H-1B Sponsorship Intelligence Platform

AI-Powered Retrieval-Augmented Generation (RAG) Application

INFO 7390: Advances in Data Science and Architecture
Northeastern University
Fall 2025

Author Abhinav Kumar Piyush

Instructor Prof. Nick Bear Brown

Table of Contents

1. Project Goals

This project implements an AI-powered Retrieval-Augmented Generation (RAG) system to analyze H-1B visa sponsorship patterns using real USCIS data. The primary goals are:

- Build a semantic search system using Pinecone vector database and OpenAI embeddings
- Develop an interactive Streamlit application for data exploration and analysis
- Create a machine learning model to predict sponsorship likelihood
- Implement a RAG-powered chatbot using GPT-4o for intelligent Q&A;
- Provide actionable insights to help international students navigate H-1B sponsorship

2. Problem Statement

International students face significant challenges when navigating the H-1B visa sponsorship landscape:

Challenge	Description
Information Asymmetry	No centralized platform to understand which companies actively sponsor H-1B visas
Lottery Uncertainty	H-1B has only ~25-30% selection rate; students need data-driven strategies
No Predictive Insights	Raw government data exists but lacks analysis on sponsorship likelihood
Time-Consuming Research	Manual research across multiple sources is inefficient

3. Solution Overview

The H-1B Sponsorship Intelligence Platform is a comprehensive RAG application that combines semantic search, machine learning predictions, and natural language processing to deliver actionable insights through an intuitive interface.

The system processes real USCIS LCA Disclosure Data (600,000+ applications), aggregates company-level statistics, and enables users to explore sponsorship patterns through interactive visualizations, side-by-side comparisons, predictive modeling, and AI-powered Q&A.;

4. Technology Stack

Component	Technology	Purpose
Frontend	Streamlit	Interactive web application
Vector Database	Pinecone	Semantic search & retrieval
LLM	OpenAI GPT-4o	Response generation
Embeddings	text-embedding-3-small	Convert text to 1536-dim vectors
ML Models	Scikit-learn, XGBoost	Sponsorship prediction
Visualization	Plotly	Interactive charts
Data Processing	Pandas, NumPy	Data cleaning & analysis
Language	Python 3.10+	Backend development

5. System Architecture

The system follows a modular three-layer architecture designed for scalability and maintainability:

Data Layer

Raw USCIS Excel files (600K+ applications) are processed through clean_data.py to produce aggregated company-level statistics. The cleaned data is stored as CSV in the data/ folder and serves as the foundation for all downstream analysis.

Application Layer

A 6-page Streamlit application provides: Dashboard (visualizations), Company Comparison, ML Predictor, Data Pipeline (interactive upload), AI Advisor (RAG chatbot), and About page. Each module is separated for clean code organization.

External Services

Pinecone handles vector storage and semantic search using 1536-dimensional embeddings with cosine similarity. OpenAI provides text-embedding-3-small for embeddings and GPT-4o for chat completions in the RAG pipeline.

6. Features

Interactive Dashboard: Visualize top 10 sponsors, salary distributions, state-wise filings, and company size breakdown with interactive Plotly charts.

Company Comparison: Compare up to 3 companies side-by-side on filings, salaries, and sponsorship scores with detailed comparison tables and winner cards.

ML Predictor: ML-based prediction of sponsorship likelihood (HIGH/MEDIUM/LOW) with confidence scores, factor explanations, and personalized recommendations.

Data Pipeline: Interactive upload of raw USCIS Excel files with configurable cleaning parameters and real-time processing progress.

AI Advisor (RAG): GPT-4o powered chatbot that retrieves relevant company data from Pinecone and generates context-aware, factually grounded responses.

Documentation: Project documentation, architecture overview, and future work roadmap.

7. Data Pipeline

Data Source

Attribute	Value
Source	U.S. Department of Labor

Dataset	LCA Disclosure Data FY2024 Q4
URL	dol.gov/agencies/eta/foreign-labor/performance
Original Size	600,000+ H-1B applications
License	Public Domain (U.S. Government Work)

Processing Steps

1. Filter H-1B Certified applications only
2. Standardize company names (e.g., GOOGLE LLC → GOOGLE)
3. Convert all wages to annual salary
4. Remove outliers ($30K - $500K range)
5. Aggregate by company with statistics (filings, avg/median salary)
6. Compute sponsorship score (0-100) and categorize by size/salary

Note: The application ships with pre-cleaned data in the data/ folder. Users can also upload raw USCIS Excel files directly via the Data Pipeline page to process custom datasets within their session—no coding required.

