A Dynamic Query Framework for Research Accessibility using OpenAI and Langchain

Louise Gabrielle L. Talip and Reginald Neil C. Recario

***Abstract*— In response to the multi-faceted challenges that researchers face in understanding academic literature, this study introduces IskolarBot – an AI-driven chatbot as a research support tool for researchers. With the use of contemporary tools such as OpenAI, Langchain, and Pinecone, researchers can use the chatbot to streamline their research processes such as extracting relevant information efficiently and fostering a deeper understanding of academic papers. The study addresses methodological gaps by introducing novel approaches to research support, enhancing scalability, and lowering accessibility barriers associated with high costs of proprietary models. This is especially beneficial in the Philippines where researchers have limited access to resources and tools. Through the RAGAS evaluation method and user usability testing, the chatbot demonstrated effectiveness in generating relevant and accurate answers, highlighting its value and impact as a research assistant. The findings also emphasize the chatbot's positive impact on user experience, with respondents indicating that they would use the system and recommend it to other researchers. This research contributes to advancing innovative technologies in the Philippine academe and empowering researchers in their academic pursuits.**

***Index Terms*— chatbot, research tool, langchain, retrieval augmented generation, large language models**

I. Introduction

In the ever-evolving realm of education and academic research, the need for innovative tools and technologies remains a persistent challenge. The rise of the digital age gave way to a wide variety of tools such as for data collection and analysis (e.g. Google Sheets, Python), collaborative writing (e.g. Overleaf), and even for plagiarism detection (e.g. Turnitin). However, these tools primarily cater to experienced researchers because they often require a level of technical expertise that can be challenging for individuals without specialized knowledge. Furthermore, while these tools excel in writing and coding tasks, they fall short in facilitating the crucial aspect of comprehending existing research literature — a cornerstone of effective research. Reading is a crucial aspect of the research process, influencing both topic selection and formulation [1]. It is a fundamental step in shaping one's own research direction and synthesizing new knowledge as it is not just a supplement to empirical methods, but a method of inquiry in itself [2]. By comprehending previous studies, researchers can actively shape the path of their own investigations. Currently, there remains a scarcity of resources for these researchers, particularly students, in initiating their academic journey, especially when it requires specialized skills and deeper technical and theoretical understanding. This gap in accessible tools and resources becomes especially apparent in countries like the Philippines, where the pursuit of research is gradually declining [3] [4]. It is imperative to foster the culture of innovation and research among scholars through accessibility and approachability. Developing entry-level tools like a research assistant chatbot tailored for academics in the Philippines could serve to bridge this gap. By allowing access to simplified academic resources, there is potential to stimulate increased interest and participation in research, potentially contributing to the nation's intellectual advancement and innovation.

A. Background of the Study

Chatbots are intelligent systems that can hold conversations with humans using natural language [5]. They have a wide range of applications, including education, information retrieval, business, and e-commerce [6]. Over the years, chatbots have been developed to automate interactions, streamlining communication between users and providing convenient access to valuable information for both businesses and individuals. They prove particularly useful in industries where immediate assistance might not be available, serving as an accessible resource. Chatbots efficiently address common inquiries such as specific restaurant operating hours, detailed airline guidelines, or government institution procedures.

The use of chatbots in education has been on the rise in recent years [7]. These AI-powered virtual assistants have demonstrated significant potential in streamlining various tasks such as student onboarding or providing instant general support to students such as Dashly and enhancing the learning experience such as Socratic, Mathway, and Duolingo. Chatbots have become integral in improving student engagement, retention, and overall educational outcomes, especially now with the rise of ChatGPT and applications powered by it.

Chatbots either execute specific tasks or engage in open-ended conversations. Many methodologies have emerged in this domain, spanning from initial hard-coded response systems to more sophisticated Artificial Intelligence-driven advancements, showcasing the evolution of techniques employed in their development [8]. With the advancements in artificial intelligence, chatbots have progressed to offer more comprehensive and dynamic answers.

Presented to the Faculty of the Institute of Computer Science, University of the Philippines Los Baños in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science

They can now be trained or prompted to understand a wider scope of topics with the aid of appropriate datasets, such as in the case of OpenAI's ChatGPT [9]. Consequently, diverse sectors, including education, have embraced these evolving technologies.

The rise of Large Language Models have been revolutionizing the landscape of natural language processing. Their ability to comprehend and generate text with a human-like fluency has led to their wide adoption across various industries [10], including education. LLMs have become indispensable tools, aiding students, educators, and researchers in a variety of tasks. They can now assist in drafting essays and automating data analysis such as ChatGPT [9] or writing code through GitHub's Copilot [11].

Due to the popularity and demand of these models, developers have also created frameworks to better automate processes and build applications on top of the OpenAI API. One of these tools is called Langchain, a framework specifically designed to facilitate the creation of said applications. Its objective is to enable developers to use any large language model and use it across multiple sources of data or seamlessly interact with multiple APIs [12].

This paper explores the potential and significance of developing a chatbot-driven solution that addresses academic inquiries and provides related literature using a proprietary set of data. Its primary focus lies in aspiring researchers in the University of the Philippines Los Baños. This system utilizes OpenAI's advanced language model functionalities integrated with the Langchain framework, coupled with Chainlit for its user interface.

B. Statement of the Problem

Guido and Mangali [13] stated that "a country's economic and intellectual wealth is associated with the productivity of its research". Their study indicated a notable connection between the scores from the Programme for International Student Assessment (PISA) and a country's research productivity. The Philippines in particular has been found to have a low research output and low PISA scores. The country falls below the average band line for Southeast Asian nations in terms of published documents, citations, self-citations, and open access [14]. Despite the country recording the lowest PISA scores among its Southeast Asian counterparts, its research performance and quality remained relatively commendable within the region.

In a study in 2015 by Wa-Mbaleka [15], one of the major impediments to extensive publication is the time constraint. The process of reviewing numerous papers is notably time-intensive. This is evident across all levels of expertise in research. Novices are susceptible to feeling overwhelmed while more experienced researchers, often occupied with demanding schedules, lack the privilege to dedicate ample time to thoroughly analyze academic papers. A recent study by Lobo [16] unraveled that students have encountered obstacles related to infrastructure, communication, and time management while undertaking their thesis work. Bueno [4] in 2019 also confirmed Wa-Mbaleka's findings in their study, but also added that the lack of solid foundation, along with the lack of resources, also contributes to the lack of interest from researchers to publish more papers. With these factors and many more, it has been increasingly difficult to find funding, which has always been a major barrier to research. The 2015 report from the UNESCO Institute of Statistics [17] [18] highlights that a mere 0.2% of the country's gross domestic expenditure is directed toward research and development, which is significantly lower than the global average. This negative cycle persists to the present day, characterized by the minimal return in the country's research output, the absence of a research culture, and time constraints, ultimately resulting in a withdrawal or reallocation of funding from the government. Consequently, this either discourages researchers from pursuing a career in the academe or compels them to pursue career opportunities abroad.

Furthermore, within the Philippine research landscape, there is a notable absence of documented evidence or data on the utilization of tools that leverage Large Language Models, specifically research assisting chatbots. The current research practices in the country also emphasize the need for enhanced research methods that can optimize the exploration and comprehension of scholarly literature. With the rise of LLM-driven tools in developed nations, it is practical to explore its potential applications within the Philippine context.

It is imperative to foster the culture of research in the Philippines as the country exhibits considerable promise for the development of research in various fields, creating significant strides despite the limited resources. This is evidenced by the development of the country's own COVID-19 test kits [19]. Moreover, the nation boasts a substantial number of over 7,000 published researchers from diverse backgrounds and disciplines, according to the ADS Scientific Index [20]. There is also a steady increase in the number of publications since the early 2000s, as indicated by Scival [21] [22]. Despite the various challenges impeding research development, it remains imperative to contribute to and facilitate the advancement of Philippine research by alleviating the burdens faced by researchers.

C. Objectives of the Study

The primary focus of this study is to develop a research assistant chatbot that facilitates researchers in comprehending scientific publications.

Specifically, this study aims to:

1) Develop and deploy a web-based application with Chainlit featuring a chatbot with advanced conversational understanding capabilities in complex academic contexts using Langchain and OpenAI to address the challenge of natural language processing in research assistance.
2) Devise streamlined techniques for content extraction from academic papers, including the retrieval of accurate, simplified answers, generation of clear summaries, keywords, and creation of proper citations, utilizing the functionalities of Langchain and OpenAI.
3) Evaluate and improve the accuracy of information retrieval from academic papers using the RAGAS library.

4) Evaluate the chatbot's effectiveness through a Usability Test where users interact with the system and provide insights, assessing its usability and performance in aiding researchers.

D. Significance of the Study

In the context of the Philippines, the research and academic landscape are often deemed inaccessible to many researchers due to the niche areas of study and steep learning curves.

There remains a notable gap in the field, specifically in the lack of conclusive research in the utilization of AI-powered chatbots as a means to enhance the accessibility and comprehension of academic literature for researchers. Existing applications like ChatPDF [23] and the AI Research Assistant by Elicit [24] exhibit limitations, handling only one PDF at a time and having restricted datasets, often excluding Philippines-based papers. In contrast, a tool incorporating custom data offers more accessibility to local resources. Previous studies have scarcely addressed this in the context of research in the Philippines, leaving a significant gap in understanding how such technologies can be used effectively. Thus, this study focuses on bridging the literature gap in the field.

There is also a notable absence of documented evidence or data on the utilization of research-assisting chatbots. This empirical gap necessitates new evidence to validate the effectiveness and impact of integrating advanced LLMs like OpenAI with innovative frameworks like Langchain. By addressing this gap, researchers can gain insights into the practical implications and benefits of leveraging AI-driven chatbots in the research process, contributing to a more evidence-based approach to research support tools.

The academic landscape in the Philippines also calls for more enhanced and effective methods that can optimize the comprehension of scholarly literature. By leveraging AI-driven chatbots, researchers can streamline their research processes, extract relevant information more effectively, and foster a deeper understanding of academic content. This methodological gap is addressed by adopting innovative approaches to research support, paving the way for more efficient and insightful research practices in the academic community. Moreover, high costs associated with proprietary models hinder accessibility, especially for entry-level scholars with limited purchasing power. Utilizing the methodology of this study improves scalability by dividing documents into smaller chunks of data, saving tokens for data retrieval and expediting the process. This approach of data segmentation optimizes efficiency by targeting specific segments containing the pertinent information needed. In the Philippines where access to GPUs is limited and the cost is steep, innovative tools play a crucial role in accessibility. The disproportionately high cost of existing applications significantly impede the involvement of aspiring individuals in the academe. By addressing these challenges, it lowers barriers and fosters greater participation in the research process.

In summary, the exploration of novel technologies, the validation of the usage of chatbots, and the adoption of innovative research methods are focal points in addressing the challenges faced by researchers in navigating academic literature. This study aims to contribute to bridging these gaps and advancing research support tools to empower researchers in their academic pursuits.

II. REVIEW OF RELATED LITERATURE

A. Evolution of Chatbot Technology

The evolution of chatbot technology has been significant, with a shift from rudimentary models to advanced intelligent systems [25]. The term "Turing Test" emerged from Alan Turing's research paper entitled "Computing Machinery and Intelligence" in the 1950s where an artificial intelligence has to pass as human according to certain benchmarks. The development of ELIZA in the Massachusetts Institute of Technology in 1996 led to the first chatbot that passed the Turing Test due to how they didn't foresee how many individuals would easily attribute human-like emotions to the program. According to Adamopouloou [26], it employed basic pattern matching and a response system based on templates. Deshpande further states that the first few chatbots operated on these predefined response pairs linked to specific inputs. Using pattern matching and string processing, it facilitated ongoing conversations between computers and humans, but it lacked the criteria to be considered an intelligent chatbot. Chatbots have then evolved from simple keyword recognition to "evaluating user input victimization through heuristical pattern matching rules" such as in the case of ALICE (Artificial Linguistic internet computer Entity) in 1995. This evolution has been driven by the maturation of AI technologies, the integration of NLP, and the recent developments of using Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) to achieve AI [27]. The prevalence of chatbots has been on the rise, finding applications across diverse fields such as marketing, education, and healthcare [26], which is clear in Figure 1 below – Scopus search results from 2000 to 2019 using keywords like "chatbot," "conversation agent," or "conversational interface".

