QUBO: A Retrieval-Augmented Generation Powered Philippine Cultural History Chatbot

A Thesis

Presented to the Faculty of Computer Science Department

College of Computing and Information Technologies

National University

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science in Computer Science with

Specialization in Machine Learning

BY:

Brix Anthony Manzanero

Christian Bongao

Kris Brian Diaz

Miguel Raphael Layos

Renz Andrei Alis

Marizkays Jamison

Thesis Adviser

January 2026

ACKNOWLEDGEMENT

First and foremost, we offer our deepest praise and gratitude to God Almighty, for the wisdom, strength, and resilience granted to us throughout this journey, making this accomplishment possible. And we would like to extend our heartfelt appreciation to Coffee Pages Coffee Shop, which became our everyday workspace and second home throughout the thesis-writing process. The welcoming atmosphere, comforting coffee, and the kindness of their staff greatly contributed to our productivity and motivation during long hours of research and writing and would like to express our sincere appreciation to everyone who supported us in the realization of this research. This study stands as a testament to the collective encouragement and assistance we received. We extend our profound gratitude to our thesis advisor, Ms. Mariskayz Jamison. Her expert guidance, constructive feedback, and unwavering patience were fundamental in shaping this project from its inception to the final manuscript. We are truly grateful for her mentorship and dedication to our academic growth. We also wish to thank our thesis subject professor, Sir Rogelio Labanan, for his valuable insights and direction during the manuscript preparation process. His guidance was instrumental in refining the quality of this work. Our gratitude also goes to the faculty and staff of the College of Computing and Information Technologies (CCIT) at National University Manila for providing a conducive learning environment and for their continued administrative support.

ABSTRACT

THIS STUDY EVALUATES THE APPLICATION OF RETRIEVAL-AUGMENTED GENERATION (RAG) IN THE DEVELOPMENT OF A DOMAIN-SPECIFIC QUESTION-ANSWERING SYSTEM FOCUSED ON PHILIPPINE CULTURAL HISTORY. UTILIZING A CURATED CORPUS OF HISTORICAL DOCUMENTS AS AN EXTERNAL KNOWLEDGE BASE, THE RESEARCH ANALYZES VARIOUS RAG CONFIGURATIONS TO ASSESS THEIR IMPACT ON RETRIEVAL PRECISION AND GENERATION QUALITY. SYSTEM PERFORMANCE WAS ASSESSED USING A DOMAIN-SPECIFIC EVALUATION DATASET DERIVED FROM THE SOURCE CORPUS, FACILITATING A CONTROLLED COMPARISON OF ARCHITECTURAL CONFIGURATIONS. RESULTS INDICATE THAT CONDITIONING GENERATIVE OUTPUTS ON RETRIEVED CONTEXT SIGNIFICANTLY IMPROVES ALIGNMENT WITH EXTERNAL DATA AND MITIGATES UNGROUNDED MODEL HALLUCINATIONS. CONSEQUENTLY, THE SYSTEM PRODUCES EVIDENCE-ALIGNED RESPONSES SUITABLE FOR ACADEMIC AND EDUCATIONAL INTERACTIONS. THE FINDINGS UNDERSCORE THE UTILITY OF RAG AS A KNOWLEDGE-ENHANCEMENT STRATEGY FOR CONVERSATIONAL AI, PARTICULARLY WITHIN CULTURALLY SENSITIVE AND INFORMATION-INTENSIVE DOMAINS.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND OF THE STUDY

The Republic of the Philippines is an archipelagic country with a wide heritage of culture and confusing histories. Preserving and understanding this legacy is essential to national identity [1]. History is considered as an important curriculum in further education, helping to promote historical awareness, critical thinking, and social understanding [2]. Readings in Philippine History and The Life and Works of Rizal are fundamental classes that teach students how to analyze and evaluate historical events. Philippine cultural history is considered as evolving, influenced by current research and different local perspectives [3]. For such, lessons in this field need learning resources that can handle complexity and cultural differences.

In the Philippines, education is quickly adopting digital technology. According to recent study, Filipino college students commonly use generative AI technologies to assist them with academic tasks [4]. According to Instructure's 2025 State of Higher Education study, around 63% of Filipino students utilize AI chatbots for text production, 58% for translation, 55% for explaining complex ideas, and 52% for summarizing academic texts. These findings indicate an important impact pattern of popular adoption of AI tools in university learning settings, which shows how generative AI is currently affecting study habits and academic assistance for Filipino students [4] [5].

While general-purpose large language models (LLMs) provide accessible and efficient automatic features, their use in education has limits. These models are trained on huge, different datasets to produce responses based on simulated language patterns instead of educational or officially restricted information [6]. As a result, the degree to that their outputs are consistent with recognized academic sources or established instructional materials still needs further investigation [7]. This lack of formal consistency causes it to be difficult to effectively identify and analyze hallucinations, since the model generates material that seems reasonable and official yet can be inaccurate or rejected by academic sources [8][9]. This concern is especially important in disciplines like Philippine cultural history, where accuracy, situational basis, and dedication to recorded historical sources are important [10] [11].

To address issues about accuracy and source alignment, current approaches emphasize the integration of outside documents into big language models' response

generation processes. Instead of depending only on inside model representations, these methods aim to match produced outputs with clearly defined documents [12]. The language model can provide more effectively accurate, accessible and consistent replies by conditioning them on relevant sources. Such document-aligned generation is particularly important in academic areas, where attachment to reliable sources and clear knowledge foundation is essential for learning and moral conduct [11] [12].

A large amount of study addressing these limitations focuses on retrieval-augmented generation (RAG). RAG combines a retrieval method and a generative model by adding relevant documents from an external knowledge source into the prompt before generating responses. This strategy allows LLMs to refer to selected sources without training[13] [14]. RAG has been tested in several kinds of contexts as a method of attaching generated answers to domain-specific information [14]. However, the success of RAG systems is dependent on aspects such as document retrieval quality and evaluation approach [15]. Despite recent developments, there is still limited scientific proof of how alternative RAG configurations work when applied to locally constructed historical information, particularly in the context of Philippine cultural history.

This study explores the application of Retrieval-Augmented Generation for grounding large language model outputs in textual information regarding Philippine cultural history. The study's goal is to verify a RAG-based system for answering questions using selected historical materials as an external source of information. Different RAG implementation techniques and assessment methods will be implemented to identify response coherence and grounding. The findings will guide the creation of a prototype chatbot to help students connect with Philippine cultural history resources. Lastly, the goal of this research is to contribute to the study of AI-assisted learning in higher education.

1.2 STATEMENT OF THE PROBLEM

- While Large Language Models (LLMs) are increasingly integrated into higher education, their reliability in providing accurate information on Philippine cultural history remains unverified. General-purpose LLMs often generate responses based on broad pre-training data, which without access to institutionally curated primary sources, may lead to historical inaccuracies, colonial biases, or culturally misaligned information. Consequently, there is a

need to determine how these models perform when constrained to domain-specific historical documents compared to their baseline, ungrounded outputs

- Although Retrieval-Augmented Generation (RAG) has been introduced as a method for grounding LLM outputs in external knowledge bases, there is limited exploration of its application using proprietary Philippine cultural history texts. The extent to which RAG improves the relevance and contextual grounding of generated responses in this domain has not been clearly established.

- Various techniques exist for implementing Retrieval-Augmented Generation, including differences. However, there is insufficient investigation into how these variations affect response quality when applied to Philippine cultural history content.

- There is a lack of systematic evaluation comparing different RAG-based configurations using multiple assessment methods. Without such evaluation, it is difficult to identify which approaches are more suitable for supporting historically grounded question answering in an academic context.

1.3 OBJECTIVES OF THE STUDY

General objective

This study aims to investigate the application of Retrieval-Augmented Generation (RAG) in grounding large language model responses using curated Philippine cultural history texts through the development and evaluation of a document-grounded question-answering system.

Specific Objectives

- To assess the effectiveness of document-grounded generation in producing responses that are contextually aligned with institutionally curated Philippine cultural history sources.

- Develop a Retrieval-Augmented Generation system using curated textual documents on Philippine cultural history as an external knowledge source.

- Explore and compare different RAG implementations, such as Embedding, retrieval techniques, reranking, and Model selection to assess their effects on response relevance and grounding.

- Evaluate the performance of RAG-based configurations using multiple assessment methods such as Hit rate, Mean Reciprocal Rank, Bleu, Rouge, Bert Score and LLM-judge to determine which approaches demonstrate stronger alignment with the provided historical sources.

1.4 SIGNIFICANCE OF THE STUDY

The results from the development and evaluation of this study will benefit the following:

For Students, developing cultural identity and historical knowledge in college requires knowing Philippine history. This is made possible by QUBO, which provides correct, source-specific solutions that help students move from recognizing dates and into critically engaging with historical stories and the formation of national identities. It allows students to navigate complex interpret differences caused by current research and regional perspectives.

For Educators and Institutions, this study provides evidence to the historical function of Philippine educators as "cultural receivers". By automating solutions to repeated course questions, the technology opens up faculty time to focus on higher-order duties like facilitating engaging classroom discussions and giving the in-depth, human-centered feedback required for expanding courses and focusing localized knowledge.

For Future Researchers, As the usage of AI tools increases in education, this study provides the foundation for developing AI tools that reduce the danger of misinformation. Researchers can use these approaches to connect AI to specific knowledge bases, such as localized historical records or specialized cultural documents, ensuring that technology is a useful instrument for society engagement.

1.5 SCOPE AND DELIMITATIONS

1.5.1 Scope of the Study

This study focuses on the design and evaluation of a Retrieval-Augmented Generation (RAG) system for document-grounded academic question answering in the domain of Philippine cultural history. The primary scope of the research is the investigation of retrieval and generation configurations that enable large language models to produce responses grounded in curated Philippine historical texts.

The system operates on user-provided Philippine cultural history documents, which serve as the exclusive knowledge source for answer generation. The scope includes document preprocessing, text extraction from PDF files, vector indexing, and retrieval-based context selection to support grounded response generation.

A retrieval pipeline integrated with a Large Language Model (LLM) is employed to generate coherent and factually supported answers strictly based on the content of the uploaded documents. The study focuses on the core stages of the RAG pipeline, namely document ingestion, retrieval, and response generation.

System performance is evaluated using retrieval metrics such as Hit Rate and Mean Reciprocal Rank (MRR), as well as automatic quality metrics including BLEU, ROUGE, and BERTScore. These are complemented by an LLM-as-a-Judge evaluation and qualitative output analysis to assess answer faithfulness, correctness, and overall response quality.

To demonstrate and operationalize the proposed RAG architecture, a web-based chatbot application is developed as a prototype interface. The chatbot serves solely as a delivery mechanism for document upload and question answering and is not intended to function as a fully deployed educational platform.

The intended users of the system are college-level students who engage with Philippine cultural history materials for academic learning. The system is positioned as a supplementary learning tool designed to support document analysis and inquiry-based learning rather than replace formal instruction or scholarly interpretation.

1.5.2 Delimitations of the Study

This study is limited to the development of a proof-of-concept RAG system rather than a fully scalable, production-ready application. The accompanying chatbot application is strictly web-based and does not provide mobile platform support.

The proposed system is designed exclusively to answer questions based on a provided corpus of Philippine cultural history documents which only supports pdf-based documents, the system does not focus on other file formats such as docx or Json. Consequently, the dataset used for evaluation is strictly domain-specific and excludes global or non-Philippine historical sources. Furthermore, while the dataset includes certain terms in Filipino, the underlying Large Language Model (LLM) is optimized for English-language processing. Therefore, the study focuses primarily on English-language retrieval and generation, and the system's proficiency in handling complex syntax or nuanced context in Filipino or other regional dialects is outside the current scope. The study does not evaluate the system's capability for multi-hop reasoning the ability to synthesize disparate facts from multiple documents to answer complex causal questions or does it implement advanced guardrails against prompt injection attacks common in public-facing LLM applications.

The study is only limited to the preprocessing of textual data from the document and does not employ image recognition for documents containing images or scanned documents. Furthermore, the study does not support highly academic materials such as full research papers that contain dense technical terminology, citations, and intricate document structures.

The evaluation of the system is confined to technical performance, particularly retrieval effectiveness and the grounding of generated responses in source documents. The study does not investigate learning outcomes, pedagogical effectiveness, user satisfaction, or large-scale user behavior. The prototype chatbot application serves only as a delivery mechanism for document upload and question answering and is not evaluated in terms of usability, user experience, or deployment scalability.

1.6 DEFINITION OF TERMS

Artificial Intelligence (AI) - Refers to computer systems designed to perform tasks that normally require human intelligence, such as language understanding, reasoning, and information retrieval.

Large Language Model (LLM) - A type of artificial intelligence model trained on large volumes of text data to generate human-like language responses. In this study, LLMs are used as the generative component of the question-answering system.

Generative AI - A category of artificial intelligence capable of producing new content such as text, summaries, or explanations based on learned patterns from training data.

Hallucination- A phenomenon in which a language model generates information that appears coherent and authoritative but is factually incorrect, unsupported, or not grounded in reliable sources.

Retrieval-Augmented Generation (RAG) - An approach that combines information retrieval and text generation by incorporating relevant external documents into the prompt of a language model to produce grounded and source-aligned responses.

Document Grounding -The process of constraining generated responses so that they are explicitly based on retrieved documents rather than solely on the internal knowledge of the language model.

Embedding - A numerical representation of text that captures semantic meaning, allowing documents and queries to be compared mathematically for similarity during retrieval.

Vector Database - A storage system used to index and retrieve text embeddings efficiently based on semantic similarity rather than exact keyword matching.

Retrieval Pipeline - The sequence of processes involved in selecting relevant documents, including preprocessing, embedding, similarity search, and optional reranking.

Reranking - A retrieval refinement technique that reorders initially retrieved documents based on deeper relevance evaluation to improve the quality of context provided to the language model.

Hit Rate - An evaluation metric that measures whether at least one relevant document is retrieved within the top-k results.

Mean Reciprocal Rank (MRR) - A retrieval evaluation metric that assesses how high the first relevant document appears in the ranked retrieval results.

BLEU (Bilingual Evaluation Understudy) - A metric used to evaluate generated text by comparing it to reference answers based on n-gram overlap.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) - An evaluation metric that measures overlap between generated responses and reference texts, commonly used for summarization and text generation tasks.

BERTScore - A semantic evaluation metric that measures similarity between generated text and reference answers using contextual embeddings from transformer models.

LLM-Judge - An evaluation approach where a language model is used to assess the relevance, coherence, and grounding quality of generated responses against source documents.

CHAPTER 2
REVIEW OF RELATED LITERATURE AND STUDIES

2.1 RELATED LITERATURES AND STUDIES

2.1.1 Preservation and Accessibility of Philippine Cultural Heritage

The Filipino culture's collective historical, literary, and social forms constitute the Philippine cultural legacy [16]

. The Philippine government considered that preserving these materials was essential in maintaining national identity and ensuring that historical information was available for public awareness, education, and academic work [17]. Eventually, programs to protect Philippine cultural heritage developed. Institutions like the National Commission for Arts and Culture (NCCA), the National Historical Commission for the Philippines (NHCP), and lots of college libraries provide significant functions in the collection, arrangement, and preservation of historical materials. Expert commentary that clarifies context, authorship, and relevance is typically associated with collected resources, such as the Philippine History Source Book, which provides structured access to primary and secondary historical texts [18], [19].

Digital technology is considered an essential method to provide cultural resources easier to access while keeping damaged copies [20]. National archives and libraries have turned unique documents, historical texts, and other materials into digital format, making these treasures available to researchers and the general public online. Platforms such as the Filipino Heritage Library, ASEAN Information Library, and the Philippine eBook Hub show this strategy by offering organized, accessible collections for academic and educational uses [21] [22]. With these improvements, issues remain. Some historical materials are incomplete, not correctly categorized, or at risk of environmental damage, and preservation resources and staff can be limited. More initiatives are required to improve the management, recording, and digital accessibility of cultural heritage, so that the Philippines' valuable historical legacy is available to future generations [23].

2.1.2 The Role of AI in Student Learning

Artificial intelligence (AI) is transforming education through improving way students learn, and technology offers the potential to address some of the biggest problems in education today [24], [25]. Higher education institutions are increasingly adopting AI-driven tools to stay up with innovation and prepare students for a fast-changing world [26]. Within the core of this change are AI areas like machine learning (ML) and natural language processing (NLP). With an increase of LLMs, students are using AI to help them understand teachings at school[27].

A personalized strategy helps reduce learning gaps, better engagement, and improve academic achievements [28], [29]. Besides, artificially intelligent technology help instructors by automated manual tasks like grading & attendance, however it also allows those to focus on higher-value instructional activities while ensuring students receive consistent and meaningful feedback [30],[31]. Artificial intelligence enables data-driven understanding into student performance, allowing educators to more effectively monitor progress and adapt teaching tactics in real time. This is especially useful in scenarios with large or different classrooms, where individual monitoring is challenging [29], [24].

In the Philippine school system, AI integration brings all potential and challenges. Local academics highlight the potential of artificial intelligence (AI) to improve individualized learning, support teachers in creating effective training, and provide scalable educational solutions. However, issues like privacy, data security, and the potential of increasing the digital divide must be addressed to ensure equal access and ethical implementation, which has resulted in the development of new ethical regulations [26], [32], [33].

2.1.3 Large Language Models (LLMs)

Large Language Models (LLMs) constitute a specialized class of Artificial Intelligence designed to process, comprehend, and generate human-like text through Natural Language Processing (NLP). As a subset of Generative AI which encompasses the creation of novel data across text, image, and audio domains LLMs focus specifically on language-oriented tasks such as translation, question answering, and content generation [33], [34].

A defining feature of these architectures is their ability to maintain contextual understanding over long sequences. By incorporating advanced memory units and

attention mechanisms, LLMs can store and retrieve relevant information to produce responses that are not only coherent but also contextually accurate [33], [34], [35]. The development of these models, particularly following the advancements in deep learning architectures in 2020, has significantly influenced educational technology.

Contemporary literature identifies several foundational architectures that drive these capabilities:

- Generative Pre-trained Transformers (GPT) Series: Developed by OpenAI, the GPT series (including GPT-3 and GPT-4) represents the most prominent class of decoder-only models. They are renowned for their ability to generate coherent, contextually appropriate text and are widely utilized in applications ranging from translation to conversational agents. GPT-4, in particular, aims to push the boundaries of reasoning and generation quality [36].

- Bidirectional Encoder Representations from Transformers (BERT): Unlike models that read text sequentially (left-to-right), BERT introduced a breakthrough in understanding language bidirectionally. This encoder-only architecture has led to significant improvements in discriminative tasks, such as sentiment analysis, named entity recognition, and text classification [37].

- Text-To-Text Transfer Transformer (T5): The T5 model proposes a unified framework where every linguistic task is treated as a text-to-text transformation. Unlike BERT (encoder-only) or GPT (decoder-only), T5 utilizes a full encoder-decoder architecture, allowing it to handle a diverse range of tasks using a consistent input-output format [38]. [39].

2.1.4 Limitations of Large Language Models

Large Language Models (LLMs) have different fundamental challenges which affect their reliability in knowledge-intensive and academic question-answering jobs. An important downside is hallucination, in which models provide proficient but objectively wrong or generated information. This happens because LLMs are trained to guess the most likely next token based on statistical patterns rather than comparing facts to an external ground truth. According to research, hallucination is an inherent feature of LLMs that cannot be completely avoided by model scaling alone [40], [41].

Another important problem is the absence of clear foundation and source attribution. LLMs automatically retain knowledge within their parameters and do not

automatically retrieve or reference external documents during inference. Since an outcome, they're incapable of properly identify sources or support correct statements, that's essential in academic and historical research. This limitation is especially challenging in subjects like cultural history, where accuracy of fact and access to primary sources are critical [40], [41], [42].

Static training data is another limitation of LLMs. They cannot absorb currently presented or domain-specific content without retraining or fine-tuning because their understanding is restricted to the data available at the time of training. Because Philippine cultural history may be disadvantaged in general-purpose training resources, this limits their effectiveness in specialized or limited fields [40], [42, [43]. [44].

Other challenges are data bias and domain coverage. Responses from general-purpose LLMs may be insufficient, broad, or culturally incorrect because to their lack of exposure to region-specific historical narratives. Transparency and true reliability may be impacted by biases in the training data [44], [45].

2.1.5 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has been proposed as a solution to key limitations of Large Language Models (LLMs), particularly their tendency to generate factually incorrect or fabricated information, commonly referred to as hallucinations. Because LLMs are probabilistic models that generate text based on learned patterns rather than verified knowledge, they may produce fluent but inaccurate responses when relevant information is missing or outdated [15], [46], [48]. Its basic process in seen in Figure 1. RAG addresses this issue by integrating an external knowledge base into the generation process, enabling models to retrieve relevant documents before producing a response. This hybrid approach allows outputs to be grounded in specific, up-to-date sources without requiring costly model retraining [46], [49].

Figure 1. Basic LLM Process

Seen in Figure 2. The RAG framework typically consists of three stages: indexing, retrieval, and generation. During indexing, documents are preprocessed, segmented into meaningful chunks, and transformed into vector representations using embedding models such as BERT or Sentence-BERT, which are then stored in a vector database [50], [51]. When a user submits a query, the retrieval stage identifies the most relevant document chunks using similarity measures such as cosine similarity [51]. These retrieved texts are then incorporated into the generation stage, where the LLM produces responses that balance linguistic fluency with factual grounding [10], [44]. This architecture is particularly beneficial in educational settings, where responses must remain accurate, verifiable, and aligned with course-specific materials [47].

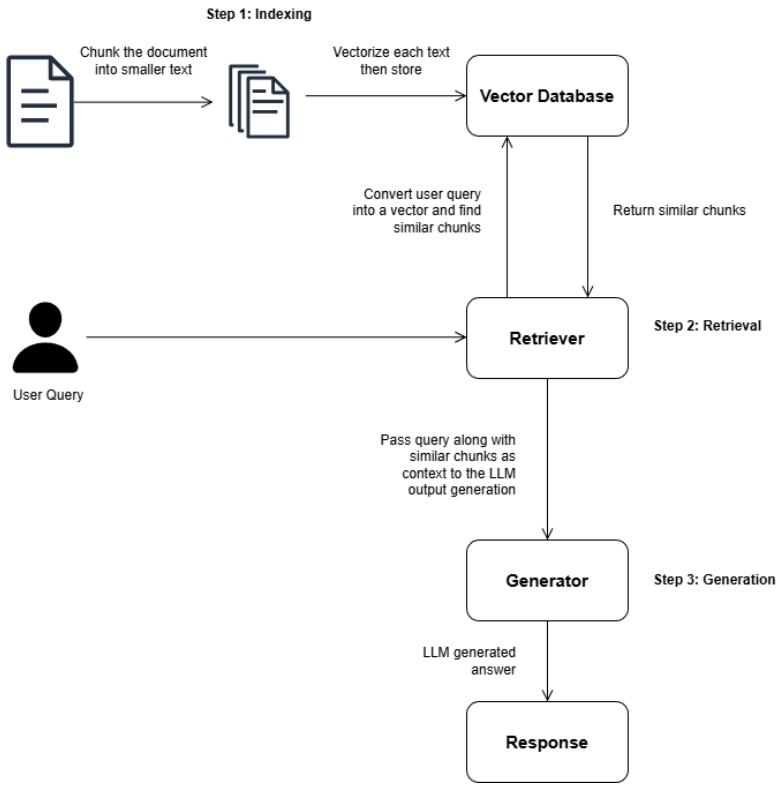


Figure 2. Basic Retrieval-Augmented Generation Process

Recent advances in RAG research focus on improving retrieval accuracy, contextual reasoning, and system reliability. Techniques such as multi-stage retrieval, semantic chunking, and adaptive context selection have been introduced to better capture user intent and highlight the most relevant information [52], [53]. Additional research emphasizes retrieval diversity, reasoning across multiple documents, and privacy-preserving mechanisms to support domain-specific and institution-controlled

knowledge bases [41], [49]. These developments have strengthened the applicability of RAG for academic and educational chatbot systems.

Despite its advantages, RAG presents several challenges. System performance is highly dependent on the quality and relevance of retrieved documents; incomplete or noisy sources can negatively affect response quality [51], [52]. Although retrieval grounding reduces hallucinations, it does not fully eliminate them, as models may still misinterpret or incorrectly synthesize retrieved content [48], [49]. RAG systems also introduce higher computational complexity due to additional processing stages such as embedding, indexing, and similarity search, increasing latency and resource requirements [40] [50], [54], [55]. Furthermore, evaluation remains difficult due to the lack of standardized metrics that clearly separate retrieval errors from generation errors, particularly in low-resource or culturally specific domains [56], [57], [58]. These limitations highlight the need for careful system design and evaluation, especially when applying RAG to culturally grounded subjects such as Philippine cultural history.

2.1.9 Document Content Extraction

The effective transformation of unstructured documents into machine-readable formats is a prerequisite for any Retrieval-Augmented Generation (RAG) pipeline. Document Layout Analysis is the process of identifying and categorizing regions of interest within a document image, such as text blocks, tables, titles, and figures. Unlike simple text extraction, which reads a file as a continuous stream of characters, DLA preserves the spatial and logical structure of the content [59].

This preservation of structure is critical for academic and historical texts, where multi-column layouts, sidebars, and footnotes carry significant semantic weight. Advanced parsing libraries utilize bounding box coordinates to map text to its physical location on the page, ensuring that the reading order is maintained and that data from tables is not serialized incorrectly.

2.1.10 Chunking

Information retrieval systems fundamentally rely on the ability to process and retrieve precise segments of information. Since Large Language Models (LLMs) and embedding models have strict token limits, entire documents cannot be processed in a single pass. To address this, texts must be divided into smaller, self-contained units known as "chunks." This process, referred to as chunking, is a critical preprocessing step that directly influences retrieval accuracy. If a chunk is too small, it may lack the

necessary context to be useful; if it is too large, it may dilute the specific information needed to answer a user's query, leading to retrieval noise [60], [61].

The most traditional approach to this segmentation is Fixed-Size Chunking. In this method, the document is split based on a predetermined count of characters, words, or tokens, for instance, creating a new chunk every 500 words. While this approach is computationally inexpensive and simple to implement, it relies on arbitrary boundaries that ignore the semantic flow of the text. A single sentence, a definition, or a coherent idea might be split across two separate chunks, leading to "context fragmentation." When the retrieval system later attempts to find an answer, it may retrieve only half of the necessary information, resulting in incomplete or hallucinated responses [62].

To mitigate the limitations of arbitrary splitting, Semantic Chunking has emerged as a more sophisticated strategy. Unlike fixed-size methods, semantic chunking divides text based on meaning rather than length. This technique utilizes embedding models to analyze the relationship between consecutive sentences by calculating the cosine similarity of their vector representations. When the system detects a significant drop in similarity between two sentences indicating a shift in topic or context, it creates a "breakpoint." This ensures that each chunk represents a coherent thought or topic, significantly improving the quality of the information provided to the LLM during generation [63]. [64]. [65].

Building upon these concepts, Hybrid Chunking seen in figure 3, represents an integration of structural constraints and semantic awareness. This approach acknowledges that while semantic coherence is ideal, technical constraints (such as token limits) must still be respected. In a hybrid system, documents are often first constrained by a maximum token limit to fit the model architecture, and then semantically refined to ensure logical coherence. Furthermore, to prevent information loss at the edges of these segments, hybrid systems frequently employ a "sliding window" technique. By overlapping the end of one chunk with the beginning of the next, the system preserves the context for sentences that might otherwise be severed, ensuring a more robust and continuous retrieval process [64], [65].

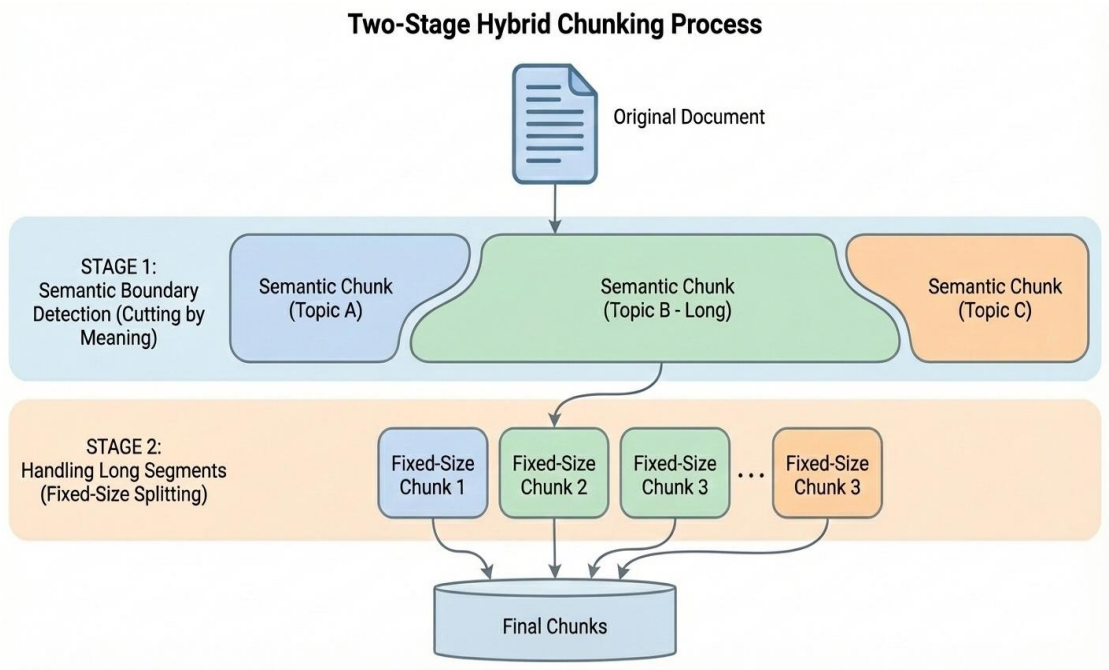


Figure 3. Two Stage Hybrid Chunking Process

2.1.11 Text Embedding

In the context of Natural Language Processing (NLP), an embedding is a numerical representation of text mapped into a high-dimensional vector space. Unlike traditional string-processing methods that rely on exact keyword matching, embeddings encode the *semantic meaning* of the text. This allows the system to understand that words like "canine" and "dog" are related, even if they share no common letters [66].

