THESIS PROPOSAL

A Retrieval-Augmented Generation Approach Using Fine-Tuned SahabatAI for Indonesian Consular Chatbot

Abstract

Background of Research: The provision of timely and accurate consular information and assistance is a critical function of the Indonesian Ministry of Foreign Affairs (MoFA) for its citizens residing or traveling abroad. Current digital platforms, such as the "Peduli WNI" website and the "SARI" chatbot, offer valuable support. However, opportunities exist to significantly enhance service delivery through the application of advanced Artificial Intelligence (AI). The advent of Large Language Models (LLMs) specifically developed for Bahasa Indonesia, notably SahabatAI, offers a promising foundation for creating more sophisticated, responsive, and context-aware consular service tools. This research addresses the need for improved accessibility and accuracy in consular information dissemination.

The Objective of Research: This research aims to design, develop, and rigorously evaluate a prototype chatbot for Indonesian consular services. The proposed system will leverage the SahabatAI LLM, augmented by a Retrieval-Augmented Generation (RAG) framework. The central objective is to enhance the accuracy, comprehensiveness, and user accessibility of consular information provided to Indonesian citizens, thereby improving their overall experience and support when abroad.

Methods: The methodology encompasses several key stages: (1) Curation of a comprehensive and authoritative knowledge base derived from official MoFA sources, including the "Peduli WNI" portal and other pertinent consular documents and regulations. (2) Implementation of a RAG architecture, wherein a retriever module identifies and fetches relevant information segments from the curated knowledge base in response to user queries. Subsequently, the SahabatAI model will generate coherent, contextually appropriate, and factually grounded responses in Bahasa Indonesia. (3) Exploration of fine-tuning the SahabatAI model on a specialized consular domain corpus, contingent upon data availability and the constraints of the research timeline. (4) Systematic evaluation of the chatbot's performance using a combination of automated metrics (e.g., ROUGE, BLEU for generative aspects; Precision, Recall, F1-score for retrieval efficacy) and human assessment (evaluating accuracy, fluency, relevance, and helpfulness). This evaluation will include

comparisons against existing MoFA solutions or appropriate baseline models.

Expected Result: The anticipated outcome of this research is a functional prototype of an AI-powered consular chatbot that demonstrates superior performance in providing relevant, reliable, and timely information compared to existing systems. This work is expected to contribute a practical AI application for the enhancement of public services in Indonesia. Furthermore, it will showcase the capabilities of the SahabatAI LLM within a specialized governmental domain and provide valuable insights into the effective design and implementation of RAG systems for government services, particularly in non-English linguistic contexts.

1. Introduction

1.1. Detailed Background of Research

The Ministry of Foreign Affairs (MoFA) of the Republic of Indonesia plays an indispensable role in safeguarding the rights, interests, and welfare of Indonesian citizens (Warga Negara Indonesia - WNI) who are located overseas. These consular services encompass a wide spectrum of support, including assistance with legal predicaments, issuance and renewal of critical documentation such as passports and visas, guidance and support during emergency situations (natural disasters, civil unrest, personal crises), and the dissemination of general information pertinent to living or traveling abroad [MoFA Consular Guidelines, 2020]. The effective delivery of these services is paramount to ensuring the security and well-being of millions of Indonesians globally.

However, the delivery of consular services faces persistent challenges. Ensuring round-the-clock accessibility across different time zones, managing potentially high volumes of inquiries, maintaining consistency and accuracy of information across diverse communication channels, and catering to the nuanced needs of various citizen demographics (e.g., migrant workers, students, tourists, business travelers) are significant operational hurdles. Human-operated services, while essential for complex cases, are resource-intensive and may not always be immediately available.

In response to these challenges, MoFA has progressively adopted digital technologies. Key initiatives include:

- The **"Peduli WNI" portal**, a web-based platform serving as a primary repository for consular information, travel advisories, and self-help resources. While comprehensive, its nature as a website often necessitates users to manually navigate and search for specific information, which can be time-consuming and

may not directly address complex or multi-faceted queries (https://peduliwni.kemlu.go.id/beranda.html).
- The **"Safe Travel" application**, a mobile platform designed to provide Indonesian citizens with travel advisories, registration facilities for overseas travel, and emergency assistance features.
- The **"SARI" (Sahabat Artifisial Migran Indonesia) chatbot**. According to news reports, SARI was specifically developed to "melengkapi celah informasi bagi pekerja migran di luar negeri" (complete information gaps for migrant workers abroad). This targeted focus suggests that while SARI is a valuable step towards interactive assistance, its scope might be specialized, and its underlying Natural Language Processing (NLP) technology may predate the capabilities of current-generation Large Language Models.

The progression from static web portals like Peduli WNI to a more targeted, interactive system like SARI indicates an institutional learning curve within MoFA. This journey reflects an increasing ambition to leverage digital tools for enhanced, citizen-centric service delivery. Peduli WNI established a foundational digital presence, primarily enabling one-way information dissemination. The introduction of SARI marked a shift towards interactive engagement, albeit potentially with limitations inherent in earlier NLP or rule-based systems, and critically, it was designed to fill "information gaps," implying that previous systems were insufficient for particular user segments or query types. The current proposal to utilize SahabatAI with a Retrieval-Augmented Generation (RAG) framework represents a significant leap towards employing cutting-edge AI, capable of more nuanced language understanding, sophisticated reasoning, and dynamic information retrieval. This evolutionary trajectory suggests that MoFA is iteratively refining its digital service capabilities, learning from past implementations, and aiming for more comprehensive and intelligent solutions. Such a trend could also signify a broader strategic impetus within the Indonesian government to adopt advanced AI, with MoFA potentially acting as a pioneering agency whose success could inspire similar initiatives across other ministries, contributing to a wider national digital transformation.

The transformative potential of Artificial Intelligence (AI), particularly Large Language Models (LLMs) and the Retrieval-Augmented Generation (RAG) technique, offers a powerful avenue to address the aforementioned challenges. These technologies can enable the development of conversational agents (chatbots) that provide more natural, efficient, accurate, and personalized information delivery.

Of particular significance to this research is **SahabatAI**, heralded as "The First Large Language Model In Bahasa Indonesia". Developed by GoToCompany, SahabatAI represents a crucial advancement in AI capabilities tailored to the Indonesian linguistic and cultural context (https://sahabat-ai.com/, https://huggingface.co/collections/GoToCompany/sahabat-ai-672af7b248f5fdfd39ae2403). The development and availability of such a national LLM are not merely technical achievements; they also reflect a strategic national interest in fostering sovereign AI capabilities. Utilizing SahabatAI for a critical public service like consular affairs aligns with potential national objectives of digital autonomy, technological self-sufficiency, and culturally attuned service delivery. Employing a national LLM can ensure better handling of linguistic nuances specific to Bahasa Indonesia, potentially improve data sovereignty depending on deployment models, and reduce reliance on foreign technology providers for critical national infrastructure. The successful application of SahabatAI within MoFA could thereby encourage further investment in and development of Indonesian-specific AI models and applications, fostering a local AI ecosystem and contributing to bridging the digital divide by offering advanced services in the national language.

This research proposes to harness the capabilities of SahabatAI within a RAG framework to develop an advanced consular service chatbot, aiming to significantly improve upon existing MoFA digital offerings.

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

The application of AI in public administration and government services has seen considerable growth, with a focus on enhancing citizen-facing solutions such as chatbots, virtual assistants, and intelligent information portals. These tools aim to improve service efficiency, accessibility, and user satisfaction.

Chatbot technology itself has undergone a significant evolution. Early systems were predominantly rule-based, relying on predefined scripts and keyword matching, making them rigid and limited in their conversational abilities. Subsequent advancements led to machine learning-based chatbots, which could learn from data to understand user intents and generate more flexible responses [Jurafsky and Martin, 2023]. The current state-of-the-art is dominated by LLM-powered conversational AI, which exhibits unprecedented capabilities in natural language understanding and generation.

LLMs are typically based on the Transformer architecture, introduced by Vaswani et al. (2017), which utilizes self-attention mechanisms to process input sequences effectively. These models are pre-trained on vast amounts of text data, enabling them to learn intricate patterns of language. Popular pre-training paradigms include

Masked Language Modeling (MLM), as seen in BERT, and Causal Language Modeling (CLM), characteristic of the GPT series. Following pre-training, LLMs can be fine-tuned on smaller, task-specific datasets to adapt them to particular applications.

The emergence of regional and national LLMs, such as SahabatAI for Bahasa Indonesia, is a critical development. These models hold potential advantages in understanding local context, cultural nuances, and linguistic idiosyncrasies that global, predominantly English-trained models might miss [Aji et al., 2022]. The use of SahabatAI is central to this proposal, aiming to leverage its specific strengths for the Indonesian context.