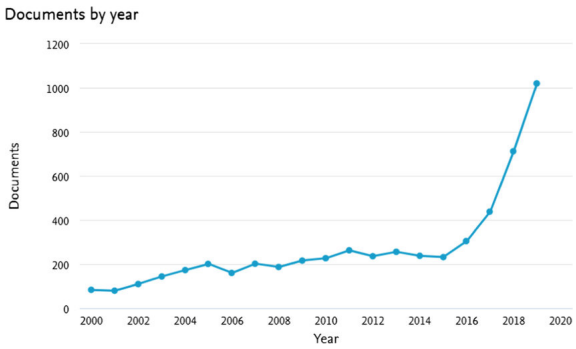


Fig. 1: Scopus search results (2000-2019) of the keywords "chatbot" or "conversation agent" or "conversational interface" [26]

Presently, the release of OpenAI's ChatGPT has sparked significant interest and concern in the AI community. Haque in 2023 [28] emphasizes the revolutionary nature of ChatGPT and its potential to automate a range of tasks such as translation, customer service, and content generation could

significantly influence and dominate industries because of its ability to generate human-like responses and answer a broader range of topics. This is because it has been trained with a model that has 175 billion parameters allowing it to generate quality responses.

B. Language Models and OpenAI

Large Language Models (LLMs) are a groundbreaking development in the field of artificial intelligence as they are a class of language models that have demonstrated exceptional performance in natural language processing tasks both in understanding and generation [29]. Moreover, LLMs are pre-trained on a vast database of text and possess the ability to generate coherent and contextually relevant text in response to user queries. These models, armed with their extensive knowledge, serve as the foundational technology for many chatbot systems.

According to MindsDB [30], OpenAI, a pioneering research organization in artificial intelligence, has played a pivotal role in advancing LLMs and making them accessible to the broader community. Their GPT-3.5 model, in particular, has garnered widespread attention for its exceptional language generation capabilities. OpenAI's infrastructure and API have enabled developers to create sophisticated chatbots and applications that leverage the power of GPT-3.5. Now that they have released GPT-4, it has been the leading model in terms of quality responses. MindsDB [30] also shows in the table below the models that are at the forefront today.

Model	Provider	Open-Source	Speed	Quality	Params	Fine-Tuneability
gpt-4	OpenAI	No	★☆☆	★★★★	-	No
gpt-3.5-turbo	OpenAI	No	★★☆	★★★☆	175B	No
gpt-3	OpenAI	No	★☆☆	★★★☆	175B	No
ada, babbage, curie	OpenAI	No	★★★	★☆☆☆	350M - 7B	Yes
claude	Anthropic	No	★★☆	★★★☆	52B	No
claude-instant	Anthropic	No	★★★	★★☆☆	52B	No
command-xlarge	Cohere	No	★★☆	★★☆☆	50B	Yes
command-medium	Cohere	No	★★★	★☆☆☆	6B	Yes
BERT	Google	Yes	★★★	★☆☆☆	345M	Yes
T5	Google	Yes	★★☆	★☆☆☆	11B	Yes
PaLM	Google	Yes	★☆☆	★★☆☆	540B	Yes
LLaMA	Meta AI	Yes	★★☆	★★☆☆	65B	Yes
CTRL	Salesforce	Yes	★★★	★☆☆☆	1.6B	Yes
Dolly 2.0	Databricks	Yes	★★☆	★★☆☆	12B	Yes

Fig. 2: Table of Large Language Models [30]

Because of this, OpenAI is the optimal model to use for the chatbot to ensure the best responses for each query. Depending on the use case of the prompt, a developer can choose from the different models that OpenAI offers – some models are more appropriate for conversational chats while others are suitable for direct answers to queries. The models are also priced differently due to the number of tokens that it can accept and generate. The data in tables I and II are from the official OpenAI website which shows the differences of these models [9].

Model Name	Tokens	Training Dataset
gpt-4-1106-preview	128,000	Up to Apr 2023
gpt-3.5-turbo-1106	16,385	Up to Sep 2021
gpt-3.5-turbo	4,096	Up to Sep 2021
gpt-3.5-turbo-16k	16,385	Up to Sep 2021
gpt-3.5-turbo-instruct	4,096	Up to Sep 2021

TABLE I: Comparison of OpenAI Models - Tokens and Training Dataset [9]

Model Name	Pricing (Input)	Pricing (Output)
gpt-4-1106-preview	$0.01 / 1K tokens	$0.03 / 1K tokens
gpt-3.5-turbo-1106	$0.0010 / 1K tokens	$0.0020 / 1K tokens
gpt-3.5-turbo	$0.002/1K tokens	$0.002/1K tokens
gpt-3.5-turbo-16k	$0.003 /1K tokens	$0.004 /1K tokens
gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens

TABLE II: Comparison of OpenAI Models - Pricing for Input and Output [9]

Another crucial component utilized in this study is OpenAI's function calling, where the model intelligently decides which tool is used for a user's query. It outputs a JSON object containing specified arguments [9]. The function call holds a collection of tools curated for specific usage based on the user input. These tools can contain functions utilizing different types of chains, different APIs and libraries, or simple Python code, enabling an extensive scope of capabilities.

C. Applications in Academia and Research

According to Shawar and Atwell [6], originally, chatbots were created for amusement and to simulate human dialogue. While this purpose remains prevalent in chatbot development, the increased traction of this technology has led to diverse applications. Subsequently, chatbot technology has found utility in various domains, including information retrieval, query resolution, facilitating evidence-based decisions, and many more. Abdelhamid [31] further explores the use of chatbots as smart teaching assistants, stating that students are much more likely to use and engage with the system than ask questions directly to their professors.

The prevalence of chatbots in education and research has also been increasing, especially now with the advancement of language models [32]. Conversational agents, which facilitate easier access to information through engaging interactions, have prompted tools to improve development processes. This includes refining user experience and design, improving frameworks and platforms, among other enhancements.

However, the integration of chatbots in education has been focused on skills such as subject-learning, reading and writing, speaking, and language translation [33]. The utilization of OpenAI's ChatGPT has significantly influenced how learners process and comprehend information. The results in Memarian's [34] study highlighted that ChatGPT offers extensive potential for various applications, including personalized and complex learning, customized teaching approaches, diverse learning activities, assessments, asynchronous communication, feedback systems, accurate research, personas, task distribution, and cognitive support.

Significant gaps and a lack of tools still persist in the field of research. Kooli [7] stated that AI technologies are envisioned to revolutionize research and education by automating laborious and repetitive tasks, assisting in data analysis, and fostering innovative modes of learning and evaluation. AI-powered research assistants would be advantageous in the academic field as it helps in fact-checking and offers quick access to pertinent information. These systems serve as efficient tools for gathering data, operating round-the-clock and processing large volumes of data swiftly. This capability enables researchers to amass high-quality data and obtain precise information at any time, thereby mitigating the potential for human error. Nonetheless, Kooli emphasizes that the primary accountability for the content and quality of the research still remains with the human researcher. A research chatbot functions solely as a tool, not as a substitute for the indispensable role of the researcher. Similar to human research assistants, these chatbots need thorough and ongoing oversight to prevent deviations.

Lund and Wang [35] states that ChatGPT holds potential for improving the academe in multiple ways, such as its advanced capabilities in literature reviews, generating text, automating summarization, and answering queries. It can also assist through extensive search and referencing features. These capabilities offer researchers the opportunity to streamline their workload, allocating more time and energy towards the creative and analytical facets of their endeavors. However, constraints within ChatGPT and the absence of an effective framework pose limitations. Hence, the creation of a structured chatbot using a framework like Langchain stands as an ideal approach to manifest these capabilities.

Lastly, according to Klimova and Seraj [33], AI-powered chatbots significantly influence students' skills with the use of text, speech, graphics, haptics, and gestures to assist them in completing their tasks. These chatbots extend a broader spectrum of services to students, enhancing their motivation and expanding the traditional learning paradigm into the digital realm. This offers a promising outlook, suggesting the potential to augment research processes, facilitating an easier, more accessible way of learning and fostering an increased interest in the field.

D. Langchain

Topsakal and Akinci [12] defines LangChain as a framework, developed by Harrison Chase, that is designed for creating applications that leverage a variety of large language models. It serves as a middleware layer, where its primary objective is to facilitate developers in seamlessly integrating various data sources and APIs from a diverse range of applications. To achieve this, LangChain offers components, which are modular abstractions and chains – adaptable pipelines customized for specific use cases. Sreeram and Sai in 2023 [36] adds that specifically, it is "a cutting-edge solution which helps us in the querying process and extracting information from PDFs. With its advanced NLP algorithms, it helps users to interact with the PDFs and makes the document search and retrieval very easy". Langchain's unique capabilities make it a valuable asset in creating a chatbot that can provide users with accurate and efficient access to different sources of data, including research papers.

The primary components from Langchain that will be used are the different chains. As stated in the official Langchain documentation [37], chains are the building blocks of Langchain, allowing the developers to generate responses from the chosen LLM using templates and the user input. Topsakal and Akinci [12] explains that these chains can be structured and sequenced in ways beyond the capabilities of standard LLM APIs. For instance, they can utilize preceding chain responses as inputs for subsequent chains. These chains may be organized linearly or amalgamate answers from multiple chains into a single conclusive chain. Additionally, some chains can be routed to different chains with varying templates based on user input.

E. Limitations and Challenges

The use of tools can only do so much for Philippine researchers, who face a multitude of challenges. These include limited funding and resources [38], exacerbated by issues of accessibility and local attitudes towards science [39]. Furthermore, academic staff in the country are often ill-prepared for their roles in research training and supervision [40] and the educational system as a whole is plagued by perennial issues [41]. These challenges collectively hinder the effectiveness of tools in supporting the research efforts of Philippine scholars. While a chatbot system would aid in simplifying the research process, there is no guarantee of the gravity of its impacts to the research community as a whole and its influence on the culture and perspective towards research in general.

Although the implementation of an AI chatbot would be advantageous and convenient within an ideal context, it is essential to acknowledge that there are also inherent limitations concerning its functionalities and the implications for its potential users. Some AI models tend to be biased in a number of complex ways, mostly stemming from various factors including training data, model specifications, and algorithmic constraints, product design, and policy decision [42]. A study by Cirillo, et. al. in 2020 [43] has proved that the majority of present biomedical AI technologies lack mechanisms to detect biases and that its design fails to consider the differences in health and disease outcomes influenced by the sex and gender of an individual, which will heavily alter the way it will suggest treatment and holds the potential to "produce mistakes or discriminatory outcomes". ChatGPT, for example, has a tendency to generate biased responses. The various shortcomings encompass superficial, inaccurate, or incorrect content, as stated in a study by Sallam in 2023 [44]. Ethical concerns and risks emerged that are associated with biased data in training and potential plagiarism concerns. A crucial aspect mentioned in the study as well was the concept of ChatGPT hallucination, a risky phenomenon that demands careful evaluation by experts or researchers. This is particularly significant due to ChatGPT's capacity to produce scientifically plausible yet incorrect content, emphasizing the necessity for vigilant scrutiny and evaluation within these domains. There have

also been a number of case studies that highlight the citation inaccuracies from ChatGPT, as well as insufficient or non-existing references [45]. These are concerning, as biases can influence the information and opinions it generates, potentially perpetuating harmful language patterns [46].

Another drawback introduced by Lopez and Qamber in 2022 [47] is a lack of human assistance in situations where the questions are too complex for the chatbot to answer or when the user prefers human contact. Salvagno [48] claims that AI chatbots such as ChatGPT possess the potential to support the writing process of scientific papers and aid in conducting literature reviews, pinpointing research inquiries, presenting an overview of a field's current status, and providing assistance in tasks like formatting and language review. Both studies confirm that chatbots can serve as tools that aid human researchers; however, they should not substitute for the expertise, judgment, and individuality of human researchers.

LLMs have shown promise in natural language processing and its integration into educational chatbots remains a multi-faceted challenge. It is evident that there is a need for tools that can help aspiring researchers effectively extract relevant content from papers to alleviate the challenges associated with writing and to foster an environment that sparks their interest in research writing.