The theoretical basis of embedding lies in Distributional Semantics, which posits that linguistic items with similar distributions have similar meanings. Modern transformer-based models operate this by converting variable-length text (sentences or paragraphs) into fixed-length vectors of real numbers[67], [68].

The process of generating these embeddings generally follows a three-stage theoretical workflow:

1. Input Preparation: Text chunks are tokenized and fed into the encoder model.
2. Vector Transformation: The model processes the tokens through multiple layers of neural networks (typically Self-Attention mechanisms). These layers

identify latent semantic features, such as syntax, context, and entity relationships.

3. Semantic Space Mapping: The final output is a dense vector. In this high-dimensional space, the geometric distance between two vectors corresponds to their semantic similarity.

This geometric property allows for Semantic Search. By calculating the distance (or angle) between a query vector and document vectors, a retrieval system can identify passages that are conceptually relevant to the user's intent, even if the specific phrasing differs. This capability is the foundational mechanism that enables Retrieval-Augmented Generation (RAG) systems to move beyond simple keyword lookups and provide context-aware answers. A visualization is seen in Figure 4 [68].

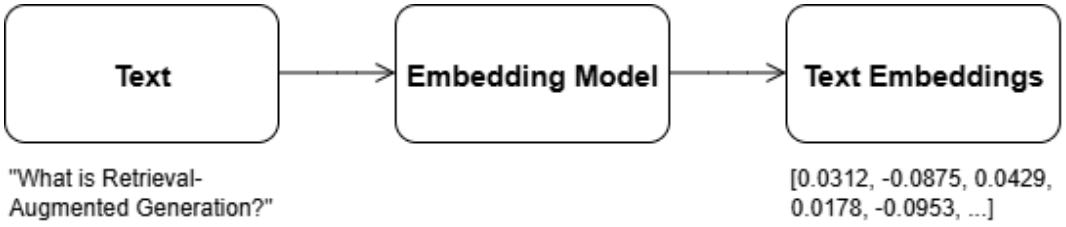


Figure 4. Text Embedding

2.1.12 Information Retrieval

Information Retrieval (IR) is the study of organizing, storing, and accessing information from sizable collections of structured or unstructured data [69]. While conventionally used to power search engines and digital libraries , modern IR has evolved into the core mechanism of Retrieval-Augmented Generation (RAG) systems. In the context of RAG, Retrieval is responsible for sourcing relevant information from a knowledge base to ground the generation of Large Language Models (LLMs) [69], [70]. The efficacy of the generation phase relies heavily on the quality and relevance of the retrieved documents. Information retrieval strategies are generally categorized into two primary approaches: Sparse Retrieval (keyword-based) and Dense Retrieval (semantic-based) [71], [72], [73].

Sparse Retrieval represents the traditional approach to information retrieval. In this model, both documents and queries are represented as sparse, high-dimensional

vectors. These vectors are termed "sparse" because most dimensions are zero, as they correspond to words in vocabulary that do not appear in the specific text segment [71], [72].

To implement this efficiently, systems utilize an Inverted Index, a data structure that maps every unique word in the corpus to a list of documents containing that word. This architecture allows the system to rapidly identify candidate documents without scanning the entire collection [72].

However, simple keyword matching is often insufficient because common words (e.g., the, is) appear frequently but carry little semantic weight. To address this, the BM25 (Best Matching 25) ranking function is widely adopted. BM25 is a probabilistic retrieval framework that improves upon simple Term Frequency (TF) by introducing two critical normalization factors. The Inverse Document Frequency (IDF) which Penalizes words that appear in many documents, as they likely contain low information content and Document Length Normalization which Prevents long documents from unfairly dominating the results simply because they contain a higher volume of words [71].

The standard BM25 scoring function for a document D and query Q containing terms q_i is expressed as:

$$\text{Score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

Equation 1. BM25

In this formulation $f(qi,D)$denotes the frequency of the term q_i within document D, while $|D|$ represents the total length of that document in words. To account for variations across the corpus, $avgdl$ signifies the average document length of the entire collection. The formula also utilizes free parameters k_1 and b to fine-tune the retrieval sensitivity; typically, k_1 is set between 1.2 and 2.0 to control term saturation, and b Is set to 0.75 to control the degree of length normalization.

The Inverse Document Frequency (IDF) component is calculated to weigh the importance of specific terms:

$$\mathrm{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

Equation 2. IDF

Where N represents the total number of documents in the corpus, and $n(q_i)$ indicates the number of documents that contain the specific query term q_i.

Dense Retrieval addresses the limitation of keyword matching by mapping queries and documents into a shared, continuous vector space using deep learning models, such as BERT. In this high-dimensional space, texts with similar meanings are located near each other, even if they do not share identical words [68].

To quantify the "closeness" or semantic relevance of two vectors in this space, Cosine Similarity is the standard metric. Unlike Euclidean distance, which measures the straight-line distance between points, Cosine Similarity measures the cosine of the angle between two vectors. This metric focuses on the orientation (semantic content) of the vector rather than its magnitude (length), making it robust to variations in document length [68].

The Cosine Similarity between two vectors A (query) and B (document) is defined as:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Equation 3. Cosine Similarity

Where, $A \cdot B$ represents the dot product of the query and document vectors, while $|A|$ and $|B|$ denote their respective Euclidean magnitudes (or norms). The resulting score ranges from -1 to 1, where a score of 1 indicates that the vectors are perfectly aligned in meaning, and a score of 0 suggests orthogonality or unrelated content. A visualization of this can be seen in figure 5.

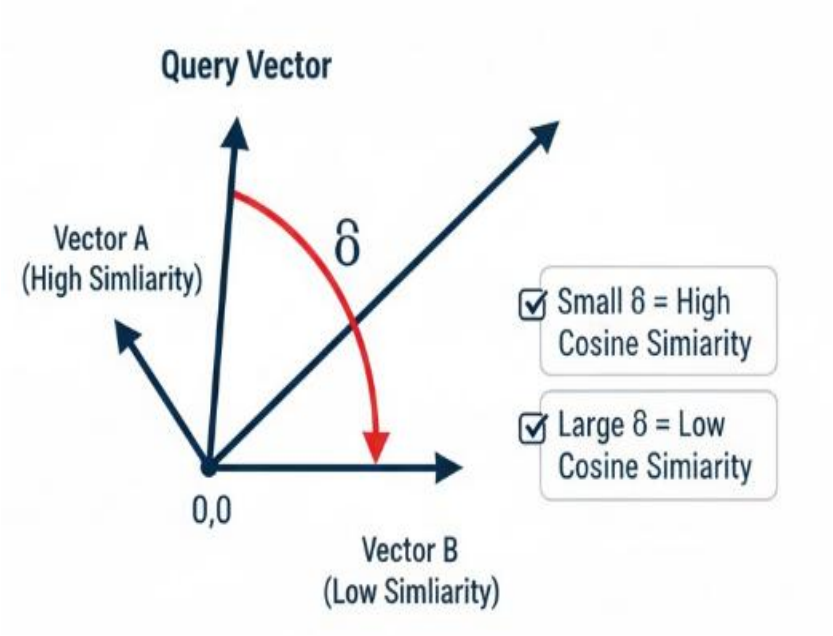


Figure 5. Cosine Similarity

Hybrid Retrieval is an advanced approach that combines sparse and dense retrieval methods to leverage their complementary strengths. While sparse retrieval excels at precise keyword matching, dense retrieval captures semantic nuance. It matches the query both literally and semantically that are ranked highest [73], [74].

Commonly, this is implemented via a Weighted Fusion approach, where relevance scores from both systems are normalized and combined [74]. The fused score for a document d is typically expressed as:

$$\text{Score}_{\text{hybrid}}(d) = \alpha \cdot S_{\text{dense}}(d) + (1 - \alpha) \cdot S_{\text{BM25}}(d)$$

Equation 4. Fused Score

Where $S_{\text{dense}}(d)$ is the semantic similarity score (e.g., Cosine Similarity), $S_{\text{BM25}}(d)$ is the keyword relevance score, and α is a weighting factor that controls the relative contribution of dense versus sparse scores.

In The final stage in high-performance retrieval pipelines is Reranking. While initial retrieval using Bi-Encoders or BM25 is fast, it trades some precision for speed by independently encoding queries and documents. To refine the results, a Cross-Encoder model is applied to the top retrieved candidates [75]. Unlike Bi-Encoders, which process the query and document separately, a Cross-Encoder jointly processes the query-document pair, allowing the model to capture rich, token-level interactions and dependencies. This results in significantly higher retrieval precision Prominent

models in this domain include the BAAI/bge-reranker series, which outputs a raw relevance score (logit) [75], [76].

To make these scores interpretable, a sigmoid transformation is often applied to map the output to a probabilistic range of $[0 - 1]$:

$$S(x) = \frac{1}{1 + e^{-x}}$$

Equation 5. Sigmoid Transformation

By reranking the top candidates (e.g., top 10) using this computationally intensive but highly accurate method, RAG systems ensure that the context provided to the generation module is of the highest possible relevance [75].

2.1.13 Prompt Engineering in Large Language Model

Prompt engineering has emerged as a critical discipline within the field of natural language processing (NLP), defined generally as the systematic process of structuring text prompts to interpret and generate specific tasks by a Large Language Model (LLM) without updating the model's weights. As LLMs such as GPT-4 and Llama 2 have become more capable, the ability to guide their outputs through "in-context learning" has shifted the paradigm from traditional fine-tuning to prompt-based interactions [46], [76].

At its core, prompt engineering leverages the probabilistic nature of transformer architectures. By optimizing the input context, developers can significantly influence the accuracy, tone, and format of the model's response [77], [78]. categorizes prompt engineering not merely as an art form but as a distinct pattern language that allows users to solve complex problems by decomposing them into modular instructions. Current literature identifies several foundational strategies used to enhance model performance.

Zero-Shot and Few-Shot Prompting: demonstrated that LLMs act as few-shot learners. In "zero-shot" scenarios, the model is given a task description without examples. In "few-shot" prompting, the model is provided with a small set of input-output demonstrations in the context window, which significantly improves performance on specific tasks by allowing the model to recognize patterns and adapt its output style accordingly [78].

Chain-of-Thought (CoT) Prompting: To address the limitations of standard prompting in complex reasoning tasks, introduced Chain-of-Thought prompting. This technique encourages the model to generate intermediate reasoning steps before arriving at a final answer. Research indicates that CoT significantly improves performance on arithmetic, commonsense, and symbolic reasoning tasks by mitigating the "black box" nature of immediate generation [79].

Instruction Tuning: highlighted the importance of aligning models with user intent. While raw language models predict the next token based on training data, instruction-tuned models are optimized to follow specific directives. Prompt engineering in this context focuses on formulating clear, unambiguous constraints to reduce hallucinations and ensure safety [80].

These diversity of studies on prompt engineering highlights the importance of creating meaningful instructions for the large language to follow, which can heavily alter the final output since modern llms can follow instructions.

2.1.14 Evaluating Retrieval-Augmented Generation Systems

The evaluation of Retrieval-Augmented Generation (RAG) systems presents a unique challenge compared to traditional Natural Language Processing (NLP) tasks. While standard metrics such as BLEU and ROUGE are effective for measuring lexical overlaps in translation or summarization, they are insufficient for RAG because they fail to capture the factual accuracy and reasoning capabilities of the model. Consequently, recent literature has shifted towards frameworks that assess the system's performance on two distinct fronts: the quality of the retrieval and quality of the generation [81], [82], [83].

In the retrieval of documents, Hit Rate (Hit@k) measures the system's probability of success [84]. It calculates the fraction of queries for which at least one relevant document appears in the top-k retrieved results. Unlike Recall, which measures total coverage, Hit Rate focuses on the binary presence of relevant information, reflecting the practical reality that an LLM often requires only a single accurate source to synthesize a correct answer. It is mathematically defined as:

$$\text{Hit Rate@}k = \frac{1}{|Q|}\sum_{i=1}^{|Q|} I\left(rel_i \in Top_k(q_i)\right)$$

Equation 6. Hit Rate@K

Where $|Q|$ is the total number of queries, and I is an indicator function that equals 1 if a relevant document rel exists in the top-k results for query q_i ,and 0 otherwise.

Complementing this is Mean Reciprocal Rank (MRR), which evaluates the ranking quality. While Hit Rate checks if the correct document exists, MRR penalizes the system if that document is buried low in the results list. It calculates the average of the reciprocal ranks of the *first* relevant document retrieved for a set of queries:

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}$$

Equation 7. Mean Reciprocal Rank

Where rank_i refers to the position of the first relevant document for the i-th query. For instance, if the relevant document appears at position 1, the score is $1/1 = 1.0$; if it appears at position 5, the score drops to $1/5 = 0.2$. This metric is crucial for RAG systems, as LLMs often prioritize information presented at the beginning of the context window [84], [85].

A significant trend in recent studies is the move away from human-only evaluation toward automated evaluation using Large Language Models. Validating thousands of RAG responses manually is computationally and financially prohibitive. Research demonstrates that advanced LLMs, when prompted with specific evaluation rubrics which focus on:

1. Context Relevance: Measures the relationship between the User Query and the Retrieved Context. It assesses whether the retrieval system successfully filtered out noise and returned only documents pertinent to the user's intent.

2. Groundedness (Faithfulness): Measures the relationship between the Retrieved Context and the Generated Response. This is critical for detecting hallucinations; it ensures that every claim made in the final answer can be inferred directly from the source documents, rather than the model's pre-trained memory.

3. Answer Relevance: Measures of the relationship between the User Query and the Generated Response. It verifies that the final output answers the specific question asked, rather than drifting into unrelated topics.

correlates highly with human judgments. This allows researchers to continuously monitor the performance of RAG pipelines during development, ensuring that improvements in retrieval algorithms (such as hybrid search or reranking) statistically translate into better answer quality [83], [84], [85], [86].

2.2 SYNTHESIS

The reviewed literature establishes a clear intersection between Philippine cultural heritage preservation, educational technology, and advances in artificial intelligence. Prior studies emphasize that Philippine cultural heritage comprising historical, literary, and social records plays a critical role in sustaining national identity and supporting education and research [16], [17]. Government agencies and cultural institutions have undertaken extensive digitization initiatives to improve access to historical materials. While these efforts have expanded public availability, the literature consistently reports challenges related to incomplete collections, unstructured data, limited preservation resources, and difficulties in effective organization and retrieval of digitized materials [18], [19], [20]. These limitations restrict the integration of cultural heritage resources into structured learning environments and highlight the need for systems that can support efficient access, contextualization, and academic use [21], [22].

In parallel, research on artificial intelligence in education demonstrates the increasing adoption of AI-driven tools to support student learning and instructional processes [23], [24]. Studies describe the use of machine learning and natural language processing to enable personalized learning, scalable academic assistance, and automated support functions [25], [26]. AI-based chatbots are frequently discussed as tools for providing on-demand explanations and learning support in higher education settings [29], [30]. However, the literature also identifies persistent concerns related to ethical use, data privacy, unequal access, and the suitability of AI systems for fact-sensitive academic domains, especially within developing educational contexts such as the Philippines [31], [32].

Existing studies on Large Language Models (LLMs) document their effectiveness in natural language generation and interaction but also highlight fundamental limitations when applied to historical and academic tasks [33], [34]. These limitations include hallucinations, lack of explicit grounding, absence of source attribution, static training data, and domain bias [40], [41]. Because LLMs generate responses based on probabilistic token prediction rather than direct consultation of verified sources, their

outputs may lack factual reliability and transparency [42], [43]. The literature consistently indicates that these constraints reduce the suitability of standalone LLMs for culturally sensitive and academically rigorous applications, such as historical inquiry and formal education [44], [45].

Research in information retrieval provides foundational methods for addressing these challenges. Traditional sparse retrieval techniques, including TF-IDF and BM25, are shown to support precise keyword-based matching, particularly for named entities, dates, and locations [71], [72]. More recent dense retrieval approaches based on text embeddings enable semantic matching beyond lexical overlap [66], [68]. However, studies report that neither sparse nor dense retrieval methods are sufficient when used independently, leading to increased interest in hybrid retrieval strategies that combine lexical and semantic signals [73], [74]. These developments form the technical basis for integrating retrieval mechanisms with generative models.

Retrieval-Augmented Generation (RAG) is presented in the literature as an architectural framework that combines LLMs with external document retrieval at inference time [15], [48]. Rather than relying solely on static model parameters, RAG systems incorporate retrieved documents into the generation process, enabling responses that are grounded in external sources [49], [50]. Studies report that RAG-based systems can reduce hallucinations and improve contextual relevance in knowledge-intensive tasks without requiring model retraining [51], [52]. Nonetheless, the literature also emphasizes that RAG performance is highly dependent on retrieval quality, document structure, chunking strategy, and system design [53], [54]. Challenges related to computational cost, evaluation complexity, reasoning limitations, and transparency remain active research concerns [55], [56].

Document ingestion and preprocessing are identified as critical components of effective RAG pipelines. Prior work highlights the importance of document layout analysis, chunking strategies, and text embedding for preserving semantic coherence and ensuring accurate retrieval [59], [60]. Fixed-size chunking is frequently associated with context fragmentation, while semantic and hybrid chunking methods are discussed as approaches that better align document segmentation with conceptual boundaries [61], [62].

Evaluation of RAG systems has also evolved in recent research. Traditional lexical overlap metrics are increasingly viewed as insufficient for assessing knowledge-grounded generation [81], [82]. As a result, studies propose evaluation frameworks

that distinguish between retrieval performance and generation quality. The RAG Triad comprising context relevance, groundedness, and answer relevance is frequently cited as a conceptual framework for evaluating system behavior [83], [84]. Information retrieval metrics such as Hit Rate and Mean Reciprocal Rank remain commonly used to assess retrieval effectiveness, while recent work explores the use of large language models as automated evaluators to support scalable assessment [85], [86].

Synthesizing these findings reveals a clear research gap. While Philippine cultural heritage resources are increasingly digitized, they are primarily treated as archival collections rather than interactive, retrieval-driven learning resources [20], [21]. At the same time, existing AI chatbots lack the grounding mechanisms required to function as reliable academic assistants in historically sensitive domains [29], [30]. Although Retrieval-Augmented Generation has been explored in various educational contexts, limited research focuses on its application to Philippine historical texts, particularly using advanced retrieval pipelines that account for document structure, semantic coherence, and evaluation rigor [57], [58].

In response to this gap, the present study investigates the design and implementation of a Retrieval-Augmented Generation system grounded in curated Philippine cultural and historical materials. Guided by the reviewed literature, the study examines the integration of hybrid chunking, hybrid retrieval, reranking mechanisms, and evaluation frameworks within a domain-specific educational chatbot. By grounding language model outputs in verified historical sources and emphasizing retrieval quality and transparency, the proposed approach seeks to support accurate, context-aware, and academically aligned student learning while addressing limitations identified in prior work.

2.3 THEORETICAL FRAMEWORK

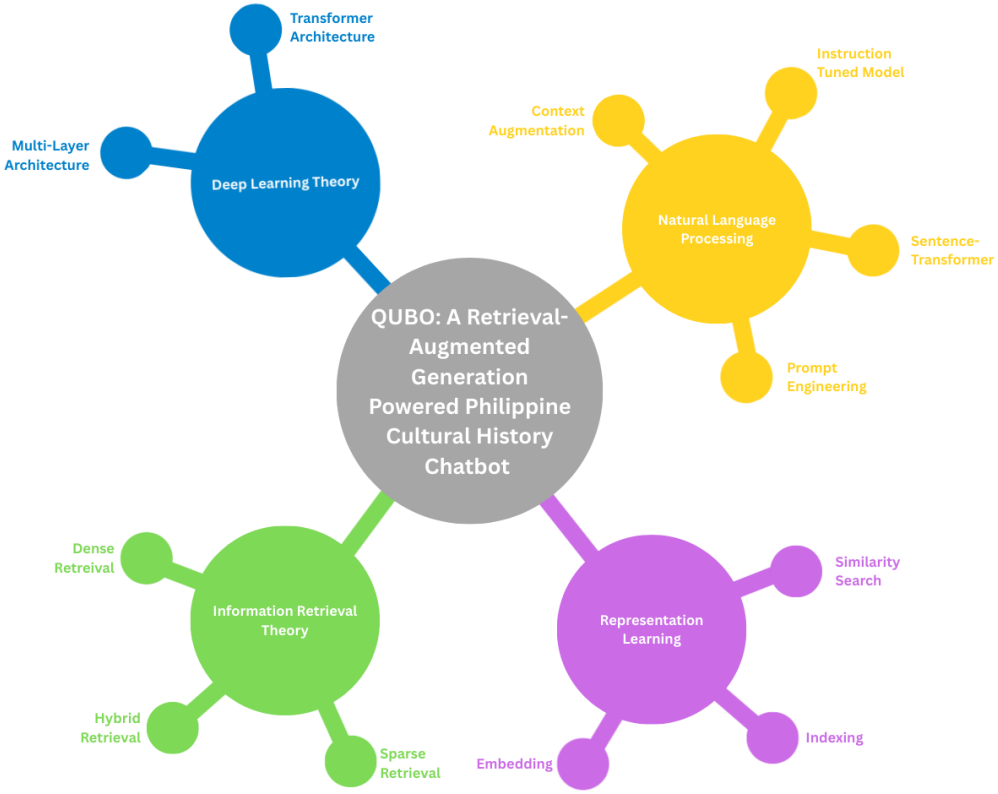


Figure 6. Theoretical Framework

The design and development of the Retrieval-Augmented Generation chatbot is guided by an integration of educational philosophy and advanced computer science. The project takes root in Natural Language Processing, Information Retrieval Theory, Representation Learning Theory, and Deep Learning as seen in Figure 6 [35].

At its core, this educational goal is brought to life through a foundation in Natural Language Processing, which governs how the system understands text, leverages large language models, and uses prompt engineering to elicit desired responses [77], [79]. To find the most relevant information for a query, the chatbot employs Information Retrieval theory, utilizing a hybrid technique that merges sparse and dense retrieval for superior accuracy [73], [74]. This process is made possible by Representation Learning, which transforms text into numerical embeddings a computer can search through using cosine similarity [61] within a specialized Faiss database [87]. The entire technological framework is powered by Deep Learning, by utilizing advanced algorithms and techniques, the capabilities required to keep up with modern standards in computing to create assistive digital tools [38].

2.4 CONCEPTUAL FRAMEWORK

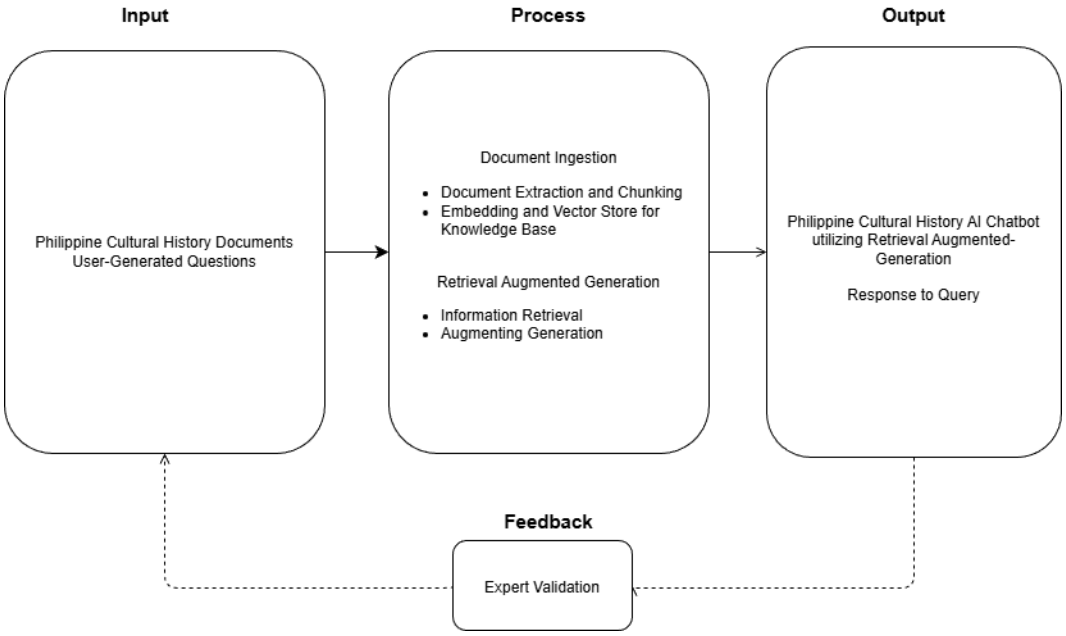


Figure 7. IPO Model

Figure 7 illustrates the study's Conceptual Framework, which uses an Input-Process-Output (IPO) model to depict the system's logical flow. The input stage requires documents to create a knowledge base for information retrieval. The process begins when a user asks a question, causing relevant information to be retrieved and augmented into the generation step. A Large Language Model then synthesizes this information to create an output grounded in the retrieved documents. This response is then submitted back to the user interface to help with the questions. Finally, feedback is collected to gather qualitative and quantitative insights for the continuous enhancement of the chatbot system.

CHAPTER 3
RESEARCH DESIGN AND METHODOLOGY

3.1 RESEARCH DESIGN

This study adopts a quantitative experimental research design to develop and evaluate a Retrieval-Augmented Generation (RAG) system for question answering in the domain of Philippine cultural history. The evaluation is conducted under controlled experimental conditions, wherein key system parameters such as retrieval depth (k) and evaluation settings are systematically varied to examine their effects on retrieval effectiveness and answer generation quality.

Multiple RAG configurations are implemented and tested to analyze how variations in retrieval and generation components influence system performance. By manipulating these configuration parameters and observing their corresponding outcomes, the study aims to establish empirical relationships between design choices and the accuracy, grounding, and reliability of generated responses.

System performance is assessed using quantitative evaluation metrics at both the retrieval and generation stages. Retrieval effectiveness is measured using Hit Rate and Mean Reciprocal Rank (MRR), while generation quality is evaluated using BLEU, ROUGE, and BERTScore, which quantify lexical overlap and semantic similarity between generated outputs and reference answers. Together, these metrics provide objective indicators of the system's ability to retrieve relevant context and generate responses aligned with verified ground truth.

To complement automated metrics, an LLM-as-Judge evaluation framework is employed to assess generated responses along higher-level qualitative dimensions, including correctness, faithfulness to retrieved evidence, and completeness. These criteria capture aspects of answer quality such as factual consistency and source grounding that are not sufficiently reflected by surface-level similarity metrics alone.

In addition, the system undergoes validation by domain experts from the fields of history, machine learning, and software engineering. This multidisciplinary expert evaluation offers a human-centered assessment of the system's accuracy, reliability, and educational suitability, thereby ensuring that the evaluation reflects both technical robustness and domain authenticity.

Overall, this quantitative experimental design seeks to identify RAG configurations that yield accurate, contextually grounded, and academically reliable responses for question answering in Philippine cultural history. The findings are intended to inform the development of an AI-based educational chatbot that prioritizes source grounding, trustworthiness, and domain fidelity in scholarly and learning-oriented applications.

3.2 RESEARCH ENVIRONMENT

The study was conducted in a controlled computational environment designed to evaluate multiple configurations of a Retrieval-Augmented Generation (RAG) system for answering questions related to Philippine cultural history. All experiments were executed on an Acer Nitro laptop equipped with an Intel Core i5 10th-generation processor and an NVIDIA RTX 3060 GPU, providing adequate computational resources for document embedding, vector indexing, and retrieval operations.

System implementation and experimentation were carried out using Python within a Jupyter Notebook environment, enabling iterative development and controlled evaluation of system components. Open-source libraries, including LangChain, were utilized to orchestrate the retrieval and generation pipelines, while large language model inference was performed through services provided by Together.AI.

The historical documents were stored locally and indexed using FAISS as the vector database to ensure consistent access, reproducibility, and efficient similarity search during retrieval. This controlled research environment ensured that all experiments were conducted under uniform conditions, allowing for fair and reliable comparison across different RAG system configurations.

3.3 DATA GATHERING PROCEDURES

This study gathered data to evaluate the performance of a Retrieval-Augmented Generation (RAG) system in answering questions related to Philippine cultural history. The primary data source was the *Philippine History Source Book: Annotated Compilation of Selected Philippine History Primary Sources and Secondary Works in Electronic Format*, published by the National Commission for Culture and the Arts (NCCA). This digitally distributed academic reference is publicly available in PDF format and contains curated primary and secondary readings spanning major periods of Philippine history.

To construct the knowledge corpus for the RAG system, the source book was divided into 34 structured PDF documents, an example of the document content in seen in Figure 8 and Figure 9. Each document contained approximately 19 pages. These documents were stored locally and served as the raw textual repository for subsequent preprocessing, indexing, and retrieval.

In addition to the corpus, an evaluation dataset was constructed to serve as the gold standard for system assessment. This dataset consisted of Question–Answer (QA) pairs explicitly linked to evidence passages within the source documents. The construction process involved identifying a relevant text segment and formulating a corresponding question and answer grounded solely in that passage. A total of 198 validated QA pairs were created and used as ground truth for evaluating retrieval and generation performance.

"A New Species of Homo in the Late Pleistocene of the Philippines"

This is an excerpt of the seminal article on the discoveries made by archaeologist Armand Salvador Mijares and his team (Florent Detroit, Philip Piper, Rainier Grun, Peter Bellwood, Maxime Aubert, Guillaume Champion, Nida Cuevas, Alexandra De Leon, Eusebio Dizon) in the Callao Cave in Peñablanca, Cagayan. This particular study details the discovery of the Right Metatarsal 3 (RMT3) of a human being in Callao Cave, Cagayan. Now designated as the *Homo luzonensis*, the date of the said remains is 67,000 years Before Present. Comparisons were made with the other human remains found in Southeast Asia.