A key challenge with LLMs is their propensity to "hallucinate" or generate factually incorrect information, especially for knowledge-intensive tasks [Ji et al., 2023]. Retrieval-Augmented Generation (RAG) has emerged as a powerful technique to mitigate this issue by grounding LLM responses in factual, up-to-date information retrieved from external knowledge bases [Lewis et al., 2020]. RAG systems combine a retriever module, which fetches relevant documents or passages from a corpus, with a generator module (an LLM), which synthesizes an answer based on the query and the retrieved context. This approach is particularly well-suited for domains like consular services, where accuracy and reliance on specific official documents are paramount.

Ethical considerations are crucial in the deployment of AI for public services. Issues such as algorithmic bias, fairness in service delivery, transparency of decision-making processes, data privacy, and accountability for AI-generated information must be carefully addressed [Floridi et al., 2018; European Commission High-Level Expert Group on AI, 2019]. Several foreign ministries and international organizations have begun exploring or implementing AI for consular support or citizen communication, providing valuable learning experiences.

Current research trends in conversational AI include the development of more robust multilingual models, techniques for low-resource language adaptation, advanced evaluation methodologies for generative models beyond simple n-gram overlap, and the increasing importance of domain-specific fine-tuning and meticulous knowledge base curation for specialized applications [Gao et al., 2023].

1.3. Gap Analysis

Despite MoFA's commendable efforts in digitalizing consular services, existing systems present certain limitations that this research aims to address:

- **Limitations of Existing MoFA Systems:**

    - **Peduli WNI Portal:** While serving as a comprehensive information repository, its static, website-based nature requires users to actively navigate and search for information. This process can be inefficient, particularly for users unfamiliar with the portal's structure or those needing quick answers to specific, potentially complex questions. It lacks interactivity and the capacity for personalized guidance based on individual user queries.
    - **SARI Chatbot:** The available information suggests SARI was developed to "melengkapi celah informasi" (complete information gaps), particularly for migrant workers. This implies that SARI may not offer comprehensive coverage of all consular topics or cater equally to all WNI demographics. Furthermore, if SARI is based on older NLP technologies (pre-LLM era), its natural language understanding capabilities, contextual awareness, and ability to handle diverse and complex user queries might be limited compared to what is achievable with current LLMs.
- **The Unmet Need:** There exists a discernible gap in providing a highly accessible, interactive, and consistently accurate 24/7 consular information service. Such a service should be capable of understanding and responding to a wide array of queries posed in natural Bahasa Indonesia, drawing its answers from the latest official MoFA knowledge base. Even with valuable resources like the Peduli WNI portal, a significant challenge often lies in the "last mile" of information delivery – ensuring that the *right* information reaches the *right* citizen at the *right* time and in an easily digestible format. The SARI chatbot represents an attempt to address this, but an LLM-RAG system can offer a substantially more sophisticated and effective solution to this persistent challenge. The Peduli WNI portal acts as a repository; however, users must actively navigate it, which can be a barrier due to cognitive load, time constraints, or varying search skills. SARI, while aiming to bridge this with an interactive layer, may be limited by its NLP capabilities and knowledge scope, as suggested by its stated purpose of filling "information gaps", potentially indicating it serves as a supplementary tool rather than a comprehensive solution. An LLM-RAG system, by combining advanced natural language understanding (via SahabatAI) with dynamic access to a broad, curated knowledge base (derived from Peduli WNI content and other official documents), directly tackles this "last mile" problem. It empowers users to ask complex questions in natural language and receive synthesized, relevant, and reliable answers. Successfully addressing this "last mile" problem in consular services can significantly improve citizen satisfaction, enhance trust in government services, and provide a model replicable across other public information services facing similar dissemination challenges.
- **How the Proposed Research Fills the Gap:**

- By leveraging **SahabatAI**, the proposed chatbot will possess strong natural language understanding and generation capabilities specifically for Bahasa Indonesia, ensuring more natural and effective user interactions.
- The **RAG framework** directly addresses the critical need for factual grounding. By retrieving information from a carefully curated MoFA knowledge base (constructed from authoritative sources like Peduli WNI and other official documents), the system ensures that responses are accurate, up-to-date, and trustworthy, rather than relying solely on the LLM's parametric knowledge which can be prone to inaccuracies or outdated information.
- This synergistic approach aims to provide a more comprehensive, reliable, interactive, and user-friendly solution than currently available through MoFA's existing digital tools, offering a qualitative leap in service delivery.

Table 1.1: Overview of Existing MoFA Digital Consular Services

Service Name	**Technology/ Platform**	**Key Features**	**Target Users**	**Data Sources (if known)**	**Identified Limitations/ Gaps (relevant to proposed research)**
Peduli WNI	Website	Information portal, travel advisories, self-registration for citizens abroad, news, FAQs.	All Indonesian citizens abroad or planning to travel	Official MoFA announcements, consular regulations, advisories	Static content, requires manual navigation/search, lacks interactivity, not optimized for conversational queries, potentially overwhelming for users seeking quick, specific answers.
Safe Travel	Mobile App	Travel advisories,	Indonesian travelers	MoFA travel advisories	Primarily focused on

		emergency contact, travel registration.			travel safety and emergency alerts, not a comprehensive consular information Q&A tool.
SARI Chatbot	Chatbot (underlying technology not publicly detailed)	Interactive Q&A, specifically mentioned to "complete information gaps for migrant workers abroad".	Primarily Indonesian migrant workers	Likely MoFA information pertinent to migrant workers	Potentially limited scope (migrant workers), may use older NLP technology limiting NLU and response generation capabilities, may not cover all consular topics comprehensively.

1.4. Problem Formulation (Research Questions)

This research seeks to answer the following key questions:

- **RQ1:** How can a Retrieval-Augmented Generation (RAG) system, utilizing the SahabatAI LLM, be effectively designed and developed to provide accurate, timely, and contextually relevant consular information for Indonesian citizens?
- **RQ2:** To what extent can the proposed RAG-SahabatAI chatbot improve upon the capabilities of existing MoFA digital consular services (e.g., SARI chatbot, Peduli WNI portal) in terms of response accuracy, relevance, comprehensiveness, and user-perceived helpfulness?
- **RQ3:** What are the optimal strategies for curating, preprocessing, and structuring a knowledge base from MoFA's official documents (including Peduli WNI content and other relevant public data) to effectively support the information retrieval component of the RAG system?

- **RQ4:** What performance metrics are most appropriate for evaluating the effectiveness of the RAG-SahabatAI consular chatbot, and how does its performance compare against relevant baseline models (e.g., SahabatAI without RAG, or potentially the SARI chatbot if a comparable evaluation framework can be established)?
- **RQ5 (Conditional upon feasibility):** If fine-tuning of SahabatAI on a specialized consular domain corpus is pursued, how does this impact the overall performance and quality of the RAG-based chatbot compared to using the pre-trained model?

1.5. Objective of Research (Research Objectives)

The primary objectives of this research are:

- **RO1:** To design and implement a functional prototype of a consular service chatbot employing a Retrieval-Augmented Generation (RAG) architecture, with the SahabatAI LLM serving as the core generation model.
- **RO2:** To curate, prepare, and structure a specialized knowledge base from official Indonesian consular service documents, including publicly available content from the Peduli WNI portal and other relevant MoFA publications, suitable for the RAG system's retrieval component.
- **RO3:** To rigorously evaluate the performance of the developed RAG-SahabatAI chatbot using a comprehensive set of quantitative metrics (e.g., ROUGE, BLEU for generation; Precision, Recall, F1-score, MRR for retrieval) and qualitative metrics (human evaluation assessing accuracy, coherence, relevance, and helpfulness).
- **RO4:** To compare the performance of the proposed RAG-SahabatAI system against relevant baselines (such as the existing SARI chatbot if a comparative assessment is feasible, and a non-RAG SahabatAI model) to empirically demonstrate its added value and improvements.
- **RO5:** To analyze the feasibility and potential impact of deploying such an advanced AI chatbot within the MoFA's consular service framework, considering the specific operational and user context in Indonesia, and to provide insights for future development.

1.6. Limitation

This research will be subject to the following limitations:

- **Scope of Consular Services:** Given the six-month timeframe for a Master's thesis, the prototype chatbot will likely focus on a well-defined subset of consular services (e.g., passport and visa information, procedures for reporting lost documents, common emergency contact procedures, frequently asked legal queries). It will not be feasible to cover the entirety of all possible consular scenarios.

