Enhancing Consular Service Delivery at the Indonesian Ministry of Foreign Affairs: A Hybrid Approach Integrating Fine-tuned SahabatAI and Retrieval Augmented Generation

1. Crafting a Compelling Thesis Title and Introduction

1.1. Formulating a Precise and Informative Thesis Title

The selection of an appropriate thesis title is paramount, as it serves as the initial point of engagement for the academic community and succinctly encapsulates the research's core focus. For a Master's thesis centered on the development of an advanced consular service chatbot for the Indonesian Ministry of Foreign Affairs (MoFA), the title must clearly articulate several key components: the specific problem domain (consular services within MoFA), the primary technologies employed (SahabatAI and Retrieval Augmented Generation - RAG), and the overarching objective (enhancement of service delivery).
Considering these elements, suitable titles could include:

- "Enhancing Consular Service Delivery at the Indonesian Ministry of Foreign Affairs: A Hybrid Approach Integrating Fine-tuned SahabatAI and Retrieval Augmented Generation"
- "Fine-tuning SahabatAI with Retrieval Augmented Generation for an Advanced Consular Service Chatbot within the Indonesian Ministry of Foreign Affairs: A Proposal"
- "Leveraging SahabatAI and RAG for Next-Generation Consular Assistance: A Case Study Proposal for the Indonesian Ministry of Foreign Affairs"

The novelty of the proposed research lies significantly in the *specific combination* of a localized Large Language Model (LLM), SahabatAI, with a RAG system tailored for a specialized governmental function within the Indonesian context. SahabatAI's inherent design for Indonesian languages and cultural nuances [1] presents a unique advantage. When augmented by RAG's capability to furnish factual, current information directly from MoFA's knowledge repositories [3], this synergy promises a significant advancement in automated consular assistance. The term "hybrid approach" effectively conveys this dual strategy of LLM fine-tuning and dynamic information retrieval. Such a title is anticipated to attract interest from academic circles focusing on applied artificial intelligence in public administration, particularly within emerging economies or concerning low-resource language technologies.

1.2. Structuring the Introduction

The introduction to the thesis proposal must compellingly establish the context, necessity, and potential impact of the research. It should be meticulously structured to guide the reader from a general understanding of the problem to the specific aims of the proposed work.
Problem Statement: The proposal should commence by clearly defining the extant challenges in the delivery of consular services that the proposed chatbot seeks to mitigate. These may include issues such as inconsistent information dissemination across various channels, delays in response times, the significant workload imposed on human consular staff, and difficulties faced by citizens in navigating complex bureaucratic procedures or accessing specific information, especially when abroad. It is pertinent to reference MoFA's existing digital transformation efforts, such as the "Benah Diri" initiative aimed at good governance and improved public service [4], and more recent technological adoptions like the Portal Peduli WNI, Safe Travel app, and the SARI (Sahabat Artifisial Migran Indonesia) chatbot.[5] While these initiatives demonstrate MoFA's commitment to modernization, the proposed SahabatAI-RAG chatbot can address potential limitations by offering a more advanced, contextually aware, and linguistically nuanced conversational interface. SahabatAI, as a national LLM initiative, is strategically positioned to support such enhancements in government services.[1] The problem statement should articulate how, despite current digital tools, there remains a clear opportunity for a sophisticated chatbot that leverages these new AI paradigms to elevate service quality.
Research Questions: Following the problem statement, specific, measurable, achievable, relevant, and time-bound (SMART) research questions must be formulated. These questions will guide the entire research endeavor. Examples include:

1. How can the SahabatAI LLM be effectively fine-tuned to comprehend and generate accurate, contextually appropriate, and empathetic responses to consular service queries in Bahasa Indonesia and relevant local dialects?
2. What is the optimal architecture for a Retrieval Augmented Generation system that integrates with a fine-tuned SahabatAI model to provide factual and current consular information sourced directly from MoFA's official knowledge bases?
3. To what extent does the proposed SahabatAI-RAG consular chatbot improve key performance indicators (e.g., response accuracy, user satisfaction, task completion rates, information accessibility) compared to existing MoFA digital channels or baseline LLM performance?
4. What are the principal ethical considerations (e.g., data privacy, bias, accountability) and practical deployment challenges associated with implementing such an AI-driven chatbot within the operational framework of the Indonesian Ministry of Foreign Affairs?

Research Objectives: The research objectives should directly correspond to the research questions, outlining the tangible outcomes the thesis aims to achieve. For instance:

1. To design and develop a prototype consular service chatbot by fine-tuning the SahabatAI LLM and integrating it with a robust RAG system.
2. To curate and prepare a specialized knowledge base from official MoFA sources for the

RAG system and a tailored dataset for fine-tuning SahabatAI.
3. To establish and apply a comprehensive evaluation framework to assess the performance, usability, and effectiveness of the developed chatbot prototype.
4. To analyze and document the ethical implications and potential challenges related to the deployment of the proposed chatbot in the MoFA context.

Significance and Contribution: This section must articulate the potential impact and value of the research. The contributions may include improving the efficiency and accessibility of MoFA's consular services, enhancing the experience for Indonesian citizens requiring assistance, advancing applied AI research within the Indonesian public sector, contributing to Indonesia's broader digital sovereignty ambitions as envisioned by initiatives like SahabatAI [1], and providing a model for leveraging localized LLMs in specialized government domains. The project's alignment with MoFA's ongoing digital transformation [4] and SahabatAI's national strategic importance [1] underscores its relevance. The significance lies not only in the technical achievement but also in its potential to contribute to public service excellence and bolster national AI capabilities.

Scope and Delimitations: The boundaries of the research must be clearly delineated. This includes specifying the range of consular services the chatbot prototype will aim to cover (e.g., passport and visa information, initial guidance on legal assistance), the particular variant of SahabatAI to be used (e.g., the Llama3 8B parameter model, given the availability of detailed information), the types of MoFA documents to be included in the RAG knowledge base, and the limitations inherent in a Master's level project (e.g., development of a prototype rather than a production-ready system).

Thesis Outline: Finally, a brief overview of the proposed structure of the thesis document should be provided, outlining the content of each chapter.

2. Literature Review: Foundations and Context

A comprehensive literature review is essential to ground the research in existing knowledge, identify theoretical underpinnings, and highlight the specific contributions of the proposed work. This review will cover advancements in LLMs for public services, the technical specifics of SahabatAI, the principles of RAG, and the current landscape of consular services and digital initiatives at MoFA.

2.1. State-of-the-Art in LLMs for Public Services and Chatbots

The application of LLMs and chatbots in government services is a rapidly evolving field. Globally, public sector organizations are exploring these technologies to enhance citizen engagement, automate information provision, and streamline administrative processes.[8] Studies have documented various deployments, ranging from chatbots answering frequently asked questions about public benefits [10] to more complex systems aiding in policy interpretation. These deployments have demonstrated potential benefits such as increased accessibility, 24/7 availability, and reduced burden on human agents.

However, challenges persist, including ensuring the accuracy and reliability of information, managing data privacy and security, mitigating algorithmic bias, and addressing the digital

divide.[10] Lessons learned often point to the need for human oversight, robust evaluation frameworks, and careful consideration of the specific socio-cultural context. For instance, the development of NIPR GPT for the U.S. Department of Defense highlights the trend towards domain-specific LLMs tailored to government terminology and nuances.[8] Within Indonesia and Southeast Asia, while AI adoption is growing, specific academic literature on LLM-powered consular chatbots may be nascent, presenting an opportunity for this thesis to make a significant contribution by addressing this gap.

2.2. Deep Dive into SahabatAI

SahabatAI represents a cornerstone technology for this research, being Indonesia's flagship open-source LLM ecosystem.[1] Its development is a collaborative effort involving prominent Indonesian technology and telecommunication companies like Indosat Ooredoo Hutchison and GoTo, with support from global entities like NVIDIA and AI Singapore.[1]

Architecture and Technical Details: SahabatAI models, including the Llama3 8B CPT Sahabat-AI v1 Instruct model pertinent to this research, are built upon the Llama3 architecture.[12] The Llama3 8B CPT Sahabat-AI v1 Instruct model is a decoder-type model with a context length of 8192 tokens and utilizes the default Llama-3-8B tokenizer.[12] A key characteristic of SahabatAI is its explicit design for the Indonesian linguistic landscape, with pre-training and instruction-tuning encompassing Bahasa Indonesia, Javanese, and Sundanese, with plans for further local language inclusion.[2]

Training Data and Pre-training: The continued pre-training data for the Llama3 8B CPT Sahabat-AI v1 base model comprises approximately 50 billion tokens, with a significant portion (55%) from the SEA-LION Pile - Indonesian dataset.[13] This extensive exposure to Indonesian text is a critical advantage. The instruction-tuned version, Llama3 8B CPT Sahabat-AI v1 Instruct, was further trained on approximately 448,000 Indonesian instruction-completion pairs, augmented by Javanese (96,000 pairs), Sundanese (98,000 pairs), and English (129,000 pairs) data.[12] This data includes synthetic instructions and publicly available, hand-curated instructions, with attention to commercially permissive licenses.[12]

Performance Benchmarks: SahabatAI has demonstrated strong performance on relevant benchmarks. For example, it scored 64.123 on the SEA HELM (BHASA) evaluation, outperforming models like Llama-3.1-8B and sea-lionv3-9B.[2] Its performance on IndoMMLU, which covers various educational levels and subjects in Indonesian contexts, further establishes its capabilities.[12] These benchmarks provide a solid baseline understanding of its linguistic and reasoning abilities in Indonesian.

