Improving Consular Question Answering in Indonesian Embassies Using Retrieval-Augmented Generation and Fine-Tuned SahabatAI

Handling consular service queries effectively is crucial for the Indonesian Ministry of Foreign Affairs (Kemlu) and its embassies, yet it remains challenging due to high query volumes and the need for accurate, context-specific information. This thesis proposes the development of a conversational AI system specifically tailored for consular query handling using Retrieval-Augmented Generation (RAG) combined with a fine-tuned SahabatAI model. The system leverages consular-specific data sources, such as the "Peduli WNI" portal and "Safe Travel" application and Kemlu website, to construct a domain-specific knowledge base that augments the model's generative capabilities with relevant, factual information. The research involves data collection and preparation focused solely on consular services, fine-tuning SahabatAI to handle consular-specific queries, and integrating RAG to enhance the accuracy and contextual relevance of responses. Performance will be evaluated based on factual accuracy, relevance, and conversational quality in Bahasa Indonesia. This study aims to demonstrate the effectiveness of combining RAG and domain-adapted LLMs in optimizing consular query handling for Indonesian citizens abroad, presenting a targeted AI solution within the public sector.

Chapter 1: Introduction

1.1. Background and Motivation

The global landscape of public administration is undergoing a significant transformation, driven by the increasing demand for more efficient, accessible, and citizen-centric government services.1 Consular services, which form a critical interface between a state and its citizens abroad, are particularly ripe for such innovation. The Indonesian Ministry of Foreign Affairs (Kemlu) bears the responsibility of serving a vast diaspora of Indonesian citizens (Warga Negara Indonesia - WNI) across the globe. This mandate involves providing a wide array of services, from passport and visa issuance to legal assistance and emergency support. The effective delivery of these services is crucial for citizen welfare and national representation.

The potential of conversational Artificial Intelligence (AI) to revolutionize public sector service delivery is increasingly recognized.[1] By providing 24/7 support, automating responses to common inquiries, and streamlining administrative tasks, conversational AI can enhance citizen satisfaction and operational efficiency. Recognizing this potential, Kemlu has already embarked on a journey of digital transformation, evidenced by initiatives such as the "Peduli WNI" web portal and the "Safe Travel" mobile application.[3] These platforms represent foundational steps towards leveraging technology for improved consular assistance.

The advent of more sophisticated AI paradigms, notably Generative AI (GenAI) and Large Language Models (LLMs), offers unprecedented opportunities for enhancing

diplomatic and consular functions.[8] These technologies possess advanced capabilities in natural language understanding, content generation, and data processing, making them suitable for tasks such as providing detailed information, assisting in crisis management, and enabling personalized citizen engagement. The global trend towards AI-powered public services, combined with Indonesia's ongoing efforts in digitalizing consular affairs, signals a propitious environment for research into advanced conversational AI solutions. This is further underscored by Kemlu's collaboration with UN Women to launch the SARI (Sahabat Artifisial Migran Indonesia) chatbot, an AI-powered tool aimed at protecting Indonesian female migrant workers by providing information and support.[10] The existence of SARI not only demonstrates Kemlu's receptiveness to AI but also highlights the practical application of such technologies in addressing specific citizen needs. This existing digital infrastructure and strategic direction towards AI adoption create a compelling case for developing a more comprehensive and intelligent query handling system for a broader range of consular services.

1.2. Problem Statement

Despite existing digital platforms, Indonesian citizens often encounter challenges in obtaining timely, accurate, and comprehensive information regarding consular services. Common inquiries related to passport renewals, visa applications, legal aid, and emergency protocols can inundate consular staff, leading to potential delays and inconsistencies in the information provided. The sheer volume and diversity of these queries necessitate a more scalable and efficient solution.

Current systems, if not leveraging advanced AI, may struggle with the complexities of natural language, failing to understand nuanced queries or provide personalized information effectively. Many existing chatbot solutions are often limited to predefined scripts or simple keyword matching, falling short of user expectations for conversational fluency and depth of knowledge. Furthermore, a critical requirement for any consular service tool in Indonesia is the ability to understand and respond proficiently in Bahasa Indonesia, catering to the linguistic needs of all citizens. The "information gap" concerning legal frameworks, employment regulations, immigration procedures, and general advice for working or living abroad remains a significant hurdle for Kemlu in effectively serving its citizens.[12] This research aims to address these gaps by proposing an intelligent conversational AI system capable of handling a wide spectrum of consular queries with high accuracy and relevance.

1.3. Research Questions

This thesis seeks to answer the following key research questions:

- RQ1: How can a conversational AI system, leveraging State-of-the-Art (SOTA) Indonesian Large Language Models and advanced NLP techniques like

Retrieval-Augmented Generation (RAG), be designed and developed to effectively handle diverse consular service queries for the Indonesian Ministry of Foreign Affairs?
- RQ2: Which specific combination of LLM (e.g., SahabatAI, NusaBERT, fine-tuned IndoBERT) and RAG architecture provides optimal performance in terms of accuracy, relevance, and factual grounding for consular information in Bahasa Indonesia?
- RQ3: What are the effective strategies for collecting, preparing, and structuring a domain-specific knowledge base from existing Kemlu resources (e.g., Peduli WNI, Safe Travel, FAQs, official documents) to support the RAG-based conversational AI?
- RQ4: How can the performance of the proposed conversational AI system be rigorously evaluated using appropriate metrics, considering both linguistic quality and the correctness of consular information provided?

1.4. Research Objectives

The primary objectives of this research are:

- RO1: To conduct a comprehensive review of SOTA conversational AI techniques, LLMs for Bahasa Indonesia, and their applications in public and consular services.
- RO2: To design and propose a novel conversational AI architecture tailored for Indonesian consular query handling, detailing the chosen LLM, RAG framework, and integration with knowledge sources.
- RO3: To develop a methodology for collecting, processing, and structuring a high-quality, domain-specific knowledge base for consular services in Bahasa Indonesia.
- RO4: To implement a prototype of the proposed conversational AI system.
- RO5: To define and apply a robust evaluation framework to assess the system's performance, focusing on accuracy, relevance, factual consistency, and user interaction quality.

1.5. Scope and Limitations of the Thesis

The scope of this thesis is defined as follows:

- The primary focus will be on developing a textual conversational AI system designed to handle informational queries related to a predefined subset of Indonesian consular services. Examples include procedures for passport and visa applications, guidelines for emergency assistance, and information on citizen registration.
- A prototype system will be developed to demonstrate the core functionalities of the proposed architecture.
- Bahasa Indonesia will be the primary language of interaction for the system.

- The creation of the knowledge base and any model fine-tuning will rely on publicly available or ethically sourced data.

The limitations of this thesis include:

- The six-month timeframe necessitates a focused scope, which will constrain the breadth of consular services covered by the prototype and the scale of data collection and model training.
- The system will primarily serve an informational role. Transactional capabilities, such as submitting applications or making payments, are beyond the scope of this prototype.
- Real-time integration with Kemlu's internal backend systems and databases is not feasible within the constraints of this Master's thesis.
- The evaluation of the system will be based on benchmark datasets created for this research and potentially limited user studies, rather than full-scale deployment and live user testing.
- The current proposal does not address voice-based interaction, focusing solely on text-based conversational AI.

1.6. Significance and Contributions

This research is poised to make significant contributions to both the field of AI and the domain of government consular services.

- **To AI in General:**
  - The study will offer valuable insights into the adaptation and application of SOTA conversational AI techniques, particularly RAG and LLMs, to a specific and relatively low-resource language (Bahasa Indonesia) within a critical, real-world domain.
  - It will contribute to a deeper understanding of the challenges and best practices associated with developing domain-specific, factually grounded chatbots, especially concerning knowledge base creation and factual consistency.
  - The research will provide a case study on the fine-tuning and evaluation of Indonesian LLMs for specialized query handling tasks, potentially informing future NLP research for Bahasa Indonesia.
- **To Government/Consular Services:**
  - The proposed system offers a blueprint for a technologically advanced solution to improve the efficiency, accessibility, and accuracy of Indonesian consular services.
  - It has the potential to significantly reduce the workload on human consular staff, enabling them to dedicate more time and resources to complex cases

and direct citizen assistance.
    - By providing 24/7 access to reliable information in Bahasa Indonesia, the system can substantially enhance the experience of Indonesian citizens seeking consular support.
    - This research directly aligns with Kemlu's stated goals of digital transformation and leveraging AI to enhance citizen services and protection.[7]

A successful prototype developed through this research could serve as a compelling model for other Indonesian government agencies. Many public sector entities face similar challenges in information dissemination and citizen service. A practical demonstration of advanced conversational AI within Kemlu, addressing a clear need with a robust technical solution, could lower perceived barriers to adoption and provide a replicable template. This, in turn, could catalyze broader AI implementation across the Indonesian public sector, leading to more efficient and responsive governance.

1.7. Thesis Structure

This thesis proposal is organized into several chapters. Chapter 2 provides a comprehensive literature review and discusses related work in conversational AI, Indonesian NLP, and consular service technologies. Chapter 3 details the proposed methodology, including the system architecture and the rationale for the chosen techniques. Chapter 4 outlines the plan for data collection and preparation, crucial for both training the AI model and building its knowledge base. Chapter 5 describes the strategy for model training, system implementation, and the performance evaluation framework. Finally, Chapter 6 discusses the expected results, presents a detailed project timeline, and analyzes the feasibility of the proposed research within the given constraints.

Chapter 2: Literature Review and Related Work

2.1. Conversational AI in Public Sector and Consular Services

The application of conversational AI in the public sector has evolved significantly, moving from basic rule-based chatbots to sophisticated LLM-powered conversational agents.1 These systems offer numerous benefits, including 24/7 availability, increased operational efficiency, substantial cost savings, improved citizen satisfaction through immediate responses, and enhanced accessibility for diverse populations.1 State-of-the-art (SOTA) features in government-focused conversational AI include advanced Natural Language Understanding (NLU), the emergence of AI "copilots" designed to assist human agents with complex tasks, multi-agent task coordination for handling intricate workflows, secure integration capabilities with legacy government systems, and Retrieval-Augmented Generation (RAG) for accessing and utilizing up-to-date external knowledge.1

Within the diplomatic and consular domain, conversational AI is being explored for various applications. These include managing high-demand services such as passport and visa applications, facilitating crisis communication by providing timely information

to affected citizens, supporting public diplomacy efforts through targeted messaging, and generally improving the dissemination of consular information.[8] The trend towards AI "copilots" in government [15] is particularly relevant. Early government chatbots were often restricted to answering simple, predefined FAQs. However, modern AI copilots offer far more advanced capabilities, including RAG, multi-agent orchestration, and secure integration with existing governmental IT infrastructure. These systems are engineered for complex operational environments and are designed to augment human decision-making rather than merely replacing simple interactions. Consular services inherently involve complex information, diverse user needs, and the necessity for accurate, context-aware responses. Therefore, a proposed system for consular query handling should aspire to these more advanced, copilot-like functionalities, making architectures like RAG highly pertinent.