Excerpt:

Hominin movement into Island Southeast Asia has always been problematic due to the lack of well-dated human remains. The humid tropical environment of Island Southeast Asia contributes to the problems of bone preservation. Early modern human remains have, however, been recovered in Niah Cave in Sarawak, Malaysian Borneo, dating to 42 ka, and from Tabon Cave in Palawan dating to 47 plus/minus 10/11 ka. Borneo is located on the sunda shelf and was possibly joined by dry land to Sumatra, Java and Peninsular Malaysia during periods of lowered sea level in the Pleistocene. The island of Palawan may have been intermittently attached to northeastern Borneo when sea levels reached their minima during the most extreme climatic phases. Thus, migrating human populations could have reached both islands without necessarily requiring a sea crossing.

To reach the rest of the Philippine archipelago and other islands in the Wallacean group (e.g., Sulawesi, Flores, Timor) that were never attached to either mainland Asia or Australasia (Sahul), open sea crossings were required. The Lake Mungo remains from Australia dating to 40 plus/minus 2 ka are evidence that modern humans were capable of making early sea crossings. Homo floresiensis, discovered on the islands of Flores, Indonesia, is another hominin that managed to cross the Wallace line. While its remains are only dated to 18-38 ka, Flores also has stone artifact assemblages suggesting that a hominin of unknown affinity reached the island more than 800 ka years ago. Our recent excavations (2007) in Callao Cave have produced what is probably one of the earliest hominin fossils east of Wallace's Line, from the island of Luzon, northern Philippines.

The Callao Cave metatarsal indicates that species of hominin crossed water gaps between Sundaland and Wallacea to reach northern Luzon by 67,000 years ago. The specimen leaves us with a conundrum—in size, it compares with modern Negrito populations, but a few morphological features are unusual. A comprehensive analysis of size and shape characteristics, including comparisons with larger samples of *Homo sapiens* and fossil species of the genus *Homo*, is now needed.

Figure 8. First Sample Document

"Soul Boats: A Filipino Journey of Self-Discovery"

The following is an excerpt of an article originally published in the multi-volume Philippine Heritage in 1978. This particular edition was taken from the book by Alfredo E. Evangelista himself entitled "Soul Boats: A Filipino Journey of Self-Discovery" published in 2001. Alfredo E. Evangelista is the former head of the Anthropology Division of the National Museum of the Philippines and was its Deputy Director until his retirement in 1989. He studied anthropology at the University Chicago and was under the tutelage of archaeologist Wilhelm Solheim and H. Otley Beyer in his excavations in the Neolithic settlements in the Bondoc Peninsula, Masbate and later in Romblon. The essay discussed some of the archaeological evidences of jar burials and coffin burials in the Late Neolithic period in the Philippines. Of particular interest is the discussion on the so-called "soul boats" or wooden coffins shaped like boats found in different archaeological sites in the Philippines.

Excerpt:

Archaeological evidence and early ethnographic accounts by Spanish missionaries indicate that the disposal of the dead in hollowed-out wood has had a long tradition in the Philippines. It developed side by side with other forms of burial including the inhumation of corpses wrapped in mats or tree bark, and interment of corpses or skeletal remains (in jar containers) in caves or under the ground or in the open air. Where only the skulls were interred the container used were ornamented square boxes of either wood or baked clay, or simply slightly deep plates of both high and low-fired clay. In a few instances, as in a Central Philippine burial cave, coffin and jar-burial were practiced simultaneously.

Generally, the dug-out coffins were provided with lids triangular in cross-section, giving them the appearance of roofed boats. Consequently, they have been called "boat coffins." Extrapolation from ethnographic data referred to these coffins as "soul boats" used by the spirit to sail to the world of the dead. Archaeological evidence appears to support this stance. Even where the bones were those of adults, in a practice termed secondary burial, the elongated shape of the coffin and its pyramidal roof were retained. Its overall length, though, was reduced, the coffin appearing at first glance to be a child's. In a rock shelter in easter Masbate Island, for example, is a sawed-off banca with human skeletal remains in it. Being of recent interment, the burial boat was left undisturbed.

It appears further that the idea of burial in a box is as old as burial in a jar, as gleaned from recent archaeological activity in the Tabon Caves of Quezon, Palawan. In one of the chambers of a cave named Manunggul, a Late Neolithic jar-burial assemblage (C14 date: ca. 1000 B.C.) was carefully excavated and studied. Almost a hundred funerary pottery vessels and their associated artifacts constituted the assemblage. An outstanding find was a magnificently

Figure 9. Second Sample Document

Ethical considerations were rigorously observed throughout the study. All textual materials were obtained from publicly accessible academic sources and were used exclusively for non-commercial research purposes. The documents were processed locally and were not distributed in full or in part. No personal, identifiable, or sensitive information was included in the dataset. Automated tools, including large language models, were used transparently and strictly to support dataset construction and system evaluation. The RAG system was designed to generate responses grounded in retrieved excerpts rather than reproducing source texts verbatim, thereby adhering to principles of intellectual property protection and fair academic use.

3.4 SYSTEM ARCHITECTURE / PROPOSED MODEL

3.4.1 Overview of the RAG System

This section introduces the overall architecture of the RAG system. A visual representation of the whole system pipeline is shown in Figure 10. It is divided into two main tasks: the Document Ingestion Flow and the Query Flow, reflecting the distinct processes of ingesting documents and performing retrieval and generation.

The document ingestion Flow focuses on preparing source documents for efficient retrieval. Raw documents are first processed to extract their textual content and are then divided into smaller, manageable sub-documents, referred to as chunks. Each chunk is stored together with its associated metadata in a content repository. In parallel, the chunks are converted into dense vector representations using an embedding model. These vectors are subsequently indexed and stored in a vector database, forming the knowledge base that supports similarity-based retrieval

The query flow begins by analyzing the user's question through an intent router, which allows the system to dynamically set retrieval parameters, then the user's query is embedded using the same embedding model employed during ingestion, ensuring consistency and compatibility between query and document representations. Retrieval is then performed using a specified retrieval technique to identify document chunks that are most relevant to the encoded query. The retrieved chunks are ranked according to their relevance scores. Once a set of top candidate chunks is identified, a structured prompt is assembled by incorporating these chunks as contextual information. This prompt is provided to the selected Large Language Model (LLM), which generates a response grounded in the retrieved document context.

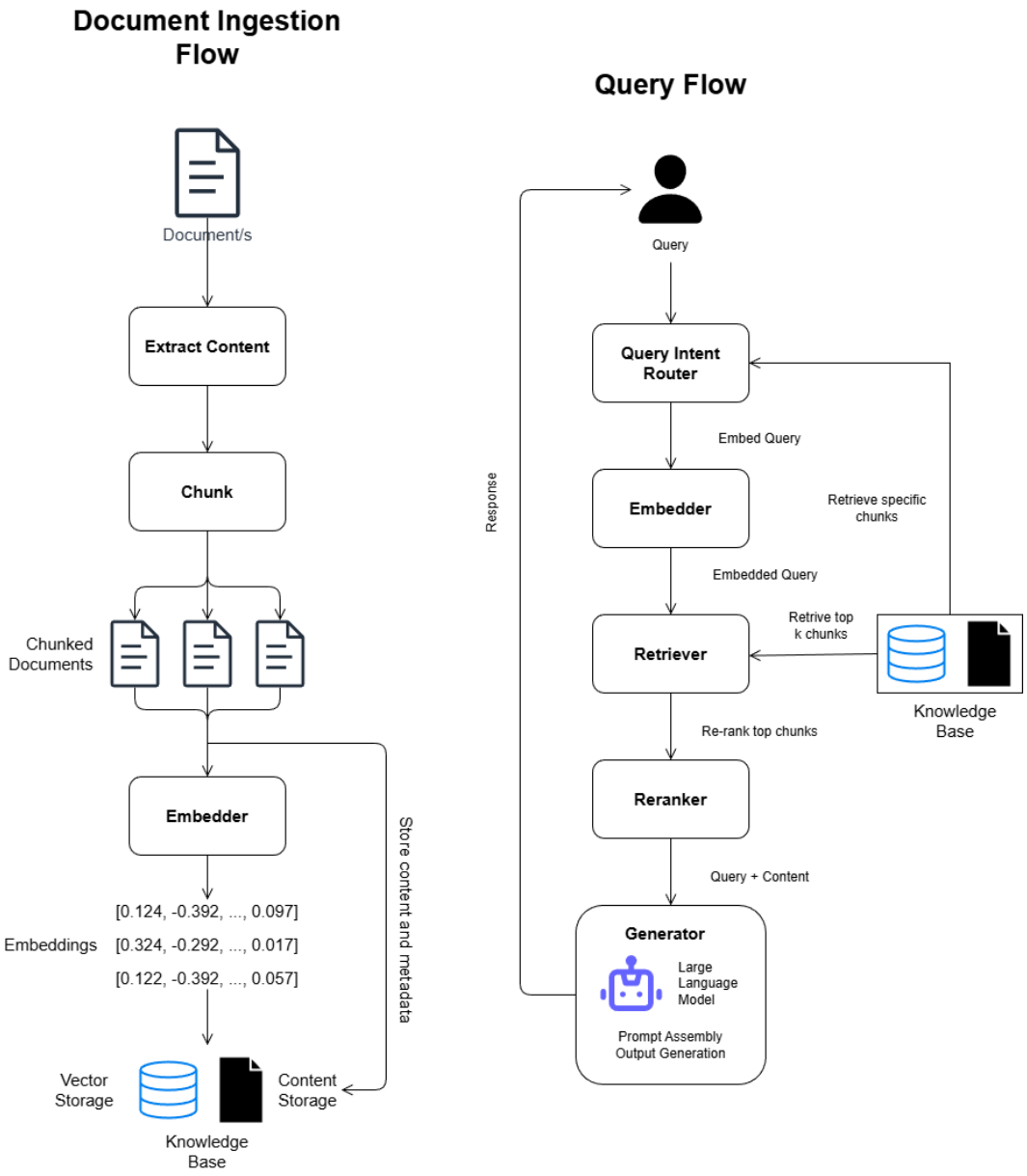


Figure 10. Overview of the RAG System

3.4.2 Ingestion Pipeline

The Ingestion Pipeline manages the transformation of external unstructured data into a structured vector index. This process ensures that source documents are cleaned, segmented, and encoded for optimal retrieval performance.

Content Extraction

The system implements an extraction strategy to handle distinct file formats. For standard textual data, the pipeline utilizes pdfplumber, a Python-based library selected for its ability to preserve document logical structures, including multi-column layouts and tabular data. The extracted of the document content are stored as text file for the chunking stage.

Chunking

To balance context preservation with retrieval precision, the system employs a Two-Stage Hybrid Chunking strategy utilizing the LangChain framework.

1 (Semantic Segmentation): These intermediate segments are further refined using the langchain_experimental library. The system calculates the cosine similarity between adjacent sentences; when similarity drops below a defined threshold, a new chunk boundary is established.The semantic chunker is set to 200 minimum token along with an 80th percentile breakpoint

2 (Token-Based constraint): The raw text is first segmented using a token-based splitter with a fixed limit of 350 tokens. This serves as a hard constraint to ensure all chunks remain within the context window limits of the embedding models.

Embedding

Following segmentation, the system transforms text chunks into high-dimensional vectors. To evaluate the impact of model architecture on retrieval performance, the study implemented two distinct embedding configurations:

- MiniLM-L6: Deployed as a baseline for lightweight, high-efficiency inference.
- Alibaba ModernBERT: Deployed to test high-capacity, long-context retention capabilities.

These models map the semantic content of each chunk into a shared vector space, enabling the system to perform dense retrieval based on conceptual alignment rather than simple keyword matching.

Vector Store

The generated embeddings are indexed in a FAISS (Facebook AI Similarity Search) vector store. The system utilizes the IndexFlatIP structure to perform exhaustive search operations, guaranteeing the identification of exact nearest neighbors without approximation errors.

- Normalization: Prior to indexing, L2 normalization is applied to all vectors to ensure unit length.
- Similarity Metric: With normalized vectors, the system employs Inner Product (IP) search, which is mathematically equivalent to Cosine Similarity, to measure the angular distance between the query and document vectors.
- Document Store (Docstore): Parallel to the vector index, a key-value Docstore is maintained to map retrieved Vector IDs back to their original raw text for context reconstruction.

3.4.3 Query Pipeline

The Query Pipeline executes the real-time processing of user inputs, involving query embedding, retrieval, reranking, and final response generation.

Query Intent Router

To address the limitations of static retrieval parameters, excessive token usage, and latency inherent in standard RAG architectures, this study implemented a Query Intent Router. This module functions as a semantic decision layer, designed to classify user inputs into discrete retrieval categories *prior* to vector database interaction. The methodology for this component involves a three-stage process: semantic classification via Large Language Models (LLM), heuristic output parsing, and dynamic execution routing.

The core classification mechanism utilizes a lightweight Large Language Model (e.g., Llama-3-8B-Turbo) instructed to analyze the semantic intent of the user's query. A structured system prompt guides the model to map the input to a predefined taxonomy seen in Figure 11.

```
"""
Analyze the user's query and classify it into one of three distinct Retrieval Categories:

        1. "GLOBAL_SUMMARY": The user wants a summary, outline, or overview of the *entire* document. (e.g.,
"Summarize the paper", "Give me the main points", "What is the doocument about?")

        2. "BROAD_SEARCH": The user asks for a list, comparison, or explanation that requires gathering many
scattered details. (e.g., "List all the themes", "What are the 5 goals?", "Compare X and Y")


        3. "SPECIFIC_SEARCH": The user asks for a precise fact, date, name, or definition. (e.g., "Who is
Rizal?", "When did he leave?", "What is the capital?")

        4. "PAGE_SPECIFIC": The user explicitly asks for content from specific pages or page ranges. (e.g.,
"Summarize page 10", "What happens on pages 15-20?", "Read the last page")

        5. "GREETING": Conversational filler, greetings, thanks, or questions NOT requiring external
information. (e.g., "Hi", "Hello", "Thank you", "Who are you?")

        6. "NONSENSE": The input is gibberish, random characters, too short to mean anything, or linguistically
incoherent. (e.g., "asdf", "sdsdsd", "..." )

        User Query: "{query}"
        OUTPUT: Output ONLY the category name. No other text.
        """
```

Figure 11. Query Intent Router Prompt

The router classifies inputs into one of six distinct categories, each triggering a specific system behavior:

- GLOBAL_SUMMARY: Defined as requests requiring a holistic overview of the document (e.g., "Summarize the thesis"). This category triggers a "JSONL Bypass," reading a subset of the document directly from file storage to ensure comprehensive coverage without semantic search noise.

- PAGE_SPECIFIC: Defined as queries containing explicit metadata constraints (e.g., "Summarize page 15"). This category enables a deterministic lookup method, extracting text chunks strictly matching the page metadata field, rather than running retrieval.

- SPECIFIC_SEARCH: Defined as fact-seeking queries targeting precise details (e.g., dates, names, definitions). This category dictates a High-Precision Strategy, minimizing

the retrieval parameter (Top k) to fetch a smaller, more concentrated set of chunks. Furthermore, the hybrid weighted fusion is adjusted to prioritize keyword matching (Sparse/BM25), ensuring the system retrieves exact term matches while reducing noise in the context window.

- BROAD_SEARCH: Defined as questions that require synthesis, thematic exploration, or multi-hop reasoning (e.g., explanations, comparisons, lists). This category dictates a High-Recall Strategy, increasing the retrieval parameter (Top k) to capture information scattered across the document. In this mode, the hybrid weighted fusion is shifted to prioritize semantic density (Dense Retrieval), allowing the system to capture conceptually related passages even if they lack exact keyword overlap.

- GREETING: Defined as phatic communication (e.g., "Hello", "Thanks") that requires no external data. The system bypasses retrieval entirely to minimize latency and computational cost.

- NONSENSE: Defined as adversarial, gibberish, or linguistically incoherent input. This serves as a quality control filter, terminating the pipeline immediately to prevent hallucinated processing.

Given the non-deterministic nature of Generative AI, the raw output from the Intent Router is processed through a Regular Expression (Regex) Validation Layer. This layer extracts the intent label even if the model generates extraneous text or punctuation. Based on the validated intent, the system dynamically alters the retrieval parameters. For SPECIFIC and BROAD intents, the system adjusts the retrieval parameters k and weighted fusion a values. For GLOBAL summary and PAGE intents, the system executes an Early Exit strategy, bypassing the vector index entirely to access the raw data files directly. A fallback mechanism defaults to SPECIFIC_SEARCH in cases of ambiguous classification to ensure system robustness.

Query Embedding

The system encodes the input using the same embedding model (MiniLM-L6 or ModernBERT) employed during the ingestion phase. This ensures that the query vector and document vectors exist within the same latent semantic space.

Retrieval

The retrieval phase is responsible for identifying the most relevant document chunks from the knowledge base in response to a user query. To optimize the system's ability to locate pertinent information, this study implements and evaluates three distinct retrieval methodologies. For each method, the system retrieves a configurable top-k set of candidate documents based on their relevance scores.

- Sparse Retrieval (BM25): The system utilizes the BM25 (Best Matching 25) algorithm, a probabilistic retrieval framework based on Term Frequency-Inverse Document Frequency (TF-IDF). This method scores documents by analyzing the overlap of exact keywords between the query and the text, making it highly effective for queries containing specific terminology or proper nouns.

- Dense Retrieval (Semantic Search): simultaneously, the system executes a dense vector search using Cosine Similarity. By encoding both the query and documents into high-dimensional vectors, this method identifies semantically related chunks based on conceptual meaning rather than literal word matches.

- Hybrid Retrieval: To leverage the complementary strengths of the previous two methods, a Hybrid Retrieval approach is employed. This technique merges with a weighted scoring mechanism controlled by a tuning parameter α. This parameter governs the balance between dense and sparse scores, allowing the system to prioritize either semantic understanding or keyword precision depending on the configuration.

Reranking

To refine the retrieval results, the results from the hybrid stage are processed through a Cross-Encoder model, specifically BAAI/bge-reranker-base. Unlike the initial bi-encoder step, this model jointly processes the query and document pair to output a raw relevance logit.

- Score Normalization: The raw logits are passed through a Sigmoid function to map the output to a probabilistic range of [0,1], providing an interpretable relevance score.

- Selection: The top documents that is set with the K parameter with the highest reranked scores are selected as the final context for the generation module.

Generation

The generation phase represents the final stage of the RAG pipeline, where the synthesized response is produced by the Large Language Model (LLM). This process consists of two key steps: Prompt Assembly and Inference.

Prompt Assembly: A structured prompt is dynamically assembled by combining the user's original query with the top-ranked context chunks retrieved from the vector store. To ensure factual accuracy, the system injects strict instructions that explicitly constrain the model to answer solely based on the provided context, thereby mitigating the risk of hallucinations.

Inference: The assembled prompt is transmitted to the hosted inference service via Together.AI. To determine the most effective generator for this specific domain, the study assesses three distinct instruction-tuned Large Language Models:

1. Mistral-Instruct (e.g., Mistral 7B)
2. OpenAI - GPT OSS 20B (referred to as GPT OSS 20B)
3. Meta Llama 70B

The selected model processes the prompt and generates the final response, which aims the Final output to be grounded in the retrieved historical data.

3.5 TOOLS AND TECHNOLOGIES

The Retrieval-Augmented Generation (RAG) system was developed using a Python-based architecture designed to support flexible document ingestion, semantic retrieval, and controlled text generation. LangChain was used as the main orchestration framework for building the RAG pipeline because it simplifies the coordination between retrieval and generation components. Its abstraction for prompt handling and chaining allowed the system to manage complex workflows without excessive implementation overhead, making it suitable for rapid experimentation.

To support diverse document formats text extraction approach was adopted. Digitally generated PDF files were processed using pdfplumber, which enables extraction of structured text, including paragraphs, tables, and layout-aware elements

For semantic indexing and retrieval, FAISS (Facebook AI Similarity Search) was selected as the vector database due to its efficiency in handling high-dimensional embeddings and large-scale similarity searches. FAISS enabled fast retrieval of relevant document chunks while remaining scalable as the dataset size increased. Several embedding models were evaluated to balance retrieval quality and computational efficiency. The all-MiniLM-L6-v2 model was chosen for general-purpose semantic retrieval because it produces compact 384-dimensional embeddings while maintaining strong performance. For documents requiring long-context understanding, Alibaba's Gte ModernBERT Base was incorporated. Unlike earlier models limited to short segments, Gte ModernBERT Base supports inputs of up to 8,192 tokens, allowing entire documents or long sections to be embedded in a single pass.

To further refine retrieval results, the BAAI/bge-reranker-base model was employed as a reranking stage. This model was selected for its cross-encoder architecture, which can process dual inputs (query and document) simultaneously. This additional step aims to improve the quality of the contextual information passed to the generation model, resulting in more accurate and coherent responses.

For the generation component, multiple open-weight language models were explored to assess performance across different scales. Mistral 7B Instruct v0.3 was used as a baseline for local inference due to its efficient architecture and strong reasoning capabilities. OpenAI GPT-OSS 20B was incorporated as a mid-sized model. Meta Llama 3 70B Instruct was evaluated as a large-scale open-weight model, selected for its stability, widespread adoption, and ability to generate high-quality, context-aware outputs for complex queries. All remote inference tasks were hosted via Together AI to ensure consistent performance.

To execute rigorous evaluation protocols, specific Python libraries were integrated into the workflow. The Hugging Face evaluate library provided the implementation for standard lexical metrics, including ROUGE-L and BERTScore (utilizing distilbert-base-uncased for semantic alignment), while SacreBLEU was employed to calculate standard BLEU scores. For the LLM-as-a-Judge framework, the assessment logic was scripted in Python, utilizing Qwen 2.5 hosted on Together AI to serve as the automated judge model. Finally, data analysis and metric aggregation were performed using Pandas and NumPy to calculate mean scores and visualize performance distributions across different experimental configurations.

To facilitate user interaction and document submission, the system interface was developed as a chatbot prototype using React.js. This framework was selected for its component-based architecture, which allows for modular development and efficient state management. The interface features a dedicated document upload module that allows users to select, validate, and preview PDF files prior to processing. Technical implementation relies on React Hooks to manage local state, file selection logic, and asynchronous network requests, ensuring responsive communication with the backend server.

The backend infrastructure was built using FastAPI, a high-performance web framework for Python. Selected for its native support for asynchronous programming, FastAPI efficiently handles concurrent requests such as document uploads and long-running generation tasks without blocking the main execution thread. It serves as the central orchestration layer, exposing RESTful endpoints that trigger the RAG pipeline, manage file I/O operations, and validate data schemas using Pydantic models. This separation of concerns ensures that the heavy computational logic remains decoupled from the client-side interface.

3.6 EXPERIMENTAL PROCEDURES / IMPLEMENTATION PLAN

The experimental knowledge base was constructed using 34 distinct documents sourced from the Philippine History Source Book: Annotated Compilation of Selected Philippine History Primary Sources and Secondary Works in Electronic Format. These documents consist of curated primary and secondary readings spanning major periods of Philippine history. Each of the 34 documents contains 19 pages of text, resulting in a total corpus of roughly 646 pages prior to processing.

Each document underwent the ingestion pipeline to prepare the unstructured data for retrieval. First, raw text was extracted using pdfplumber . Following extraction, the text underwent a specialized hybrid chunking strategy to optimize semantic integrity. LangChain Semantic Chunker (configured with a 200-token minimum and an 80th-percentile breakpoint threshold) was applied to segment text based on conceptual boundaries rather than arbitrary length. To strictly enforce the context window limits of the embedding models, these semantic segments were further refined using the TokenTextSplitter, which imposed a hard cap of 256 tokens with a 100-token overlap. Finally, these processed chunks were vectorized using two distinct embedding models (all-MiniLM-L6-v2 and Gte-ModernBERT Base) and indexed into separate FAISS vector databases to facilitate dense retrieval.11

Prior to system testing, a "Golden Standard" evaluation dataset was rigorously constructed to serve as the ground truth. A total of 198 evaluation triples (Question, Ground Truth Answer, Evidence Passage) were manually curated. For each entry, human annotators isolated a specific processed chunk and formulated a query answerable exclusively by that chunk. To ensure a stratified assessment of model capabilities, these queries were classified into three complexity levels: Factual (simple recall), Descriptive (synthesis), and Explanatory (reasoning). This dataset provided the authoritative baseline for calculating retrieval and generation performance.

To empirically determine the optimal retrieval architecture, we designed and executed a comparative study across four progressive system configurations. The study began with Configuration A (Baseline), a standard dense retrieval system using all-MiniLM-L6-v2 and FAISS. This was followed by Configuration B (Hybrid V2), which introduced a weighted fusion of dense search and sparse keyword matching (BM25) to improve lexical precision. Configuration C (Hybrid V3) upgraded the embedding model to Gte-ModernBERT Base to capture deeper semantic nuance. Finally, the Proposed System extended Configuration C by integrating the Intent Router which uses (Llama-3-8B-Turbo) and Cross-Encoder Reranker (BAAI/bge-reranker). Each configuration was evaluated using Hit Rate and Mean Reciprocal Rank (MRR) at varying retrieval depths ($k = 1, 3, 5, 10$) The configuration yielding the highest performance metrics was identified as the Optimal Retrieval Configuration for the subsequent generation phase.

Upon establishing the optimal retrieval configuration, we benchmarked the performance of three distinct Large Language Models (LLMs) to evaluate their ability to synthesize retrieved chunks into accurate answers. The models selected for this comparative study were Mistral-Instruct, GPT-OSS-20B, and Meta Llama-3-70B. To ensure a controlled experiment where the model architecture remained the sole variable, a standardized system prompt was utilized on all model inferences. This prompt strictly constrained the models to derive answers solely from the provided context, preventing the leakage of pre-trained external knowledge. Furthermore, all models were configured with a temperature setting of 0 to minimize stochasticity, ensuring that the outputs were deterministic and reproducible.

To rigorously assess the quality of the generated outputs, a multi-faceted evaluation strategy was conducted that combined traditional lexical metrics with semantic analysis using the HuggingFace evaluate library. For lexical evaluation, we employed sacrebleu to calculate standard BLEU scores, providing a baseline measure of exact n-gram precision between the generated answers and the ground and ROUGE-L (longest common subsequence) metrics.

Recognizing the limitations of exact string matching in Generative AI, we prioritized semantic analysis using BERTScore. This metric was computed utilizing the distilbert-base-uncased model, which measured the contextual embedding alignment between the candidate answer and the reference. This allowed us to quantify whether the meaning of the generated response was accurate, even when the phrasing differed from the ground truth.

Complementing these automatic metrics, we adopted an LLM-as-a-Judge approach to approximate human evaluation capabilities. We employed Qwen-2.5-72B-Instruct as a judge, providing it with a specific rubric to grade responses on a scale of 1 to 5. This grading process evaluated three dimensions: Faithfulness, ensuring the answer was derived only from the retrieved context; Completeness, assessing whether the answer addressed all parts of the query; and Correctness, verifying factual alignment with the ground truth. Finally, to identify specific model strengths and failure modes, we conducted a stratified output analysis based on the question metadata tags including Factual, Descriptive, and Explanatory types. Rather than relying solely on aggregate scores, we isolated the performance metrics for each category to compare the generated outputs against the reference text. This granular analysis allowed us to determine if models exhibited performance degradation when handling complex reasoning tasks versus simple information retrieval, providing deeper insight into the reliability of the RAG system across different query types.

Throughout the experimental process, all procedures were applied consistently across configurations, with no changes made to the underlying corpus, chunking strategy, or evaluation dataset. This ensured that observed performance differences could be attributed directly to architectural and model-level variations rather than procedural inconsistencies.

3.7 EVALUATION

The evaluation of the RAG system will be grounded in a curated Gold Standard dataset serving as the ground truth for performance benchmarking. The assessment focuses on two critical components: retrieval precision and generation quality. To measure the effectiveness of the retrieval module, the study employs rank-aware metrics, specifically Hit Rate and Mean Reciprocal Rank (MRR), which assess the system's ability to surface the most relevant documents. For generation quality, the system is evaluated using a combination of traditional lexical metrics such as BLEU and ROUGE and semantic evaluators like BERTScore. Additionally, an "LLM-as-a-Judge" framework is utilized to provide a nuanced assessment of faithfulness,

accuracy, and completeness. Finally, to ensure robust validation, a subset of outputs undergoes manual human review to verify alignment with gold reference answers and identify instances where the Large Language Model may have failed to synthesize the correct response.

3.7.1 Retrieval Evaluation

Hit Rate

Hit Rate@k is a binary success metric that measures the ability of the retriever to identify the correct context required to answer a query. Specifically, it computes the proportion of queries for which the ground-truth context defined as the document chunk containing the verified answer appears within the top-*k* retrieved results.

In this study, a *hit* is recorded if the unique identifier of the ground-truth chunk matches any of the identifiers in the top-*k* retrieved set. This metric is critical because it represents the upper bound of the system's end-to-end performance: if the retriever fails to surface the correct context, the generation component cannot produce a faithful or grounded response.

The Hit Rate@k is formally defined as:

$$\text{HitRate@k} = \frac{1}{N}\sum_{i=1}^{N} \text{hits}\,(i, k)$$

where N is the total number of evaluation queries and hits(i, k) equals 1 if the ground-truth chunk for query $\text{i}$appears in the top-k results, and 0 otherwise.

Mean Reciprocal Rank (MRR)

While Hit Rate measures the presence of relevant information, Mean Reciprocal Rank (MRR) evaluates the quality of the retriever's ranking algorithm. It accounts for the position of the first relevant document by calculating the average of the reciprocal ranks across all test queries. A high MRR indicates that the system consistently places the most relevant context at the beginning of the retrieved list, which is critical for reducing "noise" in the prompt and ensuring the language model prioritizes the most accurate data.

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

where | Q |is the total number of queries and rank_i denotes the rank position of the first relevant document for query i

3.7.2 Generation Quality Evaluation

The quality of generated responses in the RAG system was evaluated using a combination of automatic metrics and human evaluation to ensure both quantitative and qualitative assessment of performance. This dual approach captures not only the textual similarity with reference answers but also the factual correctness, fluency, and relevance of the outputs.

