THESIS PROPOSAL

Title: Enhancing Indonesian Consular Services by Integrating an Existing SahabatAI [Gemma/Llama]-based Language Model with a Retrieval-Augmented Generation Framework

Abstract
This thesis proposal outlines a six-month research project to develop an advanced AI-powered chatbot designed to improve consular services for Indonesian citizens, managed by the Ministry of Foreign Affairs (MoFA). The proposed system will leverage a Retrieval-Augmented Generation (RAG) framework, integrating an existing, fine-tuned "SahabatAI" Large Language Model (LLM)—specifically a [Gemma-based or Llama-based] instruction-tuned variant from the SahabatAI ecosystem developed by GoToCompany, Indosat Ooredoo Hutchison, and collaborators 1—with a comprehensive knowledge base derived from MoFA's official sources such as "Peduli WNI," "Safe Travel," and SARI chatbot information. The research will focus on the effective application of these state-of-the-art Indonesian LLMs within a RAG pipeline for consular queries. Key objectives include constructing a robust consular knowledge base, developing an effective RAG pipeline utilizing a selected SahabatAI model, optimizing prompt engineering for consular context in Bahasa Indonesia, and rigorously evaluating the system's performance using intrinsic, RAG-specific, and task-oriented metrics. This research aims to contribute to the digital transformation of MoFA's services by demonstrating a practical and effective way to deploy advanced, locally-developed AI tools for providing scalable, accessible, and reliable information and assistance to Indonesian citizens abroad. The proposal details the background, problem statement, literature review, methodology, evaluation plan, project timeline, and expected contributions to the fields of AI in government and NLP for Bahasa Indonesia.

Chapter 1: Introduction

1.1 Background: Consular Services and the Role of AI

Consular services form a critical lifeline for citizens residing or traveling abroad, encompassing a wide spectrum of support including emergency assistance, guidance on legal matters, facilitation of documentation processes, and the provision of general information pertinent to their stay in a foreign country. The traditional mechanisms for delivering these services, however, frequently encounter significant challenges. These include limitations in accessibility, particularly outside of standard working hours or in remote locations, the substantial workload imposed on consular staff often leading to delays, and difficulties in maintaining consistency and accuracy of information disseminated across various channels and diplomatic posts. The increasing global mobility of citizens and the complexity of international regulations further exacerbate these pressures, highlighting an urgent need for more efficient and responsive service delivery models.

Artificial Intelligence (AI), and more specifically the advancements in conversational AI driven by Large Language Models (LLMs), presents a transformative potential to address these inherent challenges in consular service provision.[5] LLMs, with their sophisticated capabilities in understanding natural language, generating human-like text, and engaging in coherent dialogue, are well-suited for developing intelligent chatbot solutions. Such AI-powered chatbots can offer 24/7 support, provide instantaneous responses to a multitude of common inquiries, and deliver personalized assistance tailored to individual user needs. The application of AI in government services globally has already demonstrated its capacity to automate routine tasks, thereby enhancing efficiency and citizen satisfaction.[6] In the consular context, AI chatbots can effectively function as a first line of support, adeptly handling a high volume of routine queries and thereby allowing human consular officers to dedicate their expertise and time to more complex, sensitive, or urgent cases that require direct human intervention.[6] This strategic augmentation of human capabilities with AI tools promises not only to improve the operational efficiency of consular services but also to enhance the overall experience for citizens seeking assistance abroad.

1.2 Indonesian Ministry of Foreign Affairs (MoFA) Context: Current Services and the Emergence of Advanced Local LLMs

The Indonesian Ministry of Foreign Affairs (MoFA) has demonstrated a proactive stance in embracing digital transformation to augment the quality and accessibility of services provided to its substantial diaspora and citizens traveling overseas.[11] This commitment is evidenced by the development and deployment of several digital platforms. Among these are the "Portal Peduli WNI" (Care for Indonesian Citizens Portal), a comprehensive web-based platform that facilitates self-reporting for citizens residing abroad, offers access to various consular services, and provides a channel for reporting cases or incidents.[11] Complementing this is the "Safe Travel" mobile application, designed for Indonesian citizens undertaking short trips abroad, which aims to provide essential travel information and assistance.[11]

A significant step in MoFA's digital journey is the development of the "Sahabat Artifisial Migran Indonesia" (SARI) AI chatbot. This initiative, undertaken in collaboration with UN Women, is specifically tailored to support and protect Indonesian female migrant workers.[11] The SARI chatbot is envisioned to possess capabilities such as language detection (potentially including regional Indonesian languages like Javanese) and the ability to offer empathetic responses.[11] The development process for SARI emphasized a human-centered design and the integration of gender bias-free data.[16]

Concurrently, the Indonesian AI landscape has seen remarkable advancements with the emergence of the "SahabatAI" ecosystem, co-initiated by leading Indonesian technology and telecommunication companies, GoTo Group and Indosat Ooredoo Hutchison, with development support from entities like AI Singapore.[1] This initiative has produced a collection of powerful LLMs, including Gemma-based and Llama-based models, which have been specifically pre-trained and instruction-tuned for Bahasa Indonesia and its various dialects (e.g., Javanese, Sundanese).[1] These models, such as Gemma2 9B CPT Sahabat-AI v1 Instruct and Llama3 8B CPT Sahabat-AI v1 Instruct, represent a significant leap in providing high-quality, locally relevant language understanding and generation capabilities.[1]

The existence of MoFA's SARI chatbot and the broader SahabatAI LLM ecosystem creates a highly opportune environment for this research. This project aims to leverage these existing advanced Indonesian LLMs to develop a more comprehensive consular assistance chatbot, building upon MoFA's digital transformation goals and the availability of powerful, locally-tuned AI models. User feedback on existing applications, such as connectivity or language issues with "Safe Travel" [13], underscores the need for technically robust and linguistically effective solutions, a challenge that the high-quality SahabatAI models are well-positioned to address.

1.3 Problem Statement: Challenges and Opportunities for AI-Enhanced Consular Assistance Using Existing Advanced Indonesian LLMs

Despite MoFA's ongoing efforts and the availability of platforms like Portal Peduli WNI and the SARI chatbot [11], the delivery of comprehensive consular services to Indonesian citizens globally faces persistent challenges. These include ensuring timely and accurate information dissemination, managing a high volume of repetitive inquiries, providing consistent support across time zones, and delivering empathetic communication, especially in sensitive situations.[20] While SARI targets a specific demographic and set of issues [16], a broader need exists for an AI assistant capable of addressing a wider range of general consular queries.

The emergence of the SahabatAI LLM ecosystem [1] presents a significant opportunity to address these challenges more effectively. These models, already pre-trained and instruction-tuned for Bahasa Indonesia and its dialects, offer a powerful foundation for a sophisticated consular chatbot. The core problem this thesis addresses is how to best *integrate and apply* these existing advanced Indonesian LLMs within a

Retrieval-Augmented Generation (RAG) framework to create an efficient, reliable, and user-centric consular assistance tool for MoFA. The technical challenge shifts from building an LLM from scratch to:

1. Selecting the most suitable SahabatAI model variant for the consular domain.
2. Curating a comprehensive and accurate MoFA knowledge base.
3. Designing an optimal RAG pipeline that effectively grounds the chosen SahabatAI model in this knowledge base.
4. Ensuring the system can generate accurate, contextually relevant, and appropriately toned responses in Bahasa Indonesia for a wide array of consular matters.

