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

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

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

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

Abstract

This study presents the development of a Retrieval-Augmented Generation (RAG) chatbot designed to enhance literature search and thesis retrieval within the CSPC Library. By integrating advanced natural language processing techniques with a robust retrieval system, the chatbot aims to provide users with efficient access to academic resources. The system leverages large language models (LLMs) to understand user queries and retrieve relevant documents from the library's database. "Findings"

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 '25).* ACM, New York, NY, USA, 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

This chapter provides an overview of the study, covering the challenges of the campus library, the study's objectives, and its significance. It defines the problem, outlines research goals, and highlights the proposed system's potential impact. The scope and limitations clarify its boundaries, while the project dictionary and notes offer essential terms and supporting details.

1.1 Background of the Problem

Large Language Models (LLMs) such as OpenAI's ChatGPT [2] and gemini [29] have made significant advancements in Natural Language Processing (NLP) by excelling in diverse tasks such as semantic search, classification, and clustering, providing more accurate, context-aware results than keyword-based approaches. [13, 38] In addition, these advancements have benefited various fields, including academic research. However, LLM responses can depend heavily on the data on which the model was trained, and they cannot retrieve real-time or external information beyond their pretrained knowledge. This makes them less effective for tasks that require up-to-date, specific institutional data, such as retrieving current academic resources in university libraries [34].

Writing an academic paper requires a deep understanding of the subject and a significant amount of related literature for credible evidence, which can be challenging and time-consuming [25]. It is essential to first visit the university library to search and gather existing related literature significant to the researcher's 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 offer limited 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 [42].

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 [9, 28, 42]. Unfortunately, most digitalized libraries today still use outdated search systems that need an exact keyword search, which can result in irrelevant materials [48]. The current search mechanism of this digital archives including the Camarines Sur Polytechnic Colleges (CSPC) library still heavily depends on traditional keyword-matching algorithms. If users do not input the exact title or precise keywords, the system returns "not found" even when relevant content exists. This limitation highlights a deeper issue in search functionality, where vague or topic-based queries cannot retrieve appropriate materials, thereby hindering access to valuable research. This inefficiency in retrieval presents a serious barrier for the academe community, particularly when conducting time-sensitive or exploratory academic work.

These challenges of university libraries in the Philippines, including the CSPC library, have shared difficulties in accessing academic resources, outdated search systems, and ineffective information retrieval that affect the efficiency of academic research. While numerous studies have also explored the integration of the emerging

Hypertext '25, June 2025, CSPC, Philippines

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

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 [10, 53].

To overcome these challenges, Retrieval-Augmented Generation (RAG) has emerged as a superior approach [30]. Unlike standalone LLMs, which require retraining and additional domain-specific data to adjust LLM weights, RAG presents an advanced approach by retrieving relevant external information to generate responses and holds significant practical implications for university libraries by improving search functionalities. Additionally, RAG ensures that the most relevant academic resources are retrieved quickly and straightforwardly, making it suitable for libraries with expansive collections of academic papers that are difficult for researchers and students to navigate [22, 56].

This thesis developed an enhanced LLM-powered chatbot with the integration of the Retrieval-Augmented Generation (RAG) technique to improve information retrieval, especially in literature search and thesis retrieval of university-owned thesis PDFs at the Camarines Sur Polytechnic Colleges (CSPC) Library. This chatbot application generates answers and retrieve relevant documents based on the user's prompt.

1.2 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. Furthermore, library policies restrict users from taking thesis books outside the premises, limiting accessibility to essential academic resources. In response to these challenges, this study aims to explore creating a chatbot that eliminates those limitations by enabling searches based on topics, keywords, or even general descriptions. Additionally, the goal is to make this accessible everywhere.

This goal, with the use of the Retrieval Augmented Generation (RAG) algorithm, will revolutionize how the academe community interacts with the CSPC library, making research faster, smarter, and more user-friendly. By bridging the gap between traditional archiving and modern AI-driven retrieval, the system ensures that valuable institutional knowledge is readily accessible to all.

1.3 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.

1.3.1 General Objective. The general objective of this study is to develop a chatbot for Camarines Sur Polytechnic Colleges (CSPC) library, using Retrieval-Augmented Generation (RAG) to enhance thesis retrieval and literature search in CSPC Library, replacing the traditional keyword-based search with a more conversational and topic-oriented search and response approach.

1.3.2 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 RAGASS and user satisfaction metrics.

1.4 Significance of the Study

The result of this study will benefit the following:

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

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

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

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

Future Developers. This study serves as a technical reference for students and developers interested in Natural Language Processing and Large Language Models. The architectural design and implementation details can guide future software development projects within the college.

1.5 Scope and Limitation

The scope of this study is to develop a chatbot for the Camarines Sur Polytechnic Colleges (CSPC) library, utilizing Retrieval-Augmented Generation (RAG) with Google Gemini LLM. The goal is to address the challenges faced by the academic community in searching and retrieving thesis literature by replacing the current yet 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

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 will depend on the quality and structure of thesis records, and on the clarity and relevance of the user's prompts. Additionally, computational efficiency may depend on the configuration and resources allocated in the cloud environment, which could affect the chatbot's real-time processing capabilities. Finally, while the RAG technique can reduce hallucination, users are advised to validate the outputs carefully as occasional inaccuracies or fabricated information may still occur.

1.6 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 scholarly documents, such as research papers and theses, to support academic work [47]. In this study, the implementation of LLMs is essential to improve the retrieval of academic literature.
- **Chatbot.** An AI-powered conversational agent designed to interact with users in natural language, providing assistance, answering queries, and facilitating access to information in a user-friendly manner [15]. In this study, chatbots will be implemented for answering questions with human-like responses.
- **CSPC Library.** The Camarines Sur Polytechnic Colleges (CSPC) Library serves as the primary academic resource center for students and faculty. It offers access to a diverse collection of books and theses inside the premises. The library has initiated steps toward digitalization, providing an online catalog for users to search materials. In this study, the CSPC Library is examined to assess its current digital infrastructure and explore enhancements to improve information retrieval and user experience.
- **Google Gemini.** A state-of-the-art Large Language Model (LLM) developed by Google, designed to understand and generate human-like text based on extensive training data [29]. In this study, Google Gemini was utilized as the core LLM for implementing the RAG technique to enhance information retrieval and response generation in the chatbot system.
- **Generative AI.** A kind of artificial intelligence that may produce original text, graphics, or code, frequently in response to a user-inputted prompt. More and more online applications and chatbots that let users enter instructions or inquiries into an input box are using its models. The AI model will produce a response in the output field that resembles a human response [11]. In this study, the implications of Generative AI in the context of education and academic integrity will be examine.
- **Large Language Models (LLMs).** AI models trained on vast text datasets to understand and generate human-like responses. They excel in natural language tasks but struggle with retrieving real-time and domain-specific information [27]. In this study, the implementation of the Large Language Model (LLM) streamlines access to information, assists in literature searches, and facilitates query handling effectively.
- **Natural Language Processing (NLP).** A subfield of artificial intelligence (AI) called natural language processing (NLP) makes it possible for computers to comprehend and understand spoken, written, or even handwritten human language. NLP is essential to enabling seamless and organic human-computer interactions as AI-driven technologies grow more pervasive in daily life [43]. In this study, NLP significantly improves machine comprehension to understand human language and improves user interaction through chatbots.
- **Retrieval-Augmented Generation (RAG).** An AI framework that enhances LLMs by incorporating an external knowledge retrieval mechanism, improving the accuracy and contextual relevance of generated responses [30]. In this study, RAG was developed for navigating and retrieving information from large amounts of academic papers.
- **Semantic Search.** A search approach that goes beyond keyword matching by understanding the intent and contextual meaning of queries to return more relevant results [36]. In this study, semantic search will significantly improve the performance of RAG in generating relevant and contextual responses due to the enhanced retrieval process by understanding user queries which is beneficial for the CSPC library that holds a large collection of academic papers.

2 Related Literature and Studies

This chapter presents the analysis of relevant literature and existing systems associated with the study. It includes a summary of related works, a synthesis of the state of-the-art technologies and methodologies, and identifies the research gaps addressed by the current study.