BLEU (Bilingual Evaluation Understudy)

The quality of the generated outputs was evaluated using the BLEU (Bilingual Evaluation Understudy) score, an automatic evaluation metric commonly used to measure the similarity between machine-generated text and reference answers. BLEU assesses generation quality by calculating the degree of n-gram overlap between the generated response and the corresponding reference text. Higher BLEU scores indicate greater lexical similarity and closer alignment with the reference answer.

In this study, BLEU was used to provide an objective, quantitative measure of how closely the generated responses matched the expected answers in terms of wording and phrase structure. Although BLEU does not capture deeper semantic meaning or factual correctness, it is useful for comparing relative performance across different RAG configurations under consistent evaluation conditions.

BERTScore

BERTScore was employed as a complementary evaluation metric to address the limitations of surface-level overlap measures. Unlike BLEU, BERTScore uses contextual embeddings from pretrained BERT models to evaluate the semantic similarity between generated responses and reference answers. This allows the metric to account for paraphrasing and variations in wording while preserving meaning.

In the context of this study, BERTScore was used to assess how well the generated answers captured the intended meaning of the reference responses, even when exact wording differed. This metric is particularly suitable for evaluating RAG-based question answering

systems, where semantically correct answers may not always exhibit high lexical overlap with reference texts.

LLM-As Judge

To address the limitations of rigid, code-based evaluation metrics, this study introduced a complementary assessment layer known as LLM-as-a-Judge, inspired by the RAGAS evaluation framework. This approach utilized the Qwen 2.5 model to evaluate the semantic quality of generated responses, with the objective of approximating human judgment at scale. The integration of an LLM-based evaluator enabled the exploration of automated qualitative analysis within Retrieval-Augmented Generation (RAG) systems, particularly for identifying nuanced response issues such as hallucinations and subtle factual inconsistencies that traditional lexical metrics (e.g., BLEU or ROUGE) often fail to capture.

The evaluation process involved providing Qwen 2.5 Instruct Turbo with a structured set of inputs, including the original user query, the retrieved contextual documents, the RAG-generated response, and a ground-truth reference answer. Based on this information, the model was instructed to assess the response across three critical qualitative dimensions: factuality, which evaluates the correctness of the information presented; completeness, which determines whether the response sufficiently addresses the full scope of the query; and faithfulness, which verifies that the generated answer is strictly grounded in the retrieved context without introducing unsupported or fabricated details.

As illustrated in Figure 11, the evaluation prompt was adapted from the RAGAS framework and explicitly instructed the model to assume the role of an expert auditor. The model was required to assign numerical scores ranging from 1 to 5 for each evaluation dimension of faithfulness, correctness, and completeness and to output the results in a strict JSON format to facilitate automated inspection and aggregation. This procedure was applied consistently across all generated responses, enabling the computation of average scores for comparative analysis.

Despite its utility, the LLM-as-a-Judge mechanism was deliberately treated as experimental and exploratory rather than authoritative. The study acknowledges that large language models remain susceptible to biases, variability in judgment, and limitations in contextual reasoning. Consequently, the scores produced by Qwen 2.5 were interpreted as auxiliary indicators intended to support and complement human evaluation, which served as the primary and final standard for assessing response quality in this research.

```
"""
### Role
You are an expert auditor. Your task is to provide a rigorous, objective evaluation of a GENERATED ANSWER.

### Inputs
**Question:** {question}

**Retrieved Context (The only source of truth):** {context}

**Reference Answer (The ideal response):** {reference}

**Generated Answer (To be evaluated):** {generated}

### Evaluation Criteria & Rubric

#### 1. Faithfulness (Groundedness)
- **Definition:** Is every claim in the answer supported by the Context?
- **Score 1:** Major hallucinations; contains information that contradicts the context.
- **Score 3:** Partially grounded; contains some claims not found in the context.
- **Score 5:** Perfectly grounded; every single sentence is supported by the provided context.

#### 2. Correctness
- **Definition:** How well does the generated answer match the factual content of the Reference Answer?
- **Score 1:** Factually wrong or completely irrelevant.
- **Score 3:** Captures the main idea but misses key nuances.
- **Score 5:** Fully accurate and matches the reference answer's meaning.

#### 3. Completeness
- **Definition:** Does the answer address all parts of the user's question?
- **Score 1:** Extremely brief or ignores most of the question.
- **Score 5:** Thorough; provides a comprehensive response to the query.

### Response Format
You must think step-by-step. First, analyze the alignment between the inputs. Then, provide the scores in the
following JSON format. Do not add conversational text outside the JSON.

{{
  "thought_process": "<your step-by-step reasoning for each metric>",
  "faithfulness": <int>,
  "correctness": <int>,
  "completeness": <int>,
  "verdict": "<final brief summary>"
}}
"""
```

Figure 12. LLM-As-Judge prompt

3.7.3 Expert Validation

The Retrieval-Augmented Generation (RAG) system integrated within the chatbot prototype underwent a rigorous expert evaluation to assess both its functional software quality and the integrity of its data processing capabilities. To ensure a holistic assessment, the evaluation panel was selected from three distinct professional domains: Software Engineers with expertise in system deployment and infrastructure; AI and Machine Learning specialists focused on algorithms

and information retrieval; and History experts specializing in cultural studies and Philippine events.

The evaluation process involved inviting these experts to a scheduled system demonstration where the core functionalities specifically file upload and RAG-based chatbot response capabilities were showcased. Prior to the session, all participants were briefed on data privacy protocols, ensuring that all collected data would be strictly utilized for research purposes. Following the demonstration, experts recorded their feedback using a structured Google Forms survey.

Designed with a dual-standard approach, the survey instrument assesses both software quality and data reliability. Metrics for software performance and usability adhere to ISO/IEC 25010, while data retrieval and content accuracy follow ISO/IEC 5259. This ensures that the system's stability is tested alongside the credibility of the AI's data. The questionnaire is tailored to the experts' respective fields and is presented in Tables 1–4

The responses from the experts were analyzed using the weighted mean to determine the overall evaluation of the system per criterion. A five-point Likert scale was used, where 5 indicates Strongly Agree and 1 indicates Strongly Disagree. The weighted mean was computed by multiplying each scale value by its corresponding frequency, summing the products, and dividing the result by the total number of respondents.

$$Weighted\ Mean = \Sigma(F \times X) \div N$$

Equation 8. Weighted Mean

The computed weighted means were then interpreted using the predefined verbal interpretation scale. Results showed that the system obtained weighted mean scores within the “Very Good” to “Excellent” range across the evaluated criteria. This indicates that the experts generally agree that the system is functional, accurate, usable, and reliable. Overall, the findings suggest that the proposed system meets the required quality standards and is acceptable for implementation.

Table 1: General Evaluation Criteria (All Experts)

Standard	Quality Characteristic	Survey Statement
ISO 25010	Functional Suitability	The system successfully processes uploaded documents without issues.
ISO 5259	Semantic Accuracy	The responses are relevant and directly address the user's questions.
ISO 5259	Completeness	The system answers questions strictly using the information from the provided files.
ISO 5259	Consistency	The system properly refuses to answer questions that are not related to the uploaded document.
ISO 25010	Operability	The interface is simple and easy to navigate.
1ISO 25010	Time Behavior	The system responds to questions at a reasonable speed.
ISO 25010	Learnability	The tool is suitable for students or researchers with little technical background.

Table 2: Evaluation Criteria for Software Engineers

Standard	Metric/ Characteristic	Survey Statement
ISO 25010	Maturity (Reliability)	The application runs smoothly and does not crash during standard operation.
ISO 25010	Performance Efficiency	The system handles tasks (like uploading files or generating answers) without excessive delays or lag.
ISO 25010	Integrity (Security)	The system handles user inputs and file uploads securely and correctly.

Table 3: Evaluation Criteria for AI / ML Experts

Standard	Metric/ Characteristic	Survey Statement
ISO 5259	Data Accuracy (Syntactic)	The system effectively prevents the AI from making up information (“hallucinating”) by sticking to the source syntax.
ISO 5259	Data Completeness	The system consistently retrieves the correct and complete information segments from the document.
ISO 25010	Interpretability	The AI synthesizes information well, creating answers that are coherent and easy to read.

Table 4: Evaluation Criteria for History Experts

Standard	Metric/ Characteristic	Survey Statement
ISO 5259	Data Credibility	The chatbot's answers are faithful and factually aligned with the valid historical text provided.
ISO 25010	Effectiveness	This tool helps reduce the time needed to analyze historical documents.
ISO 5259	Currentness/Value	The system effectively presents historical data in a way that aids student understanding (Pedagogical Value).

CHAPTER 4
RESULTS AND DISCUSSION

4.1 EXPERIMENTAL SETUP

The experiments in this study were conducted in a controlled computational environment to evaluate different configurations of a Retrieval-Augmented Generation (RAG) system for question answering in Philippine cultural history. All experiments were performed on an Acer Nitro laptop equipped with an Intel Core i5 10th-generation processor and an NVIDIA RTX 3060 GPU, providing sufficient computational capacity for local embedding and retrieval tasks. System implementation was carried out using Python within a Jupyter Notebook environment. Open-source libraries, including LangChain, were used to orchestrate the retrieval pipelines, while the Large Language Models (LLMs) were hosted on inference services provided by Together AI and Huggingface to ensure consistent high-performance inference.

The dataset used in this study consists of curated textual documents on Philippine cultural history derived from Philippine History Source Book: Annotated Compilation of Selected Philippine History Primary Sources and Secondary Works in Electronic Format. The original material from the book was manually curated and organized into 34 distinct documents, each corresponding to approximately 19 pages of content. This process transformed the source book into a structured corpus suitable for retrieval and generation experiments.

The 34 documents underwent the ingestion pipeline to convert unstructured text into a vector-searchable format. The content of each document was extracted using pdfplumber for. Then hybrid chunking is implemented using the langchain_experimental Semantic Chunker segmented text based on semantic meaning using a 200 minumum token count and 80th percentile breakpoint. Second, to strictly adhere to the model's context window, these segments were refined using TokenTextSplitter, enforcing a maximum chunk size of 256 tokens with a 100-token overlap.The processed chunks were vectorized using two distinct embedding models all-MiniLM-L6-v2 and Gte-ModernBERT Base to allow for a comparative analysis of semantic alignment in the historical domain. The resulting vectors were indexed in separate FAISS vector databases using Cosine Similarity for ranking.

To establish a reliable benchmark for both retrieval and generation performance, a "Gold Standard" evaluation dataset was constructed derived from the preprocessed documents tbe description of the dataset is seen in Table 5. A total of 198 question-answer pairs were manually

created. Each item in this dataset links a user query to its ground-truth answer and the precise corresponding text chunk generated by the ingestion pipeline.

Table 5: Evaluation Dataset

Field	Description
Question	The natural language inquiry represents a user's request for historical information. These are formulated to simulate real-world queries about Philippine history.
Answer	The ground-truth response derived directly from the source text. This serves as the "Gold Standard" reference used to evaluate the correctness of the model's generation.
Question Type	A categorical label (e.g., short fact, Descriptive, Explanatory) assigned to each query. This allows for stratified evaluation to see how the system performs on different levels of complexity.
Evidence Passage	The specific processed text chunk from the vector database that contains the factual basis for the answer. This ensures that the ground truth aligns with the granularity of the retrieval system.

With the corpus indexed in FAISS and the Gold Standard dataset established, the experimental framework was fully prepared to execute retrieval and generation evaluations. The evaluation of the RAG system was divided into two stages: retrieval evaluation and generation evaluation. In the retrieval evaluation, the 198 questions from the gold-standard dataset were issued to the retriever to assess its ability to locate relevant text chunks. Performance was measured using Hit rate, defined as the percentage of questions for which at least one relevant chunk is successfully retrieved, and Mean Reciprocal Rank (MRR), which reflects the rank position of the first relevant chunk within the retrieved list. The experiment was conducted using *k* values of 1, 3, 5, and 10 to represent the number of chunks returned per query; these results were used to determine the optimal *k* value for subsequent stages. These metrics were selected because they directly capture both the presence of relevant evidence and the efficiency with which it is retrieved.

To rigorously examine the impact of distinct system design choices, we implemented and evaluated four progressive retrieval configurations seen in Table 6.

Table 6. Retrieval Configurations

Configuration	Retrieval Strategy	Embedding Model	Indexing & Search	Reranking	Routing
Baseline	Dense Retrieval	all-MiniLM-L6-v2	FAISS (Dense Index)	None	Fixed
V2 (Hybrid)	Hybrid (Dense + Sparse) with Weighted Fusion	all-MiniLM-L6-v2	FAISS (Dense) + BM25 (Sparse)	None	Fixed
V3 (Enhanced)	Hybrid (Dense + Sparse) with Weighted Fusion	Gte Modernbert Base	FAISS (Dense) + BM25 (Sparse)	None	Fixed
Proposed	Hybrid (Dense + Sparse) with Weighted Fusion	Gte Modernbert Base	FAISS (Dense) + BM25 (Sparse)	Cross-Encoder (BAAI/bge-reranker)	Intent Routing

1. Configuration A (Baseline): Established a fundamental performance benchmark using a standard dense retrieval system with all-MiniLM-L6-v2 embeddings and FAISS vector similarity search.

2. Configuration B (Hybrid V2): Introduced a hybrid retrieval strategy combining dense vector search with sparse keyword matching (BM25) via a weighted fusion algorithm to address the limitations of purely semantic retrieval.

3. Configuration C (Hybrid V3): Maintained the hybrid architecture but upgraded the embedding model to Gte-ModernBERT Base, aiming to enhance semantic representation and capture more nuanced contextual relationships.

4. Proposed System (Adaptive): Extended Configuration C by integrating an Intent Router and a Cross-Encoder Reranker.

To ensure a fair quantitative comparison with the static baselines, the final output size k of the Proposed System was fixed to match the experimental control values (e.g., $k = 1, 3, 5, 10, 10$). However, unlike the baseline, the Proposed System dynamically adjusted its internal retrieval parameters based on the detected intent:

Broad Search: The system optimized for Recall by setting the hybrid fusion weight to $\alpha = 0.7$ (favoring dense vectors) and expanding the internal candidate pool to 50 chunks. These 30 candidates were re-scored by the Cross-Encoder to capture semantic nuance, and the top-k were returned.

Specific Search: The system optimized for precision by setting the fusion weight to $\alpha = 0.3$ (favoring keyword matches) with a tighter internal retrieval window of $2 \times k$. These candidates were also re-scored by the Cross-Encoder to filter out keyword-stuffing false positives, ensuring that the final k chunks contained the most accurate factual matches.

By strictly limiting the final output to the experimental k value, this ensures that any observed performance gain is attributable to the system's adaptive intelligence in selection and ranking, rather than simply retrieving a larger volume of data than the baselines.

Upon establishing the optimal retrieval configuration determined by the highest Hit Rate and Mean Reciprocal Rank, the system advances to the Generation Evaluation phase. In this stage, the retrieved context chunks were processed by three distinct Large Language Models to

synthesize answers: Mistral-Instruct (7B), Gpt-OSS 20B (representing the mid-sized open-weights category), and Meta-Llama-3-70B. To ensure a strictly controlled comparison, all models utilized the identical System Prompt and User Prompt structure defined in Table 7, where the context variable was populated with the specific chunks retrieved by the optimal configuration and the question was drawn directly from the Evaluation Dataset.

Table 7: LLM Prompt

System	You are an expert history assistant.
User	Answer the question using only the provided context. If the answer is not present, say: “The information is not available in the provided documents.” Context: {context} Question: {question} Answer:

The quality of the responses generated was assessed using a multi-layered evaluation framework that combined lexical, semantic, and qualitative analysis. We first calculated automatic reference-based metrics, specifically Bleu, ROUGE-L and BERTScore, to measure the linguistic similarity between the generated answer and the ground-truth reference. However, recognizing that these lexical metrics can penalize correct answers that simply use different wording, we complemented them with an LLM-as-a-Judge approach. A Qwen 2.5 model was employed as an automated evaluator to grade each response on three critical dimensions: Faithfulness (ensuring the answer is derived only from the retrieved context), Correctness (factual alignment with the ground truth), and Completeness (addressing all aspects of the user's query). To conclude the

evaluation, we conducted a stratified qualitative analysis to understand the nuanced behaviors of the system beyond aggregate numbers. This involved manually verifying a sample of responses from the best-performing model across three distinct question types: Short Fact, Descriptive, and Explanatory. This manual review allowed for the identification of specific failure modes, distinguishing between simple retrieval errors in fact-seeking queries and complex reasoning failures in explanatory tasks. While LLM-based evaluation introduces subjectivity, prior studies have shown strong correlation between LLM judgments and human evaluation when constrained by explicit criteria and zero-temperature decoding.

To ensure reproducibility and scientific rigor, the parameters detailed in Table 8 were maintained across all experiments.

Table 8: Parameters Experimental Value

Algorithm/model	Parameter	Experimental Value
LangChain Semantic Chunker	Breakpoint Threshold	80th Percentile
	Min Chunk Size	200 Tokens
LangChain Token Text Splitter	Chunk Size	256 Tokens
	Chunk Overlap	100 Tokens
sentence-transformers/all-MiniLM-L6-v2	Max Sequence Length	256 Tokens

	Dimensions	384
Alibaba-NLP/gte-modernbert-base	Max Sequence Length	8192 Tokens
	Dimensions	768
FAISS	Index Type	Cosine Similarity (L2-Normalized Euclidean IndexFlatL2)
Sparse (BM25)	K	[1, 3, 5, 10]
Dense (Vector)	K	[1, 3, 5, 10]
Hybrid Weighted Fusion	Weight (α)	[0.3, 0.5, 0.7]
	Output Candidates	[1,3,5,10]
BAAI/bge-reranker-base	Input Candidates	$[\,2K, 30]$
	Output Candidates	[1, 3, 5, 10]

Meta Llama-3-8B-Turbo	Temperature	0.0
	Max Output Tokens	16
mistralai/Mistral-7B-Instruct-v0.2	Temperature	0.0
	Max Output Tokens	1024
openai/gpt-oss-20b	Temperature	0.0
	Max Output Tokens	1024
meta-llama/Meta-Llama-3-70B-Instruct	Temperature	0.0
	Max Output Tokens	1024
Qwen 2.5	Temperature	0.0
	Max Output Tokens	1024

Every configuration was evaluated against the identical preprocessed corpus and Gold-Standard dataset; no modifications were made to the underlying data or chunking strategy between trials. Furthermore, the optimal retrieval depth k once empirically determined during the initial retrieval evaluation, was applied consistently across all subsequent generation experiments. This strict control of independent variables ensures that any observed performance differences can be attributed solely to the architectural improvements such as the introduction of hybrid fusion or reranking rather than inconsistencies in the experimental setup.

4.2 MODEL PERFORMANCE RESULT

4.2.1 Retrieval Result

The performance of the retriever was evaluated using Hit Rate and Mean Reciprocal Rank (MRR) across four retrieval depths $k = \{1, 3, 5, 10\}$ Table 9 summarizes the performance metrics of the Final Proposed Configuration compared to the Baseline.

Table 9: Retrieval Scores

Variant	Metric	k=1	k=3	k=5	k=10
Baseline	Hit Rate (%)	55.20	70.71	74.20	80.32
	MRR	0.552	0.618	0.625	0.635
Proposed	Hit Rate (%)	83.00	**91.50**	93.20	94.10
	MRR	0.83	**0.868**	0.872	0.874

The Proposed System demonstrated significance in performance, achieving a Hit Rate of 83.00% and an MRR of 0.830 at the strictest retrieval depth of $k = 1$. This indicates that for most queries, the system correctly identified the most relevant document as the top result, validating the combined efficacy of the Hybrid retrieval strategy and Cross-Encoder Reranking. Expanding the retrieval scope to $k = 3$ further improved the Hit Rate to 91.50%, with a corresponding MRR increase to 0.868. The close alignment between the Hit Rate and MRR scores confirms the

system's ranking precision; even when the correct answer was not at rank 1, it consistently appeared near the top of the list rather than being buried deep in the retrieval window.

In contrast, the Baseline (Dense-only) configuration exhibited moderate performance, achieving a Hit Rate of only 55.20% at $k = 1$. While its recall improved as the retrieval window expanded, reaching 80.32% at $k = 10$, the relatively low MRR of 0.552 suggests that relevant documents were frequently ranked lower in the list. Notably, the Proposed System's performance at $k = 1$ (83.00%) outperformed the Baseline's performance even at its most lenient setting of $k = 10$ (80.32%), demonstrating a significant improvement in both precision and recall.

Figure 12 visually illustrates the performance trajectory, revealing a distinct "elbow point" at k=3 where the system reaches optimal efficiency. The curve shows a steep ascent from k=1 to k=3, representing a significant performance gain of 8.5%. However, beyond this point, the curve flattens into a plateau of diminishing returns. Increasing the retrieval depth from k=3 to k=5 yields only a marginal improvement of 1.7% (91.50% to 93.20%). This saturation pattern suggests that retrieving more than three documents adds minimal informational value while potentially introducing irrelevant noise. Consequently, k=3 is selected as the optimal cutoff for the generation phase, balancing high recall with context window efficiency.

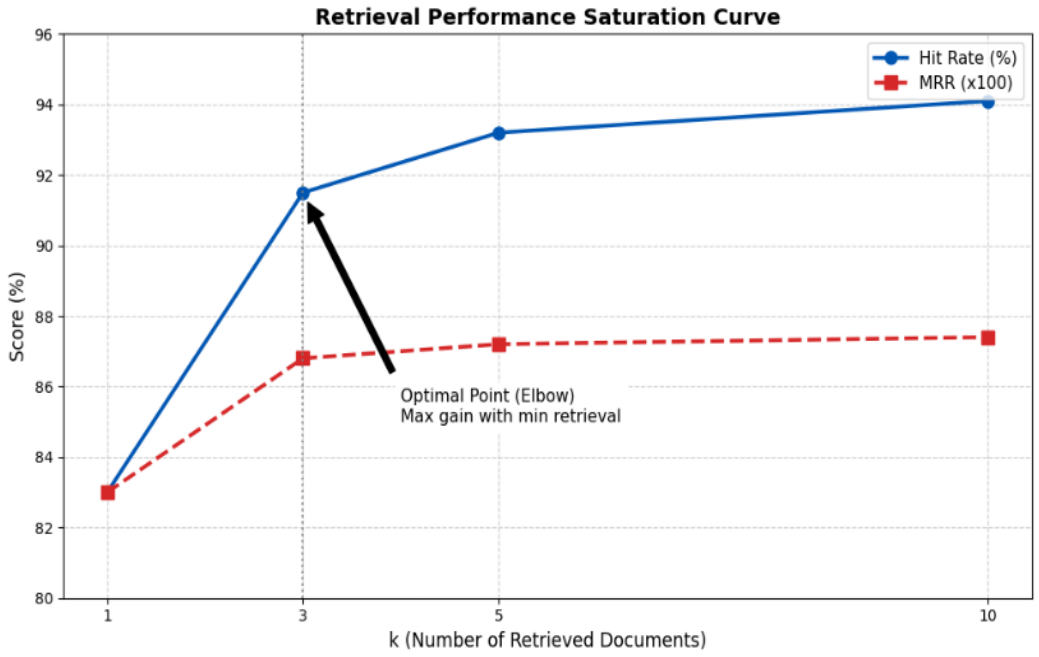


Figure 13. Retrieval Performance Saturation Curve

4.2.2 Generation Result

The generation quality was evaluated using the optimal retrieval setting k=3. using it into the Meta-Llama-3-70B Instruct Turbo. The model was tasked with answering queries.

Automatic Metrics Result

Seen in Table 10, the evaluation of the RAG system that uses the Meta, Llama model resulted in a BERT Score of 0.86. This metric quantifies semantic similarity, measuring how closely the meaning of the generated answer matches the reference text. A score at this level indicates a high degree of conceptual alignment, suggesting that the model captures the underlying intent and accuracy of the information, regardless of specific wording differences.

Regarding content coverage, the model recorded a ROUGE score of 0.56. This metric assesses content recall by analyzing the overlap of word sequences between the output and the reference. This result shows that the model retrieved and incorporated a majority of the key terms and significant details present in the reference text. It indicates that the core information is being preserved during the generation process.

The BLEU score was recorded at 29.28. This metric specifically measures exact word-for-word matching against the reference text. The score is lower than the semantic metrics, which reflects a divergence in phrasing. When viewed alongside the higher BERT score, this data suggests that while the model retains the correct meaning, it constructs sentences using different vocabulary and structure rather than strictly replicating the reference text verbatim.

Table 10: Automatic Metrics Results

Metric	Score	Range
BERT Score	0.86	0 – 1
ROUGE	0.56	0 – 1
BLEU	29.28	0 – 100

LLM-Judge Scores

To assess the semantic quality of the generated responses, the study employed Qwen 2.5 as an "LLM-as-a-Judge." Acting as an independent evaluator, Qwen 2.5 scored the outputs of the proposed systen configuration across three critical dimensions: Correctness, Faithfulness, and Completeness.

Seen in Table 10, the evaluation recorded a Faithfulness score of 4.85, the highest metric observed. This result indicates a strict adherence to the retrieved historical data and a minimal tendency to hallucinate external information. Regarding content coverage, the model achieved a Completeness score of 4.59, suggesting that the generated responses are thorough and successfully address the full scope of the user's inquiry without missing critical details. Finally, a Correctness score of 4.44 confirms that the system maintains a high standard of factual accuracy. Collectively, these results demonstrate that the system effectively synthesizes the provided context to produce answers that are both factually accurate and sufficiently comprehensive for the given query.

Table 11: LLM-As-Judge Results

Evaluation Metric	Score (1–5)
Faithfulness	4.85
Completeness	4.59
Correctness	4.44

Model Output Analysis

The following tables present the outputs generated by using the optimal retriever and Meta llama 70b. Each entry compares the original Search Query and the Expected Answer from the evaluation dataset against the Answer Generated by LLM. The results are stratified into three categories based on query complexity.

Table 12 displays Short Fact questions. These inquiries demand specific, precise entities such as names, dates, numbers, or short phrases requiring high-precision retrieval. It demonstrates the model's high accuracy on Short Fact questions, where the generated answers closely match the expected ground truth. The results show that the system effectively retrieves specific details, such as names (e.g., "Homo luzonensis") and complete lists without missing any information. Additionally, the model shows flexibility by presenting answers either as direct lists or within complete sentences. This confirms that the system is highly reliable for tasks that require precise, factual retrieval.

Table 13 presents Descriptive questions. These prompts require the model to define concepts or summarize characteristics, testing the system's ability to synthesize descriptive information. It shows that the reference answers are included in the generated answer. The model provides slightly more information than strictly asked for, but in a helpful way. the model answers in complete sentences rather than just using fragments, making the responses sound more natural and professional. This shows that the model isn't just extracting keywords; it is capturing the broader context of the document it retrieved.

Table 14 features Explanatory questions. These represent the most complex category, requiring the model to understand cause-and-effect relationships, comparisons, or procedural logic (e.g., "How" or "Why"). Table 3 highlights the model's strength in answering Explanatory questions. It successfully handles logical and often provides richer, more detailed answers than the reference by synthesizing information from multiple sources. However, the results also reveal a limitation: when a question requires abstract inferences such as interpreting the significance of an event the model tends to simply list the factual entities present rather than explaining the deeper context."

Table 15 highlights examples of undesirable outputs, illustrating specific cases or hallucinations encountered during evaluation. Table 4 highlights the system's failure modes, specifically regarding retrieval gaps and hallucinations. The most common error involves the model responding that "documents are not provided," which confirms a failure in the retrieval mechanism to locate relevant chunks, which is given since the retrieval’s hit rate doesn’t have a perfect hit rate. Additionally, there are discrepancies in response depth: questions requiring detailed explanations were occasionally met with short, factual answers. This likely stems from context noise, where the retrieved text was voluminous but substantively irrelevant, preventing the model from synthesizing a complete answer and leading to unintended stylistic variations.

Table 12: Short Factual Queries

Query	Expected Answer	Generated Answer
Where is Easter Island located	near Chile in South America	Easter Island is located near Chile in South America.
What species of Homo was established based on the findings in Callao Cave in Peñablanca, Cagayan?	Homo luzonensis	Homo luzonensis.
What should be done to counter claims made by natives after the lapse of the period prefixed for granting titles	They should be peremptorily disregarded	The counter claims the natives may seek to put in after the lapse of the period prefixed, should be peremptorily disregarded.
What were some of the key areas where Lopez Jaena identified faults and shortcomings of the Spanish colonial administration	education, economy, and colonial bureaucracy	Education, economy, and colonial bureaucracy.
What subjects are absent from the curriculum in public schools	Spanish grammar, physics, chemistry, geography, and agriculture	Spanish grammar, physics, chemistry, geography, and agriculture.

Table 13: Descriptive Queries

Query	Expected Answer	Generated Answer
What methods will be used to build a better country and life for the people?	peace, freedom, and the rule of law	The methods that will be used to build a better country and life for the people are through peace, freedom, and the rule of law, following the established Filipino way of life.
What geographical locations did the Austronesian languages expand to after originating in Taiwan?	the Philippines, Borneo, Sulawesi, Java, Sumatra, Malaya, and Oceania	The Austronesian languages expanded to the following geographical locations after originating in Taiwan: 1. The Philippines 2. Borneo 3. Sulawesi 4. Java 5. Sumatra 6. Malaya 7. Oceania (which includes islands such as those in the area of Easter Island near Chile in South America and Madagascar)

What requirements must be met for a search warrant or warrant of arrest to be issued	probable cause determined personally by the judge after examination under oath or affirmation of the complainant and the witnesses, and a particular description of the place to be searched and the persons or things to be seized	For a search warrant or warrant of arrest to be issued, the following requirements must be met: 1. There must be probable cause, 2. The probable cause must be determined by the judge (or such other responsible officer as may be authorized by law), 3. The determination of probable cause must be after examination under oath or affirmation of the complainant and the witnesses he may produce, and 4. The warrant must particularly describe the place to be searched for and the persons or things to be seized.

