THESIS PROPOSAL

Title: A Retrieval-Augmented Generation Approach with Fine-Tuned SahabatAI for Indonesian Consular Question Answering

Abstract:
The provision of efficient and accessible consular services is paramount for citizens abroad. The Indonesian Ministry of Foreign Affairs (MoFA) has initiated digital platforms like "Peduli WNI," "Safe Travel," and the "SARI" chatbot. However, there remains an opportunity to leverage advanced AI for more comprehensive and nuanced query handling. This research aims to design, develop, and evaluate an AI-powered chatbot employing Retrieval-Augmented Generation (RAG) and a fine-tuned SahabatAI (Gemma2-based) Large Language Model (LLM) for Indonesian consular question answering. The proposed methodology involves curating a domain-specific knowledge base from official MoFA sources, implementing a RAG pipeline with SahabatAI as the generator, and investigating Parameter-Efficient Fine-Tuning (PEFT) techniques to adapt SahabatAI to the consular domain. Evaluation will utilize metrics assessing retrieval accuracy, generation quality (faithfulness, relevance), and overall task performance. The expected outcome is a robust chatbot prototype demonstrating improved efficiency and accuracy in delivering consular information, offering insights into the practical application of localized LLMs and RAG for e-governance in Indonesia, and providing a feasible solution within a six-month Master's thesis timeframe.

1. Introduction

1.1. Background

Consular services represent a cornerstone of a nation's support for its citizens abroad. These services, which include the issuance of passports and visas, provision of emergency assistance, facilitation of self-reporting for citizens residing overseas, and management of case reports, are vital for ensuring the safety, well-being, and legal standing of individuals in foreign territories. The Indonesian Ministry of Foreign Affairs (MoFA) is tasked with serving a substantial global diaspora and a large number of citizens traveling internationally, which underscores the necessity for highly efficient, accessible, and responsive support mechanisms.

In response to these demands, the Indonesian MoFA has proactively embraced digital transformation to enhance its consular service delivery. This commitment is evident in its existing digital ecosystem:

- **Portal Peduli WNI:** This web-based platform serves as a central hub for Indonesian citizens abroad, offering critical features such as *Lapor Diri* (Self-Reporting), *Pelayanan Kekonsuleran* (Consular Services), and *Pengaduan Kasus* (Case Reporting). The portal has significantly streamlined processes that

previously necessitated physical visits to Indonesian embassies or consulates, allowing services to be accessed online with internet connectivity.[1]

- **Safe Travel App:** A mobile application designed for Indonesian citizens undertaking short trips abroad, although it can also be utilized by expatriates. It provides practical country-specific information (e.g., time differences, security conditions, local laws and customs, immigration requirements, health services at Indonesian missions), travel registration, notifications (appeals, advice, warnings), and crucial emergency assistance features. In critical situations, users can send their location, record video, and contact the nearest Indonesian mission.[1]
- **SARI (Sahabat Artifisial Migran Indonesia):** This AI-powered chatbot represents a significant step towards leveraging advanced technology for citizen protection. Developed in collaboration with UN Women, SARI is specifically tailored to assist and protect Indonesian female migrant workers from potential violence and exploitation.[5] Integrated within the Safe Travel application, SARI aims to deliver accessible, unbiased, and non-discriminatory information. A key feature is its designed capacity for empathetic responses and its ability to understand local languages, such as Javanese, in addition to Bahasa Indonesia.[1] The launch of SARI underscores MoFA's commitment to "digital empathy" and delivering excellent service and protection, particularly for vulnerable groups.[5] As of early 2025, SARI was reported to be undergoing testing before its official launch within the Safe Travel platform.[7]

The global landscape of public service delivery is increasingly being reshaped by advancements in Artificial Intelligence (AI), particularly Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) techniques.[9] These technologies offer transformative potential, including 24/7 citizen assistance, the capacity to handle complex and nuanced queries, multilingual support, and the ability to personalize interactions, thereby enhancing the efficiency and effectiveness of government services.[12]

A pivotal development in Indonesia's AI journey is **SahabatAI**, the nation's first foundational LLM designed explicitly for Bahasa Indonesia and its various local dialects.[14] This collaborative initiative by Indosat Ooredoo Hutchison and GoTo Group is built upon Google's Gemma2 architecture.[15] SahabatAI has been trained with an extensive dataset comprising over 640,000 instruction-completion pairs, covering Bahasa Indonesia, Javanese, and Sundanese, with plans to include other regional languages like Batak and Balinese.[15] The core objectives of SahabatAI include promoting linguistic diversity, fostering AI sovereignty for Indonesia, and enabling seamless business-to-government (B2G) and business-to-business (B2B) interactions

to significantly enhance the quality and accessibility of government services.[14] Potential use cases for SahabatAI in the public sector include simplifying applications for the national identity card (KTP), demystifying taxation processes, and streamlining procedures for official document changes related to life events such as marriage or relocation.[14] Importantly, SahabatAI models are open-source and accessible via platforms like Hugging Face, facilitating research and development.[14]

The strategic direction of MoFA, particularly evidenced by the development of SARI with its emphasis on AI for empathetic and specialized support [1], closely aligns with the broader national ambition embodied by the SahabatAI initiative to improve government services through localized LLMs.[14] SARI's focus on understanding local languages like Javanese and providing nuanced support resonates with SahabatAI's design philosophy. This convergence suggests a coordinated, top-down strategic impetus within the Indonesian government for adopting sophisticated AI solutions for citizen-centric services, with MoFA playing an active role. The proposed research seeks to build upon this existing strategic momentum by extending the application of a SahabatAI-based system to a wider array of general consular services, complementing SARI's specialized focus.

Furthermore, the existence of SahabatAI as an open-source [18] Gemma2-based LLM, extensively trained in Indonesian and its local dialects [15], provides a robust and highly relevant technological foundation for this thesis. Developing an LLM from scratch is a resource-intensive endeavor, typically requiring multi-year efforts and significant computational power, far exceeding the scope of a Master's thesis. By leveraging a pre-trained, localized, and powerful LLM like SahabatAI, this research can concentrate on the application and refinement of AI techniques—specifically RAG and fine-tuning—to address the consular service challenge. This makes the project technically feasible within the typical six-month timeframe allocated for a Master's degree thesis.

1.2. State of the Art (Theory) - Initial Overview

The theoretical underpinnings of this research draw from several key areas within artificial intelligence and its application to service delivery.
Large Language Models (LLMs), primarily constructed using transformer architectures (Vaswani et al., 2017), have demonstrated remarkable capabilities in understanding, processing, and generating human language. This makes them inherently suitable for complex Question Answering (QA) tasks, where they can interpret user queries and generate relevant, coherent responses based on their vast pre-trained knowledge or context provided at inference time (Jurafsky & Martin, 2023).
Retrieval-Augmented Generation (RAG) is a technique developed to enhance the

capabilities of LLMs, particularly in knowledge-intensive tasks (Lewis et al., 2020 [20]). RAG architectures connect LLMs to external, often dynamic, knowledge sources.[24] By retrieving relevant information from these sources and providing it as context to the LLM during answer generation, RAG mitigates common LLM limitations such as factual inaccuracies (hallucinations) and reliance on outdated training data. This grounding in external factual documents is especially critical for domains like consular services, which demand high accuracy and adherence to current policies and regulations.[21]

The application of AI in consular services is a growing trend globally. Governments are increasingly exploring AI-powered solutions, including chatbots for handling frequently asked questions, assisting with visa applications, providing information during crises, and improving overall citizen engagement.[9] For instance, Singapore's GovTech agency has deployed AI chatbots across various government departments, leading to significant reductions in call center workload and faster response times for citizen inquiries.[12] Similarly, the U.S. Department of State has outlined plans to use AI for various consular functions, including passport photo quality assessment, analysis of customer feedback, AI-driven translation services, and an enhanced search and chatbot system for its Travel.State.Gov portal.[28] These examples highlight a global shift towards leveraging AI to make consular services more efficient, accessible, and citizen-centric. Other relevant works include general AI principles (Russell & Norvig, 2016 [29]) and AI applications in diplomacy (Mostafaei et al., 2025 [9]).

1.3. Gap Analysis

Despite the Indonesian MoFA's commendable progress with digital platforms like "Peduli WNI" and "Safe Travel," and the specialized "SARI" chatbot, several gaps and opportunities remain for enhancing consular service delivery through advanced AI.

The current MoFA digital platforms, while providing valuable information and transactional services, primarily function as information repositories or form-based interaction systems.[1] The SARI chatbot, though an innovative AI application, is specifically designed to support Indonesian female migrant workers, focusing on protection against violence and exploitation.[5] This leaves a broader spectrum of general consular inquiries from the wider Indonesian citizenry—covering diverse topics like passport procedures, visa regulations for various countries, assistance for lost documents, and general emergency guidance—without a dedicated, advanced AI-powered conversational interface. These general queries can be complex, nuanced, and often require synthesizing information from multiple official sources. Current systems may not be equipped to handle such multi-turn conversational

interactions or provide comprehensive answers that require understanding implicit user needs. Furthermore, user feedback for existing applications like the Safe Travel app has indicated occasional technical difficulties, such as server connectivity issues and application crashes [4], suggesting room for improvement in the robustness and user experience of digital service channels.

The advent of SahabatAI, Indonesia's national LLM with a strong foundation in Bahasa Indonesia and local dialects like Javanese and Sundanese [15], presents a significant, largely untapped opportunity. Applying state-of-the-art RAG techniques in conjunction with a fine-tuned SahabatAI model can potentially deliver more accurate, contextually relevant, and up-to-date responses to consular queries than could be achieved with generic global LLMs or simpler rule-based chatbot technologies. The specific combination of a localized LLM like SahabatAI with advanced RAG tailored for the Indonesian consular domain remains an underexplored area of research and application.

Moreover, general-purpose LLMs or even generic RAG systems often necessitate significant adaptation to perform optimally in specialized domains such as consular services. This domain is characterized by specific terminologies, intricate regulations, evolving policies, and diverse user needs and contexts.[30] Therefore, fine-tuning SahabatAI to better understand and generate text specific to consular affairs, coupled with the meticulous curation of a dedicated and comprehensive consular knowledge base, is crucial for developing an effective and reliable AI assistant.

The existing MoFA digital tools—Peduli WNI (web-based), Safe Travel (mobile app), and SARI (chatbot integrated within Safe Travel)—while individually valuable, operate with some degree of separation in terms of user interface and scope.[1] A sophisticated RAG-based chatbot, as proposed in this research, could serve as a more unified and intelligent front-end. Such a system could potentially integrate information from, or direct users to, these existing platforms, thereby providing a more seamless and comprehensive user experience for a wider range of consular questions. The knowledge base for the RAG system would ideally be constructed by consolidating information from these diverse official MoFA sources, creating a centralized and reliable information backbone for the AI. This approach could improve the discoverability and accessibility of consular information that might currently be distributed across different platforms or document formats.

1.4. Problem Formulation (Research Questions)

This research aims to address the identified gaps by investigating the following key

research questions:

- **RQ1:** How can a Retrieval-Augmented Generation (RAG) system, leveraging the SahabatAI LLM, be effectively designed and implemented to provide accurate, reliable, and contextually relevant answers to a diverse range of Indonesian consular service queries (e.g., passport renewal, visa applications, emergency assistance procedures)?
- **RQ2:** What Parameter-Efficient Fine-Tuning (PEFT) strategies (e.g., QLoRA, instruction tuning) are most effective for adapting the SahabatAI (Gemma2) model to the specific linguistic nuances, terminology, and information synthesis requirements of the Indonesian consular domain within a RAG framework?
- **RQ3:** How does the performance of the proposed fine-tuned SahabatAI-RAG system compare against baseline models (e.g., zero-shot SahabatAI, naive RAG with non-fine-tuned SahabatAI) and potentially existing MoFA chatbot solutions, in terms of factual accuracy, relevance, coherence, and user-perceived helpfulness?
- **RQ4:** What are the key components and considerations for creating a high-quality, domain-specific knowledge base and question-answering dataset for Indonesian consular services, suitable for training and evaluating the RAG system?
- **RQ5:** What are the practical challenges, ethical considerations (e.g., data privacy, bias, safety), and resource implications for developing and potentially deploying such an AI-powered consular chatbot within the MoFA, and how can these be addressed within a 6-month Master's thesis scope?

1.5. Objective of Research (Research Objectives)

To answer the research questions, this study will pursue the following objectives:

- **RO1:** To design and develop a prototype RAG-based chatbot system for Indonesian consular services, utilizing the SahabatAI (Gemma2) LLM as the core generative model and incorporating a curated knowledge base of official consular information.
- **RO2:** To investigate, implement, and evaluate appropriate PEFT techniques (e.g., QLoRA) for fine-tuning SahabatAI using a custom-developed Indonesian consular QA dataset, aiming to enhance its domain-specific understanding and response generation capabilities.
- **RO3:** To systematically curate and prepare a comprehensive knowledge base from official MoFA documents and resources, and to construct a representative question-answering (QA) dataset covering common Indonesian consular service inquiries.

- **RO4:** To rigorously evaluate the performance of the proposed fine-tuned SahabatAI-RAG system against baseline approaches using a combination of automated metrics (e.g., ROUGE, BLEU, faithfulness, relevance) and human evaluation, focusing on accuracy, coherence, and helpfulness.
- **RO5:** To analyze the feasibility of the proposed system, identify potential challenges in its development and deployment (including ethical considerations and safety), and provide recommendations for future enhancements, all within the constraints of a six-month Master's thesis project.

1.6. Limitation

This research, while ambitious, will be conducted within certain practical and scope-related limitations:

- **Scope of Consular Services:** The investigation will concentrate on a predefined subset of frequently encountered consular services. Examples include general information on passport applications and renewals, common visa categories and their general application requirements, procedures for reporting lost or stolen documents, and general guidance for emergency assistance. Highly personalized case management, real-time crisis intervention, or services requiring significant human discretion are beyond the scope of this automated system.
- **Knowledge Base Currency:** The knowledge base underpinning the RAG system will be constructed from publicly available MoFA information and official documents accessible at the commencement and during the early stages of the project. Due to the six-month project duration, real-time, continuous updates to reflect regulatory changes that may occur during this period might not be fully incorporated into the static knowledge base used for evaluation.
- **Language Focus:** The primary operational language of the chatbot will be Bahasa Indonesia, leveraging the core strengths of the SahabatAI LLM. While SahabatAI has demonstrated some training in Javanese and Sundanese [17], developing comprehensive support and evaluation datasets for multiple local dialects is a significant undertaking and falls outside the primary scope of this thesis.
- **Fine-Tuning Depth:** Given the constraints of a six-month Master's thesis and typical academic computational resources, the fine-tuning efforts will focus on Parameter-Efficient Fine-Tuning (PEFT) techniques, such as QLoRA.[33] Full model retraining of a 9B parameter LLM is not feasible. The extent and complexity of fine-tuning will be contingent on the final size of the curated QA dataset and the accessibility of suitable computational infrastructure (e.g., high-memory GPUs). Recent studies suggest that even datasets around 1000 high-quality samples can

be effective for PEFT.[34]

- **Dataset Scale:** The custom-developed Indonesian consular QA dataset will be of a practical and manageable size for a Master's thesis project, aiming for several hundreds to a few thousand high-quality question-answer pairs. It will be representative but not exhaustive of every conceivable consular query.
- **Evaluation Scale:** While a combination of automated and human evaluation is planned, comprehensive human evaluation across a very large test set will be constrained by time and resources. Human evaluation will be conducted on a statistically relevant and diverse subset of test queries.
- **Deployment:** The primary output of this research will be a research prototype demonstrating the proposed approach's feasibility and performance. It will not be a production-ready system for immediate, large-scale deployment by MoFA.
- **SahabatAI Model Version:** The research will utilize the version of SahabatAI (e.g., Gemma2 9B CPT SahabatAI v1 Instruct [17]) that is publicly available at the commencement of the project. Any subsequent versions or proprietary models released during the project timeline are outside this study's scope.
- **Safety Alignment:** The base SahabatAI models are not pre-aligned for safety.[16] While the methodology will incorporate considerations for safety guardrails (see Section 3.4.7), achieving comprehensive safety alignment is a complex research area in itself and is beyond the primary QA-focused scope of this thesis. The focus will be on mitigating obvious risks through context grounding and basic output filtering.