III. Methodology

A. Development Tools

*1) **Hardware**:* The system was developed on an ASUS Strix GL503GE machine with the following specifications:

- **Operating System:** Microsoft Windows 10
- **Processor:** Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz, 6 Core(s), 12 Logical Processor(s)
- **Memory:** 8GB RAM
- **Storage:** 256 SSD, 1TB HDD

*2) **Software**:* The development environment utilized the following software tools:

- **Programming Language:** Python 3
- **Frontend Framework:** Chainlit
- **Backend Framework:** Langchain
- **Large Language Model:** OpenAI
- **Vectorstore:** Pinecone
- **Debugging and Logging:** Langsmith
- **Testing Library:** RAGAS

B. Framework

The framework for this chatbot primarily relies on the capabilities of Langchain, OpenAI, and Pinecone. These tools are the components used for data ingestion and data processing. User interaction is provided by Chainlit and the LLM to be used is OpenAI's GPT-3.5 Turbo. The framework starts with the preprocessing phase, where the chatbot ingests and organizes the dataset provided and saves it into a vectorstore. This preparatory stage involves data collection, PDF loading, text chunking and splitting, and storage in a structured format in a vectorstore. This establishes the foundation for retrieval later during the subsequent user interactions. The operational phase focuses on the main application features such as answering the user queries, intelligent decision-making function calls, and information retrieval mechanisms. This framework merges contemporary techniques such as that of Topsakal and Akinci in 2023 [12] and Sreeram [36] for building applications using large language models, enabling the chatbot to adeptly handle a wide array of user queries. Refer to Appendix I for a visual representation of the framework.

C. Preprocessing Phase: Data Collection

The study involved gathering research papers from diverse publications, accessing repositories through OpenAthens and downloading from platforms including arXiv, IEEE Xplore, ScienceDirect, and JSTOR. Priority was given to papers with predominantly textual content to align with the framework's emphasis on text comprehension, thereby excluding those predominantly filled with images. The collection of these datasets strictly adheres to data privacy regulations and licensing agreements associated with its respective subscriptions. None of these datasets were publicly shared and are solely intended for academic purposes.

The framework categorizes research papers into six distinct categories within Computer Science. These categories are Artificial Intelligence, Cryptography and Cybersecurity, Data Structures and Algorithms, Human-Computer Interaction, Operating Systems, and a General category. Papers are identified and assigned to these categories based on the keywords they are tagged with in the platforms where they are sourced. The publications included in the framework primarily span from 2014 to 2024, reflecting recent developments and trends in the field. However, some Human-Computer Interaction papers from the 1990s are also included to provide historical context and perspective. Each category contains a mutually exclusive set of 15 research papers, with the page count varying per category, as illustrated in Table 3.

The amassed data was stored systematically in Pinecone, a specialized vector database designed for efficient indexing and storage of vector embeddings. This facilitates efficient retrieval and similarity search processes. To ensure accessibility, each paper was organized within its respective category folder and a Pinecone serverless index was created for each category. The decision to opt for serverless indexing over pod-based indexes (non-serverless) was motivated by its inherent scalability, where the rate is dynamically calculated based on the amount of data stored and operations performed, without imposing minimum requirements [49].

Each index within Pinecone is capable of accommodating a significant number of vectors. With an index dimensionality of 1536, a single index can hold up to 2.5 million vectors, representing a remarkable capacity for storing and accessing research papers. The breakdown of vectors per category is also shown in Table 3, demonstrating the impressive capacity of the vector store. The scalability of Pinecone's indexing system is exemplified by the substantial room for expansion within each category. For instance, the General category, with its current allocation of 3,709 vectors, could potentially accommodate thousands more papers, underscoring the efficiency and scala-

Category	Pages	Vectors
Artificial Intelligence	126	625
Cryptography and Cybersecurity	151	876
Data Structures and Algorithms	269	1007
Human-Computer Interaction	108	601
Operating Systems	128	600
General	782	3,709

TABLE III: Pinecone Vectorstore Data

bility of the vector storage system in managing and retrieving vast volumes of academic papers.

D. Preprocessing Phase: Chunking and Text Splitting

The collected papers, in PDF format, were loaded using Langchain's document loader functions in order for the data to undergo a process of segmentation into manageable 'chunks'. This chunking process breaks down extensive documents into smaller sections, enabling efficient handling and analysis. This is also done to adhere to the token limitations of OpenAI's API. Post-segmentation, these chunks undergo parsing to extract essential textual content, subsequently stored as 'embeddings'. These embeddings serve as numerical representations of the text, facilitating streamlined retrieval and comparison. Each chunk is assigned numerical data based on its content relevance, aiding later retrieval of closely related chunks upon query. Alongside these embeddings, the metadata also stores the file name of the original PDF and the exact IEEE reference. The metadata provides contextual information about the paper's identity and facilitates the generation of more refined output during the operational phase. By preserving the connection between each chunk and its source document through the PDF file name and IEEE reference, the system ensures coherence and accuracy in subsequent processing stages.

In Pinecone, cosine similarity is utilized to create the vector embeddings. This measures the cosine of the angle between two vectors. The formula for cosine similarity between two vectors $(\mathbf{a})$ and $(\mathbf{b})$ is:

$$sim(a, b) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

where:

- $(\mathbf{a} \cdot \mathbf{b})$ is the dot product of vectors $(\mathbf{a})$ and $(\mathbf{b})$
- $(|\mathbf{a}|)$ and $(|\mathbf{b}|)$ are the magnitudes (or lengths) of vectors $(\mathbf{a})$ and $(\mathbf{b})$, respectively.

According to Pinecone's documentation [49], cosine similarity is employed for semantic search and document classification tasks. This metric calculates the dot product of vectors by multiplying their components and subsequently dividing the result by the product of their magnitudes, which is determined by taking the square root of the sum of the squares of the vector components. Cosine similarity offers a reliable measure of content similarity between documents as it assesses the alignment of vectors.

The RecursiveCharacterTextSplitter from the langchain_text_splitters library is utilized to divide the text into chunks, with a defined length of 1000 characters per chunk and an overlap of ±150 characters between adjacent chunks to avoid losing context, as shown in the code snippet below.

```
text_splitter =
RecursiveCharacterTextSplitter(
    separators=["\n\n", "\n"],
    chunk_size=1000,
    chunk_overlap=50
)
```

Once split, the text is stored in the Pinecone vectorstore. The embeddings model employed is the text-embedding-ada-002 from OpenAI. This model has demonstrated superior performance across various tasks, including text search, code search, and sentence similarity. Notably, it outperforms its predecessor, Davinci, in most tasks while being significantly more cost-effective, with a price reduction of 99.8% [9].

E. Operational Phase: Application Features

The operational phase involves the user interaction with the chatbot where they first choose the category they want to ask about. The vectorestore is then set up for that category and the features that a user can use are as follows:

- **OpenAI Function Calling:** Function-calling is an OpenAI tool equipped with decision-making capabilities which determines the appropriate tool to employ and the dataset to access for acquiring the proper response. This process is able to outline functions and direct the model to intelligently produce a JSON object containing the necessary arguments for initiating one or more functions. The Chat Completions API is used to generate a JSON output which then can be utilized to activate the embedded function [9]. There is no limit to the number of functions that the call can have, but careful consideration is needed with the prompts as too many tools with convoluted instructions can cause hallucination.
  The code snippet below shows the structure of a tool. Each function and their respective arguments are defined through prompting in the description field. This is a simplified version where the list contains 1 tool called 'get_answer' with 2 arguments called 'question_type' and 'semantic_keywords'.

```
tools = [
    {
      "type": "function",
      "function": {
        "name": "get_answer",
        "description": "",
        "parameters": {
          "type": "object",
          "properties": {
            "question_type": {
              "type": "string",
              "description": ""
            },
            "semantic_keywords": {
              "type": "string",
              "description": ""
            }
          },
          "required": ["question_type",
          "semantic_keywords"]
        }
      }
    }
]
```

The code snippet below shows the overview on how the OpenAI model is used to determine which tool to use.

```
client = OpenAI()
model = "gpt-3.5-turbo-0125"
response = client.chat
.completions.create(
  model=model,
  messages=messages,
  tools=tools,
  tool_choice=tool_choice,
)
```

The response of the function call dynamically determines which of the specified tools are needed to answer the user's query. It also returns the necessary arguments for each respective tool.

- **Academic Paper Summarization:** This feature offers two methods of paper summarization. Firstly, utilizing the Load Summarize Chain in Langchain to produce an overview of the whole paper. It uses the ChatOpenAI chat model by Langchain with the LLM set to gpt-3.5-turbo-0125 and temperature set to 0 for full control of the output via prompt engineering. For the summary, it ingests all of the contents of the PDF and uses the Map-Reduce chain. The code snippet is shown below.

```
chain = load_summarize_chain(
    llm,
    chain_type="map_reduce"
)
```

The Langchain documentation [37] states that the map reduce documents chain involves two main steps: Map and Reduce. Initially, it applies an LLM chain to each document chunk individually, summarizing each chunk and treating the output of each chain as a new document (the Map step). All the new documents are then passed to a separate combine documents chain to produce a single output (the Reduce step). The mapped documents can be compressed or collapsed beforehand to ensure they fit within the combine documents chain. This compression process is recursive if needed. Refer to the Appendix for the visual representation of this process.

By using the map reduce chain, the summarization process can efficiently handle large academic papers without being constrained by memory or processing power. This is crucial for scalability, allowing the system to process larger documents.

Secondly, users can specify a particular section of the paper to summarize such as the results or the methodology. The summary can be focused specifically on that section through a vectorstore similarity search and passing the related context to an LLM using the RetrievalQAWithSourcesChain, enhancing precision and relevance. This ensures that only the relevant content is used when creating the summary.

- **Related Literature Recommendation:** This feature suggests related literature based on the user's queries and parameters via a vectorstore similarity search as shown in the code snippet below.

```
documents = vectorstore
.similarity_search(
    query,
    k=10
)
```

This involves comparing the vector representation of the user's query against a database of vectors representing the academic papers. The similarity score between the query vector and each paper vector determines the relevance of the papers to the query. It returns at most 10 papers with the highest similarity scores for recommendation. Then, the metadata containing the IEEE reference is extracted and parsed into a coherent response. This method ensures that users are presented with the most relevant research.

- **Information Retrieval:** This feature includes the extraction of accurate and contextually appropriate information from the dataset, resulting in a detailed and understandable answer.

The response from the OpenAI function call allows the chain to identify the specific question type, the subject, and the semantic keywords associated with the query. The semantic keyword extraction involves analyzing the query to identify key terms and concepts that are semantically relevant to the information being sought. These keywords are then used to perform a similarity search within a vector store, retrieving chunks of context that match the query's intent.

```
    chain = RetrievalQAWithSourcesChain
        .from_chain_type(
      llm=llm,
      chain_type="stuff",
      retriever=pinecone_vectorstore
        .as_retriever(),
      chain_type_kwargs={
          "prompt": template,
      }
)
```

Leveraging the results of the vectorstore similarity search, the retrieved context chunks are processed using the RetrievalQAWithSourcesChain chain from Langchain as shown in the code above, which is designed to refine the final answer using the given context.

- **Other Tools:**
  - IEEE Reference Extractor: This tool extracts pertinent information from papers and generates references in IEEE format, enhancing the quality of the references in the output.
  - Semantic Keyword Extractor: This tool extracts keywords from user queries, improving contextual understanding in LLM-based responses.

F. Testing: RAGAS

Es, et. al. in 2023 [50] defines RAGAS (Retrieval Augmented Generation Assessment) as a framework designed for the reference-free evaluation of Retrieval Augmented Generation (RAG) systems. RAG systems, similar to the implementation of the chatbot in this study, consist of two main components: the retrieval module that fetch the context chunks from the vector database and the language generation module based on the OpenAI LLM that generates the natural language responses using the retrieved information.

In the RAGAS framework, the documents undergo evaluation through a predefined process wherein it generates questions and responses based on the dataset. The concept of 'faithfulness' is introduced within the metric and it represents the alignment of the generated responses with the ground truth. The RAG system is then rigorously tested against the ground truth by answering the same questions. This evaluation facilitates the calculation of values for other key metrics, each of which is explained and computed as follows:

*1) **Faithfulness**:* Faithfulness measures the degree to which the answer is factually correct and relevant to the given question, ensuring that the information provided is accurate and reliable.

According to the official RAGAS documentation [51], calculating the faithfulness score is given as follows:

$$faithfulness = \frac{|A|}{|B|}$$

where:

- A = number of claims in generated answer that can be inferred from given context
- B = total number of claims in the generated answer

*2) **Context Relevancy**:* This metric assesses the relevance of the retrieved context. Ideally, the retrieved context should contain only important information necessary to address the given query.

According to the official RAGAS documentation [51], calculating the context relevancy score begins by estimating the value of $|S|$ by identifying sentences within the retrieved context that are relevant for answering the provided question. The final score is computed using the following formula:

$$context\_relevancy = \frac{|S|}{|T|}$$

where:

- T = Total number of sentences in the retrieved context

*3) **Context Precision**:* Context precision evaluates the relevance and specificity of the retrieved context in relation to the question. It assesses how well the retrieved information matches the intent of the question and how accurately it addresses the specific aspects or details mentioned in the question. A higher context precision indicates a more targeted and relevant retrieval of information.