What rights are guaranteed to the accused in all criminal prosecutions?	the right to be heard by himself and counsel, to be informed of the nature and cause of the accusation, to have a speedy, impartial, and public trial, to meet the witnesses face to face, and to have compulsory process to secure the attendance of witnesses and the production of evidence	The rights guaranteed to the accused in all criminal prosecutions are: 1. To be heard by himself and counsel 2. To be informed of the nature and cause of the accusation against him 3. To have a speedy, impartial, and public trial 4. To meet the witnesses face to face 5. To have compulsory process to secure the attendance of witnesses and the production of evidence in his behalf Additionally, the accused shall be presumed innocent until the contrary is proved.
What is required for the objectives of the Local Government Code of 1991 to be realized	The proper use of devolved powers by local authorities	The proper use of devolved powers by local authorities is required for the objectives of the Local Government Code of 1991 to be realized.

Table 14: Explanatory Queries

Query	Expected Answer	Generated Answer
What happens to the funds certified by the Commission once they are approved?	The funds are released automatically upon certification by the Chairman of the Commission.	Once the funds certified by the Commission are approved, they shall be released automatically upon certification by the Chairman of the Commission.
What are the President's responsibilities regarding the executive branch and the execution of laws?	The President has control of all the executive departments, bureaus, and offices, and is responsible for ensuring that the laws are faithfully executed.	The President shall have control of all the executive departments, bureaus, and offices, and ensure that the laws be faithfully executed. Additionally, the President shall have supervision and control of all the ministries, bureaus or offices, all local governments, and all other branches or instrumentalities of the Executive Department and take care that the laws be faithfully executed.
What needs to happen for crime to come under full control?	Criminals, whether in or out of government, need to know that they will be caught, convicted, and jailed.	Crime can only come under full control when criminals—in or out of government—know we're going to catch them, convict them, and jail them.
What tasks will the national language commission be responsible for	The national language commission will undertake, coordinate, and promote researches for the development, propagation, and preservation of Filipino and other languages.	The national language commission shall undertake, coordinate, and promote researches for the development, propagation, and preservation of Filipino and other languages.
What notable event or figure is being referenced in the context of the Quirino grandstand?	The event is likely related to a significant moment in the fight for democracy, given the presence of President Cory Aquino, Cardinal Sin, and other religious leaders and fighters for democracy, suggesting a connection to a pivotal	President Cory Aquino, Cardinal Sin, and other religious leaders and fighters for democracy.

	gathering or celebration of democratic principles.	

Table 15: Failed Queries

Query	Reference Answer	Generated Answer
What is the approximate creation year of the Álbum de Filipinas	ca. 80	the information is not available in the provided documents.
What events led to the United States gaining control over the Philippine Islands, and what were the subsequent actions taken by the US government?	The destruction of the Spanish fleet in the harbor of Manila by the United States naval squadron commanded by Rear-Admiral Dewey, followed by the reduction of the city and the surrender of the Spanish forces, led to the US gaining control over the Philippine Islands. The subsequent actions taken by the US government included the extension of military government to the whole of the ceded territory, with the military commander enjoined to make known to the inhabitants of the Philippine Islands that the authority of the United States was to be exerted, and the assumption of responsibility for the government of the islands, as a result of the treaty of peace between the United States and Spain signed at Paris.	The information is not available in the provided documents.. The context provided discusses the actions taken by the US government after gaining control over the Philippine Islands, such as levying duties and taxes, regulating commercial intercourse, and eventually recognizing the independence of the Philippine Islands. However, it does not mention the events that led to the US gaining control over the Philippine Islands.
Where was the speech delivered by Manuel Quezon"	Grand Opera House, Manila	The information is not available in the provided documents.

What country was being liberated by forces and key officials in the Philippine Commonwealth	the Philippines	The information is not available in the provided documents.
What year is mentioned in the report of His Excellency President Marcos	2045	The information is not available in the provided documents

4.3 COMPARATIVE ANALYSIS

4.3.1 Retrieval

The retrieval system was evaluated using four distinct configurations to isolate the impact of Retrieval Technique, Table 10 presents the Hit Rate (%) and Mean Reciprocal Rank (MRR) for each configuration across varying retrieval depths (k=1, 3, 5, 10).

As indicated in Table 16, the Baseline configuration demonstrated the lowest retrieval sensitivity. At k=1, it recorded a Hit Rate of 55.20%, indicating that the correct document was the top result in roughly half of the queries. As the retrieval depth increased to k=10, the Hit Rate improved to 80.32%, while the MRR showed only moderate growth, stabilizing at 0.635. This suggests that the standalone dense retrieval method struggled to capture precise entity matches (e.g., dates and names), often burying relevant documents lower in the list due to vector compression.

The transition to Retrieval-V2 marked the most significant performance shift in the experiment. By implementing Hybrid Chunking, the Hit Rate at k=1 surged to 81.50%, surpassing the Baseline's best performance (k=10) with just a single retrieved chunk. Crucially, the MRR saw a massive correction, rising from 0.552 (Baseline) to 0.815, reflecting that when the system found the correct context, it was overwhelmingly likely to place it at Rank 1.

Retrieval-V3, which upgraded the embedding model to Gte Modernbert Base, yielded consistent incremental gains. The Hit Rate at k=1 rose to 82.83%, and the system achieved 91.50% at k=10. The MRR improved to 0.828 at k=1, demonstrating that the more advanced

embedding model captured semantic nuances more effectively than the standard embeddings used in V2, resulting in better vector separation for complex queries.

The final proposed configuration achieved the highest performance across all metrics. While the Hit Rate at k=1 showed a marginal improvement to 83.00%, the impact of the reranker was most visible at k=3. The Hit Rate jumped significantly to 91.50%, with a corresponding MRR of 0.868. This specific behavior indicates that the reranker successfully cleaned up the top results, promoting relevant documents that were previously buried at ranks 4 or 5 into the top 3 optimal window.

Based on these performance trajectories, the proposed at k=3 was identified as the optimal configuration for the subsequent generation stage. While increasing retrieval depth from k=3 to k=5 provided a marginal gain in Hit Rate (+1.7%), it would necessitate processing significantly more tokens. The performance curve demonstrates a clear point of diminishing returns at k=3, where the system achieves a high Hit Rate (91.50%) and a strong MRR (0.868). Consequently, k=3 was selected to balance retrieval effectiveness with computational efficiency, ensuring the language model receives sufficient context without being overwhelmed by excessive input.

Table 16: Retriever Variants Results

Configuration	Metric	k=1	k=3	k=5	k=10
Baseline (Sparse)	Hit Rate (%)	55.20	70.71	74.20	80.32
	MRR	0.552	0.618	0.625	0.635
Retrieval-V2 (Hybrid)	Hit Rate (%)	81.50	83.90	86.10	89.00
	MRR	0.815	0.824	0.829	0.833
Retrieval-V3	Hit Rate (%)	82.83	86.50	89.20	91.50

	MRR	0.828	0.841	0.848	0.852
Proposed	Hit Rate (%)	83.00	**91.50**	93.20	94.10
	MRR	0.830	**0.868**	0.872	0.874

4.3.2 Generation

To provide an objective assessment of the generated outputs, the system was evaluated using three standard automatic metrics: BLEU (n-gram precision), ROUGE-L (structural recall), and BERTScore (semantic similarity). Table 17 presents the comparative performance of the three language models when coupled with the optimal k of the proposed retriever.

Table 17: Generation Automatic Scores

Variant	Bleu	Rouge	Bert Score
Mistral-Instruct 7B	21.8	0.35	0.79
GPT OSS 20B	16.8	0.55	0.82
Meta Llama 70B	**29.28**	**0.56**	**0.86**

The Mistral-Instruct variant demonstrated a reliance on surface-level lexical matching. While it achieved a moderate BLEU score of 21.80, indicating a capability to reproduce exact phrases from the reference text, its performance in semantic dimensions was limited. The model recorded

the lowest ROUGE score (0.35) and BERTScore (0.79) among the cohorts. This disparity suggests that while the model effectively identifies specific keywords, it struggles to construct comprehensive narratives, often failing to capture the broader semantic context required for high-fidelity historical explanations.

Conversely, the OpenAI GPT-OSS 20B variant exhibited a divergent performance profile characterized by strong semantic recall but high lexical deviation. The model achieved a BLEU score of only 16.80, the lowest in the evaluation, yet recorded a competitive ROUGE score of 0.55 and a BERTScore of 0.82. This inverse relationship reflects an abstractive generative behavior; rather than prioritizing rote replication of the source text, the model tends to paraphrase content. This indicates a capacity to synthesize the underlying factual core of the reference answers while employing distinct phrasing and sentence structures.

The Meta Llama 3-70B variant demonstrated superior performance across all evaluated dimensions, establishing it as the most robust model for this application. With a BLEU score of 29.28, the model exhibited a strong ability to adhere to the precise terminology of the reference text. Crucially, this lexical precision was matched by deep semantic alignment, evidenced by a BERTScore of 0.86 and a ROUGE score of 0.56. These results confirm that the generated responses effectively encompass the full context and nuance of the gold-standard references, balancing exactness with comprehensive reasoning.

To understand the performance disparities observed in the automatic metrics, we analyzed the distinct generation behaviors of the three models, revealing a clear dichotomy between Surface-Level Extraction and Abstractive Paraphrasing seen in figure 13.

The Mistral-Instruct (7B) variant exhibited the characteristics of surface-level extraction. While it achieved a moderate BLEU score of 21.80 higher than the mid-sized GPT-OSS. Its significantly lower BERTScore (0.79) implies that this lexical precision was superficial. The model successfully copied specific keywords from the source text but failed to assemble them into a cohesive narrative. This phenomenon of "hollow precision” where keywords are present but semantically disjointed resulted in the lowest reliability profile among the tested configurations, producing answers that were grammatically correct but lacked deep contextual understanding.

In contrast, the OpenAI GPT-OSS (20B) variant displayed a strong tendency toward abstractive generation. This model created a "Paraphrasing Gap" where it achieved the lowest exact-match score (BLEU 16.80) but a highly competitive semantic validity score (BERTScore

0.82). Qualitatively, this indicates that instead of regurgitating the retrieved text verbatim, the model reformulates the information using distinct vocabulary. Although this approach penalizes the model on strict n-gram metrics like BLEU, the high BERTScore confirms that the underlying factual logic is preserved, effectively trading lexical precision for more natural, human-like phrasing.

The Meta-Llama-3 (70B) configuration effectively bridged this gap, demonstrating a "Dual Competency" likely attributable to its massive parameter scale. By simultaneously achieving the highest BLEU (29.28) and BERTScore (0.86), the model proved it could retain precise terminology from the source documents essential for historical accuracy while synthesizing them into a coherent, semantically robust narrative. This suggests that larger parameter models do not strictly tradeoff between creativity and accuracy but rather expand the capability frontier to maximize both extraction and abstraction simultaneously.

These findings highlight the essential trade-offs inherent in model selection. The superior semantic performance of Meta-Llama-3 comes at the cost of significantly higher computational overhead and inference latency compared to the 7B and 20B variants. Ultimately, this analysis underscores that in the context of RAG, BERTScore serves as the far more significant indicator of utility than BLEU. Since valid answers can vary in phrasing, the ability to comprehend underlying meaning (as demonstrated by Llama-3 and GPT-OSS) is more valuable than the simple keyword matching observed in smaller models like Mistral.

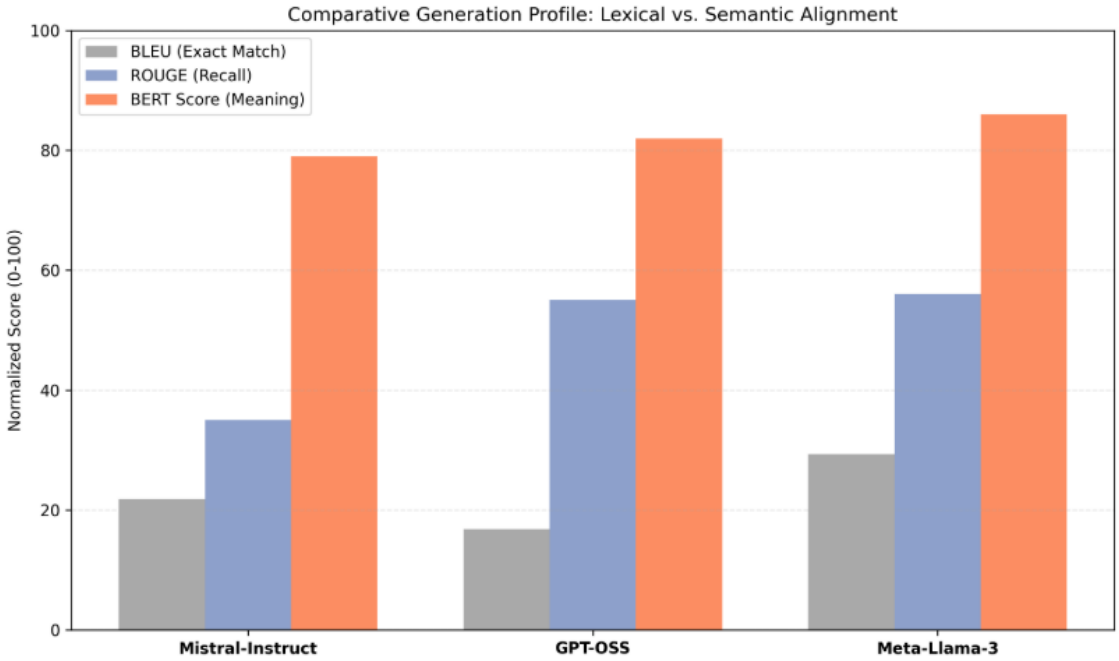


Figure 14. Comparative Generation Profile: Lexical vs. Semantic Alignment

To assess the semantic quality of the generated responses, the study employed Qwen 2.5 as the "LLM-as-a-Judge." Acting as an independent evaluator, Qwen 2.5 scored each model's outputs across three critical dimensions: Correctness, Faithfulness, and Completeness. This method provides a nuanced analysis of how well the models understood and utilized the retrieved context The results are shown in Table 18.

Mistral-Instruct achieved a Correctness score of 4.29 and a Completeness score of 3.95. However, it recorded a Faithfulness rating of 3.82, which was the lowest among the configurations tested. This performance profile indicates that while the model is generally capable of identifying relevant information, it struggles to strictly adhere to the provided context. The lower faithfulness score suggests a tendency to drift from the source material, leading to reduced reliability regarding historical accuracy.

Using The GPT OSS variant demonstrated strong semantic performance, recording a Faithfulness score of 4.46 and a Correctness score of 4.38. Its Completeness score stood at 4.23. These metrics reflect a model that produces factually accurate and grounded responses. The high faithfulness rating indicates that the model effectively limits its

generation to the retrieved context, ensuring that its answers remain faithful to the source text without significant deviations or hallucinations.

While Meta Llama (Llama-3-70B) exhibited superior performance in the Qwen 2.5 evaluation, achieving a Correctness score of 4.44 and a Completeness score of 4.59. Most notably, it attained a near-perfect Faithfulness score of 4.85. This exceptionally high rating demonstrates the model's robust ability to synthesize historical data without hallucinating external information. The combination of high correctness and faithfulness suggests this configuration is highly reliable and particularly well-suited for educational applications where accuracy is paramount.

While the automatic metrics provided a baseline for lexical performance, the LLM Judge evaluation revealed critical disparities in the trustworthiness and contextual adherence of the generated responses. The most significant finding was the divergence between "Correctness" and "Faithfulness" in the smaller models. The Mistral-Instruct variant, despite achieving a respectable Correctness score of 4.29, recorded a significantly lower Faithfulness rating of 3.82. This gap suggests a "hallucination risk," where the model frequently drifts from the provided context to import external knowledge. While the answers remained factually plausible, this lack of strict adherence undermines the utility of the RAG system in domains requiring precise evidence citation, as the model prioritized answering the question over respecting the retrieval boundaries.

In contrast, the Meta-Llama-3 configuration demonstrated a near-perfect alignment between retrieval and generation. With a Faithfulness score of 4.85, the model exhibited a robust capacity for "negative constraint the ability to limit its response strictly to the retrieved facts and refuse to hallucinate missing information. Unlike Mistral, which often truncated answers (Completeness 3.95), and GPT-OSS, which focused on broad semantic abstractions, Llama-3 successfully synthesized comprehensive answers (Completeness 4.59) without sacrificing source fidelity. This indicates that larger parameter models do not merely improve fluency but fundamentally alter the reliability profile of the system, shifting the behavior from "creative generation" to "evidence-based synthesis. A comparison of each can be seen in Figure 15.

Table 18: LLM Judge Criteria Results

Variant	Correctness	Faithfulness	Completeness
Mistral-Instruct	4.29	3.82	3.95
GPT OSS	4.38	4.46	4.23
Meta LLama	4.44	4.85	4.59

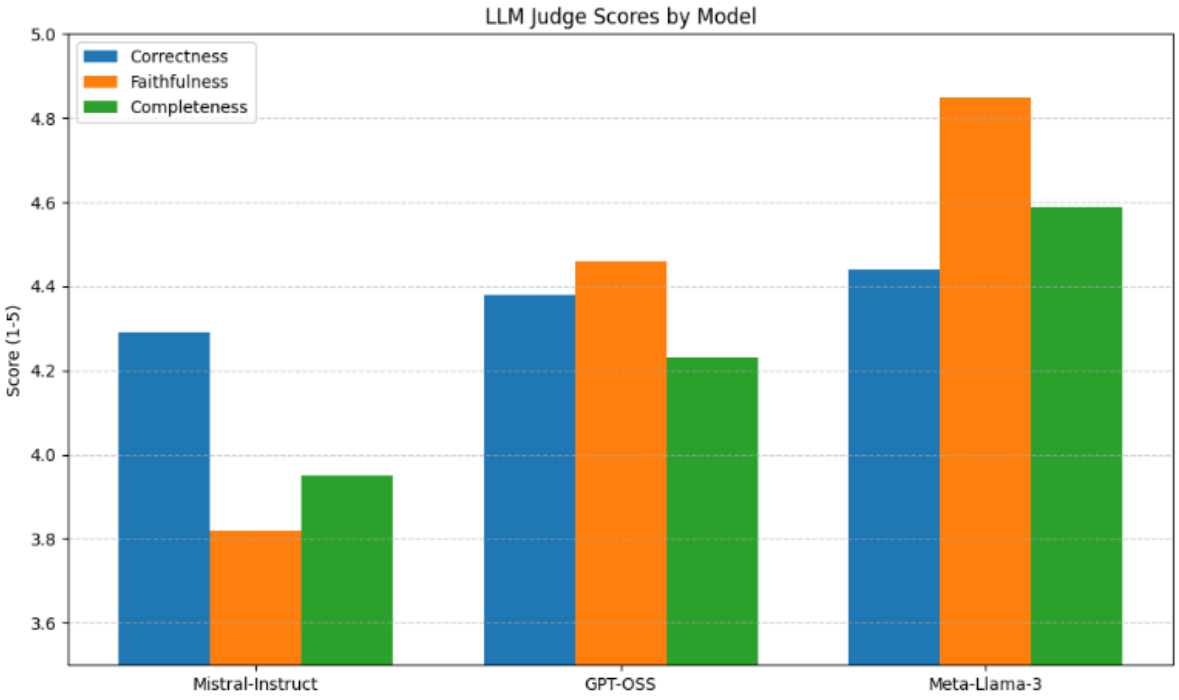


Figure 15. LLM Judge Scores by Model

4.4 DISCUSSION OF FINDINGS

The results demonstrate that the proposed retrieval-augmented generation (RAG) architecture achieves strong performance through the careful alignment of retrieval depth, embedding quality, reranking, and model scale. Quantitative evaluation of the final retrieval configuration reveals a system characterized by high early precision and a clear performance plateau. At k=1, the system already achieves a Hit Rate of 83.00% (MRR 0.830), validating the effectiveness of the Reranker. Expanding the retrieval depth to k=3 produces a substantial gain, increasing the Hit Rate to 91.50%, which represents the system's elbow point. Beyond this threshold, additional context yields diminishing returns, with only marginal improvement at k=5. This confirms that k=3 provides the optimal balance between retrieval effectiveness and computational efficiency, minimizing contextual noise while preserving sufficient evidence for generation. These findings highlight the inherent trade-offs in RAG system configuration, while a larger retrieval window (theoretically increases recall, it inadvertently introduces noise that can degrade generation quality. Conversely, a window that is too narrow risks missing the relevant document entirely. For the data, the experimentally derived $k = 3$ threshold effectively navigates this dichotomy, securing high recall while maintaining the contextual focus required for accurate generation.

A comparative analysis of the retrieval configurations highlights the critical impact of architectural enhancements. The Baseline (Dense-only) retriever, utilizing all-MiniLM-L6-v2 embeddings, exhibited limited sensitivity to precise entities, frequently ranking relevant documents lower in the candidate list due to vector compression.

The introduction of Hybrid Retrieval in Retrieval-V2 produced the most dramatic performance shift. By integrating sparse lexical signals (BM25) with semantic similarity, the system achieved significant gains in both Hit Rate and MRR. This empirical evidence suggests that traditional keyword matching remains indispensable for domain-specific retrieval. By combining the precision of sparse lexical matching with the contextual understanding of dense embeddings, the architecture effectively bridges the gap between exact-term lookups and broad conceptual search, ensuring that queries containing specific dates or proper nouns are not lost in the vector space. Retrieval-V3 delivered additional gains through improved embedding quality using Alibaba-modernbert case , demonstrating stronger discrimination for complex queries. The final integration of a BGE reranker in Retrieval-V4 proved decisive, particularly at k=3, where it consistently promoted relevant documents into the top ranks. These findings confirm that retrieval

quality is not driven by any single component but by the cumulative effect of hybrid retrieval, stronger embeddings, and post-retrieval reranking.

On the generation side, the evaluation of three models using the final retrieval configuration revealed distinct performance profiles. Mistral-Instruct (7B) demonstrated characteristics of surface-level extraction, achieving moderate lexical overlap but lower semantic alignment. In contrast, OpenAI GPT-OSS (20B) exhibited a propensity for abstractive paraphrasing, trading exact lexical matching for enhanced flow and factual synthesis. Meta-Llama-3 (70B) emerged as the superior model, demonstrating a "dual capability" that combined high lexical fidelity with deep semantic alignment. This suggests that increased model scale enables the simultaneous optimization of extraction and abstraction, allowing the system to retain precise terminology without sacrificing narrative coherence. These findings fundamentally challenge the reliance on traditional n-gram metrics for RAG evaluation. The analysis highlights that BERTScore is a far more reliable indicator of performance than ROUGE or BLEU. Because Large Language Models are inherently stochastic and capable of generating valid answers with diverse phrasing, rigid lexical matching often penalizes high-quality, natural-sounding responses. Therefore, semantic evaluation metrics must be prioritized as evaluating RAG systems.

The limitations of purely automatic metrics were further exposed through LLM-as-a-Judge evaluation using Qwen 2.5. While smaller models achieved acceptable correctness scores, they showed notable gaps in faithfulness, indicating a tendency to drift beyond the retrieved evidence. In contrast, Meta-Llama-3 achieved near-perfect faithfulness alongside high correctness and completeness, demonstrating robust adherence to retrieved historical context with minimal hallucination. This alignment between retrieval and generation reflects a shift from creative text generation toward evidence-based synthesis, a critical requirement for educational and historical applications.

Stratified output analysis of the RAG system using Meta-Llama 70 b outputs reinforces these findings across varying levels of query complexity. The system performs reliably on short factual and descriptive queries, consistently retrieving and presenting accurate information with flexible formatting. For explanatory questions, the model demonstrates strong synthesis by integrating information from multiple sources, though limitations persist when abstract reasoning or interpretive analysis is required. Identified failure modes primarily stem from retrieval gaps and context noise, underscoring the dependency of generation quality on retrieval precision rather

than language model shortcomings alone, and that it may further hallucinate if the retrieval system is not properly configured for the purpose.

The final RAG system was integrated into the chatbot QUBO. Field experts comprising AI specialists, software engineers, and a history expert were invited to participate in the system demonstration. In total, 10 experts took part in the evaluation, with the distribution of expertise illustrated in Figure 16. Following the demonstration, participants were asked to complete a survey consisting of 10 Likert-scale items ranging from 1 - 5 and a qualitative question to gather both quantitative evaluations and qualitative feedback. The result of The Likert survey scale is seen Table 19, while the Quantitative suggestions by experts are seen in Table 20.

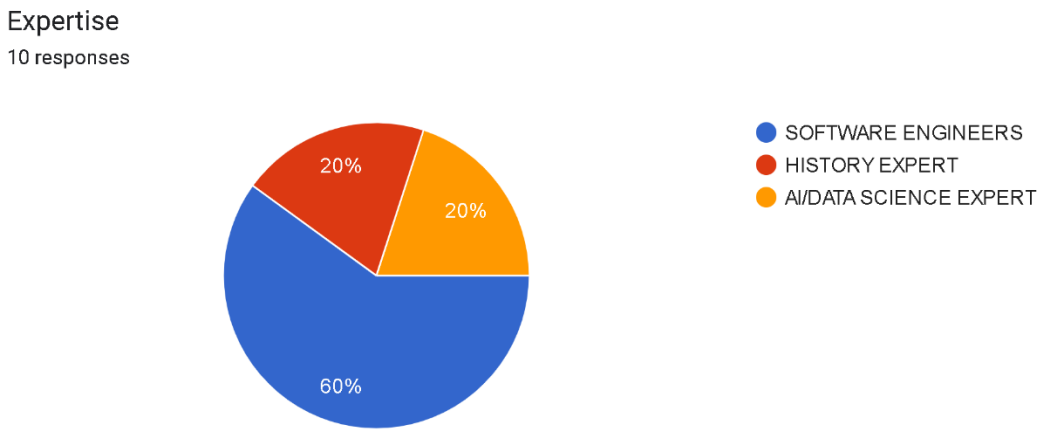


Figure 16. Experts Distribution

Table 19: Expert Validation Report

Question	Question Type	Weighted Mean	Verbal Interpretation
The system successfully processes uploaded documents without issues	General	4.7	Strongly Agree
The chatbot answers questions strictly using the information from the provided files	General	4.7	Strongly Agree

The responses are relevant and directly address the user's questions	General	4.3	Agree
The system properly refuses to answer questions not related to the uploaded document	General	4.3	Agree
The interface is simple and easy to navigate	General	4.8	Strongly Agree
The system responds to questions at a reasonable speed	General	4.2	Agree
The tool is suitable for users with little technical background s	General	4.0	Agree
The system effectively prevents the AI from making up information ("hallucinating")	AI Expert	5.00	Strongly Agree
The system consistently finds the correct information within the document.	AI Expert	4.50	Agree
The AI synthesizes information well, creating answers that are coherent and easy to read.	AI Expert	4.50	Agree
The application runs smoothly and does not crash during standard operation.	Software Engineers	4.33	Agree
The system handles tasks (like uploading files or generating answers) without excessive delays or lag.	Software Engineers	4.5	Agree
The system handles user inputs and file uploads securely and correctly.	Software Engineers	4.5	Agree
The chatbot's answers are faithful to the historical text provided.	History Expert	3.50	Moderately Agree

This tool helps reduce the time needed to analyze historical documents.	History Expert	3.00	Moderately Agree
The system may help students understand historical texts better.	History Expert	3.00	Moderately Agree
Average Score		**4.23**	

Table 20: Expert Suggestions

Expertise	**Suggestions**
AI Expert	To improve scalability, the system should be optimized to handle performance more efficiently as usage grows. This includes experimenting with different strictness levels to observe how changes affect accuracy, response time, and overall performance. In addition, implementing hard limits on document size or page count is necessary, since very large or dense files can slow down processing and negatively impact both accuracy and query speed.
Software Engineer	The system is generally functional and smooth, but several improvements can be made to enhance usability and performance. These include adding better user validation, clearer and more contextual error messages, and improving the UI/UX especially for file upload progress, citation highlighting, and input form validation. File upload speed can be further optimized by setting file size or count limits, and file management can be improved through folders and tags. Making the system accessible via a public URL instead of just localhost and validating queries before calling the API would also help optimize performance. Additionally, the system should clarify how it handles conflicting information from multiple sources. Overall, the project is already a strong undergraduate thesis, but these refinements would make it more robust and user-friendly.
History Espert	The application should be developed to function fully offline after installation, ensuring that all responses are generated strictly from the uploaded PDF

	documents. This would prevent unsupported or hallucinated answers and make the system more reliable, especially in environments without internet access. Further improvements should also be made to fix analysis issues so that document processing and answer generation become more accurate and consistent.

Overall, expert feedback was predominantly positive, particularly regarding system usability and functional reliability. Recommendations primarily focused on improving scalability, deployment flexibility, and response personalization.

Beyond the technical benchmarks, this study demonstrates the viability of utilizing open-source Artificial Intelligence for preserving Philippine cultural heritage. By achieving a BERTScore of 0.86 using open-weights models (Meta-Llama-3), we establish that high-fidelity historical question-answering systems can be built without reliance on costly, proprietary APIs. for educational institutions in the Philippines, where resource constraints often hinder the adoption of advanced technology. The findings suggest that the system may be converted into a fully local deployed RAG system in the future which would democratizes access to historical knowledge, allowing students to query complex narratives such as the causes of the Philippine Revolution and receive evidence-based answers instantly. This technological shift encourages a transition from the rote memorization of dates to a deeper, inquisitive engagement with the nation's history.

While Meta-Llama-3 (70B) proved to be the superior reasoning engine, as we used as hosting service for it, if it is to be deployed locally, it important to note that it highlights a critical trade-off between reasoning accuracy and deployment feasibility. The 70B parameter model requires significant computational resources (approximately 40GB+ of VRAM even with quantization) and yields slower inference speeds, rendering it impractical for local deployment on standard classroom hardware. In contrast, the Mistral-Instruct (7B) variant, while less semantically robust, offers a lightweight profile that is highly feasible for local, real-time inference. This creates a divergence in deployment strategy: Llama-3 is suitable for server-hosted, high-accuracy research tools, while Mistral is viable for offline, interactive classroom assistants. Furthermore, the current architecture primarily addresses single-hop queries; the system may still struggle with multi-hop reasoning that requires synthesizing conflicting accounts from distinct historical sources

(e.g., resolving the differing dates of the *Cry of Balintawak*). Future iterations must balance these computational demands with the need for sophisticated, multi-document synthesis.