1.7. Hypothesis (Optional)

The following hypotheses will guide the empirical investigation:

- **H1:** A Retrieval-Augmented Generation (RAG) system incorporating the SahabatAI LLM, fine-tuned with a domain-specific Indonesian consular QA dataset using Parameter-Efficient Fine-Tuning (PEFT), will demonstrate significantly higher factual accuracy and contextual relevance in answering consular queries compared to the baseline SahabatAI model without RAG or fine-tuning.
- **H2:** The proposed fine-tuned SahabatAI-RAG system will achieve better performance on key evaluation metrics (e.g., faithfulness, answer relevance, human-rated helpfulness) than a naive RAG system using a non-fine-tuned SahabatAI model for Indonesian consular question answering.

Table 1.1: Overview of Existing MoFA Digital Consular Platforms

Platform Name	Type	Key Features	Target Users	Reported AI/Technology Used (if known)
Portal Peduli WNI	Web Portal	Lapor Diri (Self-Reporting), Pelayanan Kekonsuleran (Consular Services), Pengaduan Kasus (Case Reporting) [1]	Indonesian citizens residing abroad [1]	Web-based platform [1]
Safe Travel App	Mobile App	Trip registration, country-specific info (security, laws, health), notifications, emergency assistance (location sharing, video recording, direct call to embassy/consulate) [1]	Indonesian citizens traveling abroad for short trips, also expatriates [1]	Mobile application [1]
SARI	Chatbot	Provides accessible, unbiased, non-discriminatory information; empathetic responses; support for female migrant workers against violence/exploitation; language detection (e.g., Javanese) [1]	Indonesian female migrant workers primarily [5]	Artificial Intelligence (AI), Natural Language Processing, integrated into Safe Travel app [1]

This table provides a consolidated view of MoFA's current digital landscape for consular services, establishing the context for the proposed research. By outlining the features and target users of existing platforms, it implicitly highlights the opportunity for a more comprehensive, AI-driven QA system that this thesis aims to develop,

thereby justifying the novelty and necessity of the proposed work.

2. Literature Review

This chapter provides a comprehensive review of the theoretical foundations and existing work relevant to the proposed research. It delves into Large Language Models (LLMs) for Question Answering (QA), Retrieval-Augmented Generation (RAG) techniques, the specifics of SahabatAI and other pertinent LLMs, the application of AI in governmental and consular services, and methodologies for evaluating such systems.

2.1. Large Language Models (LLMs) for Question Answering

LLMs have revolutionized the field of Natural Language Processing (NLP), demonstrating unparalleled capabilities in understanding and generating human-like text. Their success is largely attributable to the **transformer architecture**, first introduced by Vaswani et al. (2017). This architecture's core innovation, the **attention mechanism** (specifically self-attention and multi-head attention), allows models to weigh the importance of different words in an input sequence and capture long-range dependencies and complex contextual relationships, which are crucial for nuanced question answering (Vaswani et al., 2017).

The development of LLMs typically follows a two-stage paradigm:

- **Pre-training:** In this phase, LLMs are trained on vast and diverse text corpora, often sourced from the internet (e.g., Common Crawl, Wikipedia) and large book collections. The training objectives vary depending on the model architecture. Encoder-based models like BERT (Devlin et al., 2019) often use Masked Language Modeling (MLM), where the model predicts masked (hidden) words in a sentence. Decoder-based models, such as those in the GPT family (Radford et al., 2018, 2019; Brown et al., 2020) and Gemma (Meswani et al., 2024), typically employ Causal Language Modeling (CLM), where the model predicts the next word in a sequence. This extensive pre-training phase endows LLMs with broad linguistic understanding, grammatical proficiency, and a significant amount of factual knowledge embedded within their parameters.
- **Fine-tuning:** After pre-training, LLMs are adapted to specific downstream tasks (e.g., question answering, summarization, translation) or specialized domains using smaller, curated datasets.[36] This process refines the model's parameters to enhance its performance on the target task, making its responses more accurate and relevant to the specific application context.

Several **state-of-the-art general and multilingual LLMs** have emerged from leading research institutions and technology companies. Prominent examples include OpenAI's GPT series (e.g., GPT-3.5, GPT-4 [37]), Meta's Llama models (Touvron et al., 2023a, 2023b), Google's Gemini [38] and PaLM families, and Anthropic's Claude models. These models exhibit strong general QA capabilities. Concurrently, there has been a growing focus on developing multilingual LLMs that can perform effectively across a wide range of languages. This is particularly relevant for regions like Southeast Asia, which are characterized by rich linguistic diversity.[39]

Within the Indonesian context, the most significant development is **SahabatAI**.

- **Architecture:** SahabatAI is based on Google's Gemma2 architecture [15], with the specific publicly available instruct-tuned model being gemma2-9b-cpt-sahabatai-v1-instruct.[16] Gemma models are decoder-only transformers.[16] SahabatAI has a context length of 8192 tokens, although some evaluations have used a capped context of 4096 tokens due to inference platform limitations.[17]
- **Development and Collaboration:** This LLM is a collaborative effort, co-initiated by Indosat Ooredoo Hutchison and GoTo Group, and developed in partnership with AI Singapore.[14] The development leveraged NVIDIA's NeMo framework and NIM microservices for model training and deployment.[18]
- **Training Data & Language Capabilities:** SahabatAI has been extensively pre-trained and subsequently instruction-tuned with a strong focus on Bahasa Indonesia. The instruction-tuning dataset includes approximately 448,000 Indonesian instruction-completion pairs. Crucially, it also incorporates a significant Indonesian-dialect pool, featuring 96,000 pairs in Javanese and 98,000 pairs in Sundanese. An additional 129,000 English instruction-completion pairs were included to enhance its multilingual capabilities.[15] The training data comprised a mix of synthetic instructions and publicly available data that was hand-curated by native speakers to ensure quality and cultural relevance.[17] The initiative aims to create models that deeply understand local contexts and cultural nuances.[18]
- **Performance & Benchmarks:** SahabatAI has demonstrated strong performance, reportedly outperforming models like Llama-3.1-8B and sea-lionv3-9B on the SEA HELM (BHASA) evaluation benchmark.[15] It has been evaluated on a variety of tasks within SEA HELM (including QA, Sentiment Analysis, Toxicity Detection, Translation, Summarization, Causal Reasoning, and Natural Language Inference) and also on the IndoMMLU benchmark, which covers examination questions across various subjects and educational levels in Indonesia.[17]

- **Availability:** SahabatAI is positioned as an open-source initiative, with models accessible through Hugging Face.[14] It is released under the Gemma Community License.[17]
- **Limitations:** Like many LLMs, SahabatAI is susceptible to generating factually incorrect information (hallucinations) and can occasionally produce irrelevant content.[16] A critical point for this research is that the publicly released SahabatAI models have not undergone specific safety alignment. Developers and users are explicitly advised to conduct their own safety fine-tuning and implement appropriate security measures.[16]

To provide a broader regional context, several other **Southeast Asian LLMs** have been developed, reflecting the increasing focus on local languages. These include SEA-LION (AI Singapore), SeaLLM (Alibaba), and Sailor (SEA AI Lab & Singapore University of Technology and Design).[39] SEA-LION, for instance, was trained on 11 regional languages, including Indonesian and Javanese.[39] Alibaba's SeaLLM supports 12 regional languages, also including Indonesian and Javanese [39], while Sailor supports Indonesian among others.[39] These initiatives underscore the importance of linguistic diversity and local adaptation in the LLM landscape.

The characteristics of SahabatAI—its Gemma2 architecture, extensive training in Bahasa Indonesia and key local dialects, open-source availability, and focus on instruction following—render it an exceptionally suitable candidate for this thesis. However, its documented limitations, particularly the potential for hallucination and the lack of pre-existing safety alignment, are critical factors that must be proactively addressed within the proposed research methodology. This will involve leveraging RAG to ground responses in factual consular data and incorporating safety guardrails in the system design.

2.2. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for enhancing the capabilities of LLMs, particularly in knowledge-intensive and fact-sensitive applications. The core concept of RAG, as introduced by Lewis et al. (2020) [20], involves coupling a pre-trained LLM (the generator) with an information retrieval system (the retriever). This architecture allows the LLM to access and utilize information from external knowledge sources at inference time, rather than relying solely on the knowledge implicitly stored in its parameters during pre-training.[24]

The typical RAG pipeline operates as follows [42]:

1. **Query Encoding:** The user's input query is transformed into a dense vector

representation (embedding) using a text embedding model.
2. **Document Retrieval:** This query embedding is used to search a pre-indexed collection of documents (the knowledge base, often stored in a vector database). The retriever identifies and fetches the most relevant document chunks based on semantic similarity (e.g., cosine similarity between query and document embeddings).
3. **Context Augmentation:** The retrieved document chunks are concatenated with the original user query to form an augmented prompt.
4. **Answer Generation:** This augmented prompt is then fed to the LLM, which generates a response grounded in both the user's query and the provided contextual information.

The **benefits of RAG** are manifold. It significantly reduces the likelihood of LLM hallucination by anchoring responses to factual data retrieved from the external knowledge base.[21] RAG systems can access up-to-date information without the need for frequent and costly retraining of the entire LLM; the knowledge base can be updated independently. This also enables domain specialization by simply providing a domain-specific knowledge base. Furthermore, RAG can facilitate source attribution, allowing users to verify the information presented in the LLM's response by referring to the source documents.

RAG implementations can range from simple (**Naive RAG**) to more complex (**Advanced RAG**) techniques [45]:

- **Naive RAG:** This is the basic implementation involving direct retrieval based on query similarity and subsequent generation.[45]
- **Advanced RAG:** This encompasses a suite of more sophisticated techniques applied at various stages of the pipeline to improve performance [45]:
    - **Pre-Retrieval Strategies:** Techniques like query expansion (e.g., adding synonyms or related terms), query transformation (e.g., rephrasing the query for clarity), or generating multiple sub-queries from a complex query to retrieve a richer set of documents.
    - **Retrieval Strategies:** Moving beyond simple dense vector retrieval to include hybrid search (combining keyword-based search like BM25 with semantic search), optimizing the choice and fine-tuning of embedding models, or employing graph-based retrieval mechanisms (e.g., Graph RAG, KG-RAG, where knowledge graphs are used to enhance retrieval).[46]
    - **Post-Retrieval/Re-ranking:** After an initial set of documents is retrieved (e.g., top-N candidates), a re-ranking model is used to re-order these documents based on a more fine-grained assessment of their relevance to

the query.[30] Cross-encoder models, which jointly process the query and each candidate document, are often effective for this but are computationally more intensive than bi-encoder retrievers.
    - **Iterative/Multi-hop Retrieval:** For complex questions that require synthesizing information from multiple sources or performing multi-step reasoning, iterative retrieval techniques can be employed. This might involve decomposing the main question into sub-questions, retrieving evidence for each, and then synthesizing an answer.[20] The Collab-RAG framework, for example, proposes using a smaller language model (SLM) to decompose complex queries, with a larger LLM acting as the reader/synthesizer.[47]
    - **Fine-tuning RAG Components:** This involves training the retriever (embedding model) and/or the generator LLM specifically for the RAG task and the target domain. This can improve the alignment between the retriever and generator and enhance the generator's ability to utilize retrieved context effectively.[22]

Embedding models are crucial for the semantic retrieval component of RAG. They convert text (queries and document chunks) into numerical vectors such that semantically similar texts are closer in the vector space. State-of-the-art embedding models include Google's Gemini Embedding (models like text-embedding-004 and the experimental gemini-embedding-exp-03-07), which has shown top performance on the MTEB Multilingual benchmark.[38] Other strong open-source multilingual models like multilingual-e5-large-instruct are also widely used.[51] The performance of these models is often evaluated on benchmarks like MTEB (Massive Text Embedding Benchmark) and its multilingual extension, MMTEB, which cover various tasks and languages.[38] For Indonesian, specific resources like the Indonesian Sentence Embeddings project and its associated benchmarks (e.g., SemRel2024, Indonesian subsets of MIRACL and TyDiQA) provide valuable evaluation points.[53]

The generated embeddings are typically stored and queried using **vector databases**. These databases are optimized for efficient similarity search in high-dimensional spaces, employing Approximate Nearest Neighbor (ANN) algorithms like HNSW (Hierarchical Navigable Small World) or IVF (Inverted File Index).[54] Key features to consider when choosing a vector database include scalability, query latency, support for metadata filtering (allowing hybrid search), and ease of integration. Popular open-source options suitable for academic research include FAISS, Milvus, and Qdrant.[42]

Despite its advantages, RAG systems face several **challenges**:

- **Retrieval Quality:** The "garbage in, garbage out" principle applies; if the retriever fetches irrelevant or low-quality documents, the generator's output will likely be poor.[24] A common issue is the "lost in the middle" problem, where LLMs tend to ignore relevant information if it's buried within a long context of retrieved documents.[27]
- **Generation Quality:** Even with relevant retrieved context, the LLM might still hallucinate, fail to synthesize information coherently from multiple documents, or produce responses that are not faithful to the provided sources.[30]
- **Context Window Limitations:** LLMs have a finite input context window. Effectively summarizing and presenting a large amount of retrieved information to the LLM without exceeding this limit or losing crucial details is a challenge.
- **Evaluation Complexity:** Evaluating a multi-stage RAG pipeline is inherently complex, as it requires assessing the performance of both the retrieval and generation components, as well as their interaction.[50]
- **Safety and Bias:** RAG systems can inherit biases from the knowledge base they access or from the LLM itself. There's also a risk that even if the retrieved documents are safe and factual, a non-safety-aligned LLM might misinterpret or "twist" this information to generate unsafe or misleading outputs.[59]

Given the nature of Indonesian consular information—which often involves legal nuances, specific procedural details, and varying citizen situations—a naive RAG approach may prove insufficient. The complexity and criticality of providing accurate consular advice necessitate the exploration and implementation of advanced RAG techniques. Specifically, effective re-ranking of retrieved documents to ensure high relevance, and potentially iterative retrieval strategies for handling complex, multi-faceted queries, will be crucial for building a robust and reliable system. Furthermore, the inherent limitations of the SahabatAI model, such as its potential for hallucination [16], can be better mitigated by providing it with higher quality, more precisely retrieved context that advanced RAG components can offer.