However, the deployment of such systems is not without challenges. Ethical considerations, including algorithmic bias, the need for explainability in AI decision-making, ensuring data privacy and security, and building public trust, are paramount.[2]

2.2. Existing Consular Service Platforms in Indonesia

The Indonesian Ministry of Foreign Affairs (Kemlu) has made notable strides in digitalizing its consular services. Key existing platforms include:

- **Peduli WNI Portal:** This is a web-based platform designed to provide services and information to Indonesian citizens residing or traveling abroad.[3] While specific technological details of its underlying architecture are not extensively publicized, it serves as a central repository for consular information and citizen registration.
- **Safe Travel Application:** Developed by Kemlu, the Safe Travel mobile application offers practical information for Indonesian travelers, including destination-specific details, security conditions, legal customs, immigration requirements, and locations of Indonesian embassies or consulates.[5] The app also features emergency assistance capabilities and facilitates access to services provided by Indonesian diplomatic missions. Its design incorporates principles of dialogic communication to enhance digital diplomacy.[6] There are also indications of plans to integrate AI for features like personalized itinerary planning.[17]
- **SARI (Sahabat Artifisial Migran Indonesia) Chatbot:** A significant AI initiative by Kemlu, in collaboration with UN Women, is the SARI chatbot.[10] Integrated within the Safe Travel app, SARI is an AI-powered chatbot specifically designed to protect Indonesian female migrant workers by providing them with crucial information, support, and a safe channel to seek help without fear of stigma or prejudice.[10] The development of SARI emphasized a human-centered and participatory design process, incorporating gender bias-free data to ensure

empathetic and non-judgmental responses.[10] A notable feature of SARI is its reported ability to understand and respond in various Indonesian regional languages, catering to the diverse linguistic backgrounds of migrant workers.[12]

These existing platforms—Peduli WNI, Safe Travel, and SARI—represent more than just the current state of Kemlu's digital services; they are invaluable assets for the proposed research. The content hosted on Peduli WNI and Safe Travel, comprising official consular information, regulations, and FAQs, constitutes a rich, domain-specific knowledge source that can be leveraged to build the knowledge base for the proposed RAG system. The SARI chatbot, in particular, serves as an important precedent. Its existence signifies Kemlu's commitment to adopting AI for citizen-facing services. Furthermore, features highlighted in SARI's design, such as support for multiple Indonesian dialects and an emphasis on empathetic interaction [12], provide a useful baseline and indicate desirable characteristics for any new conversational AI system intended for Indonesian citizens. Thus, these platforms are not merely contextual background but potential data feeds and benchmarks that can inform and enrich the development of the proposed advanced consular query handling system.

2.3. State-of-the-Art Large Language Models for Bahasa Indonesia

The development of LLMs specifically tailored for Bahasa Indonesia has gained momentum, providing crucial tools for advancing NLP research and applications in the Indonesian context.

- **IndoBERT and its variants:** IndoBERT, built on the BERT architecture, was a pioneering model for Bahasa Indonesia, trained on the extensive Indo4B corpus (23.43GB) and comprising 124.5 million parameters for its base-p1 version.[19] It offers capabilities such as contextual word embeddings, masked language modeling (MLM), and next sentence prediction (NSP). IndoBERT has been successfully fine-tuned for various downstream tasks, including question answering (IndoBERT-QA, fine-tuned on a translated version of SQuAD v2.0).[21] A notable variant, SimCSE IndoBERT, focuses on generating high-quality sentence embeddings (768-dimensional vectors), which are particularly useful for semantic search and clustering tasks integral to RAG systems.[19]
- **SahabatAI:** Positioned as the first Indonesian-native LLM ecosystem, SahabatAI is a collaborative effort by Indosat Ooredoo Hutchison and GoTo.[23] This initiative has produced models like gemma2-9b-cpt-sahabatai-v1, a decoder model with an 8192-token context window, pre-trained on approximately 50 billion tokens, and supporting English, Indonesian, Javanese, and Sundanese.[24] Another significant model is llama3-8b-cpt-sahabatai-v1-instruct, which has been fine-tuned with a substantial number of instruction-completion pairs in Indonesian, Javanese, Sundanese, and English, making it adept at following

instructions.[26] SahabatAI is explicitly intended for applications like business-to-government (B2G) interactions, aiming to enhance government services.[23] However, it's important to note stated limitations, such as the potential for hallucination and the lack of specific safety alignment in current releases.[26]

- **Lazarus NLP Contributions:** The Lazarus NLP initiative has made significant contributions to Indonesian NLP resources:
  - **NusaBERT:** This model extends IndoBERT's capabilities to be multilingual and multicultural, covering Indonesian and 12 regional languages. It was fine-tuned from IndoBERT base p1 and pre-trained on approximately 16 billion tokens, demonstrating competitive performance on benchmarks like IndoNLU and NusaX.[28]
  - **IndoT5:** This is a T5-based sequence-to-sequence model specifically trained for Indonesian. It is designed for generative tasks such as summarization, question answering, and chit-chat, and was trained on the uonlp/CulturaX dataset.[28]
  - **Indonesian Sentence Embedding Models:** Lazarus NLP has also developed open-source sentence embedding models for Indonesian, which are directly applicable to the retrieval component of RAG systems.[28]
- **Other Initiatives and Benchmarks:** The Indonesian government and private sector are also exploring the development of other LLMs, such as a local version of China's DeepSeek, to promote AI innovation.[32] The NLP community relies on benchmarks like IndoNLU (covering tasks such as language modeling, sentence classification, and general NLU) [33] and NusaX (focusing on machine translation and sentiment analysis for Indonesian and its local languages) [35] to evaluate and compare model performance.

While IndoBERT is a more established model with a broader range of associated tools and research, newer models like SahabatAI and NusaBERT present distinct advantages. SahabatAI, being natively designed for Indonesian and its dialects and available in instruction-tuned versions, offers features particularly beneficial for a user-facing consular chatbot. NusaBERT's strength lies in its extended support for regional languages, which aligns with the need for accessible services, as demonstrated by the SARI chatbot's multilingual capabilities.[12] The selection of an LLM for this thesis will involve a careful trade-off between the maturity and community support of older models and the specialized, highly relevant features of newer ones. SahabatAI, with its focus on Indonesian languages and instruction-following capabilities, appears particularly promising, contingent on robust performance validation.

Table 2.1: Comparison of SOTA LLMs for Bahasa Indonesia

Feature	IndoBERT (base-p1)	SahabatAI (llama3-8b-cpt-sahabatai-v1-instruct)	NusaBERT-base	IndoT5 (base)
Base Architecture	BERT	Llama 3 (Decoder)	BERT (Encoder-based)	T5 (Encoder-Decoder)
Developer(s)	IndoBenchmark (Various Institutions)	GoTo Group, Indosat Ooredoo Hutchison, AI Singapore	LazarusNLP	LazarusNLP
Parameter Count	124.5M [20]	8B	111M [29]	Base size (similar to T5-base)
Pre-training Data	Indo4B (23.43 GB) [20]	Base model pre-trained on ~50B tokens; Instruct version fine-tuned on ~771k instruction pairs (ID, JV, SU, EN) [24]	~16B tokens (Indo Wikipedia, KoPI-NLLB, CulturaX) [29]	uonlp/CulturaX (23M Indonesian documents) [31]
Supported Languages	Bahasa Indonesia	Indonesian, Javanese, Sundanese, English [27]	Indonesian, Acehnese, Balinese, Banjarese, Buginese, Gorontalo, Javanese, Banyumasan, Minangkabau, Malay, Nias, Sundanese, Tetum [29]	Bahasa Indonesia [31]

Key Features	Contextual embeddings, MLM, NSP, Fine-tunable [20]	Instruction-tuned, Dialect support, Decoder-only [27]	Multilingual/Multicultural, Extends IndoBERT [28]	Sequence-to-sequence, Text generation tasks [31]
Performance (General)	Foundational for many Indonesian NLP tasks	High scores on SEA HELM benchmarks [26]	Competitive on IndoNLU, NusaX, NusaWrites [28]	Competitive on IndoNLG benchmark [31]
Suitability for Consular Chatbot	Good base, especially QA variants. Sentence embeddings useful for retriever.	Very high potential due to instruction-tuning, dialect support, and strong performance. Generator candidate.	Strong candidate for understanding diverse queries, especially if regional language input is a factor. Retriever/Generator.	Strong for generative aspects if fine-tuned for QA/dialogue. Generator candidate.

2.4. Key NLP Techniques for Intelligent Query Handling

The development of an intelligent query handling system relies on several core NLP techniques and architectural patterns.

- **Transformer Models:** At the heart of most modern LLMs, the transformer architecture, introduced by Vaswani et al. (2017), has revolutionized NLP. Its key innovation is the self-attention mechanism, which allows the model to weigh the importance of different words in an input sequence when processing information, leading to a superior understanding of context and long-range dependencies.[37] Transformers excel in a wide array of NLP tasks, including text generation, summarization, question answering, and sentiment analysis [37], and have even been adapted for tasks like interpreting relational keyword queries over databases.[38]
- **Retrieval-Augmented Generation (RAG):** RAG is an architectural framework that enhances the capabilities of generative LLMs by dynamically incorporating information from external knowledge sources.[15] The typical RAG process involves receiving a user query, using it to retrieve relevant document chunks from a knowledge base (often a vector database queried using semantic search), augmenting the original query with this retrieved context, and then feeding this augmented prompt to an LLM to generate a response.[39] The primary benefits of RAG include improved factual grounding (reducing LLM "hallucinations"), the ability to access fresh and domain-specific information without retraining the

entire LLM, and often being more cost-effective than extensive model retraining.[39] Key components are the retriever (which employs techniques like vector embeddings and semantic search) and the generator (the LLM itself).[39]

- **Fine-tuning LLMs:** Fine-tuning is the process of taking a pre-trained LLM and further training it on a smaller, task-specific or domain-specific dataset.[43] This adaptation allows the model to achieve better performance on the target task or to align its responses with a particular style, tone, or terminology.[44] Examples include fine-tuning IndoBERT for question answering [21] or adapting SahabatAI for specialized applications.[43]
- **Comparative Analysis (RAG vs. Fine-tuning):** The choice between RAG and fine-tuning, or a combination thereof, is a critical design decision. Research indicates that RAG often outperforms or complements fine-tuning for domain-specific chatbots, particularly when factual accuracy and access to dynamic data are crucial.[45] While fine-tuning can lead to high accuracy if substantial, high-quality domain-specific training data is available, it may suffer from knowledge cut-offs (the model only knows what it was trained on up to a certain point) and can still hallucinate facts not explicitly present in its fine-tuning dataset. Moreover, keeping a fully fine-tuned model updated with new information can be resource-intensive and costly.[43] Prompt engineering, while cost-effective, might not offer the same level of robustness for complex domains as RAG or thorough fine-tuning.[45]

The literature and practical considerations suggest that RAG and fine-tuning are not necessarily mutually exclusive approaches. For a domain like consular services, where both accurate factual recall from official documents and nuanced understanding of specific terminology and query styles are important, a synergistic approach could be most effective. An LLM could first be fine-tuned on general consular dialogue patterns or terminology in Bahasa Indonesia. This step would enhance its intrinsic understanding of the consular domain, improving its ability to interpret user intents, adopt an appropriate communication style, and understand specific jargon. Subsequently, this domain-adapted LLM could serve as the generator component within a RAG framework. In this setup, the RAG mechanism would be responsible for retrieving the most relevant and up-to-date factual information from the official consular knowledge base, while the fine-tuned LLM would excel at synthesizing this information into coherent, contextually appropriate, and stylistically fitting responses. This hybrid strategy aims to leverage the strengths of both techniques: fine-tuning for domain-specific linguistic adaptation and RAG for dynamic factual grounding, thereby mitigating the limitations each method might face when used in isolation.