4.5 SUMMARY OF THE CHAPTER

This chapter presented the experimental design, implementation, and rigorous evaluation of the proposed Retrieval-Augmented Generation (RAG) system tailored for Philippine cultural history. The study established a controlled testing environment utilizing a manually constructed "Gold Standard" dataset of 198 question-answer pairs to benchmark performance across various retrieval configurations and Large Language Models.

The retrieval evaluation demonstrated the superiority of the Proposed System, which integrates Hybrid Search (Dense + Sparse) with ModernBERT embeddings and Cross-Encoder Reranking. This configuration achieved a Hit Rate of 91.50% and a Mean Reciprocal Rank (MRR) of 0.868 at an optimal retrieval depth of $k = 3$. These findings confirm that hybrid architecture effectively resolves the limitations of standard dense retrieval by bridging the gap between semantic understanding and precise keyword matching, while reranking acts as a critical filter to maximize context relevance.

In the generation phase, comparative analysis revealed a direct correlation between model scale and reasoning fidelity. The Meta-Llama-3 (70B) model emerged as the optimal reasoning engine, exhibiting "dual competency" by achieving the highest lexical precision and semantic alignment (BERTScore 0.86). Automated evaluation via LLM-as-a-Judge further validated its performance, awarding it a near-perfect Faithfulness score of 4.85, indicating a robust ability to synthesize historical facts without hallucination capability lacking in smaller models like Mistral-Instruct (7B) and GPT-OSS (20B).

Finally, the system underwent expert validation involving AI specialists, software engineers, a history expert. The evaluation yielded a high average acceptability score of 4.23, confirming the tool's usability and relevance, though feedback highlighted the necessity for offline optimization. The chapter concludes by defining the critical trade-off between accuracy and deployment feasibility: while large-scale models offer research-grade fidelity, they impose significant computational demands, suggesting that future iterations must balance semantic depth with the hardware constraints of local educational institutions.

CHAPTER 5
SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

5.1 SUMMARY OF FINDINGS

This study examined the effectiveness of a Retrieval-Augmented Generation (RAG) system for document-grounded question answering in Philippine cultural history through a combination of quantitative experiments, automated evaluation metrics, qualitative output analysis, and expert review. The findings provide a cohesive view of how retrieval strategies, model selection, and system design decisions influence the accuracy, reliability, and practical usability of RAG-based historical question answering systems.

The experimental results demonstrate that retrieval configuration plays a decisive role in determining system performance. The baseline dense-only retrieval approach was able to surface relevant documents in many cases, but ranking inconsistencies frequently limited answer quality. In contrast, the optimized retrieval pipeline integrating hybrid dense–sparse retrieval, higher-capacity embeddings, cross-encoder reranking, and intent routing consistently surfaced the most relevant historical sources at the top of the retrieval list. This improvement in retrieval precision translated directly into more accurate, faithful, and contextually grounded answers, reinforcing the central importance of retrieval optimization in RAG systems.

Analysis of retrieval depth further revealed that increasing the number of retrieved documents does not linearly improve performance. While higher values of k increased recall, the gains beyond $k = 3$ were minimal and came at the cost of additional computation and increased context noise. Selecting an optimal retrieval depth proved essential in maintaining a balance between efficiency and answer quality, highlighting that stricter retrieval settings can often yield better outcomes than broader but less focused document inclusion.

With retrieval optimized, differences in generative model performance became more pronounced. Larger language models demonstrated a stronger ability to synthesize historically grounded responses that were both semantically accurate and clearly expressed. Meta Llama-3-70B consistently produced outputs that aligned closely with reference answers, achieving higher lexical and semantic evaluation scores and stronger qualitative judgments of faithfulness and correctness. These results suggest that model capacity matters most once reliable contextual grounding is ensured, although this benefit must be weighed against increased computational cost.

Qualitative analysis of generated answers showed that the system performed particularly well on factual and descriptive questions, accurately reproducing key entities such as names, dates, and events. For explanatory questions, the system was generally able to construct coherent multi-sentence responses that reflected causal and procedural relationships found in the source documents. When failures occurred, they were most often traced back to missing or insufficient retrieval rather than weaknesses in the language model itself. In such cases, the system appropriately refrained from generating unsupported answers, demonstrating effective hallucination mitigation through strict document grounding.

The integration of the optimized RAG system into the Qubo chatbot allowed for evaluation of its real-world applicability. Expert reviewers expressed strong overall approval, as reflected in a mean Likert score of 4.23, citing the system's historical accuracy, responsiveness, and educational relevance. At the same time, expert feedback identified practical areas for refinement, particularly in terms of scalability and usability. Suggestions included improving vector database efficiency for larger corpora, enforcing document size and page limits, strengthening query clarification mechanisms, and enhancing user interface elements such as upload feedback and citation visibility. Experts also emphasized the need for clearer handling of conflicting historical sources to improve transparency and user trust.

From a historical and scholarly perspective, reviewers highlighted the importance of offline functionality and strict reliance on uploaded documents. Ensuring that all responses are generated exclusively from verified historical texts was seen as essential in minimizing hallucinations and maintaining consistency, especially in environments with limited internet connectivity.

Taken together, the findings indicate that the proposed RAG architecture fulfills the objectives of the study. By combining optimized hybrid retrieval, reranking, question intent routing, and a high-capacity language model, the system demonstrates robust document-grounded question answering capabilities for Philippine cultural history. The results underscore the strong relationship between retrieval quality and answer faithfulness, while also showing that strict grounding and abstention from unsupported responses are key strengths for educational and scholarly applications where accuracy and evidence-based reasoning are paramount.

5.2 CONCLUSIONS

Based on the experimental evaluation and expert validation conducted in this study, it is concluded that Retrieval-Augmented Generation (RAG) provides a viable mechanism for grounding large language model outputs in external, domain-specific data sources. The integration of document retrieval into the generation process reduced the occurrence of unsupported content and improved source attribution. When applied to a corpus of Philippine cultural history documents, the developed system demonstrated the ability to retrieve relevant contextual information and generate responses aligned with the provided sources, along with refusing answers that cannot be answered,

evaluation results indicate that grounding generative responses in verified external documents contributes to improved factual consistency. Across the evaluated queries, the system exhibited a consistent capacity to retrieve appropriate context and generate responses that remained aligned with the retrieved material. These findings suggest that RAG-based architecture can enhance the reliability of domain-focused question-answering systems when compared to standalone generative approaches.

The study further emphasizes the importance of iterative system optimization within RAG pipelines. Comparative analysis across multiple configurations revealed that retrieval effectiveness and response quality are highly dependent on the selection of retrieval strategies, embedding models, and language model architectures. Despite overall improvements, occasionally undesirable outputs were still observed, which were primarily attributed to retrieval failures. These observations highlight potential areas for future enhancement. System performance was found to vary across configurations, indicating that RAG systems require careful design, tuning, and systematic evaluation to achieve stable and reproducible results.

Additionally, the results highlight a trade-off between system performance and computational cost. While the adoption of higher-capacity language models and advanced retrieval techniques led to improvements in response quality, these enhancements were accompanied by increased inference latency and operational overhead. In particular, the incorporation of a reranking model improved retrieval precision but introduced additional computational cost. Similarly, the use of large-scale models such as Meta Llama 70B demonstrated strong qualitative performance; however, such models may pose challenges for long-term deployment in resource-constrained environments. These findings underscore the

need to balance system capability with efficiency considerations in practical RAG implementations.

Expert validation provided insights into the applicability of the proposed system. Reviewer feedback indicated that the system demonstrates potential for educational and research-oriented use, while also identifying areas requiring further refinement. These observations support the conclusion that AI-assisted tools, when implemented with appropriate grounding and evaluation mechanisms, can function as supplementary resources for academic inquiry rather than authoritative sources.

Overall, this research provides foundational evidence supporting the application of Retrieval-Augmented Generation for domain-specific question answering in Philippine cultural history. The findings contribute to ongoing efforts to improve access to historical information through structured and verifiable AI-assisted systems. Continued research on the advancement of Retrieval-Augmented Generation techniques is expected to further enhance the reliability, scalability, and practical applicability of such systems.

5.3 RECOMMENDATIONS

Future research should consider extending the proposed Retrieval-Augmented Generation (RAG) system to address the limitations of the current dataset, which primarily relies on single-hop questions. While the present implementation effectively retrieves individual documents for direct queries, complex historical questions often require synthesizing information from multiple sources. The adoption of multi-hop reasoning architectures, such as GraphRAG or Chain-of-Thought (CoT) retrieval methods, would enable the system to compare, contrast, and integrate narratives from separate historical accounts, thereby supporting deeper analytical inquiry.

Further investigation is also recommended to address the trade-off between model accuracy and computational efficiency. This study observed performance disparities between large-scale models with superior reasoning capabilities and smaller models optimized for latency. Future work could explore the fine-tuning and quantization of mid-sized language models specifically on Philippine historical datasets. Techniques such as Quantized Low-Rank Adaptation (QLoRA) could allow these models to retain strong reasoning performance while remaining viable for deployment in resource-constrained environments, such as standard classroom hardware.

In addition to textual retrieval, future systems would benefit from expanding toward multimodal retrieval capabilities. Philippine cultural heritage encompasses not only written records but also visual materials, including historical photographs, maps, architectural blueprints, and artifacts. Incorporating image-aware retrieval and generation mechanisms would enable the system to answer visually grounded questions, significantly enriching the learning experience and supporting diverse forms of historical engagement.

To enhance educational applicability, future iterations should integrate adaptive difficulty mechanisms via user modeling. By identifying a user's academic background or proficiency level, the system could dynamically adjust the complexity, vocabulary, and length of its responses. This adaptation would improve usability across a wide demographic from elementary students to advanced researchers without compromising informational accuracy.

A critical direction for inclusivity involves expanding evaluation datasets and knowledge sources to include vernacular and regional languages. As the current study relied heavily on English texts, accessibility for broader audiences may be limited. Developing benchmark datasets in Filipino (Tagalog) and other regional languages is essential for evaluating cross-lingual retrieval performance and ensuring equitable access to historical knowledge. Relatedly, future research should explore code-switching aware RAG systems tailored to mixed-language usage (e.g., Taglish). By integrating cross-lingual retrieval layers, the system could interpret informal queries, retrieve relevant English or Spanish sources, and generate responses in an accessible, mixed-language format.

Future deployments may introduce transparency and trustworthiness through fine-grained citation mechanisms. Rather than listing references at the end of a response, the system should directly link specific claims to their originating passages within retrieved documents. This functionality would facilitate fact verification, help mitigate hallucination risks, and reinforce the system's utility as a reliable academic tool.

REFERENCES

[1] The Philippines' History Curriculum: Origins and Repercussions - The Peninsula Foundation, accessed on October 18, 2025, https://www.thepeninsula.org.in/2023/12/03/the-philippines-history-curriculum-origins-and-repercussions/

[2] [2] Gadaza, A. Manera, S. Santos, C. Alih, and R. Caban, “Reviving the Past, Teaching the Future: The Role of Philippine Cultural Heritage in Curriculum Development of Teacher Education Programs Focus,” International Journal on Culture, History, and Religion, vol. 7, no. SI2, pp. 80–97, Jul. 2025, doi: https://doi.org/10.63931/ijchr.v7isi2.169.

[3] Aniza Gadaza, A. Manera, R. Caban, C. Alih, Alnadzma Tulawie, and H. Picpican, “Cultural Identity and Historical Consciousness: A Study of Philippine History Instruction in Tertiary Education,” vol. 7, no. SI2, pp. 19–35, Jul. 2025, doi: https://doi.org/10.63931/ijchr.v7isi2.135.

[4] Instructure, 2025 State of Higher Education Report (Philippines: AI usage among students and educators), Instructure Holdings, Inc., 2025. [Online]. Available: https://www.bworldonline.com/technology/2025/06/12/678582/more-filipino-students-now-using-ai-for-learning/

[5] J. Bryan Sadiasa, J. Ordoñez, F. Ivan Pinar, and B. Soriaso, “The Rise of AI in Education: Exploring the Impact of Perception of Large Language Models in the Critical Thinking and Self-efficacy of Health-Allied Senior High School Students.” Available:

[6] Sowmya Vajjala, Bashar Alhafni, Stefano Bannò, Kaushal Kumar Maurya, and Ekaterina Kochmar, “Opportunities and Challenges of LLMs in Education: An NLP Perspective,” Jul. 30,2025.https://www.researchgate.net/publication/394121554_Opportunities_and_Challenges_of_LLMs_in_Education_An_NLP_Perspective

[7] J. Oche, A. G. Folashade, T. Ghosal, and A. Biswas, “A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: progress, gaps, and future directions,”

arXiv preprint arXiv:2507.18910, 2025. [Online]. Available: https://arxiv.org/abs/2507.18910

[8] Z. Ji et al., “Survey of Hallucination in Natural Language Generation,” ACM Comput. Surv., vol. 55, no. 12, Art. no. 254, pp. 1–38, Nov. 2022, doi: 10.1145/3571730.

[9] S. S. Rahman et al., “Hallucination to Truth: A review of Fact-Checking and Factuality Evaluation in large language models,” arXiv preprint arXiv:2508.03860, 205. [Online]. Available: https://arxiv.org/abs/2508.03860

[10] D. Gousopoulos and G. Petrakos, “Exploring LLMs as Educational Tools in in Cultural Heritage Heritage.” Accessed: Oct. 18, 2025. [Online]. Available: https://www.itep.gr/wp-content/uploads/2025/04/TMM_409_presentation.pdf

[11] D. Spennemann, “ChatGPT and the generation of digitally born ‘knowledge’: how does a generative AI language model interpret cultural heritage values? Texts of the Individual Essays,” Sep. 2023, doi: https://doi.org/10.13140/RG.2.2.34359.94883.

[12] E. Gumaan,, “Theoretical Foundations and Mitigation of Hallucination in Large Language Models,” rXiv (Cornell University), Jul. 2025, doi: https://doi.org/10.48550/arxiv.2507.22915.

[13] E. Kasneci et al., “ChatGPT for good? On opportunities and challenges of large language models for education,” Learn. Individ. Differ., vol. 103, Art. no. 102274, Mar. 2023, doi: 10.1016/j.lindif.2023.102274.

[14] J. Swacha and M. Gracel, “Retrieval-Augmented Generation (RAG) chatbots for Education: A survey of applications,” Appl. Sci., vol. 15, no. 8, Art. no. 4234, Apr. 2025, doi: 10.3390/app15084234.

[15] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. 34th Int. Conf. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 9459–9474.

[16] "International Day of the Intangible Cultural Heritage," UNESCO National Commission of the Philippines, Oct. 17, 2024. [Online]. Available: https://www.unesco.gov.ph/2024/10/unesco-international-intangible-cultural-heritage/

[17] "Republic Act 7356 – National Commission for Culture and the Arts," NCCA. [Online]. Available: https://ncca.gov.ph/republic-act-7356/

[18] "NCCA-NCHR introduces Philippine History Source Book," National Commission for Culture and the Arts (NCCA). [Online]. Available: https://ncca.gov.ph/2021/08/27/philippine-history-source-book/

[19] "Affiliated Cultural Agencies," National Commission for Culture and the Arts, NCCA. [Online]. Available: https://ncca.gov.ph/about-ncca-3/affiliated-cultural-agencies/

[20] "Saving the Nation's Memory: UNACOM Leads the Call to Safeguard Philippine Documentary Heritage," UNESCO National Commission of the Philippines, Apr. 15, 2025. [Online]. Available: https://www.unesco.gov.ph/2025/04/philippine-documentary-heritage/

[21] "Filipinas Heritage Library | About Us," Filipinas Heritage Library. [Online]. Available: https://www.filipinaslibrary.org.ph/about-us/

[22] "ASEAN Digital Library," ASEAN. [Online]. Available: https://asean.org/asean-digital-library/

[23] S. Wang et al., "Artificial intelligence in education: A systematic literature review," Expert Syst. Appl., vol. 252, Art. no. 124167, May 2024.

[24] UNESCO, "Artificial intelligence in education." [Online]. Available: https://www.unesco.org/en/digital-education/artificial-intelligence

[25] U.S. Department of Education, Office of Educational Technology, "Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations," Washington, DC, 2023.

[26] Ş. Gökçearslan, C. Tosun, and Z. G. Erdemir, "Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: A systematic literature review," Int. J. Technol. Educ., vol. 7, no. 1, pp. 19–39, Jan. 2024.

[27] C. Merino-Campos, "The Impact of Artificial intelligence on Personalized Learning in Higher Education: A Systematic review," Trends in Higher Education, vol. 4, no. 2, p. 17, Mar. 2025.

[28] R. Sajja, Y. Sermet, D. Cwiertny, and I. Demir, "Integrating AI and learning analytics for Data-Driven pedagogical decisions and personalized interventions in education," Technology Knowledge and Learning, Aug. 2025.

[29] K. Abdallah, A. M. Alkaabi, D. A. F. Mehiar, and Z. A. J. Aradat, "Chatbots in classrooms," in Innovative Applications of Generative AI in Education, Hershey, PA, USA: IGI Global, 2024, ch. 12, pp. 166–181.

[30] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," International Journal of Educational Technology in Higher Education, vol. 20, no. 1, Oct. 2023.

[31] Data Privacy Act of 2012, Rep. Act No. 10173, 2012 (Phil.).

[32] IEEE, "IEEE Code of Ethics," Jun. 2020. [Online]. Available: https://www.ieee.org/about/corporate/governance/p7-8.html

[33] S. Minaee et al., "Large language models: A survey," arXiv preprint arXiv:2402.06196, 2025.

[34] Zhihan Lv. Generative Artificial Intelligence in the Metaverse Era. Cognitive Robotics, 3:208–217, 2023

[35] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z.

Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, 54 Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. A Survey of Large Language Models. ArXiv, abs/2303.18223, 2023.

[36] OpenAI, "OpenAI API: Overview," OpenAI Documentation. [Online]. Available: https://platform.openai.com/docs/overview

[37] G. Jawahar, B. Sagot, and D. Seddah, "What Does BERT Learn about the Structure of Language?" in Proc. 57th Annu. Meeting of the Assoc. for Computational Linguistics, Jul. 2019, pp. 3651–3657.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[39] K. C. Sheang and H. Saggion, "Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer," in Proc. 14th Int. Conf. on Natural Language Generation, Aug. 2021, pp. 341–352.

[40] S. Alkaissi and S. McFarlane, "Artificial hallucinations in large language models: Causes, analysis, and mitigation," Journal of Artificial Intelligence Research, vol. 76, pp. 1–22, 2023.

[41] A.T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," OpenAI, Sep. 4, 2025. [Online]. Available: https://openai.com/index/why-language-models-hallucinate/.

[42] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," arXiv preprint arXiv:2401.11817, 2024.

[43] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Trans. Inf. Syst., vol. 43, no. 1, Art. no. 1, Jul. 2024.

[44] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Comput. Surv., vol. 55, no. 12, Art. no. 254, pp. 1–38, Nov. 2022.

[45] Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in Proc. Int. Conf. Learn. Represent. (ICLR), 2020.

[46] W. Zhang and J. Zhang, “Hallucination mitigation for retrieval-augmented large language models: A Survey,” Mathematics, vol. 13, no. 5, Art. no. 856, Feb. 2025, doi: 10.3390/math13050856

[47] [14] J. Swacha and M. Gracel, “Retrieval-Augmented Generation (RAG) chatbots for Education: A survey of applications,” Appl. Sci., vol. 15, no. 8, Art. no. 4234, Apr. 2025, doi: 10.3390/app15084234.

[48] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2023.

[49] K. Shuster et al., "Retrieval-augmented generation for knowledge-grounded dialogue," in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2021, pp. 59–70.

[50] S. Gupta, R. Ranjan, and S. N. Singh, "A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape, and future directions," arXiv preprint arXiv:2410.12837, 2024.

[51] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open-domain question answering," in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2021, pp. 8749–8761.

[52] W. Zhang and J. Zhang, “Hallucination mitigation for retrieval-augmented large language models: A Survey,” Mathematics, vol. 13, no. 5, Art. no. 856, Feb. 2025, doi: 10.3390/math13050856.

[53] O. M. Ayala and P. Béchard, "Reducing hallucination in structured outputs via retrieval-augmented generation," in Proc. 2024 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language, 2024.

[54] X. Lin, R. Socher, and C. Xiong, "Pre-trained Transformers for Dense Retrieval: A Survey," arXiv preprint arXiv:2104.08253, 2021.

[55] J. Kirchenbauer and C. Barns, "Hallucination reduction in large language models with retrieval-augmented generation using Wikipedia knowledge," SocArXiv, Jul. 2024.

[56] H. Soliman, H. Kotte, M. Kravcik, and N. Duong-Trung, "Retrieval-Augmented Chatbots for scalable educational support in higher education," ResearchGate, Mar. 2025.

[57] G. Lang and T. Gürpinar, "AI-Powered Learning Support: A Study of Retrieval-Augmented Generation (RAG) Chatbot Effectiveness in an Online Course," Inf. Syst. Educ. J., vol. 23, no. 2, pp. 4–13, Mar. 2025.

[58] Y. Lian, "Machine assistant with reliable knowledge: Enhancing student learning via RAG-based retrieval," arXiv preprint arXiv:2506.23026, 2025.

[59] J. Singer-Vine, pdfplumber, PyPI, 2025. [Online]. Available: https://pypi.org/project/pdfplumber/

[60] 99] “Chunking in RAG: Improving Retrieval Accuracy,” IBM Architectures, 2023. [Online]. Available: https://www.ibm.com/architectures/papers/rag-cookbook/chunking

[61] “Semantic Chunker: How to Split Documents Using Embeddings,” LangChain Python Documentation, 2025. [Online]. Available: https://python.langchain.com/docs/how_to/semantic-chunker/

[62] “Chunking Strategies for NLP and LLMs,” DataCamp Blog, 2024. [Online]. Available: https://www.datacamp.com/blog/chunking-strategies

[63] A. Smith et al., “Hybrid Chunking for Long Document Retrieval Using Embeddings,” Mathematics, vol. 13, no. 6, 2025. [Online]. Available: https://www.mdpi.com/2079-3197/13/6/151

[64] P. Verma, “S2 Chunking: A Hybrid Framework for Document Segmentation Through Integrated Spatial and Semantic Analysis,” arXiv preprint arXiv:2501.05485v1, Jan. 8, 2025. [Online]. Available: https://arxiv.org/abs/2501.05485v1

[65] R. Qu, F. Bao, and R. Tu, “Is Semantic Chunking Worth the Computational Cost?,” ResearchGate, Nov. 2024. [Online]. Available: https://www.researchgate.net/publication/386168512_Is_Semantic_Chunking_Worth_the_Computational_Cost

[66] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[67] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[68] H. S. Walsh and S. R. Andrade, “Semantic search with Sentence-BERT for design information retrieval,” in Proc. IDETC/CIE 2022, 2022, pp. V002T02A066.

[69] IBM, "What is information retrieval?" IBM Think, 2024. [Online]. Available: https://www.ibm.com/think/topics/information-retrieval

[70] M. Ashikuzzaman, "What is information retrieval and why does it matter?" Lise Edu Network, 2025.

[71] F. Rosa, R. Rodrigues, R. Lotufo, and R. Nogueira, "Yes, BM25 is a strong baseline for legal case retrieval," arXiv preprint arXiv:2105.05686, 2021.

[72] K. Sarkar and A. Gupta, "An empirical study of some selected IR models for Bengali monolingual information retrieval," 2017, *arXiv:1706.03266*. [Online]. Available: https://arxiv.org/abs/1706.03266

[73] M. Mandikal et al., “Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval,” Univ. Texas Austin, 2025

[74] H.-L. Hsu and J. Tzeng, "DAT: Dynamic Alpha Tuning for Hybrid Retrieval in Retrieval-Augmented Generation," arXiv preprint arXiv:2503.23013, Mar. 2025.

[75] BAAI, "BGE Reranker," BGE Model Documentation. [Online]. Available: https://bge-model.com/bge/bge_reranker.html

[76] L. Wu, F. Petroni, M. Josifoski, S. Riedel, and L. Zettlemoyer, “Scalable Zero-shot Entity Linking with Dense Entity Retrieval,” in Proc. EMNLP, Hong Kong, 2019, pp. 6175–6185.

[77] J. White et al., "A Prompt Pattern Language: Communicating with Large Language Models," arXiv preprint arXiv:2302.11382, 2023.

[78] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 1877–1901.

[79] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022.

[80] L. Ouyang et al., "Training language models to follow instructions with human feedback," in Adv. Neural Inf. Process. Syst., vol. 35, 2022.

[81] X. Wu, Z. Yu, H. Yang, and R. Zhang, "Beyond factuality: Evaluating reasoning in retrieval-augmented generation," arXiv preprint arXiv:2401.12345, 2024.

[82] H. Yu et al., "Evaluation of Retrieval-Augmented Generation: A Survey," arXiv preprint arXiv:2405.07437, 2024. [Online]. Available: https://arxiv.org/abs/2405.07437

[83] Salemi and H. Zamani, "Evaluating Retrieval Quality in Retrieval-Augmented Generation," in Proc. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2024.

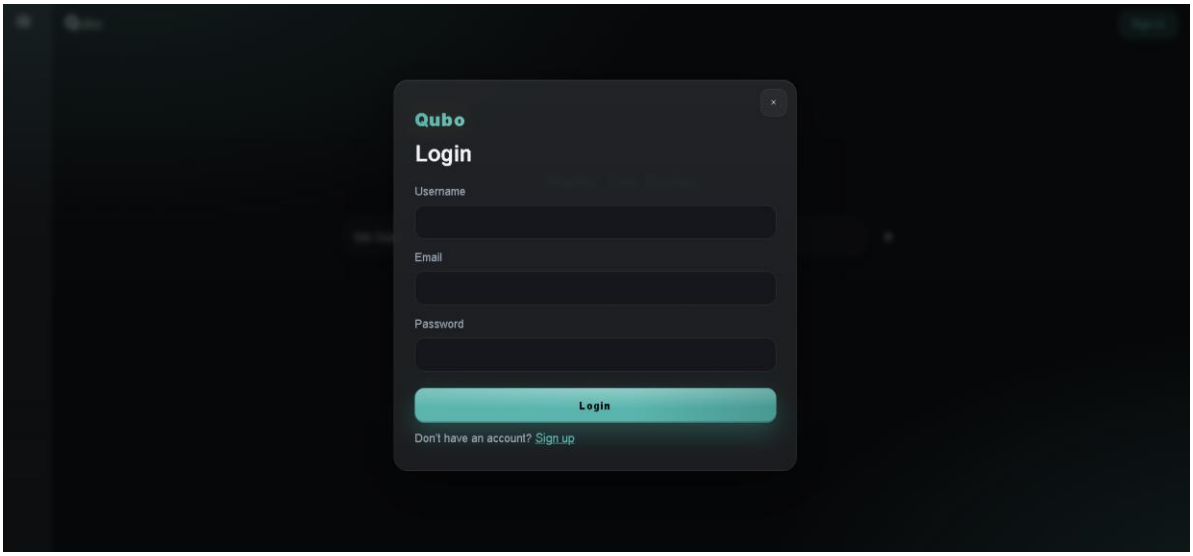
[84] “Evaluation Metrics for Retrieval-Augmented Generation (RAG) Systems,” GeeksforGeeks, 2025. [Online]. Available: https://www.geeksforgeeks.org/nlp/evaluation-metrics-for-retrieval-augmented-generation-rag-systems/

[85] T. Chen et al., “Dense X Retrieval: What Retrieval Granularity Should We Use?,” arXiv preprint arXiv:2312.06648, 2023.

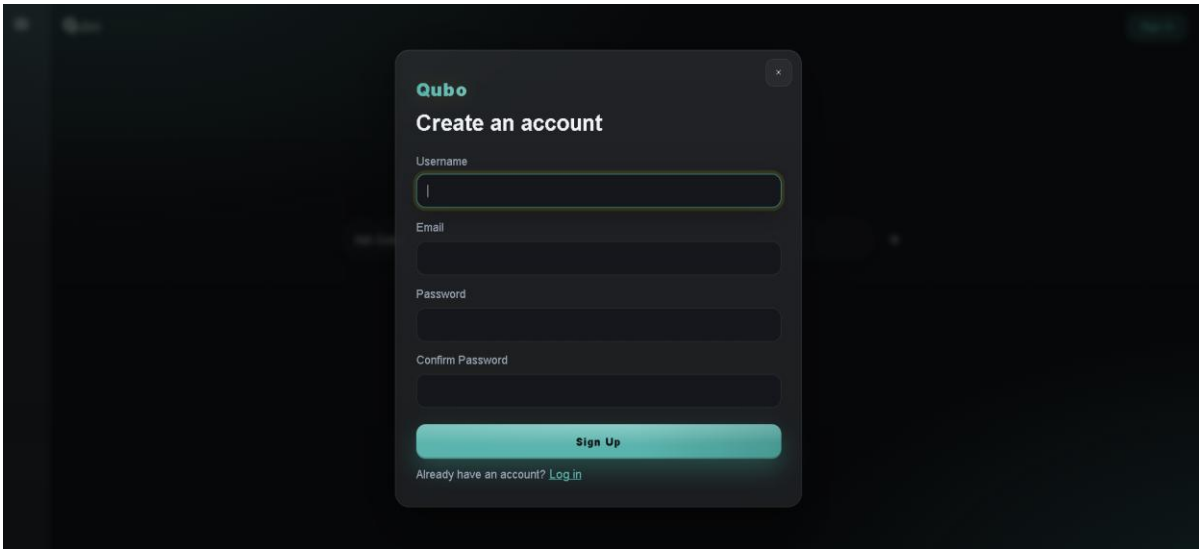
[86] J. Gu et al., "A survey on LLM-as-a-judge," arXiv preprint arXiv:2411.15594, 2024.