Open-Source Nature and Ecosystem: SahabatAI is positioned as an open-source ecosystem, with models available on platforms like Hugging Face.[12] This openness encourages collaboration and development of AI applications tailored to Indonesian needs.[1] The vision articulated by its creators, such as "putting the power of AI into the hands of everyone in Indonesia" [1], aligns with national goals of digital sovereignty and reducing the digital divide.[1]

Relevance for MoFA: SahabatAI's Indonesian-centric design, robust performance in local

languages, and the stated aim of enhancing government services make it an exceptionally suitable foundation model for the proposed MoFA consular chatbot.[1] The involvement of established international entities like AI Singapore [12] (known for its work on multilingual models for Southeast Asia, such as SEA-LION referenced in SahabatAI's model card [12]) and NVIDIA [1] (providing advanced platforms like NeMo) lends further credibility to its technical underpinnings. This suggests that SahabatAI is not merely a nascent local effort but a well-supported initiative incorporating global best practices, thereby reducing the technical risk associated with its adoption for a Master's thesis project.

2.3. Understanding Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) is an AI framework that significantly enhances the capabilities of LLMs by enabling them to access and utilize external knowledge sources during response generation.[3] This approach addresses several inherent limitations of standalone LLMs.

Core Principles: RAG operates by first retrieving relevant information from a large corpus or knowledge base in response to a user query. This retrieved information is then provided as context to the LLM, which synthesizes it to generate a comprehensive and grounded answer.[3]

Benefits for Consular Services: For consular services, RAG offers compelling advantages. It can mitigate the risk of LLM "hallucination" (generating plausible but incorrect information) by anchoring responses to official MoFA documents.[3] This is crucial for maintaining accuracy and trust. Furthermore, RAG allows the chatbot to provide up-to-date information, which is vital in a domain where policies, visa requirements, and travel advisories can change frequently. The ability to cite sources from the knowledge base can also enhance transparency and user confidence.

Key Components of a RAG System: A typical RAG system comprises three main components [15]:

1. **Knowledge Base:** A curated collection of documents and data relevant to the domain (e.g., MoFA consular policies, FAQs, service guides).
2. **Retriever:** This component is responsible for searching the knowledge base and finding the most relevant information chunks in response to a user query. This often involves techniques like dense vector retrieval using embedding models and vector databases.
3. **Generator:** This is the LLM itself (in this case, the fine-tuned SahabatAI), which takes the user query and the retrieved context as input to generate the final answer.

Challenges in Specialized Domains: While powerful, implementing RAG in specialized domains like consular services presents challenges. Optimizing the retriever to understand domain-specific jargon and accurately identify the most pertinent information from a vast knowledge base can be complex.[3] Maintaining the currency and quality of the knowledge base requires ongoing effort. A significant challenge is the "Lost in the Middle" problem, where LLMs may overlook relevant information if it is embedded within a lengthy retrieved context, particularly if not positioned at the beginning or end.[18] This necessitates careful consideration of document chunking strategies, context length management, and potentially re-ranking mechanisms to ensure that the most critical information is effectively utilized by

the LLM.

RAG vs. Fine-tuning: The relationship between RAG and fine-tuning is often discussed. While fine-tuning adapts the LLM's internal knowledge and behavior to a specific domain or task, RAG provides external, dynamic knowledge at inference time. A hybrid approach, as proposed in this thesis, aims to leverage the strengths of both: fine-tuning SahabatAI for consular language and conversational style, and using RAG to ensure factual accuracy and access to the latest information.[3]

2.4. Consular Services at MoFA: Current Landscape and Digital Initiatives

Understanding the specific context of consular services at the Indonesian Ministry of Foreign Affairs is crucial for designing an effective chatbot.

Overview of MoFA Consular Services: MoFA provides a wide array of consular services to Indonesian citizens both domestically and abroad, as well as to foreign nationals requiring services related to Indonesia. These services typically include passport issuance and renewal, visa applications, legalization of documents, provision of legal assistance and protection to Indonesian citizens overseas (Perlindungan WNI), and dissemination of travel advisories and public information.[22] The MoFA headquarters in Jakarta and its numerous embassies and consulates worldwide are key service delivery points.[22]

User Needs and Pain Points: Citizens interacting with consular services often face challenges such as navigating complex application processes, understanding detailed requirements for various documents, finding specific information quickly, and, for those abroad, overcoming language barriers or time zone differences. Common queries likely revolve around application procedures, eligibility criteria, processing times, fees, and emergency assistance.

Existing Digital Tools: MoFA has made notable strides in digitalizing its services:

- **Portal Peduli WNI:** This web-based platform is dedicated to Indonesian citizens residing abroad, offering features like Lapor Diri (Self-Reporting), Pelayanan Kekonsuleran (Consular Services), and Pengaduan Kasus (Case Reporting).[5] This portal aims to simplify processes that previously required physical presence at an embassy or consulate.
- **Safe Travel App:** A mobile application designed for Indonesian citizens traveling abroad, providing features such as GPS-based attendance, location tracking for safety, emergency reporting capabilities, lists of nearby Indonesian representative offices, and visa status checks.[5] The app requests permissions for account access, contacts, location, and camera/storage to deliver these functionalities.[25]
- **SARI (Sahabat Artifisial Migran Indonesia) Chatbot:** An AI-driven feature or chatbot, potentially integrated within the Safe Travel app, developed in collaboration with UN Women.[6] SARI is designed with an emphasis on empathy, language detection (including Bahasa Indonesia and local dialects like Javanese and Sundanese [26]), summarizing security information, and providing accessible, non-discriminatory services, particularly

for vulnerable groups like Indonesian migrant workers.[5] It signifies MoFA's early adoption of AI for citizen-facing services.

- Other digital channels include the main MoFA website (kemlu.go.id) [22], various embassy and consulate websites [23], and potentially other specialized applications like the "Kemlu Chat App" mentioned in the Google Play Store listing.[27]

MoFA's Commitment to Digital Transformation: These initiatives are part of MoFA's broader commitment to bureaucratic reform ("Reformasi Birokrasi Kemlu"), which began in 2001 with the "Benah Diri" program aimed at improving governance and public service delivery.[4] The continuous development and enhancement of digital tools indicate a proactive stance towards leveraging technology for better citizen services.

The existence of this diverse digital ecosystem (Portal Peduli WNI, Safe Travel, SARI) provides a rich contextual backdrop. The proposed SahabatAI-RAG chatbot should not be envisioned as an isolated replacement but rather as a sophisticated enhancement or a potential integration point within this ecosystem. The thesis should explore how this new chatbot can complement existing platforms, perhaps by serving as a more intelligent, conversational front-end that unifies information access or by providing deeper, context-aware responses to queries that current tools may not fully address. For instance, it could answer a general query about "self-reporting procedures" by summarizing the process and then seamlessly directing the user to the specific "Lapor Diri" feature within the Portal Peduli WNI. This implies that the system architecture for the new chatbot should ideally consider possibilities for API integration or data sharing with existing MoFA systems, adding a layer of practical complexity but also significantly enhancing its potential value and user experience.

3. Proposed Methodology: System Design and Development

This section outlines the systematic approach to designing and developing the consular service chatbot, integrating a fine-tuned SahabatAI model with a RAG pipeline.

3.1. Overall System Architecture

The proposed system will consist of several interconnected components designed to process user queries and generate accurate, context-aware responses. A high-level diagram should illustrate this architecture. The typical flow will be as follows:

1. **User Interface (UI):** A web-based chat interface where users can input their consular service-related queries in natural language (primarily Bahasa Indonesia).
2. **Query Processing:** The user's query is received by the backend system.
3. **RAG Pipeline - Retrieval:**
    - The query is first processed by the RAG retriever module.
    - This module converts the query into an embedding and searches a specialized vector database (the MoFA Consular Knowledge Base) for relevant document chunks.
    - The top-k most relevant chunks are retrieved.

4. **RAG Pipeline - Augmentation & Generation:**
    - The retrieved document chunks are combined with the original user query to form an augmented prompt.
    - This augmented prompt is then passed to the fine-tuned SahabatAI LLM.
5. **Response Generation:** The fine-tuned SahabatAI model processes the augmented prompt and generates a natural language response.
6. **Response Delivery:** The generated response is sent back to the user via the UI.

