

RESEARCH ARTICLE

Automated Question Answering for Thai Visa Services using GPT-4: System Architecture, Custom Dataset, and Performance Evaluation

Aulia Akhrian Syahidi^{1,*} 

¹Data Science and Artificial Intelligence Program, Thai-Nichi International College, Thai-Nichi Institute of Technology, Bangkok 10250, Thailand

*Corresponding author: Aulia Akhrian Syahidi, aulia.sya@tni.ac.th

Article Information

Article History:

Received: 10 June 2025

Revised: 13 October 2025

Accepted: 16 October 2025

Published: 22 June 2026

Keywords:

Automated Question Answering
GPT-4

Large Language Models
Thai Immigration Services
Visa Information System

Abstract

Thailand's immigration system includes various visa types, such as Tourist (TR), Business (Non-Immigrant B), Education (Non-Immigrant ED), ED Plus, Media (Non-Immigrant M), Official (Non-Immigrant F), Dependent/Spouse (Non-Immigrant O), SMART, Transit, Long-Term Resident (LTR), Retirement (O-A), and Long-Stay Retirement (O-X), each governed by distinct and frequently updated policies. This complexity causes confusion and difficulties for foreign nationals in understanding required documents, eligibility criteria, and policy changes, as shown by a survey involving 75 respondents in Bangkok. Without scalable and accurate solutions, these challenges can reduce service efficiency and public trust. To address this, the study employs GPT-4, a Large Language Model (LLM), optimized with prompt engineering and few-shot learning to automatically answer visa-related questions with accurate, clear, and policy-compliant responses. A simulated dataset of 1,600 question-answer pairs was created from official Thai immigration policies. GPT-4's performance was systematically evaluated using automatic metrics (BLEU, ROUGE-L, METEOR) and human assessments of accuracy, clarity, and user satisfaction, providing a robust methodological basis for evaluating its effectiveness in public service. Results show strong performance across most categories, with the highest BLEU scores observed in Tourist (0.82), Business (0.81), and Long-Term Resident (LTR) (0.78) visa-related queries. Corresponding peak ROUGE-L scores were 0.86 for Tourist, 0.84 for Business, and 0.81 for Long-Term Resident (LTR) visas, while METEOR scores peaked at 0.79 (Tourist), 0.78 (Business), and 0.76 (Long-Term Resident (LTR)). The system achieved an overall BLEU score of 0.75, ROUGE-L of 0.79, and METEOR of 0.73. Human evaluators rated GPT-4's responses with an average score of 4.5 for accuracy, 4.6 for language clarity, and 4.4 for user satisfaction (on a 5-point scale). A Pearson correlation of 0.78 between BLEU and human-rated accuracy indicates high alignment between automated and human evaluation. These results highlight GPT-4's potential in enhancing public-facing services such as immigration services by providing accurate, clear, and policy-aligned responses. Future work will focus on multilingual support, real-time policy updates, and deployment in live service environments.

1. Introduction

Public services are currently facing significant challenges in providing accurate and timely information to the public, especially in complex sectors such as immi-

gration services. According to data reported by Barber and Sciortino (2024), as a hub for intra-regional migration and a major destination country in Southeast Asia, Thailand hosts at least 5.3 million non-Thai nationals, marking an 8 percent increase from nearly 4.9

million non-Thai residents reported in the previous edition. Moreover, Thailand is a prominent destination for tourism, employment, and higher education, which significantly impacts the number of foreigners applying for various immigration services, including visa applications.

In Thailand, the immigration system includes various visa types comprising 12 categories, namely, Tourist (TR), Business (Non-Immigrant B), Education (Non-Immigrant ED), ED Plus, Media (Non-Immigrant M), Official (Non-Immigrant F), Dependent/Spouse (Non-Immigrant O), SMART, Transit, Long-Term Resident (LTR), Retirement (Non-Immigrant O-A), and Long-Stay Retirement (O-X), each governed by distinct and frequently updated policies and procedures. This regulatory complexity often leads to numerous inquiries from the public who seek precise, clear, and contextual explanations.

The complexity of regulations within Thailand's immigration system often raises numerous questions from the public, who require accurate, clear, and contextual explanations. This aligns with the findings of a survey conducted by the researchers involving 75 foreign nationals working or studying in Bangkok, Thailand. The respondents expressed complaints regarding visa procedures that were perceived as confusing, overly complicated, and lacking clear information that was easily accessible in a language they could understand. Many reported difficulties in understanding the required documents and eligibility conditions, ineffective communication of policy changes, and uncertainty in the application and renewal processes.

In addition, immigration services also face internal challenges such as heavy workloads for officers, lack of consistency in providing responses, and language barriers between officers and foreign visa applicants. These conditions may result in inaccurate information delivery, delays in application processing, and decreased user satisfaction (Betrand et al., 2023; Dongbo et al., 2023). Therefore, an Artificial Intelligence (AI)-based approach that is scalable and accurate is needed to enhance both the efficiency and quality of public services in the immigration sector.

With the rapid advancement of AI, Large Language Models (LLMs) such as the Generative Pre-trained Transformer 4 (GPT-4) have demonstrated remarkable capabilities in understanding and generating natural language across various contexts (Mai and Carson-Berndsen, 2024; Mahbub et al., 2024; Shahandashti et al., 2024). This technology has been increasingly applied in chatbots, automated Question-Answering (QA) systems, and virtual assistants across both public service and business domains (OpenAI, 2023; Brown et al., 2020). In particular, LLMs have been explored in the fields of law and public policy to simplify access to documents and provide policy-based responses (Liu et al., 2024a; Lai et al., 2024).

However, a research gap remains in the specific ap-

plication of GPT-4 within the immigration service domain, particularly in developing automated QA systems capable of comprehensively supporting various visa types. The complexity of Thailand's immigration regulations requires an LLM that excels in understanding intricate policies and generating accurate, context-aware responses. By leveraging GPT-4 through prompt engineering and few-shot learning, an automated QA system can be trained to provide answers tailored to specific visa categories. This approach has the potential to reduce the workload of immigration officers, improve the consistency of information, and deliver faster, more accurate, and policy-aligned public services.

The main objective of this study is to systematically evaluate the effectiveness of an LLM-based QA system powered by GPT-4 in providing accurate, clear, and policy-aligned responses to inquiries across 12 major Thai visa categories.

To achieve this objective, the study makes the following itemized contributions:

1. **Dataset Construction:** We built a specialized simulated dataset consisting of 1,600 question-answer pairs representing 12 major visa categories based on official Thai immigration regulations.
2. **System Architecture and Methodology:** We proposed and implemented an automated QA system architecture based on GPT-4, utilizing few-shot prompting and rule-based post-processing to enhance policy alignment and response accuracy.
3. **Comprehensive Evaluation:** We systematically evaluated GPT-4's performance using a combination of automated metrics (Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation based on longest common subsequence (ROUGE-L), and Metric for Evaluation of Translation with Explicit Ordering (METEOR)) and human assessments (accuracy, language clarity, and user satisfaction) to establish a robust methodological foundation.

The remainder of this paper is structured as follows: Section 2 reviews the related work. Section 3 details the materials and methods, including dataset construction and system architecture. Section 4 presents the experimental results. Section 5 discusses the findings and policy implications. Finally, Section 6 concludes the study and suggests directions for future work.

2. Related Works

Several previous studies have explored the application of technology to enhance the efficiency of visa application systems, either through predicting application outcomes or digitizing the verification process. One of the early studies by Swain et al. (2018) developed a data-driven classification model to predict the likelihood of

H1B visa approval in the United States with reasonably good accuracy. Their study utilized historical data encompassing job attributes, location, and salary and provided probabilistic estimates of application success. This approach proved beneficial for both applicants and companies in assessing their chances before submitting a visa application. However, a limitation they encountered was the unavailability of datasets related to the employer's nationality and the exclusion of applicant social media profiles and activities, as required under the new policy.

Another study by Yanbo and Kongjit (2020) specifically highlighted the challenges faced by Chinese nationals in applying for long-term visas to Thailand, such as the 60-day tourist, education (NON-ED), business (NON-B), and family (NON-O). Through interviews and surveys, the study identified key reasons for visa application failures, including a lack of understanding of regulations, incomplete documentation, and communication barriers. This research is significant as it reveals the complexities foreign users encounter within the Thai visa system.

In a recent development, a study conducted by Raj et al. (2023) applied data science techniques to analyze patterns in H1B work visa applications using a 2015–2019 dataset obtained from the U.S. Department of Labor. Focusing on predictive modeling, the study presented a statistical approach to identify trends and improve the prediction of application outcomes. The results demonstrate how machine learning can provide practical insights into the dynamics of visa applications in the destination country with reasonably good accuracy. However, the study recommends employing kernel-based approaches with multidimensional spaces to enhance the model's functionality in evaluating how well it fits the data for predicting future outcomes.

A recent study by Chowdhury et al. (2024) introduced a blockchain-based Self-Sovereign Identity (SSI) system using Hyperledger Indy to accelerate and secure the global visa application process. The system leverages Decentralized Identifiers (DIDs) and Verifiable Credentials (VCs) to integrate entities such as passport offices, visa centers, and immigration authorities. The study emphasizes the importance of cross-border interoperability in achieving an efficient and secure digital immigration process. However, it also acknowledges certain limitations and proposes enhancements to scalability and system performance by incorporating modern cryptographic algorithms, exploring AI-powered analytics for real-time threat detection, improving connectivity with other blockchain platforms, and integrating biometric authentication.