2.3. Fine-tuning LLMs for Domain-Specific QA and RAG

Fine-tuning pre-trained LLMs is a critical step in adapting them to specific domains or tasks, such as the nuanced requirements of consular question answering. This process adjusts the model's parameters using a smaller, domain-specific dataset, enabling it to learn the particular vocabulary, style, and knowledge patterns relevant to the target application.

A distinction is made between **Full Fine-Tuning (FFT)** and **Parameter-Efficient Fine-Tuning (PEFT)**:

- **Full Fine-Tuning (FFT):** This approach involves updating all the parameters of the pre-trained LLM. While it can lead to significant performance gains, FFT is computationally very expensive, requires large domain-specific datasets, and consumes substantial memory and time. A notable drawback is the risk of "catastrophic forgetting," where the model loses some of its general capabilities learned during pre-training as it specializes in the new task.[49]
- **Parameter-Efficient Fine-Tuning (PEFT):** PEFT methods aim to overcome the limitations of FFT by modifying only a small fraction of the model's parameters or by adding a small number of new, trainable modules while keeping the vast majority of the pre-trained model's weights frozen. This dramatically reduces computational costs, memory requirements, and training time, making fine-tuning feasible even with limited resources. Popular PEFT techniques include:
    - **LoRA (Low-Rank Adaptation)** and **QLoRA (Quantized LoRA):** LoRA introduces small, trainable low-rank matrices into the layers of the transformer model, effectively learning task-specific adaptations without altering the original weights.[35] QLoRA builds upon LoRA by further quantizing the weights of the pre-trained model (e.g., to 4-bit precision), leading to even greater memory savings. QLoRA has proven particularly effective for fine-tuning Gemma-based models like SahabatAI.[33]
    - **Adapter Layers:** These involve inserting small, trainable neural network modules (adapters) between the existing layers of the pre-trained transformer.[35] Only the adapter parameters are updated during fine-tuning.
    - **Prefix Tuning:** This method adds a small set of trainable prefix tokens to the input sequence. The model learns to modulate its behavior based on these learned prefixes, without changing the core LLM parameters.[35]

Instruction Fine-Tuning (IFT) is a specific type of fine-tuning that trains LLMs on datasets composed of (instruction, input, output) triples.[36] This helps the model learn to follow instructions better and perform specific tasks as directed. For QA tasks, this typically involves fine-tuning on (question, context, answer) examples. In the context of RAG, IFT can be applied to:

- Fine-tune the generator LLM to better utilize the retrieved context, to synthesize information from multiple documents more effectively, or to adhere to specific answer formats.
- Jointly train the retriever and generator components to improve their alignment and overall RAG performance.
- **RuleRAG** is an example where symbolic rules are introduced as demonstrations for in-context learning or as part of supervised fine-tuning data to explicitly guide both the retriever (to fetch logically related documents) and the generator (to

produce rule-consistent answers).[48]

- **Joint QA and Question Generation (QG):** Some research explores fine-tuning an LLM to perform both question answering (given context) and question generation (from context). This dual capability can facilitate self-training on synthetically generated QA pairs, thereby improving domain adaptation and context utilization.[22]

A crucial consideration for fine-tuning, especially within the constraints of a Master's thesis, is the **minimum dataset size**. Recent studies, such as the LIMA paper (Zhou et al., 2024, as cited in [34]), suggest that fine-tuning with relatively small but high-quality datasets (e.g., around 1,000 curated examples) can yield surprisingly strong performance. This is because most of the foundational knowledge is already acquired during the extensive pre-training phase.[34] For datasets with fewer than 1,000 examples, PEFT techniques are particularly recommended.[35] This finding makes the creation of a suitable fine-tuning dataset for Indonesian consular QA an achievable goal within a six-month timeframe.

Comparative studies often explore the trade-offs between fine-tuning alone, RAG alone, or a combination of both (FT+RAG):

- **RAG alone** excels in tasks requiring access to external, up-to-date, or proprietary knowledge. It is generally less prone to hallucination if the retriever is accurate and allows for easier knowledge updates by modifying the vector database rather than retraining the LLM.[43]
- **Fine-tuning alone** is effective for teaching the LLM new skills, adapting its style or tone, or instilling deep understanding of specific domain language and nuances. It can lead to higher accuracy on specialized tasks where the required knowledge can be embedded into the model's parameters.[36] However, it risks catastrophic forgetting and the model's knowledge remains static based on its training data cut-off.[49]
- **FT + RAG** is often considered the most powerful approach, combining the strengths of both.[49] Fine-tuning can make the LLM a better "reasoner" or "synthesizer" over the contextual information provided by the RAG system. For example, the generator LLM within a RAG pipeline can be fine-tuned to better handle the structure of retrieved documents, to follow specific instructions for answer generation based on the domain's requirements, or to improve its faithfulness to the provided sources.

For **fine-tuning Gemma architecture models** like SahabatAI, QLoRA is a well-established and effective PEFT method.[33] Libraries such as Hugging Face

Transformers, along with PEFT and TRL (Transformer Reinforcement Learning) libraries (specifically the SFTTrainer), provide robust tools for implementing QLoRA fine-tuning for Gemma models.[33] Keras, with backends like JAX, TensorFlow, or PyTorch, also offers support for LoRA tuning of Gemma models.[61]

The selection of QLoRA as the PEFT strategy for SahabatAI in this thesis is well-supported by its efficiency and effectiveness with Gemma-based models. The feasibility of creating a high-quality, albeit not massive (around 1000 examples), instruction dataset for Indonesian consular QA within the project timeline further strengthens this choice. This approach allows for meaningful domain adaptation without the prohibitive costs of full fine-tuning.

2.4. AI in Government and Consular Services

The adoption of Artificial Intelligence in government and public services is accelerating globally, with numerous countries implementing AI solutions to enhance efficiency, accessibility, and citizen experience.
Global Case Studies provide valuable insights into the potential and practical applications of AI:

- **Singapore:** The GovTech agency has been a pioneer, deploying AI-powered chatbots like Ask Jamie (a virtual assistant on over 70 government websites), HealthBuddy (for healthcare inquiries), and the CPF Chatbot (for financial queries). These systems utilize NLP to provide instant responses in multiple languages, reportedly achieving a 50% reduction in call center workload and 80% faster response times for common citizen inquiries.[12]
- **Japan:** The Meteorological Agency uses an AI-powered earthquake prediction system employing deep learning to analyze seismic data in real-time, improving early warning accuracy.[12]
- **European Union:** The iBorderCtrl project piloted an AI-driven border security system using facial recognition, biometric scanning, and AI lie-detection tools in Hungary, Greece, and Latvia.[12]
- **South Korea:** Seoul implemented AI-powered smart bins that identify waste types, sort recyclables, and optimize collection routes.[12]
- **Brazil:** São Paulo deployed an AI-driven smart traffic management system to adjust signals, predict congestion, and integrate with public transport.[12]
- **United States:** Various federal agencies are leveraging AI. The Food and Drug Administration (FDA) uses AI for detecting counterfeit pharmaceuticals. The Social Security Administration (SSA) employs predictive models in its Quick Disability Determinations (QDD) process to expedite claims.[13] Municipal governments are also using AI for smart energy management, urban planning,

traffic management, and water conservation.[13] Notably, the U.S. Department of State's AI Inventory for 2024 includes planned AI use cases in consular affairs, such as automated passport photo quality checks, analysis of customer feedback using NLP and LLMs, AI-assisted translation of consular content, and an enhanced search and chatbot for its Travel.State.Gov website.[28]

The application of **AI in diplomacy and embassy operations** is also gaining traction. AI is viewed as a transformative technology that can revolutionize how embassies operate and how governments serve their citizens abroad.[29] Potential applications include using predictive models for economic forecasting, data analytics for international trade, AI-assisted crisis management, and leveraging AI for public diplomacy and sentiment analysis.[9] However, challenges such as ensuring data security, addressing ethical concerns, and the need for skilled human resources to manage and interpret AI systems are also recognized.[29]

In the **Indonesian context**, MoFA's existing digital initiatives like Portal Peduli WNI, the Safe Travel app, and particularly the SARI AI chatbot, demonstrate a clear commitment to digital transformation and the adoption of AI to improve consular services.[1] This is further supported by the national SahabatAI initiative, which aims to provide a foundational Indonesian LLM for enhancing various government services.[14]

The **benefits of AI in the public sector** are compelling, including improved operational efficiency through automation of repetitive tasks, delivery of higher-quality and more responsive citizen services, potential enhancements to public safety, and better resource optimization.[10]

However, the deployment of AI, especially LLMs, in the public sector, particularly in countries like Indonesia, faces several **challenges**:

- **Data Infrastructure and Quality:** Limited data infrastructure, challenges in ensuring high data quality, and the complexity of developing sophisticated models can hinder effective AI adoption.[64]
- **Skill Gaps:** A shortage of personnel with the necessary AI and data science skills can impede the development, deployment, and maintenance of AI systems.[65]
- **Implementation Costs:** The initial investment in AI technologies, infrastructure, and training can be substantial.[65]
- **Ethical Considerations:** Ensuring transparency in AI decision-making, preventing and mitigating biases in algorithms and data, and protecting citizen privacy are paramount ethical concerns.[64]
- **Local Context Understanding:** LLMs, especially those not extensively trained on local data, may struggle with understanding specific local contexts, cultural

nuances, and specialized terminology. For example, evaluations using the IndoCareer benchmark have shown that LLMs face difficulties with questions requiring understanding of strong local contexts in Indonesian professional exams, particularly in fields like insurance and finance, and with numerical analysis tasks.[67]
- **Governance and Risk Mitigation:** Establishing robust governance frameworks, clear accountability mechanisms, and effective risk mitigation strategies are essential for responsible AI deployment.[64]

Future trends in this area point towards the emergence of "AI embassies" with increased automation and a transformation of traditional diplomatic roles, necessitating diplomats with technical expertise.[29] There will be a continued emphasis on responsible AI development, focusing on explainability, fairness, transparency, and continuous evaluation and monitoring of AI systems.[64]

While global AI advancements provide a strong impetus, the success of an AI system for Indonesian consular services will critically depend on its ability to be deeply grounded in the local context. This includes understanding Indonesian laws and regulations, specific consular procedures, the diverse needs of Indonesian citizens, and the nuances of Bahasa Indonesia (and potentially key regional dialects). Although SahabatAI represents a significant step forward by providing a base model trained on Indonesian languages [17], the knowledge base for the RAG system must be meticulously curated from official Indonesian consular documents and sources. The challenges highlighted by the IndoCareer benchmark [67], where LLMs struggled with local context, reinforce the critical importance of this domain-specific grounding. Therefore, the data collection and knowledge base curation phase of this thesis (detailed in Chapter 3) is not merely a preparatory step but a core research activity vital to the project's success. The fine-tuning process must also aim to imbue SahabatAI with a better understanding of these local consular nuances.

2.5. Evaluation of RAG and QA Systems

Evaluating the performance of complex AI systems like RAG-based question answering chatbots requires a multi-faceted approach, encompassing metrics that assess individual components as well as the end-to-end system behavior.

Component-wise vs. End-to-End Evaluation: RAG systems consist of distinct components, primarily the retriever and the generator (LLM). Evaluation can focus on the efficacy of each component in isolation (e.g., how well the retriever fetches relevant documents) or on the overall performance of the integrated system in

answering questions.[50]

Metrics for Retrieval Quality: These metrics assess how effectively the retriever identifies and ranks relevant documents from the knowledge base in response to a query:

- **Precision@k:** The proportion of the top-k retrieved documents that are relevant to the query.[68]
- **Recall@k:** The proportion of all relevant documents in the entire knowledge base that are found within the top-k retrieved documents.[68]
- **Mean Reciprocal Rank (MRR):** The average of the reciprocal ranks of the first relevant document retrieved for a set of queries. It is particularly useful when the primary goal is to find one correct answer quickly.[68]
- **Normalized Discounted Cumulative Gain (NDCG@k):** This metric evaluates the ranking quality by considering both the relevance of the retrieved documents and their positions in the ranked list. It assigns higher scores if more relevant documents are ranked higher.[68]

Metrics for Generation Quality (LLM Output): These metrics evaluate the quality of the answers generated by the LLM, given the query and the retrieved context:

- **Lexical Similarity (Reference-based):** These metrics compare the generated answer to one or more human-written reference answers:
  - **BLEU (Bilingual Evaluation Understudy):** Primarily used in machine translation, BLEU measures the n-gram precision overlap between the generated text and reference texts.[68]
  - **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Commonly used for evaluating summaries and, by extension, QA. It measures n-gram recall overlap (e.g., ROUGE-N for n-gram overlap, ROUGE-L for longest common subsequence).[68]
  - **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** Considers synonymy and stemming, performing an alignment between the generated and reference texts.[68]
- **Semantic Similarity:** Can be reference-based (comparing embedding of generated answer to reference answer) or model-based (using another model to judge similarity).
- **Factuality & Faithfulness (Groundedness):** This is a critical set of metrics for RAG systems. It assesses whether the generated answer accurately reflects the information present in the provided source documents (retrieved context) and is free from hallucinations (information not supported by the context).[49]
- **Answer Relevance:** Measures whether the generated answer is pertinent and

directly addresses the user's question.[58]

- **Coherence & Fluency:** Assesses if the answer is well-structured, grammatically correct, and easy for a human to understand.[60]
- **Completeness:** Evaluates whether the answer covers all important aspects of the question adequately.[60]
- **Perplexity:** A measure of how well the LLM predicts the sample text; lower perplexity generally indicates better language modeling performance.[70]
- **LLM-as-a-Judge:** This approach involves using a powerful, often proprietary, LLM (like GPT-4) to evaluate the outputs of another LLM based on predefined criteria (e.g., helpfulness, correctness, coherence, harmlessness) specified in a prompt.[34] This can offer a scalable alternative or complement to human evaluation.

