Visa Journey Agent: An AI-Driven Compliance Assistant for F-1 Visa Holders

Team Members: Abhay Prabhakar, Harshal Kamble, Pavithra Kannan, Jenny Nguyen, Jared Jones
Course: MSA 8770 – Text Analytics
Instructor: Dr. Soleymani
Semester: Fall 2025

Introduction

F-1 visa holders navigating U.S. immigration rules face complex, frequently changing policies. Requirements for SEVIS check-ins, CPT eligibility, OPT filing windows, job-loss notifications, and program milestones vary across institutions and are often buried within lengthy regulatory documents. Students who miss a requirement risk violating visa status, losing work authorization, or delaying graduation plans.

The **VisaJourney Agent** is a two-agent AI system designed to automate regulatory understanding and provide personalized compliance roadmaps.

- **Agent 1** extracts structured policy rules from unstructured text (e.g., USCIS/DHS guidelines) and provides summaries of legal guidelines.
- **Agent 2** transforms these rules into customized timelines and checklists based on a student's profile.

Together, they form a scalable, intelligent advisory system that reduces manual workload on international student offices and empowers students with transparent, proactive guidance.

Data Collection & Tools

The knowledge base for our agentic system was built by scraping federal and university websites that describe the requirements and procedures for obtaining an F-1 visa. Federal sources provide the official regulatory baseline, while university websites illustrate how individual institutions implement and communicate F-1 policies. This combination reflects both the legal framework and the real-world context in which students navigate the visa process. Our government sources included relevant sections of the USCIS and Department of State websites covering U.S. visas in general and the F-1 visa specifically. For university-level guidance, we used the F-1 visa resources provided by Georgia State University's International Student and Scholar Services (ISSS) office.

Data Sources

Primary Sources (USCIS Policy Manual):

- Volume 2 Part F: F-1 Student regulations (9 chapters)
- Volume 2 Part L: H-1B specialty occupations (6 chapters)
- Volume 10 Part A: Employment authorization (6 chapters)
- Volume 6 Part E: Employment-based immigrants (5 chapters)
- Volume 7 Parts A, B, E, O: Adjustment of status procedures

Supplementary Sources:

- GSU ISSS: University-specific F-1 guidance (5 pages)
- Travel.gov: International travel requirements (415 chunks)
- USCIS Newsroom: Policy alerts filtered for F-1/H-1B relevance (2024-2025)

Total Dataset: 2,828 USCIS/ISSS chunks + 415 travel.gov chunks = **3,243 documents**

Data Collection Pipeline

Web Scraping Strategy:

To build our knowledge base, we used Python scripts to collect and process data. Website HTML was retrieved using the requests library and converted to Markdown with the html_to_markdown package to enable more meaningful chunking. We then manually cleaned the text to remove irrelevant content and convert certain Markdown elements—such as tables—into natural language to produce higher-quality embeddings. Using LangChain's MarkdownTextSplitter, we generated semantically cohesive chunks. Finally, we embedded these chunks with the **all-MiniLM-L6-v2** model from **SentenceTransformers** and stored them in a **ChromaDB** vector database.

```
# Web-scraping
def download_and_clean_volume(volume_name, base_url, chapters_range, ...):
  - BeautifulSoup HTML parsing
  - Content extraction from <div class='usa-prose'>
  - Removal of navigation/headers/footers
  - HTML-to-Markdown conversion
  - Regex cleaning (removes "Skip to content", "Español" menus, etc.)
```

Quality Assurance:

- Manual verification of chapter completeness
- Keyword filtering for H-1B/F-1/OPT relevance in alerts
- Date filtering (2024-2025 only) for recent policy changes

Technical Stack

Component	Technology	Justification
Vector DB	ChromaDB (persistent)	Open-source, efficient for 3K+ docs
Embeddings	sentence-transformers/all-MiniLM-L6-v2	Fast, 384-dim, good for QA tasks
LLM	OpenAI GPT-4o-mini	Cost-effective, reliable policy summarization
Chunking	LangChain RecursiveCharacterTextSplitter	800 chars, 100 overlap for context
UI	Streamlit + ngrok	Rapid prototyping, public URL deployment
Storage	Google Drive (mounted)	Team collaboration, persistent backups

System Architecture

The Visa Journey Agent consists of two autonomous but connected modules.

Agent 1 : Regulation Retrieval + Summarize Engine

Agent 1 serves as the policy intelligence engine of the Visa Journey system. Its primary purpose is to transform unstructured immigration regulations (USCIS, DHS, SEVP, travel.gov) into a structured and searchable knowledge base, and to provide accurate, grounded answers to user policy questions.

To achieve this, Agent 1 performs three key functions:

1. Regulation Extraction and Chunking

Raw immigration policies are long and unorganized, making direct retrieval difficult. Agent 1 preprocesses these documents by splitting them into smaller, semantically coherent text chunks (e.g., CPT eligibility rules, OPT application window, H-1B filing timeline). This creates a clean set of knowledge units that can be retrieved efficiently.

2. Embedding and Vector Indexing (ChromaDB)

Each rule chunk is converted into a high-dimensional embedding using sentence-transformers/all-MiniLM-L6-v2, capturing the semantic meaning of the regulation. These embeddings, along with the original text and metadata, are stored in a vector database (ChromaDB). This allows the system to perform semantic search — retrieving relevant regulations based on meaning rather than keyword matching.