2.5. Challenges in AI Implementation for Government Services in Indonesia
The successful deployment of AI in Indonesian government services, including consular services, faces several systemic challenges that must be acknowledged. These include gaps in digital infrastructure and the need for widespread technical training.47 Skill gaps among the existing workforce can hinder the adoption and effective utilization of AI technologies, necessitating significant investment in upskilling and reskilling programs.47 Furthermore, the implementation of basic operational standards, which could extend to data governance and quality, may be uneven across different sectors and agencies, potentially impacting the data available for AI systems.47
Beyond technical and infrastructural aspects, there are societal challenges. The potential for AI misuse, such as the propagation of disinformation and misinformation, can erode public trust in both the technology and government institutions.[48] This underscores the critical need for enhanced digital literacy among the general public and government officials to enable critical assessment of AI-generated content and responsible AI usage.[48] General concerns about data privacy and security, inherent in many AI applications, are particularly acute when dealing with sensitive government and citizen data, requiring robust safeguards and transparent policies.[49]

While the primary focus of this thesis is on the technical development of a conversational AI system, these broader implementation challenges in Indonesia are crucial for contextualizing the potential real-world impact and inherent limitations of the proposed research. The project must concentrate on aspects within its control, such as ensuring model accuracy, robust data handling practices within the project's scope, and thorough documentation. Simultaneously, it must remain cognizant of external factors that could influence the eventual deployment and widespread adoption of such a system. For instance, designing the system using readily available open-source tools and well-documented models can help lower potential adoption barriers for Kemlu or other agencies, contributing to the project's practical relevance despite these wider challenges.

Chapter 3: Proposed Methodology

3.1. Overview of the Proposed System Architecture
The proposed conversational AI system for handling Indonesian consular service queries will be architected around a Retrieval-Augmented Generation (RAG) framework, leveraging a domain-adapted Indonesian Large Language Model (LLM). A high-level depiction of the system architecture is shown below:

Code snippet

```
graph TD
    A[User Interface (Web/Mobile)] --> B{Query Preprocessor};
    B --> C{Dialogue Manager (Simplified)};
    C --> D[Core NLU/NLG Engine];
    D -- User Query + Context --> E[Generator (Fine-tuned Indonesian LLM)];
    D -- User Query --> F;
    F -- Query Embedding --> G;
    G -- Retrieved Chunks --> D;
    E -- Generated Response --> C;
    C --> H{Response Postprocessor};
    H --> A;
    I[(Optional) Feedback Loop] --> G;
    I --> E;
```

Figure 3.1: High-Level System Architecture

The key modules and their interactions are as follows:

1. **User Interface (UI):** A simple web-based interface (e.g., built with Streamlit or Flask) will allow users to input their queries in Bahasa Indonesia.
2. **Query Preprocessor:** This module will receive the raw user query and perform initial cleaning, such as typo correction (if feasible with available libraries for Bahasa Indonesia) and normalization.
3. **Dialogue Manager (Simplified):** For this thesis, the dialogue manager will be kept relatively simple, focusing on single-turn Q&A or basic multi-turn interactions (e.g., handling follow-up questions by retaining minimal context from the immediate previous turn). It will orchestrate the flow between the NLU/NLG engine and the user.
4. **Core NLU/NLG Engine (RAG Pipeline):** This is the heart of the system.
    - The **Retriever** component will take the processed user query, embed it, and search the Knowledge Base for relevant document chunks.
    - The retrieved chunks, along with the original query, will form an augmented prompt.
    - The **Generator** (the fine-tuned Indonesian LLM) will take this augmented prompt and generate a coherent, factually grounded answer in Bahasa Indonesia.
5. **Knowledge Base (KB):** A vector database containing embeddings of processed Indonesian consular documents, FAQs, and other relevant information.
6. **Response Postprocessor:** This module may perform minor cleaning on the

generated response before it is displayed to the user.

7. **(Optional) Feedback Loop:** While full implementation is out of scope, the design will consider how user feedback (e.g., on answer relevance or correctness) could be collected to iteratively improve the KB or the model.

The flow of information begins with the user submitting a query through the UI. The query is preprocessed and passed to the Dialogue Manager, which then engages the Core NLU/NLG Engine. The Retriever fetches relevant context from the KB. This context, combined with the query, is used by the Generator LLM to produce an answer. The answer is then postprocessed and returned to the user via the UI.

3.2. Chosen Technique, Model, and Method: RAG with a Domain-Adapted Indonesian LLM

3.2.1. Detailed Explanation of the Chosen Approach: Retrieval-Augmented Generation (RAG)
The core of the proposed system is the Retrieval-Augmented Generation (RAG) technique. This choice is driven by RAG's demonstrated strength in producing factually grounded, up-to-date, and domain-specific responses, which are critical attributes for a consular information service where accuracy is paramount.39 RAG effectively mitigates the common LLM issue of "hallucination" by anchoring responses to information retrieved from a trusted knowledge source.

- **Retriever Component:**
  - **Input:** The user's query in Bahasa Indonesia.
  - **Process:**
    1. **Query Embedding:** The query will be transformed into a dense vector representation using a sentence transformer model optimized for Bahasa Indonesia. Candidates include models from Lazarus NLP (e.g., LazarusNLP/all-indo-e5-small-v4 [30]) or SimCSE IndoBERT [19], selected for their performance in capturing semantic meaning in Indonesian.
    2. **Similarity Search:** The query embedding will be used to search the vector index of the consular knowledge base using a similarity metric (e.g., cosine similarity).
    3. **Retrieval:** The top-k most relevant document chunks (passages) will be retrieved. The value of 'k' will be an experimental parameter.
  - **Techniques:** The primary technique will be Dense Passage Retrieval (DPR). Consideration will also be given to hybrid search approaches that combine semantic search with traditional keyword-based search if initial DPR results show limitations in retrieving specific entities or terms.
- **Knowledge Base (KB):**
  - The construction of the KB is detailed in Chapter 4. It will comprise processed

and vectorized content from official Kemlu documents, FAQs, and information from platforms like Peduli WNI and Safe Travel.
    - **Chunking Strategy:** Documents will be segmented into smaller, manageable chunks. The optimal chunking strategy (e.g., paragraph-based, fixed-size overlapping) will be determined experimentally to balance context preservation with retrieval precision.
    - **Vectorization and Storage:** Each chunk will be vectorized using the same sentence transformer model employed for query embedding. These embeddings, along with the raw text of the chunks and any associated metadata, will be stored in a vector database. For the prototype, an efficient open-source solution like FAISS (Facebook AI Similarity Search) will be utilized due to its performance and ease of integration.
- **Generator Component (LLM):**
    - **Chosen Model:** The generator will be a state-of-the-art Indonesian LLM. Prime candidates are instruction-tuned models from the **SahabatAI** family, such as **Gemma2 9B CPT Sahabat-AI v1 Instruct** or **Llama3 8B CPT Sahabat-AI v1 Instruct** [24], or a well-performing encoder-decoder model like **NusaBERT-large**.[29] The selection will prioritize models with strong performance in Bahasa Indonesia, support for local dialects (a feature of SahabatAI), and robust instruction-following capabilities. The final choice may be refined after initial experimentation if the timeline permits; otherwise, it will be based on current benchmark data and feature alignment with project needs.
    - **Domain Adaptation/Fine-tuning:** The selected LLM will undergo a targeted fine-tuning process. This will involve training on a relatively small, curated dataset of general consular question-answer pairs or dialogue snippets in Bahasa Indonesia. The goal of this fine-tuning is not to imbue the LLM with all specific consular facts (as this is the role of the RAG mechanism) but rather to adapt its language style, improve its understanding of common consular intents and terminology, and enhance its ability to synthesize information from provided context effectively.[43]
    - **Input:** The original user query concatenated or structured with the retrieved context chunks from the KB.
    - **Process:** The LLM will generate a coherent, contextually relevant, and factually grounded answer in Bahasa Indonesia. The generation process will be guided by the augmented prompt.
    - **Prompt Engineering:** Significant effort will be dedicated to crafting effective prompts. These prompts will instruct the LLM to:
        - Synthesize information primarily from the provided context.

    - Directly answer the user's specific query.
    - Maintain a helpful, respectful, and official tone appropriate for consular services.
    - If feasible and appropriate, cite the source of the information (e.g., by referring to document titles or sections if metadata is available and passed in context).
    - Handle cases where the retrieved context might be insufficient or ambiguous gracefully (e.g., by stating that specific information isn't available in the provided documents or asking for clarification).

3.2.2. Justification for the Chosen Approach

The selection of RAG with a domain-adapted Indonesian LLM is underpinned by several key advantages over alternative methodologies, especially in the context of providing reliable consular information:

- **Superiority for Factual Accuracy:** A primary concern with LLMs is their propensity to "hallucinate" or generate incorrect information. RAG is specifically designed to mitigate this by grounding the LLM's responses in factual documents retrieved from an authoritative knowledge base. For consular services, where misinformation can have serious repercussions for citizens, this factual grounding is non-negotiable.[39]
- **Handling Domain-Specific & Dynamic Knowledge:** Consular information, including regulations, procedures, and advisories, is subject to change. A RAG system allows the knowledge base to be updated independently of the LLM. New or revised documents can be processed and added to the vector store without the need for costly and time-consuming retraining of the entire LLM. This makes the system more adaptable and maintainable compared to relying solely on a statically fine-tuned model.[40]
- **Effectiveness for Query Handling:** Transformer-based LLMs possess remarkable capabilities in understanding the nuances of natural language queries.[37] RAG leverages this strength twice: first, the retriever component uses query understanding to fetch the most relevant context, and second, the generator LLM uses its understanding of both the query and the retrieved context to formulate a precise and coherent answer.
- **Comparison with Alternatives:**
    - **Pure Fine-tuning of LLMs:** While fine-tuning can adapt an LLM to the consular domain and improve its performance on specific tasks [45], it has inherent limitations. The model's knowledge is essentially frozen at the time of its last training, making it difficult to incorporate new information without retraining. Furthermore, even fine-tuned models can generate plausible but

incorrect statements if the information is not explicitly and robustly encoded in their parameters. Keeping a large, fully fine-tuned model current with evolving consular information would be a significant operational burden.
    - **Traditional NLU/Rule-Based Systems:** These systems, common in earlier generations of chatbots, lack the flexibility and sophisticated natural language understanding of LLMs. They often struggle with queries phrased in unexpected ways, out-of-vocabulary terms, or complex sentence structures, leading to brittle performance and user frustration.
    - **Simpler Retrieval Systems (e.g., TF-IDF + Basic QA):** Methods like TF-IDF or BM25 for retrieval, while useful, may not capture semantic similarity as effectively as dense retrieval methods based on sentence embeddings. Moreover, they typically lack the advanced generative capabilities of LLMs needed to produce fluid, natural, and contextually synthesized conversational responses.
- **Alignment with SOTA in Government AI:** The RAG architecture is increasingly recognized as a key feature in advanced AI copilots and conversational systems being developed and deployed in government settings, signifying its relevance and potential for high-impact applications.[15]
- **Feasibility for Master's Thesis:** The modular nature of the RAG architecture (distinct retriever and generator components) allows for focused development and evaluation within a constrained timeframe. The availability of powerful open-source tools for vector databases (e.g., FAISS), pre-trained sentence transformers, and Indonesian LLMs (like SahabatAI or NusaBERT) makes the implementation of a prototype feasible for a Master's student. The fine-tuning component, as proposed, will be targeted towards domain adaptation using a smaller, curated dataset, rather than extensive training from scratch.