This research aims to bridge the gap between the demand for enhanced consular support and current service capacities by demonstrating a practical and robust application of the SahabatAI models.

1.4 Research Aim and Objectives

Aim:
To design, develop, and evaluate a prototype AI-powered consular chatbot by integrating an existing, instruction-tuned SahabatAI [Gemma/Llama]-based Large Language Model with a Retrieval-Augmented Generation (RAG) framework, to enhance the provision of consular information and services for Indonesian citizens by the Ministry of Foreign Affairs (MoFA).

Objectives:

1. To conduct a comprehensive review of state-of-the-art (SOTA) RAG techniques and their applications in public and consular services, focusing on the optimal utilization of existing advanced Indonesian LLMs like those from the SahabatAI ecosystem.[1]
2. To curate and structure a domain-specific knowledge base from official MoFA sources (e.g., Peduli WNI, Safe Travel, kemlu.go.id, SARI-related public information, imigrasi.go.id) relevant to common consular inquiries made by Indonesian citizens.
3. To select an appropriate, existing instruction-tuned SahabatAI [Gemma/Llama]-based LLM (e.g., Gemma2 9B CPT Sahabat-AI v1 Instruct [3] or Llama3 8B CPT Sahabat-AI v1 Instruct [4]) and evaluate its suitability for generating contextually relevant, accurate, and appropriately toned responses in Bahasa Indonesia for consular topics, primarily through effective prompt engineering within the RAG framework. Minimal further fine-tuning using PEFT techniques (e.g., QLoRA) will be considered only if significant gaps are identified.
4. To implement a robust RAG pipeline that effectively integrates the selected

SahabatAI model with the curated MoFA knowledge base, employing suitable embedding models and retrieval strategies for Bahasa Indonesia.

5. To rigorously evaluate the performance of the prototype RAG-based consular chatbot in terms of accuracy, relevance, fluency, factual grounding, and its ability to handle a diverse set of simulated consular-specific queries in Bahasa Indonesia.
6. To analyze the feasibility of developing and deploying such a system leveraging existing local LLM ecosystems, discuss its potential contributions to MoFA's digital service offerings, and identify limitations and directions for future research.

1.5 Research Questions

1. Which existing instruction-tuned SahabatAI model variant (e.g., Gemma2 9B CPT SahabatAI v1 Instruct [3], Llama3 8B CPT Sahabat-AI v1 Instruct [4]) demonstrates the most optimal balance of performance on Bahasa Indonesia consular-related tasks when integrated into a RAG system, considering its existing fine-tuning and suitability for the generator role?
2. How can a comprehensive, accurate, and up-to-date knowledge base for Indonesian consular services be effectively constructed by extracting, cleaning, and structuring information from MoFA's existing digital platforms (Portal Peduli WNI, Safe Travel app information, official website documents, publicly available SARI details, imigrasi.go.id) for integration into a RAG system?
3. What is the most effective RAG architecture for a consular chatbot utilizing a selected SahabatAI model—including the choice of embedding models optimized for Bahasa Indonesia, vector database configuration, document chunking strategies, and retrieval mechanisms—to ensure the accurate and relevant retrieval of consular information?
4. To what extent can an existing SahabatAI model, when integrated into a RAG system and guided by effective prompt engineering, generate accurate, coherent, factually grounded, and (where appropriate) empathetic responses to a diverse range of simulated Indonesian consular service queries, and how does this compare to its performance without RAG?
5. What are the key technical challenges (e.g., ensuring consistent tone with a pre-tuned model, retrieval of precise information from diverse government documents, prompt sensitivity), limitations, and ethical considerations (e.g., data privacy, transparency of an AI system using a third-party model, accountability) in developing and potentially deploying an LLM-based RAG chatbot using an existing advanced Indonesian model like SahabatAI for consular services within

the MoFA context?

1.6 Scope and Delimitations of the Thesis

Scope:
The research will encompass the design, development, and evaluation of a prototype AI-powered consular chatbot. This chatbot will utilize an existing, instruction-tuned Large Language Model from the SahabatAI ecosystem (e.g., Gemma2 9B CPT Sahabat-AI v1 Instruct or Llama3 8B CPT Sahabat-AI v1 Instruct 3) as its core generative component within a RAG framework. The primary mode of interaction will be textual, exclusively in Bahasa Indonesia. The knowledge base underpinning the RAG system will be constructed from publicly accessible MoFA and related official Indonesian government information, focusing on a representative selection of common consular topics (e.g., passport services, visa information, "Lapor Diri" 11, emergency assistance). The evaluation will use automated metrics and limited human assessment.
Delimitations:
This Master's thesis project will not involve direct integration with MoFA's internal, live operational systems or confidential databases. The chatbot will operate based on a static, curated snapshot of public information. Full-scale deployment and extensive user testing with actual Indonesian citizens abroad are outside the project's boundaries. Advanced functionalities like multi-modal input or proactive notifications will not be implemented. The research will not undertake the development or extensive fine-tuning of an LLM from scratch; it will leverage existing SahabatAI models. While the SARI chatbot aims to detect regional languages 18, this project will primarily focus on standard Bahasa Indonesia, aligning with the primary language capabilities of the selected SahabatAI model for instruction-following.

1.7 Significance and Potential Contributions

The proposed research holds significant potential contributions:

- **To MoFA and Indonesian Public Service:** This thesis will deliver a tangible proof-of-concept demonstrating how existing, advanced Indonesian LLMs from initiatives like SahabatAI [1] can be practically applied to enhance consular services. It aligns with Indonesia's National AI Strategy for bureaucratic reform and public service improvement.[21] The findings can offer a blueprint for MoFA and other agencies to leverage local AI ecosystems.
- **To the Field of AI/NLP for Bahasa Indonesia:** The research will contribute to understanding the application of sophisticated RAG techniques with locally-developed, instruction-tuned LLMs like SahabatAI for a specialized domain in Bahasa Indonesia. It will provide empirical evidence on the effectiveness of these models in a real-world-like application, complementing existing benchmark evaluations (e.g., SEA HELM, IndoMMLU [1]).
- **Academic Contribution:** The thesis will serve as a documented case study and

methodology for integrating existing regional LLMs into public service RAG systems, particularly valuable for low-resource language contexts where such advanced local models are emerging.

1.8 Thesis Structure

This thesis proposal is organized into six chapters. **Chapter 1** has provided the introduction. **Chapter 2** will present a Literature Review on LLMs (focusing on the SahabatAI ecosystem), RAG techniques, AI in public/consular services, and NLP for Bahasa Indonesia. **Chapter 3** will detail the Research Methodology, including SahabatAI model selection, knowledge base construction, RAG implementation, and prompt engineering. **Chapter 4** will describe the Evaluation Plan. **Chapter 5** will lay out the Project Timeline. **Chapter 6** will discuss Expected Outcomes and Contributions.

Chapter 2: Literature Review

2.1 Advancements in Large Language Models (LLMs) for Conversational AI

The field of conversational AI has been revolutionized by Large Language Models (LLMs), which form the core of modern intelligent chatbots.