End-to-End QA Performance: These metrics assess the overall ability of the system to answer questions correctly:

- **Exact Match (EM):** The percentage of generated answers that exactly match the ground truth answer. More suitable for extractive QA or short-answer questions.
- **F1-score:** The harmonic mean of precision and recall, often used for evaluating extractive QA where answers are spans of text.
- **Human Evaluation:** Involves human assessors rating the quality of answers based on criteria like accuracy, completeness, helpfulness, fluency, and overall satisfaction. While resource-intensive, human evaluation is often considered the gold standard for assessing nuanced aspects of language generation.[68]

User-Centric Evaluation: These metrics focus on the user's experience and ability to achieve their goals with the system:

- **Task Success Rate:** The percentage of users who can successfully find the information they need or complete their intended task using the chatbot.
- **User Satisfaction (CSAT):** Measured through surveys or ratings provided by users after interacting with the system.[68]
- Other metrics like Total Users, Active Users, and Engaged Users are more relevant for monitoring production systems but can provide insights into overall utility and engagement.[72]

Benchmarking Datasets and Frameworks: While standard QA datasets like SQuAD, Natural Questions, TriviaQA, and MS MARCO exist [42], domain-specific QA often requires the creation of custom datasets. For example, ChemLit-QA serves the chemistry domain.[31] The need for high-quality, domain-specific evaluation datasets for RAG systems is well-recognized.[50] Several frameworks have been developed to

streamline RAG evaluation, such as RAGAS [50], ARES [50], DeepEval [68], UpTrain [68], and Tonic Validate.[68]

Evaluating Low-Resource Language LLMs: This presents unique challenges due to the scarcity of appropriate benchmarks and the need for culturally relevant and linguistically accurate evaluation materials and criteria.[41]

Given the high-stakes nature of consular information, where accuracy, reliability, and clarity are paramount, a multi-faceted evaluation strategy is indispensable for the proposed system. While automated metrics like ROUGE and BLEU can provide quick feedback during development iterations, they are insufficient on their own to capture the full spectrum of quality. Therefore, metrics focusing on faithfulness (groundedness to consular regulations), answer relevance, and coherence, potentially assessed using an LLM-as-a-Judge approach, will be critical. Ultimately, human evaluation, even if conducted on a smaller, representative subset of queries, will be essential for validating the practical helpfulness and trustworthiness of the AI-powered consular chatbot.

2.6. Identification of Gaps and Justification for Proposed Work

Recap of Identified Gaps:
The preceding review of literature and existing systems highlights several key gaps that this research aims to address:

1. **Lack of Comprehensive AI QA for General Consular Services at MoFA:** While the Indonesian MoFA has made strides with digital platforms like "Peduli WNI," "Safe Travel," and the specialized "SARI" chatbot [1], there is currently no comprehensive, advanced AI-powered question-answering system designed to handle the diverse range of general consular inquiries from all Indonesian citizens. SARI's focus, while important, is specific to female migrant workers and protection issues.
2. **Need for Domain-Specific Adaptation of Localized LLMs:** Powerful Indonesian LLMs like SahabatAI exist and are tailored for Bahasa Indonesia and local dialects.[14] However, their general-purpose training means they require significant domain-specific adaptation—through both curated knowledge via RAG and fine-tuning—to be effective for specialized and nuanced tasks like providing accurate consular advice. The current demonstrated use cases for SahabatAI (e.g., KTP information, taxation [14]) are generally simpler than the complex, often legally-grounded, advice required in consular affairs.
3. **Scarcity of Indonesian Consular QA Datasets and Benchmarks:** There is a notable absence of publicly available, high-quality question-answering datasets

and RAG benchmarks specifically for the Indonesian consular domain. Such resources are crucial for training, evaluating, and advancing AI solutions in this area.

4. **Limited Research on Advanced RAG with Fine-Tuned Local LLMs for Indonesian Public Sector:** While AI applications in government are growing, there is limited published research focusing on the application of advanced RAG techniques, particularly those involving fine-tuned localized LLMs like SahabatAI, to address specific challenges within the Indonesian public sector, especially in the critical domain of consular services.

Justification for the Proposed Work:
This research is justified by its potential to address these gaps and make significant contributions:

- **Novelty:** The proposed work is novel in its specific application of a fine-tuned, localized Indonesian LLM (SahabatAI, based on Gemma2) within an advanced Retrieval-Augmented Generation framework to tackle the problem of providing comprehensive and accurate answers to Indonesian consular service queries. This combination of specific local LLM technology, advanced RAG architecture, and the consular domain in Indonesia represents an unexplored research avenue.
- **Contribution:** This thesis is expected to make several valuable contributions:
    - **System Prototype:** The development of a functional prototype system will demonstrate the feasibility and potential benefits of the proposed AI-driven approach for enhancing Indonesian consular services.
    - **Domain-Specific Resources:** A key contribution will be the creation of a curated Indonesian consular knowledge base and a corresponding question-answering dataset. These resources, developed in Bahasa Indonesia, will be valuable for this research and potentially for future work in this domain by other researchers or MoFA itself.
    - **Fine-tuning Insights:** The research will provide practical insights into effective Parameter-Efficient Fine-Tuning (PEFT) strategies (specifically QLoRA) for adapting the SahabatAI (Gemma2) model for specialized question-answering tasks within a RAG context.
    - **Performance Benchmarks:** The rigorous evaluation and comparison of the proposed system against baseline models will establish performance benchmarks for AI-powered consular QA in the Indonesian context, informing future development efforts.
    - **Addressing Real-World Needs:** The research directly addresses a real-world need for improved efficiency, accessibility, and accuracy in the delivery of consular services to Indonesian citizens, potentially leading to enhanced

citizen satisfaction and better resource allocation for MoFA.

- **Timeliness:** The proposed research aligns strategically with the Indonesian MoFA's ongoing efforts towards digital transformation and its explicit interest in leveraging AI, as evidenced by the SARI initiative.[1] It also resonates with the broader Indonesian government's push for AI adoption and digital sovereignty, supported by national initiatives like SahabatAI.

Table 2.1: Comparison of State-of-the-Art RAG Techniques

Technique Name	Description	Key Advantages	Key Disadvantages /Challenges	Relevant Citations
Naive RAG	Basic pipeline: retrieve relevant documents, concatenate with query, feed to LLM for generation.	Simple to implement, provides baseline.	Prone to irrelevant retrieval, context window issues, suboptimal generation if context is noisy.	[45], Lewis et al. (2020) [20]
Re-ranking	Adds a step after initial retrieval to re-score and re-order documents for relevance before passing to LLM.	Improves quality of context fed to LLM, reduces noise.	Adds computational overhead, requires a good re-ranking model.	[30]
Iterative/Multi-hop Retrieval	Decomposes complex queries or iteratively refines retrieval based on intermediate results to gather comprehensive evidence.	Better for complex questions requiring synthesis from multiple sources.	Increased complexity, potential for error propagation, higher latency.	[20]
Self-RAG	LLM learns to	More adaptive	Requires	Asai et al. (as

	retrieve and self-critique retrieved passages, deciding if they are needed for generation.	retrieval, can decide not to retrieve if internal knowledge is sufficient.	specialized training for the LLM.	cited in [32])
Collab-RAG	Collaboration between a white-box Small Language Model (SLM) for query decomposition and a black-box LLM for reading/synthesis.	Leverages strengths of both SLMs (efficiency for decomposition) and LLMs (reasoning), can improve complex QA.	Requires managing interaction between two models, SLM decomposition quality is crucial.	[47]
Fine-tuned Retriever/Generator in RAG	Fine-tuning the embedding model (retriever) or the generator LLM on domain-specific data or for better RAG task alignment.	Can significantly improve domain adaptation, retrieval accuracy, and generation quality tailored to the RAG process.	Requires domain-specific training data, computational resources for fine-tuning.	[22]
Graph RAG / KG-RAG	Utilizes knowledge graphs or graph structures of documents for more structured and context-aware retrieval.	Can capture complex relationships between information entities, potentially leading to more precise retrieval.	Requires construction and maintenance of knowledge graphs, more complex retrieval algorithms.	[46]

This table offers a structured overview of various RAG techniques, facilitating a clear comparison of their respective advantages and disadvantages. This understanding is crucial for justifying the selection of specific advanced RAG components (such as re-ranking and potentially iterative elements for complex queries) for the proposed

system in Chapter 3, demonstrating a thorough grasp of the state-of-the-art.

Table 2.2: Overview of Relevant Indonesian and Multilingual LLMs/Embedding Models

Model Name	Developer(s)	Architecture (if known)	Key Features	Performance on Relevant Benchmarks	Availability
SahabatAI (Gemma2 9B CPT v1 Instruct)	GoTo Group, Indosat, AI Singapore	Gemma2 (Decoder-only)	9B parameters, Bahasa Indonesia, Javanese, Sundanese, English support; Instruction-tuned; 8192 context length.	Outperforms Llama-3.1-8B on SEA HELM (BHASA); Evaluated on IndoMMLU. [15]	Open-source (Hugging Face) [18]
Google Gemini Embedding (exp-03-07)	Google	Gemini-based	>100 languages supported, 3K output dimensions, longer input tokens.	Top rank on MTEB Multilingual (mean score 68.32). [38]	API (Experimental)
multilingual-e5-large-instruct	Microsoft (based on E5)	Transformer (Encoder)	Supports many languages, instruction-tuned for better embeddings.	Strong performance on MTEB, often a top open-source multilingual embedding model. [51]	Open-source (Hugging Face)
IndoBERT	Various Indonesian	BERT (Encoder)	Pre-trained specifically	Baseline for many	Open-source (Hugging

	NLP researchers (e.g., IndoNLU team)		on large Indonesian corpora.	Indonesian NLP tasks; performance varies by specific task and fine-tuning.	Face)
SEA-LION	AI Singapore	Transformer	11+ SE Asian languages including Indonesian, Javanese. Trained on 26x more SE Asian language data than Llama-2.	Good performance on regional benchmarks. [39]	Open-sourc e
SeaLLM	Alibaba DAMO Academy	Transformer	12 SE Asian languages including Indonesian, Javanese. Processes longer non-Latin text than ChatGPT.	Strong regional performance . [39]	Likely proprietary
Indonesian Sentence Embeddings Models	LazarusNLP (Community Project)	SimCSE, ConGen	Models specifically trained for Indonesian sentence embeddings.	Evaluated on Indonesian STS (SemRel202 4), MIRACL, TyDiQA Indonesian subsets. [53]	Open-sourc e (Hugging Face)

This table provides a focused comparison of LLMs and embedding models that are most pertinent to the thesis. It serves to justify the selection of SahabatAI as the core generative model due to its Indonesian focus and highlights strong candidates for the embedding model component of the RAG system. This demonstrates an awareness of

the relevant LLM landscape.

3. Method

This chapter details the systematic approach that will be undertaken to achieve the research objectives. It outlines the research design, data collection and preprocessing procedures, the proposed system architecture including the RAG pipeline and LLM fine-tuning, and the comprehensive evaluation strategy.

3.1. Research Design

This research will employ a **Design Science Research (DSR)** methodology. DSR is an appropriate paradigm for this study as it focuses on the creation and evaluation of innovative IT artifacts—in this instance, the AI-powered consular chatbot—designed to solve practical problems within a specific organizational or societal context (Hevner et al., 2004). The DSR process is inherently iterative, involving cycles of artifact construction (build), use and evaluation (evaluate), and subsequent refinement based on the evaluation findings.

The research will be structured into the following distinct, sequential phases, allowing for a systematic progression towards the research goals within the six-month timeframe:

1. **Phase 1: Foundational Work and Data Strategy (Weeks 1-4):** This initial phase involves finalizing the in-depth literature review, refining research questions and objectives, preparing any necessary ethics documentation, and meticulously mapping official MoFA data sources. The development environment will also be set up during this period.
2. **Phase 2: Knowledge Base and QA Dataset Curation (Weeks 5-10):** This phase is dedicated to the intensive collection of consular information from identified MoFA sources to build the knowledge base (KB). Concurrently, the strategy for generating the Indonesian consular Question-Answering (QA) dataset will be implemented, involving both manual crafting and potentially LLM-assisted generation of QA pairs. This phase will also include initial data cleaning and structuring.
3. **Phase 3: Baseline RAG System Implementation and Dataset Finalization (Weeks 11-14):** The core RAG pipeline will be developed using the selected SahabatAI model (initially without fine-tuning) and a chosen embedding model. The KB will be preprocessed (chunked) and indexed into the vector database. The QA dataset will be finalized, validated, and split into training, validation, and test sets. Baseline performance of this naive RAG system will be established.

4. **Phase 4: SahabatAI Fine-tuning and Advanced RAG Integration (Weeks 15-18):** Parameter-Efficient Fine-Tuning (PEFT), specifically QLoRA, will be applied to the SahabatAI model using the consular QA training set. Advanced RAG components, such as a re-ranking mechanism, will be integrated into the pipeline. The fine-tuned SahabatAI model will replace the baseline generator in the RAG system.
5. **Phase 5: Comprehensive System Evaluation and Analysis (Weeks 19-22):** Rigorous evaluation of the enhanced (fine-tuned SahabatAI + Advanced RAG) system will be conducted using the test set and defined metrics. This includes comparing its performance against the baseline systems. Ablation studies will be performed to understand the contribution of different components.
6. **Phase 6: Thesis Compilation, Reporting, and Submission (Weeks 23-26):** The final phase involves the intensive writing and compilation of the Master's thesis, documenting the entire research process, methodologies, results, discussion, conclusions, and future work.

This phased approach ensures a structured progression, with clear milestones and deliverables for each stage, facilitating project management and timely completion.

3.2. Data Collection

The quality and relevance of the data are paramount to the success of this RAG-based QA system. Data collection will focus on two primary artifacts: the knowledge base for the RAG system and the question-answering dataset for fine-tuning and evaluation.

3.2.1. Knowledge Base (KB) Creation for RAG

The KB will serve as the external source of truth for the RAG system, providing the factual information upon which the SahabatAI model will base its answers.

- **Sources:** The primary sources for KB construction will be official and publicly accessible materials from the Indonesian Ministry of Foreign Affairs and related government entities. These include:
  - The main MoFA website (kemlu.go.id) and the official websites of major Indonesian embassies and consulates general (e.g., the Indonesian Embassy in Washington D.C. [73], and others as identified through systematic search).
  - Content derived from the "Portal Peduli WNI".[1]
  - Relevant informational content available through the "Safe Travel" mobile application (if accessible in text format or through web interfaces).[1]
  - Publicly available Indonesian consular regulations, governmental decrees, official circulars, and announcements pertaining to passport issuance and

renewal, visa categories and application procedures, citizen protection guidelines, legal assistance frameworks, etc.
    - Frequently Asked Questions (FAQs) sections published on MoFA and embassy/consulate websites.
    - The official e-Visa website for Indonesia (evisa.imigrasi.go.id) for detailed information on electronic visa applications.[74]
- **Data Types:** The collected data will primarily be textual, including HTML content from web pages, text extracted from PDF documents (e.g., regulations, informational brochures, application forms), and potentially structured information from tables found on official websites.
- **Collection Methods:**
    - **Manual Curation:** Key documents, specific regulations, and targeted high-value content will be manually downloaded and reviewed.
    - **Web Scraping:** For systematic collection of textual content from designated MoFA and embassy web pages, Python libraries such as BeautifulSoup and Scrapy will be employed. Ethical web scraping practices will be strictly adhered to, including respecting robots.txt directives, limiting request frequency to avoid server overload, and ensuring that only publicly accessible information is collected.
    - **Document Parsing:** Libraries like PyPDF2, pdfminer.six, or python-docx will be used to extract text content from downloaded PDF and Microsoft Word documents.
- **Scope:** The KB will focus on information pertinent to common consular inquiries. This includes, but is not limited to: procedures for passport application and renewal; different types of visas for foreigners wishing to visit Indonesia and general requirements for Indonesians traveling to popular destinations; steps to take when important documents are lost or stolen abroad; contact information for emergency assistance; and guidelines for self-reporting for citizens residing overseas.