Table 3.1: Comparative Analysis of Proposed Approach (RAG + Fine-tuned Indonesian LLM) vs. Alternative Methods for Consular Query Handling

Criteria	Proposed Approach (RAG + Fine-tuned Indonesian LLM)	Pure LLM Fine-tuning	Traditional NLU/Rule-Based Systems	Basic Retrieval QA (e.g., BM25 + Extractive QA)
Factual Accuracy	High (grounded in external KB, reduces	Medium-High (dependent on training data	Low-Medium (limited to predefined	Medium (dependent on retriever

	hallucination [39])	quality and coverage, still prone to hallucination [45])	rules/knowledge)	accuracy, can present irrelevant info)
Handling Dynamic Data/Updates	High (KB can be updated independently of LLM retraining [40])	Low (requires model retraining for new knowledge)	Low (requires manual rule/database updates)	Medium (KB can be updated, but QA logic is static)
Scalability of Knowledge	High (can scale with KB size)	Medium (limited by model parameters and fine-tuning data size)	Low (difficult to scale complex rule sets)	High (can scale with KB size)
Robustness to Novel Queries	Medium-High (LLM's general NLU + retrieval for unseen topics if in KB)	Medium (generalizes from fine-tuning data, but struggles with out-of-distribution queries)	Low (fails on queries not matching rules)	Low-Medium (struggles with semantic variations not caught by retriever)
Development Complexity/Cost	Medium-High (requires LLM, vector DB, retriever setup; fine-tuning effort)	High (significant data preparation and GPU resources for effective fine-tuning [45])	Medium (complex rule engineering for broad coverage)	Low-Medium (simpler to implement)
Resource Requirements (Inference)	Medium-High (LLM inference + retrieval)	Medium-High (LLM inference)	Low	Low
Indonesian Language Nuance Handling	High (leverages SOTA Indonesian LLMs fine-tuned for domain [24])	High (if fine-tuned on diverse Indonesian data)	Low (typically struggles with linguistic variation)	Low (basic keyword matching or simple embeddings may miss nuances)

3.3. Core Components (Elaboration)

- **Query Understanding Module:**
  - **Input:** Raw user query in Bahasa Indonesia.
  - **Tasks:**
    - *Language Detection:* While the primary language is Bahasa Indonesia, this step ensures queries in other languages are not inadvertently processed or can be flagged.
    - *Query Normalization:* This involves correcting common typographical errors and expanding common abbreviations or colloquialisms relevant to consular services in Bahasa Indonesia. This step is crucial for improving the robustness of the retriever.
    - *Initial Intent Classification:* A lightweight classification might be performed to distinguish between informational queries (target of this system), greetings, or simple chitchat. This can help route queries appropriately or set initial dialogue parameters. This could leverage the main LLM's NLU capabilities or a smaller, faster classification model.
- **Document Retrieval Module (Retriever):**
  - As described in Section 3.2.1. The choice of an Indonesian sentence embedding model is critical. Models such as LazarusNLP/all-indo-e5-small-v4 [30] or lazarusnlp/simcse-indobert-base [19] are strong candidates due to their specific training on Indonesian text, enabling better semantic understanding for retrieval.
  - **Vector Database:** FAISS will be used for the prototype due to its efficiency, scalability for research purposes, and ease of integration with Python-based NLP pipelines. It allows for fast similarity searches in high-dimensional vector spaces.
- **Response Generation Module (Generator):**
  - As detailed in Section 3.2.1. The fine-tuned SahabatAI (Instruct version) or NusaBERT will serve as the generator.
  - **Prompting Strategies:** Emphasis will be placed on developing robust prompting strategies that explicitly instruct the LLM to:
    - Prioritize and synthesize information *only* from the retrieved context.
    - Clearly indicate if the answer cannot be found in the provided context, rather than speculating.
    - Maintain an official, empathetic, and helpful tone.
    - Handle ambiguous queries or conflicting information in the retrieved context by either asking for clarification or presenting the information cautiously.

- **Dialogue Management (Simplified):**
    - Given the six-month timeframe, a sophisticated, stateful dialogue manager is likely out of scope. The primary focus will be on effectively handling single-turn question-answering scenarios.
    - Basic context carry-over for simple follow-up questions will be explored. This could involve including the previous user query and system response as part of the input to the LLM for the current turn, allowing it to understand pronominal references or elliptical queries.
    - Strategies for clarification will be implemented. If a user's query is too ambiguous for effective retrieval, or if the retrieved context is insufficient or conflicting, the system will be designed to politely request clarification or indicate the limitations of the available information.

3.4. Integration with Existing Knowledge Sources (Conceptual)
While direct, real-time API integration with Kemlu's live internal systems is beyond the scope of this Master's thesis, the methodology assumes that content from existing public-facing knowledge sources can be obtained for building the RAG system's knowledge base. These sources include:

- The official Kemlu website (kemlu.go.id) and the websites of various Indonesian Embassies and Consulates.
- Content from the Peduli WNI portal.[3]
- Information available through the Safe Travel application, including FAQs and emergency procedures.[5]
- Publicly available government decrees, laws, and regulations pertaining to consular affairs.

If any public documentation, anonymized Q&A logs, or structured data from the SARI chatbot initiative [10] were to become accessible (e.g., through public reports or potential future collaboration, though not assumed for this thesis), such data could also be valuable for incorporation.

The primary challenge in utilizing these diverse sources lies in the conversion of their content—which may exist in various formats like HTML, PDF, or structured app data—into a clean, organized, and appropriately chunked format suitable for ingestion into the RAG system's vector knowledge base. This preprocessing step, detailed further in Chapter 4, is critical. The digital transformation efforts already undertaken by Kemlu, resulting in these platforms, mean that a substantial amount of relevant digital content already exists. The task, therefore, is less about de novo content creation and more about effectively curating, extracting, and structuring this existing wealth of information for consumption by an AI system. This makes the knowledge

management aspect a significant, yet manageable, part of the project.

Chapter 4: Data Collection and Preparation

4.1. Data Requirements for Training and Evaluation
The proposed RAG-based conversational AI system requires several distinct types of data for its development and robust evaluation:

1. **For the RAG Knowledge Base (KB):**
    - A comprehensive corpus of documents, articles, and informational snippets pertaining to Indonesian consular services. This includes, but is not limited to, laws, official regulations, procedural guidelines for various services (e.g., passport issuance/renewal, visa applications, legalization of documents), emergency protocols, information on citizen rights and responsibilities abroad, and FAQs.
    - This data must primarily be in Bahasa Indonesia to serve the target user base.
    - The quality, accuracy, and currency of this data are paramount as it forms the factual backbone of the system.
2. **For Fine-tuning the LLM (Generator):**
    - A dataset of question-answer (Q&A) pairs or short conversational dialogues specifically relevant to consular services, in Bahasa Indonesia.
    - This dataset will be used to adapt the chosen pre-trained Indonesian LLM to the specific linguistic style, common terminology, and typical query structures encountered in the consular domain.
    - For a six-month thesis, the focus will be on a high-quality, albeit potentially small (hundreds to a few thousand examples), fine-tuning dataset.
3. **For Evaluating the Retriever Component:**
    - A set of representative user queries in Bahasa Indonesia.
    - For each query, a corresponding set of "gold" or ideal document chunks from the KB that should be retrieved as relevant context. This requires manual annotation.
4. **For End-to-End Evaluation of the RAG System:**
    - A test set of diverse user queries in Bahasa Indonesia, covering various consular topics.
    - For each query, an ideal, factually correct, and well-phrased answer. This will serve as the reference against which the system's generated responses are compared. This also requires careful manual creation or validation.

4.2. Potential Data Sources
The primary data sources for constructing the knowledge base and potentially for deriving fine-tuning/evaluation data will be:

- **Official Kemlu Websites:** The main Ministry of Foreign Affairs website (kemlu.go.id) and the official websites of Indonesian Embassies and Consulates General worldwide. These are primary sources for official announcements, service information, and contact details.
- **Peduli WNI Portal:** This portal is a dedicated platform for Indonesian citizens abroad and is expected to contain structured information, FAQs, and guidelines relevant to consular assistance.[3]
- **Safe Travel Application:** Information embedded within the Safe Travel mobile application, including travel advisories, emergency contact procedures, and service details, can be extracted if accessible (e.g., through web versions or public documentation).[5]
- **SARI Chatbot Knowledge (Aspirational):** If any public documentation, anonymized Q&A pairs, or knowledge snippets from the SARI chatbot project become available, they would be highly relevant, particularly for understanding common query types related to citizen protection.[10] However, reliance on this is not assumed.
- **Publicly Available Government Decrees and Regulations:** Official government publications (e.g., "Peraturan Menteri Luar Negeri") concerning consular affairs, citizenship, immigration, etc.
- **Anonymized and Aggregated FAQs/Query Logs (Highly Aspirational):** Access to anonymized query logs from existing Kemlu helpdesks or contact centers would be invaluable for understanding real user needs and for creating realistic fine-tuning and evaluation datasets. However, obtaining such data is complex due to privacy and bureaucratic hurdles and is not a prerequisite for this thesis. The focus will remain on publicly accessible data.
- **General Chatbot Datasets (for Structural Reference):** Publicly available datasets for customer support or informational chatbots, such as the Bitext Customer Support dataset [51] or the FAQ-based dataset used for the Prambanan Temple chatbot [52], can provide structural templates and ideas for formatting Q&A pairs, even though their domain content differs.

The existing digital assets created by Kemlu (Peduli WNI, Safe Travel, and potentially SARI) signify that a considerable volume of relevant digital content is already available. The core challenge is not an absolute lack of information but rather ensuring its accessibility for this project, its current format (e.g., HTML, PDF, app-specific data structures), and its overall structure, which may not be immediately amenable to AI consumption. Therefore, data collection will heavily involve strategies for extracting and transforming this existing digital information.

4.3. Data Collection Strategy

- **Phase 1: Publicly Available Document Collection (Primary Focus):**
  - **Web Scraping:** Systematic scraping of relevant sections from official Kemlu websites, embassy/consulate websites, and the public-facing areas of the Peduli WNI portal. Python libraries such as Scrapy and BeautifulSoup will be employed for this task. Ethical scraping practices (e.g., respecting robots.txt, rate limiting) will be followed.
  - **Document Downloading:** Identifying and downloading publicly available PDF documents containing consular regulations, official guides, circulars, and reports.
- **Phase 2: Synthetic Data Generation (for Fine-tuning/Evaluation - Contingency):**
  - If the collection of naturally occurring Q&A pairs for fine-tuning proves insufficient, carefully controlled synthetic data generation will be considered. This might involve using a powerful base LLM (e.g., GPT-3.5/4, if ethically permissible and resources allow) with meticulously crafted prompts to generate plausible question-answer pairs based *only* on the collected official documents. Any synthetically generated data will undergo rigorous manual review and validation for accuracy and relevance by the researcher.
  - Another approach is to paraphrase existing FAQs from Kemlu sources to increase the diversity of phrasing in the fine-tuning dataset.
- **Ethical Considerations and Data Privacy:**
  - The project will prioritize the use of publicly available information to avoid complexities related to data privacy and access permissions.
  - No Personal Identifiable Information (PII) of citizens or consular staff will be collected or used in the knowledge base or training datasets. All data will pertain to general consular procedures, regulations, and information.
  - The principle of data minimization will be strictly adhered to, collecting only what is necessary for the system's intended functionality.[49]
  - Transparency regarding the data sources used to build the knowledge base will be maintained in the thesis documentation.