8. RAG Implementation

How It Works

The RAG pipeline follows a standard retrieve-then-generate pattern: User Query → Embedding → Pinecone Search → Top-K Results → GPT-4o → Response. Each company is stored as a vector with metadata enabling filtered retrieval.

Why Pinecone?

Pinecone was selected for its managed scalability, fast cosine similarity search, and seamless integration with OpenAI embeddings—enabling efficient semantic retrieval without infrastructure overhead or operational complexity.

Intelligent Response Generation

While Pinecone retrieves relevant context, GPT-4o synthesizes the retrieved data into coherent explanations, personalized recommendations, and natural language insights—going beyond simple keyword search to deliver actionable intelligence grounded in real data.

9. ML Model

Input Features

Feature	Type	Values
job_category	Categorical	Software Engineer, Data Scientist, Manager, Consultant, Research, Other
salary	Numerical	$60K - $250K
state	Categorical	CA, WA, NY, TX, NJ, MA, IL, Other
salary_category	Categorical	Low, Medium, High

Training Results

Model	Accuracy	AUC	CV Score
Logistic Regression	99.9%	0.68	99.9%
Decision Tree	99.9%	0.64	99.9%
Random Forest	99.9%	0.64	99.9%
XGBoost	99.9%	0.56	99.9%

Best Model: Logistic Regression was selected for its interpretability and comparable performance to more complex models.

10. Evaluation

RAG Evaluation Results

The RAG system was evaluated by testing domain-specific queries and verifying that responses were grounded in retrieved company data rather than hallucinated content:

Test Query	Expected Company	Retrieved?	Grounded?
"Top H-1B sponsors"	Amazon, Microsoft	Yes	Yes
"Highest paying companies"	Meta, Google, Apple	Yes	Yes
"Companies in California"	Google, Meta, Apple	Yes	Yes
"Consulting firms for H-1B"	Cognizant, TCS, Infosys	Yes	Yes
"Best company in Washington"	Amazon, Microsoft	Yes	Yes

Metric	Value
Grounding Rate	100% (5/5 queries)
Avg Response Time	~2-3 seconds
Retrieval Accuracy	Top-K matched expected companies

11. Outputs & Deliverables

Streamlit Application: 6-page interactive web application with dashboard, comparison, predictor, pipeline, AI advisor, and about pages.

Cleaned Dataset: Aggregated company-level H-1B data with sponsorship scores and categories stored in data/ folder.

ML Model: Trained Logistic Regression model with feature weights saved as .pkl and .json files.

Training Notebook: Jupyter notebook (H1B_Model_Training.ipynb) ready for Google Colab execution.

Documentation: README.md, CITATIONS.md, index.html project webpage, and this PDF documentation.

Source Code: Modular Python codebase with data_loader, data_pipeline, prediction_model, vector_store, and rag_agent modules.

12. Ethical Considerations

Bias/Fairness: Data reflects historical filing patterns and may over-represent large employers and major tech hubs (CA, WA, NY). Results are not a measure of company merit or quality—interpret cautiously.

Privacy: Uses only public, aggregate company-level data from U.S. Department of Labor disclosures. No personal applicant information is collected, stored, or processed.

Misuse Prevention: This tool is for informational purposes only and should NOT be used for definitive immigration decisions. Always consult a qualified immigration attorney.

Content Guardrails: The AI Advisor is designed to refuse specific legal instructions and instead offers general guidance with cited data sources.

13. Future Work

Time-Series Prediction: With multi-year data (2020-2024), predict a company's future H-1B filing volume. Identify growing vs declining sponsors and forecast industry trends.

Resume-Based Prediction: Upload a resume to predict H-1B approval likelihood. Extract skills, education, and experience via NLP, then match against successful H-1B profiles.

Real-Time Data Updates: Automatically fetch and process new USCIS quarterly releases. Provide alerts when new data is available or trends change significantly.

Job Matching Engine: Given a user's profile, recommend specific job postings from companies with high sponsorship likelihood. Integrate with job boards APIs.

14. References

Data Sources

• U.S. Department of Labor - LCA Disclosure Data FY2024

• https://www.dol.gov/agencies/eta/foreign-labor/performance

Technologies

• Pinecone Documentation: https://docs.pinecone.io/

• OpenAI API Reference: https://platform.openai.com/docs

• Streamlit Documentation: https://docs.streamlit.io/

• Scikit-learn User Guide: https://scikit-learn.org/

Academic References

• Lewis et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"

• Johnson et al. (2019). "Billion-scale similarity search with GPUs"

INFO 7390: Advances in Data Science and Architecture

Northeastern University • Fall 2025

Instructor: Prof. Nick Bear Brown | Developed by: Abhinav Kumar Piyush