The architecture should also consider modules for logging interactions (for evaluation and improvement, with ethical considerations), and potentially an administrative interface for managing the knowledge base.

3.2. Data Collection and Preparation

The success of the proposed system hinges on the quality and comprehensiveness of the data used for both the RAG knowledge base and the fine-tuning of SahabatAI.

3.2.1. Knowledge Base for RAG

The RAG knowledge base will serve as the authoritative source of information for the chatbot.

- **Identification of Sources:** A systematic effort will be undertaken to identify all official and publicly available MoFA documents pertaining to consular services. Key sources include:
    - The main MoFA website (kemlu.go.id) and its various sections.[22]
    - Dedicated consular information portals (e.g., consular.embassyofindonesia.org [24]).
    - Websites of Indonesian embassies and consulates general worldwide, which often contain localized consular information.[23]
    - Publicly accessible FAQs, policy documents, service guides, press releases, and official announcements related to consular affairs.
    - If ethically permissible and with appropriate authorization, anonymized data from internal MoFA resources could be considered, though this is likely outside the scope of a Master's thesis. The diverse range of contact points and directorates within MoFA, such as the Pejabat Pengelola Informasi dan Dokumentasi (PPID), Direktorat Informasi dan Media, Direktorat Pelindungan WNI, Direktorat Konsuler, and Direktorat Fasilitas Diplomatik [22], suggests that consular knowledge may be distributed across various internal units. This implies that information is not likely centralized in a single, easily accessible database. Therefore, collating and unifying this information will be a significant but critical task for building a comprehensive RAG knowledge base. The methodology should incorporate a phase for "knowledge mapping" of MoFA's consular services, potentially involving a thorough analysis of all public-facing materials to understand the provenance and structure of different types of information.
- **Data Acquisition and Formatting:** Strategies for acquiring these documents will include targeted web scraping (for HTML content) and automated extraction from PDF and DOCX files. All collected data will be converted into a clean, consistent text-based

format (e.g., Markdown or plain text) suitable for processing and indexing.

- **Preprocessing:** The raw text data will undergo several preprocessing steps:
    - **Cleaning:** Removal of HTML tags, JavaScript code, irrelevant boilerplate content (headers, footers, navigation menus), and handling of special characters.
    - **Structuring:** Where possible, documents will be segmented into logical sections based on their original structure (e.g., by service type, by policy).
    - **Chunking:** Documents will be divided into smaller, semantically coherent chunks of appropriate size for effective retrieval by the RAG system. This step is crucial for balancing information density and context length for the LLM.

3.2.2. Dataset for Fine-Tuning SahabatAI

A high-quality, domain-specific dataset of prompt-response pairs is essential for effectively fine-tuning SahabatAI to specialize in consular service interactions.

- **Objective:** The primary goal is to create a dataset that trains SahabatAI to:
    - Understand and accurately respond to a wide range of consular queries.
    - Handle variations in user phrasing and intent.
    - Adopt an appropriate conversational style (informative, polite, empathetic).
    - Adhere to MoFA's communication guidelines and avoid providing unauthorized advice.
- **Data Scarcity Challenge:** It is highly probable that a readily available, large-scale, high-quality conversational dataset specifically for Indonesian consular services does not exist. This necessitates a proactive approach to dataset creation.
- **Strategies for Dataset Creation:**
    - **Manual Curation:** A foundational set of question-answer (QA) pairs will be manually developed. These will be based on existing MoFA FAQs (e.g., from the FAQ section mentioned in [22]), common consular scenarios identified through analysis of MoFA websites, and hypothetical user queries.
    - **Synthetic Data Generation:** Given the likely scarcity of real data, synthetic data generation will be a key strategy.[28]
        - A powerful "teacher" LLM (e.g., GPT-4, or potentially the base SahabatAI model if its zero-shot capabilities are sufficient for this task) will be prompted with documents from the RAG knowledge base to generate diverse questions and corresponding answers.
        - Techniques such as "Answer Augmentation" (generating multiple answers for a single question), "Question Rephrasing" (generating multiple phrasings for a single question), and "New Question" generation (creating novel questions based on document content) will be explored.[29] The approach detailed in [28], where a larger model generates training data for a smaller model, is directly relevant.
    - **Data Augmentation for Low-Resource NLP:** Principles from data augmentation techniques for low-resource NLP settings [30] can be adapted. For example, constraints derived from MoFA policy documents (e.g., specific eligibility criteria,

required documents for a service) can be used to guide the generation of realistic and accurate synthetic QA pairs.
    - **Human-in-the-Loop Review:** A subset of the synthetically generated data will undergo rigorous human review and refinement. This step is crucial to ensure the quality, factual accuracy, linguistic correctness, and contextual appropriateness of the training examples, particularly ensuring alignment with MoFA's official stance and tone.
    - The design of the SARI chatbot, which emphasizes empathy and language detection [5], offers valuable guidance for the fine-tuning dataset. Consular services often involve individuals facing stressful or sensitive situations (e.g., loss of passport, need for emergency assistance, legal issues). A purely factual chatbot might be perceived as unhelpful or impersonal in such contexts. Therefore, the fine-tuning dataset must include scenarios that require empathetic responses, reassurance, and clear, supportive guidance. This will train SahabatAI to adopt a tone that is not only informative but also appropriately compassionate, aligning with the ethical considerations of AI in public service.
- **Categorization of QA Pairs:** To facilitate more targeted fine-tuning or evaluation, the created QA pairs may be categorized. Potential categorization schemes include by consular service type (e.g., passport, visa, citizen protection), by query intent (e.g., informational, procedural, problem-solving), or by required response style (e.g., factual, empathetic). This approach is inspired by research suggesting benefits from classifying QA pairs (e.g., 'Factual' vs. 'Conceptual') for fine-tuning.[20]

4. Fine-tuning SahabatAI for Consular Expertise

This section details the plan for adapting the chosen SahabatAI model to the specific requirements of consular service interactions through fine-tuning.

4.1. Selecting SahabatAI Model Variant

The primary model variant for this research will be the **Llama3 8B CPT Sahabat-AI v1 Instruct** model.[12] This choice is based on several factors:

- **Detailed Public Information:** Comprehensive technical details, including architecture, training data, and performance benchmarks, are available for this model.[12]
- **Instruction-Following Capabilities:** As an "Instruct" model, it has been specifically fine-tuned to follow instructions and engage in dialogue, making it a suitable base for a conversational chatbot.
- **Manageable Size:** An 8-billion parameter model is relatively manageable for fine-tuning within the resource constraints typically faced in a Master's thesis project, especially when using parameter-efficient techniques. While a 9-billion parameter variant of SahabatAI has been mentioned [7], the 8B variant is selected for its better documentation and presumed greater accessibility for research purposes.

4.2. Parameter-Efficient Fine-Tuning (PEFT)

Full fine-tuning of LLMs with billions of parameters demands substantial computational resources (multiple high-end GPUs and extensive training time), which is often impractical for individual academic research.[20] Therefore, Parameter-Efficient Fine-Tuning (PEFT) methods will be employed.
Chosen PEFT Method (LoRA/QLoRA):
The proposed PEFT method is QLoRA (Quantized Low-Rank Adaptation).32 QLoRA builds upon LoRA by introducing quantization, which further reduces the memory footprint of the model during fine-tuning.

- **Justification for QLoRA:** QLoRA offers significant advantages in terms of resource efficiency. For instance, fine-tuning a Llama 3 8B model with LoRA might require around 16GB of GPU memory, whereas QLoRA can reduce this to approximately 6GB.[33] This reduction makes fine-tuning feasible on more accessible hardware, including high-end consumer GPUs or cloud-based GPU instances like those available on Google Colab.[34]
- **LoRA/QLoRA Configuration:** The specific configuration for QLoRA will involve:
    - **Target Modules:** Identifying the layers of the SahabatAI model where the low-rank adaptation matrices will be injected. Common choices include attention mechanism layers (query, key, value, output projections) and potentially all linear layers, as explored in studies like.[33]
    - **Rank (r):** The rank of the decomposition matrices, a key hyperparameter determining the number of trainable parameters. A smaller rank means fewer parameters and faster training but might limit the adaptation capability. Typical values are 4, 8, 16, or 32.
    - **Alpha (α):** A scaling factor for LoRA.
    - **Dropout:** Regularization to prevent overfitting of the adapter layers.
    - **Quantization Settings:** For QLoRA, this includes the bit precision (e.g., 4-bit), quantization type (e.g., nf4), and compute data type (e.g., bfloat16 or float16).

