THESIS PROPOSAL

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

Abstract:
This research proposes the development and evaluation of an advanced Question Answering (QA) system to improve consular services within the Indonesian Ministry of Foreign Affairs (MoFA). The system will leverage Retrieval-Augmented Generation (RAG) to provide accurate and contextually relevant answers grounded in official MoFA documents. The core Large Language Model (LLM) will be SahabatAI, an Indonesian-centric model, which will be fine-tuned for the specific nuances of the consular domain using Parameter-Efficient Fine-Tuning (PEFT) techniques like LoRA/QLoRA. The research objectives include: (1) designing and implementing a RAG pipeline tailored for Indonesian consular information; (2) fine-tuning SahabatAI for enhanced performance in consular Q&A; and (3) rigorously evaluating the system's accuracy, faithfulness, relevance, and efficiency using a combination of automated metrics and qualitative analysis. Expected results include a significant improvement in the accessibility and reliability of consular information, contributing to both AI advancements for low-resource languages and the modernization of government public services. The proposed methodology encompasses corpus creation from MoFA sources, appropriate document chunking, embedding model selection, RAG pipeline construction, and a comprehensive evaluation framework. This project is designed to be feasible within a six-month Master's thesis timeframe.

Chapter 1: Introduction

1.1. Background of Research

The Indonesian Ministry of Foreign Affairs (MoFA) provides a comprehensive suite of consular services to its citizens residing or traveling abroad (Warga Negara Indonesia - WNI) and to foreign nationals requiring assistance related to Indonesia. These services encompass a wide array of critical functions, including passport issuance, visa applications, document legalization, and the protection of WNI.[1] However, navigating the complexities of these services presents significant challenges for users. The sheer volume of information, coupled with intricate regulations and procedural variations, often makes it difficult for individuals to access accurate, timely, and comprehensive guidance.[13] Common queries frequently pertain to visa application rejections, often due to incomplete documentation or insufficient financial proof [15], concerns regarding safety and security within Indonesia [13], and the specific procedures for various consular services.[5] These challenges underscore a persistent information bottleneck that can lead to user frustration and inefficiencies in service delivery.

Recognizing these challenges, MoFA has embarked on a journey of digitalization, introducing platforms such as the "Portal Peduli WNI" (Care for Indonesian Citizens Portal) and the "Safe Travel" mobile application. These initiatives aim to enhance services for WNI and improve data management capabilities.[12] Furthermore, MoFA has signaled its intent to leverage Artificial Intelligence (AI) to augment its service offerings, as exemplified by the Sahabat Artifisial Migran Indonesia (SARI) chatbot project. The SARI initiative aims to improve WNI services by summarizing security information and providing empathetic, language-aware responses.[27] This strategic direction towards technological innovation is consistent with MoFA's broader goals, as outlined in its Strategic Plan (Renstra), which emphasizes enhancing WNI protection and elevating the quality of public services through technological advancements and innovation.[20]

The advent of advanced AI technologies, particularly Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), presents a transformative opportunity to address the existing information dissemination challenges within consular services. An AI-powered Question Answering (QA) system, grounded in official MoFA documentation, can provide users with instant, accurate, and contextually aware answers to a diverse range of queries. Such a system would not only alleviate the burden on consular staff, allowing them to focus on more complex cases, but also significantly improve user satisfaction by making information more accessible and reliable. This research proposes to build upon MoFA's current digitalization efforts by developing a sophisticated, document-grounded QA system that leverages the unique strengths of an Indonesian-centric LLM and the factual grounding capabilities of RAG. The development of such a system aligns directly with MoFA's strategic trajectory towards AI adoption and represents a logical and impactful progression from basic digital platforms to more intelligent, context-aware information services. This thesis, therefore, aims to contribute to MoFA's ongoing modernization of consular services by providing a robust and effective AI solution.

1.2. State-of-the-Art (SOTA)

Recent years have witnessed remarkable advancements in LLMs, such as the GPT series (OpenAI) and Llama series (Meta AI), which have demonstrated impressive capabilities in understanding and generating human-like text, making them highly suitable for sophisticated QA applications.[31] These models, however, can sometimes produce factually incorrect or "hallucinated" information when not grounded in a specific knowledge base.[31]

To address this limitation, Retrieval-Augmented Generation (RAG) has emerged as a

SOTA technique. RAG enhances the factuality and domain-specificity of LLMs by enabling them to access and incorporate information from external, authoritative knowledge bases before generating a response.[31] This paradigm is increasingly being adopted for specialized domains, including legal and governmental Q&A systems, where accuracy and reliability are paramount.[38]

In the public sector, AI-powered chatbots and conversational AI systems are being deployed to improve citizen engagement and information provision.[46] Examples include Canada's "Pubbie" for performance measurement and insight reporting within the National Research Council [46], and various government chatbots designed to answer citizen queries.[47] These initiatives highlight a global trend towards leveraging AI for enhanced public service delivery.

A significant development for Indonesian Natural Language Processing (NLP) is the emergence of SahabatAI, an LLM ecosystem specifically designed for Bahasa Indonesia and its various dialects.[50] Co-initiated by Indonesian technology and telecommunication giants GoTo Group and Indosat Ooredoo Hutchison, with support from AI Singapore [50], SahabatAI aims to bolster Indonesia's AI sovereignty and capabilities. Its strong performance on Indonesian language benchmarks and its focus on local context make it an ideal candidate for developing advanced NLP applications within the Indonesian public sector.[52] The foundational work by Lewis et al. (2020) on RAG [37], Touvron et al. on Llama models [43] which form the basis for some SahabatAI variants, and the development of Indonesian LLMs like SahabatAI itself [50], collectively provide the SOTA foundation for this proposed research. Further relevant work includes studies on AI in government services [40], the development of multilingual embedding models [37], and research on evaluating RAG systems.[43]

1.3. Gap Analysis

Despite MoFA's ongoing digitalization efforts through platforms like "Portal Peduli WNI" and "Safe Travel" [12], there remains a significant opportunity to enhance consular information services through more advanced AI. Current systems, often relying on traditional search mechanisms or structured FAQs, may struggle to handle nuanced user queries, interpret complex official documents, and provide comprehensive, context-aware answers. This is evidenced by persistent challenges in areas such as WNI data management and the public's demand for more responsive and accurate services.[17]

General-purpose LLMs, even those with multilingual capabilities, often lack the deep, domain-specific knowledge and nuanced understanding of terminology and

regulations pertinent to Indonesian consular services. A critical limitation of standalone LLMs is their propensity to "hallucinate" or generate factually incorrect information, especially when dealing with specialized or rapidly changing information, unless they are properly grounded in an authoritative knowledge source.[31] This makes them unreliable for critical governmental applications where accuracy is paramount.

The proposed research directly addresses this gap by combining the strengths of RAG with a fine-tuned, Indonesian-centric LLM, SahabatAI. While RAG is gaining traction globally for enhancing LLM performance in specialized domains [33], its application using a dedicated Indonesian LLM like SahabatAI for the specific and complex domain of consular services represents a relatively unexplored area. Existing MoFA digital tools may not yet leverage advanced NLP techniques for deep document understanding and context-aware information retrieval.[20] The current gap lies in creating a system that is not only linguistically proficient in Bahasa Indonesia but also factually reliable and deeply aware of MoFA's specific knowledge base, as contained within its official documents and regulations.

Furthermore, while AI applications in government services are increasing, the development of robust, document-grounded QA systems for specific governmental domains in "low-resource" languages like Bahasa Indonesia is less mature compared to English. This research, therefore, has the potential to pioneer a practical and impactful application of SOTA AI techniques for the Indonesian public sector, offering a valuable case study and potentially a blueprint for similar initiatives in other governmental domains within Indonesia and other countries with similar linguistic and informational challenges. The successful implementation of this system will contribute to the broader field by demonstrating how localized LLMs, when augmented with RAG, can effectively serve complex public sector information needs.

1.4. Problem Formulation (Research Questions)

This research aims to address the challenges of providing accurate, comprehensive, and easily accessible consular information by Indonesian Embassies/Consulates. The central problem is the current information bottleneck, where users struggle to navigate complex regulations and MoFA staff are burdened with repetitive queries. The proposed solution is an AI-powered Question Answering system leveraging Retrieval-Augmented Generation (RAG) and a fine-tuned SahabatAI LLM. To guide this research, the following research questions are formulated:

- RQ1: How can a Retrieval-Augmented Generation (RAG) pipeline, utilizing official Indonesian Ministry of Foreign Affairs (MoFA) documents as its knowledge base, be effectively designed and implemented to answer consular service queries in

Bahasa Indonesia?
- RQ2: What are the optimal strategies for pre-processing and chunking diverse Indonesian consular documents (e.g., regulations, FAQs, official announcements) to maximize retrieval relevance and contextual integrity for the RAG system?
- RQ3: How can the SahabatAI LLM be effectively fine-tuned (e.g., using PEFT methods like LoRA/QLoRA) with consular-specific data to enhance its performance in understanding and generating accurate, relevant, and faithful answers to consular queries in Bahasa Indonesia?
- RQ4: What evaluation metrics and methodologies are most appropriate for assessing the performance of the proposed SahabatAI-RAG system in the Indonesian consular domain, specifically measuring answer accuracy, faithfulness to source documents, relevance to user queries, and context utility?
- RQ5: How does the performance of the proposed fine-tuned SahabatAI-RAG system compare against baseline systems (e.g., base SahabatAI without RAG, traditional keyword search over MoFA documents) in addressing Indonesian consular service queries?

1.5. Objectives of Research

The primary aim of this research is to develop and evaluate an advanced Question Answering system for Indonesian consular services. The specific objectives are:

- Objective 1: To design and develop a robust RAG pipeline tailored for Indonesian consular service information, incorporating effective document ingestion, embedding, indexing, and retrieval mechanisms.
- Objective 2: To fine-tune a selected SahabatAI LLM (e.g., Llama3-8B-CPT-SahabatAI-v1-instruct or Gemma2-9B-CPT-SahabatAI-v1-instruct) using PEFT techniques (LoRA/QLoRA) to adapt its capabilities for the specific language, terminology, and query patterns of the Indonesian consular domain.
- Objective 3: To create a comprehensive knowledge base from official MoFA documents, employing an optimized chunking strategy suitable for governmental and legal texts in Bahasa Indonesia.
- Objective 4: To implement and integrate the fine-tuned SahabatAI model within the RAG pipeline to create a functional consular QA system.
- Objective 5: To rigorously evaluate the performance of the developed system using a comprehensive set of metrics, including accuracy, F1-score, ROUGE, BLEU, METEOR (adapted for Indonesian), and RAG-specific metrics for faithfulness, answer relevance, and context relevance/precision/recall.
- Objective 6: To analyze the system's strengths and weaknesses and provide

insights into the practical application of LLMs and RAG for enhancing public services in Indonesia.

1.6. Scope and Limitations

Scope:
The research will focus on developing a QA system for a selected subset of Indonesian consular services. This selection will prioritize services with readily available, structured official documentation and high query frequency, such as passport applications, visa information for specific commonly visited countries, and procedures for common document legalization types. The primary language for both user queries and system-generated answers will be Bahasa Indonesia. While SahabatAI models support other Indonesian regional languages like Javanese and Sundanese 51, these will be considered out of scope for this initial six-month thesis to ensure project feasibility.
The knowledge base for the RAG system will be constructed exclusively from publicly available official documents sourced from the Indonesian Ministry of Foreign Affairs (Kemlu.go.id [63]), the Directorate General of Immigration website (imigrasi.go.id [2]), and potentially the websites of selected major Indonesian Embassies and Consulates (e.g., Washington, The Hague, Singapore [1]). The system is intended to address informational queries (e.g., "What are the requirements for obtaining a new passport?", "How do I apply for a tourist visa to country X?") rather than providing personalized, case-specific advice or performing transactional services.

Limitations:
The six-month timeframe allocated for this Master's thesis imposes certain limitations. The size and complexity of the knowledge base will necessarily be constrained, and the extent of LLM fine-tuning will be focused and targeted. A significant challenge may lie in accessing a comprehensive, up-to-date, and machine-readable corpus of MoFA documents; the system's performance will inherently depend on the quality, scope, and format of the data collected.63 While MoFA's SARI chatbot aims for "empathy" [27], the evaluation of such nuanced aspects in this research will be primarily qualitative, given the difficulty in robust quantitative measurement within the project's scope and timeframe. The research will concentrate on the technical development and evaluation of the QA system. The design of a user interface for deployment beyond a basic prototype for testing purposes is outside the current scope. Furthermore, the system will operate on a static, curated corpus; mechanisms for real-time updates to the knowledge base will not be implemented.

1.7. Thesis Outline

This thesis proposal is structured as follows: Chapter 1 provides an introduction to the research, including the background, problem statement, objectives, and scope.

Chapter 2 presents a comprehensive literature review covering Indonesian consular services, Large Language Models, the SahabatAI ecosystem, Retrieval-Augmented Generation, LLM fine-tuning techniques, and related work in AI for government services. Chapter 3 details the proposed methodology, including the research design, data collection and preparation strategies, the architecture of the proposed "ConsularAI-RAG" system, the SahabatAI model fine-tuning plan, and the performance evaluation framework. Chapter 4 outlines the expected results and contributions of the research to the field of AI and Indonesian government services, alongside a discussion of potential challenges. Finally, Chapter 5 concludes the proposal with a summary and directions for future work.

Chapter 2: Literature Review / Theoretical Background

2.1. Indonesian Consular Services: Domain Overview and Information Needs

Indonesian consular services, administered by the Ministry of Foreign Affairs (MoFA) and its network of Embassies and Consulates General worldwide, provide essential support to Indonesian citizens (WNI) abroad and foreign nationals interacting with Indonesia.[1] These services are diverse, encompassing passport issuance and renewal, various types of visa applications (e.g., tourist, business, diplomatic, service [7]), legalization of documents for international use [7], and comprehensive protection services for WNI (Pelindungan WNI), which includes legal aid and assistance in crisis situations.[18] The Directorate of Consular Affairs and the Directorate for the Protection of Indonesian Nationals and Indonesian Legal Entities (Dit PWNI & BHI) are key MoFA units responsible for these functions.[18] The legal framework for WNI protection is detailed in regulations such as the Minister of Foreign Affairs Regulation (Permenlu) No. 5 Tahun 2018.[22]

Despite the critical nature of these services, users often face challenges in accessing clear, accurate, and timely information. The procedures and document requirements can be complex and vary depending on the specific service and location.[13] For instance, visa applications are frequently rejected due to incomplete documentation or failure to meet financial sufficiency criteria.[15] Furthermore, WNI abroad may require urgent information related to safety, security, or legal assistance.[13] MoFA's own strategic documents (Renstra) and performance reports (LAKIP, if available) acknowledge the ongoing need to improve public service quality and enhance WNI protection, often highlighting digitalization as a key enabler.[20] Existing digital initiatives like the Portal Peduli WNI and the Safe Travel mobile application aim to improve WNI data management and service accessibility [19], but challenges related to data

completeness and the responsiveness of information services persist.[17]

The extensive range of consular services, each governed by specific regulations and procedures (e.g., visa requirements detailed in [15], document legalization processes outlined in [8]), creates a complex information landscape. This complexity, coupled with reported difficulties in WNI data management [17] and the continuous demand for responsive and accurate services [20], points to an "information bottleneck." Both citizens and potentially consular staff may find it challenging to quickly locate comprehensive and precise information. This situation underscores the need for more advanced information retrieval systems. MoFA's strategic documents themselves articulate the imperative to enhance information systems and service quality.[20] An AI-powered RAG system, as proposed in this thesis, aims directly at alleviating this information bottleneck by providing a centralized, intelligent, and document-grounded access point to the vast repository of Indonesian consular knowledge, thereby improving efficiency and user satisfaction.

2.2. Large Language Models (LLMs) for Question Answering

Large Language Models (LLMs) represent a significant breakthrough in artificial intelligence, characterized by their ability to understand, generate, and manipulate human language with unprecedented sophistication. At their core, most modern LLMs, such as those in the GPT and Llama families, are based on the transformer architecture, which employs self-attention mechanisms to process and weigh the importance of different parts of an input text sequence. These models undergo extensive pre-training on vast corpora of text and code, enabling them to learn intricate patterns, grammatical structures, and a broad range of world knowledge. This pre-training phase is followed by fine-tuning, where the model is further trained on smaller, task-specific datasets to adapt its capabilities for particular applications, such as question answering (QA).[31]

LLMs excel in Natural Language Understanding (NLU), allowing them to comprehend the meaning and intent behind user queries, and Natural Language Generation (NLG), enabling them to produce fluent, coherent, and contextually relevant answers.[31] These capabilities make them highly effective for building advanced QA systems. QA systems can be broadly categorized into extractive and generative types. Extractive QA systems identify and extract answers directly from provided text sources, while generative QA systems synthesize novel answers based on their learned knowledge and the provided context.[32] LLMs are particularly adept at generative QA. Furthermore, QA systems can be classified as open-domain, designed to handle questions on virtually any topic, or closed-domain, specializing in specific areas like

medicine or law.[32] The proposed system falls into the category of closed-domain, generative QA, focusing specifically on Indonesian consular services.

Despite their power, LLMs face several challenges. One significant issue is "hallucination," where models generate plausible-sounding but factually incorrect or nonsensical information, especially when not grounded in a reliable external knowledge source.[31] They also have limitations in their context window (the amount of text they can process at once), which can be problematic when dealing with long documents or complex queries requiring extensive background information.[31] Additionally, training and running very large LLMs can be computationally expensive. These challenges highlight the need for techniques like RAG to enhance the reliability and domain-specificity of LLM-based QA systems.

2.3. SahabatAI: An Indonesian Large Language Model Ecosystem

SahabatAI represents a landmark initiative in the Indonesian AI landscape, an open-source Large Language Model (LLM) ecosystem specifically developed for Bahasa Indonesia and its diverse local dialects, including Javanese and Sundanese.[50] This project was co-initiated by leading Indonesian technology and telecommunication companies, PT GoTo Gojek Tokopedia Tbk (GoTo) and Indosat Ooredoo Hutchison, with significant collaboration and support from AI Singapore and Tech Mahindra.[50] The development of SahabatAI is strategically aligned with Indonesia's national vision of "Golden Indonesia 2045" and is a crucial step towards advancing the nation's digital sovereignty and technological self-reliance in the field of AI.[53]

The SahabatAI model family includes variants based on robust open-source architectures such as Meta's Llama3 (e.g., Llama3-8B-CPT-SahabatAI-v1-base and the instruction-tuned Llama3-8B-CPT-SahabatAI-v1-instruct [51]) and Google's Gemma2 (e.g., Gemma2-9B-CPT-SahabatAI-v1-base and -instruct [56]). These models are primarily decoder-type architectures, optimized for text generation tasks. A key strength of SahabatAI is its deep understanding of Bahasa Indonesia, including regional slang and idiomatic expressions, which allows it to address critical context and cultural reference gaps often left by global, multilingual LLMs.[53] This linguistic and cultural attunement makes SahabatAI particularly well-suited for applications serving the Indonesian public, such as the proposed consular QA system. The models also support English, facilitating bilingual use cases if needed.[51]

The development of SahabatAI involved extensive pre-training and fine-tuning. For instance, the Llama3-8B-CPT-SahabatAI-v1-base model underwent continued

pre-training on approximately 50 billion tokens, incorporating diverse data sources such as Indonesian Wikipedia, Indonesian news articles, and text piles in Javanese and Sundanese.[51] The instruction-tuned version, Llama3-8B-CPT-SahabatAI-v1-Instruct, was further fine-tuned with a substantial dataset of around 448,000 Indonesian instruction-completion pairs, augmented by instruction sets in Javanese, Sundanese, and English.[52] SahabatAI models have demonstrated strong performance on various NLP benchmarks relevant to the Indonesian language, including SEA HELM (also known as BHASA) and IndoMMLU, excelling in tasks such as question answering, sentiment analysis, and translation.[52]

SahabatAI models are openly accessible via Hugging Face [50], and the initiative actively encourages collaboration from researchers, developers, and language enthusiasts to contribute to its enhancement and expansion.[50] This collaborative ethos and open accessibility present a unique opportunity for academic research, such as this Master's thesis, to contribute to the practical application and domain-specific adaptation of this significant Indonesian AI asset. While powerful, it is important to note that, like many LLMs, SahabatAI models can sometimes generate "hallucinations" or factually incorrect information. The commercially permissive releases have not been specifically aligned for safety, necessitating that developers and users implement their own safety fine-tuning and security measures for specific applications.[52] The primary challenge for any consular QA system in Indonesia is effective communication in Bahasa Indonesia, encompassing its nuances and cultural context. SahabatAI is explicitly designed for this purpose [50], offering a distinct advantage over generic multilingual models that may lack the required depth. The "call for collaboration" by the SahabatAI team [50] directly invites contributions such as sharing preference data and proposing new evaluation tasks, aligning perfectly with the objectives and potential contributions of a Master's thesis project focused on its application. This research, therefore, is not merely applying an existing LLM but has the potential to contribute to the development and domain-specific refinement of a strategically important national AI resource.

Table 2.3: Overview of SahabatAI Model Variants

Model Name	Base Architecture	Parameters	Key Features	Indonesian Performance Highlights (SEA HELM/BHASA - Overall	Suitability for Downstream QA Task

				ID+JV+SU)	
Llama3-8B-CPT-SahabatAI-v1-base [51]	Llama3	8B	Continued pre-trained on Indonesian & regional languages, English support, 8192 context length.	53.725 (Instruct version) [52]	High (requires instruction fine-tuning)
Llama3-8B-CPT-SahabatAI-v1-instruct [52]	Llama3	8B	Instruction-tuned for Indonesian, Javanese, Sundanese, English; ~448k Indonesian instruction pairs.	53.725 [52]	Very High
Gemma2-9B-CPT-SahabatAI-v1-base [77]	Gemma2	9B	Continued pre-trained, focused on Indonesian & regional languages, English support.	61.169 (Instruct version) [52]	High (requires instruction fine-tuning)
Gemma2-9B-CPT-SahabatAI-v1-instruct [77]	Gemma2	9B	Instruction-tuned for Indonesian, Javanese, Sundanese, English; strong performance on BHASA, IndoMMLU.	61.169 [52]	Very High

Note: Performance scores are indicative and may vary based on specific evaluation

setups. The instruct versions are generally more suitable for direct application in QA tasks.

2.4. Retrieval-Augmented Generation (RAG) for Domain-Specific QA

2.4.1. RAG Architecture and Workflow

Retrieval-Augmented Generation (RAG) is an advanced AI technique designed to enhance the output quality and factual accuracy of Large Language Models (LLMs). It achieves this by enabling LLMs to access and reference an authoritative, external knowledge base outside of their original training data before generating a response.[31] This approach directly addresses inherent limitations of LLMs, such as their knowledge being static (limited to the data they were trained on) and their propensity to "hallucinate" or generate incorrect information, particularly when faced with queries requiring specialized or up-to-date knowledge.[31]

The core workflow of a RAG system typically involves the following steps [35]:

1. **User Query Input:** The user submits a query in natural language.
2. **Information Retrieval:** The system first uses the user's query to search and retrieve relevant information snippets or documents from the external knowledge base. This knowledge base is often a collection of domain-specific documents that have been pre-processed and indexed.
3. **Prompt Augmentation:** The retrieved relevant information (context) is then combined with the original user query to create an augmented prompt.
4. **LLM Response Generation:** This augmented prompt, now rich with contextual information, is fed to the LLM. The LLM uses both its pre-trained knowledge and the provided context to generate a comprehensive and factually grounded answer.

The benefits of employing RAG are manifold. It leads to improved factual accuracy as responses are directly based on information from the provided knowledge source. It allows LLMs to utilize up-to-date information without requiring frequent and costly retraining of the entire model. RAG also enhances transparency, as the system can often provide references to the source documents used to generate the answer, allowing users to verify the information.[31] This is particularly crucial for applications in governmental or legal domains where traceability and verifiability are paramount. Finally, RAG enables the generation of highly domain-specific responses, making LLMs more effective for specialized QA tasks.

2.4.2. Embedding Models for Indonesian Language

Embedding models play a pivotal role in the retrieval phase of RAG systems. Their primary function is to convert textual data—both the documents in the knowledge base and the user queries—into dense numerical vector representations.[35] These vector embeddings capture the semantic meaning of the text, such that texts with similar meanings are located closer to each other in the high-dimensional vector space. This property enables efficient semantic search, where the system can identify documents relevant to a query based on the similarity of their vector embeddings, rather than relying solely on keyword matching.

For this research, which focuses on Indonesian consular services, the selection of an appropriate embedding model is critical. While many SOTA embedding models are English-centric, the effectiveness of the RAG system will heavily depend on a model that can accurately capture the semantics of Bahasa Indonesia. Multilingual models, such as sentence-transformers/paraphrase-multilingual-mpnet-base-v2, which supports Indonesian and is designed for tasks like semantic search, are strong candidates.[78] Recent advancements in multilingual embeddings, such as the LUSIFER architecture that aligns multilingual encoders (e.g., XLM-R) with LLM-based embedding models, show promise for enhancing performance, especially for low-resource languages like Indonesian.[37] The Massive Text Embedding Benchmark (MTEB) serves as a valuable resource for evaluating and comparing the performance of various embedding models across different languages and tasks.[78] The selection for this project will prioritize models demonstrating strong performance on Indonesian semantic similarity or retrieval tasks.

Table 2.2: Evaluation of Potential Embedding Models for Indonesian Language

Model Name	Indonesian Language Performance (Indicative from MTEB/Related Benchmarks)	Dimensionality	Suitability for RAG (Indonesian Consular Docs)
sentence-transformers/paraphrase-multilingual-mpnet-base-v2 [78]	Good general multilingual performance, supports Indonesian. Specific MTEB scores	768	High

	for 'id' to be verified.		
IndoBERT (various versions)	Specifically pre-trained on Indonesian; strong on IndoNLU/IndoLEM benchmarks.	Varies	High (if sentence embedding variants are used)
Other models from MTEB leaderboard with high 'id' scores [79]	To be investigated from live MTEB leaderboard.	Varies	Potentially High

Note: Performance needs to be verified on the latest MTEB leaderboard for Indonesian-specific retrieval/semantic similarity tasks.

2.4.3. Vector Databases and Indexing Techniques

Once text is converted into vector embeddings, a specialized database is required to store and efficiently query these embeddings. Vector databases are designed for this purpose, enabling fast similarity searches over large volumes of high-dimensional vector data.[35] They typically employ algorithms like k-Nearest Neighbors (k-NN) or Approximate Nearest Neighbors (ANN) to find the vectors in the database that are most similar (e.g., by cosine similarity or Euclidean distance) to a given query vector.

For a Master's thesis project, open-source vector database solutions like FAISS (developed by Facebook AI) or ChromaDB are often preferred due to their accessibility and ease of integration. Milvus is another powerful open-source option known for its scalability.[82] The choice of vector database will also depend on its compatibility with the selected RAG framework (e.g., LangChain or LlamaIndex). Efficient indexing strategies, such as Hierarchical Navigable Small World (HNSW) graphs or Inverted File Index (IVF), are crucial for accelerating the retrieval process, especially as the size of the document corpus grows.[36]

2.4.4. Document Chunking Strategies for Consular and Governmental Documents

Document chunking is a critical pre-processing step in RAG pipelines. It involves breaking down large documents into smaller, more manageable pieces, or "chunks".[36] This is necessary because LLMs have a limited context window (the maximum number of tokens they can process at once), and providing overly long, undifferentiated text

segments can dilute the relevant information and impair retrieval accuracy. Well-structured chunks help ensure that the retrieval system can accurately identify the most relevant parts of a document for a given query.

Consular documents, such as laws, regulations, and detailed procedural guidelines (e.g., Permenlu No. 5 Tahun 2018 on WNI Protection [22], Permenlu No. 14 Tahun 2022 on Document Legalization [74]), often possess complex structures, hierarchical organization (e.g., chapters, articles, sections), specific legal terminology, and important cross-references. Simple chunking strategies like fixed-size chunking (dividing text into uniform blocks based on character or token count) or basic sentence-based chunking [83] risk breaking apart crucial contextual information. For instance, a single sentence might not capture the full meaning of a legal clause, or a fixed-size chunk might arbitrarily split a regulation mid-article. Document-based chunking, while preserving full context, might be too coarse for lengthy official documents, exceeding LLM context limits or making precise information retrieval difficult.[85]

Therefore, for governmental and legal texts, more sophisticated chunking strategies are required. Options include:

- **Paragraph-based chunking:** Treating each paragraph as a chunk, which often aligns with distinct ideas or arguments in structured documents.[83]
- **Content-aware chunking:** This involves splitting documents based on their inherent structure, such as headings, sections, articles, or even individual Q&A pairs in FAQ documents. This respects the logical organization of the source material. For example, Permenlu documents are often structured by "BAB" (Chapter) and "Pasal" (Article), which could serve as natural chunk boundaries.[22]
- **Semantic chunking:** This experimental technique groups sentences based on semantic similarity, aiming to create chunks that are thematically coherent, even if they cross paragraph boundaries.[83] LangChain offers an implementation of a semantic chunker.[84]
- **Recursive character text splitting:** This adaptive approach uses a hierarchy of separators (e.g., paragraphs, sentences, specific markers) to find meaningful boundaries.[85]

The LegalBench-RAG benchmark specifically emphasizes the importance of precise retrieval, focusing on extracting minimal, highly relevant text segments from legal documents rather than large, imprecise chunks, as the latter can exceed context window limitations and lead to information dilution or hallucinations.[39] The choice of chunking strategy for this thesis must carefully consider these factors to preserve the

semantic integrity and legal nuances of Indonesian consular documents. An optimal strategy might involve a hybrid approach, perhaps starting with structural cues (like sections or articles) and then applying semantic or sentence-based splitting within those larger units if they are still too long. The determination of optimal chunk size and the use of chunk overlap (repeating some tokens between adjacent chunks to maintain context) will also be important considerations. This careful selection and potential adaptation of chunking strategies for Indonesian official documents could represent a valuable contribution of this research.

2.4.5. Advanced RAG: Reranking and Hybrid Approaches

To further enhance the quality of information provided to the LLM, many RAG systems incorporate a reranking step after the initial retrieval phase.[36] A reranking model takes the top-k initially retrieved document chunks and reorders them based on more nuanced relevance signals, ensuring that the most pertinent information is prioritized for the LLM. This can improve the overall accuracy of the final generated answer.

Hybrid RAG approaches represent another area of active research. These systems aim to combine the strengths of dense vector retrieval from unstructured text with information from structured knowledge sources, such as Knowledge Graphs (KGs).[38] GraphRAG, for example, constructs a KG from documents and uses it to enhance text retrieval.[38] Such hybrid methods can improve reasoning capabilities and provide more interpretable results. Other advanced techniques include iterative retrieval, where the system refines its understanding and retrieves more information in multiple steps, and multi-step QA, where complex questions are broken down into simpler sub-questions.[41] While a full implementation of complex hybrid RAG or iterative retrieval might be beyond the scope of a six-month Master's thesis, understanding these SOTA directions provides important context for the proposed research.

Table 2.1: Comparison of RAG Frameworks and Techniques

Framework	Key Features	Primary Focus	Suitability for Indonesian Context	Open-Source Status
LangChain [82]	Component chaining, data connections, model flexibility, extensive	General LLM application development, RAG	High; flexible for integrating custom components like SahabatAI and	Yes

	integrations, agents, evaluation tools.	applications.	Indonesian embedding models.	
LlamaIndex (formerly GPT Index) [82]	Data ingestion from various sources (APIs, PDFs, SQL), customizable indexing (vector, keyword, graph), query engines, response synthesis.	Connecting LLMs with private data sources, RAG.	High; strong focus on data indexing and retrieval, Python and TypeScript support.	Yes
DSPy [82]	Declarative LLM programming, automatic prompt optimization, modular architecture, evaluation.	Building and optimizing complex LLM systems.	Moderate; powerful for optimization but might have a steeper learning curve for a focused RAG project.	Yes
Haystack (by deepset AI) [31]	Flexible components for NLP pipelines (retrievers, readers, generators), technology-agnostic, evaluation tools, Prompt Explorer.	Production-ready NLP and RAG applications.	High; well-established with good support for various embedding models and LLMs. Deepset AI also offers RAG-specific guidance.	Yes

Note: The choice of framework will be further detailed in Chapter 3, based on ease of use, community support, and specific features aligning with the project's requirements for handling Indonesian consular data and integrating SahabatAI.

2.5. Fine-tuning LLMs: Techniques and Considerations (PEFT, LoRA, QLoRA for SahabatAI)

Pre-trained LLMs, while possessing broad knowledge, often require further

adaptation to excel in domain-specific tasks. Fine-tuning is the process of taking a pre-trained model and continuing its training on a smaller, specialized dataset relevant to the target domain or task.[87] This process allows the model to learn domain-specific terminology, understand nuanced query patterns, and align its response style with the desired output for the specific application, such as consular question answering. Effective fine-tuning can significantly improve performance, reduce hallucinations, and enhance the model's ability to generate accurate and contextually appropriate responses.[87]

However, fully fine-tuning very large LLMs (those with billions of parameters, like SahabatAI variants [51]) poses significant challenges. It is computationally intensive, requiring substantial GPU resources and long training times, which are often beyond the capacity of individual researchers or smaller institutions.[88] Moreover, full fine-tuning can sometimes lead to "catastrophic forgetting," where the model loses some of its general capabilities learned during pre-training.

Parameter-Efficient Fine-Tuning (PEFT) methods have emerged as a powerful solution to these challenges.[89] PEFT techniques aim to adapt LLMs to new tasks by training only a small subset of the model's parameters, or by adding a small number of new, trainable parameters, while keeping the vast majority of the original pre-trained weights frozen. This dramatically reduces the computational and memory requirements for fine-tuning.

One prominent PEFT technique is **LoRA (Low-Rank Adaptation)**.[89] LoRA works by injecting trainable rank decomposition matrices into the layers of the transformer architecture (typically the attention blocks). Instead of updating the original large weight matrices, LoRA trains these much smaller low-rank matrices. This approach significantly reduces the number of trainable parameters (often by orders of magnitude) and the GPU memory needed for fine-tuning, making it feasible to adapt large models on consumer-grade hardware. The original pre-trained weights remain frozen, allowing for multiple lightweight and portable LoRA "adapters" to be created for various downstream tasks, all built upon the same base model. Models fine-tuned using LoRA often achieve performance comparable to fully fine-tuned models.[89]

An even more resource-efficient variant is **QLoRA (Quantized LoRA)**.[89] QLoRA combines LoRA with quantization, where the weights of the pre-trained model are quantized to a lower precision (e.g., 4-bit) before applying the LoRA adapters. This further reduces the memory footprint, enabling the fine-tuning of very large models on relatively modest hardware, such as a single consumer-grade GPU (e.g., a Colab Tesla T4 [91]). Tutorials and frameworks like Hugging Face PEFT and TRL provide tools

and examples for implementing LoRA and QLoRA for various LLMs, including Llama and Gemma architectures, which are relevant to SahabatAI.[89]

For fine-tuning SahabatAI in the context of this Master's thesis, a PEFT approach, particularly QLoRA if computational resources are highly constrained, is the most viable strategy. This makes the ambitious goal of adapting a large Indonesian LLM to the specific domain of consular services achievable within the typical resource limitations of academic research. The SahabatAI initiative's call for collaboration, including the sharing of pre-training, instruction, and preference data [50], also presents an opportunity to potentially leverage or contribute to domain-specific datasets for fine-tuning, further strengthening the feasibility and impact of this research. The fine-tuning process for the SahabatAI Llama3-8B-Instruct model involved full parameter fine-tuning, on-policy alignment, and model merges [52], which is a more resource-intensive process than what is proposed here with PEFT for domain adaptation.

2.6. Related Work in AI for Government and Consular Services

The application of AI, particularly LLMs and chatbots, in public sector services is a rapidly growing field aimed at enhancing citizen engagement, improving information accessibility, and increasing operational efficiency.[32] Governments worldwide are exploring AI for tasks ranging from answering citizen queries to aiding in policy-making and data analysis.[47]

Several case studies demonstrate the use of AI in government. For instance, AI has been used to assist with tax-related queries by grounding models on tax law and training materials.[40] Another example involves using GraphRAG to generate policy proposals by analyzing YouTube data related to logistics issues.[45] The National Research Council of Canada developed "Pubbie," an intelligent agent using LLM orchestration and semantic embedding, to automate performance measurement and data management.[46] Comparative studies have also been conducted to analyze the strengths and weaknesses of existing government chatbots against general-purpose LLMs like ChatGPT.[47]

Within Indonesia, the Ministry of Foreign Affairs has already signaled its intent to use AI for WNI services through the SARI chatbot initiative, which aims to provide empathetic responses and summarize security information.[27] This local initiative provides strong contextual relevance for the proposed research.

Despite the promise of LLMs, their deployment in governmental contexts faces

challenges, including concerns about data privacy, the risk of generating "hallucinations" or factually incorrect information, ensuring the integrity of outputs, and addressing data sovereignty issues.[34] RAG systems are increasingly seen as a way to mitigate some of these risks by grounding LLM responses in authoritative, verified texts, thereby improving accuracy and reducing the likelihood of fabricated information.[38] Research from NLP conferences like ACL and EMNLP also explores the safety and efficacy of RAG-based LLMs for citizen services, including the potential for RAG to introduce new safety considerations if not carefully implemented (e.g., vulnerabilities from harmful documents in the corpus, though this is less of a concern when using curated official documents).[43]

The proposed research builds upon this existing body of work by focusing on a specific, complex governmental domain (Indonesian consular services) and leveraging a localized LLM (SahabatAI) within a RAG framework. This approach aims to address the unique linguistic and contextual needs of Indonesian public services while mitigating the common pitfalls of standalone LLMs.

Chapter 3: Proposed Methodology

3.1. Research Design

This research will adopt a **constructive research approach**, focusing on the design, development, and evaluation of a novel AI-driven Question Answering (QA) system tailored for Indonesian consular services. The primary output will be the "ConsularAI-RAG" system (or a similarly refined name), which integrates a fine-tuned SahabatAI Large Language Model (LLM) with a Retrieval-Augmented Generation (RAG) pipeline.

The development process will follow an **iterative prototyping methodology**. This involves several cycles:

1. **Baseline System Development:** An initial RAG pipeline will be constructed using a pre-trained, instruction-tuned SahabatAI model (e.g., Llama3-8B-CPT-SahabatAI-v1-instruct or Gemma2-9B-CPT-SahabatAI-v1-instruct) and a knowledge base built from a preliminary set of MoFA consular documents.
2. **Corpus Expansion and Refinement:** The knowledge base will be incrementally expanded and refined based on further data collection and optimized document chunking strategies.
3. **SahabatAI Fine-tuning:** The selected SahabatAI model will be fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) techniques (specifically LoRA or QLoRA)

on a curated dataset of Indonesian consular Q&A pairs.
4. **Integration and Testing:** The fine-tuned SahabatAI model will be integrated back into the RAG pipeline.
5. **Evaluation and Iteration:** The system will be rigorously evaluated at each major stage using both quantitative metrics and qualitative analysis. Findings from the evaluation will inform further iterations and improvements to the pipeline components, fine-tuning process, or knowledge base.

An **experimental evaluation** will be conducted to assess the performance of the developed system. This will involve comparing the ConsularAI-RAG system against defined baselines (e.g., the base SahabatAI model without RAG, a traditional keyword search system) on a curated set of consular queries. Performance will be measured using metrics relevant to QA accuracy, faithfulness to source documents, relevance of answers, and the utility of retrieved context, adapted for the Indonesian language.

3.2. Data Collection and Preparation

3.2.1. Corpus Development: Sourcing Official MoFA Consular Documents

The foundation of the RAG system is its knowledge base, which will be constructed from official documents pertaining to Indonesian consular services. The primary sources for these documents will be:

- The official website of the Indonesian Ministry of Foreign Affairs (Kemlu.go.id [63] and its sub-portals like e-ppid.kemlu.go.id).
- Websites of major Indonesian Embassies and Consulates General, particularly those in countries with significant Indonesian diaspora or high consular service demand (e.g., Washington [1], San Francisco, Los Angeles, Houston, Chicago, New York [1], The Hague [5], Lima [8], Mexico City [9], Singapore [65], Kuala Lumpur [73]).
- The official website of the Directorate General of Immigration (imigrasi.go.id [2]), which provides crucial visa and immigration information.

The types of documents to be collected will include:

- **Frequently Asked Questions (FAQs):** From MoFA and embassy websites.[5]
- **Regulations (Peraturan Menteri Luar Negeri - Permenlu):** Such as Permenlu No. 5 Tahun 2018 concerning the Protection of Indonesian Citizens Abroad [22] and Permenlu No. 14 Tahun 2022 concerning Document Legalization Procedures.[74] Other relevant Permenlu documents will also be sourced.[21]
- **Official Announcements and Circulars:** Pertaining to changes in consular procedures or policies.

- **Service Procedure Descriptions:** Outlining steps for passport applications, visa types, document legalization, etc..[8]
- **MoFA Strategic Plans (Renstra):** Such as Renstra Kemlu 2020-2024, to understand policy context and service priorities.[20]
- **Information on WNI Protection:** Including guidelines and reports from Dit PWNI & BHI.[18]

Potential challenges in data acquisition include documents primarily available in PDF format (some of which may be scanned images rather than text-searchable), the need to scrape information from various websites, and ensuring the collected information is current and official. Mitigation strategies will involve using Optical Character Recognition (OCR) tools for scanned documents, PDF-to-text conversion utilities, and developing web scraping scripts (e.g., using Python libraries like BeautifulSoup or Scrapy; Firecrawl [82] could also be explored if suitable for government sites). All collected data will be in Bahasa Indonesia.

Table 3.1: Potential Data Sources for MoFA Consular Knowledge Base

Source Type	Specific URL/Document Name (Examples)	Content Type	Language	Format	Potential Volume (Estimate)
Official MoFA Website	kemlu.go.id/faq [7], layanandiplomatik.kemlu.go.id/guest/faq [6], e-ppid.kemlu.go.id (for regulations, Renstra) [20]	FAQs, Service Procedures, Regulations (Permenlu), WNI Protection Info, Strategic Plans	Bahasa Indonesia	HTML, PDF	Medium to Large
Indonesian Embassy/Consulate Websites	kemlu.go.id/lima/pelayanan-perwakilan/legalisasi-dokumen- [8], kemlu.go.id/denhaag/pel	FAQs, Specific Service Procedures (Passport, Visa, Legalization)	Bahasa Indonesia	HTML, PDF	Medium

	ayanan-perwakilan/pelayanan-konsuler/legalisasi-dokumen-warga-negara-indonesia- [10], kemlu.go.id/files-service/... (various PDFs) [9]	, Local Announcements			
Directorate General of Immigration Website	imigrasi.go.id [2]	Visa Information, Immigration Regulations, FAQs	Bahasa Indonesia	HTML, PDF	Medium
Official Regulation Depository (e.g., BPK, JDIH)	peraturan.bpk.go.id (for Permenlu PDFs like No. 5/2018 [22], No. 14/2022 [74], No. 13/2019 [75])	Full text of Ministerial Regulations (Permenlu) related to consular affairs and WNI protection	Bahasa Indonesia	PDF	Medium

3.2.2. Data Pre-processing

Once the raw documents are collected, they will undergo a series of pre-processing steps to prepare them for ingestion into the RAG system:

- **Text Extraction:** Extracting plain text from various formats (HTML, PDF). For PDFs, tools like PyMuPDF or pdfminer.six will be used. OCR (e.g., Tesseract OCR with Indonesian language packs) will be applied to scanned documents or image-based PDFs.
- **Cleaning:** Removing irrelevant content such as HTML tags, navigation menus, headers/footers, and advertisements. Standardizing text encoding (UTF-8) and normalizing whitespace.
- **Language Verification:** Although sources are primarily Indonesian, a language identification step (e.g., using langdetect or fastText) will be applied to filter out any non-Indonesian content, ensuring the knowledge base remains focused.

- **Structuring Data (Initial):** Where possible, document structure (e.g., titles, sections, Q&A pairs from FAQs) will be identified and preserved as metadata. This can aid in more targeted chunking and retrieval. For example, FAQs from kemlu.go.id [7] can be directly parsed into question-context pairs.

3.2.3. Document Chunking Strategy

The choice of document chunking strategy is critical for the performance of the RAG system, especially when dealing with complex governmental and legal documents prevalent in the consular domain.[39] A naive fixed-size chunking approach can sever important contextual links within regulations or procedures, leading to incomplete or misleading retrieved information.

Justification: Consular documents such as Permenlu [22] are often hierarchically structured with articles, sections, and clauses that have specific legal meanings. FAQs [7], while more direct, still require that the full context of a question and its corresponding answer be kept together. Therefore, a content-aware or semantic chunking strategy is preferred over simple fixed-size or sentence-based methods.

Proposed Strategy:
A hybrid approach will be investigated:

1. **Structural Chunking (Primary):** For documents with clear structural demarcations (e.g., Permenlu with "Pasal" (Article) and "Ayat" (Clause) [22], FAQs with distinct Q&A blocks [7]), chunks will be created based on these logical units. This aims to preserve the semantic integrity of individual regulations or answers.
2. **Recursive Character Text Splitting (Secondary):** If structural chunks are still too large for the LLM's context window (e.g., a very long "Pasal"), recursive character text splitting [85] will be applied within that chunk, using a hierarchy of separators (e.g., paragraph, then sentence) to further subdivide it while attempting to maintain coherence.
3. **Semantic Chunking (Exploratory):** The potential of semantic chunking, using sentence embeddings to group semantically related sentences [83], will be explored for less structured documents or as a refinement step. This will involve using the selected Indonesian embedding model.

Chunk Size and Overlap:
The target chunk size will be determined empirically, aiming for a balance that fits within the SahabatAI model's context window (e.g., Llama3-based models have an 8192 token limit 51, but effective context for RAG is often smaller) and provides sufficient context. An overlap between chunks (e.g., 10-20% of chunk size) will be implemented to ensure that information spanning across chunk boundaries is not lost and can be effectively retrieved.

3.3. Proposed System: "KonsulAI-RAG"

The proposed system, "KonsulAI-RAG," will be an AI-powered Question Answering system designed to provide accurate and contextually relevant answers to queries about Indonesian consular services.

3.3.1. RAG Pipeline Implementation

The RAG pipeline will be implemented using an open-source framework to facilitate development and experimentation.

- **Framework Choice: LlamaIndex** [82] is selected as the primary framework. Its strong focus on data ingestion, indexing, and retrieval specifically for LLM applications, along with its robust support for various data connectors and vector databases, makes it well-suited for this project. It also offers flexibility in integrating custom LLMs like SahabatAI and embedding models. LangChain [82] will be considered as a secondary option or for specific components if LlamaIndex presents limitations for certain tasks.
- **Detailed Workflow:**
  1. **Query Input:** The user submits a query in Bahasa Indonesia.
  2. **Query Embedding:** The user query is converted into a vector embedding using the selected Indonesian embedding model.
  3. **Similarity Search:** The query embedding is used to search the vector database for the most semantically similar document chunk embeddings. The top-k relevant chunks are retrieved.
  4. **Context Reranking (Optional but Recommended):** An additional reranking step may be implemented using a cross-encoder or a simpler relevance scoring mechanism to refine the order of the retrieved chunks, prioritizing the most relevant ones.
  5. **Prompt Augmentation:** The original query and the retrieved (and potentially reranked) context chunks are formatted into a prompt for the LLM.
  6. **LLM Answer Generation:** The augmented prompt is fed to the fine-tuned SahabatAI model, which generates an answer in Bahasa Indonesia based on the provided query and context.
  7. **Answer Output:** The generated answer, along with references to the source document chunks, is presented to the user.

3.3.2. Embedding Model Selection and Implementation

Based on the literature review (Table 2.2), sentence-transformers/paraphrase-multilingual-mpnet-base-v2 [78] is a strong initial

candidate due to its multilingual capabilities including Indonesian and its proven performance in semantic search tasks. However, performance on Indonesian-specific benchmarks from MTEB [79] will be a deciding factor. If a high-performing, readily available Indonesian-specific sentence embedding model (e.g., an IndoBERT sentence variant) is identified with comparable or superior performance on relevant retrieval tasks, it will be prioritized. The chosen model will be used to generate 768-dimensional (or other dimensionality depending on the model) embeddings for all processed document chunks and incoming user queries.

3.3.3. Vector Database Setup and Indexing

For this Master's thesis, an open-source, in-memory or easily deployable vector database will be used. **FAISS (Facebook AI Similarity Search)** is a strong candidate due to its efficiency, wide adoption, and good integration with Python-based frameworks like LlamaIndex. ChromaDB is another lightweight alternative. The document chunk embeddings will be stored in the chosen vector database. An efficient indexing strategy, such as HNSW (Hierarchical Navigable Small World) or IVFADC (Inverted File Index with Product Quantization), will be implemented within FAISS to ensure fast and scalable similarity searches.

3.4. SahabatAI Model Fine-tuning for Consular Q&A

3.4.1. Base SahabatAI Model Selection

The base model for fine-tuning will be an instruction-tuned variant of SahabatAI. Based on availability and performance metrics (Table 2.3), either **Llama3-8B-CPT-SahabatAI-v1-instruct** [52] or **Gemma2-9B-CPT-SahabatAI-v1-instruct** [56] will be selected. The Llama3-based model has a context length of 8192 tokens [51], which is advantageous for RAG. The choice will also consider any specific guidance or newer releases from the SahabatAI team.

3.4.2. Fine-tuning Approach (PEFT)

Given the significant computational resources required for full fine-tuning of models with 8B or 9B parameters, and the constraints of a Master's thesis, a Parameter-Efficient Fine-Tuning (PEFT) approach is essential. **QLoRA (Quantized LoRA)** [89] will be the primary method explored. This involves:

1. Quantizing the pre-trained SahabatAI model to 4-bit precision to reduce its memory footprint.

2. Applying LoRA by adding low-rank adaptation matrices to specific layers (e.g., attention layers) of the quantized model. Only these LoRA weights will be trained. This approach has been shown to make fine-tuning of large LLMs feasible even on consumer-grade GPUs (e.g., Google Colab Tesla T4, which supports float16 and bfloat16 operations necessary for such techniques [91]). The Hugging Face PEFT library will be utilized for implementing QLoRA.[89] Key LoRA hyperparameters such as rank (r), alpha, dropout, and target modules will be configured based on best practices and available tutorials for Llama/Gemma models.[91]

3.4.3. Preparation of Fine-tuning Dataset (Consular Q&A)

A high-quality, domain-specific dataset is crucial for effective fine-tuning. For this project, a consular Q&A dataset in Bahasa Indonesia will be created using the following strategies:

1. **Extraction from MoFA FAQs:** Manually or semi-automatically extract existing question-answer pairs from the FAQ sections of MoFA and embassy websites.[5]
2. **Synthetic Q&A Generation:**
    - Select key consular documents (regulations, procedural guides).
    - Use a base LLM (e.g., the pre-trained SahabatAI-Instruct or another capable model like GPT-3.5 via API if budget allows for a small scale) with carefully engineered prompts to generate potential questions and answers based on segments of these documents. This technique is suggested in [33] for creating domain-specific datasets for RAG.
    - **Manual Verification and Refinement:** All synthetically generated Q&A pairs will be manually reviewed, corrected, and refined by individuals familiar with Indonesian language and, ideally, consular procedures to ensure accuracy, relevance, and naturalness.
3. **Focus:** The dataset will focus on informational and procedural questions relevant to the scoped consular services.
4. **Target Size:** Given the 6-month timeline, a target of a few hundred to a few thousand high-quality, verified Q&A pairs will be aimed for. While the full SahabatAI-Instruct model was fine-tuned on a much larger dataset [52], PEFT methods can achieve good domain adaptation with smaller, highly targeted datasets. The SahabatAI team's call for collaboration on data [50] might also offer avenues for dataset augmentation or sharing.

3.4.4. Training and Hyperparameter Tuning

The fine-tuning process will be conducted using the Hugging Face ecosystem, leveraging libraries such as transformers for model loading and management, PEFT for LoRA/QLoRA

implementation, and trl (Transformer Reinforcement Learning) for supervised fine-tuning (SFT) if applicable.87
Key hyperparameters to be tuned will include:

- LoRA rank (r)
- LoRA alpha
- Learning rate
- Number of training epochs
- Batch size The fine-tuning will be performed on available GPU resources (e.g., Google Colab Pro, or university HPC resources if accessible). Checkpoints will be saved regularly, and the best-performing adapter based on a validation set (a subset of the created Q&A dataset) will be selected.

3.5. Performance Evaluation Plan

A multi-faceted evaluation plan will be implemented to assess the performance of the KonsulAI-RAG system, focusing on both the components of the RAG pipeline and the end-to-end QA capabilities in Bahasa Indonesia.

3.5.1. Intrinsic Evaluation of RAG Components

The effectiveness of the retrieval component is crucial for RAG. This will be evaluated using:

- **Context Precision:** Measures the proportion of retrieved document chunks that are relevant to the input query.
- **Context Recall:** Measures the proportion of all relevant document chunks in the knowledge base that were successfully retrieved for a given query.
- **Context Relevance (or Context Utility):** A more qualitative or LLM-assisted measure of how useful the retrieved context is for answering the query. These metrics are highlighted as important in RAG evaluation frameworks like RAGAS and LlamaIndex's evaluation modules.[31] This will involve creating a small set of representative consular queries and manually annotating the relevance of retrieved chunks from the MoFA corpus.

3.5.2. End-to-End QA Evaluation Metrics (Adapted for Indonesian)

The overall QA performance will be assessed using a combination of metrics, adapted for Bahasa Indonesia:

- **Accuracy-based Metrics:**
  - **Exact Match (EM):** Percentage of generated answers that exactly match a human-generated reference answer.[59]

    - **F1-score:** Calculated based on token-level overlap (precision and recall) between the generated answer and the reference answer, providing a more nuanced measure of accuracy than EM.[59]
- **NLG Metrics:**
    - **ROUGE-L:** Measures the longest common subsequence, useful for evaluating the recall of information in answers that might be summary-like.[59]
    - **BLEU:** Assesses n-gram precision against reference answers.[59]
    - **METEOR:** Considers synonyms and stemming, offering a more semantically aware evaluation than BLEU.[59]
    - *Adaptation Note:* For BLEU and METEOR, Indonesian-specific tokenizers and potentially stemmers/lemmatizers will be necessary for meaningful results.
- **Semantic Similarity:**
    - **Answer Semantic Similarity:** Cosine similarity between the vector embeddings (generated by the chosen Indonesian embedding model) of the system-generated answer and the human-generated reference answer.[59]

3.5.3. Faithfulness and Relevance Assessment (Adapted for Indonesian)

These are critical RAG-specific dimensions:

- **Faithfulness (or Groundedness):** Measures whether the generated answer is factually consistent with and supported by the retrieved MoFA document chunks. This is crucial for minimizing hallucinations.[31] Evaluation will involve:
    - Using RAGAS framework's faithfulness metric, potentially adapted with Indonesian prompts and an LLM (like SahabatAI itself) as the judge.
    - Manual verification of a subset of answers against their cited contexts.
- **Answer Relevance:** Assesses if the generated answer directly, comprehensively, and pertinently addresses the user's query, avoiding irrelevant information.[31] This will also be evaluated using RAGAS metrics and manual review.
- **Negative Rejection/Refusal Capability:** The system's ability to correctly identify and refuse to answer out-of-scope queries or those for which no relevant information exists in the knowledge base is vital for a system dealing with official information.[94] This will be tested with a set of adversarial and out-of-domain queries.

The adaptation of evaluation frameworks like RAGAS [61] for Bahasa Indonesia is a key consideration. While these frameworks provide structured approaches, their default prompts and underlying LLMs for evaluation are often English-centric. For this thesis, adaptation will involve translating evaluation prompts into Indonesian, potentially leveraging SahabatAI itself as an evaluator for aspects like semantic similarity or

checking if an answer is supported by Indonesian context, and being mindful of linguistic nuances that might affect metric calculations. This methodological adaptation could be a valuable contribution of the thesis.

3.5.4. Qualitative Analysis

Quantitative metrics will be supplemented by qualitative analysis:

- **Manual Error Analysis:** A sample of generated answers will be manually reviewed to categorize errors (e.g., irrelevant context retrieved, poor generation quality, factual inaccuracies not caught by automated metrics, issues with Indonesian grammar or phrasing).
- **Expert Review:** If feasible, a small panel of individuals familiar with Indonesian consular procedures will be asked to evaluate the quality, accuracy, and usefulness of answers to a set of representative queries.

3.5.5. Baselines for Comparison

To demonstrate the efficacy of the proposed KonsulAI-RAG system, its performance will be compared against:

1. **Base SahabatAI Model:** The selected instruction-tuned SahabatAI model answering queries directly without RAG (i.e., relying only on its parametric knowledge).
2. **Keyword-Based Search:** A traditional information retrieval system (e.g., using BM25 or TF-IDF) operating over the same MoFA document corpus.
3. **(Optional, if time and resources permit):** A generic, powerful multilingual LLM (e.g., GPT-3.5-turbo via API for a limited set of queries) with a basic RAG setup using the same corpus, to benchmark against a non-Indonesian-centric SOTA model.

Table 3.2: Proposed Evaluation Metrics for the Consular Q&A System

Metric Category	Specific Metric	Description	How it will be Measured/Calculated (Tools/LLMs)	Relevance to Indonesian Consular QA
Retrieval Quality	Context Precision [61]	Proportion of retrieved document	Manual annotation of relevance for a	Ensures that the information provided to the

		chunks that are relevant to the query.	sample set of queries; RAGAS context_precision.	LLM is pertinent, reducing noise and improving the basis for answer generation. Crucial for complex consular regulations.
	Context Recall [61]	Proportion of all relevant document chunks in the knowledge base that were retrieved.	Manual annotation and RAGAS context_recall. Requires defining "all relevant chunks" for sample queries.	Ensures that the system does not miss critical pieces of information from MoFA documents needed to answer a query comprehensively.
Answer Faithfulness	Faithfulness / Groundedness [61]	Degree to which the generated answer is factually consistent with and supported by the retrieved context.	RAGAS faithfulness (may require prompt adaptation for Indonesian and using SahabatAI as judge); Manual verification against source documents for a sample.	Paramount for a system providing official consular information to prevent misinformation and ensure trust. Directly addresses LLM hallucination.
Answer Quality & Relevance	Exact Match (EM) [59]	Percentage of generated answers that are an exact string match to the reference answer.	Direct comparison with human-annotated reference answers.	Useful for queries with concise, factual answers (e.g., "What is the fee for X?").

	F1-score [59]	Harmonic mean of precision and recall at the token level between generated and reference answers.	Standard NLP libraries, using Indonesian tokenization.	Provides a more nuanced measure of accuracy than EM, especially for longer or slightly rephrased answers.
	ROUGE-L [60]	Measures recall of the longest common subsequence between generated and reference answers.	Standard ROUGE scripts, adapted for Indonesian.	Suitable for evaluating how well the system captures key information from reference answers, especially for explanative queries.
	Answer Relevancy [61]	Degree to which the generated answer directly and comprehensively addresses the user's query.	RAGAS answer_relevancy (Indonesian prompt adaptation, SahabatAI as judge); Manual assessment by reviewers.	Ensures that the answers are on-topic, useful, and directly satisfy the user's informational need regarding consular matters.
	Answer Semantic Similarity [61]	Semantic similarity between the generated answer and the reference answer.	Cosine similarity of embeddings from the chosen Indonesian embedding model.	Captures meaning beyond exact wording, useful for evaluating paraphrased or semantically equivalent correct answers.
System Robustness	Negative Rejection [94]	Ability to correctly identify and	Testing with a curated set of out-of-domain	Critical for maintaining system reliability

		refuse to answer out-of-scope or unanswerable (from corpus) queries.	and unanswerable queries; Manual evaluation of refusal appropriateness.	and trust, preventing the generation of speculative or incorrect answers when information is not available in MoFA documents.

3.6. Feasibility and Six-Month Timeline

This research project is designed to be feasible within a standard six-month Master's thesis timeframe. Feasibility is supported by:

- **Leveraging Existing Technologies:** The project will build upon open-source frameworks (LlamaIndex, Hugging Face ecosystem) and a pre-trained Indonesian LLM (SahabatAI), significantly reducing development time from scratch.
- **PEFT for Efficient Fine-tuning:** The use of QLoRA or LoRA for fine-tuning SahabatAI makes the model adaptation process computationally manageable with limited GPU resources.
- **Focused Scope:** The research will concentrate on a well-defined subset of Indonesian consular services, ensuring that data collection, fine-tuning dataset creation, and evaluation can be completed within the allocated time.
- **Availability of Public Data:** The knowledge base will be constructed from publicly accessible MoFA documents, mitigating risks associated with restricted data access.

Table 3.3: Six-Month Thesis Implementation Timeline

Month	Key Tasks/Activities	Deliverables/Milestones	Estimated Duration (Weeks)
1	Finalize detailed literature review; Identify and begin collection of MoFA consular documents; Obtain ethics approval (if required for Q&A dataset involving human	Comprehensive literature review document; Initial corpus of MoFA documents (raw); Ethics approval obtained.	4

	annotation).		
2	Corpus pre-processing (cleaning, structuring); Implement and experiment with document chunking strategies; Select and set up Indonesian embedding model; Set up vector database.	Cleaned and chunked document corpus (v1); Selected embedding model; Functional vector database with indexed corpus.	4
3	Develop initial RAG pipeline with base SahabatAI model; Begin creation of consular Q&A fine-tuning dataset (extraction from FAQs, initial synthetic generation).	Functional baseline RAG system (v1); Draft fine-tuning dataset (Q&A pairs).	4
4	Complete and verify consular Q&A fine-tuning dataset; Fine-tune selected SahabatAI model using PEFT (LoRA/QLoRA); Integrate fine-tuned model into RAG pipeline.	Verified fine-tuning dataset; Fine-tuned SahabatAI adapter; KonsulAI-RAG system with fine-tuned LLM (v2).	4
5	Conduct rigorous system evaluation: intrinsic RAG component testing, end-to-end QA metrics, faithfulness & relevance assessment, qualitative analysis;	Preliminary evaluation results and analysis report; System improvements implemented.	4

	Perform error analysis; Iterative improvements to system based on evaluation.		
6	Finalize all experiments and result interpretation; Complete thesis writing (introduction, literature review, methodology, results, discussion, conclusion); Prepare final report and presentation.	Draft thesis submitted; Final thesis report and presentation prepared.	4

This timeline allocates sufficient time for each critical phase, including potential iterations based on evaluation outcomes, making the project achievable within the six-month constraint.

Chapter 4: Expected Results and Contributions

4.1. Anticipated Performance of the Proposed System

The development of the "KonsulAI-RAG" system is expected to yield significant improvements in accessing and understanding Indonesian consular information. It is anticipated that the system will demonstrate high accuracy and F1-scores for informational queries within the defined scope of consular services, particularly for questions that can be directly answered from the MoFA document corpus.

A key expectation is a substantial enhancement in the **faithfulness** and **answer relevance** of the generated responses compared to baseline systems, such as a non-RAG SahabatAI model. This improvement will stem directly from the RAG architecture's ability to ground answers in specific, verifiable information retrieved from official MoFA documents, thereby mitigating the risk of LLM "hallucinations".[31]

The fine-tuning of SahabatAI using a domain-specific consular Q&A dataset is expected to further refine its capabilities. This adaptation should lead to a better understanding of the nuances of Bahasa Indonesia as used in the consular context, including specific terminologies and common query patterns. Consequently, the

fine-tuned model integrated within the RAG pipeline should produce answers that are not only factually accurate but also more contextually appropriate and natural-sounding for users seeking consular information.

While high performance is anticipated, the evaluation is also expected to identify specific challenges and limitations. These might include difficulties in handling highly dynamic information if the corpus is not frequently updated (though real-time updates are out of scope for this thesis), potential ambiguities in interpreting certain complex consular regulations even with RAG, or limitations in the breadth of knowledge if certain niche topics are not well-covered in the collected documents. The system's ability to handle out-of-scope queries through effective negative rejection will also be a critical outcome to assess.

4.2. Contribution to the Field of AI

This research is poised to make several contributions to the broader field of Artificial Intelligence, particularly in NLP and its applications:

- **RAG for Low-Resource and Domain-Specific Languages:** The project will provide valuable insights and a practical implementation of RAG for Bahasa Indonesia. While RAG is well-studied for English, its application and optimization for languages with fewer readily available NLP resources and benchmarks, especially within specific governmental domains, are less explored. This work will contribute to understanding the challenges and effective strategies for deploying RAG in such contexts.
- **Fine-tuning Indonesian LLMs for Specialised Domains:** By demonstrating the application of Parameter-Efficient Fine-Tuning (PEFT) techniques like LoRA or QLoRA to SahabatAI for the consular domain, this research will contribute to the growing body of knowledge on customizing Indonesian-centric LLMs. It will offer practical insights into dataset creation, fine-tuning methodologies, and the performance gains achievable for specialized tasks, which is crucial for advancing the utility of local LLMs like SahabatAI.[50]
- **Methodology for Domain-Specific QA Systems:** The thesis will present a comprehensive case study and methodology for building specialized QA systems in knowledge-intensive and complex domains like consular services. The approaches to data collection from official governmental sources, document chunking strategies tailored for regulatory texts, and the integration of RAG with a localized LLM can serve as a model for similar projects in other domains.
- **Evaluation Strategies in the Indonesian Context:** A significant part of the research involves adapting and applying SOTA RAG evaluation metrics (e.g., from

frameworks like RAGAS [61]) for Bahasa Indonesia. This may involve translating prompts, using SahabatAI as an evaluation model, and addressing linguistic nuances. The findings and methodologies developed for evaluation could contribute to establishing best practices for assessing RAG systems in non-English, particularly Indonesian, contexts.

4.3. Contribution to Government Services

The successful implementation of the KonsulAI-RAG system has the potential to significantly benefit Indonesian government services, particularly those provided by the Ministry of Foreign Affairs:

- **Enhanced Information Accessibility and Quality:** The primary contribution will be to provide Indonesian citizens, foreign nationals, and MoFA staff with a more efficient, accurate, and readily accessible means of obtaining consular information.[40] By grounding answers in official documents, the system can ensure consistency and reliability, reducing confusion and the effort required to find specific details within complex regulations.
- **Improved Public Service Efficiency:** By automating responses to a wide range of common and informational queries, the system can potentially reduce the workload on consular staff. This would free up human resources to focus on more complex cases requiring nuanced judgment or direct intervention, thereby improving overall service efficiency and responsiveness. This aligns directly with MoFA's strategic objectives for service excellence and WNI protection as outlined in their Renstra documents.[20]
- **Support for Digital Transformation and AI Sovereignty:** This project serves as a practical example of AI implementation within the Indonesian public sector. By leveraging an Indonesian-developed LLM like SahabatAI, it contributes to the nation's goals of achieving "AI Sovereignty" and advancing its digital capabilities as envisioned in the "Golden Indonesia 2045" plan.[53]
- **Blueprint for AI Adoption in the Indonesian Public Sector:** A successful KonsulAI-RAG system can act as a powerful case study and demonstrator for other Indonesian government agencies looking to leverage AI for service improvement. The technical challenges addressed (e.g., processing official documents in Bahasa Indonesia, ensuring factual accuracy, adapting evaluation methods) and the solutions developed will be highly relevant to other public sector domains, such as legal information dissemination, public health advisories, and tax regulation inquiries. The project's alignment with national AI strategies further amplifies its potential as a model for broader AI adoption and capacity building within the Indonesian government, potentially influencing future

investments and development in AI for public services.

4.4. Potential Challenges and Mitigation Strategies

Several potential challenges are anticipated in this research, along with strategies to mitigate them:

- **Data Availability and Quality:**
  - *Challenge:* Difficulty in obtaining a comprehensive, clean, and consistently machine-readable corpus of MoFA consular documents. Some documents may be scanned PDFs requiring OCR, or information might be scattered across multiple websites with varying formats.[63]
  - *Mitigation:* The project will initially focus on publicly available data from official MoFA and key embassy websites. Robust pre-processing scripts will be developed to handle various formats and clean the text. The scope of consular services covered will be adaptable based on the quality and quantity of data successfully collected and processed.
- **Fine-tuning Dataset Creation and Complexity:**
  - *Challenge:* Creating a high-quality, domain-specific Q&A dataset in Bahasa Indonesia for fine-tuning SahabatAI can be time-consuming and requires domain understanding.
  - *Mitigation:* A hybrid approach will be used: extracting existing Q&A pairs from MoFA FAQs and using a base LLM (with careful prompt engineering) to generate initial Q&A pairs from regulations, followed by rigorous manual verification and refinement. The focus will be on a smaller, high-impact dataset suitable for PEFT. The SahabatAI team's call for data collaboration might also be explored.[50]
- **Evaluation Nuances for Bahasa Indonesia:**
  - *Challenge:* Standard RAG evaluation metrics and tools are often English-centric. Adapting and validating these for Bahasa Indonesia, considering its linguistic characteristics, can be complex.
  - *Mitigation:* Established frameworks like RAGAS will be used as a starting point. Evaluation prompts will be carefully translated and adapted. The fine-tuned SahabatAI model itself may be leveraged as an LLM-based evaluator for certain metrics (e.g., semantic similarity, faithfulness checks in Indonesian). Qualitative analysis and manual error checking will be crucial supplements to quantitative metrics.
- **Computational Resources:**
  - *Challenge:* Fine-tuning and experimenting with large LLMs like SahabatAI (8B/9B parameters) typically require significant GPU resources, which may be

limited for a Master's thesis.
    - *Mitigation:* The primary fine-tuning approach will be QLoRA, which is designed for resource efficiency and can run on consumer-grade GPUs.[89] Cloud-based GPU resources (e.g., Google Colab Pro) will be utilized for short, intensive training runs. Experiments will be carefully scoped to fit within available computational budgets.
- **Handling Ambiguity and Out-of-Scope Queries:**
    - *Challenge:* Consular regulations can be inherently complex and sometimes ambiguous. Users may also ask queries that are outside the scope of the system's knowledge base.
    - *Mitigation:* The RAG system will be designed to primarily provide information *based only on the provided MoFA documents*. Clear disclaimers about the system's scope will be important. Robust negative rejection capabilities will be developed and evaluated to ensure the system can gracefully decline to answer questions for which it lacks relevant, authoritative information, rather than generating speculative or incorrect responses.

Chapter 5: Conclusion

This thesis proposal outlines a research project to develop "KonsulAI-RAG," an advanced Question Answering system designed to enhance Indonesian consular services. The core problem addressed is the difficulty users face in accessing accurate and comprehensive information from the Indonesian Ministry of Foreign Affairs (MoFA) due to the complexity and volume of consular regulations and procedures. The proposed solution leverages Retrieval-Augmented Generation (RAG) to ground answers in official MoFA documents, combined with a fine-tuned SahabatAI Large Language Model, which is specifically designed for Bahasa Indonesia and its nuances.

The methodology encompasses several key stages: (1) the creation of a comprehensive knowledge base from official MoFA consular documents, employing optimized document chunking strategies suitable for governmental texts; (2) the selection and fine-tuning of an appropriate SahabatAI model using Parameter-Efficient Fine-Tuning (PEFT) techniques like QLoRA to adapt it to the consular domain; (3) the design and implementation of a RAG pipeline using the LlamaIndex framework, integrating the fine-tuned SahabatAI and a suitable Indonesian embedding model; and (4) a rigorous performance evaluation using a combination of automated metrics (accuracy, F1-score, ROUGE, BLEU, METEOR, and RAG-specific metrics like faithfulness and relevance, all adapted for Indonesian) and qualitative analysis, benchmarked against baseline systems. The project is designed

to be feasible within a six-month Master's thesis timeframe.

The expected contributions of this research are multifaceted. For the field of AI, it will provide insights into applying RAG to low-resource languages like Bahasa Indonesia, demonstrate effective PEFT strategies for local LLM customization (SahabatAI), offer a case study for domain-specific QA in complex governmental contexts, and potentially contribute to the adaptation of RAG evaluation metrics for Indonesian. For Indonesian government services, the KonsulAI-RAG system aims to significantly enhance information accessibility and service quality for consular matters, supporting MoFA's digital transformation goals and contributing to the broader vision of national AI sovereignty.

Potential future work beyond the scope of this Master's thesis could include expanding the system to cover a wider range of consular services, incorporating multilingual capabilities to serve diverse user groups, developing a user-friendly interface for deployment within MoFA's existing digital platforms (like Portal Peduli WNI or a successor to the SARI chatbot), and implementing mechanisms for real-time or periodic updates to the knowledge base to ensure the information provided remains current. This research lays a strong foundation for such future developments, showcasing the transformative potential of localized LLMs and RAG in modernizing public services.

REFERENCES

- [13] Travel Canada. (n.d.). *Indonesia*. Retrieved from travel.gc.ca
- [14] U.S. Department of State. (2023). *2023 Country Reports on Human Rights Practices: Indonesia*. Retrieved from state.gov
- [50] Sahabat AI. (n.d.). *Sahabat-AI: The First Large Language Model in Bahasa Indonesia*. Retrieved from sahabat-ai.com
- [51] GoToCompany. (n.d.). *Llama3-8B-CPT-SahabatAI-v1-base*. Hugging Face. Retrieved from huggingface.co
- [52] GoToCompany. (n.d.). *Llama3-8B-CPT-SahabatAI-v1-instruct*. Hugging Face. Retrieved from huggingface.co
- [57] Supa.so Blog. (2025, February 21). *Evaluating LLMs for Bahasa Indonesia: SEA-LIONv3 vs SahabatAI-v1*.
- [31] Deepset AI. (n.d.). *Generative Question Answering*. Retrieved from docs.cloud.deepset.ai
- [35] Amazon Web Services. (n.d.). *What is Retrieval-Augmented Generation?* Retrieved from aws.amazon.com
- [76] Comtrac. (n.d.). *Navigating the risks of public LLMs: Why government agencies*

need safe, targeted AI solutions for their investigations. Retrieved from comtrac.com.au

- [34] Li, J., Liu, S., Li, Z., Dong, H., Zhang, T., & Zhang, Y. (2024). Large Language Models in Primary Health Care: Opportunities and Challenges. *Journal of Medical Internet Research, 26*, e11960148. [PMC11960148]
- [15] CPT Corporate. (n.d.). *What To Do If Your Indonesian Visa Application is Rejected*. Retrieved from cptcorporate.com
- [16] TraveloBiz. (2025, May 15). *Indonesia Updates Multiple-Entry Visa Rules: Stay Up to 180 Days Without Exit*.
- [53] The Fast Mode. (n.d.). *Indosat Ooredoo Hutchison, GoTo Launch Indonesian LLM Sahabat-AI*. Retrieved from thefastmode.com
- [54] The Jakarta Post. (2024, November 18). *AI Day drives Indonesia's AI sovereignty with Indosat leading solutions and connectivity*.
- [36] NVIDIA. (n.d.). *Retrieval-Augmented Generation*. Retrieved from nvidia.com/en-us/glossary
- [81] Writer Engineering. (n.d.). *RAG vector databases: The good, the bad, and the graph-based alternative*. Retrieved from writer.com
- [87] DataCamp. (n.d.). *Fine-tuning Large Language Models*. Retrieved from datacamp.com
- [88] Machine Learning Mastery. (n.d.). *Custom Fine-tuning for Domain-Specific LLMs*. Retrieved from machinelearningmastery.com
- [99] BytePlus. (n.d.). *Natural Language Processing in Indonesia: BytePlus ModelArk's Innovations*. Retrieved from byteplus.com
- [100] BytePlus. (n.d.). *Revolutionizing Manufacturing: Natural Language Processing's Impact in Indonesia*. Retrieved from byteplus.com
- [1] Embassy of the Republic of Indonesia in Washington D.C. (n.d.). *Contact Us*. Retrieved from kemlu.go.id/washington
- [2] Directorate General of Immigration, Republic of Indonesia. (n.d.). *Official Website*. Retrieved from imigrasi.go.id
- [3] Kementerian Luar Negeri RI. (n.d.). *Kontak Kami*. Retrieved from fe-non-production.apps.opppd2-dev.layanan.go.id/kontak
- [4] Wikipedia. (n.d.). *Ministry of Foreign Affairs (Indonesia)*.
- [5] Embassy of Indonesia in The Hague. (n.d.). *FAQ*. Retrieved from indonesia.nl/en/faq
- [23] U.S. Department of State - Special Issuance Agency. (n.d.). *Official/Diplomatic Visa Information: Indonesia*. Retrieved from travel.state.gov
- [17] VOA Indonesia. (n.d.). *Pendataan Jadi Tantangan Besar dalam Perlindungan WNI di Luar Negeri*. Retrieved from voaindonesia.com
- [24] U.S. Embassy & Consulates in Indonesia. (n.d.). *What You'll Need to Bring (for*

Notarial Services). Retrieved from id.usembassy.gov

- [25] Embassy of Indonesia, Washington D.C. Consular Section. (n.d.). *Document Legalization*. Retrieved from consular.embassyofindonesia.org
- [92] DataCamp. (n.d.). *Fine-Tune Gemma 3: A Step-by-Step Guide With Financial Q&A*.
- [91] Unsloth AI. (n.d.). *Tutorial: How to Run & Fine-tune Gemma 3*. Retrieved from docs.unsloth.ai
- [77] GoToCompany. (n.d.). *Sahabat-AI Collection*. Hugging Face. Retrieved from huggingface.co
- [52] GoToCompany. (n.d.). *Llama3-8B-CPT-SahabatAI-v1-instruct - Model Card*. Hugging Face.
- [50] Sahabat AI. (n.d.). *Sahabat-AI Website Details*. Retrieved from sahabat-ai.com
- [55] Dataloop AI. (n.d.). *Model Library: gotocompany/llama3-8b-cpt-sahabatai-v1-instruct*.
- [56] Dataloop AI. (n.d.). *Model Library: gotocompany/gemma2-9b-cpt-sahabatai-v1-base*.
- [58] Anonymous. (2025). *LUSIFER: A Robust and Efficient Framework for Zero-Shot Multilingual Embedding*. arXiv:2501.00874v1.
- [37] Anonymous. (2025). *LUSIFER: A Robust and Efficient Framework for Zero-Shot Multilingual Embedding*. arXiv:2501.00874v3.
- [85] F22 Labs. (n.d.). *7 Chunking Strategies in RAG You Need to Know*. Retrieved from f22labs.com
- [83] Analytics Vidhya. (2024, October). *Chunking Techniques to Build Exceptional RAG Systems*.
- [82] Firecrawl.dev. (n.d.). *Best Open-Source RAG Frameworks*.
- [86] IBM. (n.d.). *LlamaIndex vs. LangChain: Choosing the Right RAG Framework*. Retrieved from ibm.com/think
- [38] Anonymous. (2025). *Agentic Generative AI with RAG, Knowledge Graphs, and Vector Stores for Legal Systems*. arXiv:2502.20364v1.
- [39] Pipitone, A., & Alami, O. (2024). *LegalBench-RAG: A Benchmark for Evaluating Retrieval in Legal RAG Systems*. arXiv:2408.10343.
- [40] StateTech Magazine. (2025, February). *What Is Retrieval Augmented Generation, and How Are State and Local Agencies Using It?*
- [45] Anonymous. (2025). *GraphRAG for Policy Proposal Generation from YouTube Logistics Data*. *Electronics, 14*(7), 1241. MDPI.
- [41] Anonymous. (2024). *ToG-2: A Hybrid RAG Approach Integrating Knowledge Graphs and Unstructured Data*. arXiv:2407.10805v7.
- [42] Anonymous. (2025). *Agent-Guided Iterative Retrieval for Multi-Step Question Answering*. arXiv:2503.13275v2.

- [46] Anonymous. (2025). *Pubbie: An Intelligent Agent for Performance Measurement at NRC Canada*. arXiv:2504.10497.
- [47] Anonymous. (2023). *Comparative Analysis of Government Chatbots and Large Language Models*. arXiv:2312.02181.
- [43] Anonymous. (2025). *Safety of RAG-based LLMs*. *Proceedings of NAACL 2025*. ACL Anthology.
- [44] Li, Z., et al. (2024). Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. *Proceedings of EMNLP 2024: Industry Track*.
- [32] IBM. (n.d.). *Question Answering*. Retrieved from ibm.com/think
- [33] Anonymous. (2025). *QuIM-RAG: Question-to-question Inverted Index Matching for RAG Systems*. arXiv:2501.02702v1.
- [28] Kedutaan Besar Republik Indonesia di London. (n.d.). *Rencana Strategis KBRI London 2020-2024*. Retrieved from kemlu.go.id
- [29] Kedutaan Besar Republik Indonesia di London. (n.d.). *Rencana Strategis KBRI London 2015-2019*. Retrieved from kemlu.go.id
- [20] Kementerian Luar Negeri Republik Indonesia. (2020). *Rencana Strategis Kementerian Luar Negeri Tahun 2020-2024*. Retrieved from e-ppid.kemlu.go.id
- [30] Direktorat Asia Timur dan Pasifik, Kementerian Luar Negeri RI. (2021). *Rencana Strategis 2020-2024*.
- [6] Layanan Diplomatik Kemlu. (n.d.). *FAQ*. Retrieved from layanandiplomatik.kemlu.go.id
- [7] Kementerian Luar Negeri RI. (n.d.). *FAQ*. Retrieved from kemlu.go.id/faq
- [27] Tempo.co. (2025, February 14). *Kementerian Luar Negeri akan Gunakan AI untuk Pelayanan WNI di Luar Negeri*.
- [26] Scribd. (n.d.). *Renstra Kemlu 2020-2024 (Excerpt)*.
- [18] Jurnal Normatif Al-Azhar. (n.d.). *Perlindungan WNI di Luar Negeri Selama Pandemi Covid-19*.
- [19] Jurnal Hubungan Internasional Unair. (2020). *Tata Kelola Perlindungan Warga Negara Indonesia dalam Melakukan Peran Diplomasi Digital*.
- [22] Peraturan Menteri Luar Negeri Republik Indonesia Nomor 5 Tahun 2018 tentang Pelindungan Warga Negara Indonesia di Luar Negeri.
- [21] Direktorat Perlindungan WNI dan BHI, Kementerian Luar Negeri RI. (2020). *Rencana Strategis Direktorat PWNI BHI 2020-2024*.
- [8] KBRI Lima. (n.d.). *Legalisasi Dokumen*. Retrieved from kemlu.go.id/lima
- [9] KBRI Mexico City. (2024). *Legalisasi Dokumen*. Retrieved from kemlu.go.id
- [10] KBRI Den Haag. (2024). *Legalisasi Dokumen Warga Negara Indonesia*. Retrieved from kemlu.go.id/denhaag
- [11] KBRI Mexico City. (2024). *Informasi Visa Dinas dan Diplomatik*. Retrieved from

kemlu.go.id
- [74] Peraturan Menteri Luar Negeri Republik Indonesia Nomor 14 Tahun 2022 tentang Tata Cara Legalisasi Dokumen.
- [75] Peraturan Menteri Luar Negeri Republik Indonesia Nomor 13 Tahun 2019 tentang Tata Cara Legalisasi Dokumen.
- [7] Kementerian Luar Negeri RI. (n.d.). *FAQ (General Consular Services)*. Retrieved from kemlu.go.id/faq
- [12] KJRI Los Angeles. (n.d.). *Perlindungan WNI*. Retrieved from kemlu.go.id/losangeles
- [50] Sahabat AI. (2025, May 15). *Sahabat-AI Collaboration and Model Access*.
- [51] Hugging Face. (n.d.). *Llama3-8B-CPT-SahabatAI-v1-base Model Card Details*.
- [52] Hugging Face. (n.d.). *Llama3-8B-CPT-SahabatAI-v1-instruct Model Card Details*.
- [31] Deepset AI. (n.d.). *RAG QA Limitations and Evaluation*.
- [89] Hugging Face Blog. (2023, March 9). *Fine-tuning LLMs with TRL and PEFT*.
- [78] Hugging Face. (n.d.). *sentence-transformers/paraphrase-multilingual-mpnet-base-v2 Model Card*.
- [84] Pinecone. (2023, June 30). *Advanced Chunking Strategies for RAG*.
- [93] arXiv. (2023). *Quranic Semantic Search using RAG (Metrics Discussion)*. arXiv:2311.05120.
- [20] Kementerian Luar Negeri RI. (2020, December 7). *Rencana Strategis Kementerian Luar Negeri 2020-2024 (Details on WNI Protection and Digital Tools)*.
- [90] Hugging Face PEFT Documentation. (2024, May 15). *Conceptual Guides: LoRA*.
- [20] Kementerian Luar Negeri RI. (2020, December 7). *Rencana Strategis Kementerian Luar Negeri 2020-2024 (Challenges, Digital Strategy, Service Targets)*.
- [6] Layanan Diplomatik Kemlu. (n.d.). *FAQ (Stay Permit Services)*.
- [27] Tempo.co. (2025, February 14). *SARI Chatbot and Existing MoFA Digital Platforms*.
- [59] arXiv. (2023). *Evaluating Large Language Models: A Comprehensive Survey (Metrics for QA, Faithfulness, Relevance)*. arXiv:2307.03109.
- [27] Tempo.co. (2025, February 14). *MoFA's AI Plans for WNI Services (SARI Chatbot)*.
- [78] Hugging Face. (n.d.). *Indonesian Language Performance of paraphrase-multilingual-mpnet-base-v2*.
- [60] arXiv. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Evaluation Metrics)*. arXiv:2005.11401.
- [61] Ragas Documentation. (2024, February 8). *Evaluating RAG Pipelines with*

LlamaIndex.

- [22] Peraturan Menteri Luar Negeri Republik Indonesia Nomor 5 Tahun 2018. (2018, September 19). *Consular Protection for WNI, Case Handling, Portal Peduli WNI*.
- [21] Direktorat Perlindungan WNI dan BHI, Kementerian Luar Negeri RI. (2020, December). *Strategic Plan 2020-2024 (Challenges, Digital Strategy, Service Targets)*.
- [22] Peraturan Menteri Luar Negeri Republik Indonesia Nomor 5 Tahun 2018. (2018, September 19). *WNI Cases, Portal Peduli WNI, Service Challenges*.
- [21] Direktorat Perlindungan WNI dan BHI, Kementerian Luar Negeri RI. (2020, December). *Strategic Plan 2020-2024 (WNI Protection Challenges, Digital Strategy, Service Targets)*.
- [62] Ragas Documentation. (2025, January 23). *RAGAS Framework Evaluation Metrics*.
- [62] Ragas Documentation. (2025, January 23). *Adapting RAGAS Metrics for Non-English Languages*.

Note: Some snippet IDs were not directly citable as they were inaccessible website links.[48] Information from these was integrated based on the summary provided in the outline where applicable, or the outline indicated they were inaccessible.

APPENDICES

A.1. Sample Structure of MoFA Consular Documents

- **Permenlu (Ministerial Regulation) - Example: Permenlu No. 5 Tahun 2018 tentang Pelindungan WNI di Luar Negeri [22]:**
  - BAB I: Ketentuan Umum (General Provisions - Definitions)
  - BAB II: Pelindungan WNI (WNI Protection - Principles, Scope)
  - BAB III: Bentuk Pelindungan (Forms of Protection - Consular, Diplomatic, Legal Aid)
  - BAB IV: Penanganan Keadaan Darurat (Emergency Handling)
  - BAB V: Kelembagaan (Institutional Framework - Central, Representatives, Integrated Teams)
  - BAB VI: Sumber Daya Manusia (Human Resources)
  - BAB VII: Sistem Informasi (Information Systems - Portal Peduli WNI, Safe Travel)
  - BAB VIII: Pendanaan (Funding)
  - BAB IX: Ketentuan Penutup (Closing Provisions)
  - *Each "BAB" is typically divided into "Pasal" (Article) and "Ayat" (Clause).*
- **FAQ Document - Example: kemlu.go.id/faq [7]:**

    - Structured as a list of questions followed by direct answers.
    - Covers various topics like types of consular services, fees, document legalization procedures, passport application processes.
- **Service Procedure Description - Example: Legalisasi Dokumen KBRI Lima [8]:**
    - Outlines general requirements (e.g., personal appearance, forms, ID).
    - Details specific requirements for different document types (e.g., translated documents, certified copies, power of attorney).
    - Specifies procedures before visiting the embassy and at the embassy.
    - Lists tariffs for different services.

A.2. Detailed Project Timeline Gantt Chart
(A visual Gantt chart would be inserted here in a full thesis, detailing tasks, sub-tasks, durations, and dependencies for each of the 6 months as outlined in Table 3.3.)

A.3. List of Acronyms

Acronym	Full Form
AI	Artificial Intelligence
ANN	Approximate Nearest Neighbors
Dit PWNI & BHI	Direktorat Perlindungan Warga Negara Indonesia dan Badan Hukum Indonesia
EM	Exact Match
FAQ	Frequently Asked Questions
FAISS	Facebook AI Similarity Search
GPU	Graphics Processing Unit
HNSW	Hierarchical Navigable Small World
IKU	Indikator Kinerja Utama (Main Performance Indicator)
Kemlu	Kementerian Luar Negeri (Ministry of Foreign Affairs)

KG	Knowledge Graph
k-NN	k-Nearest Neighbors
LAKIP	Laporan Akuntabilitas Kinerja Instansi Pemerintah (Government Agency Performance Accountability Report)
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MoFA	Ministry of Foreign Affairs
MTEB	Massive Text Embedding Benchmark
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OCR	Optical Character Recognition
PEFT	Parameter-Efficient Fine-Tuning
Permenlu	Peraturan Menteri Luar Negeri (Minister of Foreign Affairs Regulation)
QA	Question Answering
QLoRA	Quantized LoRA
RAG	Retrieval-Augmented Generation
Renstra	Rencana Strategis (Strategic Plan)
SARI	Sahabat Artifisial Migran Indonesia

SOTA	State-of-the-Art
WNI	Warga Negara Indonesia (Indonesian Citizen)

A.4. Preliminary System Architecture Diagram
(A diagram illustrating the components of the KonsulAI-RAG system: User Interface -> Query Processor -> Embedding Model -> Vector Database (with MoFA Corpus) -> Retriever -> Reranker (Optional) -> Prompt Formatter -> Fine-tuned SahabatAI LLM -> Answer Generator -> User Interface. Arrows would indicate data flow.)

Works cited

1. The Embassy of the Republic of Indonesia - Washington, accessed May 15, 2025, https://kemlu.go.id/washington/tentang-perwakilan/kontak-kami
2. Welcome to the Indonesian Directorate General of Immigration Website - Direktorat Jenderal Imigrasi, accessed May 15, 2025, https://www.imigrasi.go.id/index?lang=en-US
3. Kementerian Luar Negeri RI - Portal Kemlu, accessed May 15, 2025, https://fe-non-production.apps.opppd2-dev.layanan.go.id/kontak
4. Ministry of Foreign Affairs (Indonesia) - Wikipedia, accessed May 15, 2025, https://en.wikipedia.org/wiki/Ministry_of_Foreign_Affairs_(Indonesia)
5. FAQ | Indonesia - KBRI Den Haag, accessed May 15, 2025, https://indonesia.nl/en/faq/
6. Layanan Diplomatik Terpadu Satu Pintu - Ditjen Protkons Kemlu RI, accessed May 15, 2025, https://layanandiplomatik.kemlu.go.id/guest/faq
7. FAQ - Portal Kemlu, accessed May 15, 2025, https://kemlu.go.id/faq
8. Legalisasi Dokumen - Lima, accessed May 15, 2025, https://kemlu.go.id/lima/pelayanan-perwakilan/legalisasi-dokumen-
9. KEDUTAAN BESAR REPUBLIK INDONESIA MEXICO CITY FUNGSI PROTOKOL DAN KONSULER LEGALISASI DOKUMEN Legalisasi adalah pengesahan tand, accessed May 15, 2025, https://kemlu.go.id/files-service/storage/submenu/additional_file/17297201306719 6f4219a4c_2024__LEGALISASI__Dokumen__Ind_.pdf
10. Legalisasi Dokumen Warga Negara Indonesia - Den Haag, accessed May 15, 2025, https://kemlu.go.id/denhaag/pelayanan-perwakilan/pelayanan-konsuler/legalisasi-dokumen-warga-negara-indonesia-
11. KEDUTAAN BESAR REPUBLIK INDONESIA MEXICO CITY FUNGSI PROTOKOL DAN KONSULER PERSYARATAN PEMBUATAN VISA DINAS DAN DIPLOMATIK Peme, accessed May 15, 2025, https://kemlu.go.id/files-service/storage/submenu/additional_file/1729875925671b cfd597a5b_2024__VISA__Informasi_Visa_Official___Diplomatik.pdf
12. Perlindungan WNI - Los Angeles - Portal Kemlu, accessed May 15, 2025, https://kemlu.go.id/losangeles/perlindungan-wni

13. Travel advice and advisories for Indonesia - Travel.gc.ca, accessed May 15, 2025, https://travel.gc.ca/destinations/indonesia
14. Indonesia - United States Department of State, accessed May 15, 2025, https://www.state.gov/reports/2023-country-reports-on-human-rights-practices/indonesia/
15. What To Do If Your Indonesian Visa Application is Rejected - cptcorporate, accessed May 15, 2025, https://cptcorporate.com/what-to-do-if-your-indonesian-visa-application-is-rejected/
16. Indonesia Updates Multiple-Entry Visa Rules: Stay Up to 180 Days Without Exit - travelobiz, accessed May 15, 2025, https://travelobiz.com/indonesia-multiple-entry-visa-update-180-day-stay-without-exit/
17. Pendataan Jadi Tantangan Besar dalam Perlindungan WNI di Luar Negeri - VOA Indonesia, accessed May 15, 2025, https://www.voaindonesia.com/a/pendataan-jadi-tantangan-besar-dalam-perlindungan-wni-di-luar-negeri/6016183.html
18. PERAN KEMENTERIAN LUAR NEGERI TERHADAP PERLINDUNGAN WARGA NEGARA INDONESIA DI LUAR NEGERI PADA MASA PANDEMI COVID-19, accessed May 15, 2025, https://jurnal.alazhar-university.ac.id/index.php/normatif/article/download/174/168
19. Tata Kelola Perlindungan Warga Negara Indonesia dalam Melakukan Peran Diplomasi Digital - Journal of Universitas Airlangga, accessed May 15, 2025, https://e-journal.unair.ac.id/JHI/article/download/17996/10686/73865
20. e-ppid.kemlu.go.id, accessed May 15, 2025, https://e-ppid.kemlu.go.id/storage/619/Renstra-Kemlu-2020-2024.pdf
21. kemlu.go.id, accessed May 15, 2025, https://kemlu.go.id/files-service/storage/repositori/65391/01.%20Renstra%20Dit%20PWNI%20BHI%202020-2024%20(FINAL).pdf
22. peraturan.bpk.go.id, accessed May 15, 2025, https://peraturan.bpk.go.id/Download/130822/Permenlu%20No.%205%20Tahun%202018.pdf
23. Indonesia - Travel.gov - U.S. Department of State, accessed May 15, 2025, https://travel.state.gov/content/special-issuance-agency-home/en/spec-issuance-agency/official-diplomatic-visa-information/indonesia.html
24. What you'll need to bring - U.S. Embassy & Consulates in Indonesia, accessed May 15, 2025, https://id.usembassy.gov/what-youll-need-to-bring/
25. Document Legalization - e-Consular Service KBRI WDC, accessed May 15, 2025, https://consular.embassyofindonesia.org/page/documentlegalization.html
26. Renstra Kemlu 2020 - 2024 | PDF | Bisnis | Pengelolaan Keuangan & Uang - Scribd, accessed May 15, 2025, https://id.scribd.com/document/539794959/Renstra-Kemlu-2020-2024
27. Kementerian Luar Negeri akan Gunakan AI untuk Pelayanan WNI di ..., accessed May 15, 2025, https://www.tempo.co/internasional/kementerian-luar-negeri-akan-gunakan-ai-u

ntuk-pelayanan-wni-di-luar-negeri-1207028
28. 0 RENCANA STRATEGIS (RENSTRA) KEDUTAAN BESAR REPUBLIK INDONESIA LONDON, INGGRIS TAHUN 2020 - Portal Kemlu, accessed May 15, 2025, https://kemlu.go.id/files/submenu/additional_file/4355-files-statis-pages-perwakilan-London-href-contains-pdf-5-shared.pdf
29. KEDUTAAN BESAR REPUBLIK INDONESIA LONDON KEPUTUSAN KEPALA PERWAKILAN REPUBLIK INDONESIA UNTUK KERAJAAN INGGRIS MERANGKAP REPUBLI, accessed May 15, 2025, https://kemlu.go.id/files/submenu/additional_file/4355-files-statis-pages-perwakilan-London-href-contains-pdf-10-shared.pdf
30. RENCANA STRATEGIS 2020-2024, accessed May 15, 2025, https://kemlu.go.id/files/repositori/65386/Renstra%20Dit.%20Astimpas%202020%20-%202024.pdf
31. Retrieval Augmented Generation (RAG) Question Answering, accessed May 15, 2025, https://docs.cloud.deepset.ai/docs/generative-question-answering
32. What is Question Answering? | IBM, accessed May 15, 2025, https://www.ibm.com/think/topics/question-answering
33. QuIM-RAG: Advancing Retrieval-Augmented Generation with Inverted Question Matching for Enhanced QA Performance - arXiv, accessed May 15, 2025, https://arxiv.org/html/2501.02702v1
34. Opportunities and Challenges for Large Language Models in Primary Health Care - PMC, accessed May 15, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11960148/
35. What is RAG? - Retrieval-Augmented Generation AI Explained - AWS, accessed May 15, 2025, https://aws.amazon.com/what-is/retrieval-augmented-generation/
36. What is Retrieval-Augmented Generation (RAG)? | NVIDIA, accessed May 15, 2025, https://www.nvidia.com/en-us/glossary/retrieval-augmented-generation/
37. LUSIFER: Language Universal Space Integration for Enhanced Multilingual Embeddings with Large Language Models - arXiv, accessed May 15, 2025, https://arxiv.org/html/2501.00874v3
38. Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization - arXiv, accessed May 15, 2025, https://arxiv.org/html/2502.20364v1
39. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain - arXiv, accessed May 15, 2025, http://arxiv.org/pdf/2408.10343
40. What Is Retrieval Augmented Generation, and How Are State and Local Agencies Using It?, accessed May 15, 2025, https://statetechmagazine.com/article/2025/02/what-is-rag-perfcon
41. Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation - arXiv, accessed May 15, 2025, https://arxiv.org/html/2407.10805v7
42. Knowledge-Aware Iterative Retrieval for Multi-Agent Systems - arXiv, accessed May 15, 2025, https://arxiv.org/html/2503.13275v2
43. RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models - ACL Anthology, accessed May 15, 2025,

https://aclanthology.org/2025.naacl-long.281.pdf
44. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach - ACL Anthology, accessed May 15, 2025, https://aclanthology.org/2024.emnlp-industry.66/
45. Enhancing Policy Generation with GraphRAG and YouTube Data: A Logistics Case Study, accessed May 15, 2025, https://www.mdpi.com/2079-9292/14/7/1241
46. [2504.10497] Exploring Generative AI Techniques in Government: A Case Study - arXiv, accessed May 15, 2025, https://arxiv.org/abs/2504.10497
47. How Generative-AI can be Effectively used in Government Chatbots ——A Joint Experimental Research based on Large Language Model - arXiv, accessed May 15, 2025, https://arxiv.org/pdf/2312.02181
48. accessed January 1, 1970, https://www.oecd.org/gov/digital-government/Going_Digital_Toolkit_Note_Chatbots_for_Public_Service.pdf
49. accessed January 1, 1970, https://www.capgemini.com/insights/research-library/conversational-ai-for-citizen-services/
50. Sahabat-AI, accessed May 15, 2025, https://sahabat-ai.com/
51. GoToCompany/llama3-8b-cpt-sahabatai-v1-base · Hugging Face, accessed May 15, 2025, https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-base
52. GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct - Hugging Face, accessed May 15, 2025, https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct
53. Indosat Ooredoo Hutchison & GoTo Launch Indonesian LLM, Sahabat-AI - The Fast Mode, accessed May 15, 2025, https://www.thefastmode.com/technology-solutions/38190-indosat-ooredoo-hutchison-goto-launch-indonesian-llm-sahabat-ai
54. AI Day drives Indonesia's AI sovereignty, with Indosat leading solutions and connectivity, accessed May 15, 2025, https://www.thejakartapost.com/business/2024/11/18/ai-day-drives-indonesias-ai-sovereignty-with-indosat-leading-solutions-and-connectivity.html
55. Llama3 8b Cpt Sahabatai V1 Instruct · Models - Dataloop AI, accessed May 15, 2025, https://dataloop.ai/library/model/gotocompany_llama3-8b-cpt-sahabatai-v1-instruct/
56. Gemma2 9b Cpt Sahabatai V1 Base · Models - Dataloop AI, accessed May 15, 2025, https://dataloop.ai/library/model/gotocompany_gemma2-9b-cpt-sahabatai-v1-base/
57. Evaluating LLMs for Bahasa Indonesia: SEA-LIONv3 vs SahabatAI-v1 – We SUPA AI Blog, accessed May 15, 2025, https://blog.supa.so/2025/02/21/evaluating-llms-for-bahasa-indonesia-sea-lionv3-vs-sahabatai-v1/
58. LUSIFER: Language Universal Space Integration for Enhanced Multilingual

Embeddings with Large Language Models - arXiv, accessed May 15, 2025, https://arxiv.org/html/2501.00874v1
59. arxiv.org, accessed May 15, 2025, https://arxiv.org/abs/2307.03109
60. arxiv.org, accessed May 15, 2025, https://arxiv.org/abs/2005.11401
61. Evaluating - LlamaIndex, accessed May 15, 2025, https://docs.llamaindex.ai/en/stable/module_guides/evaluating/root.html
62. Metrics - Ragas, accessed May 15, 2025, https://docs.ragas.io/en/stable/concepts/metrics/index.html
63. Portal Kemlu, accessed May 15, 2025, https://kemlu.go.id/
64. Portal Kemlu, accessed May 15, 2025, https://kemlu.go.id/portal/id
65. accessed January 1, 1970, https://kemlu.go.id/singapore/pages/layanan-konsuler/id
66. accessed January 1, 1970, https://example-kemlu.go.id/lakip_2023.pdf
67. accessed January 1, 1970, https://example-kemlu.go.id/renstra_2020-2024.pdf
68. accessed January 1, 1970, https://example.com/LAKIP_KEMLU_RI_2023.pdf
69. accessed January 1, 1970, https://example.com/LAKIP_KEMLU_PUSAT_2023.pdf
70. accessed January 1, 1970, https://example.com/RENSTRA_KEMLU_PUSAT_2020-2024.pdf
71. accessed January 1, 1970, https://e-ppid.kemlu.go.id/direktori/108/PERMENLU_Nomor_3_Tahun_2019_TTG_PANDUAN_UMUM_HUBUNGAN_LUAR_NEGERI_OLEH_PEMERINTAH_DAERAH.pdf
72. accessed January 1, 1970, https://kbriparis.fr/layanan-konsuler/paspor
73. accessed January 1, 1970, https://kemlu.go.id/kualalumpur/pages/layanan-kekonsuleran/id
74. MENTERI LUAR NEGERI REPUBLIK INDONESIA PERATURAN MENTERI LUAR NEGERI REPUBLIK INDONESIA NOMOR 14 TAHUN 2022 TENTANG TATA CARA L, accessed May 15, 2025, https://peraturan.bpk.go.id/Download/310940/PERMENLU-14-2022.pdf
75. Permenlu No. 13 Tahun 2019.pdf - Peraturan BPK, accessed May 15, 2025, https://peraturan.bpk.go.id/Download/130666/Permenlu%20No.%2013%20Tahun%202019.pdf
76. Navigating the risks of public LLMs: Why government agencies need safe, targeted AI solutions for their investigations - Comtrac, accessed May 15, 2025, https://www.comtrac.com.au/blog/navigating-the-risks-of-public-llms-why-government-agencies-need-safe-targeted-ai-solutions-for-their-investigations
77. Sahabat-AI - a GoToCompany Collection - Hugging Face, accessed May 15, 2025, https://huggingface.co/collections/GoToCompany/sahabat-ai-672af7b248f5fdfd39ae2403
78. sentence-transformers/paraphrase-multilingual-mpnet-base-v2 ..., accessed May 15, 2025, https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
79. MTEB Leaderboard - a Hugging Face Space by mteb, accessed May 15, 2025, https://huggingface.co/spaces/mteb/leaderboard
80. arxiv.org, accessed May 15, 2025, https://arxiv.org/abs/2107.08320

81. RAG vector database explained - Writer, accessed May 15, 2025, https://writer.com/engineering/rag-vector-database/
82. 15 Best Open-Source RAG Frameworks in 2025 - Firecrawl, accessed May 15, 2025, https://www.firecrawl.dev/blog/best-open-source-rag-frameworks
83. 15 Chunking Techniques to Build Exceptional RAGs Systems - Analytics Vidhya, accessed May 15, 2025, https://www.analyticsvidhya.com/blog/2024/10/chunking-techniques-to-build-exceptional-rag-systems/
84. Chunking Strategies for LLM Applications | Pinecone, accessed May 15, 2025, https://www.pinecone.io/learn/chunking-strategies/
85. 7 Chunking Strategies in RAG You Need To Know - F22 Labs, accessed May 15, 2025, https://www.f22labs.com/blogs/7-chunking-strategies-in-rag-you-need-to-know/
86. Llamaindex vs Langchain: What's the difference? - IBM, accessed May 15, 2025, https://www.ibm.com/think/topics/llamaindex-vs-langchain
87. Fine-Tuning LLMs: A Guide With Examples - DataCamp, accessed May 15, 2025, https://www.datacamp.com/tutorial/fine-tuning-large-language-models
88. Custom Fine-Tuning for Domain-Specific LLMs - MachineLearningMastery.com, accessed May 15, 2025, https://machinelearningmastery.com/custom-fine-tuning-for-domain-specific-llms/
89. Fine-tuning 20B LLMs with RLHF on a 24GB consumer GPU, accessed May 15, 2025, https://huggingface.co/blog/trl-peft
90. LoRA - Hugging Face, accessed May 15, 2025, https://huggingface.co/docs/peft/main/en/conceptual_guides/lora
91. Tutorial: How to Run & Fine-tune Gemma 3 | Unsloth Documentation, accessed May 15, 2025, https://docs.unsloth.ai/basics/tutorial-how-to-run-and-fine-tune-gemma-3
92. Fine-Tune Gemma 3: A Step-by-Step Guide With Financial Q&A Dataset | DataCamp, accessed May 15, 2025, https://www.datacamp.com/tutorial/fine-tune-gemma-3
93. Yasser Shohoud Maged Shoman Sarah Abdelazim - arXiv, accessed May 15, 2025, https://arxiv.org/abs/2311.05120
94. arxiv.org, accessed May 15, 2025, https://arxiv.org/abs/2312.05248
95. accessed January 1, 1970, https://arxiv.org/abs/2404.07126
96. arxiv.org, accessed May 15, 2025, https://arxiv.org/abs/2402.11562
97. arxiv.org, accessed May 15, 2025, https://arxiv.org/abs/2401.05563
98. accessed January 1, 1970, https://blog.langchain.dev/evaluating-rag-pipelines/
99. Revolutionizing Business Intelligence: Natural Language Processing Applications in Indonesian Technology - BytePlus, accessed May 15, 2025, https://www.byteplus.com/en/topic/427704
100. Revolutionizing Manufacturing: Natural Language Processing's Impact in Indonesia, accessed May 15, 2025, https://www.byteplus.com/en/topic/427702
101. accessed January 1, 1970, https://python.langchain.com/docs/guides/evaluation/

102. accessed January 1, 1970, https://docs.trychroma.com/guides/rag-evaluation