- **Dataset Limitations:** The availability, quality, and format of public MoFA documents for constructing the knowledge base may pose challenges. Access to internal MoFA data, such as comprehensive query logs from the SARI chatbot or detailed internal procedural manuals, is not guaranteed and will be explored through formal channels if deemed critical, but the research design will primarily rely on publicly accessible information.
- **SahabatAI Model Access and Fine-tuning:** The extent to which SahabatAI can be fine-tuned will depend on the specific terms of access to the model for fine-tuning purposes, available computational resources, and the successful curation of a sufficiently large and high-quality domain-specific dataset. All these factors must be managed within the six-month constraint. The primary research focus will be on effectively leveraging the pre-trained SahabatAI model within the RAG framework; fine-tuning is a secondary, conditional objective.
- **Technical Complexity of RAG Implementation:** Implementing and optimizing a RAG system involves the integration and tuning of multiple complex components (retriever, vector database, generator). Debugging and fine-tuning the interplay between these components can be intricate and time-consuming.
- **Evaluation Challenges:** Comprehensive evaluation of generative AI systems is inherently challenging. While automated metrics will be employed, human evaluation is crucial for assessing nuanced aspects of response quality but can be resource-intensive to organize and execute. A direct, quantitative comparison with the existing SARI chatbot might be difficult if its internal workings, training data, and evaluation benchmarks are not publicly available.
- **Timeframe:** The six-month duration necessitates a highly focused scope and a pragmatic approach to development, experimentation, and analysis. Some exploratory avenues or advanced optimizations may be deferred to future work.

1.7. Hypothesis (Optional)

The following hypotheses will be investigated:

- **H1:** A consular chatbot developed using a Retrieval-Augmented Generation (RAG) framework with the SahabatAI LLM will achieve significantly higher accuracy, relevance, and comprehensiveness in responding to user queries concerning Indonesian consular services compared to a baseline SahabatAI model operating without RAG, and potentially demonstrate improvements over the existing SARI chatbot (if comparable evaluation is possible).

- **H2:** The utilization of a curated knowledge base, derived from official MoFA sources such as the Peduli WNI portal and other authenticated documents, within the RAG system will lead to more factually grounded, reliable, and trustworthy responses than those generated by a generic LLM or an LLM without real-time access to this specific external knowledge.

2. Detail Theory / Literature Review

This chapter reviews the existing literature pertinent to the core technologies and application domain of this research. It covers Large Language Models (LLMs), the SahabatAI model, Retrieval-Augmented Generation (RAG), chatbots in public services, and MoFA's current digital service offerings.

2.1. Review of Existing Work in the Research Area

2.1.1. Large Language Models (LLMs)
LLMs represent a paradigm shift in natural language processing, built upon foundational concepts of neural networks and deep learning, particularly sequence-to-sequence architectures. The Transformer architecture, introduced by Vaswani et al. (2017) in their seminal paper "Attention Is All You Need," is central to most modern LLMs. Its core innovation, the self-attention mechanism, allows models to weigh the importance of different words in an input sequence when processing information, leading to superior understanding of long-range dependencies and context.
LLMs are typically developed through a two-stage process: pre-training and fine-tuning. **Pre-training** involves training the model on massive unlabeled text corpora (often terabytes of data) using self-supervised learning objectives. Common pre-training paradigms include:

- **Masked Language Modeling (MLM):** Used in models like BERT (Bidirectional Encoder Representations from Transformers), where the model learns to predict randomly masked words in a sentence based on their surrounding context.
- **Causal Language Modeling (CLM):** Employed by autoregressive models like the GPT (Generative Pre-trained Transformer) series, where the model learns to predict the next word in a sequence given the preceding words.

Following pre-training, LLMs acquire a broad understanding of language structure, grammar, and a vast amount of world knowledge. **Fine-tuning** then adapts these pre-trained models to specific downstream tasks (e.g., text classification, question answering, summarization) or specialized domains by training them further on smaller, labeled datasets relevant to the target application.

Despite their remarkable capabilities, LLMs face several **challenges**. These include:

- **Hallucinations:** Generating plausible but factually incorrect or nonsensical information [Ji et al., 2023].
- **Bias:** Reflecting and potentially amplifying societal biases present in their training data.
- **Computational Cost:** Requiring significant computational resources for training and, for very large models, even inference [Patterson et al., 2021].
- **Ethical Implications:** Concerns regarding misuse, lack of transparency, and accountability [Weidinger et al., 2021].

2.1.2. SahabatAI: A National LLM for Bahasa Indonesia

SahabatAI is presented as the first large language model specifically for Bahasa Indonesia, developed by GoToCompany. While detailed specifics of its architecture or the exact composition of its training data may not be fully public, its existence signifies a crucial step towards building AI resources tailored for the Indonesian language and context. The model and associated resources are accessible via platforms like Hugging Face, which facilitates its adoption by researchers and developers (https://huggingface.co/collections/GoToCompany/sahabat-ai-672af7b248f5fdfd39ae2403).

The **significance** of such national or regional LLMs cannot be overstated. Models pre-trained primarily on English data often struggle with the nuances, cultural references, and specific linguistic structures of other languages. An LLM like SahabatAI, presumably trained on a substantial corpus of Indonesian text, is expected to have a more innate understanding of Bahasa Indonesia, leading to better performance in tasks involving this language [Aji et al., 2022; Wilie et al., 2020]. This can reduce dependency on translation layers or less effective cross-lingual transfer learning from English-centric models. The development of SahabatAI aligns with a global trend of creating foundational models for various languages to ensure more equitable access to AI advancements and to support digital sovereignty [Costa-jussà et al., 2022]. Potential applications of SahabatAI are vast, spanning various sectors in Indonesia. Any reported performance benchmarks or comparative studies involving SahabatAI against other multilingual models on Indonesian NLP tasks would be highly relevant for contextualizing its capabilities.

2.1.3. Retrieval-Augmented Generation (RAG)

RAG is an architectural approach designed to enhance the factual grounding and reliability of LLMs by integrating an explicit information retrieval step into the generation process [Lewis et al., 2020]. The core concept is to combine the parametric knowledge implicitly stored in the LLM's weights with non-parametric knowledge explicitly stored in an external corpus, accessed at inference time.

A typical RAG system consists of two main components:

1. **Retriever:** This module is responsible for finding relevant information from a large knowledge base (e.g., a collection of documents, a database of FAQs) given a user query. Common retriever types include:
   - **Sparse Retrievers:** Based on lexical matching, such as TF-IDF or BM25. These are often efficient and effective baselines.
   - **Dense Retrievers:** Based on semantic similarity using learned embeddings. Models like DPR (Dense Passage Retriever) [Karpukhin et al., 2020] use dual-encoder architectures (e.g., based on BERT) to embed queries and documents into a shared vector space, enabling semantic search. Sentence-BERT and similar models are also widely used for generating these embeddings.
2. **Generator:** This is typically a sequence-to-sequence LLM (e.g., BART [Lewis et al., 2019], T5, or in this proposal, SahabatAI).

The **mechanism** involves the retriever first fetching a set of relevant documents or passages (e.g., top-k results) from the knowledge base in response to the input query. These retrieved contexts are then concatenated with the original query and fed as input to the generator LLM, which synthesizes the final answer based on both the query and the provided factual information.

Advantages of RAG are significant for knowledge-intensive tasks:

- **Improved Factual Accuracy:** By conditioning generation on retrieved evidence, RAG significantly reduces the likelihood of hallucinations and grounds responses in verifiable facts.
- **Ability to Cite Sources:** The system can potentially indicate the source documents used for generation, enhancing transparency and trustworthiness.
- **Easier Knowledge Updating:** The LLM itself does not need to be retrained to incorporate new information. Instead, the external knowledge base can be updated, and the RAG system will immediately have access to the new facts [Petroni et al., 2020].
- **Mitigation of Hallucinations:** Directly addresses one of the major weaknesses of standalone LLMs.

Applications of RAG are diverse and growing, including open-domain question answering [Lewis et al., 2020], fact verification, dialogue systems requiring factual recall, and various other knowledge-intensive NLP tasks. For consular services, where providing accurate and current information is critical, RAG offers a highly promising approach.

2.1.4. Chatbots and Conversational AI in Public Services

The use of chatbots in government and public administration has seen a steady rise, evolving from simple FAQ bots to more sophisticated conversational AI systems [Longo, 2019; Cunha et al., 2020]. The primary motivation is to improve service delivery, enhance citizen engagement, and increase operational efficiency.

Benefits of AI-powered chatbots for public administration include:

- **24/7 Availability:** Providing instant support to citizens regardless of time or day.
- **Reduced Workload:** Automating responses to common queries, freeing up human agents to handle more complex or sensitive cases.
- **Consistent Information Delivery:** Ensuring that all citizens receive standardized and accurate information.
- **Improved Citizen Engagement:** Offering a more accessible and interactive channel for citizens to obtain information and services [Misuraca et al., 2020].

Numerous **case studies** demonstrate successful chatbot implementations in various government sectors globally. For instance, chatbots are used for tax inquiries (e.g., by the IRS in the USA or HMRC in the UK), healthcare information (e.g., symptom checkers or COVID-19 information bots by health ministries), processing applications for social benefits, and collecting citizen feedback.

However, deploying chatbots for **consular services** presents specific challenges:

- **Handling Sensitive Information:** Consular queries often involve personal data and sensitive situations, requiring robust security and privacy measures.
- **Dealing with Distressed Users:** Citizens seeking consular assistance may be in distress, requiring empathetic and carefully worded responses, which can be difficult for AI.
- **Ensuring Accuracy in Critical Situations:** Misinformation in consular matters (e.g., visa requirements, emergency procedures) can have severe consequences.
- **Multilingual Support:** For countries with diverse diaspora, supporting multiple languages may be necessary, although this proposal focuses on Bahasa Indonesia.