Furthermore, we highlight the rapidly growing trend of applying LLMs for government and policy QA systems. According to Gao et al. (2023), the use of LLMs is crucial for enhancing QA systems, particularly in government contexts, where they are valuable for improving administrative efficiency. Additionally, Liu et al.

(2024a) investigated the use of LLMs to create Legal QA systems, focusing on the ability of models such as GPT to interpret legal language and provide legally accurate advice. The core of the work emphasizes the need for advanced prompt engineering to achieve reliable performance in contexts where factual accuracy is critical.

A survey conducted by Lai et al. (2024) showed that LLMs for policy QA have the potential to assist stakeholders by providing efficient legal advice, accelerating case processing, and improving decision accuracy and consistency. However, challenges remain, including limitations in processing long texts, adapting to individual cases, and addressing privacy and ethical issues. The authors recommend improving data quality, enhancing privacy protection, developing model capabilities (e.g., GPT), and establishing adequate ethical and regulatory frameworks.

Equally important, Jiang et al. (2025) introduced ARGUS (Answer Retrieval and Governance Understanding System), an LLM-based QA system customized for government domains using a hybrid retrieval strategy with LlamaIndex. ARGUS improves factual consistency and contextual relevance through a combination of graph-based entity retrieval and text retrieval. Evaluation using classical metrics and RAGAS indicators demonstrated high performance (ROUGE-1 0.68; semantic relevance 0.81). The chain-of-thought mechanism enhanced interpretability, and ablation studies showed significant improvements over the baseline. Results also indicated that hybrid retrieval outperformed pure vector or pure graph-based strategies.

While previous studies have made significant contributions to visa processing, focusing mostly on outcome prediction, root cause analysis of administrative issues, or digitalizing processes through decentralized identity technology, there remains a research gap regarding the application of LLMs, particularly using GPT-4 as the latest model. Prior studies also support the recommendation of applying LLMs as automated QA systems capable of providing context-rich and policy-specific responses to government-related queries, especially in immigration.

Our study differentiates itself by shifting the focus from predicting visa outcomes or digitalizing administrative processes to directly addressing users' information deficits through a real-time policy QA system, using a carefully constructed dataset and few-shot prompting to ensure GPT-4 strictly adheres to policy-specific facts, an essential requirement in high-risk domains such as immigration. In doing so, the research evaluates not only the technical accuracy of responses using BLEU, ROUGE-L, and METEOR metrics, but also user assessments of clarity, satisfaction, and alignment with the underlying immigration policies.

3. Materials and Methods

3.1 Dataset Construction and Splitting

To evaluate GPT-4’s ability to automatically answer immigration-related questions, we constructed a simulated dataset consisting of 1,600 question–answer (QA) pairs grounded in official Thai immigration policy sources, covering the 12 predefined visa categories included in this study. Each QA pair was derived from authoritative policy knowledge bases, including immigration policy rules issued by the Thai Immigration Bureau, visa category classifications defined by the Ministry of Foreign Affairs of Thailand, as well as visa application guidelines and consular requirements from official Royal Thai Embassies and Consulates. These sources include legal guidelines, standardized application forms, official FAQs, and publicly available visa guidance documents, ensuring legal validity, policy consistency, and practical applicability.

Each question was formulated to resemble natural conversational English. For instance, examples include: “Can I bring my spouse on a SMART Visa?” and “What are the financial requirements for an O-X visa?” The answers were derived from official and current regulations while considering the context of the inquiry, such as the length of stay, applicant type (e.g., student, retiree, or journalist), and required supporting documents.

To ensure a fair, objective, and data-leakage-free model evaluation process, the dataset consisting of 1,600 QA pairs was divided into three main subsets: a training set, a validation set, and a test set. This data split was performed using a stratified sampling approach adopted from Sharma et al. (2023) and Lu et al. (2023), based on visa categories to ensure that each visa type was proportionally represented in all subsets. The detailed data distribution is presented in Table 1.

Table 1. Dataset splitting for GPT-4 evaluation.

Dataset subset	Number of Q&A pairs	Percentage
Training set	960	60%
Validation set	320	20%
Test set	320	20%

The data split presented in Table 1 serves specific purposes in the prompt engineering process and model evaluation, as described by Dai et al. (2025). The training set was used to construct examples for few-shot prompting scenarios, enabling GPT-4 to understand the context of questions related to Thai immigration policies. The validation set was used to test and refine the prompt engineering approach prior to the final evaluation. Meanwhile, the test set was used to comprehensively assess GPT-4’s performance using both automated metrics, such as BLEU, ROUGE-L, and METEOR, and human evaluation. Table 2 summarizes the

distribution of QA pairs across the 12 main visa categories included in the simulated dataset.

Table 2. Dataset partition for evaluating GPT-4.

Visa category	Training (60%)	Validation (20%)	Test (20%)	Total Q&A Pair
Tourist (TR)	80	27	27	134
Business (Non-Immigrant B)	90	30	30	150
Education (Non-Immigrant ED)	100	33	33	166
ED Plus	90	30	30	150
Media (Non-Immigrant M)	80	27	27	134
Official (Non-Immigrant F)	70	23	23	116
Dependent/Spouse (Non-Immigrant O)	80	27	27	134
SMART	90	30	30	150
Transit	70	23	23	116
Long-Term Resident (LTR)	80	27	27	134
Retirement (Non-Immigrant O-A)	70	23	23	116
Long-Stay Retirement (Non-Immigrant O-X)	80	27	27	134
<i>Total</i>	<i>960</i>	<i>320</i>	<i>320</i>	<i>1,600</i>

3.2 System Architecture

The system architecture was designed to support an automated QA service powered by GPT-4 within the context of Thai immigration visa policies. The system consists of several key components that work in an integrated manner, from the user interface to the output evaluation stage. Each component plays a specific role in ensuring that user questions are processed efficiently and that accurate, policy-compliant answers are generated.

Fig. 1 presents a visual representation of the system architecture. This modular approach allows flexible development and enhancement of the system as needed. Each component in the architecture plays a distinct role and is organized within a modular structure. The following sections describe the function of each component.

3.2.1 User Interaction

The system interface is designed as a simple web-based chatbot (see Fig. 2) to facilitate intuitive user interaction. This component enables users to submit queries in natural English using a conversational style. The system supports real-time interaction, in which user inputs are processed and responses are generated dynamically. This approach allows users to interact with the system in a natural and user-friendly manner without requiring structured or predefined query formats.

3.2.2 Query Preprocessing

Before being sent to the model, user queries submitted through the interface undergo a multi-stage preprocessing phase to ensure that the input is clean, standardized, and ready for GPT-4 inference.

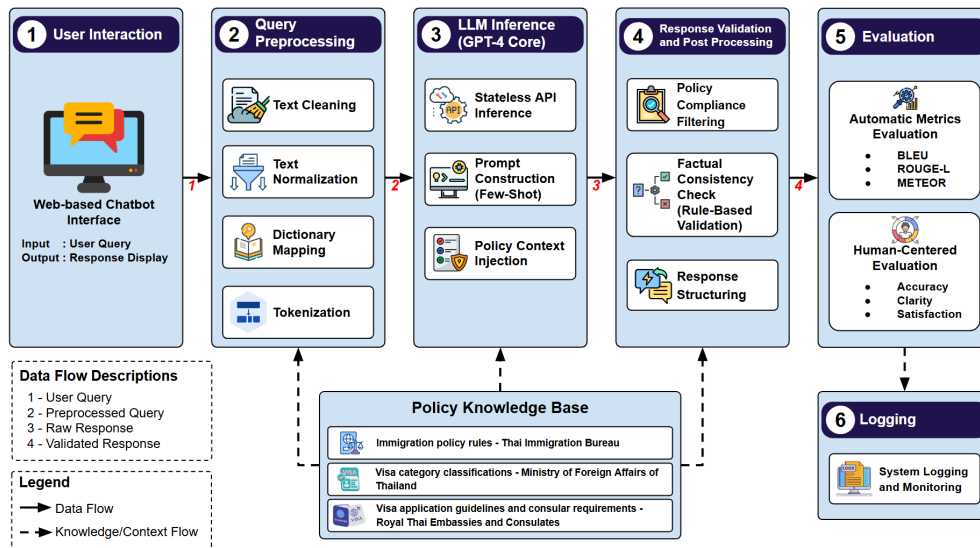


Figure 1. System architecture.