2.1 Review of Related Literature and Studies

To develop a deeper understanding of the research topic, a comprehensive review of books, scholarly articles, journals, and previous thesis projects was conducted. The findings are organized thematically to align with the key areas of the study.

2.1.1 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 [57] [58]. For instance, the recent studies of Khraisha et al. 2024 and Gartlehner et al. 2023 reveal that LLMs are capable of automating processes like systematic review, data extraction, and document screening, which demonstrate the capability and potential of LLMs in enhancing the efficiency of academic research [26] [19].

While large language models (LLMs) offer advantages for information retrieval, they also come with challenges. One major challenge is that their inefficient when applied to domain-specific

tasks that require specialized knowledge. This limitation occurs because of the models' dependency on their pre-trained knowledge, which limits them from providing factual answers for specific domains, like in Academe. Omar et al. discussed that LLMs, such as ChatGPT, serve as complementary tools in specialized scenarios but may struggle with complex queries due to a lack of exposure to field-specific training data [26]. Additionally, pre-trained LLMs encounter challenges in keeping up with constant expansions of data in various domains, which makes them incapable of updating their knowledge without extensive fine-tuning. Lucas et al. highlighted that for applications in academic and professional settings, the inability of LLMs to access current domain-specific repositories reduces their effectiveness and utility [19].

While LLMs stand at the forefront of NLP innovation, substantial limitations arise in their application to domain-specific tasks. These include real-time data retrieval, pre-trained knowledge bases, and ethical considerations surrounding data privacy. Addressing these challenges through innovative approaches like RAG can help leverage the models' capabilities, ensuring they can meet the rigorous demands of specialized applications.

2.1.2 Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) has conveyed notable progress in information retrieval (IR), especially in the context of literature search and thesis retrieval in library systems [55]. The concept integrates traditional large language models (LLMs) with external knowledge sources to enhance response relevance, richness, and correctness [12].

Lewis et al. 2020, in their influential study "Retrieval-Augmented Generation for Knowledge Intensive NLP Tasks," emphasized that RAG enables more precise responses by overcoming the inherent limitations of LLMs, particularly regarding accurate knowledge retrieval and contextual relevance. Extending this, Shuster et al. 2021, in their study "Retrieval Augmentation Reduces Hallucination in Conversation," showed that RAG reduces inconsistencies and hallucinations in LLM responses. Their findings indicated that RAG mechanisms significantly improved conversational fluency and integrity, especially in open-domain contexts, resulting in more knowledgeable and coherent outputs.

Sagi 2024, study "GENAI: RAG Use Cases with Vector DB to Solve the Limitations of LLMs," further reinforced this by demonstrating that combining vector databases with RAG significantly enhances retrieval speed and relevance. Particularly in dynamic domains like academic and business libraries, the semantic search capabilities of vector databases support continuous real-time updates, greatly improving knowledge management and the factuality of generated responses. Thus, RAG not only strengthens the retrieval capabilities of LLMs but also substantially mitigates their traditional weaknesses in consistency and factual accuracy [46].

2.1.3 Document Ingestion and Retrieval. The performance of Retrieval Augmented Generation (RAG) systems depends on efficient document use and retrieval procedures, especially when working with large, complicated datasets like academic libraries. Any type of data source, including text, video, images, and audio, can be used with retrieval-augmented generation (RAG) systems, allowing for flexible and contextually rich information retrieval. In this study, the researchers focused on utilizing PDF documents as the primary corpus for academic content extraction [31].

The effectiveness of RAG systems heavily depends on the quality of preprocessing, which involves converting unstructured PDF data into machine-readable formats suitable for embedding and semantic search [8] [7]. Tools such as PyPDF2, PyMuPDF, and pypdfium are commonly employed for this task, enabling the extraction of raw text from complex PDF layouts [3].

Sagi 2024, study "GENAI: RAG Use Cases with Vector DB to Solve the Limitations of LLMs," further reinforced this by demonstrating that combining vector databases with RAG significantly enhances retrieval speed and relevance. Particularly in dynamic domains like academic and business libraries, the semantic search capabilities of vector databases support continuous real-time updates, greatly improving knowledge management and the factuality of generated responses. Thus, RAG not only strengthens the retrieval capabilities of LLMs but also substantially mitigates their traditional weaknesses in consistency and factual accuracy [46].

Adhikari and Agarwal 2024 evaluated several PDF parsers using F1 score, BLEU-4, and local alignment across diverse document categories. Their study revealed that PyMuPDF and pypdfium consistently preserved sentence structure and layout more accurately than other tools. These capabilities are essential for maintaining the necessary semantic coherence for accurate vectorization and retrieval. They also highlighted parsing difficulties in complex documents such as scientific and patent PDFs, where rule-based tools struggled while transformer-based models demonstrated significant improvements. Moreover, efficient document ingestion and retrieval are crucial in managing large repositories such as academic libraries [3].

According to Zhang et al. 2023, automated ingestion pipelines that parse and store documents in a searchable index improve the discoverability and accessibility of scholarly content.

Techniques like optical character recognition (OCR), metadata extraction, and structured indexing are often applied to thesis repositories to facilitate retrieval operations [59]. Similarly, Karpukhin et al. 2020 emphasized the importance of pre-processing, chunking, and embedding documents for semantic search in their work on Dense Passage Retrieval (DPR), informing modern RAG pipelines [23]. Typically, the ingestion process involves multiple steps: (1) text extraction using tools like PyMuPDF or pypdfium, (2) text chunking into smaller, logical parts, and (3) embedding using models like Sentence-BERT. Finally, these vectors are stored in specialized vector databases such as FAISS, Pinecone, or FAISS for efficient retrieval during user queries. Efficient document ingestion and storage directly influence retrieval accuracy, system responsiveness, and user experience. Sagi emphasized that robust ingestion and vectorization processes ensure that relevant information can be retrieved quickly and that RAG models generate highly accurate, contextually rich responses, especially in dynamic environments like academic libraries [23].

Deepak et al. 2025, in their study "Langchain-chat with my pdf" highlighted the significance of vectorization techniques such as embedding and chunking in processing PDFs. Their research illustrated how chunking aids the RAG framework in identifying relevant sections of documents during user queries, streamlining

the management of comprehensive PDF-based information, and enhancing the system's semantic search capabilities [17].

In conclusion, the studies collectively highlight that robust preprocessing, ingestion, and vectorization processes are foundational for bridging the gap between static document repositories and real-time information retrieval, demonstrating the potential of RAG architectures in managing large collections of academic knowledge [4] [7].

2.1.4 RAG Applications in Various Domains. Beyond academic contexts, RAG frameworks are increasingly being applied to specialized domains such as legal research, medical retrieval, and scientific literature search, highlighting their wide versatility and impact.

In the academic domain, Grigoryan and Madoyan 2024, in their study "Building a Retrieval-Augmented Generation (RAG) System for Academic Papers," developed a RAG-powered system that significantly enhanced academic literature retrieval using vector search techniques like cosine similarity and HNSW indexing [20]. Similarly, Song et al. 2024 emphasized that RAG frameworks not only improve search capability but also boost academic outputs by integrating external knowledge into LLMs, leading to more accurate and efficient information retrieval for students and researchers [52]. Their findings align with those of Karpukhin et al. 2020, who also reported that better information retrieval accuracy correlates with improved search results and question-answering performance [23].

In the healthcare domain, Arzideh et al. 2024, in "MIRACLE - Medical Information Retrieval using Clinical Language Embeddings for Retrieval Augmented Generation at the Point of Care," demonstrated the effectiveness of RAG systems integrated with domain-specific clinical embeddings [8]. Their approach greatly improved clinical decision-making, supported efficient documentation workflows, and offered greater personalization in healthcare information access. Supporting this, Amugongo et al. 2024 showed that RAG systems could successfully retrieve external medical data to generate highly accurate, reliable responses, surpassing traditional LLM limitations [6].

In the legal field, Aquino et al. 2024, in their study "Extracting Information from Brazilian Legal Documents with Retrieval Augmented Generation," illustrated that RAG systems significantly optimize legal research by speeding up case law retrieval and improving the authenticity and contextual accuracy of outputs [7].

