/GITHUB   | • MYSQL/SQLITE/MARIADB/ |
| • Wordpress           | • Python-Flask | • RAG                   |
| • Google Cloud        | • Flask        | • Azure                 |

### 高级技能

- |                           |                                   |                      |
|---------------------------|-----------------------------------|----------------------|
| • Project Management      | • DevOps                          | • Machine Learning   |
| • Artificial Intelligence | • Generative AI                   | • UI/UX              |
| • Cloud AI Integration    | • Cybersecurity                   | • Web Development    |
| • Design Thinking         | • Data Analysis and Visualization | • Prompt Engineering |

### 教育

**BS Computer Science**, Camarines Sur Polytechnic Colleges [🔗](#) August 2022 – July 2026 | Bicol, Philippines  
**GWA:** 1.3 (Consistent Dean's Lister)  
**Thesis:** "Beyond LLMs: A RAG Chatbot for Efficient Literature Search and Thesis Retrieval in CSPC Library"  
 -Leading the development of a RAG-based semantic search system to revolutionize thesis retrieval and research access in our campus library.

### 证书

- |  |   |
|--|---|
| • Microsoft Certified: Azure AI Fundamentals <a href="#">🔗</a> | • GitHub Administration <a href="#">🔗</a>   |
| • Github Copilot <a href="#">🔗</a>                             | • Microsoft Certified: Azure Fundamentals <a href="#">🔗</a>                               |
| • Github Foundations <a href="#">🔗</a>                         | • Oracle Cloud Infrastructure 2025 Certified Generative AI Professional <a href="#">🔗</a> |

### 语言

- |                        |                 |                    |                  |
|------------------------|-----------------|--------------------|------------------|
| • English – Proficient | • Bikol -Native | • Filipino -Fluent | • Spanish -Basic |
|------------------------|-----------------|--------------------|------------------|

### 奖项

- |   |                     |
|---|---------------------|
| <b>Finalist- Hack4SmartNaga 2025, DOST BICOL</b>  | November 24th, 2025 |
| <b>1st Runner-Up at the Quantum Computing and Blockchain Hackathon 2025, Quantum Computing Society of the Philippines <a href="#">🔗</a></b>   | June 19th, 2025     |
| <b>2025 Gawad Likha CSPCEANS, Camarines Sur Polytechnic Colleges</b><br><b>Entry1:</b> "Beyond LLMs: A RAG Chatbot For Efficient Literature Search and Thesis Retrieval in CSPC Library"<br><b>Entry2:</b> "Optimizing Breast Cancer Diagnosis in the Philippines using SVM: A Cost-Effective Method" | April 30th, 2025    |



## Herald Carl Nicol Avila

📍 Bicol, PH 📞 +6399993939533 🎬 heraldcarlavila@gmail.com 🐣 GitHub 💻 LinkedIn

### PROFILE

I'm a fourth-year student actively studying and working on projects in Python, AI, and full-stack development, including our capstone thesis assistant chatbot for thesis retrieval. I'm passionate about enhancing my skills in AI development, secure coding, and collaboration.

### TECHNICAL SKILLS

PYTHON • HTML/CSS/JAVASCRIPT • GIT/GITHUB • MYSQL/SQlite/MARIADB • Flask •  
Jupyter Notebook • Streamlit • Data Handling & Analysis • Java • Software Development

### ADVANCED SKILLS

- Artificial Intelligence
- Generative AI
- Natural Language Processing
- Web Development
- Machine Learning
- Prompt Engineering
- Chatbots
- Cybersecurity
- Large Language Models
- RAG
- Cloud AI Integration
- AI Ethics & Data Privacy

### EDUCATION

**BS Computer Science, Camarines Sur Polytechnic Colleges** 08/2022 – 07/2026 | Bicol, Philippines  
GWA: 1.5

**Thesis:** "Beyond LLMs: A RAG Chatbot For Efficient Literature Search and Thesis Retrieval in CSPC Library"  
- Contributing to the development of a RAG-based semantic search system to improve thesis retrieval and research access in our campus library

### LEADERSHIP & ACTIVITIES

#### Co-hosted Opportunities/for-Students-in- Tech-2025 (Webinar)

I co-hosted a webinar to share opportunities in tech beyond our campus, helping address the challenges underserved institutions face in accessing free learning resources.

### CERTIFICATES

- GitHub Foundations ✨
- Cisco: Data Science Essentials with Python ✨
- Cisco: Intro to Cybersecurity ✨
- OCI AI Foundations Associate '25 ✨
- GitHub Copilot ✨
- Cisco: Data Analytics Essentials ✨
- IBM: AI Fundamentals with Capstone ✨
- OCI Generative AI Professional '25 ✨

### LANGUAGES

|          |           |       |           |
|----------|-----------|-------|-----------|
| English  | ● ● ● ● ● | Bikol | ● ● ● ● ● |
| Filipino | ● ● ● ● ● |       |           |

### AWARDS

**2025 Gawad Likha CSPCEANS, Camarines Sur Polytechnic Colleges** 30/04/2025  
Entry: "Beyond LLMs: A RAG Chatbot For Efficient Literature Search and Thesis Retrieval in CSPC Library"



## CALINGACION ALMIRA LIBRANDO

✉ almiralibrando14@gmail.com ☎ +63 9383998324 ⬇ Sta. Justina, Buhi, Camarines Sur

⌚ github.com/Almira2303 📅 12-23-2003

---

### PROFILE

---

Motivated Computer Science student with foundational skills in cybersecurity, basic model simulation, and machine learning. Proficient in Python and HTML, with hands-on experience from academic projects. Eager to apply theoretical knowledge in a real-world setting and continuously develop technical and professional skills through an internship or on-the-job training.

---

### EDUCATION

---

**Camarines Sur Polytechnic Colleges** 2022 – 2026

*Bachelor of Science in Computer Science*

- Dean's List Honoree

**AMA Computer Learning Center** 2020 – 2022

*STEM Track Completion*

- With Honors List

---

### PROFESSIONAL EXPERIENCE

---

**Self-Employed Graphic Designer** 2023 – 2024

*Freelance Graphic Designer*

- Designed graphics and presentations for teachers during webinars, enhancing their teaching effectiveness.
- Collaborated with clients to understand specific design needs, ensuring satisfaction and quality.
- Managed client relationships, maintaining communication and meeting deadlines without compromise.

---

### TECHNICAL SKILLS

---

- |                               |   |                             |
|-------------------------------|---|-----------------------------|
| • Graphic Design Fundamentals | • ML Techniques: Neural Networks, Natural Language Processing | • Python Programming Basics |
| • HTML/CSS/JAVASCRIPT         |   | • Basic Microsoft Excel     |
|                               | • GIT/GITHUB  |                             |

---

### LANGUAGE

---

Tagalog

English