- 2.1.1 Foundational LLM Architectures (e.g., Transformer)
  The Transformer architecture, with its self-attention mechanisms, remains the foundation of most SOTA LLMs, enabling nuanced understanding of context and long-range dependencies.23 This allows LLMs to perform a wide array of NLP tasks with high proficiency.
- 2.1.2 The SahabatAI Ecosystem: Advanced LLMs for Bahasa Indonesia
  A significant development for NLP in Indonesia is the "SahabatAI" ecosystem, co-initiated by Indonesian tech and telecommunication companies PT GoTo Gojek Tokopedia Tbk (GoToCompany) and Indosat Ooredoo Hutchison, with development by GoToCompany and AI Singapore.1 SahabatAI provides a collection of LLMs specifically pre-trained and instruction-tuned for Bahasa Indonesia and its various dialects, including Javanese and Sundanese.1 This initiative aims to empower Indonesians to develop AI-based services using local languages.1
  Key models within the SahabatAI collection available on Hugging Face include [4]:
  - **GoToCompany/gemma2-9b-cpt-sahabatai-v1-base**: A Gemma-2 9B model that underwent continued pre-training from the Gemma2 9B CPT SEA-Lionv3 base model with approximately 50B tokens of Indonesian,

Javanese, Sundanese, and English data.[2] It has a context length of 8192 tokens.[2]

- ○ **GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct**: This is an instruction-tuned version of the gemma2-9b-cpt-sahabatai-v1-base model. It was fine-tuned with around 448,000 Indonesian instruction-completion pairs, 96,000 Javanese pairs, 98,000 Sundanese pairs, and 129,000 English pairs.[1] It is designed for instruction-following and conversational tasks.
- ○ **GoToCompany/llama3-8b-cpt-sahabatai-v1-base**: A Llama-3 8B model that has undergone continued pre-training for the SahabatAI ecosystem.[4]
- ○ **GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct**: An instruction-tuned version of the Llama-3 8B SahabatAI base model.[4]

These models have been evaluated on benchmarks like SEA HELM (BHASA) for tasks including Question Answering, Sentiment Analysis, Translation, and Summarization in Indonesian, Javanese, and Sundanese, as well as on IndoMMLU for Indonesian examination questions and IFEval for instruction-following.[1] The availability of these locally-tuned, high-performance models significantly lowers the barrier for developing advanced NLP applications for Indonesian contexts, such as the proposed consular chatbot. This research will focus on selecting one of the SahabatAI *instruct* models as the generator in the RAG system.

- 2.1.3 Parameter-Efficient Fine-Tuning (PEFT) and Instruction Tuning in SahabatAI
The SahabatAI instruct models (e.g., Gemma2 9B CPT Sahabat-AI v1 Instruct) were created through instruction tuning on their respective base models.1 While the specific fine-tuning methodology for SahabatAI isn't fully detailed in the provided snippets, such processes typically employ Parameter-Efficient Fine-Tuning (PEFT) techniques like QLoRA (Quantized Low-Rank Adaptation) or LoRA (Low-Rank Adaptation) for efficiency.24 QLoRA, for instance, quantizes the pre-trained model to 4-bits and freezes its weights, then trains small, low-rank adapter layers.24 This makes fine-tuning large models feasible on limited hardware.
Supervised Fine-Tuning (SFT) is the process of training an LLM on a dataset of instructions and desired responses.26 The SahabatAI instruct models were fine-tuned on extensive instruction-completion pairs in multiple Indonesian languages and English.1
For this thesis, the primary approach will be to leverage the existing instruction-tuning of the selected SahabatAI model. However, if specific deficiencies in handling nuanced consular language or tasks are observed even within the RAG framework, a minimal further SFT using QLoRA on a small, highly targeted consular dataset could be considered as an extension.

2.2 Retrieval-Augmented Generation (RAG) for Knowledge-Intensive Tasks

RAG is an AI framework that enhances LLM capabilities by dynamically integrating them with external knowledge bases.[28] This addresses LLM limitations like knowledge cut-offs and hallucinations.[28]

- 2.2.1 Core Principles, Architecture, and Benefits of RAG
  RAG systems typically have a Retriever (searches external knowledge) and a Generator (an LLM that uses retrieved context to form an answer).29 Benefits include access to up-to-date information, factual grounding, reduced hallucinations, domain-specific knowledge integration without full retraining, and potential for transparency by citing sources.28
- 2.2.2 RAG in Domain-Specific Chatbots (e.g., public sector, FAQs)
  RAG is effective for domain-specific chatbots needing high accuracy, like those in the public sector using official documents or FAQs.36 The quality of the knowledge base is crucial.31 Consular services, with their factual and procedural nature based on official MoFA guidelines, are an ideal RAG use case.
- **2.2.3 Advanced RAG Techniques**
  - **Vector Databases and Semantic Search:** Documents are embedded and stored in vector databases (e.g., FAISS, Vertex AI Vector Search [32]) for semantic similarity search against query embeddings.[28]
  - **Re-ranking:** A secondary model can re-rank initially retrieved documents for better relevance.[28]
  - **Retrieval-Augmented Fine-Tuning (RAFT):** RAFT fine-tunes an LLM to better handle "open-book" in-domain settings by training it with questions, relevant "golden" documents, and irrelevant "distractor" documents. The model learns to ignore distractors and cite relevant sources.[39] This technique could be valuable for making the chosen SahabatAI model more adept at using the MoFA knowledge base, especially if retrieved chunks are noisy.
  - **RL for RAG Optimization:** Reinforcement learning can optimize RAG, for example, by deciding when to retrieve context to save costs.[37]

2.3 AI in Public Sector and Consular Services

AI is increasingly used in the public sector globally to enhance service delivery and efficiency.[6]

- 2.3.1 International Case Studies of AI in Government and Citizen Services
  Examples include Singapore's GovTech chatbots 7, Dubai's "Rammas" 7, US USCIS AI for asylum processing 6, Australia's ATO "Alex" 7, and the US Department of State's use of AI for customer feedback, translation, and an enhanced chatbot for

Travel.State.Gov.44 The "Visa-Vanguard" chatbot provides visa/immigration support for students using NLP/ML.45 These show AI's utility in automating services, FAQ handling, and document processing.6

- 2.3.2 AI Initiatives within the Indonesian Government and MoFA
  Indonesia's National AI Strategy (2020-2045) prioritizes public service reform.46 MoFA's SARI chatbot, for female migrant workers, is a key example.11 Other ministries also use chatbots (e.g., Ministry of Tourism's "TIWI", Ministry of Health's "PeduliLindung WhatsApp Chatbot" 18). AI is also explored for land certification.49 The SahabatAI LLM ecosystem 1 represents a major step in local AI capability, providing powerful tools for such government initiatives.

2.4 Natural Language Processing for Bahasa Indonesia

Developing "SahabatAI" relies on robust NLP for Bahasa Indonesia.