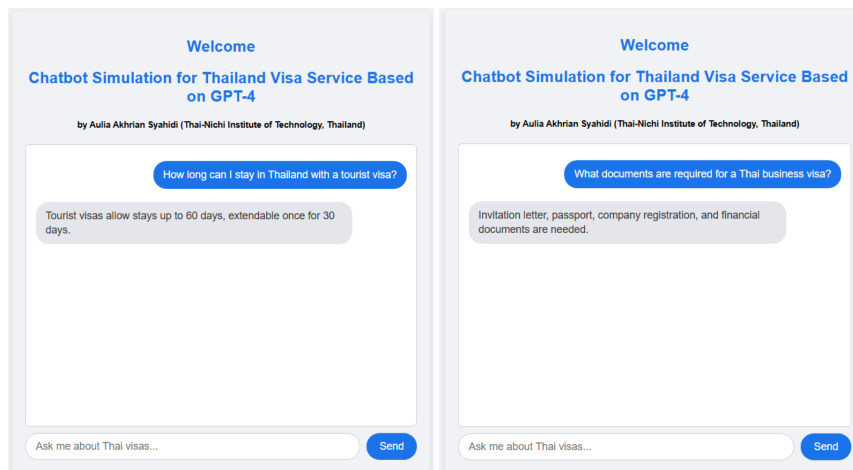


Figure 2. Web-based chatbot interface.

The process begins with a cleaning stage, which involves removing irrelevant characters such as excessive spaces, special symbols, and non-alphabetic punctuation marks using regular expressions (regex) to standardize the text format. Subsequently, normalization is performed to simplify sentence structures, correct common typographical errors, and resolve ambiguous terminology to improve contextual clarity. Ambiguous or multi-interpretable entities frequently found in user queries, such as “short stay,” “partner,” or “permanent resident,” are mapped to their official policy equivalents. For instance, “short stay” is normalized to “Single-Entry Tourist Visa” or “Visa Exemption Status,” depending on the context, through a simple dictionary-based replacement technique for specific policy-related terms. Finally, the cleaned and normalized text is tokenized in preparation for submission to the GPT-4 Application Programming Interface (API). Since GPT-4 employs the `cl100k_base` tokenizer, the input text is segmented into tokens, which serve as the

fundamental units of text processed by the LLM.

3.2.3 LLM Inference (GPT-4 Core)

The core system module in this study is based on GPT-4, a large language model (LLM) developed by OpenAI and accessed through an API endpoint in a stateless manner. GPT-4 is built upon the Transformer architecture (Yekta, 2024) and optimized using Reinforcement Learning from Human Feedback (RLHF), which aligns model outputs with human preferences regarding safety, helpfulness, and factual grounding (Barman et al., 2025). GPT-4 is trained on an extensive multi-modal dataset consisting of text and code, enabling it to generate text, translate languages, create diverse types of content, and provide informative answers to a wide range of questions. Compared with its predecessors, GPT-3 and GPT-2, GPT-4 incorporates a substantially larger and more sophisticated architecture based on a Mixture-of-Experts (MoE) design, with an estimated

1.76 trillion total parameters (OpenAI, 2023; Olivato et al., 2024). This scale and architectural complexity allow the model to learn intricate patterns from a considerably larger volume of training data. As a result, GPT-4 produces more precise, nuanced, and contextually coherent outputs across diverse tasks (Shultz et al., 2025; OpenAI, 2023).

The model adopts a decoder-only neural architecture based on self-attention and feed-forward networks, which constitute the core components of the Transformer framework. GPT-4 relies exclusively on the decoder component to process hidden-state sequences and generate output tokens autoregressively. Fundamentally, GPT-4 predicts the next token based on the preceding context, making it highly effective for producing coherent and contextually appropriate text (Uruij et al., 2025). This capability for complex language understanding and factual adherence makes GPT-4 particularly suitable for high-stakes domains requiring a high degree of accuracy, such as public immigration services.

Given the stateless nature of the API, each interaction is treated as an independent session without contextual memory. To ensure consistent and context-aware responses, this study employs a few-shot prompting strategy, in which several representative input-output examples are embedded within each API request (Brown et al., 2020; Bahrami et al., 2023; Konduru et al., 2024). This approach enables GPT-4 to infer the expected task patterns and generate outputs that conform to the desired format and semantics, even in the absence of persistent contextual information. Moreover, few-shot prompting was selected for two primary reasons: cost-effectiveness, as it avoids the need for expensive retraining, and rapid iteration, which allows flexible adaptation to policy updates or system requirements through prompt modification.

Few-shot examples are selected dynamically based on the initial identification of the visa category inferred from the user input. The selection criteria are based on visa relevance, whereby examples are drawn from the identified visa category, and topical representation, ensuring that the selected examples cover frequently asked questions and key policy topics aligned with the official guidelines of the Thai Immigration Bureau, including required documents, duration of stay, work eligibility, and extension policies.

3.2.4 GPT-4 Prompt Template Design

This template is designed to guide the GPT-4 model to assume the role of a policy authority, prioritize factual accuracy, and provide structured responses. The system prompt, shown in Table 3, defines the primary roles and rules, including instructions that govern the model's behavior, tone, and knowledge boundaries. It is sent only once per session or as part of the initial API call.

To manage uncertainty and reduce the risk of pro-

viding incorrect information, a fallback clause is included in the prompt (see Table 3, Fallback Mechanism Components). This fallback mechanism is essential for enabling self-correction in the LLM through feedback-based methods (Sun et al., 2024; Liu et al., 2024b), reflecting the precautionary principle commonly applied in the implementation of automated public services.

For example, in the Business (Non-Immigrant B) – Documents category, diverse and detailed few-shot examples were constructed to cover key aspects, including General Document Requirements (Invitation Letter), Supporting Documents from the Thai Company, Consequences of Incomplete Documents, and Required Documents After Approval. Table 4 presents the design of the few-shot examples for this category.

The final instruction sent to the GPT-4 model is a single, hierarchically organized text designed to maximize performance and policy compliance. This structure consists of three mandatory components. First, the [COMPLETE SYSTEM PROMPT] serves as the foundational guardrail, containing instructions that define the model's role (persona), rules, knowledge limitations, and fallback mechanisms, thereby establishing the model as an accurate immigration officer. Second, the instruction is followed by six dynamically selected few-shot examples ([DYNAMIC FEW-SHOT EXAMPLES: Q1–A6]). This component comprises six question-answer pairs selected from the Non-Immigrant B dataset, covering topic variations such as mandatory documents, exemptions, and visa-status conversion restrictions, thereby guiding the model toward the desired answer format and factual depth. Finally, the prompt includes the USER QUESTION: [cleaned and normalized user input], representing the question submitted by the user after preprocessing, in which typographical errors have been corrected and ambiguous terms standardized. This component is immediately followed by the ANSWER: tag, which serves as an explicit signal for GPT-4 to begin generating output that adheres to the provided instructions and examples.

An example of real-world user queries and GPT-4 outputs is presented in Table 5, illustrating variations in user input, the preprocessing actions performed (Normalization), and the final outputs generated by GPT-4 after processing through the System Prompt and relevant Few-Shot Examples.

The normalization technique was applied to bridge the gap between natural, informal language and the formal language used in immigration policies while ensuring uniform input for the GPT-4 module. This process employs a hybrid approach centered on a Rule-Based Mapping Algorithm (Dictionary Lookup) to resolve terminology-related issues (Gadag and Sagar, 2016; Bollmann et al., 2011). The algorithm maps ambiguous or abbreviated user phrases, such as replacing “Non-B” with “Non-Immigrant B Visa” or converting ‘holiday visa’ into “Tourist Visa (TR),”, to their official equivalents, which is important for maintaining factual accu-

Table 3. System prompt (main roles and rules).

Component	Content	Key Objective
System Persona	You are a professional Thai Immigration Officer responsible for providing visa information.	Establishes expertise and an authoritative tone.
Policy Rules (Knowledge Cutoff)	Your task is to answer all user questions clearly, concisely, and accurately, based solely on official Thai immigration policies.	Limits model creativity and emphasizes factual accuracy.
Clarification and Structure	Explicitly include the visa category name (e.g., Tourist (TR)) in the first response. Use formal official language (e.g., English, Indonesian, etc.)	Ensures responses are verifiable and well-structured.
Fallback Mechanism	If you are unsure, or the question falls outside the scope of the policy documents you are aware of, you MUST NOT guess or fabricate an answer. Respond with the standard phrase: "For the most accurate and up-to-date information regarding your specific case, please contact the Thai Immigration Office directly."	Ensures service safety and ethics by preventing hallucinations.

racy within the domain. In addition, a Lexical Substitution Algorithm is employed for sentence standardization (Vladika et al., 2025; Naser-Karajah and Arman, 2024; Roller and Erk, 2016). This process transforms incomplete or informal user questions (e.g., Raw User Query: "Is just the invitation letter from the office enough?") into complete and formal policy-oriented queries (e.g., Query After Normalization: "Is the Invitation Letter alone sufficient to apply for a Non-Immigrant B Visa?"), thereby preparing optimized input for accurate few-shot prompting.

Next, the GPT-4 generation parameters were carefully configured to balance factual accuracy and limited randomness, which is particularly important in the immigration policy domain that requires strict compliance. The model used was GPT-4-0314, selected for its stability and strong reasoning capabilities at the time of the study, ensuring consistent and reproducible behavior throughout the experiments. To achieve deterministic outputs, the Temperature parameter was set to a low value (0.2 on a scale from 0 to 1), thereby reducing creativity and minimizing the risk of hallucinations. Meanwhile, Top_p was maintained at 0.9 to allow limited flexibility in selecting high-probability tokens, which helps preserve natural language fluency without compromising factual accuracy. Finally, Max_tokens was limited to 256 to ensure concise and efficient responses while preventing excessive elaboration. To ensure methodological transparency, all system inference data were collected through the GPT-4 API during March 12–16, 2025.