Finally, recent advancements such as Google Gemini, a state-of-the-art LLM, demonstrate that when integrated with RAG mechanisms Prabhulal 2025, LLMs can attain improved semantic understanding and retrieval precision [41]. In parallel, vector search offers a robust foundation for developing intelligent, document-aware systems. By combining high-quality semantic embeddings with indexing, this approach ensures that responses remain accurate, transparent, and firmly anchored in domain-specific data rather than relying solely on general model knowledge.

2.1.5 Evaluation of Retrieval-Augmented Generation (RAG) Systems. The evaluation of Retrieval-Augmented Generation (RAG) systems requires more specialized approaches than traditional large language model (LLM) benchmarks. RAGAS (Retrieval-Augmented Generation Assessment Scores) provides a structured methodology for assessing retrieval precision, context relevance, and the faithfulness of generated responses (RAGAS Documentation). Studies such as those by Shuster et al. 2021 have demonstrated that retrieval quality significantly impacts user satisfaction and perceived reliability of conversational AI, particularly in academic settings. Thus, specialized evaluation frameworks are crucial for ensuring the effectiveness of RAG systems [50]. Building upon the need for specialized evaluation, metrics specifically designed for RAG models play a pivotal role. The RAGAS evaluation framework is widely utilized, emphasizing primary metrics such as Context Recall, Faithfulness, and Response Relevance to measure how well the retrieved documents support the generated response [44]. Context Precision measures the proportion of relevant chunks in the retrieved contexts, while Context Recall ensures that essential information is not omitted. Faithfulness evaluates the factual consistency between generated responses and the retrieved documents, and Response Relevance assesses whether the response addresses the user's query [7] [17].

However, though automated measures are reliable, they frequently fail to assess qualitative aspects like consistency, fluency, and general user happiness.

Sivasothy et al. 2024 noted that human assessment is still necessary to improve these systems and take into account factors that automated approaches can ignore [51].

2.2 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.

Large Language Models (LLMs) with integrated RAG techniques have greatly improved the knowledge-intensive NLP tasks, overcoming LLMs' challenges. Studies [54] and [55] underline 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. Furthermore, the study [30] highlighted the use of vector databases for continuous information adaptation integrated with RAG, greatly enhancing retrieval efficiency and relevancy of LLM outputs, which is essential for literature search and thesis retrieval in university libraries.

The application of RAG in various domains is addressed in numerous studies. For instance, the study by Arzideh et al. 2024 incorporates clinical language embeddings within RAG to improve healthcare information retrieval, while the study by Grigoryan and Madoyan 2024, "Building a Retrieval-Augmented Generation (RAG) System for Academic Papers," presents a system that enhances academic retrieval using vector search. Additionally, Aquino et al. 2024 employs RAG for effectively extracting and analyzing Brazilian legal documents, and Ryu et al. 2023 validates RAG's effectiveness in legal question-answering tasks. Moreover, Google Gemini, when integrated with a RAG mechanism and supported by vector search, can achieve enhanced semantic understanding, retrieval precision, and responses that are accurate, explainable, and grounded in domain-specific data.

The findings from these various studies demonstrate RAG's flexibility, highlighting its potential to transform how university libraries handle searches and improve access to academic papers.

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 [46] and [8], ensure the authenticity and consistency of the generated outputs of the model. Despite the effectiveness of automated metrics, human evaluation remains important in assessing coherence and user satisfaction, as mentioned in this study [7].

In summary, Retrieval-Augmented Generation (RAG) integrated in Large Language Models (LLMs) presents a groundbreaking method for improving literature searches and thesis retrieval in university libraries, especially at CSPC library. By examining the limitations and obstacles faced by traditional LLMs, the integration of RAG reveals its promise to transform research accessibility at the CSPC library.

2.3 Gap Bridge of the Study

Existing studies have extensively explored the capabilities of Retrieval-Augmented Generation (RAG) systems in various domains, including healthcare, legal research, and academic literature retrieval. However, there is a notable gap in the literature regarding the specific application of RAG systems within academic libraries, particularly in enhancing literature search and thesis retrieval processes. While previous research has demonstrated the effectiveness of RAG in improving information retrieval, there is limited implementation in the context of university libraries, where unique challenges and requirements exist.

This study aims to bridge this gap by developing a RAG-based chatbot system specifically designed for the CSPC library. By focusing on the unique challenges and requirements of academic libraries, this research seeks to contribute valuable insights into the effective implementation of RAG systems in enhancing information retrieval.

3 Methodology

This chapter discusses the specific steps and logical procedures that was employed to develop and evaluate the Retrieval-Augmented Generation (RAG)-based Large Language Model (LLM) chatbot system. This includes the research design, theoretical and mathematical framework, software and hardware tools, instruments, procedures, evaluation metrics, and a conceptual framework.

3.1 Research Design

Constructive research design focuses on designing and building technological artifacts to address real-world problems and evaluating their practical utility Lukka 2003. 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 [35].

This study adopted a construcive research design where the primary artifact was a Retrieval-Augmented Generation (RAG)-based chatbot integrated with a Large Language Model (LLM). The system was designed to revolutionize how the academe community interacts when finding thesis literature in CSPC library. It addresses the challenges faced in searching and retrieving thesis literature by replacing the current yet traditional database and keyword based search with a vector database and a 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.

3.2 Theorems, Algorithms, and Mathematical Models

This study implemented advanced machine learning techniques, natural language processing (NLP) models, and the Retrieval-Augmented Generation (RAG) pipeline, integrated with a Large Language Model (LLM) and a vector database. These components collaboratively enabled efficient information retrieval and generation in the context of literature and thesis search within the CSPC Library.

3.3 Retrieval-Augmented Generation (RAG) Pipeline

The Retrieval-Augmented Generation (RAG) pipeline is a hybrid architecture that combines information retrieval with natural language generation. It allows LLMs to access external documents during inference, thereby improving both accuracy and contextual relevance.

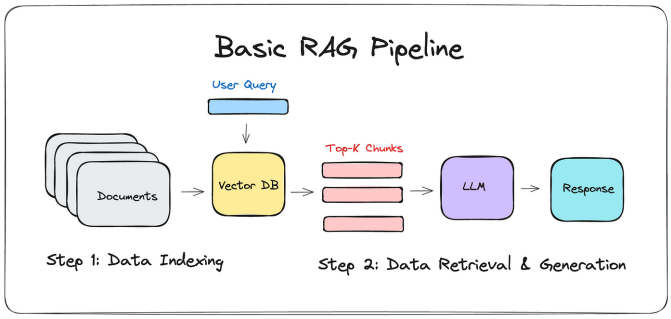

Figure 1: Basic RAG Pipeline by Dr. Julija

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

3.3.1 A. Data Indexing. The data indexing process begins with document preparation, where thesis documents are collected and preprocessed into smaller, semantically coherent chunks using a token-based method that respects academic structure (e.g., Abstract; Chapters 1–5) to preserve context. Each chunk is then converted into a dense vector using the open-source 'sentence-transformers/all-MiniLM-L6-v2' model from Hugging Face, which was chosen for its lightweight architecture and strong semantic representation capabilities. Finally, these vectors and their associated metadata are stored in FAISS, enabling efficient similarity search and scalable retrieval over the entire thesis corpus.

3.3.2 B. Retrieval and Generation. The retrieval and generation process begins with query processing, where a user's query is

embedded using the same model employed for indexing. A FAISS-backed retriever then performs a semantic search to return the top-K most relevant chunks (default $K = 6$), balancing precision and recall. Finally, during contextual generation, these retrieved chunks are provided to the Gemini 2.5-flash language model as grounded context, enabling the generation of relevant and factual responses aligned with the source documents.

3.4 Large Language Model

Large Language Models (LLMs) are cutting-edge AI systems trained on massive datasets to process and generate text, excelling in tasks like summarization, question answering, and retrieval Naveed et al. 2024. This study utilized Gemini 2.5-flash.

3.5 Gemini 2.5-flash

The large language model integrated in this project is Gemini 2.5 Flash, part of the Gemini 2.X model family introduced by Comanici et al. 2025. It delivers advanced reasoning, multimodal support, extended context windows, and agentic workflows. Its architecture optimizes factual accuracy and relevance while minimizing latency. Incorporated into a RAG framework, it enhances domain-specific retrieval for CSPC Library users, leveraging its real-time, cost-effective capabilities.