- 2.4.1 Challenges and Progress for Low-Resource Languages
  Bahasa Indonesia, while widely spoken, is often considered low-resource in NLP due to a relative scarcity of large, diverse, high-quality datasets compared to English.50 However, significant progress has been made. Early models like IndoBERT and IndoGPT provided baselines.55 More recently, adapting multilingual models or developing local ones like NusaMT-7B (LLaMA2-7B based, for low-resource Indonesian languages like Balinese and Minangkabau) 50 and the SahabatAI models 1 showcase advanced capabilities. These efforts emphasize data quality and sophisticated fine-tuning.
- **2.4.2 Existing Datasets, Benchmarks, and Models for Indonesian**
  - **Datasets/Corpora:** NusaCrowd unifies Indonesian language resources.[54]
  - **Benchmarks:**
    - **IndoCareer:** 8,834 multiple-choice questions from Indonesian professional exams, including a "Law" section relevant to consular tasks. Used to evaluate models like Gemma, Llama, and SahabatAI variants.[55]
    - **SeaExam & SeaBench:** Benchmarks for SEA languages including Indonesian, based on real-world scenarios and regional exams, evaluating instruction-following. Used to test Gemma-2 and Llama-3.1.[61] SahabatAI models also report SEA HELM (BHASA) benchmark results.[1]
    - **IndoMMLU:** Indonesian MMLU, used for SahabatAI evaluation.[1]
  - **Models:**
    - **SahabatAI Models:** The Gemma2 9B CPT Sahabat-AI v1 Instruct and Llama3 8B CPT Sahabat-AI v1 Instruct are SOTA for Indonesian, pre-trained and instruction-tuned on extensive Indonesian (and regional dialects) and English data.[1] Their performance on benchmarks like SEA

    HELM and IndoMMLU makes them prime candidates.
  - Other multilingual models (GPT-4o, other Llama/Gemma variants) also perform well on Indonesian benchmarks.[55]

The availability of strong, locally-tuned models like SahabatAI significantly enhances the feasibility and potential impact of this project.

2.5 Synthesis and Identification of Research Gaps

The literature shows advanced LLMs (SahabatAI), robust RAG techniques, growing AI adoption in public service (including Indonesia), and maturing NLP for Bahasa Indonesia. Key research gaps this thesis addresses:

- There is limited published research on *applying* existing advanced Indonesian LLMs like SahabatAI within a RAG framework specifically for the comprehensive domain of Indonesian consular services. While MoFA's SARI exists [11], and SahabatAI provides general-purpose models [1], their specific integration and evaluation for broad consular tasks using RAG is a novel area.
- A detailed framework for building and evaluating such consular chatbots using *existing local LLM ecosystems* in the Indonesian public sector context is needed. This includes optimal RAG configurations for MoFA data with SahabatAI models and appropriate evaluation metrics.
- Further investigation into the best RAG strategies (embedding models for Bahasa Indonesia, chunking, prompt engineering) when using a pre-tuned model like SahabatAI as the generator for official Indonesian documents is required. The potential benefits of techniques like RAFT [39] to enhance SahabatAI's contextual grounding with consular data warrant exploration.

This thesis aims to fill these gaps by proposing a detailed approach for a consular RAG system leveraging the SahabatAI ecosystem, offering a case study for applying local AI advancements to public service.

Chapter 3: Proposed Research Methodology

This chapter outlines the systematic approach to design, develop, and prepare for the evaluation of the consular chatbot, leveraging an existing SahabatAI model.

3.1 Overall System Design: Architecture of the RAG-based Consular Chatbot

The proposed architecture will be a RAG pipeline:

```
+-----------------+    +-----------------+    +---------------------+
+-------------------+    +-----------------+
| User Interface | --> | Query Processor | --> | RAG Module | --> | Response Generator|
--> | User Interface |
| (Text Input) | | (Preprocessing) | | | (SahabatAI Model) | | (Text Output) |
+-----------------+    +-----------------+    +----------+----------+
+-------------------+    +-----------------+
| ^
| |
| Query | Retrieved
| | Context
                                   v |
                              +-----------+-----------+
| Retriever | Generator |
| | (SahabatAI |
| | Model e.g., |
| | Gemma2 9B |
| | Instruct) |
                              +-----------+-----------+
| ^
| |
                                   v |
                              +---------------------+
| Knowledge Base |
| (MoFA Consular Info)|
| (Vector DB + Raw |
| Text Store) |
                              +---------------------+
```

Figure 3.1: Proposed Architecture for the Consular Chatbot

- **User Interface (UI):** Simple text-based interface.
- **Query Processor:** Basic cleaning of user queries.
- **RAG Module:**
  - **Retriever:** Embeds query, searches Knowledge Base for relevant chunks.
  - **Generator (SahabatAI Model):** A selected instruction-tuned model from the

SahabatAI ecosystem (e.g., GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct [3] or Llama3 8B CPT Sahabat-AI v1 Instruct [4]). It receives the query and retrieved context to generate a response.
- **Knowledge Base (KB):** Curated MoFA consular information (vector DB + raw text).
- **Response Generator:** The output from the SahabatAI model.

This modular design allows focused development and testing of each component.[29]

3.2 Data Acquisition and Knowledge Base Construction

The KB's quality is paramount for RAG effectiveness.[31]

- 3.2.1 Identification of Data Sources
  Primary sources:
  - MoFA website (kemlu.go.id).
  - Portal Peduli WNI (peduliwni.kemlu.go.id).[11]
  - Safe Travel app public information.[11]
  - Publicly available SARI chatbot information.[11]
  - Directorate General of Immigration website (imigrasi.go.id) for passport/visa info.[64]
  - MoFA FAQs, press releases, consular announcements, relevant regulations. Data will be publicly available, non-sensitive, adhering to Indonesia's PDP Law.[67]
- **3.2.2 Techniques for Data Extraction, Cleaning, and Structuring for RAG**
  - **Extraction:** Web scraping tools (e.g., BeautifulSoup, Requests, Puppeteer [69], Firecrawl [70], OneFileLLM [70]) for websites; PDF text extraction libraries (PyMuPDF, pdfminer.six) for documents.
  - **Cleaning:** Remove HTML, boilerplate; normalize text.
  - **Structuring for RAG:** Segment into coherent **chunks** (experiment with size/strategy).[31] Extract Q&A pairs from FAQs.[30] Store chunks with **metadata** (source, date, topic).[31] Formats: JSON or Markdown.
- 3.2.3 Development of a Sample Consular Knowledge Base (Indonesian context)
  Focus on a subset of common consular topics: passport renewal, visa info, "Lapor Diri" 11, emergency guidelines, Portal Peduli WNI services. All content in Bahasa Indonesia.

3.3 LLM Selection and Application Strategy for the Generator

The generator will be an existing model from the SahabatAI ecosystem.

- 3.3.1 Justification for SahabatAI Model Choice
  The selection will be from the instruction-tuned SahabatAI models, likely GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct 1 or GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct.4 Criteria:
  1. **Performance on Indonesian Benchmarks:** Reported scores on SEA HELM, IndoMMLU, IFEval.[1]
  2. **Instruction Tuning Quality:** Effectiveness for conversational tasks and following complex instructions in Bahasa Indonesia, Javanese, and Sundanese.[1]
  3. **Open-Source Availability and Licensing:** Gemma Community License.[1]
  4. **Computational Requirements for Inference:** Feasibility for a prototype.
  5. **Context Window Size:** SahabatAI Gemma2 9B has 8192 tokens.[2]
  6. **Community Support and Ecosystem:** Hugging Face integration.[2]

  The chosen model is already adapted for Indonesian, reducing the need for extensive fine-tuning from scratch.
- 3.3.2 Application Approach: Leveraging Existing Instruction-Tuning and Prompt Engineering
  The primary strategy will be to use the selected SahabatAI instruct model "as-is" for generation, relying on its existing strong instruction-following capabilities in Bahasa Indonesia.1 The main effort will be focused on robust prompt engineering to guide the model effectively within the RAG context for consular queries.
  Minimal Further Fine-tuning (Contingency): If the selected SahabatAI model, even with sophisticated prompting and RAG, shows specific, consistent deficiencies in handling highly specialized consular terminology, tone, or complex reasoning unique to this domain, a minimal and highly targeted Supervised Fine-Tuning (SFT) using QLoRA 24 could be explored. This would only be undertaken if deemed essential and feasible within the project timeline.
- 3.3.3 Training Data Preparation (for potential minimal fine-tuning)
  If minimal SFT is pursued, a small, high-quality dataset of instruction-response pairs specific to nuanced consular scenarios in Bahasa Indonesia would be curated.
  Format: "Messages" format (system, user, assistant roles).24
  Content Examples: Queries requiring understanding of specific MoFA procedures not easily captured by general instruction tuning, or needing a very particular empathetic yet official tone.
  Data Sources/Generation: Derived from MoFA documents, simulated complex queries. LLM-assisted generation (e.g., using a larger model like GPT-4 via API for

drafts) followed by rigorous human review for accuracy and appropriateness in Bahasa Indonesia, similar to data strategies for NusaMT-7B 58 or RAFT.39
The primary goal is to leverage the existing strengths of SahabatAI models, making extensive new fine-tuning a secondary option.