3.2.2. Question-Answering (QA) Dataset Generation

This dataset is crucial for two purposes: (1) fine-tuning the SahabatAI LLM to adapt it to the consular domain and improve its instruction-following and synthesis capabilities, and (2) evaluating the performance of the final RAG system. The methodology for creating this dataset will follow best practices outlined in recent literature 42:

1. **Defining Objectives:** The dataset aims to encompass a diverse range of consular queries, including factual questions (e.g., "What is the validity period of an Indonesian passport?"), comparative questions (e.g., "What is the difference between a B211A visa and a Visa on Arrival?"), and exploratory or procedural questions (e.g., "How do I report a lost passport while in Germany?"). The

evaluation goals for using this dataset include assessing factual accuracy, retrieval relevance of the RAG system, and the comprehensiveness and helpfulness of the generated answers.[75]

2. **Identifying Document Sources:** The primary source material for generating QA pairs will be the curated consular knowledge base (developed in Section 3.2.1).
3. **Developing Evaluation Queries (Question Generation):**
    - **Manual Creation:** A significant portion of the questions will be manually crafted by the researcher to simulate realistic user inquiries. This process will involve formulating questions that vary in complexity and cover different aspects of the consular services defined in the scope. Input from individuals with experience in consular matters or expatriate needs may be sought if feasible to enhance the ecological validity of the questions.
    - **Synthetic QA Generation (LLM-assisted):** To augment the dataset and achieve the target size efficiently, a powerful LLM (e.g., GPT-3.5-turbo, GPT-4, or another high-performing model accessible through university resources or APIs) will be used to generate additional question-answer pairs based on selected chunks of text from the curated consular KB.[34] Prompt engineering will be critical here; prompts will be designed to instruct the LLM to generate relevant, answerable questions and to extract or synthesize concise answers strictly from the provided document excerpts. Techniques such as the D-Naive approach (generating QA pairs directly from document chunks) might be explored for this purpose.[34]
4. **Preparing Ideal Responses (Answer Generation/Extraction):**
    - For manually created questions, the ideal answers will be meticulously written by the researcher or directly extracted and verified from the official KB documents.
    - For LLM-generated QA pairs, the synthetically generated answers will undergo a rigorous manual review and editing process. Each answer will be checked for factual accuracy, completeness, and faithfulness to the source document(s) from the KB.
5. **Pairing Queries with Document Passages (Grounding):** Each question-answer pair in the dataset will be explicitly linked to the specific source document(s) or passage(s) within the KB that contain the information required to answer the question. This grounding is essential for evaluating the RAG system's ability to retrieve relevant context and for certain fine-tuning strategies that require context alongside the question and answer.

- **Dataset Size:** The target is to create a high-quality dataset of at least 1,000 QA pairs. Research indicates that PEFT can yield good results with datasets of this magnitude for models in the 7B-9B parameter range.[34] If time and resources

permit, expanding the dataset to around 3,000 pairs would be beneficial for more robust fine-tuning and evaluation.[58] The final dataset will be divided into standard splits: training (e.g., 70%), validation (e.g., 15%), and testing (e.g., 15%).

- **Language:** The dataset will be primarily in Bahasa Indonesia to align with SahabatAI's strengths and the target user base. While SahabatAI has some training in Javanese and Sundanese [17], creating extensive, high-quality QA datasets in these regional dialects is likely beyond the scope of a six-month Master's thesis. The focus will remain on Bahasa Indonesia to ensure depth and quality. However, if common consular phrases or terms in these dialects are identified in official MoFA communications, a very small, targeted test set might be considered for exploratory evaluation if time allows, acknowledging the challenges of low-resource language dataset creation.[41]

3.2.3. Ethical Considerations and Data Privacy

Ethical considerations are paramount throughout the data collection and usage process:

- **Publicly Available Data:** The knowledge base will be constructed exclusively from publicly accessible government information. No personal or private citizen data will be collected, stored, or used in this research.
- **Synthetic Data Transparency:** If LLM-assisted synthetic QA generation is employed, this methodology will be transparently and thoroughly documented in the final thesis. This documentation will include details of the LLM used for generation, the prompting strategies employed, and the validation process applied to the synthetic data.[78] The inherent limitations of synthetic data, such as the potential for introducing subtle biases or not perfectly reflecting real user query patterns, will be acknowledged and discussed.[82]
- **Bias Mitigation in QA Dataset:** During the manual creation of QA pairs and the validation of synthetically generated ones, conscious efforts will be made to avoid introducing or perpetuating biases. Questions will be formulated to be neutral and to cover a diverse range of scenarios fairly, without making assumptions about user demographics or backgrounds, unless directly relevant to a specific consular regulation (e.g., age requirements for a passport).
- **Responsible AI Principles:** The entire development process will adhere to principles of responsible AI, emphasizing the creation of a system that is accurate, reliable, and avoids generating harmful, misleading, or inappropriate outputs.[64]

The systematic collection, cleaning, structuring of Indonesian consular documents into a usable knowledge base, and the subsequent creation of a validated, domain-specific QA dataset in Bahasa Indonesia are, in themselves, significant undertakings. Given the general scarcity of such specialized resources for the

Indonesian language, these data artifacts will represent a key contribution of this thesis, directly impacting the performance of the RAG system and the efficacy of the fine-tuning process. The quality and comprehensiveness of this data will be a critical determinant of the overall success of the research.

3.3. Data Preprocessing

Once the raw data for the knowledge base and QA dataset has been collected, several preprocessing steps will be necessary to prepare it for use in the RAG system and for LLM fine-tuning.

3.3.1. Text Cleaning

Raw text data, especially from web scraping or PDF extraction, often contains noise and artifacts that need to be removed:

- **Markup Removal:** HTML tags (e.g., , <p>, <a>), JavaScript code snippets, and CSS style definitions will be stripped from content scraped from web pages to retain only the core textual information.
- **Text Normalization:** This includes converting all text to a consistent case (typically lowercase, though the impact on Indonesian language models will be considered), handling or removing special characters and symbols that are not part of standard language, and normalizing or removing excessive whitespace (e.g., multiple spaces, redundant line breaks).
- **Standardization (if necessary):** Common entities like dates, numbers, or specific official terminologies might require standardization to ensure consistency across the knowledge base, although this will be applied judiciously to avoid altering the original meaning of official texts.

3.3.2. Document Chunking for RAG Indexing

To make the knowledge base searchable and usable by the RAG system, the collected documents need to be divided into smaller, manageable segments or "chunks." The choice of chunking strategy is critical as it impacts retrieval relevance and the context provided to the LLM.56

- **Strategy Selection:** Several chunking strategies will be considered and evaluated:
    - **Fixed-Size Chunking:** This straightforward method involves splitting text into chunks of a predefined number of tokens or characters, often with a specified overlap between consecutive chunks (e.g., 500 tokens per chunk with a 50-token overlap).[84] While simple to implement, it risks breaking sentences or semantic units, potentially harming context.
    - **Recursive Character Text Splitting:** This adaptive approach uses a predefined list of separator characters (e.g., paragraph breaks \n\n, sentence

terminators ., then single spaces) in a hierarchical order.[84] It attempts to split along the most semantically meaningful boundaries first, making it generally superior to fixed-size chunking for preserving context. This is often a good default strategy.
    - **Document-Based or Section-Based Chunking:** For documents with clear intrinsic structures, such as official regulations divided into articles and sections, or FAQs with distinct question-answer blocks, chunking can be performed at these logical boundaries.[84] This approach is highly effective for preserving the semantic integrity of structured content and is anticipated to be very relevant for many Indonesian consular documents.
    - **Content-Aware Chunking (Advanced, for consideration):** More advanced techniques might involve using NLP models to identify semantic breaks in the text. While potentially offering the best semantic coherence, implementing and fine-tuning such models might be beyond the primary scope but could be discussed as future work.
- **Justification of Chosen Strategy:** The final chunking strategy will be chosen and justified based on an analysis of the typical structure and nature of Indonesian consular documents. A hybrid approach might be adopted, for example, using section-based chunking for highly structured regulatory documents and recursive character splitting for less structured web content or FAQs. The goal is to maximize contextual coherence within each chunk.
- **Chunk Size and Overlap:** Experimentation will be conducted with different chunk sizes (e.g., ranging from 256 to 1024 tokens) and overlap percentages (e.g., 10-20% of chunk size). The optimal chunk size will balance the need for providing sufficient context to the LLM against the risk of including irrelevant information or exceeding the input limits of the embedding model or the LLM itself. Overlapping chunks help ensure that information is not lost at chunk boundaries.[85]

3.3.3. Metadata Extraction and Association

For each generated chunk, relevant metadata will be extracted or assigned and stored alongside the chunk text and its embedding in the vector database. This metadata may include:

- Source URL or original document filename.
- Document title or a descriptive name.
- Original document type (e.g., "Regulation," "FAQ," "Web Page").
- Date of publication or last modification of the source document (if available).
- Section headings or other structural information from the original document. This metadata is valuable for several reasons: it can be used to filter search results during retrieval (e.g., retrieve only information from regulations published after a

certain date), to provide source attribution for the answers generated by the chatbot, and for debugging and analysis of the RAG system's behavior.

3.3.4. Structuring the Dataset: Q&A Pairs, Document Snippets, and Metadata

To effectively power a RAG-based consular Q&A system, the sourced knowledge must be meticulously structured. This involves transforming raw information into formats that are optimized for retrieval and subsequent generation by an LLM.

1. **Question-Answer (Q&A) Pairs:** A significant portion of the dataset should consist of well-defined Q&A pairs. These can be extracted directly from existing FAQs on MoFA and embassy websites and from analyses of common inquiries received by consular staff.
2. **Document Snippets (Chunks):** Lengthy official documents, such as Permenlu, Renstra, and international agreements, need to be segmented into smaller, semantically coherent and contextually complete snippets or "chunks." This process is critical for effective retrieval, as LLMs have context window limitations, and providing overly long, undifferentiated text can dilute relevance.[82] The chunking strategy must be carefully chosen to preserve the meaning and integrity of legal articles, procedural steps, and policy statements.
3. **Metadata Enrichment:** Each Q&A pair and document chunk must be tagged with rich metadata to facilitate accurate retrieval and contextual understanding by the AI system. Essential metadata fields include:
    - chunk_id or qa_pair_id: A unique identifier.
    - original_document_id: A reference to the source document (e.g., "Permenlu No. 5 Tahun 2018," "FAQ_Kemlu_Visa").
    - document_type: Classification of the source (e.g., Regulation, FAQ, Guideline, News Report, Strategic Plan).
    - text_snippet_indonesian: The actual textual content of the chunk or answer in Bahasa Indonesia.
    - text_snippet_english: The English translation, if available or deemed necessary for multilingual capabilities.
    - keywords: A list of relevant keywords (e.g., "paspor hilang," "visa kerja," "legalisasi ijazah," "kekerasan PMI," "lapor diri").
    - category: A primary categorization (e.g., Passport Services, Visa Services, Document Legalization, WNI Protection, Emergency Assistance, SARI-Specific Support).
    - sub_category (optional): More granular categorization (e.g., "WNI Protection - Human Trafficking," "Visa Services - Schengen").
    - target_audience (optional): Specifying if the information is for WNI umum (general public), PMI wanita (female migrant workers), MoFA staff, etc.

- related_questions: Examples of user questions that this chunk or Q&A pair could effectively answer.
- source_url: If applicable, the URL of the source document or webpage.
- publication_date: Date of the original document's publication.
- last_verified_date: Date when the information in the chunk was last verified for accuracy and currency.
- empathy_level_required (for SARI): A tag indicating if the response should incorporate a specific level or style of empathy.

Table 2: Sample Dataset Snippets for Consular QA

chunk_id	original_document_id	document_type	text_snippet_indonesian	keywords	category	related_questions	last_verified_date
PERMENLU_5_2018_ART8_A	Permenlu No. 5 Tahun 2018	Regulation	"Pelindungan Kekonsuleran...paling sedikit meliputi: a. melindungi kepentingan Negara dan WNI di Negara Penerima berdasarkan ketentuan peraturan perundang-undangan..."	pelindungan WNI, kekonsuleran, kepentingan negara	WNI Protection	"Apa saja bentuk pelindungan kekonsuleran untuk WNI?"	2024-10-01

FAQ_KEMLU_PASPOR_003	kemlu.go.id/faq/paspor	FAQ	"T: Berapa lama proses pembuatan Paspor Diplomatik dan Paspor Dinas? J: Paling lambat 4 (empat) hari kerja setelah semua persyaratan dilengkapi..."	paspor diplomatik, paspor dinas, waktu proses	Passport Services	"Berapa lama bikin paspor dinas?", "Proses paspor diplomatik berapa hari?"	2025-01-15
SARI_KB_EMPATHY_001	Modul Pelatihan SARI - Empati	Guideline	"Saya memahami ini pasti situasi yang sangat sulit bagi Anda. Mari kita coba cari solusi bersama. Bisakah Anda ceritakan lebih lanjut apa yang terjadi?"	empati, dukungan, kekerasan, pekerja migran, lapor	WNI Protection - Migrant Worker Support	"Saya takut dan butuh bantuan.", "Saya mengalami masalah."	2025-03-01

RENSTRA_KEMLU_2024_DIGI	Renstra Kemlu 2020-2024, Bab IV	Strategic Plan	"Transformasi digital untuk mewujudkan pelayanan tepat, mudah, murah dan akurat melalui Portal Peduli WNI dan Safe Travel."	transformasi digital, Portal Peduli WNI, Safe Travel	MoFA Strategy - Digitalization	"Apa strategi digital Kemlu untuk layanan WNI?"	2024-11-01
LEGAL_DOC_IJAZAH_001	Laman Legalisasi Dokumen Kemlu.go.id	Web Content	"Untuk legalisasi ijazah yang diterbitkan di Indonesia dan akan digunakan di luar negeri, dokumen harus dilegalisasi terlebih dahulu oleh Kemendikbud..."	legalisasi, ijazah, dokumen, persyaratan, Kemendikbud	Document Legalization	"Bagaimana cara legalisasi ijazah untuk dipakai di luar negeri?"	2025-02-20

This structured dataset, particularly the sample table, concretely illustrates how diverse consular information can be transformed into a machine-usable format. For MoFA, this clarifies the data preparation lifecycle, emphasizing the crucial role of

comprehensive metadata for retrieval accuracy and the inclusion of specialized content, like empathetic dialogue examples, to meet SARI's unique objectives.

The dataset requires careful curation. Beyond factual information, it must incorporate extensive examples of empathetic phrasing, active listening prompts, culturally sensitive language, and dialogue flows designed to guide and support users in distress. This may necessitate the generation of new content or the careful adaptation of existing materials, drawing from the insights gained during human-centered design processes involving consultations with WNI. This specialized content is vital for training or fine-tuning the LLM to deliver genuinely supportive and non-judgmental interactions.