The output generated by GPT-4 is subsequently passed to a post-processing phase, where rule-based filters are applied. These filters verify that the response explicitly mentions the correct visa type and adheres to standardized formatting for public-service communication. All input and output interactions are logged systematically to support performance evaluation, iterative prompt refinement, and documentation of failure cases. This modular system architecture provides flexibility and scalability as immigration policies evolve and new visa categories are introduced.

3.2.5 Response Validation and Post Processing

The post-processing phase is designed to ensure the quality, clarity, and regulatory compliance of the answers generated by GPT-4. After the model produces an initial response, the output undergoes a refinement stage aimed at removing redundant phrases, normalizing formatting, and aligning the response structure with the communication standards used in official immigration documents. This process includes the use of concise yet complete phrasing, consistent terminology, and a polite and professional tone appropriate for public-service contexts.

During this phase, the system also applies a series of rule-based filtering mechanisms to ensure that GPT-4 outputs comply with policy requirements and maintain structural consistency before being presented to the end user. The mechanism consists of three validation levels designed to address potential inconsistencies or ambiguities that may arise from the LLM.

The first level involves two mandatory checks: Visa Category Mention Validation and Logical Consistency Checking. Visa Category Mention Validation uses Regular Expressions (Regex) to detect normalized visa-category keywords (e.g., Tourist (TR) or Non-Immigrant B) within the first 20 tokens of the response. If no visa category is detected, the appropriate visa category is automatically inserted at the beginning of the output to ensure clarity. Meanwhile, Logical Consistency Checking operates through a predefined key-value schema containing core policy constraints (e.g., "O-X Visa holders are not permitted to work").

The second level is Factual Inconsistency Handling. If Logical Consistency Checking detects a conflict, such as when the GPT-4 output contains phrases implying permission or uncertainty (e.g., "may be allowed" or "not sure"), while the corresponding policy constraint specifies "PROHIBITED," the system immediately overrides the conflicting statement. This mechanism prevents the generation of responses that are legally contradictory or inconsistent with established

Table 4. Few-shot example for Business (Non-Immigrant B).

Question Category	User Input (Q)	Reference Output (A)
Main Mandatory Requirement	What key documents must my company in Thailand prepare for my Business Visa?	To support your Non-Immigrant B (Business) application, the Thai company must provide: an official Invitation/Sponsorship Letter specifying the purpose and duration of your stay, Proof of Company Registration (e.g., Affidavit of the Registrar), and an updated List of Shareholders.
Obligations of the Thai Company	Besides the invitation letter, what other mandatory documents must the sponsoring company include?	The sponsoring company must provide the audited Financial Statements (Balance Sheet and Profit and Loss Statement) for the most recent year. These documents must match those submitted to the Thai tax authorities.
Indonesian Passport Document Procedure	If I am an Indonesian citizen, does my passport need a minimum validity for the Business Visa?	Yes. To apply for a Non-Immigrant B, your passport must have a minimum validity of six (6) months from your planned date of arrival in Thailand. Passports with less than six months' validity will be rejected.
Consequences of Completeness	What happens if I forget to include the work permit proof from the Thai company?	If the work permit proof from the sponsoring company is incomplete, your Non-Immigrant B application will be delayed or rejected. Complete documentation from the Thai company is an absolute requirement for approval.
Post-Approval Documents	After obtaining the Business Visa, do I need to bring additional documents upon arrival in Thailand?	Yes. After the visa is issued, upon arrival you may be asked to present a Guarantee Letter from your company and proof of sufficient funds to cover living expenses during your stay.
Status Conversion Restriction	Can I change from a Tourist Visa to a Business Visa in Thailand?	Generally, this is not allowed. You must apply for a Non-Immigrant B from outside Thailand, through a Thai Embassy or Consulate. Status changes from a Tourist (TR) within the country are only permitted under limited emergency circumstances.

immigration policies. For example, ambiguity regarding the work eligibility of O-X Visa holders is replaced with the definitive statement: “No, holders of the Long-Stay Retirement (O-X) Visa are not permitted to work in Thailand.”

The final level is the Fallback Trigger, which ensures that model uncertainty is safely redirected to official authorities. This mechanism uses string matching or regex-based detection to identify uncertainty markers generated by GPT-4, such as ‘I am not sure’ or incomplete fallback phrases. If a trigger condition is detected, the system automatically appends a standardized fallback clause to the end of the output. The clause states: “For the most accurate and up-to-date information regarding your specific case, please contact the Thai Immigration Office directly.” This mechanism ensures that users are referred to an official source whenever the system cannot provide a definitive answer.

The implementation of rule-based filtering improves system reliability by correcting uncertainties and inconsistencies in GPT-4 outputs. One representative example involves questions regarding employment eligibility under the Long-Stay Retirement (O-X) Visa category. A GPT-4 response containing uncertainty (e.g., “I am not sure about working”) is immediately processed through the Logical Consistency Checking mechanism. The uncertain statement is replaced with the corresponding policy constraint (‘Not permitted to work in Thailand’), while the Fallback Trigger simultaneously appends the standardized redirection clause.

For queries related to Education (ED) Visa docu-

mentation, the system performs an additional formatting correction. The Visa Category Mention Validation mechanism automatically inserts the full visa-category label (“Education Visa (Non-Immigrant ED)”) at the beginning of the response when the original output contains only the abbreviated term “ED visa.”

Through these mechanisms, the system ensures that every response remains factually accurate and structurally consistent with Thai immigration service standards, regardless of variations in the quality of the model’s initial output.

3.2.6 Logging and Evaluation

To ensure reliable tracking of every interaction and facilitate auditing, the system utilizes an SQLite database as a lightweight embedded repository for recording all user input–output pairs, inference parameters, and timestamps. This setup ensures that every interaction can be fully traced through an audit trail, supporting post-mortem evaluation and failure analysis.

The recorded input–output pairs also serve as the foundation for systematic performance evaluation and quantitative analysis of GPT-4 responses. All interactions are stored within the logging system for tracking and subsequent analysis. The system evaluation is conducted using two complementary approaches. First, automatic evaluation is applied to 320 QA pairs from the test dataset using three metrics: BLEU, ROUGE-L, and METEOR.

BLEU measures the similarity of n-grams between

Table 5. Real-world usage examples (User Queries) and GPT-4 output results.

Raw User Query	User Query After Normalization	Expected Output (GPT-4)
What are the required documents from the Thai side?	What mandatory documents from the Thai side are needed to apply for a Non-Immigrant B Visa?	The mandatory documents from the Thai side include the Official Invitation Letter, Proof of Company Registration (Affidavit of the Registrar), and the audited Financial Statement (Balance Sheet and Profit and Loss Statement) for the most recent year. These documents must match those submitted to the Thai tax authorities.
Is just the invitation letter from the office enough?	Is the Invitation Letter alone sufficient to apply for a Non-Immigrant B Visa?	No, it is not sufficient. In addition to the Official Invitation Letter, you must include other supporting documents from the sponsoring company, such as Proof of Company Registration and the audited Financial Statement.
Can I upgrade my Tourist Visa to a Business Visa at Bangkok Immigration?	Is it possible to change a Tourist Visa (TR) to a Non-Immigrant B Visa at the Bangkok Immigration Office?	Generally, this is not allowed. Applications for a Non-Immigrant B Visa must be made from outside Thailand. Changing visa status within the country is only permitted in very limited emergency situations.
I want to start working, but my visa is still a Tourist Visa. What should I do?	What is the procedure to change my current Tourist Visa to a Work Visa (Non-Immigrant B)?	The standard procedure is to apply for a Non-Immigrant B Visa from outside Thailand. You cannot directly convert a Tourist Visa (TR) to a Work Visa from within Thailand.

model-generated answers and reference answers, providing a quantitative assessment of word- and phrase-level precision (Syahidi and Kiyokawa, 2025; Gupta et al., 2024; Onan and Alhumyani, 2024; Wolk and Marasek, 2014). The BLEU score is calculated using Equation (1).

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N W_n \log P_n \right) \quad (1)$$

where BP represents the Brevity Penalty applied to outputs that are excessively short; N denotes the maximum n-gram length considered (typically 4); W_n is the weight assigned to each n-gram precision; and P_n denotes the modified n-gram precision.

ROUGE-L evaluates similarity based on the Longest Common Subsequence (LCS) between model-generated answers and reference answers, emphasizing sentence continuity and contextual consistency (Citarella et al., 2025; Syahidi et al., 2025; Onan and Alhumyani, 2024). Equation (2) is used to compute ROUGE-L.

$$ROUGE-L = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (2)$$

where β is a parameter used to balance recall and precision (typically set to 1), R_{LCS} represents the recall score based on the longest common subsequence, and P_{LCS} denotes the corresponding precision score.