4.4. Data Preprocessing and Annotation Plan

- **Knowledge Base Preprocessing (for RAG):**
  - **Format Conversion:** Convert all collected data (scraped HTML, PDFs, etc.) into a consistent clean text format (e.g., plain text files, JSON, or CSV structures).[53] Tools like pdfminer.six or PyMuPDF for PDF extraction will be used.

    - **Cleaning:** Remove irrelevant HTML tags, navigation menus, advertisements, boilerplate text (headers, footers), and other noise. Correct any obvious OCR errors if scanned documents are part of the corpus.
    - **Segmentation/Chunking:** This is a critical step. Long documents will be segmented into smaller, semantically coherent chunks. The ideal chunk should be self-contained enough to answer a specific type of query but not so long as to dilute the information or exceed the context window of the LLM. Strategies will include:
        - Paragraph-based splitting.
        - Section-based splitting (if documents have clear structural headings).
        - Fixed-size overlapping chunks (e.g., 256-512 tokens with an overlap of 50-100 tokens) as a fallback or for unstructured text. The optimal chunk size and overlap will be an experimental parameter.
    - **Metadata Tagging (Recommended):** Where possible, associate metadata with each chunk, such as the source URL, original document title, section heading, and last updated date (if available). This metadata can be invaluable for traceability, source attribution in responses, and potentially for filtering or prioritizing retrieved chunks.
- **Fine-tuning Dataset Preprocessing:**
    - Ensure Q&A pairs or conversational turns are formatted consistently (e.g., JSON objects with "question" and "answer" keys, or a specific conversational format if using models like DialoGPT-style fine-tuning).
    - Perform text cleaning and normalization similar to KB preprocessing.
- **Annotation (Primarily for Evaluation Sets):**
    - **Gold Answers:** For the end-to-end evaluation query set, ideal answers will be manually crafted or validated. These answers should be factually correct, comprehensive yet concise, and well-phrased in Bahasa Indonesia.
    - **Relevant Chunks:** For a subset of queries intended to evaluate the retriever, the specific document chunks from the KB that contain the necessary information to answer the query will be manually identified and labeled.
    - This annotation process will be time-intensive. The strategy will be to create a smaller, high-quality annotated evaluation set rather than a large, noisy one. The researcher will perform the primary annotation, potentially with cross-validation if a peer is available.

4.5. Sample Planned Dataset

Table 4.1: Proposed Dataset Specification (Illustrative Examples)

Data Type	Source/Format Example	Raw Data Snippet (Bahasa Indonesia)	Processed Chunk (for Vector DB)	Metadata Example
KB: Official Web Page	URL: https://kemlu.go.id/portal/id/read/XYZ/konsuler/perpanjangan-paspor-habis-berlaku	"Paspor yang masa berlakunya habis dapat diperpanjang. Proses perpanjangan paspor memerlukan dokumen asli seperti paspor lama, KTP, dan Kartu Keluarga. Biaya perpanjangan adalah Rp XXX.XXX,-. Waktu proses standar adalah 3-5 hari kerja setelah semua dokumen lengkap."	"Perpanjangan paspor habis berlaku. Dokumen: paspor lama asli, KTP, KK. Biaya: Rp XXX.XXX,-. Proses: 3-5 hari kerja."	{"source_url": "...", "title": "Perpanjangan Paspor Habis Berlaku", "category": "Paspor", "last_crawled": "YYYY-MM-DD"}
KB: PDF Document	File: PERMENLU_No_X_Tahun_YYYY.pdf, Section 3.1	"Warga Negara Indonesia yang kehilangan paspor di luar negeri wajib segera melaporkan kepada Perwakilan Republik Indonesia terdekat untuk mendapatkan Surat Perjalanan Laksana Paspor (SPLP)."	"Kehilangan paspor di luar negeri: WNI wajib lapor ke Perwakilan RI terdekat untuk SPLP."	{"source_file": "PERMENLU_No_X...", "section": "3.1", "topic": "Kehilangan Paspor"}

Fine-tuning Q&A Pair	Format: JSON {"question": "...", "answer": "..."}	N/A	N/A	N/A
Example		{"question": "Bagaimana cara perpanjang paspor yang sudah habis masa berlakunya?", "answer": "Untuk memperpanjang paspor yang habis masa berlakunya, Anda perlu menyiapkan dokumen asli seperti paspor lama, KTP, dan Kartu Keluarga. Anda dapat mengajukan permohonan di kantor imigrasi atau Perwakilan RI jika berada di luar negeri. Ada biaya administrasi yang perlu dibayarkan dan prosesnya biasanya memakan waktu beberapa hari kerja."}		
Evaluation Q&A Pair	Format: JSON {"query_id": "...", "query_text": "...", "ideal_answer": "...",	N/A	N/A	N/A

	"relevant_doc_ids": ["..."]}			
Example		{"query_id": "PASPOR_EXT_001", "query_text": "Paspor saya mau habis, gimana cara perpanjangnya ya?", "ideal_answer": "Untuk perpanjangan paspor yang akan habis masa berlakunya, Anda perlu membawa paspor lama asli, KTP, dan Kartu Keluarga. Biaya perpanjangannya adalah Rp XXX.XXX,- dan prosesnya sekitar 3-5 hari kerja. Pengajuan bisa dilakukan di kantor imigrasi atau Perwakilan RI.", "relevant_doc_ids": ["chunk_id_paspor_perpanjangan_01", "chunk_id_paspor_biaya_02"]}		

This table illustrates the transformation of data from its raw form into structured formats suitable for the different components of the RAG system and its evaluation. It underscores the importance of careful processing and annotation to ensure data

quality.

4.6. Knowledge Base Creation for RAG

The creation of a high-quality, comprehensive knowledge base (KB) is fundamental to the success of the RAG system. The process will follow established best practices 53 and involve these detailed steps:

1. **Identify Knowledge Sources:** As outlined in Section 4.2, this involves pinpointing official Kemlu websites, relevant portals (Peduli WNI, Safe Travel), public documents, and regulations.
2. **Data Extraction and Ingestion:** Develop and utilize automated scripts (Python with Scrapy/BeautifulSoup for web content, PDF text extraction libraries for documents) to gather the raw textual data from these identified sources.
3. **Preprocessing & Cleaning:** This stage is crucial for removing noise and standardizing the data (as detailed in Section 4.4). This includes removing HTML artifacts, irrelevant boilerplate text, handling special characters, and potentially normalizing variations in terminology if consistent patterns are identified.
4. **Chunking Strategy:** Documents will be segmented into smaller, semantically meaningful chunks.
    - Initial experiments will compare fixed-size chunking (e.g., 200-300 words with 50-word overlap) with more semantic approaches like splitting by paragraphs or logical sections identified by HTML tags (e.g., <h2>, <p>).
    - The goal is to create chunks that are self-contained enough to answer specific types of queries but not so large as to introduce excessive noise or exceed LLM context limits during generation.
5. **Embedding Generation:** Each cleaned and chunked piece of text will be converted into a dense vector embedding using the selected Indonesian sentence transformer model (e.g., from Lazarus NLP or SimCSE IndoBERT). These embeddings capture the semantic meaning of the text chunks.
6. **Vector Store Indexing:** The generated embeddings, along with their corresponding raw text chunks and any extracted metadata (source, title, etc.), will be loaded into a vector database (FAISS for this project). An efficient index will be built to enable fast similarity searches based on query embeddings.
7. **Knowledge Base Updating Strategy (Conceptual):** While the automated, continuous updating of the KB is beyond the scope of a 6-month Master's thesis, the design will acknowledge its importance for a production system. In a real-world deployment, a mechanism for periodically re-scraping sources, identifying changes, and updating the vector store with new or modified information would be essential to maintain the currency and accuracy of the chatbot's knowledge.[40]

The process of creating and refining the KB will likely be iterative. Initial attempts at data extraction, cleaning, and chunking might not be perfect. The performance of the retriever component (evaluated in Chapter 5) will provide direct feedback on the quality of the KB. For instance, if the retriever frequently fails to find relevant chunks or retrieves noisy ones, it will indicate a need to revisit and refine the preprocessing scripts, chunking logic, or even the source selection. This iterative loop of KB construction, retriever testing, and KB refinement is key to building an effective RAG system.

Chapter 5: Training, Implementation, and Evaluation Plan

5.1. Model Training Strategy

The training strategy involves two main components: ensuring an effective retriever and adapting the generator LLM for the consular domain.

- **Retriever Training (Selection and Configuration):**
  - The primary approach will be to use high-quality, pre-trained sentence transformer models specifically designed or well-adapted for Bahasa Indonesia. Models from Lazarus NLP, such as LazarusNLP/all-indo-e5-small-v4 [30], or lazarusnlp/simcse-indobert-base [19], are strong candidates as they have been trained on Indonesian text and are designed for tasks like semantic similarity and information retrieval.
  - Fine-tuning the sentence transformer itself is generally a complex task requiring large, specialized datasets (e.g., Indonesian Semantic Textual Similarity datasets) and significant computational resources. Given the 6-month timeframe, this is likely out of scope. The focus will be on selecting the best available pre-trained model and optimizing its usage within the RAG pipeline (e.g., through appropriate chunking of the knowledge base).
- **Generator (LLM) Fine-tuning:**
  - **Base Model Selection:** An instruction-tuned Indonesian LLM will be prioritized to reduce the fine-tuning burden. Candidates include **SahabatAI Llama3 8B CPT Sahabat-AI v1 Instruct** [26] or **SahabatAI Gemma2 9B CPT Sahabat-AI v1 Instruct** [24], due to their native Indonesian design, instruction-following capabilities, and support for regional dialects. If these prove difficult to access or work with, **NusaBERT-large** [29] (fine-tuned from IndoBERT) could be considered, though it might require more effort to elicit instruction-following behavior if not already adapted for it.
  - **Fine-tuning Dataset:** A curated dataset of question-answer pairs and/or conversational snippets specific to Indonesian consular services (as described in Chapter 4) will be used. The size will likely be in the range of

hundreds to a few thousand high-quality examples.
    - **Fine-tuning Objective:** The primary goal is not to teach the LLM all consular facts (RAG handles this) but to:
        - Adapt its language style to be appropriate for official consular communication (formal, empathetic, clear).
        - Familiarize it with common consular terminology and query patterns in Bahasa Indonesia.
        - Improve its ability to follow instructions for synthesizing answers strictly from the provided RAG context.
        - Enhance its ability to generate responses in a consistent and desired format.
    - **Technique:** Supervised Fine-Tuning (SFT) will be employed, where the model learns to generate target responses given input prompts (which, during inference, will be the augmented query + context).
    - **Hyperparameters:** Key hyperparameters such as learning rate, batch size, and number of training epochs will be carefully chosen. Initial values will be based on best practices for the selected LLM architecture (e.g., referring to training details provided for SahabatAI models if available [24]) and then potentially tuned through limited experimentation on a validation set if time permits.
    - **Frameworks:** The Hugging Face transformers library, along with PyTorch, will be the primary frameworks for implementing the fine-tuning process.