4.3. Instruction Tuning Strategy

The fine-tuning process, using the curated consular service dataset, will focus on adapting SahabatAI to excel in the following aspects:

- **Consular Domain Language:** Mastering the specific terminology, jargon, and procedural language used in MoFA consular services.
- **Conversational Style:** Developing a natural, polite, clear, and helpful conversational style in Bahasa Indonesia. Given Indonesia's linguistic diversity and the nature of consular services (which may involve citizens accustomed to regional languages or code-switching), the ability to understand and respond appropriately to such inputs would be beneficial, although the primary focus will be Bahasa Indonesia. The fine-tuning data should include examples reflecting typical user language. (Further exploration of code-switching capabilities could draw on research like [16], if sufficient relevant data can be curated).
- **Task-Specific Responses:** Accurately generating information and guidance for defined consular tasks, such as detailing passport renewal steps, explaining visa application

requirements, or providing initial information on seeking legal aid.

- **Empathetic Communication:** Incorporating empathetic phrasing and tone, particularly for queries related to sensitive situations (e.g., emergencies, distress, protection cases). This aligns with the design goals of MoFA's SARI chatbot.[5]
- **Adherence to Constraints and Safety:** Training the model to strictly avoid providing legal or medical advice, speculative information, or opinions. It should be trained to clearly state when a query is outside its scope or when information is unavailable in its knowledge base, and to direct users to official MoFA channels or human agents when necessary.[35]

A significant advantage in this fine-tuning endeavor is SahabatAI's strong foundation in the Indonesian language. The Llama3 8B CPT Sahabat-AI v1 base model was pre-trained on a substantial corpus of Indonesian text, with the SEA-LION Pile - Indonesian dataset constituting 55% of its ~50 billion token pre-training data.[13] This extensive prior exposure means the model already possesses a robust understanding of Indonesian grammar, vocabulary, and common usage. Consequently, the fine-tuning process can concentrate more on imparting domain-specific knowledge related to consular services and shaping its conversational style, rather than teaching the model basic Indonesian. This could potentially lead to more effective adaptation with a smaller, high-quality fine-tuning dataset compared to what might be required for a generic LLM with less initial exposure to Indonesian.

4.4. Technical Implementation Details

The fine-tuning process will leverage established frameworks and tools from the open-source AI community:

- **Frameworks and Libraries:**
  - **Hugging Face Transformers:** For loading the SahabatAI model and tokenizer.[34]
  - **Hugging Face PEFT (Parameter-Efficient Fine-Tuning):** For implementing QLoRA.[34]
  - **bitsandbytes:** For 4-bit quantization required by QLoRA.[34]
  - **trl (Transformer Reinforcement Learning) library:** Specifically, the SFTTrainer (Supervised Fine-tuning Trainer) for managing the instruction tuning process.[34]
- **Hyperparameter Tuning:** Systematic experimentation will be conducted to determine optimal hyperparameters for fine-tuning, including:
  - Learning rate
  - Batch size
  - Number of training epochs
  - QLoRA-specific parameters (rank, alpha)
- **Computational Resources:** Based on available data, fine-tuning the original SahabatAI Llama3 8B CPT Sahabat-AI v1 Instruct took approximately 4 hours on 8x NVIDIA H100-80GB GPUs.[12] QLoRA significantly reduces these requirements; fine-tuning Llama 3 8B with QLoRA (rank 8) has been reported to need as little as 6GB of GPU memory [33], making it feasible on a single high-end GPU or cloud instances. The exact time will depend on dataset size and chosen hyperparameters.

- **Low-Annotation Budget Strategies:** If the curated fine-tuning dataset proves to be relatively small, strategies for low-annotation budget scenarios will be considered. This might involve techniques such as merging the domain-specific dataset with a larger, general-purpose Indonesian instruction dataset to maintain broader conversational abilities, or applying more intensive data augmentation techniques as discussed in Section 3.2.2. Research such as [36] offers insights into fine-tuning under such constraints.

5. Implementing Retrieval Augmented Generation (RAG)

The RAG component is critical for ensuring the chatbot provides accurate, up-to-date, and verifiable information grounded in official MoFA sources.

5.1. Designing the RAG Pipeline

The RAG pipeline will involve several key stages from query ingestion to context provision for the LLM.

- **Retriever Selection:**
  - The primary retrieval method will be **dense vector retrieval**, which uses semantic similarity between query embeddings and document chunk embeddings.
  - **Embedding Model Choice:** Selecting an appropriate embedding model is crucial. Options include:
    - High-performing multilingual models (e.g., from the MTEB benchmark [18]).
    - Indonesian-specific embedding models, if available and performant.
    - An advanced, though more resource-intensive, option would be to fine-tune an embedding model on MoFA-specific text, similar to the "in-house embedding model" approach described in [37], to better capture the nuances of consular terminology. For a Master's thesis, using a strong pre-trained model is likely more feasible.
  - **Vector Database:** A vector database will be used to store and efficiently query the embeddings of the MoFA document chunks. Suitable open-source options include ChromaDB (as used in [39]), FAISS, or Qdrant. The choice will depend on ease of use, scalability, and integration capabilities.
- **Document Processing:**
  - **Chunking Strategy:** This is a critical parameter influencing retrieval quality.[18] Documents will be segmented into manageable chunks. The optimal chunk size (e.g., number of tokens or sentences) and the degree of overlap between chunks will be determined through experimentation. Overlap helps ensure that semantic context is not lost at chunk boundaries.
  - **Metadata:** Each chunk will be associated with relevant metadata, such as the source document title, original document URL, date of last modification (if available), and type of consular service it pertains to. This metadata can be used

for filtering search results or providing context to the user.

- **Information Retrieval:**
  - **Similarity Metric:** Cosine similarity is a common choice for comparing vector embeddings.
  - **Top-k Retrieval:** The number of document chunks (k) to retrieve for each query will be a tunable parameter. Retrieving too few might miss crucial information, while too many can overwhelm the LLM or lead to the "Lost in the Middle" problem.[18]
  - **Re-ranking (Optional but Recommended):** To improve the relevance of the context provided to the LLM, especially if k is large, a re-ranking step can be implemented. This could involve a simpler model or heuristic to re-order the initial top-k retrieved chunks, placing the most likely relevant ones in positions where the LLM is more likely to attend to them (e.g., at the beginning or end of the context window). This can help mitigate the "Lost in the Middle" issue where LLMs might ignore information buried in long contexts.[18]

5.2. Building and Maintaining the MoFA Consular Knowledge Base

The knowledge base is the heart of the RAG system.

- **Population:** The vector database will be populated by processing all collected and preprocessed MoFA consular documents. This involves generating embeddings for each text chunk and storing them along with their metadata.
- **Maintenance and Updates:** Consular information (policies, procedures, fees, advisories) is dynamic. A strategy for keeping the knowledge base current is essential for the chatbot's long-term reliability. For a prototype, this might involve manual updates and re-indexing of modified documents. For a production system, automated pipelines for monitoring changes in source documents and updating the vector database would be necessary. This aspect is critical for maintaining the accuracy and trustworthiness of the chatbot.

The SARI chatbot's function of summarizing "security information, dangers, and crime threats" [5] implies that a truly comprehensive consular chatbot might need to access dynamic, frequently updated information sources. These could include travel advisories (MoFA would issue its own, analogous to the US State Department's advisories mentioned in [40]), emergency contact lists, or real-time alerts. While static policy documents form the core of the knowledge base, integrating such dynamic data feeds (e.g., via APIs or regular scraping of specific MoFA pages) would significantly enhance the chatbot's utility, particularly in citizen safety contexts, mirroring features of the Safe Travel app.[25] For a Master's thesis, the primary focus may be on semi-static documents due to complexity, but the proposal should acknowledge the importance of dynamic data and suggest its integration as a key area for future development.

5.3. Prompt Engineering for RAG

The way the LLM is prompted to use the retrieved context significantly impacts the quality of

the generated response. Effective prompt engineering is crucial.[35]

- **Instruction Design:** Prompts will be carefully crafted to instruct the fine-tuned SahabatAI model on how to utilize the provided document snippets. Key instructions will include:
    - **Grounding:** Explicitly stating that the answer must be based *solely* on the information contained within the retrieved documents.
    - **Handling Insufficient Context:** Providing clear instructions on what the model should output if the retrieved documents do not contain the answer to the query (e.g., "Based on the information available, I cannot answer that question. Please refer to the official MoFA website or contact a consular officer for assistance," adapting the suggestion from [42]).
    - **Source Attribution (Optional):** Instructing the model to cite the source document(s) or section(s) from which the information was derived. This enhances transparency and allows users to verify the information.
    - **Explaining Context:** Briefly explaining to the LLM the nature of the provided documents (e.g., "The following are excerpts from official MoFA consular service manuals...") can help it better contextualize the information.[42]
    - **Conciseness and Relevance:** Guiding the model to synthesize information from multiple chunks if necessary, and to provide a concise and relevant answer to the specific user query.

5.4. Frameworks

To streamline the development of the RAG pipeline, established open-source frameworks will be considered:

- **LangChain:** A popular framework for developing applications powered by language models, with extensive support for RAG components, including document loaders, text splitters, embedding model integrations, vector store integrations, and retrieval chains.[15]
- **LlamaIndex:** Another widely used framework specifically designed for building context-augmented LLM applications, offering similar capabilities to LangChain with a focus on data indexing and retrieval.