3.4 RAG Framework Implementation

- **3.4.1 Selection and Implementation of Embedding Models for Indonesian** Research and select sentence embedding models effective for Bahasa Indonesia (e.g., multilingual models from sentence-transformers or Indonesian-specific ones if available and superior). The quality of embeddings is crucial for retrieval.[37]
- **3.4.2 Vector Database Setup and Indexing Strategy** Use an open-source vector database like FAISS for storing and searching document chunk embeddings. Indexing parameters will be chosen based on dataset size and search needs.
- **3.4.3 Integration of Retriever with the Selected SahabatAI Generator**
  1. User query embedded.
  2. Vector DB search retrieves top-k relevant document chunks.
  3. Retrieved chunks + original query form an augmented prompt.
  4. Augmented prompt fed to the selected SahabatAI instruct model for response generation. **Consideration of RAFT:** If initial RAG performance with the chosen SahabatAI model needs improvement in handling noisy or complex retrieved MoFA documents, the RAFT methodology [39] will be explored. This would involve fine-tuning the SahabatAI model with training data including both relevant and distractor documents, teaching it to cite sources and ignore irrelevant context. This would be a more involved fine-tuning step than the minimal SFT mentioned earlier.
- **3.4.4 Prompt Engineering Strategies for Consular Queries with SahabatAI Models** Effective prompt engineering is critical, especially when relying on a pre-existing instruction-tuned model. Prompts will:
  - Define the chatbot's persona (official, helpful, empathetic MoFA assistant).
  - Instruct the SahabatAI model to base answers on provided context.
  - Specify Bahasa Indonesia output.
  - Guide handling of insufficient context (avoid hallucination).
  - Encourage appropriate tone. Example prompt structure (similar to previous, but emphasizing guidance for the pre-tuned SahabatAI model):

```
System: Anda adalah asisten konsuler virtual dari Kementerian Luar Negeri Republik
Indonesia, didukung oleh model SahabatAI. Tugas Anda adalah menjawab pertanyaan
pengguna terkait layanan konsuler Indonesia secara akurat, jelas, dan jika situasinya
membutuhkan, dengan empati, berdasarkan informasi yang disediakan. Selalu jawab
```

dalam Bahasa Indonesia.

User: {user_query}

Konteks Dokumen Resmi yang Relevan:

: {retrieved_chunk_1}
: {retrieved_chunk_2}
...
: {retrieved_chunk_k}

Assistant:
```
Iterative refinement of prompts will be key.

### 3.5 Ethical Considerations and Data Privacy

- **Data Privacy:** Adherence to Indonesia's PDP Law (Law No. 27 of 2022).[67] KB from public MoFA/government data only. No private citizen data.
- **Bias Mitigation:** SahabatAI models are developed with attention to local languages and contexts [1], which may help. Fine-tuning data (if any) and prompts will promote neutrality. The SARI chatbot's aim for gender-bias-free data is a good reference.[16]
- **Transparency and Explainability:** Inform users they are interacting with an AI (SahabatAI-powered system).[6] RAG allows citing sources for transparency.[34]
- **Accountability and Human Oversight:** The prototype is assistive, not a replacement for human officers.[6] Defer complex/critical cases.
- **Alignment with Ethical Guidelines:** Consider UNESCO's Ethical AI recommendations and Indonesian AI ethics frameworks (e.g., Ministerial Circular No. 9/2023).[73]

---

## Chapter 4: Evaluation Plan and Metrics

Evaluation will assess the SahabatAI-RAG system's performance.

### 4.1 Intrinsic Evaluation of the Selected SahabatAI Model

- **Baseline Capabilities:** Assess the chosen SahabatAI instruct model's standalone
```

performance on a set of consular-style prompts (without RAG) for coherence, fluency in Bahasa Indonesia, relevance, and general factual correctness. This establishes a baseline before RAG integration.
- **Perplexity (if further fine-tuning is done):** If minimal SFT is performed, compare perplexity on a consular domain test set before and after this additional tuning.

4.2 RAG System Performance Metrics

- 4.2.1 Retrieval Accuracy and Relevance
  Using a test set of consular queries:
  - **Context Precision@k / Hit Rate@k:** Proportion of top-k retrieved chunks relevant to the query (manual judgment).
  - **Context Recall@k:** Proportion of all relevant chunks retrieved (harder, estimate on subset).
  - **Mean Reciprocal Rank (MRR):** For ranking the first relevant document.
  - **Normalized Discounted Cumulative Gain (NDCG@k):** Evaluates ranking quality with graded relevance. The framework by Zając et al. (2024) for RAG in education (evaluating if correct/sufficient info was retrieved) is relevant.[33]
- **4.2.2 Generated Response Quality (SahabatAI model with RAG)**
  - **Faithfulness / Factual Accuracy (Groundedness):** How well the response reflects retrieved context, avoiding contradiction/hallucination.
    LLM-as-a-judge [37] and human evaluation.
  - **Answer Relevance:** How well the response addresses the query using context (human-evaluated).
  - **Fluency and Coherence:** Naturalness, grammar in Bahasa Indonesia (human-evaluated). ROUGE/BLEU as indicative.
  - **Human Evaluation (Overall Quality):** Panel of Indonesian speakers rating accuracy, completeness, helpfulness, clarity, tone.
  - **NLU Rate / Negative Signals:** Track query understanding, repetition, abandonment.[74] The multi-perspective framework by Patalinghug et al. (2024) for LGU chatbots offers a good checklist.[75]

4.3 Task-Oriented Evaluation
Define representative consular tasks (e.g., "procedure for lost passport in X country").
- **Task Completion Rate (TCR):** Percentage of tasks where the chatbot provides accurate, complete, actionable information (human judgment).[33]

4.4 User-Centric Evaluation (if feasible)
Small-scale pilot with 5-10 Indonesian speakers role-playing.
- **User Satisfaction Surveys:** SUS or custom Likert scales for helpfulness, ease of

use, trust.[76]

- **Qualitative Feedback:** Open-ended questions/interviews.

4.5 Baselines for Comparison

1. **Generic Multilingual LLM (similar size to selected SahabatAI model, No RAG):** E.g., a standard Gemma or Llama model without SahabatAI's specific Indonesian CPT/SFT. This shows the value of SahabatAI's localization.
2. **Selected SahabatAI Instruct Model (No RAG):** The chosen SahabatAI model prompted directly. This isolates the RAG contribution.
3. **Keyword-Based Search + Simple Template (Optional):** Traditional IR baseline.