3.4. Data Processing (Proposed System Architecture)

The proposed AI-powered consular chatbot will be built around a RAG architecture, integrating several key components. The selection of each component is based on its suitability for the Indonesian language, the consular domain, and feasibility within a Master's thesis.

3.4.1. Core Large Language Model (LLM)

- **Model:** The primary LLM for answer generation will be **SahabatAI**, specifically the gemma2-9b-cpt-sahabatai-v1-instruct model or the latest stable version available at the project's commencement.[16]
- **Justification:** This choice is driven by several factors:
  - **Indonesian Language Proficiency:** SahabatAI is pre-trained and instruction-tuned extensively on Bahasa Indonesia, including local dialects like Javanese and Sundanese, making it uniquely suited for understanding and generating text in the target language of Indonesian consular services.[15]
  - **Architecture for PEFT:** Its Gemma2 architecture is known to be amenable to Parameter-Efficient Fine-Tuning techniques like QLoRA, which is crucial for adapting the model within the resource constraints of this project.[33]
  - **Open-Source Availability:** SahabatAI is open-source and accessible via Hugging Face, facilitating its use in academic research.[14]
  - **Instruction Following:** The instruct version is specifically tuned to follow instructions, which is beneficial for a QA chatbot that needs to adhere to prompts and utilize provided context.[17]
  - **Manageable Size:** At 9 billion parameters, it is a powerful model, yet its size is manageable for fine-tuning using PEFT with readily available academic

resources (e.g., Google Colab Pro with high-RAM GPUs, or university High-Performance Computing clusters).

3.4.2. Embedding Model for Retrieval

The choice of embedding model is critical for the RAG system's ability to retrieve relevant document chunks.

- **Selection Criteria:** The selection will be based on:
  - Demonstrated performance on multilingual or specifically Indonesian text retrieval benchmarks (e.g., MTEB, MMTEB, results from the Indonesian Sentence Embeddings project [53]).
  - Embedding dimensionality (balancing richness of representation with computational cost).
  - Maximum input context length.
  - Computational efficiency for generating embeddings.
  - Ease of integration with Python-based RAG frameworks.
  - Support for Bahasa Indonesia.
  - References:.[38]
- **Candidate Models:**
  - **Google's Gemini Embedding (gemini-embedding-exp-03-07):** This model has shown state-of-the-art performance on the MTEB Multilingual leaderboard and supports over 100 languages, including Bahasa Indonesia. Its high dimensionality (3072) and longer input token capacity are advantageous.[38] Access via API might be a consideration.
  - **multilingual-e5-large-instruct:** A strong, widely-used open-source multilingual embedding model known for good performance across many languages and tasks.[51] It is readily available on Hugging Face.
  - **Models from the "Indonesian Sentence Embeddings" project:** If models from this project [53] demonstrate competitive performance on relevant Indonesian retrieval benchmarks (e.g., MIRACL Indonesian subset, TyDiQA Indonesian subset), they would be strong candidates due to their specific focus on Indonesian.
- **Justification:** The final selection will be based on a comparative assessment of these candidates, prioritizing models with proven strong performance in Indonesian or similar low-resource language contexts for retrieval tasks, and considering practical aspects like accessibility and integration effort.

3.4.3. Vector Database

A vector database is required to store and efficiently query the embeddings of the document chunks from the consular knowledge base.

- **Selection Criteria:**

    - Open-source license, suitable for academic research.
    - Ease of setup, configuration, and use within a Master's project timeframe.
    - Good Python compatibility and integration with RAG frameworks.
    - Sufficient performance and scalability for the anticipated dataset size (potentially tens of thousands to hundreds of thousands of vectors, depending on chunking).
    - Support for metadata filtering alongside vector similarity search (for hybrid search capabilities).
    - References:.[54]
- **Candidate Systems:**
    - **FAISS (Facebook AI Similarity Search):** A highly efficient library for similarity search, often used in research. It requires more manual setup for a full database solution but offers excellent performance.[42]
    - **Qdrant:** An open-source vector database known for its performance, filtering capabilities, and ease of use. It offers a good balance of features for this project.
    - **Milvus:** A popular open-source, scalable vector database designed for AI applications.[49]
- **Justification:** The choice will lean towards a system that balances powerful features with ease of deployment and management for a single researcher within the thesis timeline. Qdrant or Milvus might offer a more complete out-of-the-box solution compared to building a system around raw FAISS, unless specific performance needs dictate otherwise.

3.4.4. RAG Pipeline Implementation

The RAG pipeline will be implemented in two main stages: indexing and retrieval/generation.

- **A. Indexing Pipeline (Offline Process):**
    1. **Load Documents:** Ingest the curated Indonesian consular documents (from Section 3.2.1).
    2. **Preprocess & Chunk:** Apply text cleaning and document chunking strategies (as defined in Section 3.3.2).
    3. **Generate Embeddings:** Use the selected embedding model (Section 3.4.2) to create a vector embedding for each document chunk.
    4. **Store in Vector DB:** Store these embeddings, along with the corresponding chunk text and extracted metadata (Section 3.3.3), in the chosen vector database (Section 3.4.3).
- **B. Retrieval and Generation Pipeline (Online Process):**
    1. **User Query Input:** The system receives a user's query in Bahasa Indonesia.
    2. **Query Encoding:** The user query is encoded into a vector embedding using

the same embedding model employed for document chunk encoding.

3. **Semantic Search (Retrieval):** A similarity search (e.g., using cosine similarity) is performed against the vectors in the database to retrieve the top-K most relevant document chunks. The value of K (e.g., 3 to 5 chunks) will be an empirically tuned hyperparameter.
4. **Re-ranking (Advanced RAG Component):** To enhance the quality of retrieved context, a re-ranking step will be implemented.
    - **Method:** A cross-encoder model (e.g., a multilingual model from the Sentence Transformers library, or one fine-tuned on a relevant task like Semantic Textual Similarity or Natural Language Inference) will be used to re-score the initial top-K [query, chunk] pairs. The chunks will then be re-ordered based on these more nuanced relevance scores. Simpler methods like BM25 re-ranking could serve as a fallback if cross-encoder integration proves too complex within the timeframe.
    - **Rationale:** This step aims to address limitations where pure vector similarity might not perfectly capture true relevance for complex queries, thereby providing higher-quality context to the generator LLM.[30]
5. **Contextual Prompt Construction (Augmentation):** The original user query is combined with the content of the (re-ranked) retrieved document chunks. This forms a comprehensive prompt for the SahabatAI LLM. Effective prompt engineering strategies will be explored to clearly instruct the LLM on how to use the provided context (e.g., "Based only on the following documents, answer the question...").
6. **Answer Generation:** The augmented prompt is fed to the (potentially fine-tuned, see Section 3.4.5) SahabatAI model to generate a natural language answer in Bahasa Indonesia.
7. **Source Attribution (Optional but Desirable):** Where feasible, the system will attempt to link parts of the generated answer back to the specific retrieved source documents or chunks, enhancing transparency and allowing users to verify information.

- **Frameworks/Libraries:** To streamline the development and integration of these components, popular Python-based RAG frameworks such as **LangChain** or **LlamaIndex** will be considered.[26] These frameworks provide abstractions and tools for building complex LLM applications, including RAG pipelines.

3.4.5. SahabatAI Fine-tuning (PEFT)

To adapt SahabatAI more closely to the specific language, style, and knowledge patterns of Indonesian consular services, PEFT will be employed.

- **Objective:** Enhance SahabatAI's ability to follow consular-specific instructions,

accurately synthesize information from retrieved consular documents, and generate responses that are stylistically appropriate for the domain.

- **Technique: QLoRA (Quantized Low-Rank Adaptation)**.[33] This technique is selected due to its proven effectiveness and efficiency with Gemma2-based models like SahabatAI, significantly reducing memory and computational demands compared to full fine-tuning, making it viable for this project.
- **Training Data:** The training split of the custom-curated Indonesian consular QA dataset (from Section 3.2.2) will be used. The data will be formatted in an instruction-following style, for example: {"instruction": "Jawab pertanyaan berikut berdasarkan konteks konsuler yang diberikan.", "input": "Pertanyaan: [Pertanyaan Pengguna]\nKonteks:", "output": "[Jawaban Ideal]"}
- **Procedure:**
  1. Load the pre-trained gemma2-9b-cpt-sahabatai-v1-instruct model.
  2. Configure QLoRA parameters (e.g., LoRA rank r, lora_alpha, lora_dropout, target modules within the Gemma2 architecture where LoRA matrices will be applied) based on established best practices for Gemma models and available computational resources.[33]
  3. Utilize the Hugging Face transformers library, in conjunction with the peft (Parameter-Efficient Fine-Tuning) library and trl (Transformer Reinforcement Learning) library, specifically the SFTTrainer (Supervised Fine-tuning Trainer), for the fine-tuning process.[33]
  4. Train the model for a predetermined number of epochs, carefully monitoring the loss on the validation set to prevent overfitting and to select the best performing checkpoint. Hyperparameters such as learning rate, batch size, and weight decay will be tuned based on initial experiments and literature recommendations.[35]
- **Expected Outcome:** The fine-tuning process will yield a set of LoRA adapter weights. When loaded with the base SahabatAI model, these adapters will specialize the model for Indonesian consular question answering within the RAG framework.

3.4.6. User Interface (for demonstration and evaluation)

A simple, functional web-based chatbot interface will be developed to allow for interactive querying of the system.

- **Technology:** Python libraries such as **Streamlit** or **Gradio** will be used due to their ease of use for rapidly creating interactive AI/ML web applications.
- **Purpose:** This interface will serve multiple purposes:
  - Demonstrating the capabilities of the developed chatbot.
  - Facilitating qualitative testing and debugging during development.

    - Collecting data for human evaluation of the system's responses.

3.4.7. Safety and Ethical Guardrails

Given that SahabatAI is not pre-aligned for safety 17, and the sensitive nature of consular information, implementing safety and ethical guardrails is essential.

- **Input Sanitization:** Basic mechanisms to filter user input for obviously malicious patterns or common prompt injection attempts will be explored.[87]
- **Output Moderation:**
    - **Content Filtering:** Keyword-based filters will be implemented to detect and block clearly inappropriate, offensive, or harmful language if inadvertently generated by the LLM.
    - **Canned Responses for Sensitive/Out-of-Scope Queries:** The system will be designed to identify queries that are potentially harmful, unethical, request illegal actions, or fall far outside the defined consular domain. For such queries, instead of attempting to generate an answer (which could be speculative or unsafe), the chatbot will default to a polite refusal (e.g., "Maaf, saya tidak dapat menjawab pertanyaan tersebut.") or a generic statement indicating its limitations.[89]
    - **Prompt Engineering for Safety:** The system prompts provided to SahabatAI will include explicit instructions to:
        - Base answers strictly on the provided retrieved context.
        - Maintain a factual and neutral tone.
        - Avoid speculation or generating information not present in the source documents.
        - Refrain from generating harmful, biased, or discriminatory content.
- **Limiting Hallucinations:** The RAG architecture itself is a primary mechanism to reduce hallucinations by grounding responses in retrieved documents. This will be reinforced by strong prompting.
- **Bias Mitigation:** The primary source of information (MoFA documents) is assumed to be official and relatively unbiased. However, the QA dataset creation process (especially if involving synthetic generation) will be carefully reviewed to minimize the introduction of unintended biases. The performance of the system will be qualitatively assessed for any signs of biased responses.
- **Transparency:** The user interface will clearly state that the user is interacting with an AI assistant and that its knowledge is based on official MoFA information up to a specific point in time. This manages user expectations and promotes responsible use.
- **References for Guardrail Techniques:**.[59]

The development of the RAG pipeline is an iterative process. The performance of each

component—embedding model selection, chunking strategy, retriever configuration, re-ranker effectiveness, and the LLM's generation quality—interacts with and influences the others. The methodology must therefore allow for some degree of iterative testing and refinement. For example, if initial retrieval results are poor, adjustments to the chunking strategy or the embedding model might be necessary. Similarly, the fine-tuning of SahabatAI might alter how it best consumes and synthesizes retrieved context, potentially necessitating adjustments to the prompting strategies. Ablation studies will be important to understand the specific contributions of different components (e.g., the impact of fine-tuning, the effect of the re-ranker).

3.5. Data Analysis (Evaluation Strategy)

A comprehensive evaluation strategy will be employed to assess the performance of the proposed fine-tuned SahabatAI-RAG system for Indonesian consular question answering. This will involve comparisons against several baseline models and the use of a diverse set of metrics targeting different aspects of the system's performance.

3.5.1. Baseline Models for Comparison

To contextualize the performance of the proposed system, the following baselines will be established:

- **B1: Zero-Shot SahabatAI:** The pre-trained gemma2-9b-cpt-sahabatai-v1-instruct model will be prompted to answer consular questions directly, without access to the RAG knowledge base. This baseline will measure the LLM's inherent knowledge and capabilities regarding Indonesian consular topics based solely on its pre-training and instruction tuning.
- **B2: Naive RAG + SahabatAI (No Fine-Tuning):** A basic RAG pipeline will be implemented using the pre-trained (non-fine-tuned) SahabatAI as the generator. This will involve a standard retriever (e.g., dense vector search without advanced re-ranking) and the curated consular knowledge base. This baseline helps isolate the impact of the RAG architecture itself, before fine-tuning the generator.
- **B3: (Potentially) Existing SARI Chatbot (Qualitative Comparison):** Direct API access or the ability to run the SARI chatbot on the custom test set is unlikely to be available for a Master's thesis. However, a qualitative comparison can be made. Based on SARI's publicly described capabilities (e.g., focus on female migrant workers, empathetic responses, Javanese language support [1]) and the performance of the proposed system on similar types of queries (if applicable from the test set), a discussion of their differing strengths and target areas can be included.
- **B4: State-of-the-Art RAG Baselines (from Literature, for Context):** While direct replication might be out of scope, the performance of the proposed system

can be contextualized by citing reported results from open-source RAG systems or research papers evaluating RAG on other domain-specific QA tasks (e.g., general performance levels reported in studies like [31]). This helps position the achieved results within the broader RAG research landscape.

3.5.2. Evaluation Metrics

The evaluation will utilize the test split of the curated Indonesian consular QA dataset and will cover retrieval quality, generation quality, and end-to-end QA performance.

