Design and Development of a RAG-Based Chatbot for Japan Visa Requirements

Kenji Sugino

July 2025

# Abstract

This thesis presents the design, implementation, and evaluation of a chatbot system that assists users with Japan visa requirements using Retrieval-Augmented Generation (RAG). The system enhances answer accuracy by combining neural generation with document retrieval and incorporates a custom-trained intent classification model to improve relevance. The solution targets the travel industry and simplifies access to visa-related information.

Figure 3.1: Architecture of the RAG-based chatbot system

# 1. Introduction

## 1.1 Background

Travelers seeking visa information often face inconsistent, outdated, or inaccessible official websites. Japan, in particular, has visa policies that vary by nationality, purpose, and length of stay.

## 1.2 Problem Statement

Traditional chatbots struggle to provide up-to-date and context-sensitive visa information. Static rule-based systems or generic language models often hallucinate or fail to handle real-world questions.

## 1.3 Objective

The goal of this research is to develop a RAG-based chatbot enhanced with a custom-trained intent classifier to deliver accurate and contextually relevant visa answers.

# 2. Related Work

* **RAG (Lewis et al., 2020)**: Combines retrieval with generation for knowledge-intensive tasks.
* **Intent Classification**: Used in customer service to route queries and improve accuracy.
* **Chatbots in Travel**: Applied in airlines, booking platforms, but underutilized in immigration assistance.

# 3. Methodology

## 3.1 System Overview

The chatbot pipeline consists of: - A user query input - An intent classification model - A filtered retrieval from a vector store - A generator model (e.g., LLaMA or Mistral) that creates the final response

## 3.2 Data Collection

* Scraped FAQs from Japan embassies
* Reddit (r/JapanTravel, r/immigration)
* Public visa datasets and ChatGPT-generated questions

## 3.3 Preprocessing

* Text cleaned and chunked into ~300-token documents
* Embedded using sentence-transformers
* Stored in FAISS for retrieval

## 3.4 Intent Classifier

* Trained on 1,000+ labeled questions across 8 visa categories
* Models used:
  + TF-IDF + Logistic Regression (baseline)
  + DistilBERT fine-tuned (final model)

# 4. Implementation

## 4.1 Tech Stack

* Python (Hugging Face, FAISS, Flask)
* Kotlin (optional for mobile front-end)
* Google Colab for model training
* Together.ai for optional inference hosting

## 4.2 RAG Pipeline

1. User submits a question
2. Classifier predicts the intent
3. Documents matching that intent are retrieved
4. Generator formulates a response

## 4.3 Interface

* Basic chat UI via Streamlit or FastAPI
* Supports English input; future version may support multilingual

# 5. Results

## 5.1 Evaluation Metrics

| Metric | Logistic Classifier | DistilBERT Classifier |
| --- | --- | --- |
| Accuracy | 84% | 91% |
| F1 Score | 0.81 | 0.88 |

## 5.2 Retrieval Quality

* Without classifier: ~68% relevant docs
* With classifier filter: ~86% relevant docs

## 5.3 Qualitative Examples

**Q:** What visa do I need if I’m attending a 2-week conference? **Without Classifier:** Long-term work visa info (incorrect)  
**With Classifier:** Short-stay business visa (correct)

# 6. Challenges

* Class imbalance during training
* Overlap between tourist and student visa questions
* Keeping data up to date with recent policy changes

# 7. Conclusion

This research demonstrates that integrating a custom intent classifier into a RAG-based chatbot improves visa-related question answering in the travel industry. Future work includes temporal-awareness models, multilingual support, and integration with embassy data APIs.

# References

Lewis et al. introduced the RAG architecture [1].

[1] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *arXiv preprint arXiv:2005.11401*, 2020.