Chapter 5: Project Timeline and Feasibility

A six-month plan for the Master's thesis.

5.1 Detailed 6-Month Work Plan

- **Month 1: Literature Review, SahabatAI Model Selection, Setup (Weeks 1-4)**
  - Weeks 1-2: In-depth literature review (RAG, SahabatAI ecosystem [1], AI in public service). Finalize selection of the specific SahabatAI instruct model.
  - Weeks 3-4: Development environment setup. Detailed MoFA data source identification.
  - Deliverables: Chapters 2 & 3. Functional dev environment.
- **Month 2: Knowledge Base Construction (Weeks 5-8)**
  - Weeks 5-6: Web scraping/PDF extraction for MoFA data.[69]
  - Weeks 7-8: Data cleaning, structuring, chunking. If minimal SFT is planned for SahabatAI, begin curating this small dataset.
  - Deliverables: Initial MoFA consular KB. Draft of minimal SFT dataset (if any).
- **Month 3: Retriever Implementation & Initial SahabatAI Integration (Weeks 9-12)**
  - Weeks 9-10: Implement retriever (embedding model for Indonesian, vector DB like FAISS). Index KB.
  - Weeks 11-12: Integrate selected SahabatAI model with the retriever. Develop initial prompt engineering strategies. (If minimal SFT or RAFT is pursued, this is when that focused fine-tuning would occur, using QLoRA [24]).
  - Deliverables: Functional Retriever. Initial RAG prototype with SahabatAI model.
- **Month 4: RAG Pipeline Testing, Evaluation Setup & Iteration (Weeks 13-16)**
  - Weeks 13-14: End-to-end testing. Refine prompts for SahabatAI.
  - Weeks 15-16: Preliminary evaluations (retrieval, generation quality). Debug and iterate. Prepare human evaluation protocols.

    - Deliverables: RAG prototype v1. Preliminary Evaluation Report.
- **Month 5: Comprehensive Evaluation & Analysis (Weeks 17-20)**
    - Weeks 17-18: Implement any refinements from initial feedback. Conduct comprehensive evaluation (automated, human, task-oriented).
    - Weeks 19-20: Analyze all results.
    - Deliverables: Refined RAG prototype v2. Comprehensive Evaluation Results (Chapter 4 basis).
- **Month 6: Thesis Writing, Finalization & Submission (Weeks 21-24)**
    - Weeks 21-24: Write all thesis chapters. Prepare presentation. Review, revise, submit.
    - Deliverables: Final Thesis. Presentation.

(Gantt Chart similar to original, but "LLM Fine-tuning" in Month 3 is now "Initial SahabatAI Integration & Optional Minimal Tuning")

This timeline is feasible due to leveraging existing, powerful SahabatAI models, reducing the LLM development burden.[1] The focus shifts to KB quality, RAG pipeline robustness, and effective prompting.

5.2 Resource Requirements

- **Computational Resources:**
    - **GPU Access:** Needed for embedding generation, vector search, and inference with the SahabatAI 9B/8B model. If minimal SFT/RAFT is done, GPU requirements for QLoRA are manageable (e.g., Colab Pro, university HPC).[24]
    - **CPU and RAM:** Standard development machine.
- **Software and Libraries:** Python, PyTorch, Hugging Face ecosystem (Transformers, Datasets, PEFT, TRL), sentence-transformers, FAISS.
- **Datasets and Models:** Public MoFA data. Access to selected SahabatAI model from Hugging Face.[4]
- **Human Resources:** Student researcher. Native Bahasa Indonesia speakers for data validation/evaluation.

5.3 Risk Assessment and Mitigation Strategies

- **Risk 1: Difficulty Acquiring/Structuring Consular Data:** (Same as original)
    - Mitigation: Early start, focus on core services, automated tools + manual review.[69]
- **Risk 2: Suboptimal Performance of SahabatAI Model for Specific Consular Nuances (even with RAG):**
    - Description: The existing instruction-tuning of SahabatAI might not perfectly

cover all specific terminologies, tones, or complex reasoning required for some consular queries.
    - Mitigation: Intensive prompt engineering. As a contingency, conduct minimal, targeted SFT/RAFT if critical gaps persist and time allows.[39]
- **Risk 3: RAG System Underperformance (Retrieval or Generation Integration):** (Same as original)
    - Mitigation: Experiment with chunking/embeddings, refine prompts. Consider RAFT for the SahabatAI model to better use retrieved context.[39]
- **Risk 4: Time Constraints:** (Same as original, but slightly reduced by not building LLM from scratch)
    - Mitigation: Focused scope, prioritize core RAG functionality, monitor progress.
- **Risk 5: API/Access Issues for SahabatAI Models (Unlikely as they are on Hugging Face):**
    - Description: Unexpected changes to access terms for SahabatAI models.
    - Mitigation: Rely on current open access via Hugging Face [4] and Gemma Community License.[1] Have an alternative open-source multilingual model as a distant fallback.

Chapter 6: Expected Outcomes and Contributions

6.1 Novel Contributions to AI in Government and Consular Services

1. **A Functional Prototype Demonstrating Application of SahabatAI for Consular Services:** A working RAG-based chatbot prototype integrating an existing advanced Indonesian LLM from the SahabatAI ecosystem [1] to address MoFA consular inquiries.
2. **Methodological Framework for Applying Local LLM Ecosystems to Public Service RAG:** Documented approach for leveraging existing, regionally-focused LLMs (like SahabatAI) within a RAG framework for a specific public service domain in a low/mid-resource language context.
3. **Practical Insights for MoFA on Utilizing Local AI Capabilities:** Insights on the benefits, challenges, and best practices for MoFA and other Indonesian public sector entities in adopting locally developed AI model ecosystems like SahabatAI.

6.2 Advancements in LLM Application for Bahasa Indonesia

1. **Empirical Performance Data of SahabatAI in a Consular RAG System:** Results on how well a specific SahabatAI instruction-tuned model performs within a RAG pipeline for Indonesian consular tasks, complementing its existing general benchmark scores.[1]

2. **Curated Resources for Indonesian Consular NLP:** The developed consular knowledge base and the fine-tuning/evaluation datasets (if any further tuning is done) tailored for applying SahabatAI to this domain.

6.3 Limitations of the Proposed Research

(Largely same as original, but emphasizing that the LLM itself is from an external initiative)

- Prototype nature; limited scope of services; simulated evaluation.
- Dependence on publicly available MoFA data.
- Performance is tied to the capabilities and characteristics of the chosen existing SahabatAI model.
- Focus on standard Bahasa Indonesia.

6.4 Directions for Future Work

(Largely same as original)

1. Expansion of KB and service coverage.
2. Integration with MoFA internal systems.
3. Multi-modal capabilities.
4. Longitudinal user studies.
5. Further specialized fine-tuning of SahabatAI models for even more specific government tasks.
6. Continuous learning and adaptation of the RAG system with SahabatAI.
7. Advanced conversational AI features.

This revised proposal now accurately reflects the plan to use the existing SahabatAI models and focuses the research effort on their effective integration and application in the consular domain via RAG.

(References section would be updated to include specific SahabatAI model cards from Hugging Face and any related publications by GoToCompany/Indosat/AI Singapore if available, in addition to the previously listed academic and official sources.)