- **A. Retrieval Quality Evaluation:** The effectiveness of the retrieval component (including the embedding model and any re-ranking mechanism) will be assessed independently of the LLM generator. The ground truth for this evaluation will be the document passages manually linked to each question in the test set.
    - Metrics: **Precision@K**, **Recall@K**, **Mean Reciprocal Rank (MRR@K)**, and **Normalized Discounted Cumulative Gain (NDCG@K)**.[68] K will be varied (e.g., K=1, 3, 5) to understand performance at different retrieval depths.
- **B. Generation Quality Evaluation:** This assesses the quality of the answers generated by SahabatAI (both baseline and fine-tuned versions) when provided with the query and retrieved context.
    - **Automated Metrics (Reference-based):** These will be used if the ground truth answers in the test set are suitable for direct n-gram comparison.
        - **ROUGE (ROUGE-1, ROUGE-2, ROUGE-L):** To measure n-gram overlap with reference answers, focusing on recall.[68]
        - **BLEU:** To measure n-gram precision against reference answers.[68]
        - **METEOR:** For a more nuanced lexical similarity assessment, considering synonyms and stemming.[68]
    - **LLM-as-a-Judge (Reference-free or Reference-assisted):** A powerful external LLM (e.g., GPT-4 API if university resources permit, or a strong open-source model like Prometheus 2 7B [34]) will be prompted to evaluate the generated answers based on specific criteria. This approach is valuable for assessing qualitative aspects that are hard for n-gram metrics to capture.[34]
        - **Faithfulness/Groundedness:** Does the answer accurately and exclusively reflect information from the provided retrieved context? (This is paramount for RAG).
        - **Answer Relevance:** Does the answer directly and comprehensively address the user's question?
        - **Coherence/Fluency:** Is the answer well-written, grammatically correct, and easy to understand?
        - **No Hallucination:** Does the answer avoid inventing facts or information not present in the provided context?.[68]

- **C. End-to-End QA Performance (Human Evaluation):** A subset of the test set (e.g., 50-100 diverse questions, selected to represent various query types and consular topics) will be used for human evaluation. At least two human evaluators (ideally, the researcher and another individual trained on the criteria) will rate the answers generated by the proposed system and key baselines.
    - Criteria for Human Evaluation:
        - **Factual Accuracy:** Is the information provided in the answer factually correct according to official Indonesian consular guidelines and the curated knowledge base?
        - **Completeness:** Does the answer provide all the necessary and relevant information required to address the user's query adequately?
        - **Helpfulness:** Overall, how helpful is the answer in resolving the user's query or guiding them towards a solution?
        - **Clarity and Understandability:** Is the answer presented clearly, concisely, and in a manner that is easy for a typical user to understand?
    - A Likert scale (e.g., 1 to 5) will be used for rating each criterion. Inter-rater reliability (e.g., using Cohen's Kappa or Krippendorff's Alpha) will be calculated if multiple independent evaluators are used.
- **D. User Satisfaction (Optional, if a small user study is feasible):** If time permits, a small user study involving a few representative users interacting with the chatbot prototype could provide qualitative feedback.
    - Metrics could include:
        - **Task Completion Rate:** Can users successfully find the answer to predefined mock consular queries using the chatbot?
        - **CSAT (Customer Satisfaction Score):** A simple post-interaction survey asking users to rate their satisfaction with the chatbot's performance.[68]

3.5.3. Evaluation Protocol

- The curated QA test set will be the primary instrument for all quantitative evaluations.
- **Ablation Studies:** To understand the specific contributions of different components of the proposed system, ablation studies will be conducted:
    - Compare the full proposed system (Fine-tuned SahabatAI + Advanced RAG) against B1 (Zero-Shot SahabatAI without RAG) to show the overall impact of RAG and fine-tuning.
    - Compare the full proposed system against B2 (Naive RAG + non-fine-tuned SahabatAI) to show the impact of advanced RAG components and fine-tuning.
    - If a re-ranking component is implemented, compare the performance of the

RAG system with and without the re-ranker.
    - Compare the RAG system using the fine-tuned SahabatAI generator against the RAG system using the non-fine-tuned SahabatAI generator (with the same retrieved context) to specifically isolate the impact of fine-tuning the generator.
- **Statistical Significance:** Where appropriate for the sample sizes and data distributions, statistical significance tests (e.g., paired t-tests, Wilcoxon signed-rank test) will be used to compare the metric scores between different system configurations.
- **Qualitative Analysis:** In addition to quantitative metrics, a qualitative analysis of example good and bad responses generated by the system will be performed. This will help identify common error patterns, understand the system's strengths and weaknesses in handling different types of queries, and provide insights for future improvements.

Evaluating the Fine-Tuned Generator within RAG: The primary goal here is to determine if fine-tuning SahabatAI specifically improves its ability to leverage the retrieved consular context effectively.[49] This will be measured by comparing metrics like Faithfulness, Answer Relevance, and human-evaluated Accuracy when the fine-tuned generator is used within the RAG pipeline versus when the non-fine-tuned generator is used with the *exact same retrieved context*. The evaluation dataset should include questions that specifically test the model's capacity to synthesize information from multiple retrieved documents or to follow complex instructions based on the provided context.

3.6. Data Visualisation (Planned for final thesis report)

To effectively communicate the findings of the research, the final thesis report will include various data visualizations:

- **Bar charts:** These will be used to compare key performance metrics (e.g., ROUGE scores, Faithfulness scores from LLM-as-a-Judge, human-evaluated Accuracy percentages) across the different system configurations (Baselines vs. Proposed System, and ablation study variants).
- **Tables:** Comprehensive tables will summarize all quantitative results, including mean scores, standard deviations (if applicable from multiple runs or evaluators), and potentially confidence intervals.
- **Error Analysis Examples:** Illustrative examples of good and bad responses from different models will be presented to support qualitative analysis and highlight specific system behaviors.

- **Learning Curves (if applicable):** If fine-tuning involves multiple epochs, learning curves showing training and validation loss/metrics over epochs might be included to demonstrate the training process.
- **Conceptual Diagrams:** Diagrams illustrating the proposed RAG architecture and the data flow will be used to enhance clarity.

3.7. Data Validity and Reliability

Ensuring the validity and reliability of the data and evaluation processes is crucial for the credibility of the research findings:

- **Knowledge Base Accuracy:** Information for the KB will be sourced exclusively from official MoFA channels. Where possible, information on key topics will be cross-referenced from multiple official sources to ensure accuracy and consistency.
- **QA Dataset Quality:**
  - **Human Validation:** All manually created QA pairs and a significant sample (or all, if feasible) of LLM-assisted synthetically generated QA pairs will be manually validated by the researcher. This validation will check for question clarity, answer correctness according to the KB, relevance of the question to the consular domain, and absence of ambiguity.
  - **Inter-Rater Reliability:** If multiple human annotators are involved in the creation of the QA dataset or in the human evaluation of system responses, inter-rater reliability measures (e.g., Cohen's Kappa for two raters, Fleiss' Kappa for more than two) will be calculated on a subset of the data to ensure consistency in judgments.
- **Evaluation Consistency:**
  - Standardized prompts and detailed scoring rubrics will be developed and used for both LLM-as-a-Judge evaluations and human evaluations to ensure consistency in how responses are assessed.
  - If human evaluators are involved (beyond the researcher), they will be trained on the evaluation criteria and rubrics.
- **Reproducibility:** All steps of the research methodology, including data collection parameters, preprocessing scripts, model versions used, specific hyperparameter settings for fine-tuning and RAG components, and evaluation scripts, will be meticulously documented. Random seeds will be fixed for all experiments involving stochastic processes (e.g., model training, data shuffling) to enhance the reproducibility of the results. Code developed for the project will be managed using version control (e.g., Git) and will be made available (e.g., via GitHub) as per university policy, to the extent possible.

3.8. Tentative Timeline (Gantt Chart for 6 months)

The following Gantt chart outlines the planned activities and their durations over a six-month period:

Task ID	Task Name	Duration (Weeks)	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
Phase 1: Foundation & Planning		**4**	WW WW					
1.1	In-depth Literature Review Finalization	2	WW					
1.2	Refine RQs, ROs; Ethics Prep	2	WW					
1.3	MoFA Data Source Mapping; Dev Env Setup	2	WW					
Phase 2: Data Collection & Preparation		**6**		WW WW WW				
2.1	Intensive Data Collection (KB Content)	3		WW W				
2.2	Knowledge Base V1 Curation & Cleaning	2		WW				

2.3	Design QA Dataset Strategy; Start Generation	3		WW W				
Phase 3: Baseline System & Dataset Refinement		**4**			WW WW			
3.1	Data Preprocessing (Chunking); Index KB	2			WW			
3.2	Implement Naive RAG (Baseline B2)	2			WW			
3.3	Complete & Validate QA Dataset (Train/Val/Test)	2			WW			
Phase 4: Fine-tuning & Advanced RAG		**4**				WW WW		
4.1	SahabatAI PEFT (QLoRA) Fine-tuning	3				WW W		
4.2	Implement Adv. RAG components;	2				WW		

	Integrate FT LLM							
Phase 5: Evaluation & Analysis		**4**					WW WW	
5.1	Automated Metrics & LLM-as-a-Judge Eval	2					WW	
5.2	Human Evaluation; Analyze Results; Ablations	3					WW W	
Phase 6: Thesis Writing & Submission		**4**						WW WW
6.1	Write Chapters 1-3 (Intro, LitRev, Method)	2						WW
6.2	Write Chapters 4-5 (Results, Discussion, Conc)	2						WW
6.3	Final Revisions & Submission	1						W

This timeline is ambitious but feasible for a Master's thesis, assuming dedicated effort and access to necessary resources (e.g., computational power for fine-tuning, APIs if needed). Regular meetings with the thesis supervisor will be crucial for monitoring progress and addressing any challenges that arise.

Table 3.1: Proposed RAG System Architecture Components

Component	Selected Technology/Model	Justification for Choice
Core LLM	SahabatAI (Gemma2 9B CPT SahabatAI v1 Instruct) [17]	Strong Bahasa Indonesia & local dialect capabilities; Gemma2 architecture suitable for PEFT (QLoRA); Open-source; Instruction-tuned; Manageable size for academic fine-tuning.
Embedding Model	Candidate: multilingual-e5-large-instruct [51] or Google Gemini Embedding [38] (pending accessibility)	High performance on multilingual/Indonesian retrieval benchmarks; Good balance of dimensionality, context length, and efficiency; Strong support for Bahasa Indonesia.
Vector Database	Candidate: Qdrant or Milvus [49]	Open-source; Good Python compatibility; Sufficient performance for expected dataset size; Supports metadata filtering; Balances features with ease of deployment for thesis timeframe.
Retriever	Semantic similarity search (e.g., cosine) in Vector DB	Standard and effective method for initial document retrieval based on query-chunk semantic closeness.
Re-ranker	Cross-encoder model (e.g., from Sentence Transformers library)	Improves relevance of top-K retrieved chunks beyond pure vector similarity, providing higher quality context to the generator LLM.
Generator	Fine-tuned SahabatAI (using QLoRA)	Adapts the base SahabatAI model to the specific nuances, terminology, and

		information synthesis requirements of the Indonesian consular domain.
Orchestration	LangChain or LlamaIndex [26]	Simplifies integration of RAG components, streamlining development of the pipeline.

This table provides a clear, consolidated view of the proposed technical stack, justifying each component choice based on the research requirements and available technologies.

Table 3.2: Evaluation Metrics for Consular QA RAG System

Evaluation Aspect	**Specific Metric**	**Description/Formula**	**Relevance to Consular QA**
Retrieval Quality	Precision@K, Recall@K	Proportion of relevant documents among top-K retrieved; Proportion of all relevant documents retrieved in top-K. [68]	Ensures the system finds the correct consular regulations/information.
	MRR@K, NDCG@K	Rank-aware metrics for retrieval effectiveness. [68]	Assesses how quickly and accurately relevant information is surfaced.
Generation Quality (Automated)	ROUGE (1, 2, L), BLEU, METEOR	N-gram overlap and lexical similarity with reference answers. [68]	Measures fluency and similarity to ideal answers, useful for iterative development.
Generation Quality (LLM-as-Judge)	Faithfulness/Groundedness	Assesses if the answer accurately reflects *only* information from the provided retrieved context. [60]	Critical for trust; ensures answers are based on official MoFA data, not hallucinated.

	Answer Relevance	Assesses if the answer directly and fully addresses the user's question. [60]	Ensures the chatbot provides useful and on-topic responses.
	Coherence/Fluency	Assesses if the answer is well-written, grammatically correct, and easy to understand. [60]	Important for user experience and clear communication of consular information.
	No Hallucination	Assesses if the answer avoids making up facts not present in the context. [68]	Essential for providing reliable and trustworthy consular advice.
End-to-End QA (Human Eval)	Factual Accuracy	Is the information factually correct according to official consular guidelines/KB?	Paramount for providing correct consular information.
	Completeness	Does the answer provide all necessary information for the query?	Ensures users receive comprehensive guidance.
	Helpfulness	Overall, how helpful is the answer in resolving the user's query?	Reflects the practical utility of the chatbot for citizens.
	Clarity/Understandability	Is the answer presented clearly and easy to understand?	Ensures information is accessible to a general audience.

This table clearly outlines the comprehensive evaluation framework, linking specific metrics to the core objectives of building an effective and trustworthy AI-powered consular chatbot.

References

(This section will be populated with full citations for all academic papers and key resources mentioned, following a consistent academic style like ACM or IEEE, as appropriate for AI/NLP research. For brevity in this proposal outline, only key authors/sources and snippet IDs are mentioned in the text. The final thesis will have a complete bibliography.)