METEOR evaluates semantic quality by considering stemming, synonym matching, and penalties associated with incorrect word order, thereby providing a more sensitive assessment of semantic similarity and linguistic flexibility. In addition, this metric combines precision and recall with different weights and incorporates a fragmentation penalty (Mahyoub et al., 2025; Onan and Alhumyani, 2024). METEOR is calculated using Equation (3).

$$METEOR = (1 - Penalty) \cdot F_{mean} \quad (3)$$

where F_{mean} is defined by Equation (4).

$$F_{mean} = \frac{10PR}{P + 9R} \quad (4)$$

where P denotes unigram-based precision, R denotes unigram-based recall, and $Penalty$ represents the fragmentation penalty applied when matching words appear in an order different from that of the reference. The metric F_{mean} corresponds to a weighted harmonic mean that assigns nine times greater importance to recall than precision.

Second, human evaluation is conducted by three independent evaluators who assess the generated answers based on content accuracy with respect to Thai immigration policies, language clarity, and user satisfaction. This evaluation is performed after the automatic testing phase and during selected stages of the system development cycle, such as major version releases or model updates, to ensure consistent response quality. A five-point rating scale is employed, where 1 indicates very poor performance and 5 indicates excellent performance.

By combining automatic and human evaluation methods, the system can assess not only the technical correctness of generated responses but also their practical relevance and comprehensibility for end users (Liu et al., 2023; Sottana et al., 2023).

3.3 Example Prompt and Output Results

To evaluate the effectiveness of the GPT-4-based question-answering system in the context of Thai visa information services, the system was tested using a variety of questions representing the 12 major visa categories. This evaluation employed prompt engineering

with a few-shot learning approach, as described in Section 3.2.3, in which the model was provided with several example question-answer pairs to guide response generation. This approach enables GPT-4 to identify question patterns and understand the specific context of complex visa regulations.

Table 6 presents representative examples of the prompts used, together with two summary responses generated by GPT-4 for each visa category. The results indicate that, when provided with well-designed and normalized prompts, the model can generate informative responses that align with official regulations and address the information needs of visa applicants.

4. Experimental Results

4.1 Evaluation Results per Visa Category

To assess the ability of the GPT-4-based system to provide accurate and relevant responses to visa-related questions, an automatic evaluation was conducted using three widely adopted Natural Language Processing (NLP) metrics, namely BLEU, ROUGE-L, and METEOR. This evaluation was performed to measure the model's consistency in understanding the complex context of immigration policies and generating responses that are both factually accurate and clearly communicated. Given the diversity of Thai immigration services and the distinct regulations associated with different visa categories, evaluating performance at the visa-category level is particularly important. Table 7 presents the average BLEU, ROUGE-L, and METEOR scores calculated from two responses generated by GPT-4 for each question across the 12 major visa categories. These results provide a comprehensive overview of the model's performance and highlight the visa categories for which the system generates the most accurate responses.

Table 7. Average automatic evaluation scores per visa category.

Visa category	BLEU	ROUGE-L	METEOR
Tourist (TR)	0.82	0.86	0.79
Business (Non-Immigrant B)	0.81	0.84	0.78
Education (Non-Immigrant ED)	0.75	0.80	0.73
ED Plus	0.73	0.78	0.72
Media (Non-Immigrant M)	0.69	0.74	0.68
Official (Non-Immigrant F)	0.70	0.75	0.69
Dependent/Spouse (Non-Immigrant O)	0.76	0.79	0.74
SMART	0.74	0.78	0.72
Transit	0.71	0.76	0.69
Long-Term Resident (LTR)	0.78	0.81	0.76
Retirement (Non-Immigrant O-A)	0.77	0.80	0.75
Long-Stay Retirement (Non-Immigrant O-X)	0.72	0.76	0.70

The results presented in Table 7 indicate that the Tourist (TR) category achieved the highest performance, with BLEU, ROUGE-L, and METEOR scores of 0.82, 0.86, and 0.79, respectively. These values represent the highest scores among all evaluated visa categories. This result suggests that visa categories supported by extensive and publicly available documentation allow the model to generate responses that closely align with the reference answers. Other categories, including Business (Non-Immigrant B) and Long-Term Resident (LTR), also demonstrated strong performance, achieving BLEU scores of 0.81 and 0.78, respectively. These findings indicate that the system can provide clear and accurate responses for visa categories with well-defined policy requirements.

To provide an overall assessment of system performance, the average values of the three automatic evaluation metrics were calculated across all visa categories. These aggregate scores offer a general indication of the quality and consistency of GPT-4-generated responses in different visa-related contexts. The results show that the system achieved an average BLEU score of 0.75, an average ROUGE-L score of 0.79, and an average METEOR score of 0.73. These results demonstrate that the proposed system maintains a consistent level of performance across a broad range of Thai immigration visa categories.

4.2 Human Evaluation

In addition to the automated evaluation, a manual evaluation was conducted by three independent evaluators with backgrounds in Thai immigration policy and computational linguistics. The purpose of this evaluation was to obtain a broader perspective on the quality of responses generated by the GPT-4-based system.

4.2.1 Evaluation Procedure

The manual evaluation procedure involved assessing 100 responses generated by GPT-4. These responses were evaluated using a five-point Likert scale, where 1 indicated very poor performance and 5 indicated very good performance. The evaluation was based on three dimensions: Accuracy, which assessed whether the response was consistent with official Thai immigration regulations (Q: Does the answer align with official Thai immigration regulations?); Language Clarity, which evaluated the clarity and comprehensibility of the language used (Q: Is the language used clear and easy to understand?); and User Satisfaction, which measured the extent to which the response was perceived as helpful and satisfactory (Q: Does the answer help or satisfy the user's inquiry?).

4.2.2 Evaluation Results

The manual evaluation results demonstrate strong performance of the GPT-4-based question-answering (QA)

system across all evaluation dimensions, as summarized in Fig. 3.

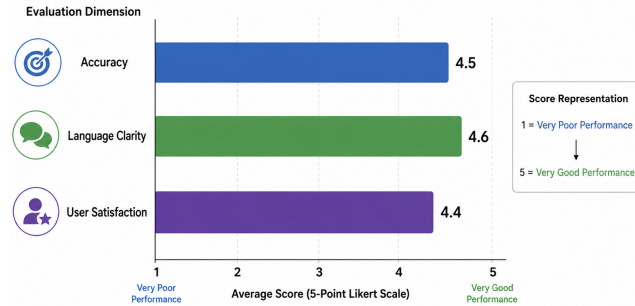


Figure 3. Manual evaluation results by independent evaluators across three dimensions.

The high score obtained for Accuracy (4.5) indicates that the combination of few-shot prompting and rule-based post-processing contributed to generating responses that were largely consistent with official immigration policies. Given the policy-oriented nature of the domain, this result suggests that the system was able to adhere to factual constraints associated with the 12 visa categories included in this study. The findings provide evidence that the combination of system prompts and rule-based validation mechanisms helped reduce the occurrence of factually inconsistent responses.

The highest score was observed for Language Clarity (4.6), indicating that the system effectively communicated complex immigration regulations in a manner that was easy to understand. This result suggests that GPT-4 can transform policy-related information into user-friendly responses while maintaining the intended meaning of the original regulations. The findings also support the objective of reducing communication barriers associated with immigration procedures and policy interpretation.

The User Satisfaction score of 4.4 further reflects the positive perception of the generated responses. This result suggests that users considered the responses helpful, relevant, and sufficiently informative for addressing their inquiries. The combination of factual consistency and clear language likely contributed to the favorable satisfaction ratings observed during the evaluation.

Taken together, the consistently high scores across Accuracy, Language Clarity, and User Satisfaction indicate that the proposed system can generate responses that are generally accurate, understandable, and useful within the context of Thai immigration services. These findings provide additional evidence supporting the suitability of GPT-4 for policy-oriented question-answering applications and highlight its potential for assisting users in navigating complex immigration regulations.

4.3 Correlation Analysis

To examine the relationship between automatic evaluation metrics and human judgments, Pearson correlation coefficients were calculated, with the results presented in Table 8.

Table 8. Correlation between automatic evaluation metrics and human judgments.

Metric Pair	Pearson Correlation (r)
BLEU vs Accuracy (Human)	0.78
ROUGE-L vs Language Clarity (Human)	0.72
METEOR vs User Satisfaction (Human)	0.75

The correlation analysis revealed strong positive relationships between the automatic evaluation metrics and the corresponding human evaluation dimensions. All correlation coefficients exceeded the threshold of $r \geq 0.70$, indicating substantial agreement between automated and manual assessments. Among the evaluated metrics, BLEU demonstrated the strongest correlation with Accuracy, suggesting that n-gram overlap is closely associated with factual consistency in responses related to immigration policies. ROUGE-L showed the strongest relationship with Language Clarity, indicating that structural similarity between generated and reference answers is related to human perceptions of comprehensibility. In addition, METEOR exhibited a strong correlation with User Satisfaction, reflecting the importance of semantic similarity in shaping user perceptions of response quality.

These findings suggest that the selected automatic evaluation metrics provide meaningful indicators of response quality within the immigration policy domain. The observed correlations support the use of BLEU, ROUGE-L, and METEOR as complementary evaluation measures for monitoring the performance of LLM-based question-answering systems. Furthermore, the results indicate that automatic evaluation can provide a practical and efficient assessment framework while reducing the need for extensive manual evaluation in future system development and maintenance.