3. Retrieval-Augmented Generation (RAG) Policy Summaries

When the user asks a policy question, Agent 1 queries the vector database to retrieve the most relevant regulation chunks. These chunks are passed to an LLM (GPT-4o-mini), which produces a structured and grounded summary that includes key rules, required actions, and a risk level. Because the summaries are generated from retrieved policy text, the answers remain accurate, explainable, and consistent with official guidelines.

Agent 2 : Checklist Generator

Agent 2 serves as the system's **planning and coordination component**, connecting the policy information extracted by Agent 1. Whereas Agent 1 focuses on retrieving immigration rules and interpreting policy language through a retrieval-augmented question-answering process, Agent 2 translates those rules into practical guidance that reflects the student's academic timeline, milestones, and visa requirements.

At its core, Agent 2 relies on a structured set of date-driven computations. Using the student's program start date and expected graduation, it calculates key regulatory periods such as when CPT becomes available, when the OPT filing window opens, and how long the student has been in valid F-1 status. These calculations are built on **rule-based logic,** ensuring that timeline results remain consistent.

An important responsiblity of Agent 2's roles are **policy integration**. When policy updates or regulatory statements appear in Agent 1's outputs, they are incorporated into Agent 2's checklist so that the student receives action items connected directly to official guidance.

Overall, Agent 2 forms the system's **decision-support layer**. It not only assembles a chronological roadmap for CPT/OPT compliance but also personalizes that roadmap by embedding relevant immigration rules, deadlines, and tasks. This makes Agent 2 essential for moving the system beyond simple question-answering and toward a more adaptive, **student-specific advisory tool**.

Implementation

Technologies Used

Component	**Technology**
Embeddings	all-MiniLM-L6-v2
Vector DB	ChromaDB
Backend Logic	Python + DateUtil
Front-end UI	Streamlit
Deployment	Colab + Ngrok

Key Innovations

- **Two-agent architecture for continuous policy updates**
- **Fully automated CPT/OPT timeline calculation**
- **User-specific compliance checklists**
- **Real-time chat assistant in Streamlit**

UI Interface

The UI is designed to feel natural and conversational while quietly handling complex logic in the background. The chat screen maintains **short-term memory** using Streamlit's session state, allowing the assistant to remember what the user asked earlier in the same conversation and respond with context. Beyond this, we introduced **long-term memory** through a persistent JSON store that saves key user details such as profile information and important notes the user wants the system to remember across sessions.

This means that when a student returns to the app later, the assistant can recall their previous inputs and provide more personalized guidance. The interface also includes a dedicated timeline and checklist page that updates automatically based on both the user profile and the long-term memory store. Combined with a debug panel showing Agent-1's extracted rules, the UI makes the system feel both intelligent and transparent while keeping everything simple for the user.

Results & Evaluation

Evaluation Summary

To assess the accuracy of Agent-1's policy retrieval and summarization, we designed an evaluation framework consisting of four representative immigration questions: H-1B requirements, H-1B filing timeline, specialty occupation criteria, and the effect of full-time CPT on OPT eligibility. For each query, we manually defined a set of *must-have* keywords derived directly from authoritative USCIS guidelines. These keywords represent the essential policy concepts that a correct answer should contain.

Agent-1 retrieves relevant policy documents from the Chroma vector store and generates a structured summary using the LLM. We then compute a simple **recall metric**, defined as the proportion of expected keywords present in the generated answer. This provides a transparent and interpretable measure of how well Agent-1 is grounding its answers in the underlying knowledge base.

The evaluation results are as follows:

- Three of the four questions (**H-1B requirements**, **H-1B filing timeline**, and **CPT → OPT eligibility**) achieved a perfect recall score of **1.0**, meaning every expected policy keyword was successfully included in the generated answer.

- The query regarding “specialty occupation” achieved a recall score of **0.5**, with one of the two expected policy components captured. This reflects partial grounding and highlights a realistic limitation of summarization-based RAG systems when dealing with conceptual definitions.

Overall, the evaluation demonstrates that Agent-1 consistently retrieves the correct policy fragments from the knowledge base and produces accurate, grounded summaries. These results confirm that the RAG pipeline is functioning as intended and is capable of answering diverse immigration-related questions without relying on predefined static templates.

Conclusion & Future Work

Conclusion

The Visa Journey Agent demonstrates that AI-driven systems can dramatically simplify immigration compliance for students. By combining automated legal text extraction (Agent 1) with personalized timeline computation (Agent 2), the tool provides reliable, proactive guidance that traditionally requires human advisors.

One way to improve our system would be to use a more powerful embedding model. **all-MiniLM-L6-v2** is relatively small—about 22 million parameters with a 384-dimensional embedding size. Adopting a larger model, such as **EmbeddingGemma** or a **BERT variant fine-tuned for sentence similarity**, would likely produce richer embeddings and improve retrieval quality in our RAG-based question answering.

We could also broaden the scope of our knowledge base by incorporating additional federal sources and especially more university websites. This would increase the system’s applicability and make the tool relevant to a wider range of international students.

Future Enhancements

- Add a **“Document Upload”** feature for PDFs (I-20, job offers, etc.)
- Integrate calendar notifications (Google Calendar API)
- Expand beyond F-1 to J-1, and permanent residency workflows
- Add real-time USCIS rule fetching via web crawlers
- Extend to mobile app for student accessibility

Contribution

Data Collection + Cleaning: Jenny & Jared

Agent 1 + Agent 2: Abhay & Pavithra

Streamlit UI: Harshal