2.1.5. Existing MoFA Digital Services (Peduli WNI, SARI)

A closer examination of MoFA's existing digital platforms provides context for the proposed research.

- **Peduli WNI:** This portal (https://peduliwni.kemlu.go.id/beranda.html) serves as a central hub for information. Its content likely includes FAQs, guides on consular procedures (passports, visas, legalizations), travel advisories, news updates relevant to WNI abroad, and contact information for Indonesian embassies and consulates. The structure of this content (e.g., HTML pages, downloadable PDFs) will inform the data collection and preprocessing strategies for building the RAG system's knowledge base. The organization of information, headings, and paragraph breaks will be important for effective document chunking.

- **SARI Chatbot:** The Antaranews article states SARI's purpose is to "melengkapi celah informasi bagi pekerja migran di luar negeri." This targeted objective suggests its knowledge base and conversational flows might be optimized for issues pertinent to Indonesian migrant workers (e.g., employment contracts, remittance procedures, rights protection, specific emergency contacts for this group). While its underlying technology is not detailed, its existence indicates MoFA's recognition of the value of interactive AI. SARI can serve as an important baseline for comparison, at least qualitatively, and understanding its limitations can help highlight the advancements offered by the proposed LLM-RAG system. If any public-facing interaction logs or common query types addressed by SARI could be ethically accessed or inferred, they would be valuable for designing test cases.

2.2. Comparison of Various Methods, Models, and Approaches

The choice of an LLM-RAG architecture for the proposed consular chatbot is based on a comparative analysis of different technological options.

LLMs vs. Traditional NLP Chatbots:

- **Traditional NLP Chatbots:** These systems typically rely on rule-based approaches, keyword matching, or intent classification using earlier machine learning models (e.g., SVMs, Naive Bayes) with limited Natural Language Understanding (NLU).
  - *Pros:* Highly interpretable, behavior is predictable, relatively less data-intensive for simple tasks.
  - *Cons:* Brittle (fail on unseen queries or slight variations), hard to scale to many intents or complex dialogues, poor handling of nuanced language, require extensive manual effort for rule creation and maintenance.
- **LLM-based Chatbots:** These leverage the advanced NLU and Natural Language Generation (NLG) capabilities of LLMs.
  - *Pros:* More flexible and robust in understanding diverse user inputs, better generalization to unseen queries, can engage in more human-like and coherent conversations, capable of complex reasoning and summarization.
  - *Cons:* Can hallucinate facts if not properly grounded, less interpretable ("black box" nature), computationally more intensive, require large pre-trained models.

For a consular chatbot requiring nuanced understanding of diverse queries and the ability to generate informative, contextually relevant responses, LLM-based approaches offer significantly greater potential than traditional methods.

RAG vs. Other LLM Grounding/Augmentation Techniques:
Several techniques exist to provide LLMs with external knowledge or to adapt them to specific domains. RAG is chosen over alternatives due to its specific advantages for this project:

- **Full Fine-tuning on Domain Data:** This involves retraining the entire LLM or a significant portion of its parameters on a large, domain-specific corpus (e.g., all available consular documents and Q&A pairs).
  - *Pros:* Can deeply embed domain knowledge into the model's parameters, potentially leading to very fluent and knowledgeable responses within that domain.
  - *Cons:* Requires massive, high-quality domain-specific datasets, which may not be available for consular services. Extremely computationally expensive and time-consuming. Knowledge becomes static (embedded in weights) and requires complete retraining for updates. Susceptible to "catastrophic forgetting" of general knowledge if not done carefully [McClelland et al., 1995].
- **Prompt Engineering with Extensive Context (In-Context Learning):** Providing large amounts of relevant information directly within the LLM's input prompt at inference time, hoping the model uses this context to answer.
  - *Pros:* Conceptually simple, does not require model retraining.
  - *Cons:* Severely limited by the LLM's context window size (e.g., a few thousand tokens for many models). Inefficient for large knowledge bases, as only a small fraction can fit. The challenge of selecting the *most* relevant context to include in the prompt still remains (which RAG's retriever addresses systematically).
- **Retrieval-Augmented Generation (RAG):**
  - *Pros:* Dynamically incorporates external knowledge at inference time, allowing access to vast and up-to-date information. The knowledge base can be easily updated without retraining the LLM. Significantly reduces hallucinations by grounding responses in retrieved facts. Can cite sources, enhancing transparency. More scalable and cost-effective for knowledge updates than full fine-tuning [Lewis et al., 2020].
  - *Cons:* Adds the complexity of a retriever component. The overall system performance is heavily dependent on the quality of retrieval (if irrelevant documents are fetched, the generator may still produce poor answers).

Given the need for factual accuracy, the ability to access a potentially large and evolving knowledge base of consular information, and the desire to mitigate hallucinations, RAG emerges as the most suitable technique for the proposed consular chatbot. It offers a balance between leveraging the power of pre-trained LLMs like SahabatAI and ensuring responses are grounded in verified, domain-specific information.

Different RAG Component Choices:

- **Retrievers:**
  - *Sparse Retrievers (e.g., BM25, TF-IDF):* Based on lexical overlap. Fast and often strong baselines. Good for keyword-heavy queries.
  - *Dense Retrievers (e.g., DPR, Sentence-BERT based):* Use neural embeddings for semantic similarity. Better at capturing meaning and handling paraphrased queries but computationally more intensive for indexing and querying.
  - *Hybrid Approaches:* Combine sparse and dense methods to leverage the strengths of both. The choice will involve evaluating trade-offs in accuracy, speed, and implementation complexity. Dense retrievers are generally preferred for their semantic capabilities if resources allow.
- **Generators:** The choice of LLM is critical. This proposal focuses on SahabatAI due to its specialization in Bahasa Indonesia. The size and specific architecture of the chosen SahabatAI variant (if multiple exist) will impact generation quality, fluency, and computational requirements.

Analysis of AI Applications in Consular Services by Other Governments/International Bodies: A review of AI adoption by other foreign ministries or international organizations (e.g., UN agencies, EU bodies) for consular or citizen support services is instructive. For example:

- Some countries have deployed chatbots for visa applications, providing information on requirements and application status [Marr, 2020].
- Others use AI for crisis communication, disseminating alerts and information to citizens in affected areas [Chatbots Life, 2019].
- The technologies employed vary, from simpler rule-based systems to more advanced machine learning-driven platforms. The extent to which LLMs and RAG are currently used in operational consular systems globally is an area of active development and less documented in public literature compared to other government sectors. This review helps benchmark the proposed work, understand common challenges (e.g., data security, user trust), and identify best practices or innovative features that could be adapted.

2.3. Identification of Gaps and Justification for the Proposed Work

The preceding review and analysis highlight specific gaps that this research aims to fill:

- **Gaps in MoFA's Current Services:** As detailed in Section 1.3, MoFA's current digital consular services, Peduli WNI and the SARI chatbot, while valuable, have limitations in terms of interactivity, comprehensiveness, natural language understanding capabilities, and consistent factual grounding across a wide range of consular topics. There is a clear need for a more advanced, reliable, and user-friendly solution.
- **Novelty of the Proposed Approach:** While RAG is an established technique and LLMs are increasingly common, the specific application of a Bahasa Indonesia-native LLM like **SahabatAI combined with a RAG framework to the Indonesian consular service domain represents a novel contribution.** Much of the existing research and application of advanced conversational AI is focused on English or other resource-rich languages. This project pioneers such an application in the Indonesian context for a critical government service.

The **justification** for this research stems from its potential to significantly enhance consular service delivery for Indonesian citizens. The proposed system aims to provide:

- **Greater Accuracy and Reliability:** By grounding responses in an official, curated knowledge base via RAG.
- **Improved Accessibility:** Offering a 24/7 interactive channel that can understand natural language queries in Bahasa Indonesia.
- **Enhanced User Experience:** Providing quicker, more relevant, and comprehensive answers compared to manual website navigation or potentially limited existing chatbots.

The **contribution to the field** will be multi-fold:

- A practical demonstration of applying state-of-the-art AI (LLMs and RAG) to improve public administration in a non-English, developing country context.
- Specific insights into the challenges and best practices for developing and deploying such systems for specialized government services, particularly with national LLMs like SahabatAI.
- An open-source prototype (if feasible) or detailed methodology that could inform similar initiatives in Indonesia or other countries.