5. Discussion

The experimental results presented in Section 4 provide strong evidence regarding the performance and reliability of the GPT-4-based system in answering questions related to Thai immigration visa services. These findings, derived from both automatic and manual evaluations, demonstrate that the designed system holds significant potential in generating accurate, clear, and satisfactory responses for users across various visa categories.

5.1 Strengths Based on Automated Evaluation

The automated evaluation results demonstrate that the proposed GPT-4-based question-answering system can maintain a consistently high level of performance across multiple visa categories. The overall BLEU, ROUGE-L, and METEOR scores indicate that the generated responses exhibit substantial lexical similarity, structural consistency, and semantic alignment with the reference answers. These findings suggest that GPT-4 can effectively interpret immigration-related questions and generate responses that are consistent with the intended policy information.

Among the evaluated visa categories, the Tourist (TR) category achieved the strongest performance. This outcome may be attributed to the availability of extensive and standardized public documentation, including visa requirements, eligibility criteria, and application procedures. The abundance of publicly accessible information likely provides more consistent linguistic patterns and policy descriptions, allowing the model to generate responses that closely align with the reference answers. Similar observations have been reported in previous studies, where LLM performance tended to improve when tasks were associated with well-documented and frequently occurring information domains.

Strong performance was also observed for the Business (Non-Immigrant B) and Long-Term Resident (LTR) categories. These results suggest that the proposed prompt engineering strategy can effectively guide GPT-4 in handling visa categories that involve multiple eligibility conditions, supporting documents, and procedural requirements. The ability to maintain relatively high scores across these categories indicates that the model can preserve contextual consistency while generating detailed policy-oriented responses.

In contrast, comparatively lower scores were observed for categories such as Media (Non-Immigrant M) and Long-Stay Retirement (Non-Immigrant O-X). One possible explanation is that these visa categories involve more specialized regulations and less frequently referenced policy information. Consequently, the available information may exhibit greater variability in terminology and procedural details, increasing the difficulty of generating responses that closely match the reference answers. This finding highlights the importance of domain-specific knowledge representation when deploying LLM-based systems in specialized administrative contexts.

The average BLEU, ROUGE-L, and METEOR scores further indicate that the system performs consistently across a diverse range of visa-related scenarios. The relatively strong ROUGE-L score suggests that GPT-4 can maintain coherent information flow and response structure, which is particularly important when communicating procedural information to visa applicants. Likewise, the METEOR score reflects

the model's ability to preserve semantic meaning even when alternative wording is used. Taken together, these findings suggest that the proposed system can generate responses that are not only factually relevant but also linguistically coherent and semantically appropriate for immigration-related question-answering tasks.

5.2 Findings from Manual Evaluation

The manual evaluation results provide a complementary perspective to the automated metrics by assessing response quality from a human-centered viewpoint. While automatic evaluation focuses primarily on lexical and semantic similarity, manual assessment captures aspects related to factual appropriateness, comprehensibility, and perceived usefulness. The consistently high scores obtained across all evaluation dimensions suggest that the proposed system can generate responses that align well with both policy requirements and user expectations.

The strong performance observed in the Accuracy dimension indicates that GPT-4 was generally able to produce responses that were consistent with official Thai immigration regulations. This finding is particularly important in the context of visa-related services, where inaccurate information may lead to misunderstandings or inappropriate decision-making by applicants. The result suggests that the combination of domain-specific prompt design, few-shot examples, and rule-based validation mechanisms contributed to maintaining factual consistency in the generated responses.

Language Clarity received the highest evaluation score among all assessed dimensions. This result suggests that the system was effective in transforming policy-oriented information into responses that were easy to understand and communicate. Immigration regulations often contain technical terminology and procedural requirements that may be difficult for applicants to interpret. The ability of GPT-4 to present such information in a clear and structured manner may therefore improve accessibility and facilitate information retrieval for users seeking visa-related guidance.

The User Satisfaction results further support the practical value of the proposed system. The relatively high satisfaction score indicates that users generally perceived the generated responses as useful and relevant to their information needs. The findings suggest that response quality is influenced not only by factual correctness but also by the clarity, completeness, and overall readability of the generated content. Consequently, the combination of accurate information and understandable language appears to contribute positively to the overall user experience.

Although the evaluation results were highly encouraging, qualitative feedback provided by the evaluators also highlighted opportunities for improvement. In some cases, responses lacked specific procedural details, such as processing timelines or additional administra-

tive requirements. These observations suggest that future enhancements could focus on incorporating more detailed procedural information and expanding domain-specific knowledge resources to further improve response completeness.

Overall, the manual evaluation findings indicate that the proposed GPT-4-based system achieves a balance between factual consistency and communicative effectiveness. The alignment between the manual and automated evaluation results provides additional evidence that the system can generate responses that are not only relevant to immigration policies but also understandable and useful from the perspective of end users.

5.3 Correlation Analysis Between Automatic and Manual Evaluation

The correlation analysis provides additional insights into the relationship between automatic evaluation metrics and human perceptions of response quality. While automated metrics offer a scalable and objective assessment framework, human evaluation remains essential for capturing factors such as factual appropriateness, readability, and user satisfaction. Therefore, examining the degree of agreement between these two evaluation approaches is important for determining whether automatic metrics can serve as meaningful indicators of real-world system performance.

The results presented in Table 8 show strong positive correlations between all metric-assessment pairs, indicating substantial alignment between automatic and manual evaluations. In particular, the strong relationship between BLEU and Accuracy suggests that lexical overlap remains an important indicator of factual consistency within the immigration policy domain. Unlike open-ended conversational tasks, immigration-related responses often rely on standardized terminology, official policy statements, and procedural descriptions. As a result, higher lexical similarity between generated and reference answers is likely to reflect greater adherence to policy-related information.

A similarly strong correlation was observed between ROUGE-L and Language Clarity. This finding suggests that response structure and information organization play an important role in shaping human perceptions of comprehensibility. Since ROUGE-L emphasizes the longest common subsequence between generated and reference responses, the metric may capture the extent to which key information is presented in a logical and coherent order. This characteristic is particularly relevant in public-service applications, where procedural information must be communicated clearly to avoid misunderstanding.

The observed relationship between METEOR and User Satisfaction further highlights the importance of semantic similarity in evaluating question-answering systems. Unlike BLEU and ROUGE-L, METEOR incorporates synonym matching and partial lexical varia-

tions, making it more sensitive to differences in wording while preserving meaning. The strong correlation observed in this study suggests that users tend to respond positively to answers that convey the intended information effectively, even when the wording differs from the reference responses.

Taken together, these findings indicate that BLEU, ROUGE-L, and METEOR capture complementary aspects of response quality and exhibit meaningful alignment with human judgments. The results support the use of automatic evaluation metrics as practical indicators for monitoring the performance of LLM-based question-answering systems in policy-oriented domains. Although human evaluation remains indispensable for assessing nuanced aspects of response quality, the strong correlations observed in this study suggest that automatic metrics can provide an efficient and scalable mechanism for routine system evaluation and performance tracking.

Furthermore, the observed agreement between automated and human assessments provides additional confidence in the validity of the evaluation framework adopted in this study. This finding is particularly important for future research involving large-scale policy-oriented question-answering systems, where extensive human evaluation may be costly, time-consuming, and difficult to maintain consistently.

5.4 System Advantages

Several advantages of the GPT-4-based QA system were identified in this study. First, the system demonstrated adaptability across visa categories by successfully answering questions related to 12 different types of Thai visas. Second, it generated responses in natural and communicative English, making the information more accessible and easier to understand for non-expert users. Third, the system exhibited modularity through prompt engineering techniques, allowing contextual guidance toward specific visa types using a few-shot learning approach.

These findings suggest that the GPT-4-based system can provide accurate, well-structured, and user-centered responses for visa-related services. The alignment observed between automated and manual evaluation results further indicates the potential for efficient and scalable validation of public policy question-answering systems. Overall, the findings suggest that GPT-4 can flexibly interpret complex policy-related inquiries without requiring explicit fine-tuning, making it a potentially valuable tool for supporting communication in public services.

5.5 Limitations and Challenges

Although the results demonstrate encouraging performance, several limitations of the proposed system were identified. First, the system relies heavily on the format

and quality of the few-shot prompts used, which directly influence the accuracy and relevance of the generated responses. Second, the system may encounter difficulties when answering highly specific questions, such as those involving policy exceptions or rapidly changing regulations, and may occasionally generate generalized responses.

Third, GPT-4 does not have real-time access to official immigration policy databases, making it unable to incorporate newly issued regulations or policy updates unless such information is manually included in the prompt. In addition, generated responses may not always contain important advisory statements, such as recommending users verify information with immigration authorities, which may create risks if the information is used without further confirmation.

The variation in performance observed across visa categories also highlights the importance of domain-specific adaptation, particularly for visa types associated with limited or less frequently referenced information. Future improvements may include fine-tuning with localized datasets, integrating Retrieval-Augmented Generation (RAG) techniques for real-time policy updates, and providing multilingual support to improve accessibility. Qualitative error analysis may also help identify failure patterns and support continuous system improvement. Furthermore, incorporating user-specific contextual information, such as nationality, travel history, or purpose of visit, may improve response relevance and user satisfaction while maintaining consistency with evolving immigration policies.