Works cited

1. gemma-2-9b-cpt-sahabatai-instruct Model by Gotocompany - NVIDIA NIM APIs, accessed May 19, 2025, https://build.nvidia.com/gotocompany/gemma-2-9b-cpt-sahabatai-instruct/modelcard
2. GoToCompany/gemma2-9b-cpt-sahabatai-v1-base - Hugging Face, accessed May 19, 2025, https://huggingface.co/GoToCompany/gemma2-9b-cpt-sahabatai-v1-base

3. GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct - Hugging Face, accessed May 19, 2025, https://huggingface.co/GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct
4. GoToCompany (PT GoTo Gojek Tokopedia Tbk) - Hugging Face, accessed May 19, 2025, https://huggingface.co/GoToCompany
5. Case Studies of How Generative Artificial Intelligence is Changing Work | Congress.gov, accessed May 19, 2025, https://www.congress.gov/crs-product/IN12537
6. ash.harvard.edu, accessed May 19, 2025, https://ash.harvard.edu/wp-content/uploads/2024/02/artificial_intelligence_for_citizen_services.pdf
7. 12 Global Government Agencies That Use Chatbots - Blog - V-Soft Consulting, accessed May 19, 2025, https://blog.vsoftconsulting.com/blog/15-governments-agencies-that-use-chatbots
8. 10 ways AI-powered public services are driving innovation in government - Rocket.Chat, accessed May 19, 2025, https://www.rocket.chat/blog/innovation-in-government
9. AI use cases in government that improve the citizen experience - The Future of Commerce, accessed May 19, 2025, https://www.the-future-of-commerce.com/2025/02/20/examples-of-ai-use-cases-in-government/
10. How Governments are Using AI: 8 Real-World Case Studies, accessed May 19, 2025, https://blog.govnet.co.uk/technology/ai-in-government-case-studies
11. Foreign Ministry to Utilize AI-Based Services for Indonesian Citizens ..., accessed May 19, 2025, https://en.tempo.co/read/1975389/foreign-ministry-to-utilize-ai-based-services-for-indonesian-citizens-abroad
12. Peduli WNI - Kemenlu, accessed May 19, 2025, https://peduliwni.kemlu.go.id/
13. Safe Travel - Apps on Google Play, accessed May 19, 2025, https://play.google.com/store/apps/details?id=id.go.kemlu.safetravel
14. Safe Travel – Apps on Google Play, accessed May 19, 2025, https://play.google.com/store/apps/details?id=id.go.kemlu.safetravel&hl=en_GB
15. Indonesian Ministry Of Foreign Affairs And UN Women Launch Humanist AI SARI Chatbots, Equipped With Various Regional Languages - VOI, accessed May 19, 2025, https://voi.id/en/news/477218
16. Indonesia, UN Women Launch AI Chatbot SARI to Protect Female Migrant Workers - Jakarta Daily, accessed May 19, 2025, https://www.jakartadaily.id/international/16214999370/indonesia-un-women-launch-ai-chatbot-sari-to-protect-female-migrant-workers
17. MMPTF Key Results Infosheet - IOM Indonesia, accessed May 19, 2025, https://indonesia.iom.int/sites/g/files/tmzbdl1491/files/documents/2025-04/key-results-infosheet_mmptf_eng.pdf
18. Ministry Of Foreign Affairs Ready To Launch SARI Chatbot, This Is A List Of Ministries That Already Have Chatbots - VOI, accessed May 19, 2025,

https://voi.id/en/technology/460202
19. Migration MPTF Final Report, accessed May 19, 2025, https://mptf.undp.org/sites/default/files/documents/2025-02/final_report_2024_migration_mptf_indonesia.pdf
20. Indonesia Diplomacy in Protecting the Rights of Indonesian Migrant Domestic Workers in Malaysia (2017-2022) - CORE, accessed May 19, 2025, https://core.ac.uk/download/617687082.pdf
21. Opportunities in Indonesia, accessed May 19, 2025, https://vm.ee/sites/default/files/documents/2025-03/Indonesia_Strategy_ICT.pdf
22. Opportunities in Indonesia, accessed May 19, 2025, https://vm.ee/sites/default/files/documents/2025-03/Indonesia_Strategy_Smart%20Industry%20and%20Electronics.pdf
23. What is Retrieval Augmented Generation(RAG) in 2025? - Glean, accessed May 19, 2025, https://www.glean.com/blog/rag-retrieval-augmented-generation
24. Fine-Tune Gemma using Hugging Face Transformers and QloRA ..., accessed May 19, 2025, https://ai.google.dev/gemma/docs/core/huggingface_text_finetune_qlora
25. Fine-tune Gemma in Keras using LoRA | Google AI for Developers - Gemini API, accessed May 19, 2025, https://ai.google.dev/gemma/docs/core/lora_tuning
26. The Ultimate Guide to Fine-Tune LLaMA 3, With LLM Evaluations ..., accessed May 19, 2025, https://www.confident-ai.com/blog/the-ultimate-guide-to-fine-tune-llama-2-with-llm-evaluations
27. LLaMA 2 Fine Tuning: Building Your Own LLaMA, Step by Step - Run:ai, accessed May 19, 2025, https://www.run.ai/guides/generative-ai/llama-2-fine-tuning
28. What is Retrieval-Augmented Generation (RAG)? | Google Cloud, accessed May 19, 2025, https://cloud.google.com/use-cases/retrieval-augmented-generation
29. A Comprehensive Review of Retrieval-Augmented Generation (RAG): Key Challenges and Future Directions - arXiv, accessed May 19, 2025, https://arxiv.org/pdf/2410.12837?
30. What is Retrieval Augmented Generation (RAG)? - Aisera, accessed May 19, 2025, https://aisera.com/blog/retrieval-augmented-generation-rag/
31. RAG for Chatbots: Best Practices & SearchUnify Virtual Assistant, accessed May 19, 2025, https://www.searchunify.com/sudo-technical-blogs/best-practices-for-using-retrieval-augmented-generation-rag-in-ai-chatbots/
32. Vertex AI RAG Engine overview - Google Cloud, accessed May 19, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/rag-engine/rag-overview
33. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications, accessed May 19, 2025, https://www.mdpi.com/2076-3417/15/8/4234
34. Retrieval-Augmented Text Generation: Methods, Challenges, and Applications[v1], accessed May 19, 2025, https://www.preprints.org/manuscript/202504.0443/v1
35. Retrieval Augmented Generation: An Evaluation of RAG-based Chatbot for Customer Support - UTUPub, accessed May 19, 2025,