Leveraging an existing instruction-tuned model like SahabatAI-Instruct is a strategic choice to manage the fine-tuning effort within the thesis timeline. These models have already undergone extensive training to understand and follow instructions in Indonesian.[26] This allows the thesis research to concentrate on the more targeted task of adapting the model to the specific nuances of the consular domain—its language, style, and typical interaction patterns—rather than expending significant resources on teaching basic instruction-following from a non-instruct base model. This focused approach makes the fine-tuning task more feasible and directly relevant to improving performance on consular queries.

5.2. Implementation Details

- **Programming Language:** Python will be the primary programming language due to its extensive ecosystem of libraries for NLP, machine learning, and data processing.
- **Key Libraries and Frameworks:**
    - **Hugging Face transformers:** For accessing, loading, and fine-tuning

pre-trained LLMs and sentence transformer models.
    - **PyTorch:** As the underlying deep learning framework for transformers.
    - **FAISS (Facebook AI Similarity Search):** For creating and querying the efficient vector index for the knowledge base.
    - **LangChain or LlamaIndex:** These frameworks provide high-level abstractions and tools for building RAG pipelines, potentially simplifying the development and integration of the retriever, generator, and knowledge base components. Their use will be evaluated for accelerating prototype development.[45]
    - **Scikit-learn:** For calculating various performance evaluation metrics.
    - **Pandas and NumPy:** For data manipulation and numerical operations.
    - **Streamlit or Flask:** For developing a simple web-based user interface to demonstrate the prototype's functionality.
- **Hardware Resources:**
    - **Fine-tuning:** Access to GPUs is essential for LLM fine-tuning. This will likely involve using services like Google Colab Pro (with its T4 or A100 GPUs), university High-Performance Computing (HPC) clusters, or short-term rentals of cloud GPUs (e.g., NVIDIA H100s as used in SahabatAI training [24], though more modest GPUs like V100/A100 would also be suitable for fine-tuning).
    - **Inference:** For the prototype demonstration, inference might be feasible on a moderate GPU or even a powerful CPU, depending on the final size of the chosen LLM and whether techniques like quantization are explored (though quantization itself adds complexity and is a secondary consideration).
- **Version Control:** Git and GitHub will be used for managing code, tracking changes, and facilitating collaboration (if any).

5.3. Performance Evaluation Metrics

A multi-faceted evaluation approach will be adopted to assess the performance of the proposed system comprehensively, covering the retriever, the generator, the end-to-end RAG system, and conversational quality.

5.3.1. Retriever Evaluation:

These metrics assess the ability of the retriever component to find relevant document chunks from the knowledge base given a user query.

- **Hit Rate:** The percentage of queries for which at least one relevant ("gold") document chunk is successfully retrieved within the top-k results.[55]
- **Mean Reciprocal Rank (MRR):** Measures the average reciprocal of the rank at which the first relevant document is retrieved. Higher MRR indicates relevant documents are found closer to the top.[55]
- **Precision@K:** The proportion of retrieved documents in the top K that are relevant.[55]

- **Recall@K:** The proportion of all relevant documents in the knowledge base that are retrieved in the top K.[55]
- **Normalized Discounted Cumulative Gain (NDCG@K):** A ranked-based metric that evaluates the quality of the retrieved list by considering both the relevance and the position of the documents. It assigns higher scores if more relevant documents are ranked higher.[55]

5.3.2. Generator Evaluation (LLM Output Quality with RAG Context):
These metrics assess the linguistic quality of the answers generated by the LLM when provided with the query and retrieved context, typically by comparing them to human-written reference answers.

- **BLEU (Bilingual Evaluation Understudy):** Measures n-gram precision overlap between generated and reference answers.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Measures n-gram recall, word sequence, and word pair overlap. Variants like ROUGE-L (LCS) are common.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** Considers synonymy and stemming along with precision and recall.
- **Perplexity:** An intrinsic measure of language model fluency; lower perplexity generally indicates better fluency (though less directly correlated with factual accuracy in RAG).

5.3.3. End-to-End RAG System Evaluation (Focus on Factual Accuracy and Relevance):
These are crucial metrics for a consular chatbot, assessing the overall quality and correctness of the final answers.

- **Answer Relevance:** Assesses how well the generated answer addresses the specific user query. This can be evaluated by human annotators on a Likert scale or using model-based evaluation if a suitable evaluation LLM is available.[55]
- **Factual Correctness / QA Correctness / Groundedness / Faithfulness:** This is a critical set of metrics. It determines if the information presented in the generated answer is factually accurate according to the consular domain knowledge and, importantly, if it faithfully reflects the information present in the retrieved context documents. This is often evaluated as a binary classification (Correct/Incorrect, Factual/Hallucinated, Faithful/Unfaithful) by human annotators comparing the answer to the retrieved context and ground truth knowledge.[55]
- **Context Relevance (for RAG):** Evaluates how relevant the retrieved context (that was fed to the generator) actually was to the user's query. This helps diagnose whether issues are with retrieval or generation.[55]
- **Absence of Hallucination:** Specific checks and metrics to identify instances where the LLM generates information not supported by the retrieved context or

known facts. This is a key aspect of factual correctness.[55]

5.3.4. Conversational Quality Metrics (Primarily Qualitative or Human-Evaluated):
These metrics assess the user experience aspects of the conversation.

- **Conversation Relevancy:** Does the chatbot maintain focus on the user's topic throughout the (potentially short) conversation?.[57]
- **Knowledge Retention (if multi-turn interactions are explored):** Does the chatbot remember information provided in earlier turns of the same conversation?.[57]
- **Role Adherence/Persona Consistency:** Does the chatbot maintain an appropriate persona (e.g., official, helpful, empathetic) consistent with a government service?.[57]

5.3.5. Task-Oriented Metrics (Adapted for Informational Tasks):
While traditional Goal Completion Rate (GCR) is often for transactional bots, it can be adapted for informational ones.

- **Informational Goal Completion Rate:** For a defined set of informational queries, what percentage of the time did the user receive the specific information they were seeking in a complete and accurate manner? This will likely require human judgment or carefully designed test cases.[58]
- **Task Completion Rate (Informational):** Defined as the percentage of queries for which the chatbot provides a response that is deemed complete and accurate by human evaluators for the specific informational need expressed in the query.[59]

For a consular chatbot, the emphasis during evaluation must be overwhelmingly on factual accuracy, groundedness, and the absence of hallucination.[55] While linguistic fluency (measured by BLEU/ROUGE) is desirable, providing grammatically perfect but factually incorrect consular advice would be detrimental. Therefore, metrics like QA Correctness, Faithfulness, and specific Hallucination Detection will be prioritized in the evaluation protocol.

Table 5.1: Selected Performance Evaluation Metrics and Their Relevance to Consular Chatbot

Metric Name	Category	Description	How Measured	Relevance to Consular Chatbot
Hit Rate	Retrieval	% queries where ≥1 relevant	Automated (requires	Ensures the system can find

		document is retrieved.	annotated query-doc pairs)	potentially relevant information.
MRR	Retrieval	Average reciprocal rank of the first relevant retrieved document.	Automated (requires annotated query-doc pairs)	Indicates if relevant info is ranked highly by the retriever.
Precision@K (Retrieval)	Retrieval	Proportion of top K retrieved docs that are relevant.	Automated (requires annotated query-doc pairs)	Measures precision of retrieved context.
NDCG@K (Retrieval)	Retrieval	Considers relevance and rank of retrieved docs.	Automated (requires annotated query-doc pairs)	Holistic measure of retrieval ranking quality.
ROUGE-L	Generation Quality	LCS-based overlap between generated and reference answers.	Automated (requires reference answers)	Measures content overlap and fluency.
Answer Relevance	End-to-End	How relevant the final answer is to the query.	Human Evaluation (Likert scale) / Model-graded	Core measure of usefulness.
Factual Correctness / Faithfulness	End-to-End (Critical)	Does the answer accurately reflect info in retrieved context and known facts? Is it free of hallucination?	Human Evaluation (Binary: Correct/Incorrect; Faithful/Unfaithful) [55]	Paramount for providing reliable consular information.

Informational Goal Completion	Task-Oriented	Did the user receive the specific, complete, and accurate information they sought?	Human Evaluation (based on predefined informational tasks/queries)	Measures overall effectiveness for user's informational needs.
Role Adherence	Conversational Quality	Does the bot maintain an official and helpful persona?	Human Evaluation (Likert scale)	Important for user trust and appropriateness of a government service.

5.4. Experimental Setup and Baselines for Comparison

To rigorously assess the performance of the proposed RAG-based system, a set of baselines will be established for comparison:

- **Baseline 1: Simple Retrieval System:** This system will use a standard retrieval method (e.g., BM25 for keyword search, or the same dense retriever as the RAG system) but will directly display the top-k retrieved document chunks to the user without any generative LLM processing. This baseline helps isolate the contribution of the generative component.
- **Baseline 2: Fine-tuned LLM without RAG:** The chosen Indonesian LLM, after domain adaptation/fine-tuning (as described in 5.1), will be used to answer queries directly without access to the external RAG knowledge base. This will demonstrate the impact and necessity of the RAG component for factual grounding and accessing specific consular details.
- **Baseline 3 (Optional, if feasible):** If time and resources permit, implementing the RAG pipeline with an alternative SOTA Indonesian LLM (e.g., if SahabatAI is chosen, compare against NusaBERT in the same RAG setup) would provide insights into the impact of the LLM choice itself.

Experiments will include:

- **Ablation Studies:** To understand the contribution of different components. For example:
  - Evaluating the RAG system with and without the generator LLM fine-tuning to quantify the impact of domain adaptation.
  - Comparing different sentence embedding models for the retriever.
  - Experimenting with different document chunking strategies for the knowledge base and observing their effect on retrieval and end-to-end performance.
- **Performance Comparison:** The proposed RAG system will be benchmarked

against the baselines using the suite of evaluation metrics defined in Section 5.3.

5.5. Benchmarking (Contextual)

The primary benchmark for this thesis will be the custom evaluation dataset developed specifically for Indonesian consular service queries. This dataset, with its curated queries and gold-standard answers/relevant documents, will provide the most direct measure of the system's performance on its intended task.

While the final RAG system itself is domain-specific and not designed for general NLP benchmarks like IndoNLU [33] or NusaX [35], the performance of the underlying Indonesian LLM chosen for the generator component on these standard benchmarks (as reported in existing literature or its model card) will be cited. This helps establish the general linguistic competence and capabilities of the selected LLM in Bahasa Indonesia before its application within the specialized RAG framework.