Official government information, such as that contained within the Peduli WNI portal and other MoFA documents, constitutes a unique and authoritative "knowledge moat." This verified, domain-specific information is precisely what consular services must rely upon. The RAG methodology is exceptionally well-suited to leverage this "knowledge moat." General-purpose LLMs, even a national one like SahabatAI in its pre-trained state, might not have ingested all the specific, detailed, and frequently updated consular information during their initial training. RAG allows the system to dynamically tap into this authoritative knowledge base at query time, ensuring that the responses generated are not only linguistically fluent but also factually grounded in official MoFA sources. This direct linkage to verified information significantly enhances the trust and reliability of the service, which are paramount considerations for consular assistance. This characteristic underscores a key strength of RAG for government applications: the ability to synergize the advanced linguistic capabilities of LLMs with the verified, domain-specific knowledge that governments uniquely possess and are mandated to disseminate. This model can serve as a template for other government agencies seeking to leverage their own "knowledge moats" for improved public service delivery.

Furthermore, the combination of SahabatAI and the RAG framework specifically addresses two critical challenges often encountered in the application of AI for public services in Indonesia. Firstly, it tackles the **linguistic barrier** by employing a model designed and trained for Bahasa Indonesia, ensuring nuanced understanding and culturally appropriate communication. Secondly, it addresses the **factual accuracy barrier** – a major concern with standalone LLMs which can "hallucinate" – by using RAG to ground responses in verified official data drawn from sources like Peduli WNI. This synergistic approach, therefore, offers a robust solution that simultaneously meets both the language-specific and knowledge-intensive requirements of providing effective consular services. This combination provides a practical and powerful template for developing trustworthy AI services in diverse linguistic contexts, particularly where reliance on official knowledge bases is critical, and demonstrates a tangible way to harness national LLM initiatives for concrete public benefit.

Table 2.1: Comparative Analysis of LLM Augmentation Techniques for Chatbots

Technique	**Principle of Operation**	**Pros**	**Cons**	**Data Requirements**	**Computational Cost**	**Suitability for Indonesian Consular Chatbot**
Standard LLM Prompting	LLM generates response	Simple to implement, no	Prone to hallucination,	None beyond pre-traine	Low (inference).	Low: Lacks factual

(Zero/Few-shot)	based solely on its pre-trained knowledge and the input prompt.	additional training needed.	knowledge may be outdated or generic, limited by context window for providing extensive background.	d LLM.		grounding and up-to-date domain knowledge essential for consular services.
Prompt Engineering (Extensive Context)	Provide significant relevant domain information directly in the LLM's input prompt.	No model retraining, can use latest information if manually inserted.	Severely limited by LLM context window size, inefficient for large knowledge bases, manual selection of context is difficult and not scalable.	Curated context snippets per query.	Moderate (inference, depending on context size).	Medium-Low: Impractical for a comprehensive consular knowledge base due to context limits and manual effort.
Full Fine-tuning on Domain Data	Re-train LLM (or parts of it) on a large corpus of domain-specific text (e.g., consular documents, Q&A pairs).	Can deeply embed domain knowledge, potentially high fluency in domain language.	Requires very large, high-quality domain dataset, computationally very expensive, knowledge becomes static until next re-training, risk of	Massive domain-specific corpus.	Very High (training), Low-Medium (inference).	Medium: Potentially high quality but very resource-intensive and difficult to keep updated. Data availability for comprehe

			catastrophic forgetting.			nsive fine-tuning is a major constraint.
Retrieval-Augmented Generation (RAG)	**LLM generates response based on input prompt AND relevant documents dynamically retrieved from an external knowledge base.**	**Reduces hallucination, uses up-to-date knowledge from corpus, scalable knowledge base, can cite sources, balances parametric and non-parametric knowledge.**	**Adds complexity of retriever, performance depends on retrieval quality, potential latency from retrieval step.**	**Curated external knowledge base (documents), smaller Q&A dataset for retriever tuning/evaluation.**	**Medium-High (retriever indexing & query, LLM inference).**	**High: Optimal choice. Directly addresses need for factual accuracy from official sources (e.g., Peduli WNI), allows knowledge updates, leverages SahabatAI for Bahasa Indonesia generation.**

Table 2.2: Selected SOTA AI Applications in Governmental Public/Consular Services

Country/Organization	Service Area	AI Technology Used	Key Features/Capabilities	Reported Impact/Outcomes (if available)	Source/Reference
Singapore Gov (AskJamie)	General Govt. Services Info	Virtual Assistant (NLP, ML)	Answers queries on various govt services,	Improved citizen access to info, reduced	

			directs users to resources.	calls to hotlines.	
US Citizenship and Immigration Services (EMMA)	Immigration Services	Virtual Assistant (NLP)	Answers questions about immigration services, forms, policies. Navigates users to relevant website sections.	Enhanced user experience, 24/7 availability.	
Estonia (various e-services)	Multiple Govt. Services	AI integrated into digital services (e.g., proactive services)	Automated processes, personalized service delivery.	High efficiency, citizen satisfaction.	[e-Estonia, n.d.]
Australian Taxation Office (ATO)	Tax Information	Virtual Assistants (Alex, an internal one)	Answers tax-related queries, assists with online services.	High volume of queries handled, improved compliance.	
UK Foreign, Commonwealth & Development Office (FCDO)	Travel Advice (experimental)	Exploring AI for analyzing travel risks, social media monitoring.	Early-stage exploration for enhanced travel advisories.	Potential for more dynamic risk assessment.	[UK Parliament, 2023]
Proposed MoFA System	*Indonesian Consular Services*	*LLM (SahabatAI) with RAG*	*Q&A on consular procedures, documentation, emergency support in Bahasa*	*Expected: Improved accuracy, accessibility, user satisfaction.*	*This Proposal*

			Indonesia, grounded in official MoFA knowledge (Peduli WNI etc.).		

3. Method

This chapter outlines the methodology that will be employed to achieve the research objectives. It details the research design, data collection and preprocessing procedures, the model development process focusing on the RAG architecture with SahabatAI, and the evaluation strategy.

3.1. Research Design

A **phased, iterative approach** will be adopted for this research, suitable for the development and evaluation of an AI prototype within a six-month Master's thesis timeframe. This approach allows for systematic progress, incorporation of learnings at each stage, and flexible adaptation to challenges.

- **Phase 1: Literature Review & Requirement Finalization (Month 1)**
  - Conduct an in-depth review of literature on LLMs, RAG architectures, SahabatAI, chatbot evaluation, and AI in consular/public services.
  - Refine the research questions, objectives, and scope based on literature findings and initial feasibility assessments.
  - Finalize the specific subset of consular services to be covered by the prototype.
- **Phase 2: Data Collection & Preparation (Months 1-2)**
  - Identify and gather relevant documents from official MoFA sources, primarily the "Peduli WNI" portal, but also other publicly available handbooks, regulations, and FAQs.
  - Preprocess the collected data: clean text, segment documents into appropriate chunks for retrieval, and extract relevant metadata.
  - Begin development of a gold-standard Question-Answer (Q&A) dataset for evaluation purposes.
- **Phase 3: RAG System Development & SahabatAI Integration (Months 2-4)**
  - Select and implement the retriever module (e.g., dense retriever using sentence embeddings).
  - Set up a vector database (e.g., FAISS) to store and index the embeddings of

the knowledge base chunks.

  - Integrate the pre-trained SahabatAI model as the generator component.
  - Develop the end-to-end RAG pipeline, including prompt engineering to guide SahabatAI in generating responses based on retrieved context.
  - Conduct initial unit testing of retriever and generator components.
- **Phase 4: (Optional, if feasible and beneficial) SahabatAI Fine-tuning (Month 3-4, concurrent with Phase 3)**
  - If a sufficiently large and high-quality consular-specific dataset (e.g., Q&A pairs, domain-specific texts) is curated from Phase 2, and if access and resources permit:
    - Explore limited fine-tuning of SahabatAI to enhance its performance on consular-specific language or tasks.
    - Evaluate the impact of fine-tuning compared to using the base pre-trained model.
- **Phase 5: Evaluation & Analysis (Months 4-5)**
  - Conduct comprehensive evaluations of the RAG-SahabatAI chatbot using both automated metrics and human assessment.
  - Compare the system's performance against defined baselines (e.g., non-RAG SahabatAI, potentially SARI if a comparative framework is possible).
  - Analyze the results, identify strengths and weaknesses, and draw conclusions regarding the research questions.
- **Phase 6: Thesis Writing & Refinement (Months 5-6)**
  - Document the entire research process, methodology, development, results, and discussion in the thesis.
  - Refine the thesis based on feedback and prepare for submission.

The **research paradigm** is primarily **experimental and developmental**. A prototype system will be designed, built, and empirically evaluated to test the hypotheses and answer the research questions concerning the efficacy of the proposed RAG-SahabatAI approach for Indonesian consular services.

Justification: This phased and iterative design is well-suited for an engineering-focused Master's thesis. It allows for managing the complexity of building an AI system by breaking it into manageable stages. The iterative nature permits adjustments based on findings in earlier phases, which is crucial for research involving novel applications of emerging technologies like SahabatAI in a specific domain. The focus is on developing a functional and evaluable prototype that demonstrates the core capabilities and potential of the proposed solution within the given timeframe.