The choice of framework will depend on factors such as ease of integration with SahabatAI, flexibility in customizing components, and community support.

6. Evaluation Framework: Measuring Success

A robust evaluation framework is essential to assess the performance, effectiveness, and usability of the developed SahabatAI-RAG consular chatbot. This framework will encompass metrics for the RAG pipeline, the fine-tuned LLM, and user-centric aspects. The evaluation should be an iterative process, with findings from initial evaluations informing further refinements.[9]

6.1. Metrics for RAG Performance

Evaluating the RAG component focuses on the quality of retrieved context and how well the

LLM grounds its responses in that context.

- **Retrieval Quality:**
    - **Context Precision@k:** Measures the proportion of the top-k retrieved document chunks that are actually relevant to the user's query. A higher precision indicates less noise in the retrieved context.[18]
    - **Context Recall:** Measures the proportion of all relevant document chunks in the entire knowledge base that were successfully retrieved among the top-k. This is harder to measure without exhaustive ground truth but can be estimated on a curated test set.[18]
    - **Context Relevance:** A qualitative or LLM-as-judge assessment of how well the content of the retrieved chunks pertains to the user's query. This can involve human annotators rating relevance on a scale or using another LLM to score relevance.[18]
    - **Mean Reciprocal Rank (MRR):** Assesses the rank of the first relevant document retrieved. Higher MRR indicates that relevant information is found quickly.[18]
- **Generation Quality (Groundedness & Faithfulness):**
    - **Faithfulness / Factual Consistency:** This critical metric evaluates whether the chatbot's generated answer accurately reflects the information present in the retrieved context and avoids contradicting it or introducing external, unverified information.[18] This can be assessed through human evaluation or by using an LLM-as-judge approach, where another LLM compares the generated answer against the provided context.
    - **Answer Relevance:** Determines if the generated answer, given the retrieved context, directly and comprehensively addresses the user's specific question.[18]
- **Tools for RAG Evaluation:** Open-source libraries like **Ragas** [18], which provides a suite of metrics for faithfulness, answer relevance, context precision, and context recall, will be considered. Other tools like Arize Phoenix or Quotient AI offer more comprehensive platforms for RAG evaluation and monitoring.[18]

6.2. Metrics for Fine-tuned SahabatAI (with RAG context)

This evaluates the overall quality of the chatbot's final responses, which are generated by the fine-tuned SahabatAI model using the context from the RAG pipeline.

- **Task-Specific Accuracy:** For a predefined set of consular tasks (e.g., "What are the requirements for renewing an Indonesian passport?", "How do I report a lost document abroad?"), the percentage of queries for which the chatbot provides a factually correct and complete answer will be measured. This requires a curated test set with ground-truth answers.
- **Response Quality (Human and Automated Evaluation):**
    - **Fluency & Coherence:** Assesses whether the chatbot's responses are grammatically correct, well-structured, and easy to understand in Bahasa Indonesia.
    - **Helpfulness:** Evaluates whether the answer provides the information the user

was seeking and aids them in their task or understanding.[44]
    - **Conciseness:** Measures whether the answer is to the point and avoids unnecessary verbosity or irrelevant details.
    - **Empathy:** For queries involving sensitive situations, human evaluators will assess the appropriateness and perceived empathy of the chatbot's tone and language.
- **Automated Metrics (for reference):** While often limited for generative tasks, metrics like BLEU, ROUGE, and METEOR can be used to compare generated responses against reference answers, particularly those from the fine-tuning dataset or a manually created gold-standard test set.[18] These provide a quantitative measure of lexical overlap.
- **Reduction in Hallucinations:** Instances where the chatbot generates factually incorrect or fabricated information, especially when no relevant context is retrieved or when the query is out-of-domain, will be tracked and measured.[44]

6.3. User-Centric Evaluation

Understanding the user experience is paramount for a public-facing chatbot.

- **Usability Testing:** A small group of representative users (e.g., students, MoFA staff role-playing as citizens) will interact with the chatbot prototype to perform predefined tasks. Observations of their interactions, think-aloud protocols, and post-task interviews will provide qualitative insights into ease of use, clarity, and any pain points.
- **Satisfaction Surveys:** After interacting with the chatbot, users will complete satisfaction surveys incorporating Likert scales and open-ended questions to assess aspects like perceived usefulness, ease of use, trust in the information provided, satisfaction with the interaction, and overall helpfulness.[46]
- **Task Completion Rate:** The percentage of users who successfully achieve their intended goal (e.g., finding specific information, understanding a procedure) using the chatbot will be measured for a set of common consular tasks.
- **A/B Testing (if feasible):** If resources permit, the performance of the SahabatAI-RAG chatbot could be compared against a baseline system. This baseline could be the existing SARI chatbot, a generic LLM (like GPT-3.5) without domain-specific fine-tuning or RAG, or even the experience of finding information through MoFA's traditional website search.

6.4. Benchmarking

To objectively assess the improvements achieved through fine-tuning and RAG integration, a comparative benchmarking approach will be adopted.

- **Curated Test Set:** A comprehensive test set of diverse consular queries will be developed. This set should cover various service types, query complexities (simple factual questions to more nuanced procedural inquiries), and potential edge cases.
- **Comparative Analysis:** The performance of the fully developed fine-tuned SahabatAI+RAG system will be compared against:
    1. The base SahabatAI model (e.g., Llama3 8B CPT Sahabat-AI v1 Instruct) without

any additional fine-tuning or RAG.
2. The base SahabatAI model augmented with the RAG pipeline but without domain-specific fine-tuning.
3. The domain-specifically fine-tuned SahabatAI model without the RAG pipeline (relying only on its internal knowledge).
4. If feasible and resources allow, comparison with other readily available Indonesian LLMs or prominent general-purpose LLMs (e.g., GPT-3.5/4 via API, with appropriate cost considerations) on the same test set.

6.5. Evaluation Protocol

A clear and detailed evaluation protocol will be established, specifying:

- The methodology for creating test datasets (for RAG and end-to-end chatbot performance).
- Guidelines for human evaluators, including scoring rubrics and training, if human annotation is used for metrics like faithfulness, relevance, or empathy.[46]
- Statistical methods for analyzing quantitative data and determining significance (e.g., t-tests, ANOVA for comparing mean scores from satisfaction surveys or accuracy rates across different system versions).[46]

The evaluation process should not be viewed as a single, final step but as an integral part of an iterative development cycle. Public sector AI initiatives, particularly citizen-facing applications like chatbots, benefit greatly from continuous improvement based on feedback and performance monitoring.[9] While a Master's thesis will typically culminate in a prototype and its initial evaluation, the methodology should also propose a framework for ongoing evaluation and refinement. This includes mechanisms for systematically collecting user feedback, identifying common failure modes or areas where the chatbot underperforms, and using this data to iteratively improve the fine-tuning dataset, the RAG knowledge base content or structure, or the prompt engineering strategies. This demonstrates a mature understanding of AI system development lifecycles.

The following table summarizes the key evaluation metrics proposed:

Metric Category	Specific Metric	Definition	How to Measure	Relevance to MoFA Chatbot
RAG Retrieval	Context Precision@k	Proportion of top-k retrieved documents relevant to the query.	Manual annotation on test set; Automated tools (e.g., Ragas).	Ensures LLM receives high-quality, relevant information, reducing noise.
	Context Recall	Proportion of all relevant documents retrieved.	Manual annotation on test set (challenging for large KBs); Ragas.	Ensures comprehensive information is available to the LLM for

				answering.
	Context Relevance	Degree to which retrieved context pertains to the query.	Human evaluation (Likert scale); LLM-as-judge.	Filters out irrelevant context, improving LLM focus.
	Mean Reciprocal Rank (MRR)	Average of the reciprocal ranks of the first relevant document.	Automated calculation on test set.	Indicates how quickly relevant information is found.
RAG Generation	Faithfulness / Factual Consistency	Extent to which the generated answer accurately reflects the retrieved context without contradiction.	Human evaluation; LLM-as-judge (e.g., Ragas, DeepEval).	Critical for trust; ensures answers are grounded in official MoFA sources.
	Answer Relevance	Extent to which the answer directly addresses the user's query, given the context.	Human evaluation; LLM-as-judge (e.g., Ragas).	Ensures the chatbot provides useful and pertinent responses.
LLM Response Quality	Task-Specific Accuracy	Percentage of correctly answered queries for defined consular tasks.	Evaluation against ground-truth answers on a curated test set.	Measures core competency in providing correct consular information.
	Fluency & Coherence	Grammatical correctness, naturalness, and logical flow of the response.	Human evaluation (Likert scale).	Ensures responses are understandable and professional.
	Helpfulness	Degree to which the answer provides the information the user needs.	Human evaluation (Likert scale).	Measures the practical value of the chatbot's responses.
	Empathy (for relevant query types)	Appropriateness and sincerity of empathetic tone in sensitive situations.	Human evaluation (Likert scale).	Important for user experience in potentially stressful consular interactions.
	Reduction in	Frequency of	Manual review of	Key indicator of

	Hallucinations	generating fabricated or factually incorrect information.	responses, especially for OOD queries or failed retrievals.	reliability and trustworthiness.
User-Centric	User Satisfaction	Overall user contentment with the chatbot interaction.	Post-interaction surveys (Likert scales, open-ended questions).	Direct measure of user acceptance and perceived quality.
	Task Completion Rate	Percentage of users successfully achieving their goals using the chatbot.	Observation during usability tests; Self-reported in surveys.	Indicates the chatbot's effectiveness in helping users.
	Perceived Ease of Use	How effortless users find interacting with the chatbot.	Usability testing; Questionnaires (e.g., System Usability Scale - SUS).	Impacts user adoption and willingness to use the chatbot.