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P.,... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly, 28*(1), 75-105.
- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed. draft).
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N.,... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459-9474.
- Meswani, M., et al. (2024). Gemma: Open Models Based on Gemini Research and Technology. *Google DeepMind*.
- Mostafaei, H., Kordnoori, S., Ostadrahimi, M., & Banihashemi, S. S. A. (2025). Applications of artificial intelligence in global diplomacy: A review of research and practical models. *Sustainable Futures, 9*, 100486. [9]
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.
- Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach* (3rd ed.). Pearson Education Limited. [29]
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y.,... & Lample, G. (2023a). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T.,... & Lample, G. (2023b). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y.,... & Liu, P. (2024). LIMA: Less Is More for Alignment. *arXiv preprint arXiv:2305.11206*. (Cited in [34])

Works cited

1. Foreign Ministry to Utilize AI-Based Services for Indonesian Citizens Abroad, accessed May 15, 2025, https://en.tempo.co/read/1975389/foreign-ministry-to-utilize-ai-based-services-for-indonesian-citizens-abroad
2. Portal Peduli WNI Mobile – Apps on Google Play, accessed May 15, 2025, https://play.google.com/store/apps/details?id=id.go.kemlu.peduliwni.mobile&hl=en_GB
3. Safe Travel - Apps on Google Play, accessed May 15, 2025, https://play.google.com/store/apps/details?id=id.go.kemlu.safetravel
4. Safe Travel – Apps on Google Play, accessed May 15, 2025, https://play.google.com/store/apps/details?id=id.go.kemlu.safetravel&hl=en_GB
5. SARI App: Indonesia and UN Women's Effort to Protect ... - RRI.co.id, accessed May 15, 2025, https://www.rri.co.id/internasional/1465199/sari-app-indonesia-and-un-women-s-effort-to-protect-female-migrant-workers
6. Indonesia, UN Women Launch AI Chatbot SARI to Protect Female Migrant Workers - Jakarta Daily, accessed May 15, 2025, https://www.jakartadaily.id/international/16214999370/indonesia-un-women-launch-ai-chatbot-sari-to-protect-female-migrant-workers
7. After Being Built, AI Sarl Chatbot Made By The Indonesian Ministry Of Foreign Affairs Is Undergoing Testing - VOI, accessed May 15, 2025, https://voi.id/en/news/460102
8. Migration MPTF Final Report, accessed May 15, 2025, https://mptf.undp.org/sites/default/files/documents/2025-02/final_report_2024_migration_mptf_indonesia.pdf
9. (PDF) Applications of Artificial Intelligence in Global Diplomacy: A ..., accessed May 15, 2025, https://www.researchgate.net/publication/390329776_Applications_of_artificial_intelligence_in_global_diplomacy_A_review_of_research_and_practical_models
10. Using Large Language Models responsibly in the civil service: a guide to implementation - Bennett Institute for Public Policy, accessed May 15, 2025, https://www.bennettinstitute.cam.ac.uk/publications/using-llms-responsibly-in-the-civil-service/
11. Selecting the Right LLM for eGov ExplanationsThis project has received funding from the European Union's Horizon research and innovation programme under grant agreements no 101094905 (AI4GOV) and 101092639 (FAME). - arXiv,

accessed May 15, 2025, https://arxiv.org/html/2504.21032v1
12. How Governments are Using AI: 8 Real-World Case Studies, accessed May 15, 2025, https://blog.govnet.co.uk/technology/ai-in-government-case-studies
13. AI Use Cases in the U.S. Government, accessed May 15, 2025, https://www.snowflake.com/en/fundamentals/ai-us-government/
14. Sahabat-AI, accessed May 15, 2025, https://sahabat-ai.com/
15. Sahabat AI: The Friend Indonesia Needs for a Digital Future - Twimbit, accessed May 15, 2025, https://twimbit.com/about/blogs/sahabat-ai-the-friend-indonesia-needs-for-a-digital-future
16. Gemma2 9b Cpt Sahabatai V1 Instruct GGUF · Models - Dataloop AI, accessed May 15, 2025, https://dataloop.ai/library/model/gmonsoon_gemma2-9b-cpt-sahabatai-v1-instruct-gguf/
17. GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct · Hugging Face, accessed May 15, 2025, https://huggingface.co/GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct
18. Indonesia Launches Sahabat-AI: A Groundbreaking Local Language Model Initiative, accessed May 15, 2025, https://theoutpost.ai/news-story/indonesia-launches-sahabat-ai-a-groundbreaking-local-language-model-initiative-8298/
19. GoTo Launches Sahabat-AI, Enhancing its Leadership in Indonesia's Technology Sector, accessed May 15, 2025, https://www.gotocompany.com/en/news/press/goto-launches-sahabat-ai-enhancing-its-leadership-in-indonesias-technology-sector
20. arxiv.org, accessed May 15, 2025, http://arxiv.org/pdf/2503.09516
21. arXiv:2410.17952v2 [cs.CL] 24 Jan 2025, accessed May 15, 2025, https://arxiv.org/pdf/2410.17952?
22. \ours: Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains - arXiv, accessed May 15, 2025, https://arxiv.org/html/2410.17952v1
23. Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization - arXiv, accessed May 15, 2025, https://arxiv.org/html/2502.20364
24. What Is Retrieval Augmented Generation, and How Are State and ..., accessed May 15, 2025, https://statetechmagazine.com/article/2025/02/what-is-rag-perfcon
25. How RAG is Transforming Federal Website Search - RIVA Solutions, accessed May 15, 2025, https://rivasolutionsinc.com/insights/how-retrieval-augmented-generation-rag-is-transforming-federal-website-search/
26. RAG Tutorial: A Beginner's Guide to Retrieval Augmented Generation - SingleStore, accessed May 15, 2025, https://www.singlestore.com/blog/a-guide-to-retrieval-augmented-generation-rag/

27. Leveraging long context in retrieval augmented language models for medical question answering - PMC, accessed May 15, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12048518/
28. Department of State AI Inventory 2024, accessed May 15, 2025, https://2021-2025.state.gov/department-of-state-ai-inventory-2024/
29. AI Embassies: A New Frontier in Cyber Domain - Journal of Cyberspace Studies, accessed May 15, 2025, https://jcss.ut.ac.ir/article_100581_a63c21762f4b96d4392cdac251b8c001.pdf
30. Advanced RAG Techniques - DataCamp, accessed May 15, 2025, https://www.datacamp.com/blog/rag-advanced
31. Mean Answer correctness scores of RAG and baseline models. Evaluated on... | Download Scientific Diagram - ResearchGate, accessed May 15, 2025, https://www.researchgate.net/figure/Mean-Answer-correctness-scores-of-RAG-and-baseline-models-Evaluated-on-a-test-dataset_fig4_390021100
32. Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains - GitHub, accessed May 15, 2025, https://github.com/cognitivetech/llm-research-summaries/blob/main/retrieval-augmented-rag/Self-Improving-Retrieval-Augmented-Generation-for-Adapting-Large-Language-Models-to-Specialized-Domains_2410.17952v1.md
33. Fine-Tune Gemma using Hugging Face Transformers and QloRA | Google AI for Developers, accessed May 15, 2025, https://ai.google.dev/gemma/docs/core/huggingface_text_finetune_qlora
34. Beyond QA Pairs: Assessing Parameter-Efficient Fine-Tuning for Fact Embedding in LLMs, accessed May 15, 2025, https://arxiv.org/html/2503.01131v1
35. Fine-Tuning LLMs with Small Data: Guide - Dialzara, accessed May 15, 2025, https://dialzara.com/blog/fine-tuning-llms-with-small-data-guide/
36. Fine-tuning large language models (LLMs) in 2025 - SuperAnnotate, accessed May 15, 2025, https://www.superannotate.com/blog/llm-fine-tuning
37. Top 9 Large Language Models as of May 2025 | Shakudo, accessed May 15, 2025, https://www.shakudo.io/blog/top-9-large-language-models
38. State-of-the-art text embedding via the Gemini API - Google ..., accessed May 15, 2025, https://developers.googleblog.com/en/gemini-embedding-text-model-now-available-gemini-api/
39. Multilingual/Bilingual Large Language Models (LLMs): Tailoring AI Applications for Southeast Asia - ABI Research, accessed May 15, 2025, https://www.abiresearch.com/blog/multilingual-bilingual-large-language-models-llms
40. Speaking in Code: Contextualizing Large Language Models in ..., accessed May 15, 2025, https://carnegieendowment.org/research/2025/01/speaking-in-code-contextualizing-large-language-models-in-southeast-asia?lang=en
41. Mind the (Language) Gap: Mapping the Challenges ... - Stanford HAI, accessed May 15, 2025, https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-

of-llm-development-in-low-resource-language-contexts
42. Retrieval Augmented Generation: Everything You Need to Know About RAG in AI - WEKA, accessed May 15, 2025, https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/
43. RAG vs Traditional QA - GeeksforGeeks, accessed May 15, 2025, https://www.geeksforgeeks.org/rag-vs-traditional-qa/
44. RAG vs. Traditional Search: A Comparative Analysis - Signity Software Solutions, accessed May 15, 2025, https://www.signitysolutions.com/blog/rag-vs-traditional-search-engines
45. RAG techniques - IBM, accessed May 15, 2025, https://www.ibm.com/think/topics/rag-techniques
46. LLM-KG4QA: Large Language Models and Knowledge Graphs for Question Answering - GitHub, accessed May 15, 2025, https://github.com/machuangtao/LLM-KG4QA
47. arxiv.org, accessed May 15, 2025, https://arxiv.org/pdf/2504.04915
48. RuleRAG: Rule-Guided Retrieval-Augmented Generation with Language Models for Question Answering - ResearchGate, accessed May 15, 2025, https://www.researchgate.net/publication/384936998_RuleRAG_Rule-Guided_Retrieval-Augmented_Generation_with_Language_Models_for_Question_Answering
49. Knowledge Injection in LLMs: Fine-Tuning vs. RAG - Zilliz blog, accessed May 15, 2025, https://zilliz.com/blog/knowledge-injection-in-llms-fine-tuning-and-rag
50. Can LLMs Be Trusted for Evaluating RAG Systems? A Survey of Methods and Datasets, accessed May 15, 2025, https://arxiv.org/html/2504.20119v2
51. MMTEB: Massive Multilingual Text Embedding Benchmark - arXiv, accessed May 15, 2025, https://arxiv.org/html/2502.13595v1
52. LUSIFER: Language Universal Space Integration for Enhanced Multilingual Embeddings with Large Language Models - arXiv, accessed May 15, 2025, https://arxiv.org/html/2501.00874v3
53. LazarusNLP/indonesian-sentence-embeddings - GitHub, accessed May 15, 2025, https://github.com/LazarusNLP/indonesian-sentence-embeddings
54. How to Choose the Right Vector Database for Your RAG Architecture | DigitalOcean, accessed May 15, 2025, https://www.digitalocean.com/community/conceptual-articles/how-to-choose-the-right-vector-database
55. Vector databases explained: Use cases, algorithms and key features - Instaclustr, accessed May 15, 2025, https://www.instaclustr.com/education/vector-database/vector-databases-explained-use-cases-algorithms-and-key-features/
56. How do you handle document preprocessing for multimodal RAG? - Milvus, accessed May 15, 2025, https://milvus.io/ai-quick-reference/how-do-you-handle-document-preprocessing-for-multimodal-rag
57. RAG, AI Agents, and Agentic RAG: An In-Depth Review and Comparative Analysis, accessed May 15, 2025, https://www.digitalocean.com/community/conceptual-articles/rag-ai-agents-age

ntic-rag-comparative-analysis
58. Best Practices for Building a QA Dataset to Evaluate RAG Quality? : r/LangChain - Reddit, accessed May 15, 2025, https://www.reddit.com/r/LangChain/comments/1iq8vtb/best_practices_for_building_a_qa_dataset_to/
59. RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models - arXiv, accessed May 15, 2025, https://arxiv.org/html/2504.18041v1
60. How we are doing RAG AI evaluation in Atlas - ClearPeople, accessed May 15, 2025, https://www.clearpeople.com/blog/how-we-are-doing-rag-ai-evaluation-in-atlas
61. Fine-tune Gemma in Keras using LoRA | Google AI for Developers - Gemini API, accessed May 15, 2025, https://ai.google.dev/gemma/docs/core/lora_tuning
62. RAG vs. Fine-Tuning: How to Choose - Oracle, accessed May 15, 2025, https://www.oracle.com/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/rag-fine-tuning/
63. RAG vs Fine Tuning LLMs: The Right Approach for Generative AI - Aisera, accessed May 15, 2025, https://aisera.com/blog/llm-fine-tuning-vs-rag/
64. Public sector data stewardship for the AI era | Elastic Blog, accessed May 15, 2025, https://www.elastic.co/blog/public-sector-data-stewardship-for-the-ai-era
65. How Supervised Learning is Transforming Technology in Indonesia - BytePlus, accessed May 15, 2025, https://www.byteplus.com/en/topic/424585
66. 7 actions that enforce responsible AI practices - Huron Consulting, accessed May 15, 2025, https://www.huronconsultinggroup.com/insights/seven-actions-enforce-ai-practices
67. Cracking the Code: Multi-domain LLM Evaluation on Real-World Professional Exams in Indonesia - arXiv, accessed May 15, 2025, https://arxiv.org/html/2409.08564v2
68. RAG Optimization: Metrics, & Tools for Enhanced LLMs Performance - SearchUnify, accessed May 15, 2025, https://www.searchunify.com/blog/rag-optimization-metrics-tools-for-enhanced-llms-performance/
69. Top Metrics to Monitor and Improve RAG Performance - Galileo AI, accessed May 15, 2025, https://www.galileo.ai/blog/top-metrics-to-monitor-and-improve-rag-performance
70. Best Practices in RAG Evaluation: A Comprehensive Guide - Qdrant, accessed May 15, 2025, https://qdrant.tech/blog/rag-evaluation-guide/
71. LLM Evaluation: Key Metrics, Best Practices and Frameworks - Aisera, accessed May 15, 2025, https://aisera.com/blog/llm-evaluation/
72. Chatbot Analytics: 15 Core Metrics to Track - Sprinklr, accessed May 15, 2025, https://www.sprinklr.com/blog/chatbot-analytics/
73. The Embassy of the Republic of Indonesia - Washington, accessed May 15, 2025, https://kemlu.go.id/washington/tentang-perwakilan/kontak-kami

74. General Information - The Official eVisa website for Indonesia - Imigrasi, accessed May 15, 2025, https://evisa.imigrasi.go.id/front/info/evoa
75. Building a RAG Evaluation Dataset: A Step-By-Step Guide Using Document Sources, accessed May 15, 2025, https://magnimindacademy.com/blog/building-a-rag-evaluation-dataset-a-step-by-step-guide-using-document-sources/
76. Evaluating and Enhancing RAG Pipeline Performance Using Synthetic Data, accessed May 15, 2025, https://developer.nvidia.com/blog/evaluating-and-enhancing-rag-pipeline-performance-using-synthetic-data/
77. LEGAL-UQA: A Low-Resource Urdu-English Dataset for Legal Question Answering - arXiv, accessed May 15, 2025, https://arxiv.org/html/2410.13013v1
78. Chatbots and Ethical Considerations: Navigating Privacy ... - SmythOS, accessed May 15, 2025, https://smythos.com/ai-agents/chatbots/chatbots-and-ethical-considerations/
79. How to Write Thesis with AI Ethically - Yomu AI, accessed May 15, 2025, https://www.yomu.ai/blog/how-to-write-thesis-with-ai-ethically
80. Synthetic data created by ... - Environmental Factor - April 2025, accessed May 15, 2025, https://factor.niehs.nih.gov/2025/4/feature/ai-data-ethics
81. Leveraging LLMs for Synthetic Data Generation - Deepchecks, accessed May 15, 2025, https://www.deepchecks.com/leveraging-llms-synthetic-data-generation/
82. royalsociety.org, accessed May 15, 2025, https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Synthetic_Data_Survey-24.pdf
83. Generating Synthetic Training Data for Supervised De-Identification ..., accessed May 15, 2025, https://www.mdpi.com/1999-5903/13/5/136
84. 7 Chunking Strategies in RAG You Need To Know - F22 Labs, accessed May 15, 2025, https://www.f22labs.com/blogs/7-chunking-strategies-in-rag-you-need-to-know/
85. Effective Chunking Strategies for RAG - Cohere Documentation, accessed May 15, 2025, https://docs.cohere.com/v2/page/chunking-strategies
86. Optimizing RAG systems with fine-tuning techniques | SuperAnnotate, accessed May 15, 2025, https://www.superannotate.com/blog/rag-fine-tuning
87. LLM Security: Top 10 Risks & Best Practices to Mitigate Them - Cohere, accessed May 15, 2025, https://cohere.com/blog/llm-security
88. LLM Guardrails: Secure and Controllable Deployment - Neptune.ai, accessed May 15, 2025, https://neptune.ai/blog/llm-guardrails
89. What are LLM Guardrails? Essential Protection for AI Systems - DigitalOcean, accessed May 15, 2025, https://www.digitalocean.com/resources/articles/what-are-llm-guardrails