5.6 Implications for Public Sector Services

This study provides insights into the potential application of large language models in public-sector services, particularly in immigration visa information services. The use of automated question-answering systems may enhance service efficiency by reducing the need for repetitive information requests handled by immigration officers. Such systems may also improve information accessibility by assisting international visa applicants in understanding visa requirements without requiring in-person consultations.

In addition, this study serves as a case example of applying GPT-based systems in regulatory domains that require a high degree of factual consistency. However, practical deployment should be accompanied by appropriate governance mechanisms, including human review, regular policy updates, and transparent auditing processes to help identify and address potential limitations and biases.

5.7 Policy Significance and Real-World Impact

The findings of this study highlight the potential role of GPT-4-based question-answering systems in supporting

public-service innovation. By providing responses that were generally evaluated as accurate, clear, and useful, the proposed system has the potential to address some limitations associated with visa information services, including inconsistent information delivery and language-related communication barriers.

From an administrative perspective, the system may assist in handling routine and complex information requests, thereby supporting immigration officers in providing information services. In terms of service consistency, the use of standardized prompts and policy-oriented examples may contribute to more uniform responses across different user inquiries, which is important for maintaining public confidence in information services.

Overall, the results suggest that GPT-4-based question-answering systems may serve as useful tools for supporting immigration information services when combined with carefully designed prompts and validation mechanisms. While further validation is required before large-scale deployment, the findings indicate the potential of such systems to improve information accessibility and support communication between government agencies and visa applicants.

6. Conclusion and Future Works

This study successfully evaluated the application of GPT-4 as an automated QA system for Thai immigration visa services across 12 main visa categories. Using a few-shot prompt engineering approach and a simulated dataset of 1,600 QA pairs based on official policy guidelines, the system demonstrated high performance, achieving BLEU, ROUGE-L, and METEOR scores of up to 0.82, 0.86, and 0.79, respectively, for the Tourist (TR) category, as well as average human evaluation scores of 4.5-4.6 for accuracy and clarity. These results indicate the strong potential of GPT-4 to enhance the responsiveness, efficiency, and accessibility of digital immigration services.

Nevertheless, several limitations were identified. First, the evaluation was conducted using a simulated dataset and did not involve real-world user interactions. Second, the system does not have real-time connectivity with official immigration information systems, which may limit its ability to reflect newly issued policy updates. Third, generated responses may not always include important advisory statements or official disclaimers. Finally, the effectiveness of the system remains dependent on the design and maintenance of prompt templates, which may limit scalability and automated adaptation.

Beyond its technical contributions, this study demonstrates the potential application of large language models in public-service information systems. The results indicate that GPT-4 can support the delivery of consistent and understandable immigration informa-

tion, thereby assisting users in navigating visa-related requirements and procedures.

Future research may focus on multilingual support to improve accessibility, integration with official immigration APIs to provide real-time policy updates, domain-specific fine-tuning to enhance response precision, and the deployment of a prototype chatbot in operational environments with continuous user evaluation. Additional research may also explore governance and auditing mechanisms to improve transparency, accountability, and responsible AI deployment, as well as the use of Elasticsearch for centralized log management and real-time monitoring of system performance.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Bahrami, M., Mansoorizadeh, M., and Khotanlou, H. (2023). Few-shot learning with prompting methods. In *Proceedings of the 6th International Conference on Pattern Recognition and Image Analysis*, pages 1–5, Qom, Iran. DOI: 10.1109/IPRIA59240.2023.10147172.
- Barber, S. and Sciortino, R. (2024). Thailand migration report 2024. Technical report, United Nations Network on Migration in Thailand, Bangkok, Thailand. Accessed: 2025-03-18, <https://thailand.un.org/en/285915-thailand-migration-report-2024>.
- Barman, K. G., Lohse, S., and de Regt, H. W. (2025). Reinforcement learning from human feedback in LLMs: Whose culture, whose values, whose perspectives? *Philosophy & Technology*, 38(35):1–26. DOI: 10.1007/s13347-025-00861-0.
- Betrand, C. U., Ekwealor, O. U., and Onyema, C. J. (2023). Artificial intelligence chatbot advisory system. *International Journal of Intelligent Information Systems*, 12(1):1–9. DOI: 10.11648/j.ijiis.20231201.11.
- Bollmann, M., Petran, F., and Dipper, S. (2011). Rule-based normalization of historical texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42, Hissar, Bulgaria. <https://aclanthology.org/W11-4106/>.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901, Vancouver, Canada. <https://dl.acm.org/doi/abs/10.5555/3495724.3495883>.
- Chowdhury, M. S. R., Khan, N. H., Singha, D., Ahmed, T., Chowdhury, F. H., and Sarkar, M. I. (2024). Streamlining global visa processing and document verification: A self-sovereign identity system for secure and efficient visa processing using hyperledger indy. In *Proceedings of the 2nd International Conference on Information and Communication Technology*, pages 120–124, Dhaka, Bangladesh. DOI: 10.1109/I-CICT64387.2024.10839648.
- Citarella, A. A., Barbella, M., Ciobanu, M. G., Di Marco, F., De Biasi, L., and Tortora, G. (2025). Assessing the effectiveness of ROUGE as unbiased metric in extractive vs. abstractive summarization techniques. *Journal of Computational Science*, 87:102571. DOI: 10.1016/j.jocs.2025.102571.
- Dai, W., Cheng, Y., Aldino, A. A., Tsai, Y.-S., Gašević, D., and Chen, G. (2025). Evaluating the capability of large language models in characterising relational feedback: A comparative analysis of prompting strategies. *Computers and Education: Artificial Intelligence*, 8:100427. DOI: 10.1016/j.caeai.2025.100427.
- Dongbo, M., Miniaoui, S., Fen, L., Althubiti, S. A., and Alsenani, T. R. (2023). Intelligent chatbot interaction system capable for sentimental analysis using hybrid machine learning algorithms. *Information Processing & Management*, 60(5):103440. DOI: 10.1016/j.ipm.2023.103440.
- Gadag, A. and Sagar, B. M. (2016). Paraphrase generator using dictionary lookup for Kannada language. In *Proceedings of the 2nd International Conference on Next Generation Computing Technologies*, pages 164–168, Dehradun, India. DOI: 10.1109/NGCT.2016.7877408.
- Gao, S., Gao, L., Li, Q., and Xu, J. (2023). Application of large language model in intelligent Q&A of digital government. In *Proceedings of the 2nd International Conference on Networks, Communications and Information Technology*, pages 24–27, New York, USA. DOI: 10.1145/3605801.3605806.
- Gupta, A., Rastogi, A., Malhotra, H., and Rangarajan, K. (2024). Comparative evaluation of large language models for translating radiology reports into Hindi. *Indian Journal of Radiology and Imaging*, 35(1):88–96. DOI: 10.1055/s-0044-1789618.