7. Ethical Considerations and Responsible AI Deployment

The deployment of an AI-powered chatbot for consular services, while offering significant benefits, also necessitates careful consideration of ethical implications and adherence to responsible AI principles. This is particularly important given the sensitive nature of consular interactions and the potential impact on individuals.

7.1. Data Privacy and Security

The chatbot may handle Personally Identifiable Information (PII) if users include personal details in their queries related to specific cases or circumstances.

- **Protection Measures:** Robust measures must be in place to protect this data. This includes secure data transmission (HTTPS), encryption of any stored interaction logs, and access controls. The system design should prioritize data minimization, collecting only necessary information.
- **Compliance with MoFA Policies:** The system must comply with MoFA's existing data privacy policies, such as those outlined for the Safe Travel app [25], and any national data protection regulations in Indonesia.
- **Anonymization/Pseudonymization:** If user interaction data is logged for the purpose of system improvement or evaluation, techniques for anonymizing or pseudonymizing this data should be implemented to protect user identities.

- **Transparency in Data Use:** Users should be informed about how their data will be used, stored, and protected, typically through a clear and accessible privacy notice. Federal agencies in other contexts are advised to maintain strict control over data transmission and storage, especially PII.[8]

7.2. Bias Mitigation

Algorithmic bias can arise from various sources and lead to unfair or discriminatory outcomes.

- **Data Bias:** Biases may be present in SahabatAI's original pre-training data or in the specific dataset curated for fine-tuning. Similarly, the official MoFA documents used to build the RAG knowledge base could inadvertently reflect historical biases.
- **Auditing and Mitigation Strategies:**
    - Careful curation of the fine-tuning dataset to ensure diverse representation and avoid perpetuating stereotypes.
    - Regular audits of the RAG knowledge base to identify and address biased language or outdated information.
    - Testing the chatbot's responses across a diverse range of user personas and scenarios to detect biased outputs.[44]
    - Implementing fairness metrics during evaluation.[44]
    - The concern about AI systems making errors or exhibiting biases that lead to unjust profiling or infringement on rights is a significant ethical challenge.[11]

7.3. Transparency and Explainability

Users should have a degree of understanding about how the chatbot operates and the basis for its responses.

- **AI Identification:** The chatbot interface should clearly indicate that the user is interacting with an AI system, not a human.
- **Source Citation (RAG):** A key benefit of RAG is the ability to trace information back to its source. The chatbot should, where feasible, cite the specific MoFA document(s) or sections from which its answer is derived. This allows users to verify the information and builds trust.
- **Avoiding "Black Box" Issues:** While full explainability of LLM decision-making is an ongoing research challenge, providing source attribution is a practical step towards transparency, addressing concerns about the opacity of AI systems.[11]

7.4. Human Oversight and Escalation

The AI chatbot should not be the sole and final point of contact for all consular matters, especially critical or complex ones.

- **Clear Escalation Paths:** The system must provide a clear and easy way for users to escalate their query to a human MoFA staff member if the chatbot cannot resolve their issue, if the situation is too complex, or if the user prefers human interaction.[10]
- **Defining Scope:** The chatbot's operational scope should be clearly defined, and it should be programmed to recognize queries that fall outside this scope or require

human judgment, proactively offering escalation.

7.5. Accountability

Mechanisms for accountability are crucial in case of errors or negative impacts.

- **Responsibility for Information:** While the chatbot generates responses, MoFA ultimately remains accountable for the accuracy and appropriateness of the information provided through its official channels. The lack of transparency in AI decision-making can obscure accountability.[11]
- **Error Reporting and Correction:** A system for users to report errors or problematic responses from the chatbot should be implemented. This feedback is vital for continuous improvement and for correcting inaccuracies in the knowledge base or model behavior.

7.6. Accessibility

The chatbot interface and its content should be designed to be accessible to all users, including those with disabilities, adhering to relevant web accessibility standards (e.g., WCAG).

7.7. Alignment with Ethical AI Frameworks

The development and deployment should align with established ethical AI principles and frameworks.

- **International and National Guidelines:** Consideration should be given to guidelines such as UNESCO's Recommendation on the Ethics of Artificial Intelligence (RAM AI) [11], which emphasizes fairness, transparency, accountability, and human rights. Any national AI ethics guidelines developed in Indonesia should also be consulted.
- **Building on MoFA's Ethical Precedents:** The Indonesian Ministry of Foreign Affairs has prior experience with ethical AI considerations through the SARI chatbot initiative. SARI was developed in collaboration with UN Women and specifically designed to provide accessible, unbiased, and non-discriminatory information and support to vulnerable groups, particularly female migrant workers, with an emphasis on empathy and a human rights perspective.[6] This demonstrates an existing institutional awareness and commitment to gender-responsive and ethically sound AI within MoFA. The proposed consular chatbot should leverage and extend these established principles, ensuring that its design and operation are consistent with MoFA's values and past practices in responsible AI deployment. This includes robust measures for data privacy, proactive bias prevention, and ensuring the technology serves all citizens equitably and respectfully.

8. Thesis Proposal Structure and Timeline

This section provides guidance on structuring the formal thesis proposal document and developing a realistic timeline for the research project.

8.1. Guidance on Structuring the Thesis Proposal Document

A Master's thesis proposal typically follows a standard academic structure. The following sections are recommended, with an emphasis on the depth and focus appropriate for this level of research:

1. **Title Page:** Including the proposed thesis title, student's name, supervisor(s), department, and university.
2. **Abstract (or Executive Summary):** A concise summary (typically 250-300 words) of the entire proposal, covering the problem, objectives, methodology, expected outcomes, and significance.
3. **Chapter 1: Introduction:**
    - Background and Motivation
    - Problem Statement
    - Research Questions
    - Research Objectives
    - Significance and Contribution of the Research
    - Scope and Delimitations of the Study
    - Thesis Outline
4. **Chapter 2: Literature Review:**
    - State-of-the-Art in LLMs for Public Services and Chatbots
    - Detailed Review of SahabatAI (Architecture, Training, Performance)
    - Principles and Applications of Retrieval Augmented Generation (RAG)
    - Overview of MoFA Consular Services and Existing Digital Initiatives
    - Identification of Research Gaps
5. **Chapter 3: Proposed Methodology:**
    - Overall System Architecture (with diagram)
    - Data Collection and Preparation:
        - Knowledge Base for RAG (sources, acquisition, preprocessing)
        - Dataset for Fine-tuning SahabatAI (creation strategies, synthetic data, human review)
    - Fine-tuning SahabatAI:
        - Model Variant Selection
        - PEFT Method (QLoRA justification and configuration)
        - Instruction Tuning Strategy
        - Technical Implementation Details (frameworks, tools)
    - Implementing RAG:
        - Pipeline Design (retriever, document processing, embedding model, vector DB)
        - Knowledge Base Construction and Maintenance Strategy
        - Prompt Engineering for RAG
        - Technical Implementation Details (frameworks, tools)
6. **Chapter 4: Evaluation Plan:**
    - Metrics for RAG Performance (retrieval quality, generation quality)

    - Metrics for Fine-tuned SahabatAI (task accuracy, response quality)
    - User-Centric Evaluation Plan (usability testing, satisfaction surveys)
    - Benchmarking Strategy
    - Overall Evaluation Protocol
7. **Chapter 5: Ethical Considerations:**
    - Data Privacy and Security
    - Bias Mitigation
    - Transparency and Explainability
    - Human Oversight and Escalation
    - Accountability
    - Accessibility
8. **Chapter 6: Timeline and Project Management:**
    - Detailed Research Timeline (Gantt chart recommended)
    - Key Milestones and Deliverables
    - Risk Assessment and Mitigation Strategies
9. **Chapter 7: Expected Outcomes and Contributions:**
    - Reiteration of the specific deliverables (e.g., prototype chatbot, datasets, evaluation report).
    - Summary of anticipated contributions to knowledge and practice.
10. **Resources Required:**
    - Computational resources (GPU access, software)
    - Data access (MoFA documents, potential API access)
    - Supervisory support
11. **Bibliography/References:** A comprehensive list of all cited academic papers, technical reports, and official documents.