3.2. Data Collection

The quality and comprehensiveness of the knowledge base are critical for the success of the RAG system. Data collection will focus on authoritative and publicly accessible sources.

- **Primary Data Sources for Knowledge Base:**
  - **Official MoFA Website - "Peduli WNI" Portal:** This will be the principal source. Content to be targeted includes:
    - Frequently Asked Questions (FAQs) sections.
    - Guides and informational articles on various consular services (e.g., passport renewal, visa applications, legal assistance, emergency procedures).
    - Announcements and official statements relevant to WNI abroad.
    - Descriptions of services offered by Indonesian embassies and consulates. The content will be systematically scraped from https://peduliwni.kemlu.go.id/beranda.html and its sub-pages.
  - **Publicly Available MoFA Documents:**
    - Consular affairs handbooks or manuals for citizens (if available online).
    - Publicly released consular regulations and decrees.
    - Press releases from MoFA related to citizen services and protection.
  - **Relevant Indonesian Laws and Regulations:** Key legislative texts pertaining to citizenship, immigration, and consular matters that are publicly accessible.
- **Secondary/Supplementary Data (Exploratory, subject to ethical and practical considerations):**
  - **Anonymized Query Logs or FAQs from SARI:** If MoFA is willing and able to share anonymized query data or a list of common questions addressed by the existing SARI chatbot, this would be invaluable for understanding common user needs and for creating realistic test queries. Formal requests and ethical approvals would be prerequisites.
  - **Public Forums and Social Media:** Online platforms where Indonesian citizens discuss consular experiences or ask related questions (e.g., expatriate forums, specific social media groups). This source would be used cautiously, primarily to understand the *types* of questions users ask and the language they use, not for sourcing factual content for the knowledge base.
- **Strategy for Knowledge Base Creation:**
  - Systematic web scraping of the Peduli WNI portal and other relevant MoFA web pages using tools like Scrapy or BeautifulSoup.
  - Manual collection and digitization (e.g., OCR for PDFs if necessary) of relevant documents not available in easily machine-readable formats.

    - The focus will be on collecting textual data that provides factual information, procedures, requirements, advice, and contact details related to Indonesian consular services.
- **Sample Dataset for Evaluation:**
    - A **gold-standard Question-Answer (Q&A) dataset** will be manually created. This dataset will consist of:
        - Representative consular questions covering a diverse range of topics (e.g., "Bagaimana cara memperpanjang paspor Indonesia di luar negeri?", "Apa yang harus saya lakukan jika kehilangan paspor saat bepergian?", "Informasi visa untuk negara X?").
        - Manually verified, ideal answers sourced directly from the official documents collected for the knowledge base.
    - This Q&A dataset will serve multiple purposes:
        - Testing the retriever's ability to find relevant passages.
        - Providing reference answers for automated generation metrics (e.g., ROUGE, BLEU).
        - Serving as the basis for human evaluation of the chatbot's responses.
    - The questions will be designed to reflect realistic user queries, varying in complexity and specificity.

3.3. Data Preprocessing

Once collected, the raw data must be preprocessed to make it suitable for ingestion into the RAG system.

- **Knowledge Base Preprocessing:**
    - **Cleaning:** Removing HTML tags, JavaScript code, CSS styles, irrelevant boilerplate text (e.g., headers, footers, navigation menus) from scraped web content. Standardizing character encodings.
    - **Formatting:** Converting all data into a consistent plain text format or structured format like JSON, where each entry might represent a document or a passage.
    - **Segmentation/Chunking:** This is a critical step for RAG. Large documents will be broken down into smaller, semantically coherent passages or chunks (e.g., paragraphs, logical sections, or fixed-size overlapping segments). The optimal chunking strategy (e.g., chunk size, overlap) will be determined experimentally, as it significantly impacts retrieval performance and the context provided to the LLM. Chunks should be small enough to fit within the LLM's context window along with the query, yet large enough to contain meaningful information.

    - **Metadata Extraction:** For each chunk, relevant metadata will be stored, such as the source document URL or name, original section heading, date of publication (if available), and any other contextual information. This metadata is crucial for traceability, potentially enabling the chatbot to cite its sources, and for filtering or prioritizing information during retrieval.
- **Query Preprocessing (for user input to the chatbot):**
    - Basic text cleaning may be applied to user queries, such as lowercasing, removing excessive punctuation or special characters, and normalizing whitespace. The extent of query preprocessing will depend on the robustness of the chosen retriever model to variations in input.
- **Vectorization/Embedding Generation:**
    - To enable semantic search, both the preprocessed knowledge base chunks and incoming user queries will be converted into dense vector representations (embeddings).
    - A pre-trained sentence transformer model will be used for this purpose. Options include:
        - Multilingual models like sentence-transformers/paraphrase-multilingual-mpnet-base-v2 which have good performance across many languages, including Bahasa Indonesia.
        - Potentially, a Bahasa Indonesia-specific sentence embedding model if a high-quality one is available.
        - Exploring the use of embeddings derived from SahabatAI itself, if its architecture supports generating sentence-level embeddings suitable for retrieval tasks.
    - The generated embeddings for all knowledge base chunks will be stored in a vector database for efficient similarity search.

3.4. Data Processing (Model Development & RAG Implementation)

This section details the development of the core RAG system, integrating SahabatAI.

3.4.1. SahabatAI LLM Integration
The primary generator model for this research will be the pre-trained SahabatAI LLM. Access to the model will likely be through the Hugging Face transformers library or other official distribution channels provided by GoToCompany. The focus will be on leveraging its existing generative capabilities in Bahasa Indonesia to synthesize answers based on the query and retrieved context. Different model sizes or variants of SahabatAI, if available, may be explored based on performance and computational constraints.
3.4.2. SahabatAI Fine-tuning (Conditional and Limited Scope)

While the primary approach relies on the pre-trained SahabatAI within the RAG framework, limited-scope fine-tuning will be explored as a secondary, conditional objective. This is contingent upon:

1. **Data Availability:** Successful curation of a sufficiently large (e.g., thousands of examples) and high-quality dataset of consular-specific Q&A pairs or domain-specific texts from the data collection phase (Section 3.2).
2. **Time and Resources:** The feasibility of conducting fine-tuning experiments within the 6-month thesis timeline and with available computational resources.
3. **Model Accessibility:** Confirmation that the chosen SahabatAI model variant can be practically fine-tuned (e.g., availability of fine-tuning scripts, manageable model size).

If pursued, fine-tuning could involve:

- **Standard fine-tuning:** Further training SahabatAI on a text generation task using consular-specific Q&A pairs, where the input is the question (and potentially context) and the target is the desired answer.
- **Instruction fine-tuning:** If a dataset of instructions and corresponding desired outputs for consular tasks can be created, this paradigm could be used to better align the model with the specific requirements of the chatbot [Ouyang et al., 2022].

The hypothesis is that domain-specific fine-tuning might adapt SahabatAI's language style, improve its understanding of consular terminology, and potentially enhance its ability to synthesize information from retrieved contexts more effectively for this specific domain. However, the RAG architecture itself is designed to provide domain specificity through retrieved documents, so the added benefit of fine-tuning needs to be carefully evaluated against its cost.

3.4.3. RAG Architecture Design and Implementation

The RAG system will comprise two main interconnected components: a retriever and a generator.

- **Retriever Component:**
    - **Choice of Retriever:** The primary candidate is a **dense retriever**. This involves using a sentence transformer model (as described in Section 3.3) to compute embeddings for all knowledge base chunks and for incoming user queries. Similarity search (e.g., cosine similarity or dot product) will then be performed to find the most relevant chunks.
        - A **sparse retriever** like BM25 (Okapi BM25) will be considered as a strong baseline or for potential use in a hybrid retrieval system (combining scores from dense and sparse retrievers) if initial results with dense retrieval

alone show limitations.

  - **Vector Database:** A vector database/library such as FAISS (Facebook AI Similarity Search) [Johnson et al., 2019] will be implemented. FAISS allows for efficient storage of high-dimensional vectors and fast similarity search (e.g., k-Nearest Neighbors - kNN) even with large numbers of vectors. Other options like Milvus or Pinecone could be considered depending on ease of setup and features, but FAISS is often suitable for research prototypes.
  - **Retrieval Mechanism:** When a user submits a query, it will be embedded using the same sentence transformer model. This query embedding will then be used to search the vector database, and the top-k most semantically similar document chunks will be retrieved. The value of 'k' (number of chunks to retrieve) will be a hyperparameter to be tuned.
- **Generator Component:**
  - **SahabatAI** will serve as the generator.
  - **Input to Generator:** The input prompt for SahabatAI will be carefully constructed. It will typically include:
    1. The original user query.
    2. The content of the top-k retrieved document chunks.
    3. Specific instructions to the model (prompt engineering) guiding it to synthesize a comprehensive and accurate answer based *only* on the provided context and the query, to maintain an official tone, and to respond in Bahasa Indonesia. For example: *"Anda adalah asisten konsuler AI. Berdasarkan informasi berikut [retrieved_chunks], jawab pertanyaan pengguna ini: [user_query]. Jawab hanya berdasarkan informasi yang diberikan."*
  - **Prompt Engineering:** Significant effort will be dedicated to designing effective prompts. This is crucial for controlling the LLM's output, ensuring it focuses on the retrieved context, minimizes hallucination, and adheres to the desired response style. Different prompt structures will be experimented with.
- Integration (End-to-End Pipeline):
  A Python-based pipeline will be developed to orchestrate the entire process:
  1. User submits a query in Bahasa Indonesia.
  2. Query is preprocessed and embedded.
  3. Retriever fetches top-k relevant chunks from the vector database.
  4. The query and retrieved chunks are formatted into a prompt for SahabatAI.
  5. SahabatAI generates a response.
  6. The response is presented to the user. Consideration will be given to including references to the source documents from which information was retrieved, if

feasible, to enhance transparency.