Chapter 6: Expected Results, Timeline, and Feasibility

6.1. Expected Outcomes and Deliverables

Upon successful completion, this thesis is expected to yield the following outcomes and deliverables:

- **Comprehensive Literature Review:** A thorough review of existing research on conversational AI, its application in public and consular services, SOTA NLP techniques, and relevant LLMs for Bahasa Indonesia.
- **Detailed System Architecture Design:** A complete architectural blueprint for a Retrieval-Augmented Generation (RAG) based conversational AI system specifically designed for handling Indonesian consular service queries. This will include specifications for the retriever, generator, and knowledge base components.
- **Curated Consular Knowledge Base:** A structured and processed knowledge base containing relevant information on Indonesian consular services, derived from official and public sources, and prepared for use within the RAG system.
- **Domain-Adapted Indonesian LLM:** A fine-tuned version of a selected SOTA Indonesian LLM, adapted to the linguistic style, terminology, and common query patterns of the consular domain.
- **Working Prototype System:** A functional prototype of the conversational AI system, capable of understanding queries in Bahasa Indonesia and generating factually grounded responses based on the consular knowledge base. This prototype will include a basic user interface for demonstration.
- **Rigorous Performance Evaluation:** A detailed report on the evaluation of the prototype, including quantitative results for the defined metrics (retrieval accuracy, generation quality, factual correctness, etc.) and qualitative analysis of

its strengths and weaknesses.
- **Final Master's Thesis Document:** The complete written thesis, documenting all aspects of the research, methodology, implementation, evaluation, and conclusions, adhering to academic standards.

6.2. Potential Challenges and Mitigation Strategies

- **Data Scarcity/Quality for Fine-tuning/KB:**
  - *Challenge:* Difficulty in obtaining a sufficient volume of high-quality, structured Q&A data specific to Indonesian consular services for fine-tuning the LLM, or comprehensive, easily parsable documents for the knowledge base. Publicly available information might be scattered, in varied formats, or not always up-to-date.
  - *Mitigation:*
    - Focus on robust and intelligent preprocessing techniques for the documents collected for the KB to extract maximum value.
    - Employ careful, rule-based cleaning and structuring of web-scraped content.
    - If natural Q&A data is scarce, consider synthetic data generation for fine-tuning with extreme caution, ensuring rigorous manual validation of any generated pairs against official sources.
    - Start with a narrower, well-defined subset of consular topics for which good quality data is more readily available, and expand if time permits.
- **Computational Resources:**
  - *Challenge:* LLM fine-tuning and even large-scale RAG experimentation can be computationally intensive, requiring access to GPUs.
  - *Mitigation:*
    - Prioritize the use of pre-trained instruction-tuned models (e.g., SahabatAI Instruct versions) to minimize the extent of fine-tuning required.
    - Leverage available academic resources such as university HPC clusters or cloud computing credits (e.g., Google Colab Pro for limited experimentation).
    - Optimize model sizes for inference if possible (e.g., exploring quantization post-training, though this is a secondary goal due to added complexity).
- **Indonesian Language Nuances:**
  - *Challenge:* Bahasa Indonesia has various informal expressions, colloquialisms, and potential (though out of primary scope) regional influences that might affect query understanding.
  - *Mitigation:*
    - Select LLMs known to be trained on diverse Indonesian corpora (e.g.,

SahabatAI, NusaBERT which explicitly mention broader Indonesian language data [24]).
        - Incorporate varied phrasings and potential informal queries into the test set to assess robustness.
        - The fine-tuning data, even if small, can include examples of common query variations.
- **Time Constraints (6 months):**
    - *Challenge:* The scope of developing a full-fledged conversational AI system is ambitious for a six-month Master's thesis.
    - *Mitigation:*
        - Strict prioritization of core RAG functionality (retrieval and grounded generation).
        - Limit the number of consular service topics covered by the prototype to a manageable set.
        - Focus on one primary Indonesian LLM for the generator, rather than extensive comparative studies of multiple LLMs if time is short.
        - Keep the dialogue management component simplified, focusing on single-turn or very basic multi-turn interactions.
        - Utilize existing frameworks like LangChain or LlamaIndex to accelerate development of the RAG pipeline.
- **Evaluation Subjectivity:**
    - *Challenge:* Human evaluation of aspects like answer relevance, coherence, and factual correctness can be subjective.
    - *Mitigation:*
        - Develop clear, objective, and detailed rubrics for all human evaluation tasks.
        - If possible, involve a second evaluator for a subset of the data to check for inter-annotator agreement, even if the primary evaluation is done by the researcher.
        - Focus human evaluation on the most critical aspects, particularly factual accuracy and relevance.

6.3. Project Timeline (Detailed 6-month plan)
The proposed research will be conducted over a six-month period, with the following indicative timeline:

Table 6.1: Detailed 6-Month Thesis Timeline

Month	Key Tasks & Activities	Deliverables/Milestones

Month 1	- Finalize comprehensive literature review. - Refine and solidify the research methodology and system architecture. - Identify and catalogue primary public data sources for consular information (Kemlu websites, portals). - Draft Chapters 1, 2, and 3 of the thesis.	- Approved Thesis Proposal. - Initial draft of Chapters 1-3.
Month 2	- Commence data collection: web scraping of official sites, downloading relevant documents. - Develop and test data preprocessing scripts (cleaning, format conversion). - Begin structuring the initial knowledge base: initial document chunking strategies. - Design the schema for the fine-tuning and evaluation datasets.	- Collection of raw documents for KB. - Functional data preprocessing scripts. - Initial set of processed text chunks.
Month 3	- Implement and test the retriever module: select sentence embedding model, set up vector database (FAISS), implement retrieval logic. - Evaluate retriever performance on a small, annotated dataset. - Select the primary Indonesian LLM for the generator component. - Begin fine-tuning the selected LLM on the curated consular Q&A dataset (if applicable, or adapt instruct model).	- Working retriever module. - Initial fine-tuned/domain-adapted generator LLM (v1). - Preliminary retriever evaluation report.
Month 4	- Integrate the retriever and generator components into the full RAG pipeline (potentially using	- End-to-end RAG prototype (v1). - Completed evaluation dataset. - Draft of Chapter 4.

	LangChain/LlamaIndex). - Develop a basic user interface (e.g., Streamlit) for prototype demonstration. - Prepare and annotate the full evaluation dataset (queries, gold answers, relevant document IDs). - Draft Chapter 4 (Data Collection and Preparation).	
Month 5	- Conduct comprehensive system evaluation: run automated metrics, perform human evaluation for factual accuracy, relevance, and conversational quality. - Analyze evaluation results, identify system strengths and weaknesses. - Perform iterative refinements on the prototype if time permits (e.g., prompt adjustments, minor KB tweaks). - Draft Chapter 5 (Training, Implementation, and Evaluation).	- Complete set of evaluation results. - Analysis of system performance. - Draft of Chapter 5.
Month 6	- Complete thesis writing: finalize all chapters, incorporate feedback. - Prepare figures, tables, and references. - Conduct final proofreading and formatting. - Thesis submission and preparation for defense.	- Final, submission-ready Master's Thesis document.

6.4. Feasibility Analysis

- **Technical Feasibility:** The technical implementation of the proposed system is considered feasible. State-of-the-art Indonesian LLMs such as SahabatAI and NusaBERT are becoming increasingly accessible, with some models available through platforms like Hugging Face.[24] Open-source frameworks like LangChain and LlamaIndex significantly simplify the development of RAG pipelines.

Pre-trained sentence transformers for Bahasa Indonesia are available from initiatives like Lazarus NLP [28], and efficient vector database solutions like FAISS are well-established. The core technologies are mature enough for a Master's level project.

- **Data Feasibility:** The primary data for the knowledge base will be sourced from publicly available Indonesian government websites and documents related to consular services. While the quantity of this information is substantial, the main challenge, as identified, lies in its extraction, cleaning, and structuring. The feasibility hinges on the ability to effectively process this public data. The creation of smaller, high-quality datasets for fine-tuning and evaluation through manual effort and potentially limited, carefully validated synthetic generation is also deemed achievable within the timeframe.
- **Time Feasibility:** The six-month timeline is undeniably ambitious for a project of this nature. However, it is rendered feasible through careful and strategic scoping. Key measures include:
    - Focusing on informational queries rather than transactional capabilities.
    - Limiting the prototype to a defined subset of consular services.
    - Prioritizing a simplified dialogue management system.
    - Leveraging pre-trained and, crucially, instruction-tuned LLMs to reduce the burden of extensive training from scratch. The detailed month-by-month plan outlined in Section 6.3 provides a structured pathway to completion.
- **Resource Feasibility:** The project assumes access to standard academic computational resources. This includes a reasonably powerful personal computer for development and testing, and access to GPU resources (e.g., via Google Colab Pro, university HPC facilities, or limited cloud credits) for the LLM fine-tuning phases, which are typically short but intensive. Software requirements are met by open-source libraries.
- **Skill Feasibility:** The project demands strong skills in Python programming, Natural Language Processing, and machine learning, particularly with transformer-based models and deep learning frameworks. These are skills generally expected of a Master's degree candidate in Computer Science, Artificial Intelligence, or a related technical field undertaking a research thesis of this nature. The student will also develop expertise in RAG architectures and Indonesian NLP resources.

Chapter 7: Conclusion

This thesis proposal outlines a plan to develop and evaluate an intelligent conversational AI system for handling consular service queries for the Indonesian Ministry of Foreign Affairs. The proposed system, centered around a Retrieval-Augmented Generation (RAG) architecture and a domain-adapted Indonesian Large Language Model, aims to address the challenges of

providing timely, accurate, and accessible consular information in Bahasa Indonesia. The research will involve a comprehensive literature review, meticulous data collection and preparation from official Kemlu sources, the implementation of a RAG pipeline leveraging SOTA Indonesian LLMs like SahabatAI or NusaBERT, and a rigorous evaluation focusing on factual accuracy, relevance, and conversational quality. The project is designed to be technically feasible within a six-month Master's thesis timeframe by strategically scoping the work and leveraging existing open-source tools and pre-trained models.

The expected contributions are significant, offering a potential pathway to enhance the efficiency and effectiveness of Indonesian consular services, thereby improving citizen experience. Furthermore, this research will contribute to the growing body of knowledge on applying advanced AI techniques to specific languages like Bahasa Indonesia and to the unique context of public sector service delivery. By demonstrating a robust and well-evaluated prototype, this work aims to not only fulfill academic requirements but also to provide a valuable blueprint for future AI adoption within Kemlu and potentially other Indonesian government agencies. The successful completion of this thesis will underscore the transformative potential of conversational AI in modernizing government services and fostering better citizen engagement.