- Jiang, S., Xie, X., Tang, R., Wang, X., Sun, K., Li, G., Xu, Z., Xue, P., Li, Z., and Fu, X. (2025). ARGUS: Retrieval-augmented QA system for government services. *Electronics*, 14(12):2445. DOI: 10.3390/electronics14122445.
- Konduru, K. K., Natalia, F., Sudirman, S., and Al-Jumeily, D. (2024). Evaluating few-shot prompting approach using GPT4 in comparison to BERT-variant language models in biomedical named entity recognition. In *Proceedings of the 17th International Conference on Development in eSystem Engineering*, pages 340–345, Khorfakkan, United Arab Emirates. DOI: 10.1109/DeSE63988.2024.10911889.
- Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. (2024). Large language models in law: A survey. *AI Open*, 5:181–196. DOI: 10.1016/j.aiopen.2024.09.002.
- Liu, B., Zhu, Z., Ai, Q., Liu, Y., and Wu, Y. (2024a). LeDQA: A chinese legal case document-based question answering dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5385–5389, Boise, United States. DOI: 10.1145/3627673.3679154.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-EVAL: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. DOI: 10.18653/v1/2023.emnlp-main.153.
- Liu, Y., Meng, R., Bhat, M. M., Joty, S., Xiong, C., Zhou, Y., and Yavuz, S. (2024b). Modeling uncertainty and using post-fusion as fallback improves retrieval augmented generation with LLMs. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 69–82. ACM. DOI: 10.18653/v1/2024.knowllm-1.7.
- Lu, Y., Xie, H., Zhang, J., Jin, Y., Feng, Y., and Gong, Y. (2023). Fractal theory based stratified sampling for quality assessment of remote-sensing-derived geospatial data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:7100–7111. DOI: 10.1109/JSTARS.2023.3287347.
- Mahbub, T., Dghaym, D., Shankarnarayanan, A., Syed, T., Shapsough, S., and Zualkernan, I. (2024). Can GPT-4 aid in detecting ambiguities, inconsistencies, and incompleteness in requirements analysis? A comprehensive case study. *IEEE Access*, 12:171972–171992. DOI: 10.1109/ACCESS.2024.3464242.
- Mahyoub, M., Wang, Y., and Khasawneh, M. T. (2025). GPT-4o in radiology: In-context learning based automatic generation of radiology impressions. *Natural Language Processing Journal*, 11:100145. DOI: 10.1016/j.nlp.2025.100145.
- Mai, L. and Carson-Berndsen, J. (2024). Enhancing conversation smoothness in language learning chatbots: An evaluation of GPT4 for ASR error correction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 11001–11005, Seoul, South Korea. DOI: 10.1109/ICASSP48485.2024.10447641.
- Naser-Karajah, E. and Arman, N. (2024). Toward automated Arabic synonyms extraction using Arabic lexical substitution. *IEEE Access*, 12:174455–174463. DOI: 10.1109/ACCESS.2024.3485502.
- Olivato, M., Putelli, L., Arici, N., Gerevini, A. E., Lavelli, A., and Serina, I. (2024). Language models for hierarchical classification of radiology reports with attention mechanisms, BERT, and GPT-4. *IEEE Access*, 12:69710–69727. DOI: 10.1109/ACCESS.2024.3402066.
- Onan, A. and Alhumyani, H. (2024). Knowledge-enhanced transformer graph summarization (KETGS): Integrating entity and discourse relations for advanced extractive text summarization. *Mathematics*, 12(23):3638. DOI: 10.3390/math12233638.
- OpenAI (2023). GPT-4 technical report. Technical report, OpenAI. DOI: 10.48550/arXiv.2303.08774.
- Raj, P. B. A. S., Piri, J., Eluri, S. B., and S, S. R. (2023). Work visa analysis using machine learning techniques. In *Proceedings of the 3rd International Conference on Artificial Intelligence and Smart Energy*, pages 616–621, Coimbatore, India. DOI: 10.1109/ICAIS56108.2023.10073837.
- Roller, S. and Erk, K. (2016). PIC a different word: A simple model for lexical substitution in context. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1121–1126, San Diego, California. DOI: 10.18653/v1/N16-1131.
- Shahandashti, K. K., Sivakumar, M., Mohajer, M. M., Belle, A. B., Wang, S., and Lethbridge, T. (2024). Assessing the impact of GPT-4 turbo in generating defeaters for assurance cases. In *Proceedings of the IEEE/ACM 1st International Conference on AI Foundation Models and Software Engineering*, pages 52–56, Lisbon, Portugal. DOI: 10.1145/3650105.3652291.
- Sharma, A., Sharma, A., and Bhattacharjee, V. (2023). Breast cancer prediction: Impact of stratified sampling approach on classifier accuracy. In *Proceedings of the International Conference on Artificial Intelligence and Smart Communication*, pages 1076–1078, Greater Noida, India. DOI: 10.1109/AISC56616.2023.10084954.

- Shultz, T. R., Wise, J. M., and Nobandegani, A. S. (2025). Text understanding in GPT-4 versus humans. *Royal Society Open Science*, 12:241313. DOI: 10.1098/rsos.241313.
- Sottana, A., Liang, B., Zou, K., and Yuan, Z. (2023). Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. DOI: 10.18653/v1/2023.emnlp-main.543.
- Sun, J., Zhang, Z., Wang, X., Ji, X., and Zhang, Y. (2024). Fallback prompting guides large language models for accurate responses in complex reasoning. *Journal of Networking and Network Applications*, 4(3):109–117. <https://ieeescience.org/jpapers/185>.
- Swain, D., Chakraborty, K., Dombé, A., Ashture, A., and Valakunde, N. (2018). Prediction of H1B visa using machine learning algorithms. In *Proceedings of the International Conference on Advanced Computation and Telecommunication*, pages 1–7, Bhopal, India. DOI: 10.1109/ICACAT.2018.8933628.
- Syahidi, A. A. and Kiyokawa, K. (2025). Automatic text generation in Banjar language using GPT-4 for low-resource language preservation. In *Proceedings of the IEEE 7th Symposium on Computers and Informatics*, pages 238–243, Kuala Lumpur, Malaysia. DOI: 10.1109/ISCI65687.2025.11167521.
- Syahidi, A. A., Kiyokawa, K., and Nuchitprasitchai, S. (2025). A fine-tuned GPT-4-based question answering system for e-government services using a custom-built dataset. In *Proceedings of the IEEE 7th Symposium on Computers and Informatics*, pages 232–237, Kuala Lumpur, Malaysia. DOI: 10.1109/ISCI65687.2025.11166787.
- Uruj, S., Goswami, R., Shetty, S. D., Venkatesan, K., and Ramanujam, K. (2025). Comparative analysis of GPT-4 and LLaMA 3.2 integration with speech processing models for enhancing human-robot interaction and motion control in real-world applications. *IEEE Access*, 13:127170–127182. DOI: 10.1109/ACCESS.2025.3590592.
- Vladika, J., Meisenbacher, S., and Matthes, F. (2025). Lexical substitution is not synonym substitution: On the importance of producing contextually relevant word substitutes. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, pages 1–8, Porto, Portugal. DOI: 10.48550/arXiv.2502.04173.
- Wolk, K. and Marasek, K. (2014). Enhanced bilingual evaluation understudy. *Lecture Notes on Information Theory*, 2(2):191–197. DOI: 10.48550/arXiv.1509.09088.
- Yanbo, H. and Kongjit, C. (2020). Learning of root cause failure analysis through long-term visa application for Chinese in Thailand. In *Proceedings of the Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering*, pages 324–329, Pattaya, Thailand. DOI: 10.1109/ECTIDAMT-NCON48261.2020.9090712.
- Yekta, M. M. J. (2024). The general intelligence of GPT-4, its knowledge diffusive and societal influences, and its governance. *Meta-Radiology*, 2:100078. DOI: 10.1016/j.metrad.2024.100078.

Table 6. Sample prompts and summary responses generated by GPT-4 for 12 Thai visa categories.

Visa Category	Sample Question	Summary Response
Tourist (TR)	Q1: How long can I stay in Thailand with a tourist visa? Q2: Can I apply for a tourist visa online?	A1: "Tourist visas allow stays up to 60 days, extendable once for 30 days." A2: "Yes, e-visa application is available for eligible countries via Thai e-Visa portal."
Business (Non-Immigrant B)	Q1: What documents are required for a Thai business visa? Q2: Can I change from tourist visa to business visa in Thailand?	A1: "Invitation letter, passport, company registration, and financial documents are needed." A2: "Generally, this is not allowed. You must apply for a Non-Immigrant B from outside Thailand, through a Thai Embassy or Consulate. Status changes from a Tourist (TR) within the country are only permitted under limited emergency circumstances."
Education (Non-Immigrant ED)	Q1: Can I get an ED visa for studying Thai language? Q2: What is the minimum study duration for ED visa eligibility?	A1: "Yes, ED visas are issued for accredited language schools approved by the Thai Ministry of Education." A2: "At least 4 hours per week of instruction, typically in a full-semester program."
ED Plus	Q1: Can I extend my ED Plus visa? Q2: Does ED Plus visa allow part-time work?	A1: "Yes, based on continued enrollment and valid documentation from the institution." A2: "Yes, students may work part-time up to 20 hours/week with proper permission."
Media (Non-Immigrant M)	Q1: What documents are needed for a media visa? Q2: How long is a media visa valid for?	A1: "Press card, employer letter, passport, and background check are required." A2: "Usually valid for 90 days, extendable depending on assignment and approval."
Official (Non-Immigrant F)	Q1: Who qualifies for Non-Immigrant F visa? Q2: Is diplomatic note required for F visa?	A1: "Foreign government officials, embassy staff, and international agency personnel." A2: "Yes, official correspondence from the applicant's government or organization is mandatory."
Dependent/Spouse (Non-Immigrant O)	Q1: Can my spouse stay with me on an O visa? Q2: What documents are required for dependent visa application?	A1: "Yes, with marriage certificate and financial proof of support." A2: "Marriage/birth certificates, sponsor's visa copy, and proof of relationship."
SMART	Q1: What is the benefit of the SMART visa for tech entrepreneurs? Q2: How can I qualify for SMART S visa as a startup founder?	A1: "Up to 4-year stay, no work permit required, and family members can also work." A2: "Must be endorsed by relevant Thai agency and have registered startup or incubation project."
Transit	Q1: Do I need a visa for a 12-hour layover in Thailand? Q2: How long is a transit visa valid?	A1: "Not if you stay in the airport's transit zone and hold onward tickets." A2: "Valid for 30 days, with a permitted stay of up to 30 days for transit purposes."
Long-Term Resident (LTR)	Q1: What are the qualifications for the LTR visa? Q2: Can my family join me under LTR visa?	A1: "Applicable to high-income individuals, retirees, or skilled professionals meeting income and asset thresholds." A2: "Yes, spouse and children (under 20) can be included as dependents."
Retirement (Non-Immigrant O-A)	Q1: What is the age requirement for O-A visa? Q2: Can I re-enter Thailand multiple times with O-A visa?	A1: "Applicants must be 50 years or older with specific financial qualifications." A2: "Yes, if a multiple-entry visa is issued or re-entry permits are obtained."
Long-Stay Retirement (Non-Immigrant O-X)	Q1: Can I work on an O-X visa? Q2: What financial documents are needed for O-X visa?	A1: "No, O-X visa holders are not permitted to work in Thailand." A2: "Bank statement showing 3 million THB in Thai bank or 1.8 million with annual income."