3.6 Materials and Statistical Tools / Evaluation Methods

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

3.7 Research Materials

This section contains the dataset, hardware, and software requirements for the development of the RAG chatbot system.

3.7.1 Dataset. The study utilized a dataset consisting of all available undergraduate thesis PDFs (initially 290+ pdfs) from multiple CSPC departments, sourced from the CSPC library. Additionally, the system was designed to ingest newly published theses by allowing admin to upload new PDF files.

3.7.2 Hardware. To support the development of RAG chatbot system, the researchers will use hardware components that meet or exceed the specifications outlined in 1. These components were selected to ensure efficient ingestion of a large corpus of PDFs, as well as to handle computationally intensive tasks like embedding generation.

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)

A modern multi-core CPU enables efficient data processing and model inference, ensuring smooth query execution. At least 16 GB of RAM is recommended to manage large-scale embeddings and real-time retrieval operations effectively.

A 1 TB SSD is preferred due to its high read/write speeds, which significantly enhance data indexing and retrieval. Given the resource-intensive nature of embedding computations and AI-driven text generation, a high-performance GPU, such as an NVIDIA RTX 3090 or better, is crucial for accelerating deep learning inference and vector operations.

Table 2: Software Requirements

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

3.7.3 Software. Python 3.10 or later serves as the core programming language due to its comprehensive support for machine learning and natural language processing. FAISS is used as the vector database to facilitate fast and accurate semantic search. The system leverages Gemini 2.5-flash as its LLM via the Google Generative AI API, and sentence-transformers/all-MiniLM-L6-v2 to transform preprocessed text chunks into semantically rich vector representations. The Streamlit framework is used to build an interactive user interface.

Document parsing and extraction are managed through the PyMuPDF library, ensuring accurate and efficient retrieval of textual data from PDF files. NumPy supports numerical operations, while LangChain manages the orchestration of LLMs during query interpretation and response generation.

3.8 Instrument

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

RAGAS (Retrieval-Augmented Generation Assessment Suite). toolkit was utilized to automatically evaluate the quality of system outputs using metrics such as context precision, faithfulness, and answer relevance [49]. Furthermore, a context recall metric was included, as recommended for evaluating retrieved chunks. These instruments ensured a rigorous and balanced evaluation of the proposed system from both system-level and user perspectives [32].

Survey. Instruments serve as data collection tools across different areas and provide an effective way to gather information. They are useful when seeking insights into the attributes, preferences, opinions, or beliefs of a specific group. To meet the study objectives, the researchers will conduct a survey among employed librarians and CSPC students to evaluate the proposed RAG chatbot using a user-centered method that measures users' level of agreement on the chatbot's quality and performance. The researchers will develop

questionnaires to assess users' satisfaction with answers, likelihood to use the chatbot again, ease of reading and understanding the output, and confidence in the information retrieved by the system. The respondents of the study are all from the CSPC including 3 employees of Library, 1 faculty, and the 7 students will serve as a representative of the whole population.

3.8.1 Statistical Test. The system's technical 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 [5, 21, 33].

Additionally, to asssess 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, utilizing a 5-point Likert scale to ensure consistent measurement.

Table 3: Likert Scale for User Level of Agreement

Scale	Range	Level of Agreement	Description
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.

3 shows that RAG chatbot system will use 5-point Likert Scale to determine users Level of Agreement based on the user's experience to the system's response quality and performance. The first column showed the scale that the system level of agreement 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 "Strongly Agree" which means that the user of the RAG Chatbot completely agrees with the described criteria or finds its quality and performance excellent, scale 4 with a range 3.40-4.19 described as "Agree", the user of the RAG Chatbot agrees with the described criteria 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 "Disagree", the user of the RAG Chatbot disagrees with the criteria description but not as intensely as "Strongly Disagree", and Scale 1 is described as "Strongly Disagree", the user of the RAG chatbot completely disagrees with the criteria description or finds the system as low quality and low performance.

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.

4 shows that RAG chatbot system will use 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 will be 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.

5 shows that RAG chatbot system will use 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

Scale	Range	Level of Understanding	Description
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.

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 understand 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.

7 shows that RAG chatbot system will use 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 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 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 "Uncofident", 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.

To analyze user questionnaire responses, the researchers used the Weighted Mean to summarize Likert-scale data. It captures perceptions of user satisfaction with answers, likelihood of using again, ease of understanding outputs, and confidence in accuracy. This method enables consistent comparison across items and respondents, supporting an evidence-based assessment of the chatbot's overall usability and performance.

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

Where:

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

3.9 Procedures

The procedures encompassed the collection and preprocessing of academic data, vector-based indexing, retrieval using semantic search, LLM-based response generation, and multi-metric evaluation using RAGAS, and user-centered evaluation.

Each stage was designed to ensure the integrity, replicability, and effectiveness of the system in addressing the research objectives. By detailing the technical and methodological steps, this section served as a transparent and structured guide for future researchers seeking to replicate or build upon this study.

(1) **Data Preprocessing** - The collected PDF thesis documents underwent a preprocessing phase to extract and clean the textual content.
   (a) Text Extraction: PyMuPDF was used to convert PDF files into structured plain text.
   (b) Cleaning: Non-informative characters and formatting were removed.
   (c) Text Chunking: Text was segmented into manageable chunks to enhance semantic search accuracy.

(2) **Indexing and Vector Embedding** - The preprocessed text chunks were transformed into vector representations and indexed for efficient retrieval.
   (a) Vector Embedding: Each text chunk will be embedded using gemini-embedding-001.
   (b) Cleaning: FAISS will store the vectorized content along with metadata such as document titles, authors, and section headers.

(3) **Query Handling and Semantic Retrieval** - User queries were processed to retrieve relevant document chunks from the vector database.
   (a) Query Encoding: The user's natural language query is encoded using the same embedding model applied during indexing to maintain compatibility in the latent space.
   (b) Similarity Search: The encoded query is matched against stored vectors to retrieve the top-K relevant chunks (default $K = 6$). For exploratory or synthesis-oriented queries, K may be adaptively increased to improve coverage.

(4) **Response Generation** - The Gemini 2.5-flash language model will process the augmented input to generate a response that is factually aligned with the source documents.

(5) **Output Presentation** - The system will display the generated response via a user interface that includes metadata such as the source thesis title and section, encouraging transparency and academic integrity.

(6) **Performance Evaluation**
   (a) Automated Evaluation: Metrics from the RAGAS framework, Context Precision, Context Recall, Answer Relevance, and Faithfulness, will be calculated.

(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.

3.10 Evaluation Metrics

The researchers used a framework called **RAGAS** that comprised specific metrics to assess Retrieval-Augmented Generation (RAG)-based architectures, thereby ensuring precise measurements of both retrieval quality and generation fidelity [40]. This framework evaluated the model's performance using the following metrics: *Context Precision*, *Context Recall*, *Response Relevance*, and *Faithfulness*. Each metric was essential in addressing the system's retrieval and generation performance.

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

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}} \tag{2}$$

where Precision@k is 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. Here, K indicates the cutoff for the top results evaluated. The denominator normalizes the metric by accounting for the total number of relevant items within the top K retrieved results. This weighted approach ensures that relevant items retrieved earlier in the ranking contribute more significantly to the final score, making the metric especially meaningful for library retrieval tasks.

The precision at each position k, denoted as Precision@k, is computed as follows:

$$\text{Precision@k} = \frac{\text{true positives@k}}{\text{true positives@k + false positives@k}} \tag{3}$$

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.

Context Recall. Context Recall was used to evaluate the comprehensiveness of the retrieval system in capturing all relevant information necessary to answer a query. It measured the proportion of relevant chunks successfully retrieved by the RAG chatbot within the CSPC Library, ensuring minimal omission of important academic content.

$$\text{Context Recall} = \frac{\text{Number of relevant claims supported by retrieved chunks}}{\text{Total number of relevant claims in the reference answer}} \tag{4}$$

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 document chunks,
- *Total number of relevant claims in the reference answer* represents all the factual claims present in the ground truth answer that ideally should be covered by the retrieval process.