References
1

Works cited

1. www.moveworks.com, accessed May 10, 2025, https://www.moveworks.com/us/en/resources/blog/benefits-of-conversational-ai-in-government#:~:text=Within%20the%20public%20sector%2C%20conversational,process%20ambiguity%20and%20friction%20points.
2. Benefits of Conversational AI in Government | Moveworks, accessed May 10, 2025, https://www.moveworks.com/us/en/resources/blog/benefits-of-conversational-ai-in-government
3. MEMBANGUN PERTUMBUHAN YANG KUAT DAN MENGOPTIMALKAN PELUANG BISNIS - IDX, accessed May 10, 2025, https://www.idx.co.id/StaticData/NewsAndAnnouncement/ANNOUNCEMENTSTOCK/From_EREP/202505/0eea9519b0_9162c74a20.pdf
4. Portal Peduli WNI Mobile - Apps on Google Play, accessed May 10, 2025, https://play.google.com/store/apps/details?id=id.go.kemlu.peduliwni.mobile
5. Safe Travel - Apps on Google Play, accessed May 10, 2025, https://play.google.com/store/apps/details?id=id.go.kemlu.safetravel
6. (PDF) THE EFFECT OF DIALOGIC COMMUNICATION IN SAFE TRAVEL

APPLICATION ON DIGITAL DIPLOMACY OF THE MINISTRY OF FOREIGN AFFAIRS OF THE REPUBLIC OF INDONESIA - ResearchGate, accessed May 10, 2025, https://www.researchgate.net/publication/328976205_THE_EFFECT_OF_DIALOGIC_COMMUNICATION_IN_SAFE_TRAVEL_APPLICATION_ON_DIGITAL_DIPLOMACY_OF_THE_MINISTRY_OF_FOREIGN_AFFAIRS_OF_THE_REPUBLIC_OF_INDONESIA
7. Kementerian Luar Negeri akan Gunakan AI untuk Pelayanan WNI di Luar Negeri | tempo.co, accessed May 10, 2025, https://www.tempo.co/internasional/kementerian-luar-negeri-akan-gunakan-ai-untuk-pelayanan-wni-di-luar-negeri-1207028
8. arxiv.org, accessed May 10, 2025, https://arxiv.org/pdf/2401.05415
9. Artificial Intelligence in Diplomacy: Transforming Global Relations and Negotiations, accessed May 10, 2025, https://trendsresearch.org/insight/artificial-intelligence-in-diplomacy-transforming-global-relations-and-negotiations/
10. Indonesia, UN Women Launch AI Chatbot SARI to Protect Female Migrant Workers - Jakarta Daily, accessed May 10, 2025, https://www.jakartadaily.id/international/16214999370/indonesia-un-women-launch-ai-chatbot-sari-to-protect-female-migrant-workers
11. indonesia.iom.int, accessed May 10, 2025, https://indonesia.iom.int/sites/g/files/tmzbdl1491/files/documents/2025-04/key-results-infosheet_mmptf_eng.pdf
12. Indonesia Launches AI-Powered App to Support Migrant Workers - RRI, accessed May 10, 2025, https://rri.co.id/en/national/1465164/indonesia-launches-ai-powered-app-to-support-migrant-workers
13. 10 Best Conversational AI Platforms in 2025 - Botpress, accessed May 10, 2025, https://botpress.com/blog/conversational-ai-platforms
14. Top 10 Conversational AI Platforms for 2025: Detailed Guide + Comparison - Emitrr, accessed May 10, 2025, https://emitrr.com/blog/conversational-ai-platform/
15. AI CoPilots: State of the Art Features for the Federal Government ..., accessed May 10, 2025, https://greystonesgroup.com/ai-copilots-state-of-the-art-features-for-the-federal-government/
16. Privacy Policy - Safe Travel, accessed May 10, 2025, https://bo-safetravel.kemlu.go.id/privacy-policy
17. Indonesia website adopts AI integration to become a smart travel companion - TTG Asia, accessed May 10, 2025, https://www.ttgasia.com/2025/04/01/indonesia-website-adopts-ai-integration-to-become-a-smart-travel-companion/
18. Bertepatan dengan peringatan Hari Kartini (21/4), Kementerian Luar Negeri bekerja sama dengan Badan PBB untuk Kesetaraan Gender dan Pemberdayaan Perempuan (UN Women) menyelenggarakan Dialog Publik “Teknologi Digital dan Wajah Pelindungan Perempuan Pekerja Migran Indonesia”. Kegiatan ini bertujuan untuk mendorong diskusi tentang penggunaan kecerdasan buatan (AI) untuk

memastikan migrasi aman bagi perempuan pekerja migran. - Portal Kemlu, accessed May 10, 2025, https://kemlu.go.id/berita/kementerian-luar-negeri-dan-un-women-memperkuat-pelindungan-perempuan-pekerja-migran-indonesia-melalui-inovasi-chatbot-ai-sari?type=publication
19. Simcse Indobert Base · Models · Dataloop, accessed May 10, 2025, https://dataloop.ai/library/model/lazarusnlp_simcse-indobert-base/
20. indobert-base-p1 - PromptLayer, accessed May 10, 2025, https://www.promptlayer.com/models/indobert-base-p1
21. Indobert QA · Models · Dataloop, accessed May 10, 2025, https://dataloop.ai/library/model/rifky_indobert-qa/
22. Indobert-QA - PromptLayer, accessed May 10, 2025, https://www.promptlayer.com/models/indobert-qa
23. Sahabat-AI, accessed May 10, 2025, https://sahabat-ai.com/
24. Supa-AI/gemma2-9b-cpt-sahabatai-v1-instruct:q5_1 - Ollama, accessed May 10, 2025, https://ollama.com/Supa-AI/gemma2-9b-cpt-sahabatai-v1-instruct:q5_1
25. Indosat Ooredoo Hutchison & GoTo Launch Indonesian LLM, Sahabat-AI - The Fast Mode, accessed May 10, 2025, https://www.thefastmode.com/technology-solutions/38190-indosat-ooredoo-hutchison-goto-launch-indonesian-llm-sahabat-ai
26. Llama3 8b Cpt Sahabatai V1 Instruct · Models - Dataloop AI, accessed May 10, 2025, https://dataloop.ai/library/model/gotocompany_llama3-8b-cpt-sahabatai-v1-instruct/
27. GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct · Hugging Face, accessed May 10, 2025, https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct
28. Welcome to Lazarus NLP! - Lazarus NLP, accessed May 10, 2025, https://lazarusnlp.github.io/
29. LazarusNLP/NusaBERT-base · Hugging Face, accessed May 10, 2025, https://huggingface.co/LazarusNLP/NusaBERT-base
30. LazarusNLP (Lazarus NLP) - Hugging Face, accessed May 10, 2025, https://huggingface.co/LazarusNLP
31. LazarusNLP/IndoT5: T5 Language Models for the ... - GitHub, accessed May 10, 2025, https://github.com/LazarusNLP/IndoT5
32. Indonesia to develop new DeepSeek - Theinvestor, accessed May 10, 2025, https://theinvestor.vn/indonesia-to-develop-new-deepseek-d14607.html
33. IndoNLU Benchmark Dataset | Papers With Code, accessed May 10, 2025, https://paperswithcode.com/dataset/indonlu-benchmark
34. indonlp/indonlu · Datasets at Hugging Face, accessed May 10, 2025, https://huggingface.co/datasets/indonlp/indonlu
35. IndoNLP/nusax: High-quality parallel resource on ... - GitHub, accessed May 10, 2025, https://github.com/IndoNLP/nusax
36. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages - ar5iv, accessed May 10, 2025, https://ar5iv.labs.arxiv.org/html/2205.15960

37. What is a Transformer Model? - IBM, accessed May 10, 2025, https://www.ibm.com/think/topics/transformer-model
38. (PDF) Transformer-based Ranking Approaches for Keyword Queries over Relational Databases - ResearchGate, accessed May 10, 2025, https://www.researchgate.net/publication/390142904_Transformer-based_Ranking_Approaches_for_Keyword_Queries_over_Relational_Databases
39. What is Retrieval-Augmented Generation (RAG)? | Google Cloud, accessed May 10, 2025, https://cloud.google.com/use-cases/retrieval-augmented-generation
40. What is RAG? - Retrieval-Augmented Generation AI Explained - AWS, accessed May 10, 2025, https://aws.amazon.com/what-is/retrieval-augmented-generation/
41. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications, accessed May 10, 2025, https://www.mdpi.com/2076-3417/15/8/4234
42. RAG in AI: Enhancing Accuracy and Context in AI Responses - Acceldata, accessed May 10, 2025, https://www.acceldata.io/blog/how-rag-in-ai-is-transforming-conversational-ai
43. How can LLMs be fine-tuned for specialized domain knowledge ..., accessed May 10, 2025, https://discuss.huggingface.co/t/how-can-llms-be-fine-tuned-for-specialized-domain-knowledge/141989
44. What is AI Chatbot Fine-Tuning? - Beyondspace, accessed May 10, 2025, https://www.beyondspace.studio/qanda/what-is-ai-chatbot-fine-tuning
45. Comparative Analysis of RAG Fine-Tuning and Prompt Engineering ..., accessed May 10, 2025, https://www.scribd.com/document/827113694/Comparative-Analysis-of-RAG-Fine-Tuning-and-Prompt-Engineering-in-Chatbot-Development
46. studenttheses.uu.nl, accessed May 10, 2025, https://studenttheses.uu.nl/bitstream/handle/20.500.12932/48393/Master_thesis_Rowan_Woering_6570941.pdf?sequence=1&isAllowed=y
47. AI adoption for work safety faces challenges in Indonesia - Asia News Network, accessed May 10, 2025, https://asianews.network/ai-adoption-for-work-safety-faces-challenges-in-indonesia/
48. AI and the Challenges of Democracy in Indonesia - Jakarta Globe, accessed May 10, 2025, https://jakartaglobe.id/opinion/ai-and-the-challenges-of-democracy-in-indonesia
49. Chatbots and Data Privacy: Ensuring Compliance in the Age of AI - SmythOS, accessed May 10, 2025, https://smythos.com/ai-agents/chatbots/chatbots-and-data-privacy/
50. How AI Chatbots Improve Local Government Services - Velaro, accessed May 10, 2025, https://velaro.com/blog/ai-chatbots-in-local-government-engaging-citizens-across-cities-and-states
51. Training Dataset for chatbots/Virtual Assistants - Kaggle, accessed May 10, 2025, https://www.kaggle.com/datasets/bitext/training-dataset-for-chatbotsvirtual-assi

stants
52. (PDF) in Chatbot-based Information Service using RASA Open ..., accessed May 10, 2025, https://www.researchgate.net/publication/363181311_in_Chatbot-based_Information_Service_using_RASA_Open-Source_Framework_in_Prambanan_Temple_Tourism_Object
53. How To Create A Custom Knowledge Base Chatbot - CustomGPT.ai, accessed May 10, 2025, https://customgpt.ai/custom-knowledge-base-chatbot/
54. RAG in Customer Support: Enhancing Chatbots and Virtual Assistants, accessed May 10, 2025, https://www.signitysolutions.com/blog/rag-in-customer-support
55. RAG Evaluation Metrics Starter Kit - Arize AI, accessed May 10, 2025, https://arize.com/blog-course/rag-evaluation/
56. Testing Your RAG-Powered AI Chatbot - HatchWorks, accessed May 10, 2025, https://hatchworks.com/blog/gen-ai/testing-rag-ai-chatbot/
57. Top LLM Chatbot Evaluation Metrics: Conversation Testing ..., accessed May 10, 2025, https://www.confident-ai.com/blog/llm-chatbot-evaluation-explained-top-chatbot-evaluation-metrics-and-testing-techniques
58. 10 Key Metrics to Evaluate your AI Chatbot Performance - Inbenta, accessed May 10, 2025, https://www.inbenta.com/articles/10-key-metrics-to-evaluate-your-ai-chatbot-performance/
59. The best chatbot metrics to get a true performance measure - WhosOn, accessed May 10, 2025, https://www.whoson.com/chatbots-ai/the-best-chatbot-metrics-to-get-a-true-performance-measure
60. 12 Essential Chatbot Performance Metrics & KPIs for 2025 - Calabrio, accessed May 10, 2025, https://www.calabrio.com/wfo/contact-center-ai/key-chatbot-performance-metrics/
61. aclanthology.org, accessed May 10, 2025, https://aclanthology.org/2023.findings-acl.868.pdf