3.5. Data Analysis (Evaluation Strategy)

A comprehensive evaluation strategy involving both automated metrics and human assessment is essential to rigorously analyze the performance of the developed RAG-SahabatAI consular chatbot.

- **Performance Metrics:**
    - Retrieval Component Evaluation:
      The effectiveness of the retriever in fetching relevant document chunks is crucial for the overall RAG system. Using the manually curated gold-standard Q&A dataset (from Section 3.2), where each question is mapped to its ground-truth relevant document(s)/chunk(s):
        - **Precision@k:** The proportion of retrieved chunks (among the top-k) that are relevant. P@k=k∣{relevant retrieved chunks}∩{top-k retrieved chunks}∣
        - **Recall@k:** The proportion of all relevant chunks in the knowledge base for a given query that are successfully retrieved in the top-k results. R@k=∣all relevant chunks for query∣∣{relevant retrieved chunks}∩{top-k retrieved chunks}∣
        - **Mean Reciprocal Rank (MRR):** The average of the reciprocal ranks of the first relevant document retrieved for a set of queries. The rank is ranki1 if the first relevant document is at rank i, and 0 if no relevant document is retrieved. MRR=∣Q∣1∑i=1∣Q∣ranki1
        - **F1-score@k:** The harmonic mean of Precision@k and Recall@k, providing a balanced measure. F1@k=2⋅P@k+R@kP@k⋅R@k
    - Generation Component Evaluation (and Overall End-to-End System):
      Evaluating the quality of the generated responses.
        - **Automated Metrics (using the gold-standard Q&A dataset with reference answers):**
            - **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Measures n-gram overlap between the generated response and reference answers. ROUGE-N (unigrams, bigrams), ROUGE-L (longest common subsequence) will be used [Lin, 2004].
            - **BLEU (Bilingual Evaluation Understudy):** Measures n-gram precision between generated and reference answers, commonly used in machine translation but also applied to generation tasks [Papineni et al., 2002].
            - **METEOR (Metric for Evaluation of Translation with Explicit

ORdering): Considers unigram matching based on stemming and synonymy, generally correlates better with human judgment than BLEU. While useful, these metrics have known limitations for evaluating semantic correctness and factual accuracy in generative models.

- Human Evaluation (Crucial for nuanced assessment):
  A sample of user queries (from the test set, or newly crafted representative queries) will be presented to the chatbot, and the generated responses will be evaluated by human assessors (e.g., fellow students, researchers, or ideally, individuals familiar with consular matters, if possible).
  - **Criteria (rated on a Likert scale, e.g., 1-5):**
    1. **Accuracy/Factuality:** How factually correct is the response with respect to official MoFA information? (Evaluators may be provided with the retrieved context or reference documents).
    2. **Relevance:** How relevant is the response to the user's query?
    3. **Coherence & Fluency:** Is the response well-structured, grammatically correct, and easy to understand in Bahasa Indonesia?
    4. **Completeness/Comprehensiveness:** Does the response adequately address all aspects of the user's query?
    5. **Helpfulness:** Overall, how helpful is the response in addressing the user's need?
  - **Method:** Evaluators will be provided with clear guidelines and scoring rubrics. Inter-annotator agreement (e.g., using Fleiss' Kappa or Krippendorff's Alpha) will be calculated if multiple evaluators are used for the same responses.

- Baselines for Comparison:
  To demonstrate the effectiveness of the proposed RAG-SahabatAI system, its performance will be compared against:
  1. **Non-RAG SahabatAI:** The SahabatAI model queried directly with consular questions without any retrieved context. This will highlight the contribution of the RAG architecture in improving factual grounding and relevance.
  2. **Existing SARI Chatbot (if feasible):** If ethical and practical means to interact with and evaluate SARI on a common set of test questions can be established, a qualitative comparison (and quantitative if SARI provides evaluable outputs) will be attempted. This comparison might be limited due to the black-box nature of SARI and potential unavailability of its internal data.
  3. **Keyword-based Search System:** A simple baseline could be a keyword search implemented over the curated knowledge base (similar to a basic

search function on the Peduli WNI website), to show the advantage of semantic understanding provided by the RAG system.

- **Statistical Significance:** Where applicable (e.g., comparing mean scores from human evaluation or automated metrics across different systems), appropriate statistical tests (e.g., t-tests, ANOVA) will be used to determine if observed differences in performance are statistically significant.

3.6. Data Visualisation

To effectively communicate the findings of the evaluation, various data visualization techniques will be employed:

- **Bar charts:** To compare performance metrics (e.g., ROUGE scores, BLEU scores, human evaluation scores for accuracy, relevance, etc.) across the different systems being evaluated (RAG-SahabatAI, Non-RAG SahabatAI, SARI if possible, keyword search).
- **Line graphs:** To show retrieval performance metrics (e.g., Precision@k, Recall@k) as 'k' (number of retrieved documents) varies. This can help in tuning the retriever.
- **Radar charts:** To provide a multi-dimensional comparison of different systems across several human evaluation criteria (accuracy, relevance, fluency, completeness, helpfulness).
- **Tables:** Summarizing detailed quantitative results and statistical significance.
- **Qualitative Examples:** Presenting selected examples of good and bad responses generated by the system, along with the corresponding user queries and (if applicable) retrieved contexts. This will help to illustrate the system's capabilities, common error types, and areas for future improvement.

3.7. Data Validity and Ethical Considerations

Ensuring data validity and adhering to ethical principles are paramount throughout this research.

- **Knowledge Base Validity:** The factual accuracy of the chatbot heavily relies on the validity of its knowledge base.
  - **Source Authentication:** The knowledge base will be constructed exclusively from official MoFA channels (primarily the Peduli WNI portal) and other authenticated public government documents. Information from unverified sources will not be included.
  - **Up-to-dateness:** While the prototype development within the 6-month timeframe may use a static snapshot of the knowledge base, the RAG architecture is designed for easy updates. The thesis will discuss a conceptual

protocol for periodically refreshing the knowledge base to reflect changes in consular policies or information, even if full implementation of an automated update mechanism is beyond the scope.

- **Bias Mitigation:**
    - **LLM Bias:** SahabatAI, like any LLM, may inherit biases from its training data. While using a Bahasa Indonesia-specific model may mitigate some cultural biases present in predominantly English-trained models, biases related to gender, ethnicity, or other attributes could still exist.
    - **Document Bias:** The official consular documents themselves might inadvertently contain biases.
    - **RAG as a Mitigator:** The RAG approach, by forcing the LLM to ground its responses in specific retrieved documents, can help mitigate some types of LLM-generated biases or fabrications, as the response is more constrained by the provided factual context. However, if the source documents themselves are biased, RAG will reflect that. This limitation will be acknowledged.
- **Data Privacy:**
    - **Public Data Focus:** The primary knowledge base will be built from publicly available information.
    - **User Interaction Data:** The prototype will be tested using simulated queries or queries from the developed Q&A test set. If any real user interaction is involved for pilot testing (e.g., with a very small, controlled group of testers), all data will be fully anonymized, and informed consent will be obtained. No personally identifiable information (PII) of actual citizens seeking consular help will be collected or used.
- **Transparency and Explainability:**
    - **Source Citation:** The RAG system offers an inherent advantage in transparency. The research will explore implementing a feature where the chatbot can (optionally or upon request) cite the source document(s) or passages from the knowledge base (e.g., specific URLs from Peduli WNI) that were used to generate its response. This allows users to verify the information if needed.
- **Responsible AI Use:** The development and proposed application will adhere to established principles of responsible AI, including fairness, accountability, and transparency [Floridi et al., 2018; IEEE, 2019]. The limitations of the AI system (e.g., not being a substitute for human consular officers in complex or emergency situations) will be clearly communicated.
- **Feasibility within Thesis Timeline:** The ambitious nature of developing a full RAG system with a national LLM like SahabatAI, potentially including fine-tuning, and conducting robust evaluation within a six-month Master's thesis framework

presents a significant feasibility challenge. This necessitates a very focused scope (as outlined in Section 1.6), pragmatic choices in technology and implementation (e.g., prioritizing well-established off-the-shelf components for the retriever or vector database where appropriate), and a readiness to descope or simplify more complex aspects (such as extensive fine-tuning or highly elaborate evaluation protocols) if time constraints become critical. Meticulous project planning, clear prioritization of objectives, and an agile approach to development will be crucial for successful completion. This pragmatic consideration is vital for the credibility of the proposal, acknowledging that success will depend not only on technical proficiency but also on effective execution strategy and realistic scoping within academic time limits.

Table 3.1: Proposed Data Corpus Specification for RAG System

Data Source	**Data Type**	**Estimated Volume/Number of Documents**	**Format**	**Planned Preprocessing Steps**
Peduli WNI Portal (e.g., https://peduliwni.kemlu.go.id/) - FAQs section	FAQ (Question-Answer pairs)	~50-200 FAQs	HTML	Scraping, HTML cleaning, text extraction, structuring into Q&A format, chunking if answers are long.
Peduli WNI Portal - Articles/Guides (e.g., passport, visa, legal aid)	Informational Articles, Guides	~20-50 articles/guides	HTML	Scraping, HTML cleaning, text extraction, semantic segmentation into coherent chunks (paragraph or section-based), metadata extraction (URL, title).
Peduli WNI Portal - Announcements	News, Advisories	~50-100 relevant items	HTML	Scraping, HTML cleaning, text extraction,

/News		(dynamic)		chunking, metadata extraction (URL, title, date).
Publicly available MoFA Circulars/Regulations (if found online)	Regulation Text, Official Procedures	~5-15 documents	PDF, DOCX	Manual download, text extraction (OCR if needed for PDFs), cleaning, segmentation into logical sections/articles, metadata extraction (document title, number, date).
Relevant Indonesian Laws (e.g., on Citizenship, Immigration)	Legal Text	~2-5 key laws	PDF, HTML	Text extraction, cleaning, segmentation by articles/clauses, metadata extraction.

Table 3.2: Proposed Evaluation Metrics and Benchmarks

Evaluation Aspect	**Metric**	**Definition/Formula**	**Justification for Use**	**Target Benchmark/Baseline for Comparison**
Retrieval Component	Precision@k	Proportion of top-k retrieved docs that are relevant.	Measures accuracy of retrieval.	Compare different retriever settings (e.g., k value, embedding model).
	Recall@k	Proportion of all relevant docs retrieved in	Measures completeness of retrieval.	Compare different retriever

		top-k.		settings.
	MRR (Mean Reciprocal Rank)	Average reciprocal rank of the first relevant doc.	Focuses on getting one good answer quickly.	Compare different retriever settings.
	F1-score@k	Harmonic mean of P@k and R@k.	Balanced measure of precision and recall.	Compare different retriever settings.
Generation Component (Automated)	ROUGE-N, ROUGE-L	N-gram overlap (N=1,2) and Longest Common Subsequence with reference answers.	Standard for summarization and generation tasks; measures lexical overlap.	RAG-SahabatAI vs. Non-RAG SahabatAI.
	BLEU	N-gram precision with brevity penalty against reference answers.	Standard for MT, adapted for generation; measures similarity to human references.	RAG-SahabatAI vs. Non-RAG SahabatAI.
	METEOR	Unigram matching (stemming, synonymy) with alignment.	Correlates better with human judgment than BLEU for some tasks.	RAG-SahabatAI vs. Non-RAG SahabatAI.
Overall System (Human Evaluation)	Accuracy/Factuality (Likert 1-5)	Degree of factual correctness against official sources.	Critical for trustworthy consular info.	RAG-SahabatAI vs. Non-RAG SahabatAI; Qualitative comparison with SARI (if possible).
	Relevance	Degree to which	Essential for	RAG-SahabatAI

	(Likert 1-5)	response addresses the query.	user satisfaction.	vs. Non-RAG SahabatAI; Qualitative comparison with SARI.
	Coherence/Fluency (Likert 1-5)	Grammatical correctness, readability, naturalness of language (Bahasa Indonesia).	Important for user experience.	RAG-SahabatAI vs. Non-RAG SahabatAI.
	Completeness (Likert 1-5)	Degree to which response comprehensively answers the query.	User need fulfillment.	RAG-SahabatAI vs. Non-RAG SahabatAI.
	Helpfulness (Likert 1-5)	Overall perceived utility of the response.	Holistic measure of system quality.	RAG-SahabatAI vs. Non-RAG SahabatAI; Qualitative comparison with SARI.

References (Illustrative - student to replace with actual full citations)

Accenture. (2017). *Chatbots in Government: Exploring the use of Chabot technologies in the Public Sector*. Accenture.

Aji, A. F., et al. (2022). NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022).*

ATO (Australian Taxation Office). (2021). *ATO annual report 2020–21.*

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.*

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*.

Berry, M. J. A., et al. (2022). *Public Sector Use Cases for Artificial Intelligence*. Deloitte Insights.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P.,... & Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Brynjolfsson, E., & McAfee, A. (2017). *The Business of Artificial Intelligence*. Harvard Business Review.

Chatbots Life. (2019). *How Governments Are Using Chatbots To Connect With Citizens*. Retrieved from

Costa-jussà, M. R., et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Cunha, M. A., et al. (2020). Chatbots in the public sector: a systematic literature review. *Government Information Quarterly, 37*(3), 101480.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering, 22*(5), 811-817.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*.

DiploFoundation. (2021). *AI in Diplomacy*. Retrieved from

e-Estonia. (n.d.). *AI in Estonia*. Retrieved from

European Commission High-Level Expert Group on AI. (2019). *Ethics Guidelines for Trustworthy AI*.

Floridi, L., Cowls, J., Beltramini, M., Saunders, D., & Vayena, E. (2018). An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *AI and Society, 33*(4), 689-707.

Gao, T., Yao, X., & Chen, D. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.

IEEE. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y.,... & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys, 55*(12), 1-38.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data, 7*(3), 535-547.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed. draft). Prentice Hall.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S.,... & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lewis, P., Denoyer, L., & Riedel, S. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Naman, G.,... & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*.

Longo, L. (2019). Chatbots in the public sector: A disruptive technology for democratic engagement. *Proceedings of the 20th Annual International Conference on Digital Government Research*.

Marr, B. (2020). *The Amazing Ways Governments Around The World Are Using*

Artificial Intelligence (AI) And Big Data. Forbes.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review, 102*(3), 419.

Ministry of Foreign Affairs (MoFA) Consular Guidelines. (2020). *Panduan Layanan Konsuler* (Illustrative title, student to find actual document).

Misuraca, G., et al. (2020). *Artificial intelligence in public services: A_I Watch. European Landscape Report*. Publications Office of the European Union.

OECD. (2019). *Hello, World! Artificial Intelligence and its Use in the Public Sector*. OECD Publishing.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P.,... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of...*source *2002)*.

Patterson, D., et al. (2021). Carbon Emissions and Large Neural Network Training. *arXiv preprint arXiv:2104.10350*.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2020). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (Note: This Petroni et al. is about embeddings, the RAG context update advantage is more from Lewis et al., 2020 and general RAG literature).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M.,... & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Journal of Machine Learning Research, 21(140), 1-67.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP 2019)*.

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval, 3*(4), 333-389.

SahabatAI Website. https://sahabat-ai.com/
SahabatAI Hugging Face Collection.
https://huggingface.co/collections/GoToCompany/sahabat-ai-672af7b248f5fdfd39ae2403
Peduli WNI Portal. https://peduliwni.kemlu.go.id/beranda.html
Antaranews Article on SARI Chatbot.
https://kl.antaranews.com/berita/31537/aplikasi-chatbot-sari-melengkapi-celah-informasi-bagi-pekerja-migran-di-luar-negeri
Sai, A. B., et al. (2022). A Survey of Evaluation Metrics for Dialogue System. *arXiv preprint arXiv:2201.06985*.

Sheng, E., et al. (2021). Societal Biases in Language Generation: Progress and Challenges. *AI and Ethics, 1*(4), 437-450.

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Singapore Government. (n.d.). *AskJamie*. Retrieved from

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*.

Thorne, J., et al. (2021). The FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*.

UK Parliament. (2023). *AI in the FCDO*. House of Commons Foreign Affairs Committee Report.

UNDP. (2021). *Artificial Intelligence for Development: A Primer for Public Sector Leaders*. United Nations Development Programme.

USCIS. (n.d.). *EMMA - Your Virtual Assistant*. Retrieved from

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,... & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Weidinger, L., et al. (2021). Ethical and social risks of harm from Language Models. *arXiv preprint arXiv:2112.04359*.

Wilie, B., et al. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2020)*.

World Bank Group. (2020). *Artificial Intelligence in the Public Sector: A Maturity Model*.

Zhang, T., et al. (2023). A Survey on Instruction Tuning for Large Language Models. *arXiv preprint arXiv:2308.10792.*