8.2. Developing a Realistic Research Timeline

A detailed and realistic timeline is crucial for managing a multi-faceted Master's thesis project. The project can be broken down into the following phases, with estimated durations (assuming a typical 9-12 month active research period for a Master's thesis, adjustable based on specific program requirements):

- **Phase 1: Foundational Work and Literature Review (Months 1-2)**
    - In-depth literature review on LLMs, RAG, SahabatAI, consular services.
    - Refinement of research questions and objectives.
    - Initial exploration of MoFA public data sources.
    - **Deliverable:** Detailed literature review chapter draft; finalized research questions.
- **Phase 2: Data Collection and Preparation (Months 2-4)**
    - Systematic collection of MoFA documents for the RAG knowledge base.
    - Preprocessing and cleaning of RAG source documents.
    - Development and curation of the initial fine-tuning dataset for SahabatAI (manual creation and initial synthetic data generation).
    - Human review and refinement of the fine-tuning dataset.
    - **Deliverable:** Processed RAG knowledge base (initial version); curated fine-tuning

dataset (v1).

- **Phase 3: System Development - Fine-tuning SahabatAI (Months 4-6)**
  - Setup of the fine-tuning environment (software, libraries, GPU access).
  - Implementation of QLoRA fine-tuning scripts.
  - Conducting initial fine-tuning runs with SahabatAI Llama3 8B.
  - Hyperparameter tuning and iterative refinement of the fine-tuned model.
  - Initial qualitative assessment of the fine-tuned model.
  - **Deliverable:** Fine-tuned SahabatAI model (v1); fine-tuning scripts and logs.
- **Phase 4: System Development - RAG Implementation (Months 5-7)**
  - Selection and setup of vector database and embedding model.
  - Chunking and embedding of the RAG knowledge base documents.
  - Development of the RAG retrieval pipeline (using LangChain or LlamaIndex).
  - Integration of the RAG pipeline with the fine-tuned SahabatAI model.
  - Development of prompt engineering strategies for RAG.
  - **Deliverable:** Integrated SahabatAI-RAG chatbot prototype (v1).
- **Phase 5: Evaluation (Months 7-9)**
  - Development of evaluation datasets and protocols (for RAG metrics, LLM metrics, user studies).
  - Conducting automated and human-based evaluations of the RAG pipeline.
  - Conducting evaluations of the end-to-end chatbot (task accuracy, response quality).
  - Planning and conducting user-centric evaluations (usability tests, satisfaction surveys).
  - Data analysis and interpretation of evaluation results.
  - **Deliverable:** Evaluation report; refined chatbot prototype based on initial findings.
- **Phase 6: Thesis Writing and Submission (Months 9-12)**
  - Drafting all chapters of the thesis.
  - Incorporating feedback from supervisor(s).
  - Revisions and finalization of the thesis document.
  - Preparation for thesis defense.
  - **Deliverable:** Completed Master's thesis document.

Milestones:

- End of Month 2: Literature review complete.
- End of Month 4: Data collection and preparation complete.
- End of Month 7: Initial SahabatAI-RAG prototype developed.
- End of Month 9: Evaluation phase substantially completed, results analyzed.
- End of Month 12 (or per program deadline): Thesis submission.

Given the complexity involving LLM fine-tuning, RAG implementation, substantial data sourcing from potentially disparate government sources, and a multi-faceted evaluation, this timeline is ambitious. It is crucial to acknowledge potential delays, such as difficulties in parsing MoFA documents, unexpected challenges in the fine-tuning process, or slower-than-anticipated RAG retrieval performance. To manage these risks, a "Minimum

Viable Product" (MVP) scope for the prototype should be defined early on. This MVP could focus on a core set of 2-3 high-frequency consular services (e.g., passport renewal information, visa inquiry for a specific country, emergency contact information). Other services or more advanced features could be designated as stretch goals or avenues for future work. This approach ensures that a tangible and evaluable outcome can be achieved within the typical Master's thesis timeframe.

9. Conclusion and Future Directions

This research plan outlines a comprehensive approach for a Master's thesis focused on enhancing consular service delivery at the Indonesian Ministry of Foreign Affairs through the novel integration of a fine-tuned SahabatAI LLM and a Retrieval Augmented Generation system.

9.1. Summarizing the Proposed Research Plan

The core objective is to develop and evaluate a prototype chatbot capable of understanding and responding to consular queries in Bahasa Indonesia with accuracy, contextual relevance, and appropriate empathy. This will be achieved by:

1. Leveraging **SahabatAI**, Indonesia's national LLM, specifically the Llama3 8B Instruct variant, due to its strong foundation in Indonesian languages and instruction-following capabilities.
2. Employing **Parameter-Efficient Fine-Tuning (PEFT)**, likely QLoRA, to adapt SahabatAI to the specific linguistic nuances, conversational style, and knowledge domain of consular services using a carefully curated dataset (potentially including synthetically generated data).
3. Implementing a **Retrieval Augmented Generation (RAG)** pipeline to ground the chatbot's responses in official MoFA documents, thereby enhancing factual accuracy and providing access to up-to-date information.
4. Establishing a multi-faceted **evaluation framework** to assess RAG performance, the quality of the fine-tuned LLM's responses, and the overall user experience.
5. Addressing key **ethical considerations** related to data privacy, bias, transparency, and accountability.

The research aims to contribute to improved public service delivery, advance applied AI research in the Indonesian context, and provide a practical case study for leveraging localized LLMs in specialized government functions.

9.2. Highlighting Potential Challenges and Limitations

Several challenges and limitations are anticipated:

- **Data Accessibility and Quality:** Obtaining comprehensive, clean, and consistently formatted official MoFA documents for the RAG knowledge base may be challenging. Similarly, creating a high-quality, diverse fine-tuning dataset for a specialized domain like consular services with limited existing resources will require significant effort, including careful synthetic data generation and human review.

- **Computational Resources:** While QLoRA significantly reduces resource demands, fine-tuning and experimenting with LLMs still require access to adequate GPU resources, which might be a constraint for an individual Master's student.
- **Evolving Nature of LLMs:** SahabatAI is described as an evolving ecosystem.[1] The specific model versions, their capabilities, and access mechanisms might change during the course of the research, requiring adaptability.
- **Scope of a Master's Project:** The outcome will be a prototype, not a production-ready, fully scalable system. Certain complexities of real-world deployment (e.g., integration with MoFA's internal IT infrastructure, handling massive concurrent user loads) will be beyond the scope.
- **Evaluation Nuances:** Quantitatively evaluating nuanced aspects of conversation, such as empathy or the ability to handle very complex, multi-turn dialogues, remains a challenge in LLM research and will rely heavily on qualitative human assessment.
- **Low-Resource Language Challenges:** While SahabatAI supports Bahasa Indonesia, Javanese, and Sundanese [2], if the chatbot is intended to handle queries in other Indonesian local dialects extensively, the "low-resource" nature of these languages in terms of available training data could pose significant challenges.[49] The primary focus will remain on Bahasa Indonesia.

9.3. Suggesting Avenues for Future Research

This Master's thesis can lay the groundwork for several promising future research directions:

- **Expansion of Services and Scope:** Extending the chatbot's capabilities to cover a wider range of consular services or even other functions within MoFA or different Indonesian government agencies.
- **Deeper Integration with MoFA Systems:** Exploring secure API-based integration with MoFA's internal databases and existing digital platforms (e.g., Portal Peduli WNI, Safe Travel) to provide more personalized and transactional services (e.g., checking application status directly, submitting forms via the chatbot).
- **Longitudinal User Studies:** Conducting longer-term studies on user adoption, impact on consular staff workload, and overall effectiveness of the chatbot in a real or simulated operational environment.
- **Advanced RAG and Fine-tuning Techniques:** Investigating more sophisticated RAG techniques such as dynamic RAG (where retrieval is triggered iteratively during generation [19]), self-correcting RAG, or advanced fine-tuning methods like Reinforcement Learning from Human Feedback (RLHF) to further enhance performance and alignment.
- **Multimodal Capabilities:** Exploring the integration of multimodal capabilities, allowing the chatbot to process or generate information in formats beyond text (e.g., understanding queries related to uploaded documents for verification, providing visual aids).
- **Continuous Learning and Adaptation:** Developing mechanisms for the chatbot to learn and adapt from new information and user interactions over time, while ensuring

robust ethical safeguards and human oversight to prevent performance degradation or the learning of undesirable behaviors.

- **Tailored Deployment for Overseas Missions:** MoFA's bureaucratic reform initiative ("Reformasi Birokrasi Kemlu") aims to improve services at both the central ministry ("Pusat") and Indonesian representative offices ("Perwakilan RI") abroad.[4] Consular needs and frequently asked questions can vary significantly depending on the specific country or region due to local laws, cultural contexts, and the demographic profile of the Indonesian diaspora. Future research could investigate methods for efficiently adapting or specializing the core SahabatAI-RAG chatbot for different Perwakilan RI. This might involve creating smaller, targeted fine-tuning datasets or RAG knowledge bases specific to each mission, building upon the foundational model developed in this thesis, to provide more localized and relevant assistance.