https://www.utupub.fi/bitstream/10024/178860/1/Retrieval_Augmented_Generation_An_Evaluation_of_RAG_based_Chatbot_for_Customer_Support.pdf
36. How Retrieval Augmented Generation (RAG) is Transforming Federal Website Search - RIVA Solutions, accessed May 19, 2025, https://rivasolutionsinc.com/insights/how-retrieval-augmented-generation-rag-is-transforming-federal-website-search/
37. arxiv.org, accessed May 19, 2025, https://arxiv.org/abs/2401.06800
38. RAG chatbot: What it is, benefits, challenges, and how to build one - Tonic.ai, accessed May 19, 2025, https://www.tonic.ai/guides/rag-chatbot
39. RAFT: Adapting Language Model to Domain Specific RAG - arXiv, accessed May 19, 2025, https://arxiv.org/html/2403.10131v1
40. arxiv.org, accessed May 19, 2025, https://arxiv.org/pdf/2403.10131
41. [2403.10131] RAFT: Adapting Language Model to Domain Specific RAG - arXiv, accessed May 19, 2025, https://arxiv.org/abs/2403.10131
42. Ethical and Responsible AI Adoption in Government - REI Systems, accessed May 19, 2025, https://www.reisystems.com/roadmap-to-transformation-the-next-generation-of-government-operations-with-ethical-and-responsible-ai-adoption/
43. United States Citizenship and Immigration Services – AI Use Cases - Homeland Security, accessed May 19, 2025, https://www.dhs.gov/ai/use-case-inventory/uscis
44. Department of State AI Inventory 2024, accessed May 19, 2025, https://2021-2025.state.gov/department-of-state-ai-inventory-2024/
45. (PDF) Visa-Vanguard: An Nlp-Based, Aipowered Intelligent Chatbot ..., accessed May 19, 2025, https://www.researchgate.net/publication/389174173_Visa-Vanguard_An_Nlp-Based_Aipowered_Intelligent_Chatbot_for_Streamlining_International_Student_Support_Services
46. Indonesia to launch AI for food, social aid in August 2025 - Tech in Asia, accessed May 19, 2025, https://www.techinasia.com/news/indonesia-launch-ai-food-social-aid-august-2025
47. Artificial Intelligence at the Core of Indonesia's Digital Future - OpenGov Asia, accessed May 19, 2025, https://opengovasia.com/2025/04/21/artificial-intelligence-at-the-core-of-indonesias-digital-future/
48. Indonesia: Comprehensive Regulation for Responsible AI - OpenGov Asia, accessed May 19, 2025, https://opengovasia.com/2025/01/08/indonesia-comprehensive-regulation-for-responsible-ai/
49. (PDF) Implementation of AI-Based Chatbots to Enhance Efficiency ..., accessed May 19, 2025, https://www.researchgate.net/publication/391396953_Implementation_of_AI-Based_Chatbots_to_Enhance_Efficiency_and_Transparency_in_Land_Certification_in_Indonesia
50. arxiv.org, accessed May 19, 2025, https://arxiv.org/html/2410.07830v1

51. [2410.07830] NusaMT-7B: Machine Translation for Low-Resource Indonesian Languages with Large Language Models - arXiv, accessed May 19, 2025, https://arxiv.org/abs/2410.07830
52. NusaMT-7B: Machine Translation for Low-Resource Indonesian Languages with Large Language Models - arXiv, accessed May 19, 2025, http://www.arxiv.org/pdf/2410.07830
53. Leveraging Retrieval-Augmented Generation for Swahili Language Conversation Systems, accessed May 19, 2025, https://www.mdpi.com/2076-3417/15/2/524
54. NusaCrowd: Open Source Initiative for Indonesian NLP Resources - ACL Anthology, accessed May 19, 2025, https://aclanthology.org/2023.findings-acl.868.pdf
55. Cracking the Code: Multi-domain LLM Evaluation on Real-World Professional Exams in Indonesia - arXiv, accessed May 19, 2025, https://arxiv.org/html/2409.08564v2
56. Cracking the Code: Multi-domain LLM Evaluation on Real-World Professional Exams in Indonesia - ACL Anthology, accessed May 19, 2025, https://aclanthology.org/2025.naacl-industry.69.pdf
57. [Literature Review] NusaMT-7B: Machine Translation for Low-Resource Indonesian Languages with Large Language Models - Moonlight, accessed May 19, 2025, https://www.themoonlight.io/review/nusamt-7b-machine-translation-for-low-resource-indonesian-languages-with-large-language-models
58. arxiv.org, accessed May 19, 2025, https://arxiv.org/pdf/2410.07830
59. arxiv.org, accessed May 19, 2025, https://arxiv.org/pdf/2409.08564
60. arXiv:2504.01667v2 [cs.CL] 3 Apr 2025, accessed May 19, 2025, https://arxiv.org/pdf/2504.01667?
61. aclanthology.org, accessed May 19, 2025, https://aclanthology.org/2025.findings-naacl.341.pdf
62. Do Language Models Understand Honorific Systems in Javanese? - arXiv, accessed May 19, 2025, https://arxiv.org/html/2502.20864v1
63. arXiv:2411.17637v2 [cs.CL] 1 Mar 2025, accessed May 19, 2025, https://arxiv.org/pdf/2411.17637?
64. Welcome to the Indonesian Directorate General of Immigration Website - Direktorat Jenderal Imigrasi, accessed May 19, 2025, https://www.imigrasi.go.id/index?lang=en-US
65. Ditjen Imigrasi Logo & Brand Assets (SVG, PNG and vector) - Brandfetch, accessed May 19, 2025, https://brandfetch.com/imigrasi.go.id
66. Hubungi – Direktorat Jenderal Imigrasi, accessed May 19, 2025, https://www.imigrasi.go.id/hubungi/lokasi-kami
67. Data Protection & Privacy 2025 - Indonesia - Global Practice Guides, accessed May 19, 2025, https://practiceguides.chambers.com/practice-guides/data-protection-privacy-2025/indonesia/trends-and-developments
68. National data protection authority in Indonesia, accessed May 19, 2025, https://www.dlapiperdataprotection.com/?t=authority&c=ID

69. Best Open-source Web Scraping Libraries in 2025 - Firecrawl, accessed May 19, 2025, https://www.firecrawl.dev/blog/best-open-source-web-scraping-libraries
70. Top 5 Open Source Data Scraping and Ingestion Tools for RAG - Athina AI Hub, accessed May 19, 2025, https://hub.athina.ai/top-5-open-source-scraping-and-ingestion-tools/
71. Building a RAG Evaluation Dataset: A Step-By-Step Guide Using Document Sources, accessed May 19, 2025, https://magnimindacademy.com/blog/building-a-rag-evaluation-dataset-a-step-by-step-guide-using-document-sources/
72. How AI Chatbots Improve Local Government Services - Velaro, accessed May 19, 2025, https://velaro.com/blog/ai-chatbots-in-local-government-engaging-citizens-across-cities-and-states
73. Indonesia | Global AI Ethics and Governance Observatory - UNESCO, accessed May 19, 2025, https://www.unesco.org/ethics-ai/en/indonesia
74. 12 Essential Chatbot Performance Metrics & KPIs for 2025 - Calabrio, accessed May 19, 2025, https://www.calabrio.com/wfo/contact-center-ai/key-chatbot-performance-metrics/
75. Designing a Performance Metrics Evaluation Framework for NLP-Driven Chatbots in Local Government Unit | Journal of Information Systems Engineering and Management, accessed May 19, 2025, https://jisem-journal.com/index.php/journal/article/view/8535
76. Utilization chatbot for Indonesian tourism: a post-pandemic solution of information accessibility | Fibriasari | JPPI (Jurnal Penelitian Pendidikan Indonesia), accessed May 19, 2025, https://jurnal.iicet.org/index.php/jppi/article/view/3791
77. Evaluation of the Effect of Chatbot in Improving Customer Interaction and Satisfaction in Online Marketing in Indonesia - ResearchGate, accessed May 19, 2025, https://www.researchgate.net/publication/390009108_Evaluation_of_the_Effect_of_Chatbot_in_Improving_Customer_Interaction_and_Satisfaction_in_Online_Marketing_in_Indonesia
78. Building ChatBots: Powering Business Operations With AI | One Beyond, accessed May 19, 2025, https://one-beyond.com/building-chatbots-with-ai/