According to the official RAGAS documentation [51], this metric is computed using the question, the ground truth, and the contexts, following this formula:

$$context\_precision@K = \frac{\sum_{k=1}^{K}(Precision@k * v_k)}{L}$$

$$context\_precision@K = \frac{M}{M+N}$$

where:

- K = total number of chunks in the contexts and $v_k \in \{0, 1\}$ is the relevance indicator at rank k.
- L = total number of relevant items in the top K results
- M = true positives @ k
- N = false positives @ k

*4) **Context Recall**:* Context recall measures the effectiveness of the retrieval process in capturing and presenting all relevant information necessary to answer the question comprehensively. It assesses the system's ability to retrieve and include all pertinent details and nuances from the available context, ensuring that no essential information is overlooked or omitted. A higher context recall score indicates a more thorough and comprehensive retrieval of relevant information.

According to the official RAGAS documentation [51], it is computed using the ground truth and the retrieved context. The formula for calculating context recall is as follows:

$$context\_recall = \frac{|A|}{|B|}$$

where:

- A = the GT sentences that can be attributed to the context
- B = the number of sentences in GT

*5) **Answer Relevancy**:* Answer relevancy assesses the degree to which the answer provided by the system directly addresses the question in the generated test set. It evaluates whether the answer is informative and directly related to the content of the question. A high answer relevancy score indicates that the provided answer effectively addresses the query and is closely aligned with the information sought by the user.

According to the official RAGAS documentation [51], this metric is computed using the question, the context and the answer. It is defined as the mean cosine similarity of the original question to a number of artificial questions, which where generated (reverse engineered) based on the answer:

$$answer\_relevancy = \frac{1}{N}\sum_{i=1}^{N} cos(E_{g_i}, E_o)$$

$$answer\_relevancy = \frac{1}{N}\sum_{i=1}^{N} \frac{E_{g_i} \cdot E_o}{||E_{g_i}||||E_o||}$$

where:

- E_{g_i} = embedding of the generated question
- E_o is the embedding of the original question
- N is the number of generated questions, which is 3 default

G. Testing: Usability Test

Usability testing stands as a crucial methodology for ensuring the intuitiveness and the ease of use of applications. This approach includes several key steps, including participant selection, test script design, outcome interpretation, and creating recommendations. By systematically assessing the application's usability through real user interactions, usability testing provides valuable insights into user behavior and preferences, enabling informed design decisions and iterative improvements. Tailoring the testing method to the unique features of the application is crucial for ensuring accurate assessment and actionable results [52].

Three fundamental aspects are considered:

1) **User Demographic**: It is important to narrow down the target audience as they represent the intended user base, thereby facilitating more insightful feedback.
2) **Prior Knowledge/Experience in Research and Existing Technologies**: Assessing participants' prior knowledge and experience in research methodologies and technologies serves as a baseline for evaluating their familiarity with similar applications. This provides valuable context for interpreting the feedback and identifying areas for improvement.
3) **End-to-End User Experience and User Impression**: Evaluating the end-to-end user experience and impression of the application is essential in assessing the real-world usability of the application.

Refer to the appendix for the complete usability form created and used in this study.

IV. Results and Discussion

A. RAGAS

In the RAGAS implementation, the questions for the test set are generated based on a specific distribution. These questions encompass three categories: simple, reasoning, and multi-context. "Simple" questions allow direct retrieval of answers from the paper with little to no inference or additional context needed. "Reasoning" questions require deeper comprehension and inference, while "multi-context" questions demand background knowledge about various topics covered in the paper.

The provided code snippet illustrates the distribution percentages for generating these questions where 20% of the dataset is allocated for simple questions, 45% is allocated for reasoning questions, and 35% for multi-context questions. While testing multiple adjustments to these distributions, the evaluation of RAGAS consistently yields an 80% or higher score in the metric of answer relevancy.

```
distributions = {
    simple: 0.20,
    reasoning: 0.45,
    multi_context: 0.35
}
```

This distribution strategy prioritizes testing its reasoning abilities due to the increasing complexity of user queries as they ask more questions to the chatbot. Subsequently, multi-context questions are emphasized to assess the system's capability to retrieve relevant information across different contexts. Simple questions, requiring basic factual recall, receive lower priority as they are expected to be answered accurately even with limited contextual information.

The complete list of questions generated by the RAGAS library for each dataset is provided in the Appendix. While the test size was initially set to 15, as shown in the code snippet below, the final count of generated questions ranged between 10 and 15. This variance occurred primarily due to limitations within the test set generator. The generator operates within predefined constraints, and despite efforts to set the test size to 15, it may encounter situations where it cannot produce the full quota of questions. Factors contributing to this limitation include the complexity of the dataset, intricacies within the text such as images, charts, and code, and the inherent constraints of the algorithm. As a result, the final number of questions fluctuated within the specified range.

```
testset = generator
    .generate_with_langchain_docs(
        chunked_documents,
        test_size=15,
        distributions=distributions
    )
```

In the following figures below, each question is denoted by a letter. The columns in the table display the context relevancy, context precision, context recall, faithfulness, and answer relevancy for each category.

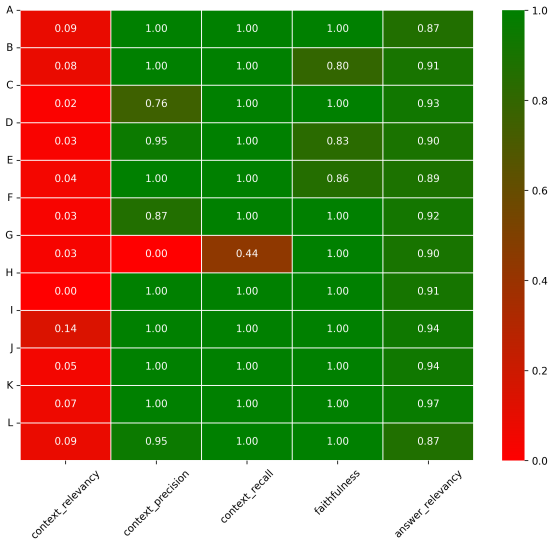


Fig. 3: Heatmap of the results for the Artificial Intelligence dataset

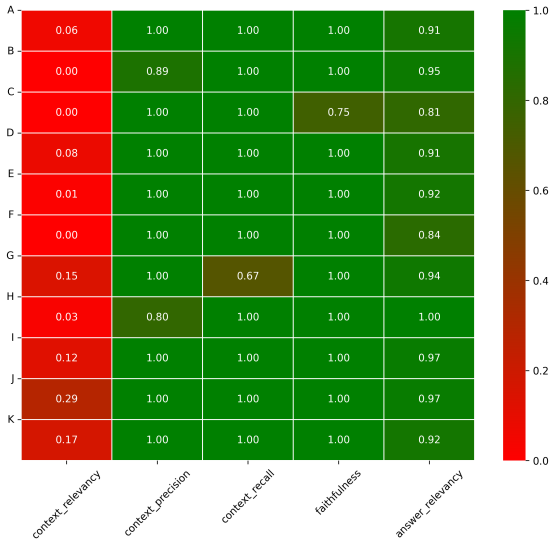


Fig. 5: Heatmap of the results for the Data Structures and Algorithms dataset

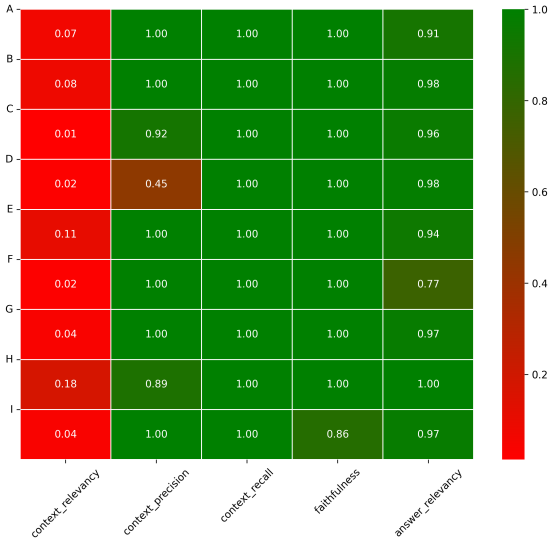


Fig. 4: Heatmap of the results for the Cryptography and Cybersecurity dataset

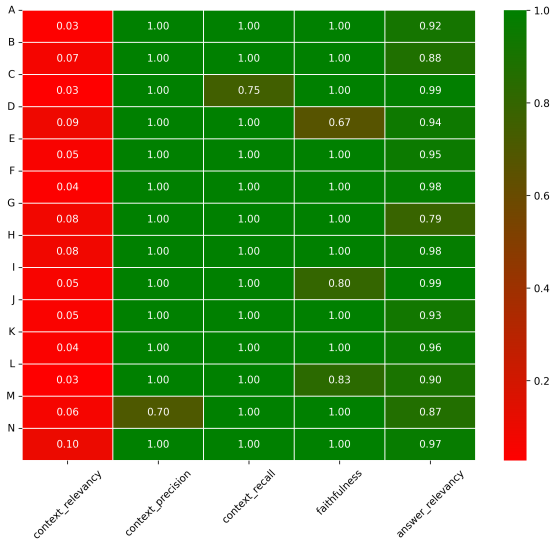


Fig. 6: Heatmap of the results for the Human-Computer Interaction dataset

*1) **Evaluation Metrics Overview**:* In assessing the performance of the application using the RAGAS implementation, a comprehensive set of metrics has been employed. This includes context relevancy, context precision, context recall, faithfulness, and answer relevancy. However, the pivotal metrics driving the evaluation are answer relevancy and context-related metrics. These metrics serve as fundamental indicators of the efficacy of the RAG implementation, highlighting its proficiency in both information retrieval and response construction.

*2) **Answer Relevancy Evaluation**:* The evaluation of answer relevancy across the various categories reveals promising outcomes. In the realm of Artificial Intelligence (AI), the metric of answer relevancy ranges impressively between 87% and 97%. Similarly, in Data Structures and Algorithms, the relevancy score stands between 81% and 100%. Human-Computer Interaction (HCI) exhibits a range of 79% to 99%, Operating Systems (OS) demonstrate relevancy scores varying from 84% to 100%, and Cryptography and Cybersecurity ranges from 77% to 100%, respectively.

*3) **Context Metrics Evaluation**:* An intriguing observation pertains to context relevancy, where the highest achieved score reaches only 29%. This discrepancy stems from the implementation of the similarity search algorithm.

```
results = vectorstore
    .similarity_search_with_score(
        query,
        k=10
    )
```

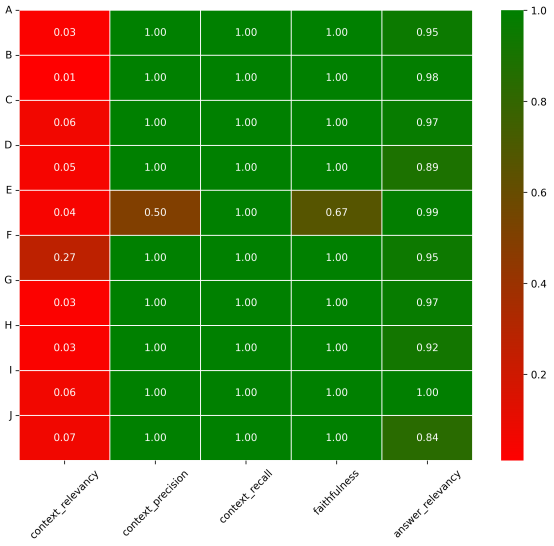


Fig. 7: Heatmap of the results for the Operating Systems dataset

The provided code snippet illustrates the vectorstore's attempt to retrieve 10 documents for every query. This implementation was made to ensure that the LLM receives ample context related to the user query. However, the chosen value of 10 exceeds the algorithm's default return, which is typically set to k=4. This adjustment was intended provide the LLM with the most comprehensive context available.

For all of the generated questions, despite only requiring the first few returned documents for an answer, the algorithm retrieves up to 10 documents. This led to a surplus of similar or associated but unused documents, which adversely affected the results of the metric.

Nevertheless, it's essential to note that despite the challenges in Context Relevancy, the Context Precision and Context Recall metrics yield exceptionally high results. These metrics are vital indicators of the system's ability to retrieve and utilize the correct data to formulate responses accurately. High Context Precision indicates the system's proficiency in matching retrieved information with the intent of the question, accurately addressing specific aspects or details mentioned in the query. Similarly, high Context Recall signifies the system's effectiveness in the retrieval process, capturing and presenting all relevant information necessary to comprehensively answer the question.