[87] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

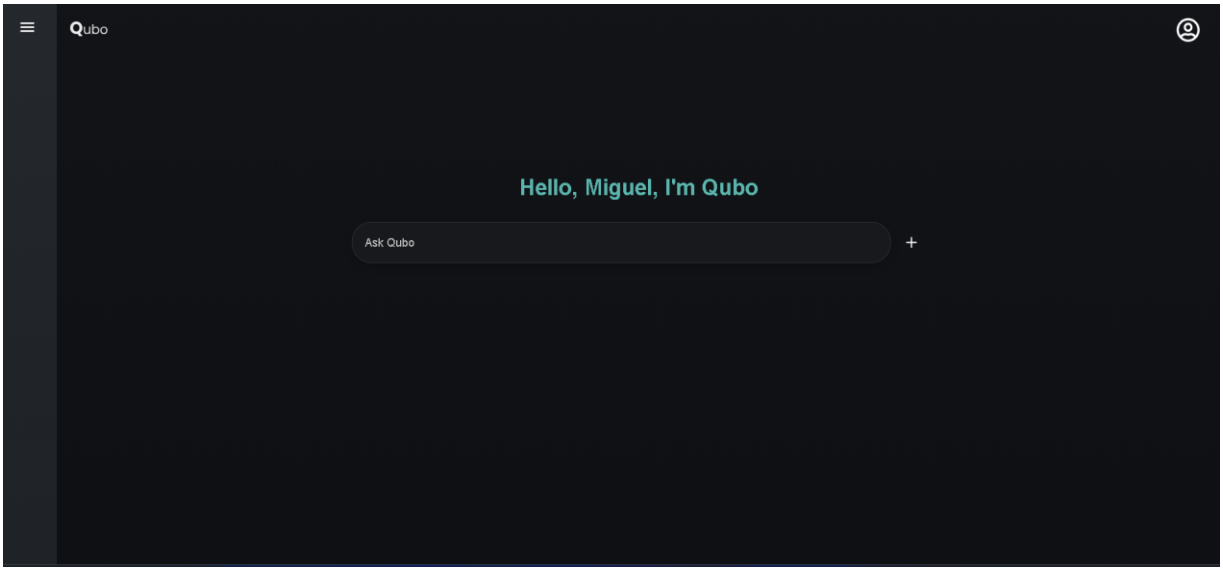
APPENDICES



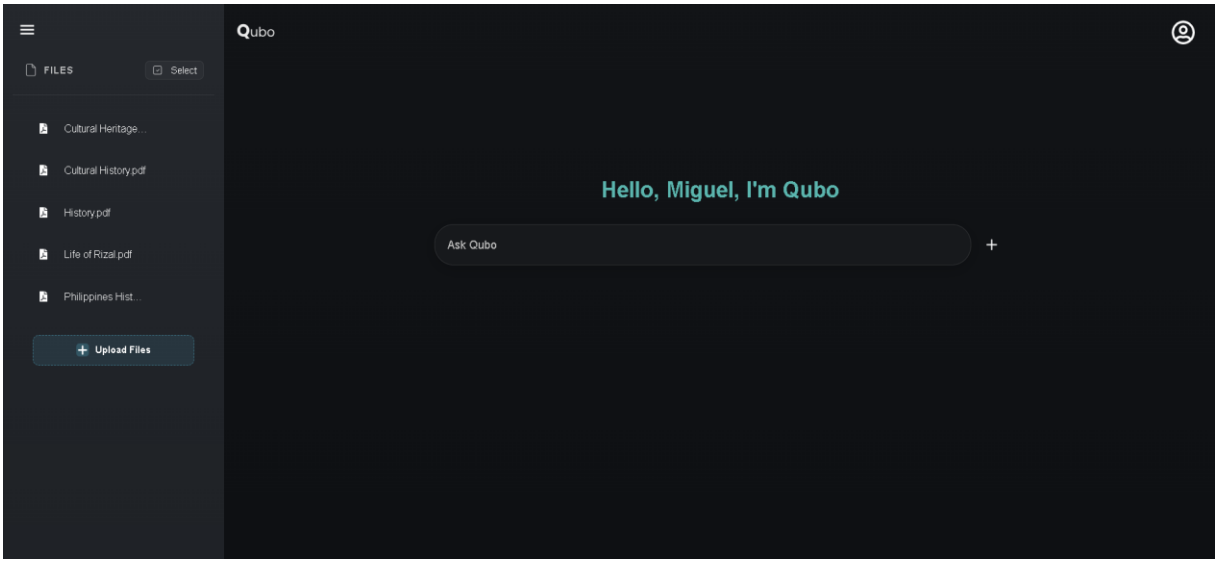
Login



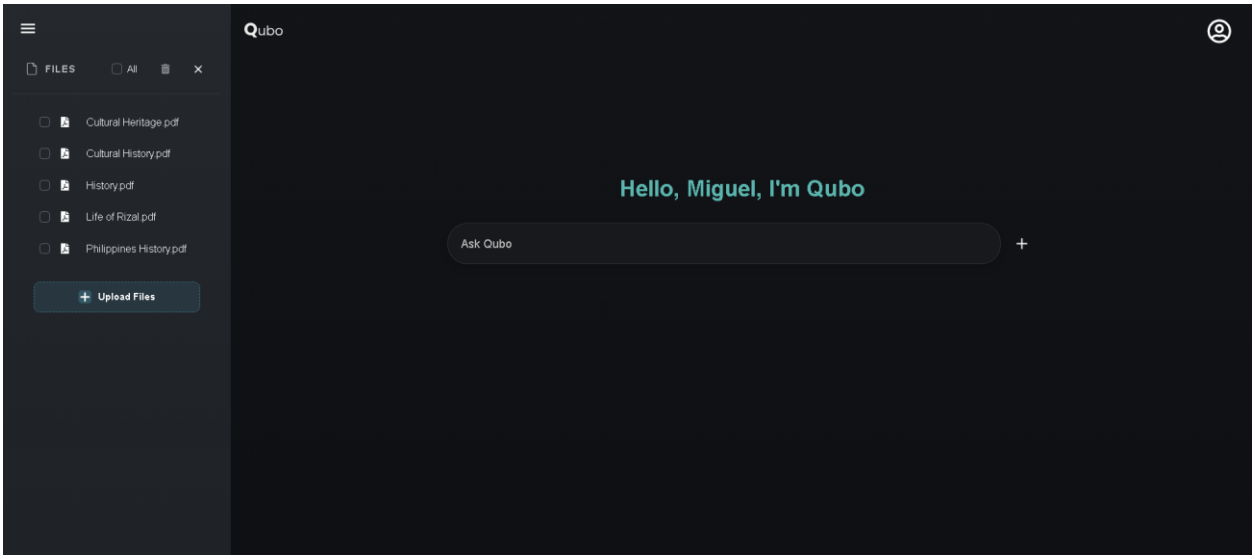
Sign Up



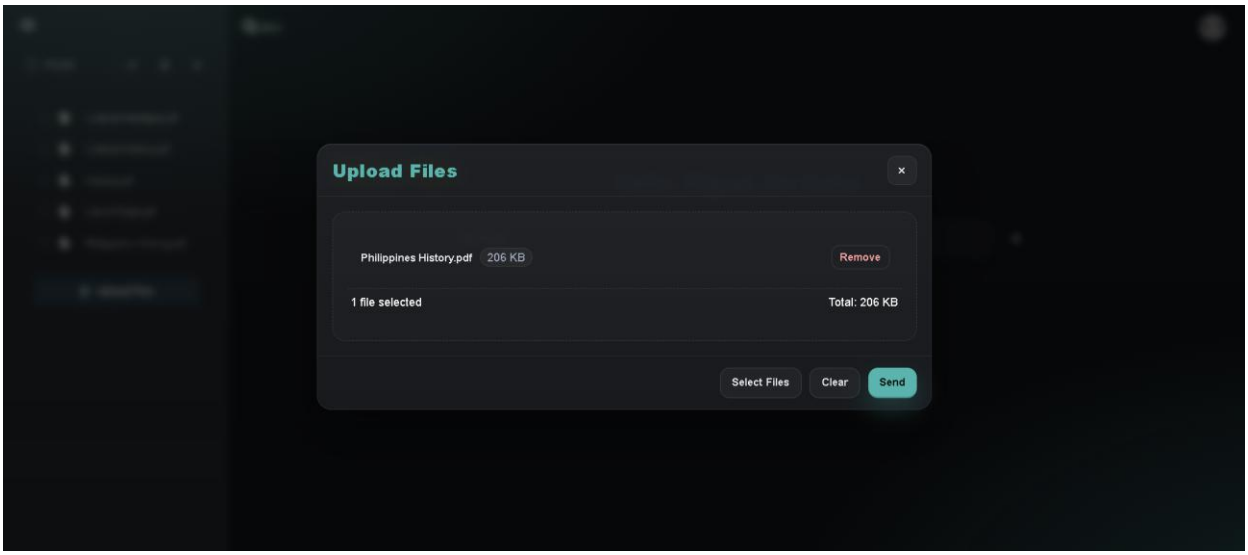
Chat



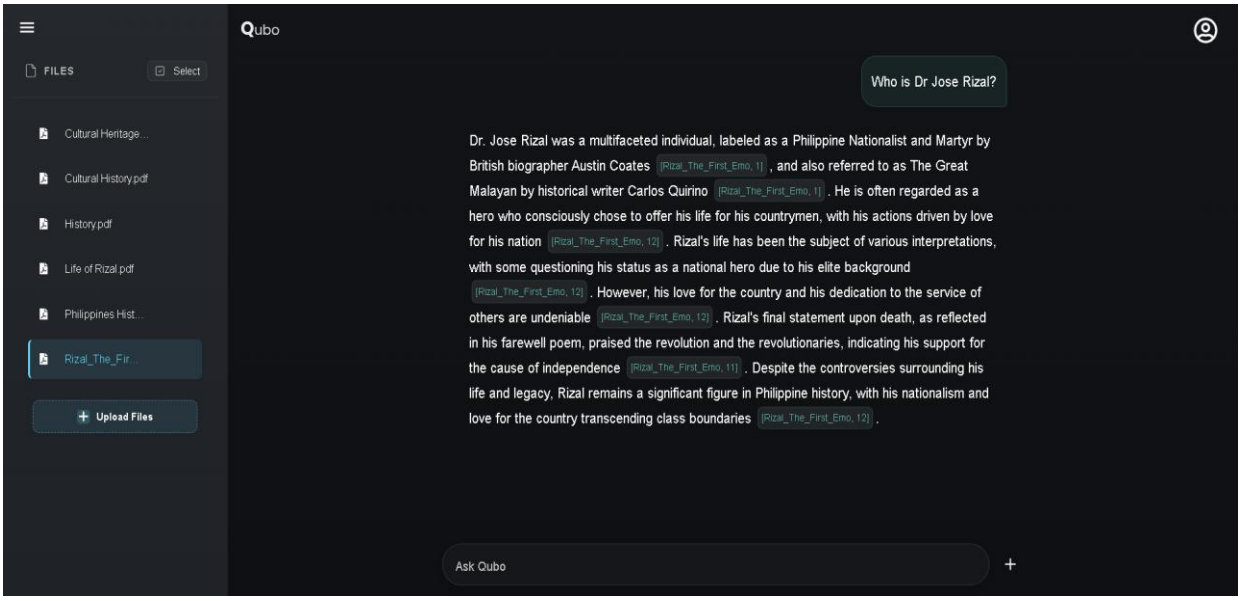
File List



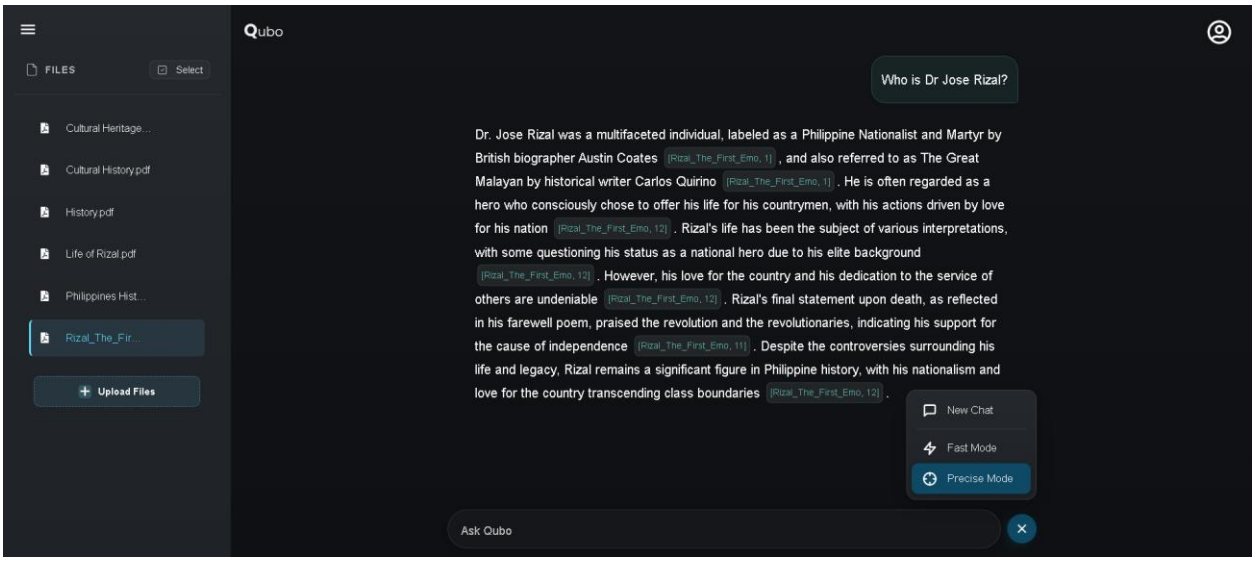
Multi Select



File Upload



Question Chat



Chat Settings

APPENDIX A.
LETTERS AND DOCUMENTATIONS

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

Date: August 30, 2025

PROF. MS. MARIZKAYS JAMISON
Faculty, CCIT

Dear Sir/Madam:

Greetings!

We are students pursuing a degree in **Bachelor of Science in Computer Science** with **Specialization in Machine Learning** currently enrolled in the course **CCTHESS1- Thesis 1.**

We are writing to humbly request your service and expertise to serve as an advisor for our study with the working title:

AI Assistant with Reliable Knowledge: Supporting Student Learning Through Lecture Notes via RAG-Based Retrieval

Your knowledge and insights will be valuable and greatly enrich our work in the entire cycle of Capstone/Thesis Project.

Thank you for being so considerate, and we hope you can fulfill our request. God Bless.

Respectfully yours,

Alis, Renz Andrei C.
Bongao, Christian I.
Diaz, Kris Brian V.
Layos, Miguel Raphael
Manzanero, Brix Anthony G.

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISING CONTRACT

As **Capstone/Thesis Topic Adviser**, it is my duty and responsibility as stated in College Research Manual Guidelines and Policies to:

1. Ensure that the project proposed by the students conforms to the standard of the College and has an immediate or potential impact on the research thrust of the university.
2. Guide the Capstone/Thesis Project of the students in the following tasks while in the proposal stage:
   a) Defining the research problems/objectives in clear specific terms
   b) Building a working bibliography for the research
   c) Identifying variables and formulating a hypothesis if any
   d) Determining research design, population to be studied, research environment, instruments to be used, and the data collection procedures
3. Meet the team regularly (NOTE: the team must fill out the Adviser Consultation Monitoring Sheet) to answer questions and resolve impasses and conflicts.
4. Point out errors in the development of work, in the analysis, or the documentation. The adviser must remind the Proponents/Researchers to do their work properly.
5. Review all deliverables thoroughly at every stage of the Capstone/Thesis Project to ensure that they meet the department's standards. The adviser may also require his/her Proponents/Researchers to submit Progress Reports regularly.
6. Recommend the Proponents/Researchers for Proposal and Oral Presentation. The adviser should not sign the Recommendation for Oral Presentation form if he/she believes that the Proponents/Researchers are not yet ready for Proposal and Oral Presentation, respectively. Thus, if the Proponents/Researchers fail in the Proposal or Oral Presentation, it is partially the adviser's fault.
7. Clarify points during the Proposal and Oral Presentation.
8. Ensure that all required revisions are incorporated into the appropriate documents and/or software.
9. Keep informed of the schedule of Capstone / Thesis Project activities, required deliverables, and deadlines.
10. Recommend to the Proposal and Oral Presentation panel the nomination of his/her Research / Capstone Project for an award.

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

I acknowledge that I have read and understood those mentioned above and swear to abide by the same as an adviser for:

CAPSTONE/THESIS PROJECT TITLE

Capstone/Thesis Advisees

Alis, Renz Andrei C.
Signature over Printed Name/Date

Bongao, Christian I.
Signature over Printed Name/Date

Diaz, Kris Brian V.
Signature over Printed Name/Date

Layos, Miguel Raphael
Signature over Printed Name/Date

Manzanero, Brix Anthony G.
Signature over Printed Name/Date

CONSULTATION HOURS	
Monday	
Tuesday	4:00 - 6:00
Wednesday	8:00 - 6:00
Thursday	
Friday	4:30 - 6:00
Saturday	

SIGNED BY

Ms. Marizkays Jamison
Capstone/Thesis Topic Adviser

Mr. Rogel Labanan
Capstone/Thesis Instructor

Mr. Eliseo Ramirez
Program Chair, IT/CS Department

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISER'S CONSULTATION

Researcher/Proponent's Name *(Lastname, First Name, MI)*	Section	Signature	Consultation No.
Alis, Renz Andrei C.	COM222		
Bongao, Christian I.	COM223		
Diaz, Kris Brian V.	COM223		1
Layos, Miguel Raphael	COM223		
Manzanero, Brix Anthony G.	COM223		
Research Course/Subject		BSCS Thesis 1	

Topic Adviser	Ms. Marizkaye Jamison Signature over Printed Name	**Date**	02/09/2025
		Time	4:00 pm
		Venue	F2F – Room 501

Consultation Details

PROJECT TITLE	*Briefly write the description of your consultation (filled out by the GROUP).*
AI Assistant with Reliable Knowledge: Supporting Student Learning Through Lecture Notes via RAG-Based Retrieval	Title Revision – Paper Revisions

Remarks/Action Points

Indicate whether the consultation has sufficiently solved the concern or needs a follow-up. (This section must be filled out by the TOPIC ADVISER.)

- Title Revision:
  From: AI Assistant with Reliable Knowledge: Supporting Student Learning Through Lecture Notes via RAG-Based Retrieval

  To: **QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes**
- Gave AI Assistant (Chatbot) a Name "QUBO" (Question Bot)
- Paper draft revisions

Reminder: *Produce triplicate copy of this form (Instructor, Adviser and Student). Fill-up all blank fields (____) prior printing the form. Other section must be filled out by the group and adviser.*

Adviser Consultation Form

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISER'S CONSULTATION

Researcher/Proponent's Name *(Lastname, First Name, MI)*	Section	Signature	Consultation No.
Alis, Renz Andrei C.	COM222		
Bongao, Christian I.	COM223		
Diaz, Kris Brian V.	COM223		
Layos, Miguel Raphael	COM223		2
Manzanero, Brix Anthony G.	COM223		
Research Course/Subject		BSCS Thesis 1	

Topic Adviser	Ms. Marizkays Jamison Signature over Printed Name	**Date**	03/09/2025
		Time	5:00 pm
		Venue	F2F – Room 501

Consultation Details

PROJECT TITLE	*Briefly write the description of your consultation (filled out by the GROUP).*
QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes	Our adviser told us to show the chatbot's limits, unique RAG-Based features and discuss future improvements

Remarks/Action Points

Indicate whether the consultation has sufficiently solved the concern or needs a follow-up. (This section must be filled out by the TOPIC ADVISER.)

- Your project needs more clarity on file support. Right now, it looks like the chatbot is mainly for PDFs, but you must specify if it can also handle other formats like Word or text files. Also, decide what the chatbot should do when the information being asked is not in the uploaded file should it return an error, or give a default response? Another point is whether it will be capable of reading images (OCR) and how many files can be uploaded at once.
- Unlike existing chatbots that rely on internet sources, your system is designed to give answers only from uploaded PDFs through RAG. Make sure to explain the difference between raw RAG and optimized RAG so your audience can see how your version is more accurate and reliable.
- In terms of originality, highlight that the system is localized for your context and not just a copy of existing tools. The features you need to stress human-like responses, and your improved RAG model. Also, don't forget to provide proof that the system is indeed powered by LLM and RAG.
- For your presentation, the last slides should talk about future improvements, like support for more file types, better image-to-text conversion, improved memory and context awareness, and possible multi-language support.

Reminder: *Produce triplicate copy of this form (Instructor, Adviser and Student). Fill-up all blank fields (____) prior printing the form. Other section must be filled out by the group and adviser.*

Adviser Consultation Form

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISER'S CONSULTATION

Researcher/Proponent's Name *(Lastname, First Name, MI)*	Section	Signature	Consultation No.
Alis, Renz Andrei C.	COM222		
Bongao, Christian I.	COM223		
Diaz, Kris Brian V.	COM223		3
Layos, Miguel Raphael	COM223		
Manzanero, Brix Anthony G.	COM223		
Research Course/Subject		BSCS Thesis 1	

Topic Adviser	Ms. Marizkays Jamison Signature over Printed Name	**Date**	05/09/2025
		Time	5:30 pm
		Venue	F2F – Room 501

Consultation Details

PROJECT TITLE	*Briefly write the description of your consultation (filled out by the GROUP).*
QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes	We showed prototype and revised paper

Remarks/Action Points

Indicate whether the consultation has sufficiently solved the concern or needs a follow-up. (This section must be filled out by the TOPIC ADVISER.)

- The chatbot should show the uploaded PDFs on the side panel, so users can easily see and access the files they've added.
- The search bar should be placed at the bottom so users can easily find the uploaded PDF files and visible.
- When the user clicks on a PDF, the selected file should open on the screen so they can view its contents.
- In the sources section of chat both the author and the link of each uploaded PDF should be displayed.
- Revise the background of the study

Reminder: *Produce triplicate copy of this form (Instructor, Adviser and Student). Fill-up all blank fields (____) prior printing the form. Other section must be filled out by the group and adviser.*

Adviser Consultation Form

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISER'S CONSULTATION

Researcher/Proponent's Name *(Lastname, First Name, MI)*	Section	Signature	Consultation No.
Alis, Renz Andrei C.	COM222		
Bongao, Christian I.	COM223		
Diaz, Kris Brian V.	COM223		
Layos, Miguel Raphael	COM223		4
Manzanero, Brix Anthony G.	COM223		
Research Course/Subject		BSCS Thesis 1	

Topic Adviser	Ms. Marizka Jamison Signature over Printed Name	**Date**	12/09/2025
		Time	6:00 pm
		Venue	F2F – Faculty

Consultation Details

PROJECT TITLE	*Briefly write the description of your consultation (filled out by the GROUP).*
QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes	Chapter 2 Revisions and Discussion of Methodology

Remarks/Action Points

Indicate whether the consultation has sufficiently solved the concern or needs a follow-up. (This section must be filled out by the TOPIC ADVISER.)

- *Chapter 2 Revisions*
  - *Contextual order of studies*
- *Image recognition budget constraints*
- *Methodology*
  - *User Testing*
  - *Student Evaluation*
  - *If possible, to conduct survey students from different universities (any year level)*

Reminder: *Produce triplicate copy of this form (Instructor, Adviser and Student). Fill-up all blank fields (____) prior printing the form. Other section must be filled out by the group and adviser.*

Adviser Consultation Form

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISER'S CONSULTATION

Researcher/Proponent's Name (Lastname, First Name, MI)	Section	Signature	Consultation No.
Alis, Renz Andrei C.	COM222		
Bongao, Christian I.	COM223		
Diaz, Kris Brian V.	COM223		
Layos, Miguel Raphael	COM223		5
Manzanero, Brix Anthony G.	COM223		
Research Course/Subject		BSCS Thesis 1	

Topic Adviser	Ms. Marizkaris Jamison Signature over Printed Name	**Date**	16/09/2025
		Time	4:30 pm
		Venue	F2F

Consultation Details

PROJECT TITLE	*Briefly write the description of your consultation (filled out by the GROUP).*
QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes	Model and Prototype Recommendations and Improvements

Remarks/Action Points

Indicate whether the consultation has sufficiently solved the concern or needs a follow-up. (This section must be filled out by the TOPIC ADVISER.)

- *UI design*
  - *File uploads*
  - *Drag and drop specific document*
  - *Font Readability*
- *Features*
  - *Input Multiple files*
  - *File Selection for Referencing*
  - *File Organization*
- *Chatbot Response Limitations and restrictions*

Reminder: *Produce triplicate copy of this form (Instructor, Adviser and Student). Fill-up all blank fields (____) prior printing the form. Other section must be filled out by the group and adviser.*

Adviser Consultation Form

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISER'S CONSULTATION

Researcher/Proponent's Name *(Lastname, First Name, MI)*	Section	Signature	Consultation No.
Alis, Renz Andrei C.	COM222		
Bongao, Christian I.	COM223		
Diaz, Kris Brian V.	COM223		6
Layos, Miguel Raphael	COM223		
Manzanero, Brix Anthony G.	COM223		

Research Course/Subject	BSCS Thesis 1		
Topic Adviser	Ms. Marizkeys Jamison Signature over Printed Name	**Date**	29/09/2025
		Time	3:00 pm
		Venue	F2F

Consultation Details

PROJECT TITLE	*Briefly write the description of your consultation (filled out by the GROUP).*
QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes	Advise on the Research Design to be used

Remarks/Action Points
Indicate whether the consultation has sufficiently solved the concern or needs a follow-up. (This section must be filled out by the TOPIC ADVISER.)

- *Discussion on Methodology and Research design*

Reminder: *Produce triplicate copy of this form (Instructor, Adviser and Student). Fill-up all blank fields (____) prior printing the form. Other section must be filled out by the group and adviser.*

Adviser Consultation Form

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISER'S CONSULTATION

Researcher/Proponent's Name *(Lastname, First Name, MI)*	Section	Signature	Consultation No.
Alis, Renz Andrei C.	COM222		
Bongao, Christian I.	COM223		
Diaz, Kris Brian V.	COM223		7
Layos, Miguel Raphael	COM223		
Manzanero, Brix Anthony G.	COM223		
Research Course/Subject		BSCS Thesis 1	

Topic Adviser	Ms. Marizkays Jamison Signature over Printed Name	**Date**	30/09/2025
		Time	5:00 pm
		Venue	F2F

Consultation Details

PROJECT TITLE	*Briefly write the description of your consultation (filled out by the GROUP).*
QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes	Chapter 1 Revisions and Fix Formatting of Paper

Remarks/Action Points

Indicate whether the consultation has sufficiently solved the concern or needs a follow-up. (This section must be filled out by the TOPIC ADVISER.)

- *Revisions for Introduction, Background of the Study and Limitations*
- *Fix paper formatting issues*
- *Fix citation format in paper set to IEEE*

Reminder: *Produce triplicate copy of this form (Instructor, Adviser and Student). Fill-up all blank fields (____) prior printing the form. Other section must be filled out by the group and adviser.*

Adviser Consultation Form

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

ADVISER'S CONSULTATION

Researcher/Proponent's Name *(Lastname, First Name, MI)*	Section	Signature	Consultation No.
Alis, Renz Andrei C.	COM222		
Bongao, Christian I.	COM223		
Diaz, Kris Brian V.	COM223		
Layos, Miguel Raphael	COM223		8
Manzanero, Brix Anthony G.	COM223		
Research Course/Subject		BSCS Thesis 1	

Topic Adviser	Ms. Marizkays Jamison Signature over Printed Name	**Date**	09/10/2025
		Time	6:00 pm
		Venue	F2F - 413

Consultation Details

PROJECT TITLE	*Briefly write the description of your consultation (filled out by the GROUP).*
QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes	Checking of Final Paper revisions and Prototype

Remarks/Action Points

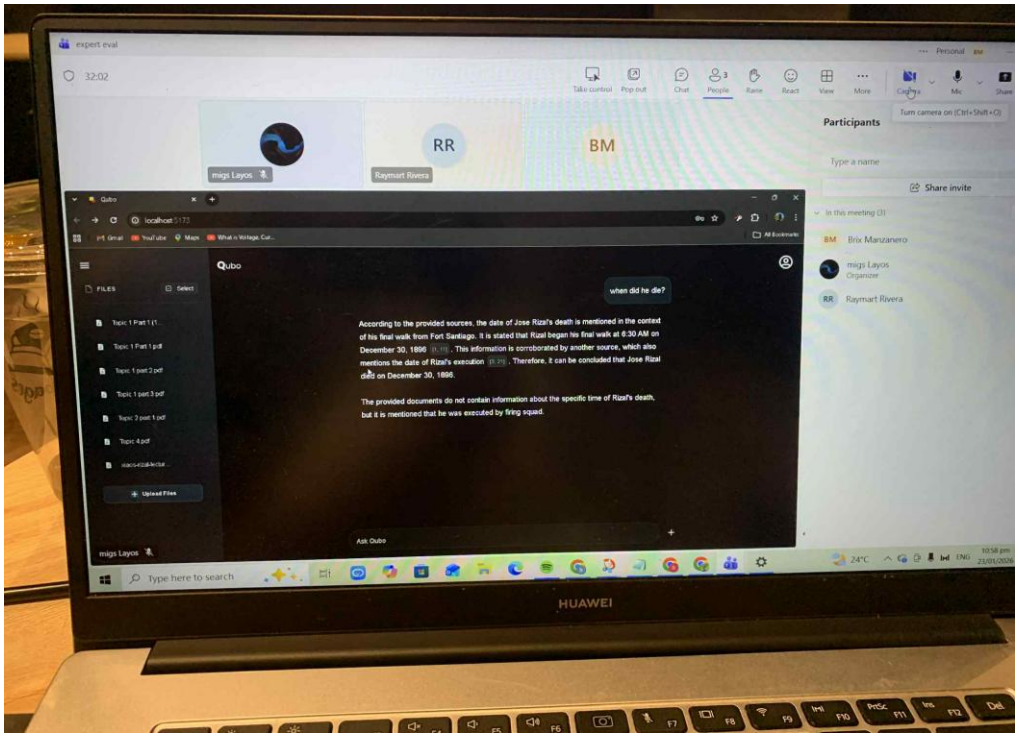
Indicate whether the consultation has sufficiently solved the concern or needs a follow-up. (This section must be filled out by the TOPIC ADVISER.)

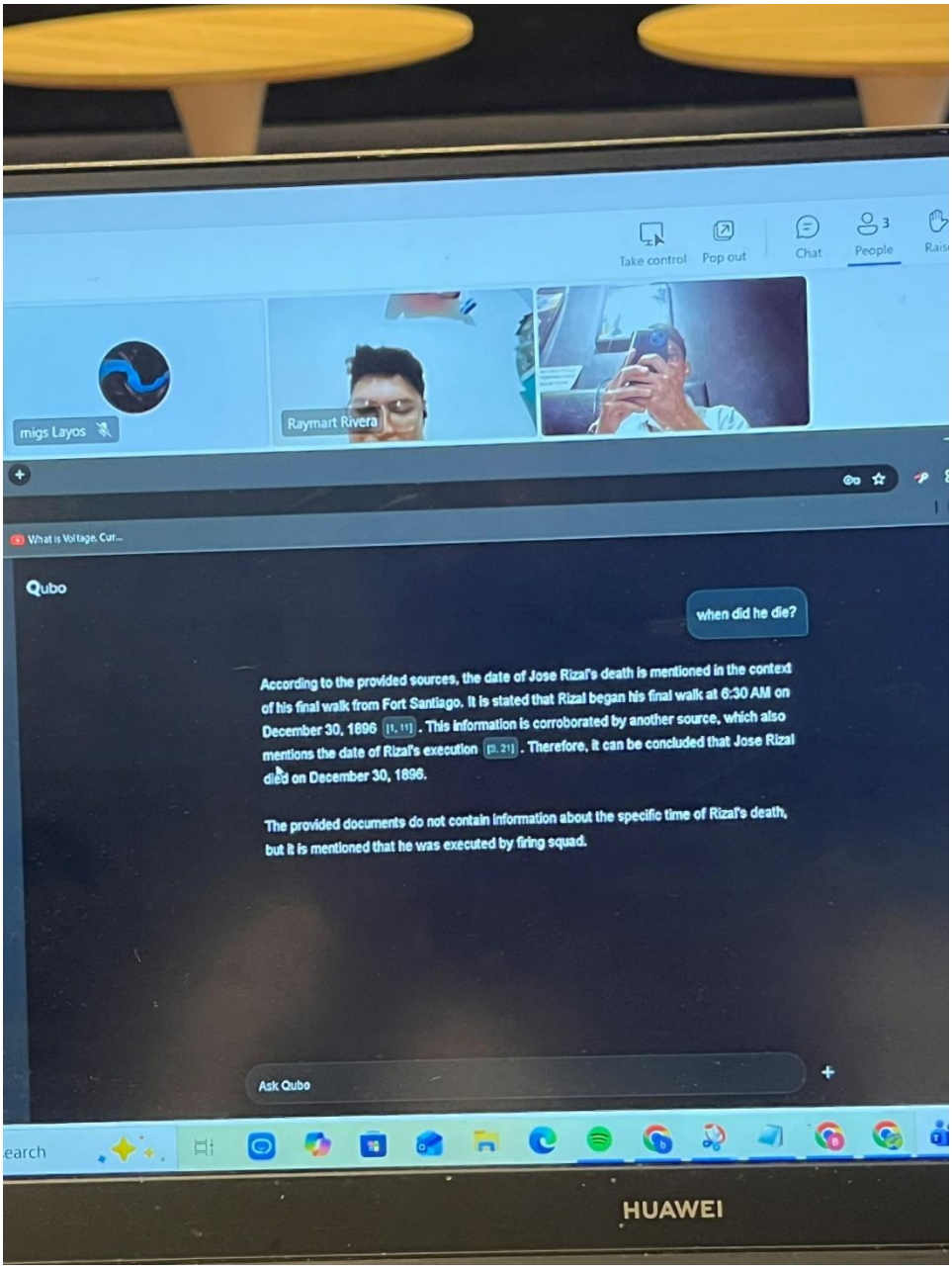
- *Checking of Final Paper and Prototype checking*

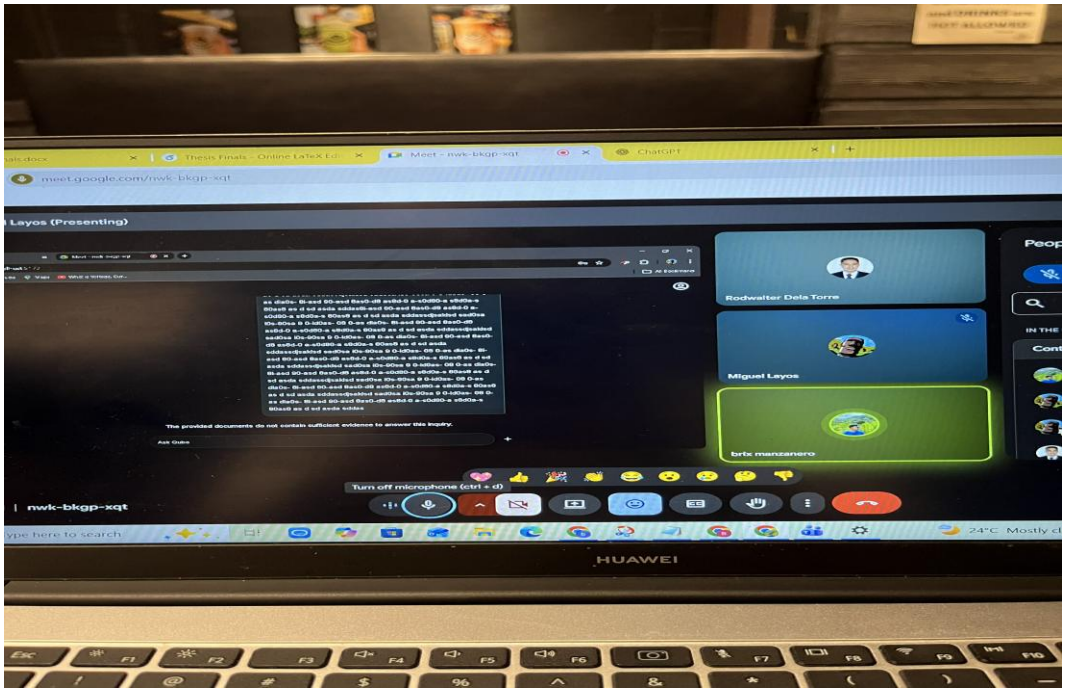
Reminder: *Produce triplicate copy of this form (Instructor, Adviser and Student). Fill-up all blank fields (____) prior printing the form. Other section must be filled out by the group and adviser.*

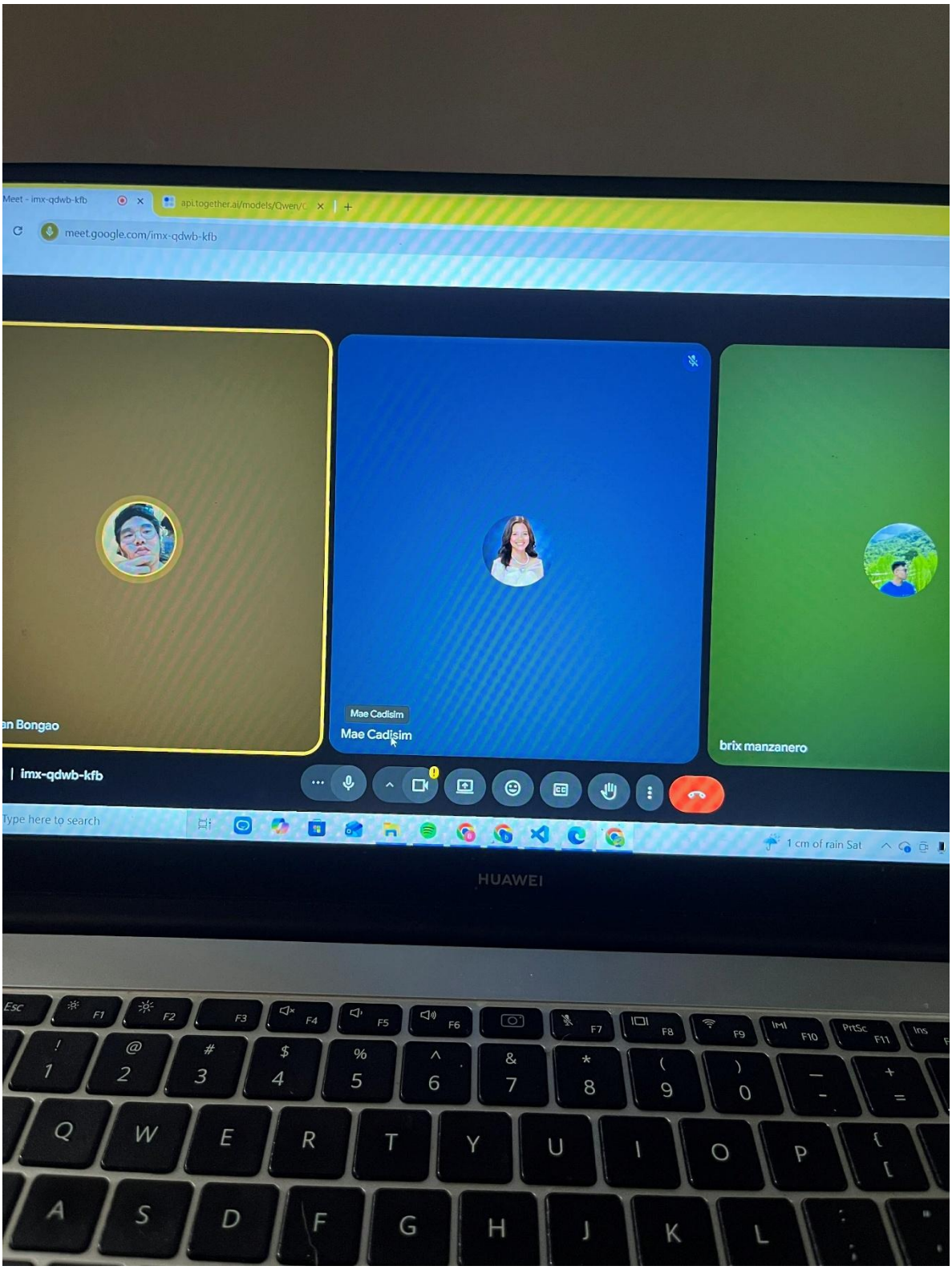
Adviser Consultation Form

APPENDIXB.
TRANSCRIPT OF INTERVIEWS AND PHOTO DOCUMENTATION


Qubo
when did he die?
According to the provided sources, the date of Jose Rizal's death is mentioned in the context of his final walk from Fort Santiago. It is stated that Rizal began his final walk at 6:30 AM on December 30, 1896 . This information is corroborated by another source, which also mentions the date of Rizal's execution . Therefore, it can be concluded that Jose Rizal died on December 30, 1896.
The provided documents do not contain information about the specific time of Rizal's death, but it is mentioned that he was executed by firing squad.
Ask Qubo
Participants
HUAWEI


Take control
Pop out
Chat
People
migs Layos
Raymart Rivera
What is Voltage, Cur...
Qubo
when did he die?
According to the provided sources, the date of Jose Rizal's death is mentioned in the context of his final walk from Fort Santiago. It is stated that Rizal began his final walk at 6:30 AM on December 30, 1896 [1, 11]. This information is corroborated by another source, which also mentions the date of Rizal's execution [3, 21]. Therefore, it can be concluded that Jose Rizal died on December 30, 1896.
The provided documents do not contain information about the specific time of Rizal's death, but it is mentioned that he was executed by firing squad.
Ask Qubo
HUAWEI

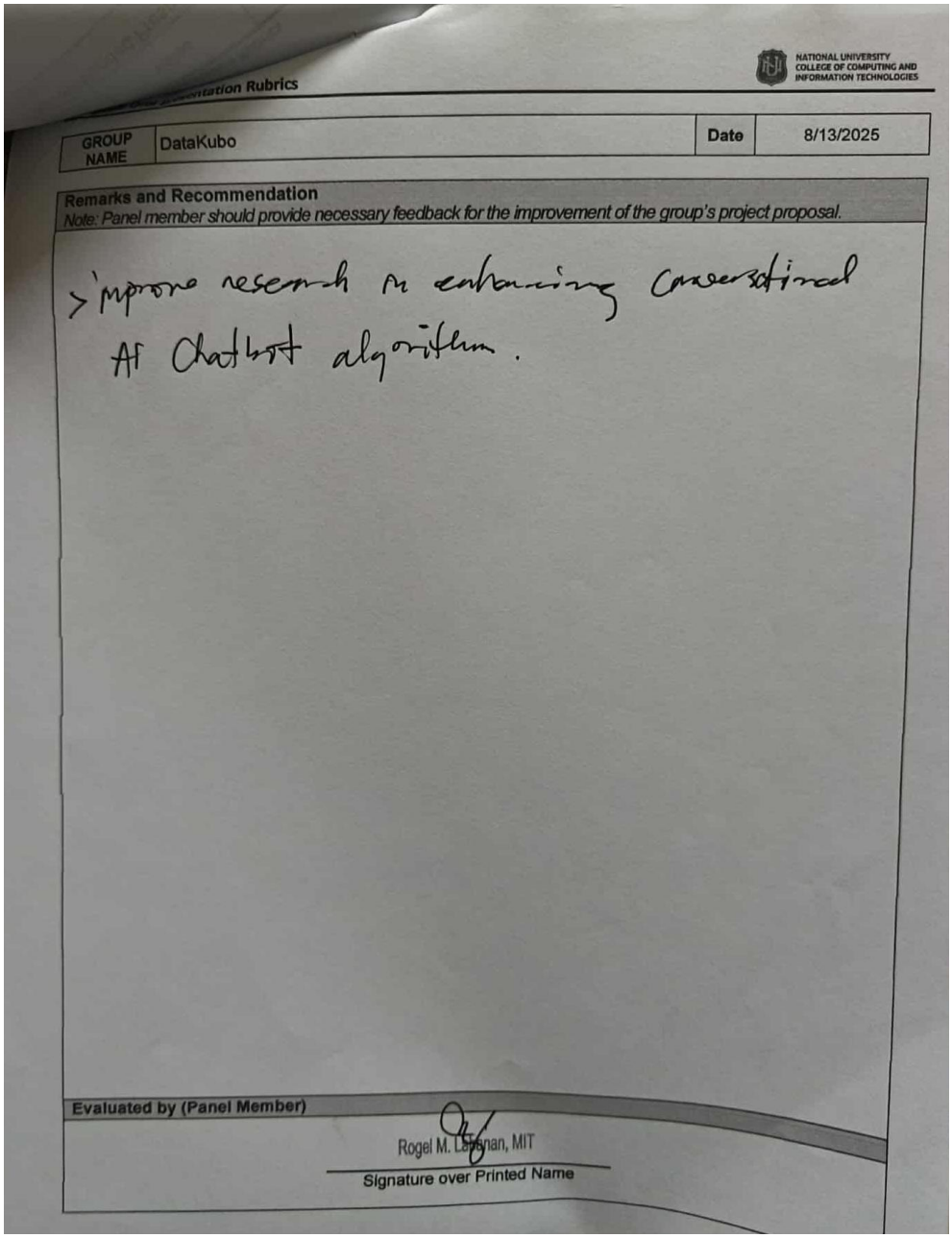

meet.google.com/nwk-bkgp-xqt
Layos (Presenting)
Rodwalter Dela Torre
Miguel Layos
brix manzanero
Turn off microphone (ctrl + d)
nwk-bkgp-xqt
HUAWEI


Meet - imx-qdwb-kfb
meet.google.com/imx-qdwb-kfb
Mae Cadisim
Mae Cadisim
brix manzanero
imx-qdwb-kfb
Type here to search
1 cm of rain Sat
HUAWEI


QUBO


Graduating Students
Clearance Requirements
REMINDER
Clearance Requirements for Graduating Students, proceed to
Graduating Students
Clearance Requirements
EXIT

APPENDIX C.
PANEL'S LIST OF RECOMMENDATIONS



...entation Rubrics

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

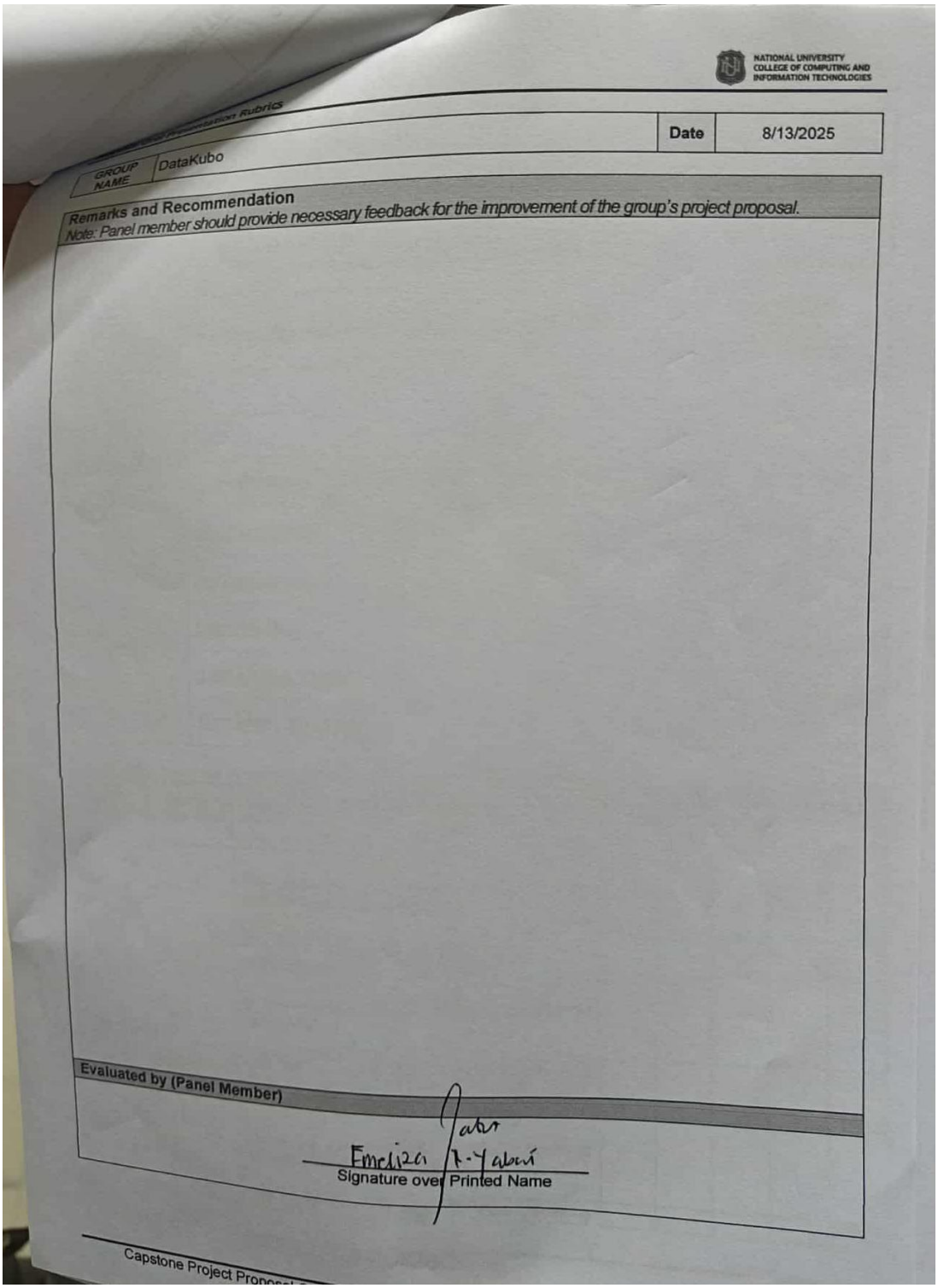
GROUP NAME	DataKubo	Date	8/13/2025

Remarks and Recommendation
Note: Panel member should provide necessary feedback for the improvement of the group's project proposal.

> Improve research in enhancing conversational AI Chatbot algorithm.

Evaluated by (Panel Member)

Rogel M. Lapanan, MIT
Signature over Printed Name



NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

Capstone Project Proposal Presentation Rubrics

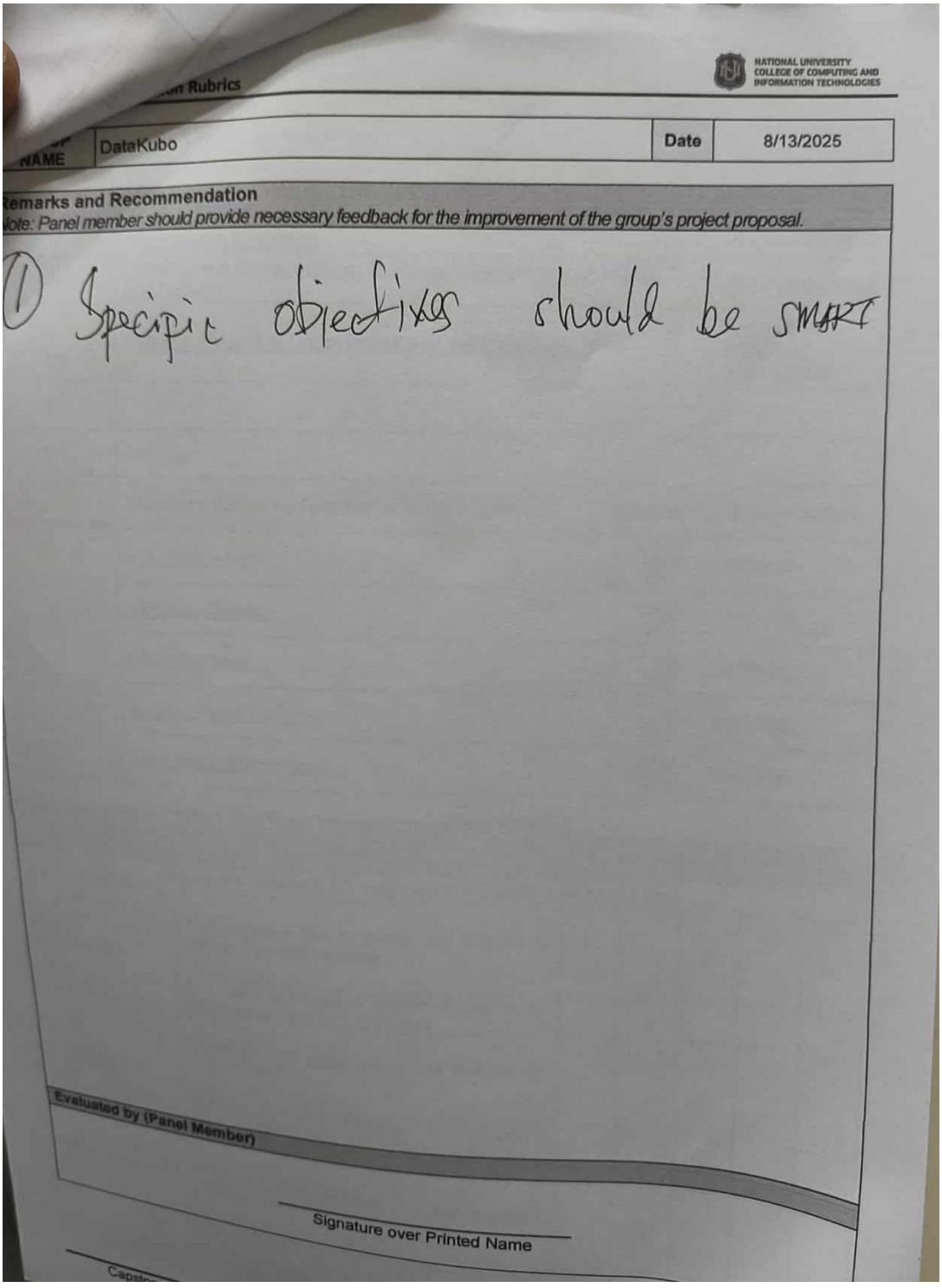
GROUP NAME	DataKubo	Date	8/13/2025

Remarks and Recommendation
Note: Panel member should provide necessary feedback for the improvement of the group's project proposal.

Evaluated by (Panel Member)

Emelita T. Yabut
Signature over Printed Name

Capstone Project Proposal



Rubrics

NATIONAL UNIVERSITY
COLLEGE OF COMPUTING AND
INFORMATION TECHNOLOGIES

NAME	DataKubo	Date	8/13/2025

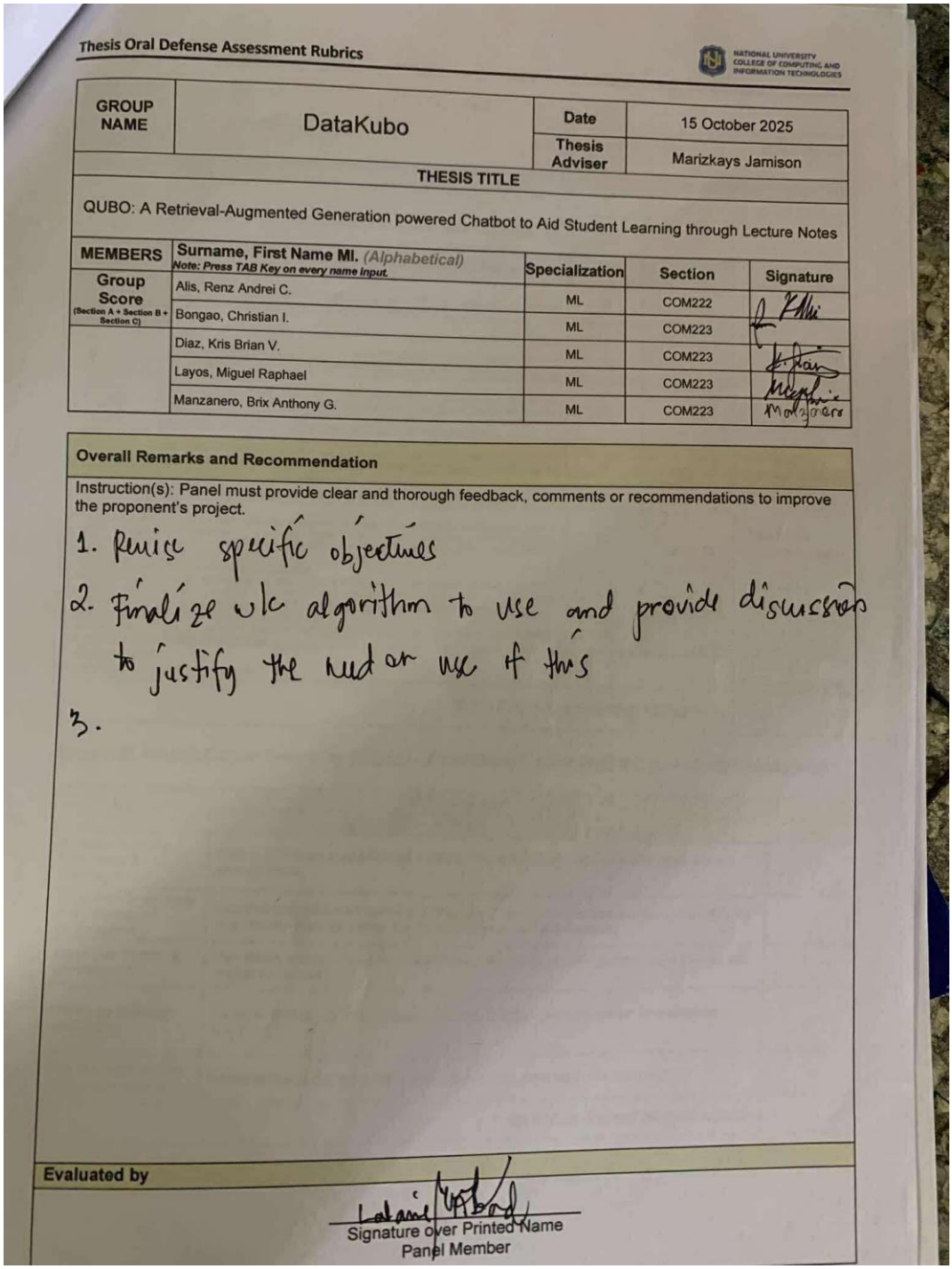
Remarks and Recommendation
Note: Panel member should provide necessary feedback for the improvement of the group's project proposal.

1) Specific objectives should be SMART

Evaluated by (Panel Member)

Signature over Printed Name

Capst...



Thesis Oral Defense Assessment Rubrics

NATIONAL UNIVERSITY COLLEGE OF COMPUTING AND INFORMATION TECHNOLOGIES

GROUP NAME	DataKubo	Date	15 October 2025
		Thesis Adviser	Marizkays Jamison

THESIS TITLE

QUBO: A Retrieval-Augmented Generation powered Chatbot to Aid Student Learning through Lecture Notes

MEMBERS	Surname, First Name MI. *(Alphabetical)* Note: Press TAB Key on every name input.	Specialization	Section	Signature
Group Score (Section A + Section B + Section C)	Alis, Renz Andrei C.	ML	COM222	
	Bongao, Christian I.	ML	COM223	
	Diaz, Kris Brian V.	ML	COM223	
	Layos, Miguel Raphael	ML	COM223	
	Manzanero, Brix Anthony G.	ML	COM223	

Overall Remarks and Recommendation

Instruction(s): Panel must provide clear and thorough feedback, comments or recommendations to improve the proponent's project.

1. Revise specific objectives
2. Finalize w/c algorithm to use and provide discussion to justify the need or use of this
3.

Evaluated by

Signature over Printed Name
Panel Member