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

Response Relevance. Response Relevance was a critical metric used to evaluate how well the RAG chatbot's generated answer addressed the specific query posed by users in the CSPC Library. This metric ensured that the chatbot provided focused, comprehensive, and directly applicable responses to academic inquiries, minimizing irrelevant or incomplete information that could hinder research efficiency.

$$\text{Response Relevance} = \frac{1}{N} \sum_{i=1}^{N} \cos(E_{g_i}, E_o) \tag{5}$$

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

This metric works on the idea that if the chatbot's response sufficiently answers the original query, then questions generated from that response will semantically align with the original question, this involves generating multiple artificial questions, embedding both the response-generated questions and the original query into vector representations, and calculating the mean cosine similarity to measure alignment, which ensures that the retrieved academic information closely matches the research needs of CSPC Library users.

Faithfulness. Faithfulness is a critical metric for evaluating the factual consistency of the RAG chatbot's generated responses with respect to the retrieved context from the CSPC Library. This metric ensures that all claims made in the chatbot's answer are directly supported by the information present in the retrieved documents, thereby minimizing hallucinations and maintaining academic integrity.

$$\text{Faithfulness} = \frac{\text{Number of claims in the response supported by retrieved context}}{\text{Total number of claims in the response}} \tag{6}$$

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, regardless of whether they are supported by the context.

A faithfulness score of 1.0 indicates that all claims in the response are grounded in the retrieved context, while lower scores reveal the presence of unsupported or hallucinated information. In the context of academic literature search and thesis retrieval, maintaining high faithfulness is essential to ensure that the chatbot's answers are trustworthy and factually accurate, directly reflecting the content of the CSPC Library's resources.

3.11 Conceptual Framework

The conceptual framework served as the foundational blueprint for the RAG-based chatbot system. It emphasized the end-to-end interaction of modules required to support intelligent, accurate, and efficient academic document retrieval. As illustrated in 2, the system followed a cyclical process beginning with data collection and ending with system evaluation and refinement. The arrows were used solely to visually indicate the step-by-step flow of each component within the chatbot framework; they did not signify any technical operation or special relationship beyond showing the direction of the process.

This visualization helps guide readers through the sequence of the system stages, ensuring clarity at the outset.

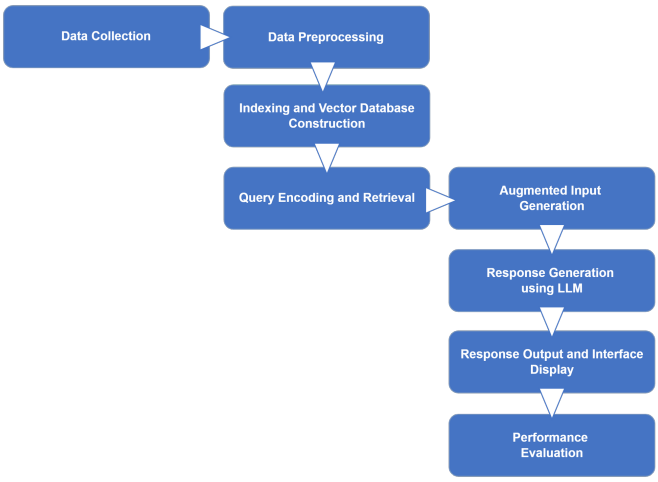

Figure 2: Conceptual Framework of the RAG-Based Chatbot System

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

Data Pre-processing. Tools like PyMuPDF were used to extract plain text from the collected PDFs. The extracted content underwent cleaning and normalization to remove non-informative characters, followed by segmentation into semantically meaningful text chunks.

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

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

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

Response Generation, The response generation stage represented the culmination of the RAG pipeline, where gemini-1.5-flash transformed augmented academic context into coherent, factually grounded answers that addressed user research inquiries.

Response Output and Interface Display. In this section, streamlit was used to create an intuitive web interface that presented the chatbot's responses alongside relevant metadata, such as source thesis titles and sections.

Performance Evaluation. Lastly, the system's effectiveness was evaluated using both automated metrics and human-centered assessments. Automated evaluation employed the RAGAS framework and for user-centered evaluation, a structured questionnaire was administered to gather feedback on usability, accuracy, and overall satisfaction. This dual-layered evaluation ensured that the RAG-based chatbot system not only met technical performance benchmarks but also aligned with user expectations and academic research needs.

4 Results and Discussion

This chapter discusses the results and evaluation of the Retrieval-Augmented Generation (RAG) chatbot developed for efficient literature search and thesis retrieval at the Camarines Sur Polytechnic Colleges (CSPC) Library.

4.1 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), enabling section-aware retrieval.

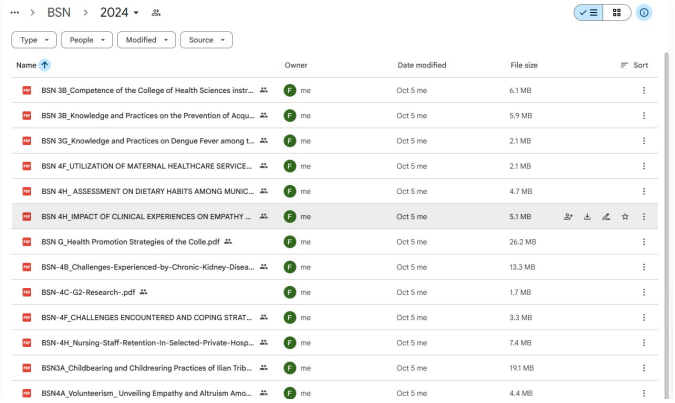

Figure 3: CSPC Thesis PDF Sample

Upon agreement on project scope and data handling, library personnel granted the researchers to gain access to the digital copies of undergraduate thesis papers. This composes of theses from different departments.

4.2 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.

Figure 4 shows the chunk analysis and statistics respectively. Chunk statistics indicated a total chunks count of 38,127 and total token count of 11,849,783. With an average of 311 tokens per chunk. The chunking strategy was effective in breaking down lengthy thesis documents into manageable, semantically coherent pieces suitable for embedding and retrieval.

4.3 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-transformers/all-MiniLM-L6-v2 (HuggingFace), chosen for its efficiency and strong semantic performance; when cloud embeddings were available, Gemini could be used as an alternative for multilingual scenarios. FAISS stored vectors alongside source/page metadata to preserve traceability. This enabled natural language queries to retrieve semantically relevant thesis segments beyond exact keyword matching.

Source	Page	Char_count	Word_count	Token_count	Is_empty
data/revmedia/BSHM 4B_Pili Pulp	0	673	86	187	False
data/revmedia/BSHM 4B_Pili Pulp	1	1043	138	264	False
data/revmedia/BSHM 4B_Pili Pulp	2	1162	157	294	False
data/revmedia/BSHM 4B_Pili Pulp	3	508	71	141	False
data/revmedia/BSHM 4B_Pili Pulp	4	512	72	145	False
data/revmedia/BSHM 4B_Pili Pulp	5	604	86	166	False
data/revmedia/BSHM 4B_Pili Pulp	6	1383	189	307	False
data/revmedia/BSHM 4B_Pili Pulp	7	1003	140	211	False
data/revmedia/BSHM 4B_Pili Pulp	8	1994	301	483	False
data/revmedia/BSHM 4B_Pili Pulp	9	685	96	193	False
data/revmedia/BSHM 4B_Pili Pulp	10	745	99	196	False
data/revmedia/BSHM 4B_Pili Pulp	11	572	69	152	False
data/revmedia/BSHM 4B_Pili Pulp	12	1062	182	294	False
data/revmedia/BSHM 4B_Pili Pulp	13	1035	177	279	False
data/revmedia/BSHM 4B_Pili Pulp	14	503	76	130	False
data/revmedia/BSHM 4B_Pili Pulp	15	568	91	165	False
data/revmedia/BSHM 4B_Pili Pulp	16	868	154	234	False

(a) Chunk Analysis

Metric	Value
Total Chunks	38127
Total Tokens	11,849,783
Avg Characters/Chunk	1323
Avg Words/Chunk	180
Avg Tokens/Chunk	311
Min Tokens	1
Max Tokens	1799
Median Tokens	335

(b) Chunk Statistics

Figure 4: Chunk Analysis & Statistics

4.4 Query Encoding and Retrieval

Queries were embedded using the same model as indexing to ensure consistency. The FAISS-backed retriever returned the top-K relevant chunks (default $K = 6$), balancing precision and recall. Diversity-enhancing strategies (e.g., MMR) were used for broader queries to avoid redundant chunks.

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 $K = 6$ produced a good balance of focused context and cross-thesis coverage.

4.5 Augmented Input and Generation

Retrieved chunks were concatenated with the user query into a structured context with lightweight citation markers. This supported grounded, traceable answers and reduced hallucination risk.

Prompt templates guided the model to answer strictly from provided context, with safeguards (token monitoring, truncation) to maintain input quality.

4.6 Response Generation with Gemini 2.5-flash

The Gemini 2.5-flash model generated answers grounded in retrieved context. The system was configured with temperature=0 to ensure deterministic outputs suitable for academic use.

Generated content was parsed into clean text for display. While RAG significantly reduced hallucinations, occasional inaccuracies were observed when context was insufficient; users were advised to validate critical findings.

4.7 Interface and Usage Observations

The Streamlit interface supported conversational exploration with session-based history and safety filters for disallowed queries. Cached chains ensured responsive interactions.

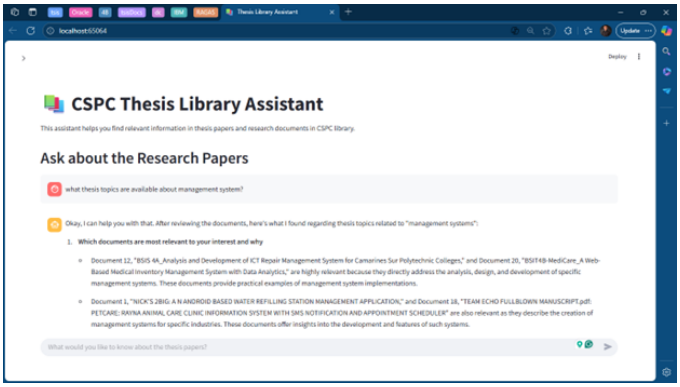

Figure 5: User Interface

Generated responses appeared as Markdown with citations and structured text. When queries violated safety parameters, clear warnings were shown. Deterministic settings improved consistency and user trust.

4.8 Model Evaluation

This section evaluated the CSPC Library RAG chatbot using four core metrics: Answer Relevancy, Context Precision, Context Recall, and Faithfulness. Together, they capture accuracy, coverage, and grounding of responses. The evaluation follows established academic practices, enabling concise, reliable measurement of retrieval quality and generation within the literature search workflow of the system effectively.

4.9 Result

This section presents the findings through tables, figures, and subsequent discussion. Prior to evaluation, a systematic data processing pipeline was applied: 290+ undergraduate thesis PDFs from the CSPC Library were processed into segmented meaningful text chunks, and embedded using Hugging Face's Embeddings. These chunks were indexed in FAISS for efficient semantic retrieval, enabling the RAG chatbot to generate contextually relevant and factually grounded responses for user queries. This process ensured that the evaluation was conducted on high-quality, well-structured academic data. In addition to the system-level RAGAS metrics, a complementary user-centered evaluation was performed using a 5-point Likert scale questionnaire, responses were summarized via weighted mean and interpreted using predefined agreement ranges (see 3) to align technical performance with perceived usability and satisfaction.

Table 8: RAG System Evaluation Metrics using RAGAS Framework

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

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

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

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

Visualization of RAG System Evaluation Metrics

The figures below illustrate the evaluation metrics of the RAG system using various visualization techniques, including bar charts, box plots, heatmaps, and radar charts.

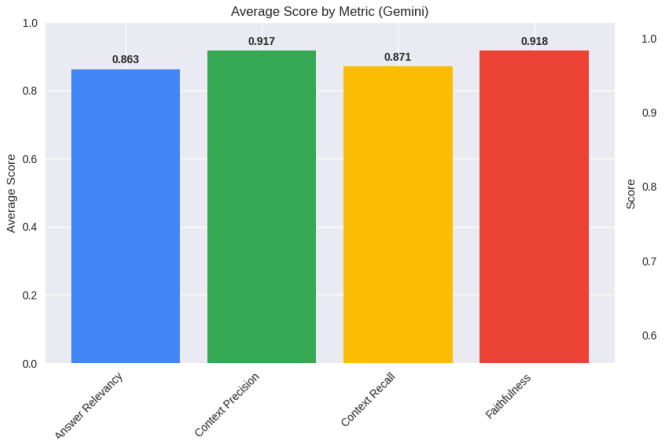

Figure 6: Bar Chart of RAG System Evaluation Result

The bar graph shows that the overall evaluation of the RAG system demonstrates a strong performance in all four metrics. The highest scores are observed in Faithfulness (0.918) and Context Precision (0.917), indicating that the system effectively grounds its responses in accurate and relevant information retrieved from the source documents. These results suggest that the system minimizes hallucinations, maintains real information during response, and focuses on the most pertinent contextual segments during retrieval. These scores prove that the RAG model is well-optimized for generating trustworthy and accurate responses.

The faithfulness result as the score means that the system consistently produces outputs that accurately reflect the underlying source material, which is helpful for users seeking reliable information. The high context precision score indicates that the retrieved passages are highly relevant to the user's information needs, minimizing the inclusion of unnecessary or loosely related content. This is particularly important in academic contexts where precision is critical. The context recall score, while slightly lower, still demonstrates that the system captures a substantial portion of relevant information, though there may be room for improvement in ensuring that all pertinent details are included. Finally, the answer relevancy score indicates that the responses generated by the system generally align well with user queries, although there may be occasional instances where the answers could be more comprehensive or directly address the user's intent.

These results, visualized using a bar chart, further confirm the effectiveness of the designed RAG pipeline. By using metrics such as faithfulness, context precision, context recall, and answer relevancy, the evaluation demonstrates robust grounding, accurate retrieval, and answers well the different types of queries. Overall, the findings indicate that the system reliably meets information needs and provides actionable assistance to users who primarily seek accurate, relevant, and well-cited academic content from the CSPC Library's thesis collection, thereby supporting accurate literature search and informed research decision-making for students and researchers.

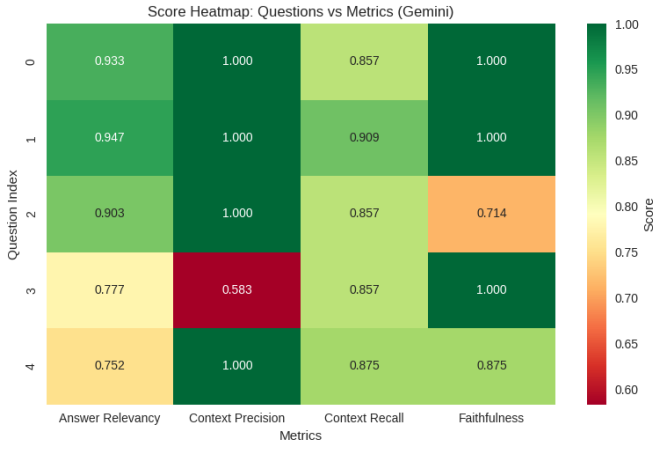

Figure 7: Heatmap of RAG System Evaluation Result

The heatmap shows the question level performance of the RAG system across the four evaluation metrics. Most of the scores are ranging from 0.75 to 1.00, indicating a generally strong performance. The dark green cells represent high-quality outputs, while the mid-range yellow tones and the single red cell indicates low Context Precision for Question 3 which highlights the area where the system's performance could be improved. Overall, the system demonstrates strong answer alignment, with Questions 0 to 2 achieving high Answer Relevancy scores above 0.90. Meanwhile, Questions 3 and 4 show slightly reduced relevancy, suggesting occasional omissions. These patterns indicate that the system generally maintains high standards but may need targeted refinements for more complex queries.

Among the RAG metrics implemented, Context precision is excellent for four of the five questions with each scoring 1.00, while Question 3's value signals low context selection, despite that, Context recall remains consistently high, indicating stable retrieval depth across queries. Faithfulness is similarly strong, with only Question 2 dipping slightly. Altogether, the heatmap highlights a reliable RAG system with minor, clearly identifiable areas for improvement.

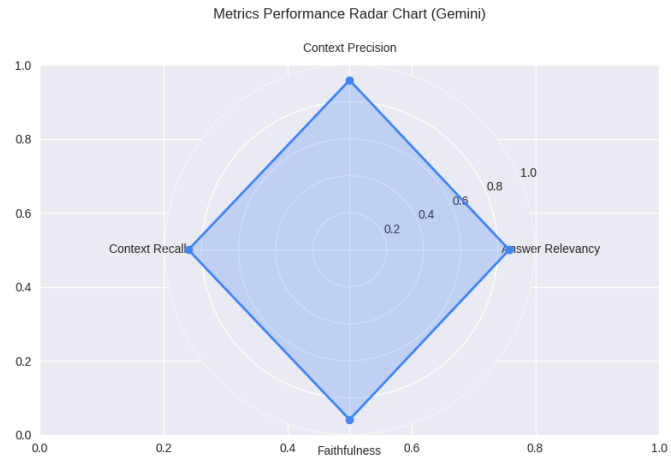

Figure 8: Radar Chart of RAG System Evaluation Result

The radar chart shows that the RAG system demonstrates a consistently high and well-balanced performance across the four evaluation metrics: Faithfulness, Context Precision, Context Recall, and Answer Relevancy. The nearly symmetrical shape of the plot indicates that no metric falls below an acceptable range, with Faithfulness and Context Precision forming the strongest extensions. This suggests that the system reliably grounds its answers in retrieved evidence and selects context that is highly relevant to the user's query, effectively minimizing hallucinations and maintaining strong alignment with source documents.

However, the Context Recall and Answer Relevancy metrics, while still the RAG system show good performance in these areas, the Radar Chart indicates that there is room for improvement to further enhance the system's ability to retrieve all relevant information and generate answers that fully meet user expectations.

Focusing on these metrics could lead to even more comprehensive and satisfactory responses in future iterations of the system.

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 Agreement on Chatbot Response Quality and Performance

9 shows the results of the user-centered evaluation of the CSPC Library RAG chatbot using 5-point likert scale survey questions that allows the respondents to evaluate and choose the level of agreement with the chatbot's response quality and performance.

Table 9: User Agreement: Chatbot Response Quality and Performance

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

The result of the evaluation of the RAG chatbot using user-centered evaluation method indicate a generally positive reception from users across various criteria. Here's the breakdown of the findings. In terms of the chatbot's question and answering performance, users strongly agreed (weighted mean: 4.3) that the system performed well in answering user questions, indicating that the chatbot meets user expectation in getting right answers. Users also strongly agreed (weighted mean: 4.5) that the chatbot provide answers relevant to the questions provided by the users, indicating that the system effectively interprets user intent and provide relevant answers based on the questions. Furthermore, users strongly agreed (weighted mean: 4.5) that the chatbot gives clear and easy-to-understand answers. This means the chatbot not only gives correct responses but also explains them in a way that users can easily follow. Moreover, the chatbot helped users find the answers they were looking for. With a (weighted mean of 4.3), users strongly agreed that the chatbot's replies were useful and matched their questions well. This shows that the system supports users in getting the help they need. In the same way, the chatbot provided enough information to help users, with a weighted mean of 4.2. This shows that users strongly agreed and felt the chatbot gave complete and useful answers during their interaction. Lastly, the chatbot was quick to reply, with users agreeing (weighted mean: 4.1) that it responded without delay. This means the system was able to give answers fast, helping users get the information they needed right away. Overall, the respondents showed agreement across the measured areas, with an average weighted mean of 4.3 (Strongly Agree). This indicates that users found the chatbot's answers to be correct, relevant, clear, and mostly complete, and that the chatbot responded quickly enough to be useful These findings suggest the chatbot works well for its main task of helping users find information and understand answers. Minor improvements could focus on making responses more complete and slightly faster to raise overall satisfaction even more. Furthermore, according to Følstad et al. 2021, user-centered evaluation has been key within several disciplines at the roots of current chatbot research, particularly in understanding users' needs, motivations, and experiences with chatbot interactions. Thus, it is advisable to utilize this method to assess system effectiveness and user satisfaction before deployment to ensure the RAG chatbot meets actual user expectations and provides satisfactory support for thesis retrieval tasks in the CSPC Library context.

User Feedback on RAG chatbot's Effectiveness and Usability

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

Table 10: User Feedback on RAG chatbot's Effectiveness and Usability

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

The results for satisfaction with answers, likelihood to use again, ease of reading and understanding, and confidence in information accuracy show generally positive user feedback. And, according to Kaushal and Yadav 2022 and Okonkwo and Ade-Ibijola 2021, these aspects of chatbots that deliver clear, useful, and readable responses greatly improve user satisfaction. In addition, Choudhury and Shamszare 2023 and Zhang et al. 2024 found that trust and factual accuracy are essential for encouraging continued use and building user confidence in AI chatbots. After considering these established determinants, the detailed breakdown is as follows. In terms of user satisfaction with answers, users were satisfied (weighted mean: 4.1), indicating that the chatbot's replies met users' needs and were generally acceptable. Regarding likelihood of reuse,

users were very likely to use the chatbot again (4.3), suggesting strong perceived utility. Users also found the responses very easy to read and understand (4.5), demonstrating clear and user-friendly output. Confidence in the chatbot's information was moderately strong (3.8), implying general trust with some expectation for accuracy improvements. Overall, the respondents provided positive feedback across all measures with the overall weighted mean of (4.2), showing that the users strongly agree that the chatbot delivers useful, relevant, clear, and mostly complete answers, supports continued usage intention, and yields satisfactory user experience, with factual confidence identified as a targeted area for further enhancement.

5 Summary of Findings, Conclusions, and Recommendations

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

5.1 Summary

Finding relevant thesis literature in a university library like in CSPC can be difficult for many students and researchers. Most people have a hard time finding the exact thesis they need because the current library website only allows searches by the exact title. If a user does not know the precise title, it becomes a struggle to locate the right documents. Making things even harder, library rules do not allow theses to be taken out of the building, which means users must visit the library in person to access important academic materials. Because of these challenges, this study explored creating a 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 Retrieval-Augmented Generation (RAG) along with a state of the art Large Language Model (LLM). The process involved preprocessing and converting 290+ undergraduate thesis papers into digital embeddings and be store on FAISS vector database, which allows the chatbot to understand and search for relevant information based on a user's question in natural language. The chatbot retrieves and shows the most fitting parts of the theses and uses the Gemini 2.5-flash model to generate accurate and appropriate responses. All the system steps from preparing the thesis files to designing a simple user interface work together to make searching faster and more effective. The system was tested using different automated metrics from RAGAS evaluation framework including context precision, context recall, answer Relevance and faithfulness. Additionally, a user questionnaire with a 5-point likert scale to assess a human-centered performance.

The results were promising, with the RAG-based chatbot achieving high performance in both precision and answer relevancy. Specifically, the system scored 0.818 for Context Precision, demonstrating strong capability to return highly relevant content, and context recall reached 0.721, showing that the chatbot successfully gathers most of the necessary supporting content, though some supporting details were still missed. Answer relevancy reached 0.737, indicating that responses were generally useful and addressed user queries appropriately. While faithfulness at 0.585 highlighted that while most answers were grounded in the source documents, there is room for improvement in ensuring all generated content is directly traceable to original texts.

The deployment of the chatbot was observed to make thesis search more intuitive, significantly lowering entry barriers for users and reducing time spent locating needed materials. Moreover, it became evident to the researchers that several factors influenced system performance, such as thesis data quality, and user prompts that significantly impairs the chatbot's accuracy.

5.2 Findings

The following are the key findings from the study:

(1) By integrating a document ingestion and retrieval module, all thesis documents in PDF format are standardized and divided using thesis-aware boundaries, enriched with metadata, embedded, and indexed in FAISS. This makes them discoverable through semantic search rather than exact keyword matching, resulting in much better retrieval of abstracts, authors, chapters, and themes compared to traditional databases. The pipeline's token-based chunking and content-type tagging (such as abstract, methodology, or results) really improve context alignment and ranking, so queries like "give me the complete abstract" or "find theses related to nursing" return relevant sections directly, whereas traditional catalogs usually require exact titles or strict keyword fields.
(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 generative LLM, user queries are matched to semantically relevant thesis chunks, which consistently yields more precise answers to intent-driven questions (such as complete abstracts, author-focused lookups, or chapter-specific content) than the current yet traditional keyword-based search that depends on exact titles and rigid field matches. The RAG pipeline grounds answers in retrieved chunks, so the chatbot can compose context-aware responses tied to real thesis passages, while the thesis-aware chunking and metadata (abstract, methodology, results, and chapter tags) improve ranking and reduce irrelevant matches.
(3) Using the RAGAS as the RAG evaluation framework the system sustained high results in Context Precision (0.9167), Context Recall (0.8711), Answer Relevancy (0.8625), and Faithfulness (0.9179). Together these values indicate a robust retrieval-and-generation pipeline: high precision and recall show the FAISS-backed retrieval reliably returns the necessary thesis passages, while the strong answer relevancy demonstrates the LLM composes useful responses from those sources. The notably high faithfulness score suggests that generated outputs are, in most cases, directly grounded in retrieved documents, which materially reduces the likelihood of unsupported or hallucinated assertions. The evaluation still highlights the usual dependencies and limitations, chiefly the quality and structure of ingested thesis PDFs, OCR/formatting errors, and user prompt variability which

remain important targets for continued optimization even as overall performance is strong.

5.3 Conclusions

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

(1) It is evident that the integration of a thesis-aware document ingestion pipeline is a critical first step in modernizing thesis discovery. By standardizing documents, applying content-aware chunking, and enriching the data with metadata before indexing, the system overcomes the limitations of traditional keyword-based catalogs. This approach makes thesis content more accessible and discoverable through semantic search, allowing users to find relevant information based on intent rather than exact phrasing.
(2) By implementing a semantic search and thesis retrieval system with RAG orchestrated in LangChain, FAISS for approximate nearest-neighbor search, using all-MiniLM-L6-v2 for vector embeddings and Google Gemini 2.5-flash as the generative LLM, user queries are matched to semantically relevant thesis chunks, which consistently yields more precise answers to intent-driven questions than the current yet traditional keyword-based search that depends on exact titles and rigid field matches. By combining semantic search with the generative capabilities of Google Gemini 2.5-flash, the RAG chatbot offers a robust solution for efficient and accurate thesis retrieval in the CSPC Library.
(3) The evaluation confirms the RAG-based chatbot is an effective tool, successfully retrieving relevant context and providing useful answers. However, it also reveals areas for improvement, particularly in ensuring complete information retrieval and strict faithfulness to the source texts. The system's performance is fundamentally linked to the quality of the ingested documents, highlighting a need for ongoing optimization to enhance its reliability.

5.4 Recommendations

(1) Future work should focus on improving the faithfulness and recall of generated responses. This may involve refining the chunking and retrieval process, integrating more advanced embedding models, or incorporating additional post-processing checks to ensure factual consistency.
(2) Increasing the diversity and volume of ingested thesis documents, and exploring the inclusion of other academic materials (e.g., journal articles, conference papers), can further improve retrieval coverage and system robustness.
(3) To complement automated metrics, future evaluations should include human-in-the-loop assessments, such as expert reviews or user feedback surveys, to better capture subjective aspects of response quality and user satisfaction.
(4) Continued development of the chatbot interface, including user-friendly features like advanced filtering, citation export, and personalized recommendations, will enhance usability and adoption within the academic community.
(5) For broader impact, consider deploying the system on scalable infrastructure and integrating it with other institutional platforms, ensuring accessibility for all CSPC stakeholders.

Acknowledgments

Acknowledgements go here. Delete enclosing begin/end markers if there are no acknowledgements.

References

[1] Mohamed Aboelmaged, Shaker Bani-Melhem, Mohd Ahmad Al-Hawari, and Ifzal Ahmad. 2024. Conversational AI Chatbots in library research: An integrative review and future research agenda. *Journal of Librarianship and Information Science* (2024), —4440.
[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
[3] 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. arXiv:2410.09871v2 https://arxiv.org/abs/2410.09871v2
[4] U. Allu, B. Ahmed, and V. Tripathi. 2024. Beyond Extraction: Contextualising Tabular Data for Efficient Summarisation by Language Models. https://doi.org/10.36227/techrxiv.170792474.42605726/v1
[5] Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. 2024. A Statistical Framework for Ranking LLM-Based Chatbots. https://doi.org/10.48550/arXiv.2412.18407 arXiv:2412.18407 [cs.CL]
[6] L. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel. 2024. Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review. https://doi.org/10.20944/preprints202407.0876.v1
[7] I. Aquino, M. Santos, C. Dorneles, and J. Carvalho. 2024. Extracting Information from Brazilian Legal Documents with Retrieval Augmented Generation. In *SBBD Estendido*. 280–287. https://doi.org/10.5753/sbbd_estendido.2024.244241
[8] K. Arzideh, H. Schäfer, A. Idrissi-Yaghi, B. Eryılmaz, M. Bahn, C. Schmidt, and R. Hosch. 2024. MIRACLE - Medical Information Retrieval Using Clinical Language Embeddings for Retrieval Augmented Generation at the Point of Care. *Research Square* (2024). https://doi.org/10.21203/rs.3.rs-5453999/v1
[9] Yahya Aydin. 2021. Comparing University Libraries in Different Cities in Turkey with regards to Digitalisation and the Impact of the COVID-19 Pandemic. *Information Society/Információs Társadalom (InfTars)* 4 (2021).
[10] 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).
[11] Aras Bozkurt. 2024. GenAI et al. Cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. , 10 pages.
[12] 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*, Vol. 38. 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728
[13] Mark Chen et al. 2021. Evaluating large language models trained on code. (2021). arXiv:2107.03374 [cs.LG]
[14] Avishek Choudhury and Hamid Shamszare. 2023. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. *Journal of Medical Internet Research* 25 (2023), e47184.
[15] 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*, Vol. 11. MDPI, 2417.
[16] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).
[17] 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 (2025), 1–9. https://doi.org/10.55041/ijsrem42403
[18] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing* 103, 12 (2021), 2915–2942.
[19] Gerald Gartlehner, Laura Kahwati, Roxanne Hilscher, Ivan Thomas, Susan Kugley, Kristina Crotty, and Rebecca Chew. 2023. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. (2023). https://doi.org/10.1101/2023.10.02.23296415 Preprint.

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

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

[22] 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.

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

[24] 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.

[25] 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.

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

[27] 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 (2024), tzae022.

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

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

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

[31] S. Li, Y. Zhang, Z. Fang, K. Meng, R. Tian, H. He, and S. 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 (2023), 6249–6260. https://doi.org/10.1021/acs.jcim.3c01442

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

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

[34] 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.

[35] Kari Lukka. 2003. *The Constructive Research Approach*. Publications of the Turku School of Economics and Business Administration. https://www.researchgate.net/publication/247817908_The_Constructive_Research_Approach Accessed: 2025-05-26.

[36] 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* (2024).

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

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

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

[40] I. OUBAH and S. 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 (2024). https://doi.org/10.47191/etj/v9i07.09

[41] 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: December 8, 2025.

[42] 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.

[43] 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 (2024), 31–39.

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

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

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

[47] Malik Sallam. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, Vol. 11. MDPI, 887.

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

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

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

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

[52] S. Song, C. Yang, X. Li, H. Shang, Z. Li, and Y. Chang. 2024. TravelRAG: A Tourist Attraction Retrieval Framework Based on Multi-layer Knowledge Graph. *ISPRS International Journal of Geo-Information* 13, 11 (2024), 414. https://doi.org/10.3390/ijgi13110414

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

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

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

[56] 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

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

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

[59] L. Zhang, X. Chen, and M. Li. 2023. Automated Document Ingestion for Academic Knowledge Repositories. *Journal of Digital Libraries* 24, 1 (2023), 12–26.

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

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