4) ***Discrepancies and Limitations****:* Several discrepancies surface within the metric results, notably falling below the 70% threshold. Notable examples include the Context Recall metric for question G in the Data Structures and Algorithms table, the Faithfulness metric for question D in the HCI table, the Faithfulness and Context Precision metric for question E in the OS table, and the Context Precision metric in the Cryptography and Cybersecurity table.

One contributing factor to these disparities is the occurrence of rate limit errors encountered during the utilization of the ChatOpenAI model. The rapid influx of requests per minute (RPM) originating from the RAGAS library exceeded the capacity of the OpenAI API. This overload occurred due to the asynchronous execution of multiple requests per question in the table, wherein metric evaluation represents one request.

It is crucial to emphasize that despite discrepancies in certain metrics, the overall accuracy of Answer Relevancy remains intact. In some instances, retrieved content provides supplementary information, contributing to a broader understanding of the queried topic. Additionally, limitations inherent within the algorithm may impact the generation and evaluation of certain questions.

B. User Usability Test

The application was deployed on Render with 0.1 CPU and 512 MB of memory allocation. This configuration negatively impacted the latency and speed of the application compared to its performance when run locally.

A total of 16 respondents from the Institute of Computer Science participated in the usability testing of the application. Convenient sampling was utilized for participant selection due to its practicality and accessibility, enabling the recruitment of participants from within the Institute's community.

The evaluation encompassed several key aspects, including the user's demographic information, their prior experience with research and research-specific tools, and their perceptions of the application's performance across various metrics. These metrics included Content, Timeliness, Ease of Use, and Functionality. Respondents were asked to rate their agreement with statements related to each metric on a five-point scale, ranging from "Strongly Disagree" to "Strongly Agree."

1) ***User Demographics****:* The age distribution of the respondents varied, as shown in Figure 8. Specifically, 62.5% were 24 years old, constituting the largest demographic group. The remaining respondents were predominantly in their early twenties, with 12.5% of respondents being 21 years old and 6.3% each for ages 22, 23, 25, and 35.

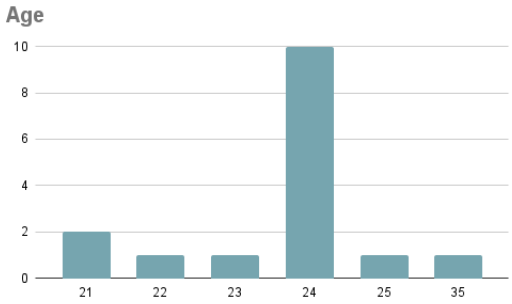


Fig. 8: Distribution of Age among the Respondents

In terms of gender distribution as shown in Figure 9, the majority of respondents identified as male, accounting for 87.5% of the respondents. One respondent (6.25%) identified as female, while another preferred not to disclose their gender.

Among the respondents, the distribution across university classifications varied, as shown in Figure 10. The largest group consisted of seniors, comprising 43.75% of the respondents. Juniors and the faculty and staff represented the next largest

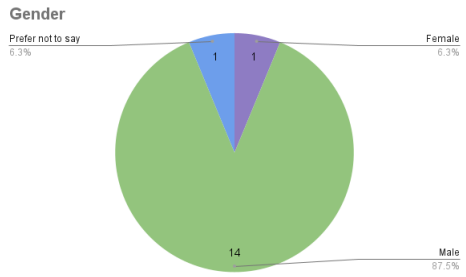


Fig. 9: Distribution of Gender among the Respondents

groups, with three (18.75%) respondents each. 12.5% classified as alumni and 6.25% as a graduate student. There were no respondents classified as freshmen or sophomores.

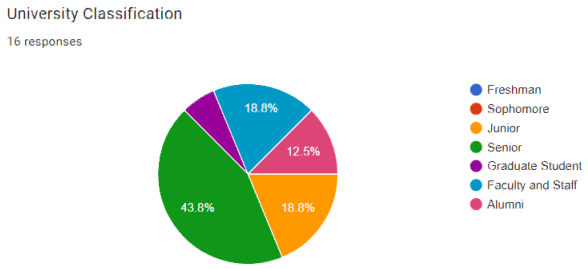


Fig. 10: Distribution of University Classification among the Respondents

*2) **Prior Knowledge/Experience in Research and in Existing Technologies**:* The majority of respondents (68.75%) are very to extremely familiar with chatbots or virtual assistants in general, with a significant portion, almost 50% using them every day, 43.8% of which use it multiple times a day. This indicates a high level of comfort and engagement with this technology among the respondents, suggesting that they may be receptive to using similar technologies for research assistance. The complete distribution for the respondents' answers are shown in Figures 11 and 12.

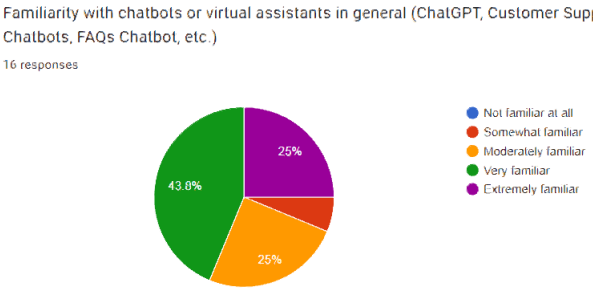


Fig. 11: Familiarity with chatbots or virtual assistants

43.75% are very to extremely familiar with academic databases or literature search engines, but most (56.25%) are moderately familiar to not familiar at all. Additionally, the majority (62.5%) rarely or occasionally use these resources

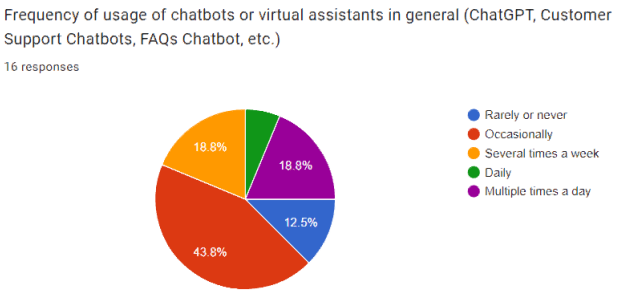


Fig. 12: Frequency of usage with chatbots or virtual assistants

and only 18.8% use it several times a week, indicating a low level of engagement or access to research-related information retrieval tools. The complete distribution for the respondents' answers are shown in Figures 13 and 14.

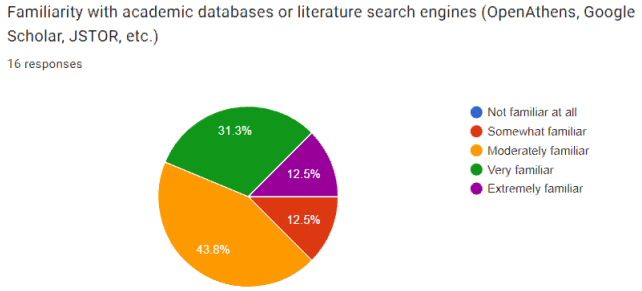


Fig. 13: Familiarity with academic databases

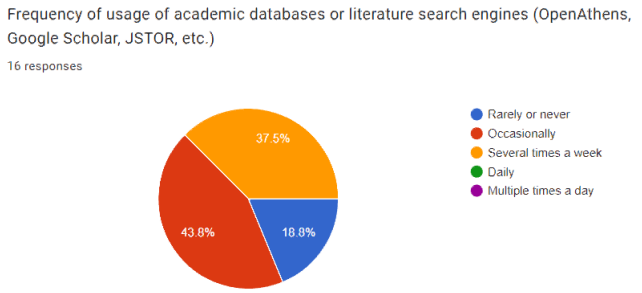


Fig. 14: Frequency of usage with academic databases

Additionally, most respondents (56.3%) have reported that they have not used any research-assisting chatbots while the remaining 43.7% have used the likes of ChatPDF, Scholarcy, and ChatGPT. Only a small proportion (37.5%) have used chatbots specifically designed for extracting information from research papers, with ChatGPT being the most commonly used. This is expected as chatbot assistants have only been prevalent for a year as of the publishing of this study. The complete distribution for the respondents' answers are shown in Figures 15 and 16.

100% of the respondents reported that they have conducted research as part of coursework, with 25% being actively

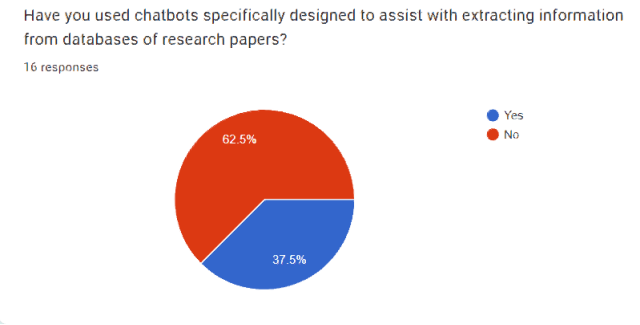


Fig. 15: Usage of research-assisting chatbots

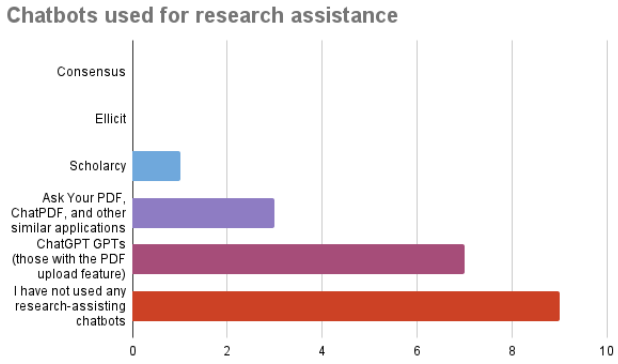


Fig. 16: Existing chatbots used

engaged in ongoing research projects, another 25% have presented at conferences, and 12.5% have published research papers in academic journals. The complete distribution for the respondents' answers are shown in Figures 17.

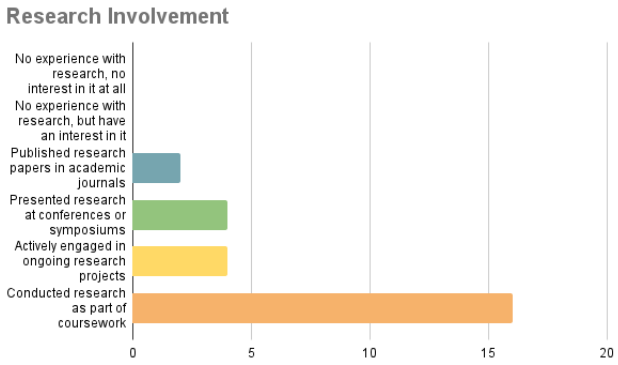


Fig. 17: Research Involvement

The most common problem encountered in research by respondents include time constraints and deadlines (93.8%), as shown in Figure 18. This is then followed by finding relevant research articles or literature, understanding concepts, methodologies, or analyses, citation management and formatting, and limited access to scholarly articles or databases (81.3% for each). The identified problems emphasize common challenges faced by researchers which highlights the importance of tools and technologies that can address these issues effectively, such as chatbots designed to assist with literature search and information extraction.

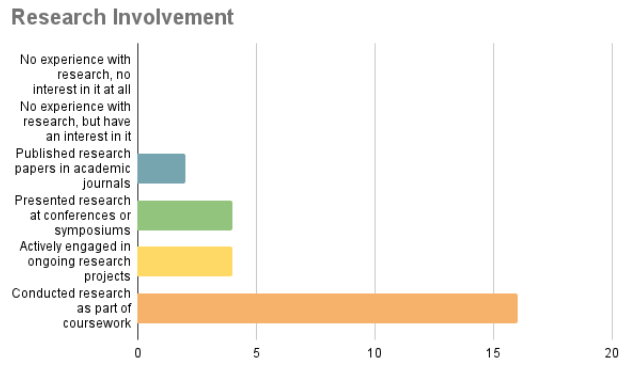


Fig. 18: Problems encountered in research process

*3) **End-to-End User Experience and User Impression**:* All (100%) of the respondents either strongly agree (50%) or agree (50%) that the chatbot provided relevant information from the publications. This indicates that users perceive the application as effective in extracting and presenting relevant information from the given dataset, contributing to a positive user experience and enhancing the application's utility for research purposes.

All (100%) of the respondents agree that the chatbot's responses were comprehensive and well-informed, with 56.25% strongly agreeing and 43.73% agreeing, indicating a high level of satisfaction with the depth and accuracy of the information provided by the application.

A significant proportion of respondents (87.5%) either agree (75%) or strongly agree (12.5%) that the chatbot effectively addressed their queries. However, a small percentage (12.5%) expressed neutrality, suggesting room for improvement in addressing user queries more effectively.

All (100%) respondents either strongly agree (56.25%) or agree (43.75%) that the format of the chatbot's responses was clear and easy to follow. This indicates that the app effectively presents information in a user-friendly manner.

43.75% of the respondents agree or strongly agree that they encountered inaccuracies or errors in the information provided by the chatbot. However, a notable proportion of respondents (56.25%) disagree, strongly disagree, or are neutral with this statement, indicating mixed perceptions regarding the accuracy of the app's responses.

Overall, the data suggests generally positive perceptions of the app's content, with the majority of respondents expressing satisfaction with the comprehensiveness, effectiveness, and clarity of the chatbot's responses. However, there are some concerns regarding the accuracy of the information provided, which may warrant further investigation and improvement.

The majority of respondents (87.5%) either strongly agree or agree that the chatbot provided timely responses to their queries while the remaining 12.5% expressed neutrality. This suggests that users generally perceive the app as responsive and efficient in addressing their information needs in a timely manner.

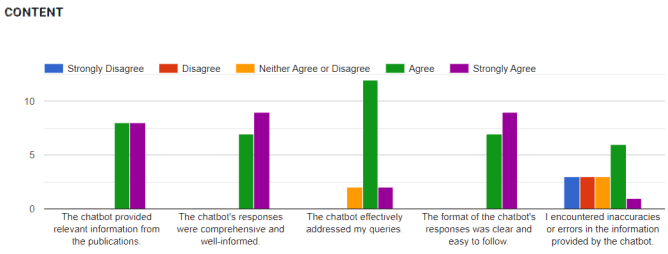


Fig. 19: Results for Content

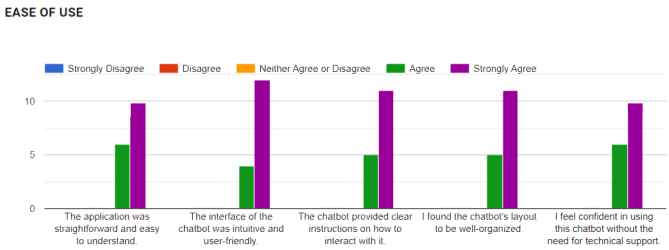


Fig. 21: Results for Ease of Use

The majority of respondents (75%) either strongly agree or agree that they did not experience significant delays in receiving information from the chatbot. However, 6.25% of the respondents expressed neutrality and a small percentage (18.75%) expressed disagreement or strong disagreement with this statement, indicating that some users may have encountered delays during their interactions with the app.

A majority of respondents (93.75%) either strongly agree or agree that the chatbot's response time was consistent throughout their interaction with the remaining 6.25% strongly disagreeing. This suggests that the app maintains a consistent level of responsiveness, contributing to a positive user experience and minimizing frustration due to unpredictable response times. However, due to latency issues with the deployment, it might have affected the consistency of the results for other respondents.

Overall, the data indicates generally positive perceptions of the app's timeliness, with the majority of respondents expressing satisfaction with the responsiveness and consistency of the chatbot's responses. However, there are some users who may have experienced delays, highlighting potential areas for improvement in optimizing the app's performance and responsiveness.

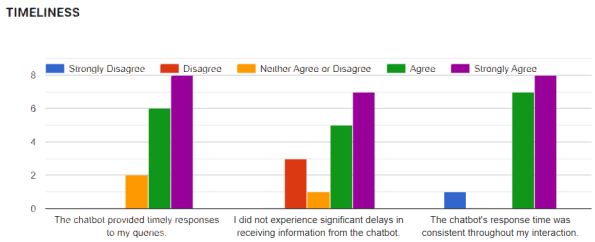


Fig. 20: Results for Timeliness

The data indicates a unanimous agreement among respondents regarding the ease of use of the application. For all aspects related to ease of use, such as the straightforwardness and clarity of the application, the intuitiveness of the interface, the provision of clear instructions, the organization of the layout, and the confidence in using the chatbot without technical support, all respondents either strongly agreed or agreed. This suggests that users found the application to be highly user-friendly, intuitive, and easily navigable, contributing to a positive user experience and indicating successful design and implementation of the user interface.

All of the respondents (100%) strongly agree or agree that the chatbot's efficiency in retrieving information demonstrates its effectiveness as a research assistant and that interacting with the chatbot provided valuable insights into its capabilities as a research assistant. This indicates that users perceive the chatbot as adept at being an assistant and that the respondents' positive experience is indicative of its potential as a research support tool.

Nearly 81.25% of respondents either strongly agree or agree that the chatbot's current features are diverse and meet the needs of a research assistant, while a small percentage (18.75%) remain neutral. This suggests that the majority of users perceive the chatbot's features as comprehensive and aligned with the requirements of research tasks.

The majority of respondents (93.75%) either strongly agree or agree that they would confidently utilize and recommend the application and 6.25% expressed neutrality, indicating a high level of confidence in the chatbot's functionality and suitability for research-related tasks.

Overall, the data reveals overwhelmingly positive perceptions and experiences regarding the functionality of the chatbot as a research assistant, with the majority of respondents expressing satisfaction with its efficiency, capabilities, features, and usability.

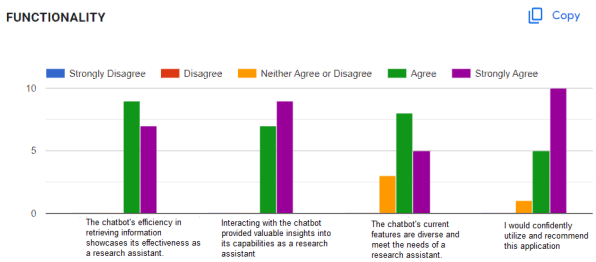


Fig. 22: Results for Functionality

V. SUMMARY AND CONCLUSION

The study aimed to address challenges faced by researchers in navigating academic literature by developing an AI-driven chatbot as a research support tool. It aimed to evaluate the effectiveness and usability of the chatbot built on a RAG framework, alongside OpenAI function calling and Langchain. Data collection methods included the use of the RAGAS library for an objective evaluation and a user usability test to assess its usability among its target users.

The RAGAS test results demonstrated the chatbot's effectiveness in generating answers to a range of queries, including simple, reasoning-heavy, and multi-context questions. Although the context relevancy metric scored lower due to the surplus of unused retrieved documents, the system achieved high scores in answer relevancy. This indicates that, despite some inefficiencies in document retrieval, the chatbot reliably provided accurate and relevant answers to user queries.

The usability test was conducted with 16 participants from the University of the Philippines Los Baños Institute of Computer Science, focusing on various aspects of the chatbot's performance, including content quality, timeliness, ease of use, and functionality. It further revealed overwhelmingly positive feedback on the chatbot's content, ease of use, and functionality, showcasing its value as a research assistant. Participants responded positively across several metrics: content quality, ease of use, and functionality. While some users noted inconsistencies in the metric of timeliness – noting occasional inconsistencies in response times, overall feedback was favorable. This positive reception is particularly noteworthy given that the testing was conducted on a limited-resource hosted site (0.1 CPU and 512MB).

Overall, these findings suggest that the chatbot addresses several gaps in literature, empirical evidence, and methodology by providing an effective, usable, and user-friendly tool for research assistance. The results highlight the potential of AI-driven chatbots in enhancing research processes by offering precise, relevant, and comprehensive literature support to individuals in the academe.

VI. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

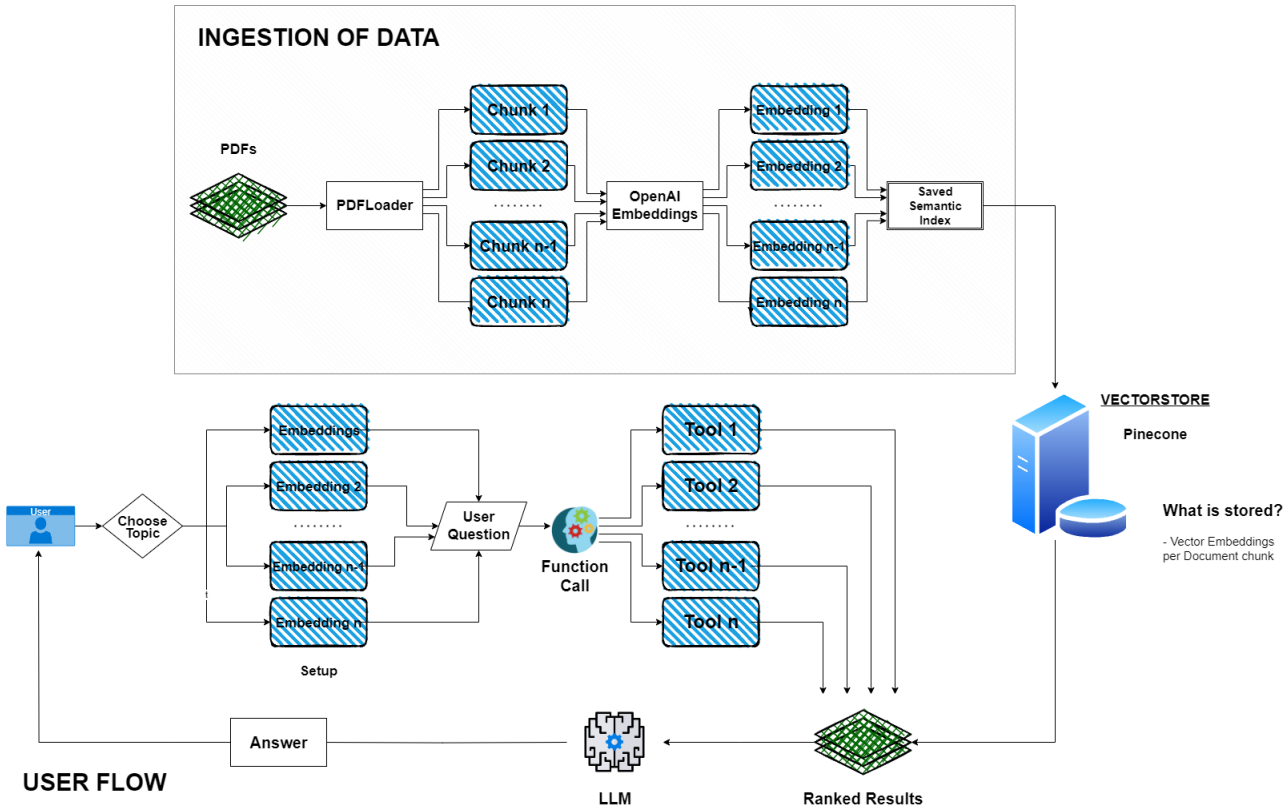
A. Limitations

Several limitations were encountered that may impact the scope and effectiveness of the implemented chatbot system. Firstly, the dataset used in this study is limited due to financial constraints in storing it in a vectorstore. This limitation restricts the diversity and comprehensiveness of the available research papers, affecting the available information for its users. Secondly, the chatbot's conversation history is constrained to the last 10 messages, leading to a loss of context over time. This limitation could hinder the continuity of interactions because it requires a frequent chat resets by users to maintain coherence. Additionally, the number of user tests conducted during the evaluation phase was also limited. This constraint may have implications for the depth of insights gained into user experiences and preferences. Lastly, while the study employed the RAGAS evaluation method, there is a need to evaluate it against other methods. Despite being widely adopted, the lack of empirical evidence to validate RAGAS as the optimal evaluation method poses a limitation on the robustness of the approach of this study.

B. Recommendations

To address the limitations and enhance the functionality of the chatbot system, several recommendations are proposed. Firstly, it is recommended to enhance the chatbot's knowledge base by including a wider variety of research papers, as well as increasing the number of ingested papers in the vectorstore. This would enhance the system's capability in providing comprehensive and diverse responses. Secondly, improvements could be made to the chatbot's capabilities to support the processing of complex content types such as LaTeX equations, images, charts, code snippets, and tables. This would enable more versatile interactions and facilitate the retrieval of a wider array of information. Additionally, the functionality of the chatbot could be diversified by integrating additional tools and features, such as image retrieval and support for various reference citation styles (APA, MLA, Chicago, etc.). These would cater to a wider scope of user needs and preferences, further enhancing the utility of the chatbot system. Lastly, user experience could be improved by implementing enhancements to the user interface. This could include introducing a more expansive chat history feature that allows access to previous conversations. It could also provide options to adjust the language model's response temperature or response length, allowing users greater control and customization over their interactions.

APPENDIX I
APPLICATION FRAMEWORK



APPENDIX II
CHAINS

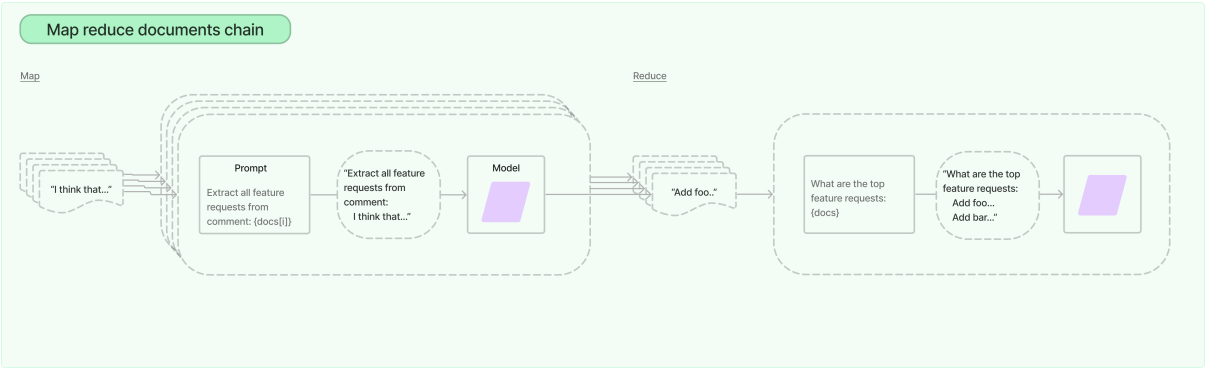


Fig. 23: Visual Representation of the Map Reduce Chain from Langchain [37]

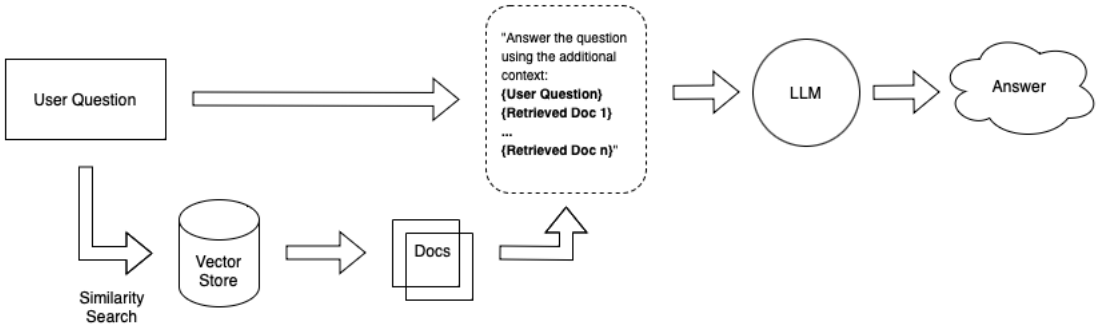


Fig. 24: Visual Representation of the Conversational Chain [37]

APPENDIX III
APPLICATION FEATURES

A. *Welcome Message*

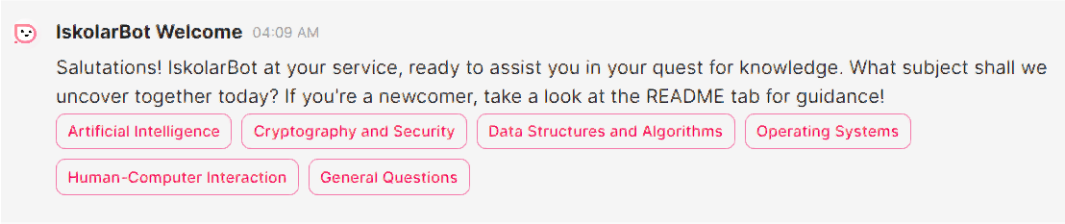


Fig. 25: Initial Welcome Message with Action to setup the datasets

B. *Summary*

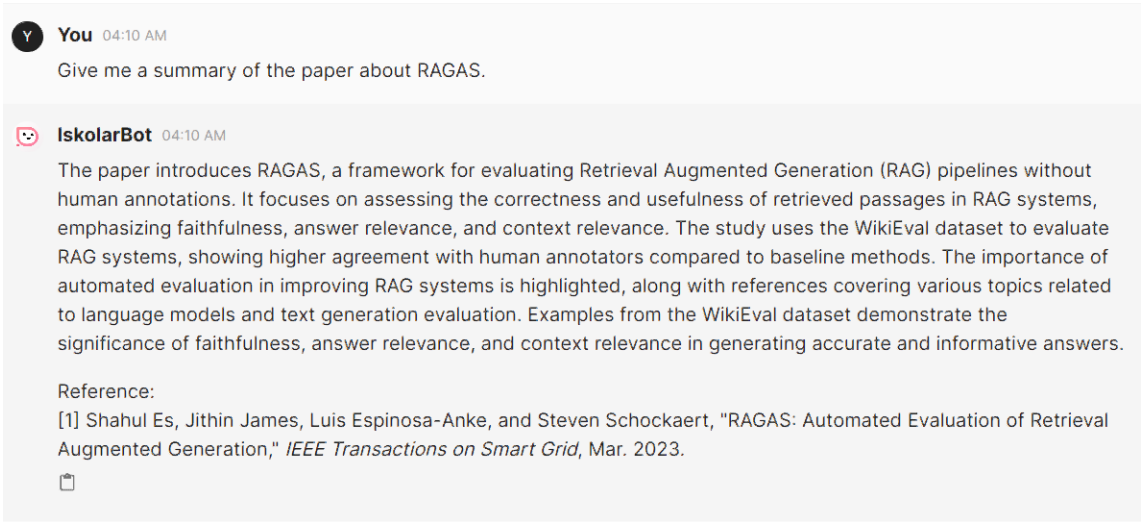


Fig. 26: Screenshot of a sample summary query

C. Related Literature

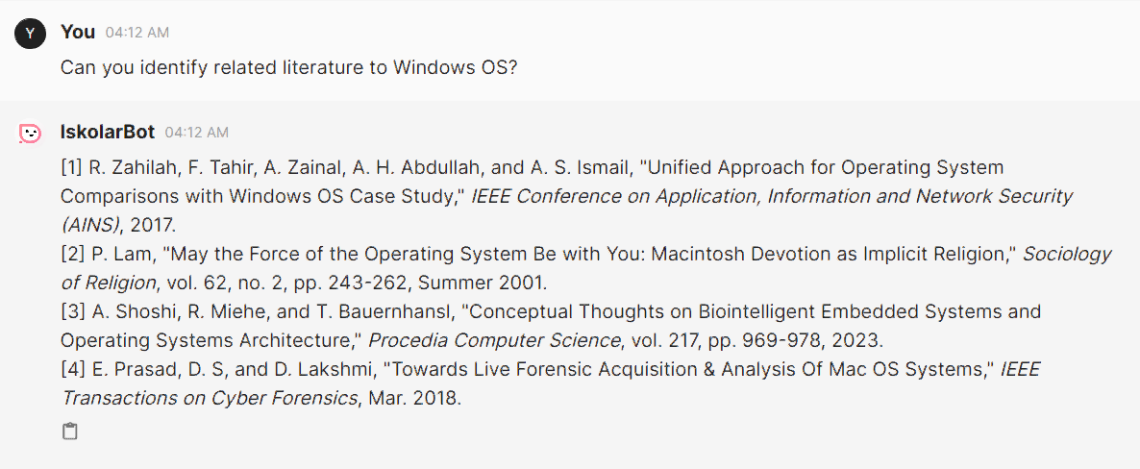


Fig. 27: Screenshot of a sample related literature query

D. Query

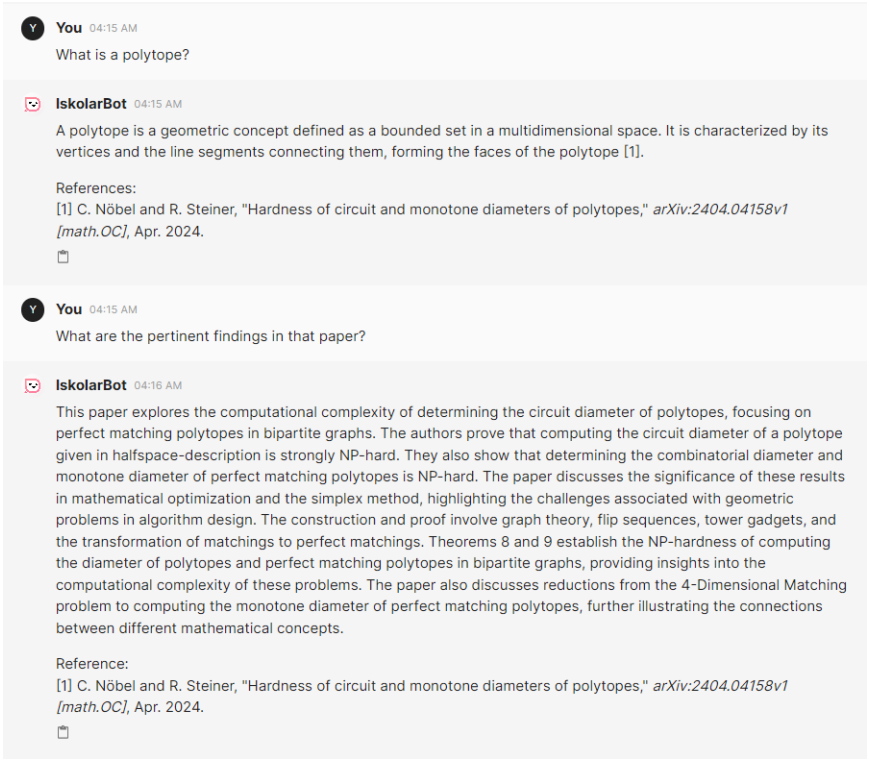


Fig. 28: Screenshot of a sample user query

APPENDIX IV
USER USABILITY TEST QUESTIONS

Experience with Technology

- Familiarity with chatbots or virtual assistants in general (ChatGPT, Customer Support Chatbots, FAQs Chatbot, etc.)
- Frequency of usage of chatbots or virtual assistants in general (ChatGPT, Customer Support Chatbots, FAQs Chatbot, etc.)
- Familiarity with academic databases or literature search engines (OpenAthens, Google Scholar, JSTOR, etc.)
- Frequency of usage of academic databases or literature search engines (OpenAthens, Google Scholar, JSTOR, etc.)
- Have you used chatbots specifically designed to assist with extracting information from databases of research papers?
- Which chatbots have you used for assisting you in research?

Experience with Research

- Research Involvement
  - No experience with research, no interest in it at all
  - No experience with research, but have an interest in it
  - Conducted research as part of coursework
  - Presented research at conferences or symposiums
  - Published research papers in academic journals
  - Actively engaged in ongoing research projects
- For those with any level of experience with research, what are the problems that you typically encounter?
  - Difficulty in finding relevant research articles or literature on a specific topic
  - Challenges in understanding complex academic concepts, research methodologies, or statistical analyses
  - Needing assistance with citation management and formatting for research papers
  - Synthesizing information from multiple sources
  - Limited access to scholarly articles or databases
  - Technical difficulties with research tools or equipment
  - Limited funding or resources for conducting research
  - Seeking guidance on writing research proposals
  - Time constraints and deadlines
  - Publishing and disseminating research findings effectively
  - I have no problems when it comes to the whole research process

Content

- The chatbot provided relevant information from the publications.
- The chatbot's responses were comprehensive and well-informed.
- The chatbot effectively addressed my queries.
- The format of the chatbot's responses was clear and easy to follow.
- I encountered inaccuracies or errors in the information provided by the chatbot.

Timeliness

- The chatbot provided timely responses to my queries.
- I did not experience significant delays in receiving information from the chatbot.
- The chatbot's response time was consistent throughout my interaction.

Ease of Use

- I found the chatbot to be unnecessarily complex.
- The interface of the chatbot was intuitive and user-friendly.
- I think that I would need the support of a technical person to use this chatbot.
- I found the chatbot's layout to be well-organized.
- The chatbot provided clear instructions on how to interact with it.

Functionality

- The chatbot's efficiency in retrieving and presenting information showcased its effectiveness as a reliable research assistant, making it a valuable tool for academic inquiries.
- Interacting with the chatbot provided valuable insights into its capabilities and potential applications.
- The chatbot's diverse array of features enhanced my user experience, demonstrating its versatility and utility. I would confidently utilize and recommend this application

ACKNOWLEDGMENT

Many thanks to...

REFERENCES

[1] A. Xie, "The effect of reading in shaping undergraduates' tolerance to ambiguity," *International Journal of Language, Literature and Linguistics*, vol. 5, no. 2, pp. 51–56, Jun 2019. [Online]. Available: http://www.ijlll.org/index.php?m=content&c=index&a=show&catid=59&id=515

[2] L. Katan and C. A. Baarts, "Inquiry-based reading – towards a conception of reading as a research method," *SageJournals*, vol. 19, no. 1, pp. 58–75, 2018. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1474022218760261

[3] J. T. Liwag, "Does the philippines value scientific research?" 2020, accessed: Nov. 1, 2023. [Online]. Available: https://www.cnnphilippines.com/life/culture/2020/10/21/scientific-research-philippines-barriers.html

[4] D. C. Bueno, "Research and publish or hibernate: Analysis of the limiting factors towards scientific productivity among professorial lecturers in the philippines," *CC The Journal: A Multidisciplinary Research Review*, vol. 14, p. Page range, October 2019. [Online]. Available: https://www.academia.edu/40029215/Research_and_Publish_or_Hibernate_Analysis_of_the_Limiting_Factors_towards_Scientific_Productivity_among_Professorial_Lecturers_in_the_Philippines

[5] P. Suta, X. Lan, B. Wu, P. Mongkolnam, and J. H. Chan, "An overview of machine learning in chatbots," *International Journal of Mechanical Engineering and Robotics*, vol. 9, no. 4, pp. 502–510, April 2020. [Online]. Available: https://pdfs.semanticscholar.org/7038/fecc82293642d64563cd73b04298073dbac6.pdf

[6] B. A. Shawar and E. Atwell, "Chatbots: Are they really useful?" *Journal for Language Technology and Computational Linguistics*, vol. 22, no. 1, pp. 29–49, July 2007. [Online]. Available: https://www.semanticscholar.org/paper/Chatbots%3A-Are-they-Really-Useful-Shawar-Atwell/8d8284bfba7ebcb4e2575d864ec7c16ea6a168f0

[7] C. Kooli, "Chatbots in education and research: A critical examination of ethical implications and solutions," *Sustainability*, vol. 15, no. 7, p. 5614, 2023. [Online]. Available: https://www.mdpi.com/2071-1050/15/7/5614

[8] R. Agarwal and M. Wadhwa, "Review of state-of-the-art design techniques for chatbots," *SN Computer Science*, vol. 1, no. 246, July 2020. [Online]. Available: https://link.springer.com/article/10.1007/s42979-020-00255-3

[9] OpenAI, "Introducing chatgpt," 2022. [Online]. Available: https://openai.com/blog/chatgpt

[10] T. Alqahtani, H. Badreldin, M. Alrashed, A. Alshaya, S. Alghamdi, K. bin Saleh, S. Alowais, O. Alshaya, I. Rahman, M. Al Yami, and A. Albekairy, "The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research," *Research in Social and Administrative Pharmacy*, vol. 19, no. 8, pp. 1236–1242, August 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1551741123002802?casa_token=I5hiMfxgWKQAAAAA:k2tlWqIGGs2mgovK_wcoTj9ZVqjp19mRoGf9guxIVO_3TJfst3VjGQtO9iq7HayHLzGnbgRoB5M

[11] GitHub, "Github copilot," 2021. [Online]. Available: https://github.com/features/copilot

[12] O. Topsakal and T. C. Akinci, "Creating large language model applications utilizing langchain: A primer on developing llm apps fast," *International Conference on Applied Engineering and Natural Sciences*, vol. 1, no. 1, pp. 1050–1056, July 2023. [Online]. Available: https://www.researchgate.net/publication/372669736_Creating_Large_Language_Model_Applications_Utilizing_LangChain_A_Primer_on_Developing_LLM_Apps_Fast

[13] R. M. Guido and G. Mangali, "Research productivity as performance dynamics of pisa among southeast asian countries," *Asia Pacific Journal of Multidisciplinary Research*, vol. 8, no. 4, pp. 76–90, November 2020. [Online]. Available: https://ieeexplore.ieee.org/document/10109453

[14] L. Crisostomo, R. M. Guido, and M. Villanueva, "Business technology and innovation research among southeast asian countries: Examining philippine performance in technoinnovation research," *International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management*, pp. 1–6, December 2022. [Online]. Available: https://ieeexplore.ieee.org/document/10109275?denied=

[15] S. Wa Mbaleka, "Factors leading to limited faculty publications in philippine higher education institutions," *International Forum*, vol. 18, no. 2, pp. 121–141, November 2015. [Online]. Available: https://www.researchgate.net/publication/290395664_Factors_Leading_to_Limited_Faculty_Publications_in_Philippine_Higher_Education_Institutions

[16] J. Lobo, ""behind research excellence": Students' challenges and barriers in the completion of an undergraduate thesis in the case of a premier local college in pampanga, philippines," *International Journal of Disabilities Sports and Health Sciences*, vol. 6, no. 1, pp. 38–52, February 2023. [Online]. Available: https://dergipark.org.tr/en/pub/ijdshs/issue/75824/1230630

[17] U. I. of Statistics, "Country report: Philippines," 2015. [Online]. Available: https://uis.unesco.org/en/country/ph?theme=science-technology-and-innovation

[18] J. Lim, "The philippines' scientific research is lagging, and it's due to a lack of government support," 2023, accessed: Nov. 14, 2023. [Online]. Available: https://www.cnnphilippines.com/life/culture/2020/10/21/scientific-research-philippines-barriers.html

[19] E. R. Deyro, "How a team of filipino scientists developed a covid-19 test kit," March 2020, accessed: Nov. 14, 2023. [Online]. Available: https://www.cnnphilippines.com/life/culture/2020/3/13/covid-test-kit-scientists.html?fbclid

[20] A. Index, "Scientific index 2024: Rankings for scientists in philippines," 2024, accessed: Dec. 2, 2023. [Online]. Available: https://www.adscientificindex.com/?country_code=ph

[21] J. Cornelio, "The state of research in the philippines," November 2023, accessed: Nov. 23, 2023. [Online]. Available: https://www.rappler.com/voices/thought-leaders/opinion-state-of-research-philippines/

[22] SciVal, "Scival: Comprehensive research analytics solution," 2023. [Online]. Available: https://www.scival.com/home

[23] ChatPDF, "Chatpdf: Your pdf ai," 2023.

[24] Elicit, "Elicit: Ai solutions," 2023. [Online]. Available: https://elicit.com/

[25] A. Agrahari, H. Shaikh, A. Pal, A. L. Yadav, and A. Singhal, "A survey of various chatbot implementation techniques," *International Journal of Research in Engineering, Science and Management*, vol. 1, no. 12, December 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:212484172

[26] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827020300062

[27] B. Borah, D. Pathak, P. Sarmah, B. Som, and S. Nandi, *Survey of Textbased Chatbot in Perspective of Recent Technologies*, June 2019, pp. 84–96.

[28] A. Haque, "A brief analysis of "chatgpt" – a revolutionary tool designed by openai," *EAI Endorsed Transactions on AI and Robotics*, vol. 1, no. 1, March 2023. [Online]. Available: https://pdfs.semanticscholar.org/5e59/9b60cf4bbe22bb8254dc33a80dde0b56241c.pdf

[29] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," April 2023. [Online]. Available: https://arxiv.org/abs/2304.02020

[30] MindsDB, "Navigating the llm landscape: A comparative analysis of leading large language models," 2022 2023. [Online]. Available: https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-...language-models

[31] S. Abdelhamid, "Using chatbots as smart teaching assistants for first-year engineering students," July 2020. [Online]. Available: https://www.semanticscholar.org/paper/Using-Chatbots-as-Smart-Teaching-Assistants-for-Abdelhamid/2a0079fba5001991701c770979480fb854f3c358

[32] A. Folstad, T. Araujo, E. L.-C. Law, P. B. Brandtzaeg, S. Papadopoulos, L. Reis, M. Baez, G. Laban, P. McAllister, C. Ischen, R. Wald, F. Catania, R. M. von Wolff, S. Hobert, and E. Luger, "Future directions for chatbot research: an interdisciplinary research agenda," *Computing*, vol. 103, p. 2915–2942, October 2021. [Online]. Available: https://link.springer.com/article/10.1007/s00607-021-01016-7#citeas

[33] B. Klimova and P. M. I. Seraj, "The use of chatbots in university efl settings: Research trends and pedagogical implications," *Frontiers in Psychology*, vol. 14, March 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1131506/full

[34] B. Memarian and T. Doleck, "Chatgpt in education: Methods, potentials, and limitations," *Computers in Human Behavior: Artificial Humans*, vol. 1, no. 2, December 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949882123000221

[35] B. Lund and T. Wang, "Chatting about chatgpt: How may ai and gpt impact academia and libraries?" *Library Hi Tech News*,

January 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4333415
[36] A. Sreeram and P. J. Sai, "An effective query system using llms and langchain," *International Journal of Engineering Research and Technology (IJERT)*, vol. 12, no. 6, June 2023. [Online]. Available: https://www.ijert.org/an-effective-query-system-using-llms-and-langchain
[37] LangChain, "Langchain documentation - introduction," 2023. [Online]. Available: https://python.langchain.com/docs/get_started/introduction
[38] A. M. Rouhi, "Scientists defy dire conditions: Researchers in the philippines use various strategies to keep some good work going," *Chemical and Engineering News*, vol. 74, pp. 57–61, November 1996. [Online]. Available: https://api.semanticscholar.org/CorpusID:155243704
[39] K. I. Navarro and M. McKinnon, "Challenges of communicating science: perspectives from the philippines," *Environmental Science, Political Science, Sociology*, February 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211551375
[40] A. Calma, "Challenges in preparing academic staff for research training and supervision : The case of the philippines," *International Journal of Educational Management*, vol. 28, pp. 705–715, August 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:142518133
[41] L. M. Coloquit, "Philippine education in the modern world: A trench for global academic success or another year of educational failure?" *Journal of English Education and Linguistics*, January 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:253542522
[42] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models," April 2023. [Online]. Available: https://arxiv.org/abs/2304.03738
[43] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha, and N. Mavridis, "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," *Digital Medicine*, vol. 3, no. 81, June 2020. [Online]. Available: https://www.nature.com/articles/s41746-020-0288-5
[44] M. Sallam, "The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations," *Healthcare*, February 2023. [Online]. Available: https://www.medrxiv.org/content/10.1101/2023.02.19.23286155v1
[45] F. Kitamura, "Chatgpt is shaping the future of medical writing but still requires human judgment," *Radiological Society of North America*, February 2023. [Online]. Available: https://pubs.rsna.org/doi/abs/10.1148/radiol.230171?journalCode=radiology
[46] W. Hariri, "Unlocking the potential of chatgpt: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing," *ArXiv*, March 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257952074
[47] T. Lopez and M. Qamber, "The benefits and drawbacks of implementing chatbots in higher education," March 2022. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1673913/FULLTEXT01.pdf
[48] M. Salvagno, F. S. Taccone, and A. G. Gerli, "Can artificial intelligence help for scientific writing?" *Critical Care*, vol. 27, no. 75, February 2023. [Online]. Available: https://ccforum.biomedcentral.com/articles/10.1186/s13054-023-04380-2
[49] The vector database to build knowledgeable ai. Pinecone. [Online]. Available: https://docs.pinecone.io/home
[50] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," *IEEE Transactions on Smart Grid*, March 2023.
[51] Ragas documentation. Ragas. [Online]. Available: https://docs.ragas.io/en/stable/index.html
[52] S. W. Black, "Current practices for product usability testing in web and mobile applications," 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:53352875

Louise Gabrielle L. Talip Loys is a senior BS Computer Science student studying at the University of the Philippines Los Banos. She is a member of the Parliament: UPLB Debate Society and the Alliance of Computer Science Students UPLB. She has also been an member of the ICS competitive programming club, UPLB Eliens, and the ICS CTF club, CLK_TCK. Her adviser is Prof. Reginald Neil C. Recario. She has been one of the Top 50 CAS Students, the CAS Outstanding Student in the Sciences in 2021, and a EAA Undergraduate Research Grant awardee. She primarily focuses on research related to Artificial Intelligence, specifically in Generative AI, and she has experience in using technological stacks for web development.