By addressing the outlined research plan, this thesis has the potential to make a valuable contribution to the field of AI in public administration, demonstrating a practical and innovative approach to enhancing consular services through the strategic application of localized LLMs and advanced AI techniques.

Works cited

1. The Evolving Journey of Sahabat-AI - Twimbit, accessed May 10, 2025, https://twimbit.com/about/blogs/the-evolving-journey-of-sahabat-ai
2. Sahabat AI: The Friend Indonesia Needs for a Digital Future - Twimbit, accessed May 10, 2025, https://twimbit.com/about/blogs/sahabat-ai-the-friend-indonesia-needs-for-a-digital-future
3. RAG vs Fine-Tuning: Choosing the Right Approach for Building LLM-Powered Chatbots, accessed May 10, 2025, https://www.techaheadcorp.com/blog/rag-vs-fine-tuning-difference-for-chatbots/
4. Reformasi Birokrasi Kemlu, accessed May 10, 2025, https://fe-non-production.apps.opppd2-dev.layanan.go.id/kebijakan/reformasi-birokrasi-kemlu
5. Foreign Ministry to Utilize AI-Based Services for Indonesian Citizens ..., accessed May 10, 2025, https://en.tempo.co/read/1975389/foreign-ministry-to-utilize-ai-based-services-for-indonesian-citizens-abroad
6. SARI App: Indonesia and UN Women's Effort to Protect ... - RRI.co.id, accessed May 10, 2025, https://www.rri.co.id/internasional/1465199/sari-app-indonesia-and-un-women-s-effort-to-protect-female-migrant-workers
7. Indosat, GoTo unveil open source AI model to bridge Indonesian ..., accessed May 10, 2025, https://www.capacitymedia.com/article/2e0zedgennwfbgusft5hd/news/article-indosat-goto-unveil-open-source-ai-model

8. Key LLM Trends 2025: Transforming Federal Agencies & Beyond - TechSur Solutions, accessed May 10, 2025, https://techsur.solutions/key-llm-trends-for-2025/
9. AI in Government: A Strategic Framework for Digital Transformation - REI Systems, accessed May 10, 2025, https://www.reisystems.com/ai-in-government-a-strategic-framework-for-digital-transformation/
10. AI-Powered Rules as Code: Experiments with Public Benefits Policy ..., accessed May 10, 2025, https://digitalgovernmenthub.org/publications/ai-powered-rules-as-code-experiments-with-public-benefits-policy/?utm_source=dbn&utm_medium=social&utm_campaign=other
11. AI Governance and Ethics: Lessons from the U.S. Visa Revocation Policy, accessed May 10, 2025, https://moderndiplomacy.eu/2025/03/11/ai-governance-and-ethics-lessons-from-the-u-s-visa-revocation-policy/
12. GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct · Hugging Face, accessed May 10, 2025, https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct
13. GoToCompany/llama3-8b-cpt-sahabatai-v1-base · Hugging Face, accessed May 10, 2025, https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-base
14. Sahabat-AI, accessed May 10, 2025, https://sahabat-ai.com/
15. How to Build a Retrieval-Augmented Generation Chatbot - Anaconda, accessed May 10, 2025, https://www.anaconda.com/blog/how-to-build-a-retrieval-augmented-generation-chatbot
16. LLM Approach: Prompting, Fine-tuning, AI Agents, & RAG Systems - Analytics Vidhya, accessed May 10, 2025, https://www.analyticsvidhya.com/blog/2025/04/llm-approach/
17. RAFT: Adapting Language Model to Domain Specific RAG - arXiv, accessed May 10, 2025, https://arxiv.org/html/2403.10131v1
18. Best Practices in RAG Evaluation: A Comprehensive Guide - Qdrant, accessed May 10, 2025, https://qdrant.tech/blog/rag-evaluation-guide/
19. RbFT: Robust Fine-tuning for Retrieval-Augmented Generation against Retrieval Defects, accessed May 10, 2025, https://arxiv.org/html/2501.18365v1
20. arxiv.org, accessed May 10, 2025, https://arxiv.org/pdf/2503.01131
21. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge - arXiv, accessed May 10, 2025, https://arxiv.org/html/2403.01432v2
22. Kementerian Luar Negeri RI - Portal Kemlu, accessed May 10, 2025, https://fe-non-production.apps.opppd2-dev.layanan.go.id/kontak
23. The Embassy of the Republic of Indonesia - Washington, accessed May 10, 2025, https://kemlu.go.id/washington/tentang-perwakilan/kontak-kami
24. General Information - e-Consular Service KBRI WDC, accessed May 10, 2025, https://consular.embassyofindonesia.org/page/generalinformation.html

25. Privacy Policy - Safe Travel, accessed May 10, 2025, https://bo-safetravel.kemlu.go.id/privacy-policy
26. mptf.undp.org, accessed May 10, 2025, https://mptf.undp.org/sites/default/files/documents/2025-02/final_report_2024_migration_mptf_indonesia.pdf
27. Kemlu GO - Apps on Google Play, accessed May 10, 2025, https://play.google.com/store/apps/details?id=id.go.kemlu.app
28. Fine-tune LLMs with synthetic data for context-based Q&A using Amazon Bedrock - AWS, accessed May 10, 2025, https://aws.amazon.com/blogs/machine-learning/fine-tune-llms-with-synthetic-data-for-context-based-qa-using-amazon-bedrock/
29. Synthetic Data Generation Strategies for Fine-Tuning LLMs - Scale AI, accessed May 10, 2025, https://scale.com/blog/synthetic-data-fine-tuning-llms
30. DALE: Generative Data Augmentation for Low-Resource Legal NLP - ACL Anthology, accessed May 10, 2025, https://aclanthology.org/2023.emnlp-main.528/
31. CoDa: Constrained Generation based Data Augmentation for Low-Resource NLP - arXiv, accessed May 10, 2025, https://arxiv.org/html/2404.00415v1
32. Quantum-PEFT: Ultra parameter-efficient fine-tuning - arXiv, accessed May 10, 2025, https://arxiv.org/html/2503.05431v1
33. A step-by-step guide to fine-tuning LLaMA 3 using LoRA and QLoRA, accessed May 10, 2025, https://rabiloo.com/blog/a-step-by-step-guide-to-fine-tuning-llama-3-using-lora-and-qlora
34. The Ultimate Guide to Fine-Tune LLaMA 3, With LLM Evaluations ..., accessed May 10, 2025, https://www.confident-ai.com/blog/the-ultimate-guide-to-fine-tune-llama-2-with-llm-evaluations
35. 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
36. arxiv.org, accessed May 10, 2025, https://arxiv.org/pdf/2401.09168
37. Reinforcement Learning for Optimizing RAG for Domain Chatbots - arXiv, accessed May 10, 2025, https://arxiv.org/html/2401.06800v1/
38. arxiv.org, accessed May 10, 2025, https://arxiv.org/abs/2401.06800
39. GRASP: Municipal Budget AI Chatbots for Enhancing Civic Engagement - arXiv, accessed May 10, 2025, https://arxiv.org/html/2503.23299v1
40. Indonesia Travel Advisory - Travel.gov, accessed May 10, 2025, https://travel.state.gov/content/travel/en/traveladvisories/traveladvisories/indonesia-travel-advisory.html
41. Emulating Retrieval Augmented Generation via Prompt Engineering for Enhanced Long Context Comprehension in LLMs - arXiv, accessed May 10, 2025, https://arxiv.org/html/2502.12462v1
42. Prompt engineering for RAG - OpenAI Developer Forum, accessed May 10, 2025,

https://community.openai.com/t/prompt-engineering-for-rag/621495
43. RAG systems: Best practices to master evaluation for accurate and reliable AI. | Google Cloud Blog, accessed May 10, 2025, https://cloud.google.com/blog/products/ai-machine-learning/optimizing-rag-retrieval
44. LLM Evaluation: Benchmarks to Test Model Quality - Label Your Data, accessed May 10, 2025, https://labelyourdata.com/articles/llm-fine-tuning/llm-evaluation
45. Fine-Tuning LLMs: A Guide With Examples - DataCamp, accessed May 10, 2025, https://www.datacamp.com/tutorial/fine-tuning-large-language-models
46. Protocol for human evaluation of generative artificial intelligence chatbots in clinical consultations - PMC, accessed May 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11922213/
47. Protocol for human evaluation of generative artificial intelligence chatbots in clinical consultations | PLOS One, accessed May 10, 2025, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0300487
48. Challenges of multi-task learning in LLM fine-tuning - IoT Tech News, accessed May 10, 2025, https://iottechnews.com/news/challenges-of-multi-task-learning-in-llm-fine-tuning/
49. Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts | Stanford HAI, accessed May 10, 2025, https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts