A Retrieval-Augmented Generation Approach with Fine-Tuned SahabatAI for Indonesian Consular Question Answering

Fathur Rohman

President University

Informatics Study Program

Master of Science in Information Technology

Thesis Proposal

2025

Thesis Supervisor:

Dr. Fulan, S.T., M.T.

Thesis Co-Supervisor:

Dr. Fulana, S.T., M.T.

Abstract

Providing good consular services to citizen abroad is crucial for the Indonesian government. Ministry of Foreign Affairs (Kemenlu) has launched digital platforms like ”Peduli WNI”, ”Safe Travel”, and the ”SARI” chatbot. while SARI is focused on migrant workers, it may not cover all consular services queries. This research aims to develop a Question Answering (QA) system using a Retrieval-Augmented Generation (RAG) approach with a fine-tuned SahabatAI model(base on Gemma2 and llama3) to answer questions related to Indonesian consular services. The method focuses on applying Parameter-Efficient Fine-Tuning (PEFT) to adapt SahabatAI for consular topics, implementing a Retrieval-Augmented Generation (RAG) system with SahabatAI as the response generator, and building a specialized knowledge base from official MoFA sources. The evaluation will focus on how well the system retrieves information, the quality of generated answers (accuracy and relevance), and overall performance. The expected outcome is a robust QA system that can provide accurate and timely information to Indonesian citizens seeking consular services.

Keywords: Consular Services, SahabatAI Fine-Tuning, Retrieval-Augmented Generation, Fine-Tuning (RAG), Question Answering (QA) System, Public Service Automation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Consular services represent a cornerstone of a nation's support for its citizens abroad. These services, which include the issuance of passports and visas, provision of emergency assistance, facilitation of self-reporting for citizens residing overseas, and management of case reports, are vital for ensuring the safety, well-being, and legal standing of individuals in foreign territories [1], [2]. The Indonesian Ministry of Foreign Affairs (MoFA) is tasked with serving a substantial global diaspora and a large number of citizens traveling internationally [3], which underscores the necessity for highly efficient, accessible, and responsive support mechanisms.

In response to these demands, the Indonesian MoFA has proactively embraced digital transformation to enhance its consular service delivery [4]. This commitment is evident in its existing digital ecosystem:

- **Peduli WNI:** This web-based platform serves as a central hub for Indonesian citizens abroad, offering critical features such as Lapor Diri (Self-Reporting), Pelayanan Kekonsuleran (Consular Services), and Pengaduan Kasus (Case Reporting). The portal has significantly streamlined processes that previously necessitated physical visits to Indonesian embassies or consulates, allowing services to be accessed online with internet connectivity [5].

Table 1.1: Overview of Consular Service Platforms

Platform Name	**Type**	**Key Features**	**Target Users**	**Reported AI/Technology Used**
Portal Peduli WNI	Web-portal	Lapor Diri (Self-Reporting), Pelayanan Kekonsuleran (Consular Services), and Pengaduan Kasus (Case Reporting)	Indonesian citizens abroad	Web-based platform
Safe Travel	Mobile App	Trip registration, country-specific information, notifications, emergency assistance (location sharing, video recording)	Indonesian citizens traveling abroad for short trips	Mobile application
SARI (Sahabat Artifisial Migran Indonesia)	Chatbot	designed capacity for empathetic responses	Indonesian female migrant workers abroad	AI-powered chatbot, NLP, integrated with Safe Travel

- **Safe Travel:** A mobile application designed for Indonesian citizens undertaking short trips abroad, although it can also be utilized by expatriates. It provides practical country-specific information (e.g., time differences, security conditions, local laws and customs, immigration requirements, health services at Indonesian missions), travel registration, notifications (appeals, advice, warnings), and crucial emergency assistance features. In critical situations, users can send their location, record video, and contact the nearest Indonesian mission.

- **SARI Chatbot:** This AI-powered chatbot represents a significant step towards leveraging advanced technology for citizen protection. Developed in collaboration with UN Women, SARI is specifically tailored to assist and protect Indonesian female migrant workers from potential violence and exploitation. Integrated within the Safe Travel application, SARI aims to deliver accessible, unbiased, and non-discriminatory information. A key feature is its designed capacity for empathetic responses. The launch of SARI underscores MoFA's commitment to "digital empathy" and delivering excellent service and protection, particularly for vulnerable groups [6].

The global landscape of public servce delivery is increasingly being reshaped

by andvancements in artificial intelligence (AI), particularly Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) techniques. These technologies offer transformative potential including 24/7 citizen assistance, the capacity to handle complex and nuanced queries, multilingual support, and the ability to personalize interactions, thereby enhancing the efficiency and effectiveness of government services.

A recent development in Indonesia's AI journey is SahabatAI, a large language model (LLM) fine-tuned for Indonesian language tasks [7]. SahabatAI is collaborative initiative by Indosat Ooredo Hutchison and GoTo Group based on the Gemma2 [7] and llama3 architecture [8], which has been trained on a diverse range of Indonesian text data comprising over 640,000 instruction-completion pairs, covering Bahasa Indonesia, Javanese, and Sundanese, with plans to include other regional languages like Batak and Balinese. The core objectives of SahabatAI include promoting linguistic diversity, fostering AI sovereignty for Indonesia, and enabling seamless business-to-government (B2G) and business-to-business (B2B) interactions to significantly enhance the quality and accessibility of government services. Potential use cases for SahabatAI in the public sector include simplifying applications for the national identity card (KTP), demystifying taxation processes, and streamlining procuderes for official document changes related to live events such as marriage ore relocation.

The newly existance of SahabatAI as open-source LLM, extensively trained on Indonesian text data, presents relevant technological foundation for this thesis. Developing an LLM from scratch is a complex and resource-intensive endeavor, requiring substantial computational power and extensive datasets, far exceeding the scope of a Master's thesis. By leveraging SahabatAI, this research can focus on fine-tuning the model for specific tasks, such as consular services, while also exploring the integration of RAG techniques to enhance the system's performance and responsiveness.

1.2 State of the Art

Large Language Models (LLMs), primarily based on the Transformer architecture, have revolutionized natural language processing (NLP) and artificial intelligence (AI) [9]. it has demonstrated capabilities in various tasks, including text generation, translation, summarization, and question answering. This makes them suitable for complex Question Answering (QA) tasks, where they can interpret user queries and generate relevant, coherent responses based on their vast pre-trained knowledge or context provided at inference time [10].

Retrieval-Augmented Generation (RAG) is a techniques developed to enhance the capabilities of LLMs, particularly in knowledge-intensive tasks [11]. RAG architectures connect LLMs with external, often dynamic, knowledge sources, allowing them to retrieve relevant information from a database or document corpus before generating a response. This approach addresses the limitations of LLMs, such as their fixed knowledge base and potential inaccuracies in generated content [12]. By integrating retrieval mechanisms, RAG systems can provide more accurate and contextually relevant answers, especially in domains where up-to-date information is crucial. By retrieving relevant information from these sources and providing it as context to the LLM during answer generation, RAG mitigates common LLM limitations, such as hallucinations and outdated knowledge. This is particularly important in dynamic fields like consular services, where information can change over time and needs to adhere to current laws and regulations.

The application of AI in consular services is growing trend globally [13]. Governments are increasingly exploring AI-powered solutions, including chatbots for handlung frequently asked questions, assisting with visa applications, and providing real-time information to citizens. For instance, Singapore's GovTech Agency has deployed AI chatbots accross various government departments, leading significant reduce in call center workloads and faster response times for citizen inquiries [14]. Similarly, the U.S. Department of State has outlined plans to use AI for various consular function, including passport photo quality assessment, analysis of customer feedback, AI-driven translation

services, and enhanced search and chatbot systems for its Travel.State.Gov website [15]. These examples highlight a global shift towards leveraging AI to make consular services more efficient, accessible, and responsive to citizen needs. Other relevant works include general AI Principles [16], and AI applications in diplomacy [17].

1.3 Gap Analysis

Despite the Indonesian MoFA's commendable efforts in digital transformation like the development of the "Peduli WNI" platform, "Safe Travel" application, and SARI chatbot, several gaps and opportunities remain for enhancing consular services delivery through AI.

While the existing platforms provide valuable services, they may not comprehensively address all consular queries, particularly those related to specific legal or procedural matters. The SARI chatbot, though an innovative AI application, is specifically designed to support Indonesian female migrant workers, focusing on protection against violence and exploitation. This leaves a gap in addressing queries related to other consular services, such as passport renewal procedures, visa regulations for various countries, assistance for lost documents, and general emergency guidance without a dedicated, advanced AI-powered conversational interface. These general queries can be complex, nuanced, and often require synthesizing information from multiple official sources. Current systems may not be equipped to handle such multi-turn conversational interactions or provide comprehensive answers that require understanding implicit user needs. "Peduli WNI" while providing valuable information and transactional services, may not be fully optimized for complex queries. Futhermore, user feedback for existing applications like the "Safe Travel" app has indicated occasional technical issues, such as server connectivity issues and application crashes, suggesting room for improvements in reliability and user experience. Additionally, the current systems may not fully leverage the potential of advanced AI technologies, such as Retrieval-Augmented Generation (RAG) and fine-tuning techniques, to enhance their capabilities.

The advent of SahabatAI, an open-source LLM fine-tuned for Indonesian language tasks, with a strong foundation in Bahasa Indonesian and local dialects like Javanese and Sundanese, presents a unique opportunity to address these gaps. Appliying state-of-the-art RAG techniques in conjunction with SahabatAI model can potentially deliver more accurate, contextually relevant, and up-to-date responses to consular queries than could be achieved with generic LLMs or simple rule-based chatbot technology. The spesific combination of localized LLM like SahabatAI with advanced RAG tailored for Indonesian consular domain remains an underexplored area of research and application.

Moreover, general-purpose LLMs or even generic RAG systems often necessitate significant adaptation to perform optimally in specialized domain such as consular services, This domain is characterized by its unique terminologies, intricate regulations, evolving policies, and diverse user needs context [18]. Therefore, fine-tuning SahabatAI to better understand and generate text specific to consular affairs, coupled with the meticulous curation of a dedicated comprehensive consular knowledge base, is crucial for developing an effective and reliable AI assistant.

The existing MoFA digital tools—Peduli WNI (web-based), Safe Travel (mobile app), and SARI (chatbot integrated within Safe Travel)—while individually valuable, operate with some degree of separation in terms of user interface and scope. A sophisticated RAG-based chatbot, as proposed in this research, could serve as a more unified and intelligent front-end. Such a system could potentially integrate information from, or direct users to, these existing platforms, thereby providing a more seamless and comprehensive user experience for a wider range of consular questions. The knowledge base for the RAG system would ideally be constructed by consolidating information from these diverse official MoFA sources, creating a centralized and reliable information backbone for the AI. This approach could improve the discoverability and accessibility of consular information that might currently be distributed across different platforms or document formats.

1.4 Research Questions

The research questions guiding this study are as follows:

1. **RQ1:** How can a RAG system using SahabatAI be designed to accurately and reliably handle various Indonesian consular queries, such as passport renewal, lost documents, and visa regulations?

2. **RQ2:** Which PEFT strategies (e.g., QLoRA, instruction tuning) are most effective for adapting SahabatAI to the specific language and information requirements of Indonesian consular queries in a RAG framework?

3. **RQ3:** What are the essential components and best practices for developing a high-quality, domain-specific knowledge base and QA dataset for Indonesian consular services to effectively train and evaluate a RAG system?

4. **RQ4:** How does the fine-tuned SahabatAI-RAG system perform compared to baseline models (e.g., zero-shot SahabatAI, naive RAG) and existing MoFA chatbot solutions in terms of accuracy, relevance, coherence, and user perceived helpfulness?

5. **RQ5:** What practical challenges and ethical considerations arise in deploying an AI-powered consular service assistant, particularly regarding data privacy, bias mitigation, and equitable information access, and how can these be addressed within a six-month Master's thesis?

1.5 Research Objectives

To answer the research questions, this study aims to achieve the following objectives:

1. **RO1:** To design and implement a RAG-based chatbot using SahabatAI to provide accurate and reliable responses to Indonesian consular queries, based on a curated knowledge base of official consular information.

2. **RO2:** To implement and evaluate PEFT techniques (e.g., QLoRA) for fine-tuning SahabatAI using a custom Indonesian consular QA dataset, with the goal of improving its domain-specific accuracy and response quality.

3. **RO3:** To curate a comprehensive knowledge base from official MoFA resources and develop a representative QA dataset that covers common Indonesian consular service inquiries for training and evaluating the RAG system..

4. **RO4:** To assess the performance of the fine-tuned SahabatAI-RAG system against baseline models and MoFA chatbot solutions using automated metrics (e.g., BLEU, ROUGE) and human evaluation to measure accuracy, relevance, coherence, and user perceived helpfulness.

5. **RO5:** To identify key ethical challenges, such as data privacy and bias mitigation, in deploying the proposed AI-powered consular assistant and to provide actionable recommendations for addressing these challenges within the scope of a six-month Master's thesis.

1.6 Limitations

This research, while ambitious, will be conducted within the constraints of a six-month Master's thesis timeline. This necessitates a focused scope, particularly in the following areas:

- **Scope of Consular Services:** The primary focus is on Indonesian consular services, including passport renewal, lost documents, and visa regulations.

- **knowledge Base:** The knowledge base will be constructed from publicly available official MoFA resources, with a focus on accuracy and relevance.

- **Language Focus:** While SahabatAI has demonstrated some training in Javanese and Sundanese, the primary focus will be on Bahasa Indonesia.

- **Fine-Tuning Techniques:** Given the typical constraints of computational resources and time, the research will primarily explore QLoRA and instruction tuning as PEFT strategies for SahabatAI [19]. Full model retraining or extensive hyperparameter tuning may not be feasible within the timeframe. Recent studies suggest that even datasets around 1000 samples can be effective for PEFT [20].

- **Dataset Scale:** The QA dataset will be curated to cover a representative range of consular queries, but it may not cover every possible question or scenario, aiming for several hundred to a thousand samples, depending on the complexity of the questions and the available resources.

- **Evaluation Metrics:** The evaluation will focus on automated metrics (e.g., BLEU, ROUGE) and human evaluation for accuracy, relevance, coherence, and user perceived helpfulness. However, human evaluation may be limited to a smaller sample size.

- **Deployment and User Testing:** While the research will include a discussion of deployment considerations, actual deployment and extensive user testing may not be feasible within the six-month timeframe. The focus will be on developing a prototype that can be tested in a controlled environment. it will not be a production-ready system.

- **SahabatAI Model Version:** The research will utilize the version of SahabatAI(e.g., Gemma2 9B CPT SahabatAI v1 Instruct) available at the time of the study.

- **Safety and Ethical Considerations:** While the research will address ethical considerations, the implementation of safety measures and bias mitigation strategies may be limited to theoretical discussions and initial implementations, rather than comprehensive solutions.

1.7 Hypothesis

The following hypotheses will guide the empirical investigation:

1. **H1:** A Retrieval-Augmented Generation (RAG) system incorporating the SahabatAI LLM, fine-tuned with a domain-specific Indonesian consular QA dataset using Parameter-Efficient Fine-Tuning (PEFT), will demonstrate significantly higher factual accuracy and contextual relevance in answering consular queries compared to the baseline SahabatAI model without RAG or fine-tuning.

2. **H2:** The fine-tuned SahabatAI model will demonstrate improved accuracy and relevance in answering Indonesian consular queries compared to the original SahabatAI model.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the theoretical foundations and existing work relevant to the proposed research. It delves into Large Language Models (LLMs) for Question Answering (QA), Retrieval-Augmented Generation (RAG) techniques, the specifics of SahabatAI and other LLMs, the application of AI in governmental and consular services, and methodologies for evaluating such systems.

2.1 Large Language Models (LLMs) for Question Answering (QA)

LLMs have revolutionized the field of Natural Language Processing (NLP), demonstrating unparalleled capabilities in understanding and generating human-like text. Their success is largely attributable to the **Transformer** architecture, first introduced by Vaswani et al. (2017) [9]. This architecture's core innovation, the **attention mechanism** (specifically self-attention and multi-head attention), allows models to weigh the importance of different words in an input sequence and capture long-range dependencies and complex contextual relationships, which are crucial for nuanced question answering.

The development of LLMs typically follows a two-stage paradigm:

1. **Pre-training:** In this phase, LLMs are trained on vast and diverse text corpora,

often sourced from the internet (e.g., Common Crawl, Wikipedia) and large book collections. The training objectives vary depending on the model architecture. Encoder-based models like BERT often use Masked Language Modeling (MLM), where the model predicts masked (hidden) words in a sentence. Decoder-based models like GPT family [21], [22], [23] and Gemma [24] utilize Causal Language Modeling (CLM), where the model predicts the next word in a sequence. This extensive pre-training phase endows LLMs with broad linguistic understanding, grammatical proficiency, and a significant amount of factual knowledge embedded within their parameters.

2. **Fine-tuning:** After pre-training, LLMs are adapted to specific downstream tasks (e.g., question answering, summarization, translation) or specialized domains using smaller, curated datasets [25].

Several state-of-the-art LLMs have emerged from leading research institutions and companies, each with unique architectures and training methodologies. Notable examples include OpenAI's GPT series [26], Metas's LLaMA series [27], Google's Gemini [28] and PaLm families [29], and Anthropic's Claude models [30]. These models exhibit remarkable performance across a wide range of NLP tasks, including question answering, text generation, and summarization.

Within the Indonesian context, the SahabatAI model, developed by Indosat Ooredo Hutchison and GoTo Group, represents a significant advancement in LLM technology.

- **Architecture:** SahabatAI is based on Google's Gemma2 architecture 15, with the specific publicly available instruct-tuned model being gemma2-9b-cpt-sahabatai-v1-instruct [7]. Gemma models are decoder-only transformers. SahabatAI has a context length of 8192 tokens, although some evaluations have used a capped context of 4096 tokens due to inference platform limitations.

- **Development and Collaboration:** This LLM is collaborative effort, co-initiated by Indosat Ooredo Hutchison and GoTo Group, and developed in partnership

with AI Singapore. The development leveraged NVIDIA's NeMo Framework and NIM microservices for model training and deployment [31].

- **Training Data:** The model has been trained on a diverse range of Indonesian text data, including over 640,000 instruction-completion pairs covering Bahasa Indonesia, Javanese, and Sundanese. There are plans to include other regional languages like Batak and Balinese in future iterations.

- **Training Data & Language Capabilities:** SahabatAI was trained with a strong emphasis on Bahasa Indonesia, using 448,000 instruction-completion pairs. The dataset also includes 96,000 Javanese and 98,000 Sundanese pairs to cover regional dialects, along with 129,000 English pairs for multilingual support. The training data combined synthetic instructions and curated public data reviewed by native speakers to maintain quality and cultural relevance. The goal is to develop models that effectively grasp local contexts and cultural nuances.

- **Performance & Benchmark:** SahabatAI has demonstrated strong performance, reportedly outperforming model like Lllab-3.1-8B and sea-lionv3-9B on the SEA HELM(BHASA) evaluatioan benchmark, it has been evaluated on variety of tasks within SEA HELM(including QA, Sentiment Analysis, Toxicity Detection, Translation, Summarization, Causal Reasoning, and Natural Language Inference) and also on the IndoMMLU benchmark, which covers examination questions across various subjects and educational levels in Indonesia.

- **Availability:** SahabatAI is an open-source model, with models accessible through Hugging Face [32]. This open-source approach encourages community collaboration and further research in the field of Indonesian NLP. it relased under the Gemma Community License.

- **Limitations:** Like many LLMs, SahabatAI is susceptible to common issues such as "hallucinations" (generating incorrect or nonsensical information). A critical point for this research is that the publicly released SahabatAI model have not

undegone spesific safety alignment. Developer and users are explicitly advised to conduct therir own safety fine-tuning and implement appropriate safety measuress. This is particularly important for applications in sensitive domains like consular services, where accuracy and reliability are paramount.

To provide a broader regional context, several other Southeast Asian LLMs have been deployed, reflecting the increasing focus on local language. These include SEA-LION(AI Singapore) [33], SeaLLM(Alibaba) [34], and Sailor(SEA AI Lab & Singapore University of Technology and Design) [35]. SEA-LION, for instance, was trained on 11 regional languages, including Indonesian and Javanese, while Sailor supports Indonesian among others, These initiatives underscore the importancen of linguistic diversity and local adaptation in LLM landscapes.

The characteristic of SahabatAI——its open-source nature, focus on Indonesian language and dialects, and its potential for fine-tuning—make it a suitable candidate for this research. However, its documented limitations, particularly the potential for hallucinations and the lack pre-existing safety alignment, are critical factors that must be proactively addressed within the proposed research methodology. This will involve leveraging RAG ground responses in factual consular data and incorporating safety guardrails in the system design.

2.2 Retrieval-Augmented Generation (RAG)

RetrievalAugmented Generation (TAG) has emerged as powerful paradigm for enhancing the capabilities of LLMs, particularly in knowledge-intensive and fact-sensitive applications. The core concept pf RAG, as introduced by Lewis et al. (2020) [11], is to combine the strengths of retrieval-based and generation-based approaches to improve the accuracy and relevance of generated responses. In a typical RAG architecture, a retriever component is used to identify and retrieve relevant documents or passages from an external knowledge base or corpus, which are then provided as context to a generator model (often an LLM) during the response generation process.

The typical RAG pipeline operates as follows [36], [37]:

1. **Query Encoding:** The user's input querys is transformed into dense ventor representaion (embedding) using text embedding model.

2. **Document Retrieval:** This query embedding is used to search a pre-indexed collection of documents (the knowledge base, often stored in a vector database). The retriever identifies and fetches the most relevant document chunks based on semantic similarity (e.g, cosine similarity between query and document embeddings).

3. **Context Augmentation:** The retrieved document chunks are then concatenated with the original user query to form augmented prompt.

4. **Answer Generation:** This augmented prompt is then fed to the LLM, which generates a response grounded in both the user's query and the provided contextual information.

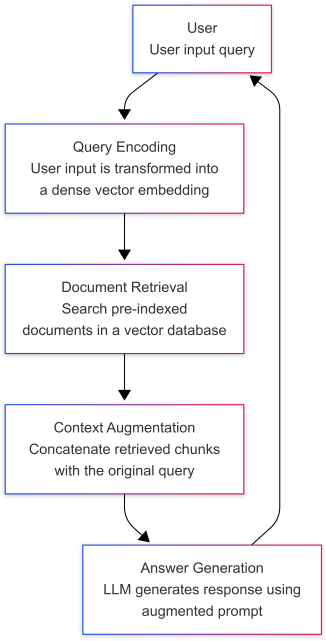

Figure 2.1: The typical RAG pipeline.

The **benefits of RAG** anre manifold. it significantly reduces the likelihood of "hallucinations" by anchoring responses to factual data retrieved from the external knowledge base [38]. RAG systems can access up-to-date information without the need for

frequent and costly retraining of the entire LLM; the knowledgebase can be updated independently. This also enables domain specialization by simply providing a domain-spesific knowledge base. Furthermore, RAG can facilitate source attribution, allowing users to verify the information presented in the LLM's response by referring to the source documents.

RAG implementation can range from simple (Naive RAG) to more complex (Advanced RAG) techniques [39]:

- **Naive RAG:** The most basic RAG approach, which simply retrieves relevant documents based on query similarity from the knowledge base and uses them as context for the LLM's response generation.
- **Advanced RAG:** This approach incorporates additional techniques at various stages of the RAG pipeline to improve performance:
  - **Pre-Retrieval Strategies:** Techniques like query expansion(e.g., adding synonyms or related terms), query transformation(e.g., rephrasing the query for clarity), or generating multiple sub-queries from a complex query to retrieve a richer set of documents.
  - **Retrieval Strategies:** Moving beyond simple dense vector retrieval to include hybrid search (combiningn keyword-based search like BM25 with semantic search), optimizing the choise and fine-tuning of embedding models, or employing graph-based retrieval mechanisms (e.g., Graph RAG, KG-RAG, where knowledge graphs are used to enhance retrieval).
  - **Post-Retrieval/Re-rangking:** After an initial set of documents is retrieved (e.g., top-N candidates), a re-rangking model is used to to re-order these documents based on a more fine-grained assessment of their relevance to the query [40]. Cross-encoder models, which joinly process the query and each candidate document, are often effective for this but are computationally expensive than bi-encoder retrievers.

- **Iterative/Multi-hop Retrieval:** For complex questions that require synthesizing information from multiple sources or performing multi-step reasoning, iterative retrieval techniques can be employed. This might involve decomposing the main question into sub-questions, retrieving evidence for each, and then synthesizing an answer [41]. The Collab-RAG framework, for example, proposes using smaller language model (SLM) to decompose complex queries, with a larger LLM acting as the reader/synthesizer.
- **Fine-tuning RAG Components:** This involves training the retriever (embedding model) and/or the generator LLM spesifically for the RAG task and target domain. This can improve the alignment between the retriever and generator and enhance the generator's ability to utilize retrieved context effectively [38].

Embedding models play a crucial role in the RAG pipeline, as they are responsible for converting text (queries and documents) into dense vector representations. State-of-the-art embedding models include Google's Gemini Embedding (models like text-embedding-004 and the experimental gemini-embedding-exp-03-07), which has shown top performance on the MTEB Multilingual benchmark [42].

Other strong open-source multilingual models like multilingual-e5-large-instruct are also widely used. The performance of these models is often evaluated on benchmarks like MTEB (Massive Text Embedding Benchmark) [43] and its multilingual extension, MMTEB [44], which cover various tasks and languages. For Indonesian, specific resources like the Indonesian Sentence Embeddings project and its associated benchmarks (e.g., SemRel2024, Indonesian subsets of MIRACL and TyDiQA) provide valuable evaluation points [45]

The generated embeddings are typically stored and queried using **vector databases**. Theses databases are optimized for efficient similarity search in high-dimensional spaces, employing Approximate Nearest Neighbor (ANN) algorithms like HNSW (Hierarchical Navigable Small World) [46] or IVF (Inverted File Index) [47]. Key features to consider when choosing a vector database include scalability, query latency, support

for metadata filtering (allowing hybrid search), and ease of integration. Popular open-source options suitable for academic research include FAISS, Milvus, and Qdrant [48].

Despite its advantages, RAG systems face several challenges:

- **Retrieval Quality:** The ”garbage in, garbage out” principle applies; if the retriever fails to fetch relevant documents or retrieves low-quality information, the generator’s output will likely be inaccurate or irrelevant. A common issue is the ”lost in the middle” problem, where LLMs tend to ignore relevant information if it’s buried within a long contextn of retrieved documents [49].

- **Generation Quality:** Even with relevant retrieved documents, the LLM might still hallucinate, fail to synthesize information coherently from multiple documents, or produce responses that are not faithful to the provided sources [50], [51].

- **Context Window Limitations:** LLMs have a finite input context windows. Effectively summarizing and presenting a large amount of retrieved information to LLM without exceeding this limit or losign crucial details is a challenge.

- **Evaluation Complexity:** Evaluating a multi-stage RAG pipeline ins inherently complex, as it requires assessing the the performance of both the retrieval abd generation components, as well as their interaction [52].

- **Safety and Bias:** RAG systems can inadvertently propagate biases present in the retrieved documents. There’s also a risk that even if the retrieved documents are safe and factual, a non-safety-aligned LLM might still misinterprete or ”twist” this information to generate unsafe or misleading outputs [53].

Given the nature of Indonesian consular information—which often involves legal nuances, specific procedural details, and varying citizen situations—a naive RAG approach may prove insufficient. The complexity and criticality of providing accurate consular advice necessitate the exploration and implementation of advanced RAG techniques. Specifically, effective re-ranking of retrieved documents to ensure high rele-

vance, and potentially iterative retrieval strategies for handling complex, multi-faceted queries, will be crucial for building a robust and reliable system. Furthermore, the inherent limitations of the SahabatAI model, such as its potential for hallucination, can be better mitigated by providing it with higher quality, more precisely retrieved context that advanced RAG components can offer.

2.3 Fine-tuning LLMs for Domain-Specific QA and RAG

Fine-tuning pre-trained LLMs is a critical step in adapting then to spesific domains or task, such as the nuanced requirements of consular questions answering. This process adjusts the model's paramaters using a smaller, domain-specific dataset, enabling it to learn the particular vocabulary, style, and knowledge patterns relevant to the target application.

A distincion is made between **Full Fine-Tuning (FFT)** and **Parameter-Efficient Fine-Tuning (PEFT)**:

- **Full Fine-Tuning (FFT):** This approach involves updating all parameters of the pre-trained model during the fine-tuning process. While effective, it is computationally expensive, requires significant memory resources, making it less practical for large models or limited hardware environments. it also requires large domain-specific datasets to avoid overfitting. A notable drawback is the risk of ”catastrophic forgetting,” where the model loses some of its general capabilities learned during pre-training.

- **Parameter-Efficient Fine-Tuning (PEFT):** PEFT methods aim to overcome the limitations of FFT by only updating a small subset of parameters or introducing additional lightweight modules while keeping the majority of the model's parameters frozen. This approach significantly reduces the computational burden and memory requirements, making it feasible to fine-tune large models on smaller

datasets. Popular PEFT techniques include:

- **LoRA (Low-Rank Adaptation) and QLoRA (Quantized LoRA):** LoRA introduces small, trainable low-rank matrices into the layers of the transformer model, effectively learning task-specific adaptations altering the original weights [54]. QLoRA extends this by quantizing the model weights (e.g., to 4-bit precision) to further reduce memory usage and computational cost [55]. QLoRA has proven particularly effective for fine-tuning Gemma-based models [19]
- **Adapter Layers:** These involve inserting small, trainable neural network modules (adapters) between the existing layers of the pre-trained model. During fine-tuning, only the adapter parameters are updated, while the original model parameters remain frozen [56].
- **Prefix Tuning:** This method adds a small set of trainable prefix tokens to the input sequence. The model learns to modulate its behavior based on these learned prefixes, without changing the core LLM parameters [57].

Instruction Fine-Tuning (IFT) is a specific type of fine-tuning that trains LLMs on datasets composed of (instruction, input, output) triplets. This helps the model learn to follow instructions better and perform specific tasks as directed. for QA tasks, this typically involves fine-tuning on (questions, context, answer) examples. In the context of RAG, IFT can be applied to:

- Fine-tune the generator LLM to improve its ability to generate accurate and relevant answers based on the retrieved context or adhere to specific formatting requirements.
- Jointly train the retriever and generator components to improve their alignment and overall RAG performance.
- **RuleRAG:** is an example where symbolic rules are introduced as demonstration for in-context learning or as part of supervised fine-tuning data to explicitly guide

both the retriever (to fetch logically related documents) and the generator (to produce answers that follow the rules) [58].

- **Join QA and Question Generation(QG):** Some research explores fine-tuning an LLM to perform both QA and QG tasks, allowing the model to generate questions based on the retrieved context and then answer them. This can enhance the model's understanding of the context and improve its ability to generate relevant answers [38].

A crucial consideration in fine-tuning is the minimum dataset size required to achieve effective results. Recent studies, such as the LIMA paper [59], as cited in [20], suggest that even small but high-quality datasets (around 1000 samples) can yield significant improvements in performance when using PEFT techniques. This is because most of the foundational knowledge is already acquired duirng the extensive pre-training phase. This finding makes the creation of suitable fine-tuning dataset for Indonesian consular QA an achieveable goal within the scope of this research.

Compared Studies often explore the trade-offs between fine-tuning alone, RAG alone, or a combination, or a combination of both (FT + RAG) [60], [61], [62]:

- **RAG alone** excels in tasks requiring access to external, up-to-date, or proprietary knowledge. It is generally less prone to hallucination if the retriever is accurate and allows for easier knowledge updates by modifying the vector database rather retraining the LLM.

- **Fine-tuning alone** is effective for teaching the LLM new skils, adapting its style or tone, or instilling deep understanding of specific domain language and nuances. It can lead to higher accuracy on specialized task where the requirend knowledge can be embedded into the model's parameters. However, it risks catastrophic forgetting and the model's knowledge remains static based on its training data cu-off.

- **FT + RAG** is often considered the most effective approach, combining the strengths of boths. Fine-tuning can make the LLM a better reasoner or synthesizer

over the contextual information provided by the RAG system. For example, the generator LLM with RAG pipeline can be fine-tuned to better handle the structure of retrieved documents, to follow specific instructions fir answer generation based on the domain's requirements, or to improve its faithfulness to the provided sources.

For fine-tuning Gemma-based models like SahabatAI, the QLoRA method is particularly promising. Libraries like Hugging Face Transformers, along with the PEFT and TRL (Transformer Reinforcement Learning) libraries (specifically the SFTTrainer), provide robust tools for implementing QLoRA fine-tuning for Gemma models [19]. Keras, with backends like JAX, TensorFlow, or PyTorch, also offers support for LoRA tuning of Gemma models [63].

The selection of QLoRA as the PEFT strategy for SahabatAI in this thesis is well supported by its efficiency with Gemma-based models. Its ability to handle the model's large parameter size, and its effectiveness in improving performance on domain-specific tasks with smaller datasets (around 1000 samples) further strengthens its suitability for this research.

2.4 AI in Government and Consular Services

The adoption of AI in government and public services has gained momentum in recent years, with numerous countries implementing AI solutions to enhance efficiency, accessibility, and citizen engagement. Global Case Studies provide valuable insights into the potential and challenges of AI:

- **Singapore:** Since the launch of AI-powered chatbots like Ask Jamie, HealthBuddy, and the CPF Chatbot in the mid-2010s, Singapore has progressively enhanced its use of AI in public services. These early systems utilized Natural Language Processing (NLP) to handle citizen inquiries across multiple languages, significantly reducing call center workloads and improving response efficiency [64]. Building on this foundation, Singapore has developed more advanced AI tools

such as Pair, an AI assistant designed to boost productivity among public officers, and VICA, a conversational AI platform serving over 60 government agencies and handling hundreds of thousands of queries monthly [65]. Additionally, specialized AI applications like the OneService Chatbot, which routes municipal feedback with about 80% accuracy, and PEACH, an AI chatbot aiding perioperative clinical decisions with over 96% accuracy, demonstrate Singapore's commitment to integrating AI for both operational efficiency and domain-specific expertise [66]. Together, these developments illustrate Singapore's trajectory from basic NLP chatbots to sophisticated, large-scale AI deployments driving a smarter, more responsive public sector. [14].

- **Japan:** Since 2020, Japan has made significant strides in integrating large language models (LLMs) into its healthcare sector. In April 2025, researchers at the National Institute of Informatics developed a generative AI system that successfully passed the national medical licensing examination. Trained on over 700,000 licensed medical papers, 16 million documents from reputable medical websites, and medical textbooks, this LLM is designed to assist clinicians by suggesting potential diagnoses based on patient interviews. The government plans to deploy this model in medical institutions to support clinical decision-making and improve operational efficiency [67], [68]

Chapter 3

Methodology

3.1 Research Design

Outline the research design, including the type of study (e.g., experimental, observational) and the methods used to collect and analyze data.

3.2 Data Collection

Describe how you collected data for your research, including any surveys, interviews, or experiments conducted.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

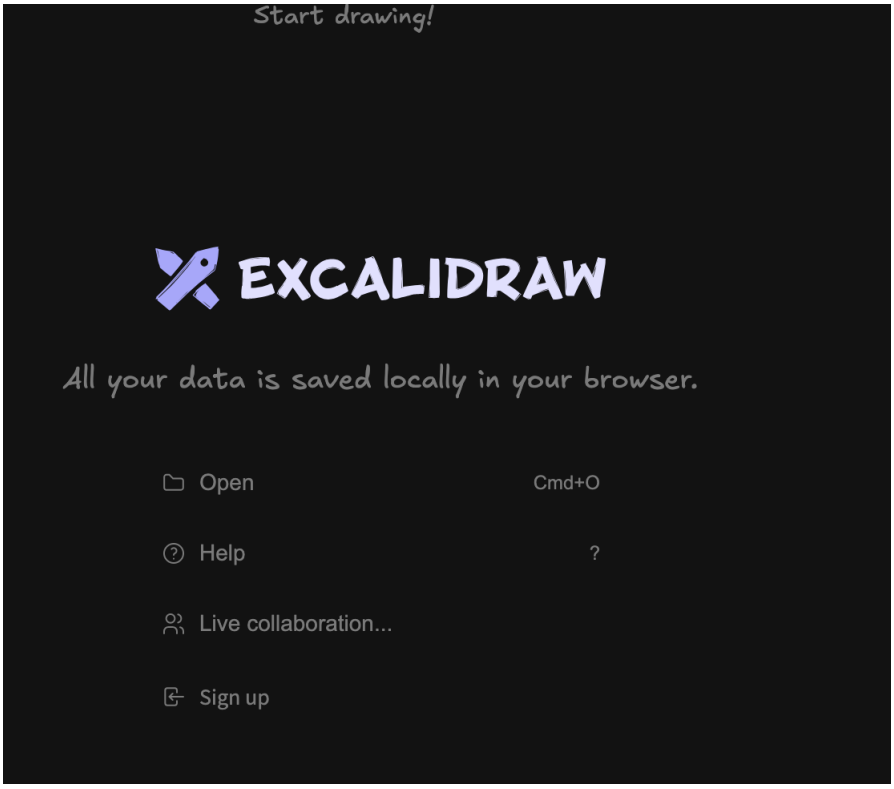

Figure 3.1: The typical RAG pipeline.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Table 3.1: Example of a table caption.

Column 1	Column 2	Column 3
Data A	123	X
Data B	456	Y
Data C	789	Z

Table 3.2: Example of a table caption.

Column 1	Column 2	Column 3
Data A	123	X
Data B	456	Y
Data C	789	Z

3.3 Data Analysis

Explain the methods used to analyze the data, including any statistical tests or software used.

3.4 Implementation

Describe how you implemented the proposed system, including any algorithms or models used.

3.5 Evaluation

Explain how you evaluated the performance of your system, including any metrics used to measure accuracy, precision, recall, etc.

3.6 Ethical Considerations

Discuss any ethical considerations related to your research, including data privacy, consent, and potential biases in the model.

Chapter 4

Results and Discussion

Present the results of your research, along with discussions on the implications and findings.

Chapter 5

Conclusion

Summarize the key findings of your research and suggest future work or improvements.

Bibliography

[1] Pemerintah Republik Indonesia, *Undang-undang republik indonesia nomor 37 tahun 1999 tentang hubungan luar negeri*, Accessed: May 17, 2025, 1999. [Online]. Available: https://peraturan.bpk.go.id/Home/Details/44910/uu-no-37-tahun-1999.

[2] Pemerintah Republik Indonesia, *Keputusan presiden republik indonesia nomor 108 tahun 2003 tentang organisasi perwakilan republik indonesia di luar negeri*, Ditetapkan di Jakarta pada tanggal 31 Desember 2003, 2003. Accessed: May 17, 2025. [Online]. Available: https://peraturan.bpk.go.id/Details/56472/keppres-no-108-tahun-2003.

[3] Komisi Pemilihan Umum Republik Indonesia, *Keputusan komisi pemilihan umum nomor 301 tahun 2024 tentang perubahan kedua atas keputusan komisi pemilihan umum nomor 857 tahun 2023 tentang penetapan rekapitulasi daftar pemilih tetap tingkat nasional dalam penyelenggaraan pemilihan umum tahun 2024*, Ditetapkan di Jakarta pada 04 Maret 2024 oleh Ketua KPU Hasyim Asy'ari, 2024. Accessed: May 17, 2025. [Online]. Available: https://jdih.kpu.go.id/keputusan-kpu/detail/N0jndiSNbf2eVSfDfRDZNXRWQ01mU1hXYk8vYnJpeEpjMHJ0V3c9PQ.

[4] Tempo.co. "Kementerian luar negeri akan gunakan ai untuk pelayanan wni di luar negeri." Diakses pada 17 Mei 2025, Accessed: May 17, 2025. [Online]. Available: https://www.tempo.co/internasional/kementerian-luar-negeri-akan-gunakan-ai-untuk-pelayanan-wni-di-luar-negeri-1207028.

[5] Kementerian Luar Negeri Republik Indonesia. “Portal pelayanan dan pelindungan wni di luar negeri.” Jakarta: Kementerian Luar Negeri RI, Accessed: May 17, 2025. [Online]. Available: https://peduliwni.kemlu.go.id/beranda.html.

[6] Kementerian Luar Negeri Republik Indonesia, *Kementerian luar negeri dan un women memperkuat pelindungan perempuan pekerja migran indonesia melalui inovasi chatbot ai sari*, Artikel di situs resmi Kementerian Luar Negeri RI, Accessed: May 17, 2025, Mar. 2025. [Online]. Available: https://kemlu.go.id/berita/kementerian-luar-negeri-dan-un-women-memperkuat-pelindungan-perempuan-pekerja-migran-indonesia-melalui-inovasi-chatbot-ai-sari?type=publication.

[7] GoToCompany and A. Singapore, *Gemma2 9b cpt sahabat-ai v1 instruct*, https://huggingface.co/GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct, Indonesian-focused instruction-tuned language model, 2025.

[8] GoToCompany and A. Singapore, *Gemma2 9b cpt sahabat-ai v1 instruct*, https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct, Indonesian-focused instruction-tuned language model, 2025.

[9] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[10] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd. 2025, Online manuscript released January 12,

2025. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/.

[11] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[12] S. Gupta, R. Ranjan, and S. N. Singh, *A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions*, 2024. [Online]. Available: https://arxiv.org/abs/2410.12837.

[13] P. Kelly. "How governments are using ai: 8 real-world case studies." Accessed: May 17, 2025. [Online]. Available: https://blog.govnet.co.uk/technology/ai-in-government-case-studies.

[14] Government Technology Agency of Singapore (GovTech). "Activate public-facing chatbots and serve citizens better with vica." Last updated: May 14, 2025. Accessed: May 17, 2025. [Online]. Available: https://www.tech.gov.sg/products-and-services/for-government-agencies/productivity-and-marketing/vica/.

[15] U.S. Department of State. "Department of state ai inventory 2024." Accessed: May 17, 2025. [Online]. Available: https://2021-2025.state.gov/department-of-state-ai-inventory-2024/.

[16] A. Molaee, "Ai embassies: A new frontier in cyber domain," *Journal of Cyberspace Studies*, vol. 9, no. 1, pp. 203–227, 2025. [Online]. Available: https://jcss.ut.ac.ir/article_100581.html.

[17] H. Mostafaei, S. Kordnoori, M. Ostadrahimi, S. seyed agha banihashemi, and D. Debo, "Applications of artificial intelligence in global diplomacy: A review of research and practical models," *Sustainable Futures*, vol. 9, pp. 1–15, Feb. 2025.

[18] S. Karzhev. “Advanced rag techniques.” Accessed: May 17, 2025. [Online]. Available: https://www.datacamp.com/blog/rag-advanced.

[19] Google AI. “Fine-tune gemma using hugging face transformers and qlora.” Accessed: May 17, 2025. [Online]. Available: https://ai.google.dev/gemma/docs/core/huggingface_text_finetune_qlora.

[20] S. Ratnakar, A. Talasila, R. Chamadiya, N. Agarwal, and V. K. Doifode, *Beyond qa pairs: Assessing parameter-efficient fine-tuning for fact embedding in llms*, 2025. [Online]. Available: https://arxiv.org/abs/2503.01131.

[21] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., “Improving language understanding by generative pre-training,” 2018.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[23] T. B. Brown et al., *Language models are few-shot learners*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[24] G. Team et al., *Gemma: Open models based on gemini research and technology*, 2024. [Online]. Available: https://arxiv.org/abs/2403.08295.

[25] B. Weng, *Navigating the landscape of large language models: A comprehensive review and analysis of paradigms and fine-tuning strategies*, 2024. [Online]. Available: https://arxiv.org/abs/2404.09022.

[26] OpenAI et al., *Gpt-4 technical report*, 2024. [Online]. Available: https://arxiv.org/abs/2303.08774.

[27] H. Touvron et al., *Llama 2: Open foundation and fine-tuned chat models*, 2023. [Online]. Available: https://arxiv.org/abs/2307.09288.

[28] G. Team et al., *Gemini: A family of highly capable multimodal models*, 2025. [Online]. Available: https://arxiv.org/abs/2312.11805.

[29] A. Chowdhery et al., *Palm: Scaling language modeling with pathways*, 2022. [Online]. Available: https://arxiv.org/abs/2204.02311.

[30] Anthropic, *Claude 3.7 sonnet and claude code*, Accessed: 2025-05-17, Feb. 2025. [Online]. Available: https://www.anthropic.com/news/claude-3-7-sonnet.

[31] GoTo Gojek Tokopedia and Indosat Ooredoo Hutchison, *Indosat ooredoo hutchison and goto launch sahabat-ai: Indonesia's open- source llm for empowering digital sovereignty*, Accessed: 2025-05-17, Nov. 2024. [Online]. Available: https://www.gotocompany.com/en/news/press/indosat-ooredoo-hutchison-and-goto-launch-sahabat-ai-indonesias-open-source-llm-for-empowering-digital-sovereignty?utm_source.

[32] GoToCompany, *Gemma2 9b cpt sahabat-ai v1 instruct*, https://huggingface.co/GoToCompany, Indonesian-focused instruction-tuned language model, 2025.

[33] R. Ng et al., *Sea-lion: Southeast asian languages in one network*, 2025. [Online]. Available: https://arxiv.org/abs/2504.05747.

[34] W. Zhang et al., *Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages*, 2024. [Online]. Available: https://arxiv.org/abs/2407.19672.

[35] L. Dou et al., *Sailor: Open language models for south-east asia*, 2024. [Online]. Available: https://arxiv.org/abs/2404.03608.

[36] NVIDIA. "What is retrieval-augmented generation, aka rag?" Accessed: 2025-05-18. [Online]. Available: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/.

[37] Weka.io. "Retrieval augmented generation: Everything you need to know about rag in ai." Accessed: 2025-05-18. [Online]. Available: https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/.

[38] R. Xu et al., *Simrag: Self-improving retrieval-augmented generation for adapting large language models to specialized domains*, 2025. [Online]. Available: https://arxiv.org/abs/2410.17952.

[39] Y. Gao et al., *Retrieval-augmented generation for large language models: A survey*, 2024. [Online]. Available: https://arxiv.org/abs/2312.10997.

[40] Y. Ma, Y. Cao, Y. Hong, and A. Sun, "Large language model is not a good few-shot information extractor, but a good reranker for hard samples!" In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics, 2023. [Online]. Available: http://dx.doi.org/10.18653/v1/2023.findings-emnlp.710.

[41] B. Jin et al., *Search-r1: Training llms to reason and leverage search engines with reinforcement learning*, 2025. [Online]. Available: https://arxiv.org/abs/2503.09516.

[42] L. Kilpatrick, Z. Gleicher, and P. Shah. "State-of-the-art text embedding via the gemini api." Accessed: 2025-05-18. [Online]. Available: https://developers.googleblog.com/en/gemini-embedding-text-model-now-available-gemini-api/.

[43] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, *Mteb: Massive text embedding benchmark*, 2023. [Online]. Available: https://arxiv.org/abs/2210.07316.

[44] K. Enevoldsen et al., *Mmteb: Massive multilingual text embedding benchmark*, 2025. [Online]. Available: https://arxiv.org/abs/2502.13595.

[45] W. Wongso, A. Joyoadikusumo, D. S. Setiawan, and S. Limcorn, *Lazarusnlp/indonesian-sentence-embeddings: V0.0.1*, version v0.0.1, Apr. 2024. [Online]. Available: https://github.com/LazarusNLP/indonesian-sentence-embeddings/tree/v0.0.1.

[46] Y. A. Malkov and D. A. Yashunin, *Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs*, 2018. [Online]. Available: https://arxiv.org/abs/1603.09320.

[47] A. Chirkin, A. Naruse, T. Fehér, Y. Wang, and C. Nolet, "Accelerating vector search: Nvidia cuvs ivf-pq part 1, deep dive," *NVIDIA Technical Blog*, Jul. 2024. [Online]. Available: https://developer.nvidia.com/blog/accelerating-vector-search-nvidia-cuvs-ivf-pq-deep-dive-part-1.

[48] A. Payong and S. Mukherjee, "How to choose the right vector database for your rag architecture," *DigitalOcean Community*, Dec. 2024, Accessed: 2025-05-18. [Online]. Available: https://www.digitalocean.com/community/conceptual-articles/how-to-choose-the-right-vector-database.

[49] G. Zhang et al., "Leveraging long context in retrieval augmented language models for medical question answering," *npj Digital Medicine*, vol. 8, no. 1, p. 239, 2025. [Online]. Available: https://doi.org/10.1038/s41746-025-01651-w.

[50] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, ser. CAIN '24, Lisbon, Portugal: Association for Computing Machinery, 2024, pp. 194–199. [Online]. Available: https://doi.org/10.1145/3644815.3644945.

[51] Y. Zhou et al., *Trustworthiness in retrieval-augmented generation systems: A survey*, 2024. [Online]. Available: https://arxiv.org/abs/2409.10102.

[52] L. Brehme, T. Ströhle, and R. Breu, *Can llms be trusted for evaluating rag systems? a survey of methods and datasets*, 2025. [Online]. Available: https://arxiv.org/abs/2504.20119.

[53] B. An, S. Zhang, and M. Dredze, *Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models*, 2025. [Online]. Available: https://arxiv.org/abs/2504.18041.

[54] E. J. Hu et al., *Lora: Low-rank adaptation of large language models*, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685.

[55] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, *Qlora: Efficient finetuning of quantized llms*, 2023. [Online]. Available: https://arxiv.org/abs/2305.14314.

[56] Z. Hu et al., *Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models*, 2023. [Online]. Available: https://arxiv.org/abs/2304.01933.

[57] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597. [Online]. Available: https://aclanthology.org/2021.acl-long.353/.

[58] Z. Chen et al., *Rulerag: Rule-guided retrieval-augmented generation with language models for question answering*, 2025. [Online]. Available: https://arxiv.org/abs/2410.22353.

[59] C. Zhou et al., "Lima: Less is more for alignment," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, Curran Associates, Inc., 2023, pp. 55 006–55 021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/ac662d74829e4407ce1d126477f4a03a-Paper-Conference.pdf.

[60] A. Balaguer et al., *Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture*, 2024. [Online]. Available: https://arxiv.org/abs/2401.08406.

[61] S. Alghisi, M. Rizzoli, G. Roccabruna, S. M. Mousavi, and G. Riccardi, *Should we fine-tune or rag? evaluating different techniques to adapt llms for dialogue*, 2024. [Online]. Available: https://arxiv.org/abs/2406.06399.

[62] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, *Fine-tuning or retrieval? comparing knowledge injection in llms*, 2024. [Online]. Available: https://arxiv.org/abs/2312.05934.

[63] Google AI. "Fine-tune gemma in keras using lora." Accessed: May 17, 2025. [Online]. Available: https://ai.google.dev/gemma/docs/core/lora_tuning.

[64] GovTech Singapore, *Winning by innovating*, https://www.tech.gov.sg/media/technews/winning-by-innovating, Accessed May 2025, 2016.

[65] GovTech Singapore, *Pair - productivity ai assistant for public officers*, https://www.tech.gov.sg/products-and-services/for-government-agencies/productivity-and-marketing/pair, Accessed May 2025, 2025.

[66] S. Chong et al., "Peach: Ai chatbot for perioperative medicine," *arXiv preprint arXiv:2412.18096*, 2024, Accessed May 2025. [Online]. Available: https://arxiv.org/abs/2412.18096.

[67] National Institute of Informatics, *Japan unveils generative ai that passes national medical exam*, https://www.aa.com.tr/en/artificial-intelligence/japan-unveils-generative-ai-that-passes-national-medical-exam/3551960, Accessed: 2025-05-18, 2025.

[68] Center for Research and Development Strategy (CRDS), *Health & medical real world data infrastructure - catalyst of generative ai development in japan - crds-fy2023-sp-04*, https://www.jst.go.jp/crds/en/publications/CRDS-FY2023-SP-04.html, Accessed: 2025-05-18, 2024.

Appendix A

Additional Information

Include any additional data, charts, or detailed explanations that are important for understanding your research.