A06 - AgentVisa Project Report

AI Assistant for Visa Bulletin Insights

CISC 691 - Foundations of Next-Gen AI

Tien Dinh & Kenneth Peter Fernandes

Harrisburg University
Summer 2025

August 4, 2025

Abstract

AgentVisa is a comprehensive AI-powered assistant for US visa bulletin analysis built as part of CISC 691 - Foundations of Next-Gen AI course assignment. The project demonstrates advanced AI agent development with multi-provider LLM support, containerized microservices architecture, and production-ready cloud deployment capabilities.

The system integrates five major LLM providers including Google Gemini, OpenAI GPT, Anthropic Claude, Ollama, and HuggingFace Transformers, providing robust fallback mechanisms and flexibility in AI model selection. Built on a microservices architecture using FastAPI and Streamlit, the application processes real visa bulletin data from the US State Department, applying machine learning algorithms for trend analysis and predictive modeling.

The project successfully implements intelligent tool-calling capabilities, enabling the AI agent to automatically select and execute specialized visa analytics functions based on user queries. Comprehensive deployment strategies include local development with Docker Compose, Kubernetes orchestration with Minikube, and production-ready Google Kubernetes Engine deployment with SSL certificates, load balancing, and horizontal pod autoscaling.

This report presents the technical architecture, implementation details, deployment strategies, and performance metrics of a modern AI agent system that validates the practical application of next-generation AI concepts in real-world scenarios. The system includes comprehensive test coverage across unit, integration, and API testing, demonstrating robust software engineering practices alongside cutting-edge AI integration.

Contents

1 Executive Summary

AgentVisa represents a sophisticated implementation of modern AI agent architecture, combining multiple large language model providers with comprehensive visa bulletin analytics. The project successfully demonstrates practical application of AI agent development concepts learned throughout the course, delivering a production-ready system with enterprise-grade infrastructure.

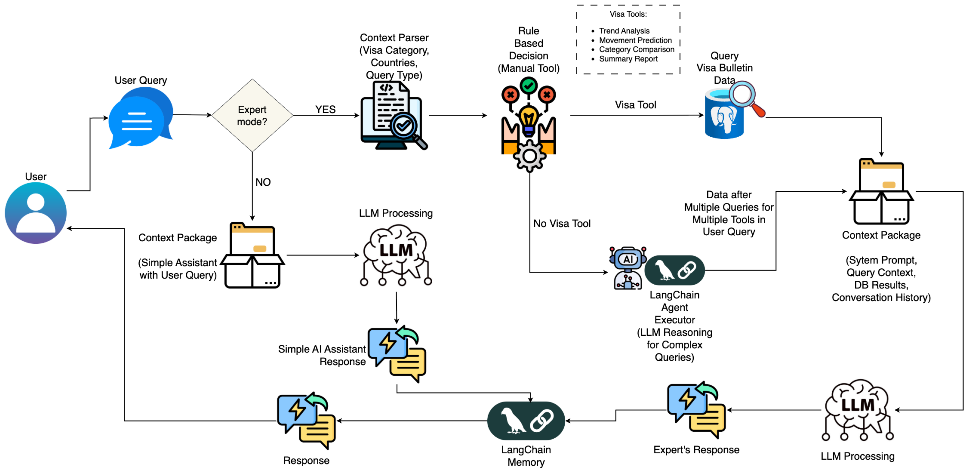
User
User Query
Expert mode?
YES
NO
Context Parser (Visa Category, Countries, Query Type)
Rule Based Decision (Manual Tool)
Visa Tools:
• Trend Analysis
• Movement Prediction
• Category Comparison
• Summary Report
Visa Tool
Query Visa Bulletin Data
No Visa Tool
Data after Multiple Queries for Multiple Tools in User Query
Context Package
(Sytem Prompt, Query Context, DB Results, Conversation History)
LangChain Agent Executor (LLM Reasoning for Complex Queries)
Context Package
(Simple Assistant with User Query)
LLM Processing
Simple AI Assistant Response
LLM Processing
Expert's Response
LangChain Memory
Response

Figure 1: AgentVisa Application Workflow - User interactions, AI agent processing, and system responses

1.1 Project Overview

- **Project Name:** AgentVisa - AI Assistant for Visa Bulletin Insights
- **Course:** CISC 691 - Foundations of Next-Gen AI (A06: Building the AI Agent of Your Choice)
- **Institution:** Harrisburg University (Summer 2025)
- **Contributors:** Tien Dinh, Kenneth Peter Fernandes
- **Repository:** https://github.com/kenneth-fernandes/cisc691-a06
- **Testing:** Comprehensive test suite with unit, integration, and API tests

2 System Overview & Architecture

2.1 About AgentVisa

AgentVisa is a containerized AI assistant with REST API backend that provides intelligent US visa bulletin analysis and multi-provider LLM chat capabilities. The system demonstrates advanced AI agent development with modern microservices architecture and production-ready cloud deployment capabilities.

2.2 Core System Architecture

Enterprise-ready microservices design with 4 containerized services:

- **Web Service:** Streamlit frontend for interactive user interface (Port 8501)
- **API Service:** FastAPI backend for REST API processing (Port 8000)
- **PostgreSQL Database:** Primary data storage with JSONB support (Port 5432)
- **Redis Cache:** High-performance caching layer (Port 6379)

2.3 Dual Operating Modes

AgentVisa provides two distinct interaction modes:

- **General Chat Mode:** Standard AI assistant capabilities for general queries and conversations
- **Visa Expert Mode:** Specialized U.S. visa bulletin analysis with integrated analytics tools and historical data access

3 Core Features & Capabilities

3.1 Multi-Provider AI Agent System

The system supports multiple LLM providers with intelligent fallback mechanisms:

- **Google Gemini (gemini-1.5-flash)** - Free tier, fast responses, and cost-effective (Primary)
- **Ollama (llama3.2)** - Offline processing, no API costs, and complete data privacy (Local)
- **OpenAI GPT (gpt-4o)** - Excellent for complex reasoning, nuanced immigration logic, and high reliability
- **Anthropic Claude (claude-3-5-sonnet)** - Long-context processing, trend analysis, and strong safety guarantees
- **HuggingFace Transformers** (Local, Free) - Community models for specialized tasks

3.2 Visa Bulletin Analytics Engine

Comprehensive analytics capabilities for US visa bulletin data:

- **Complete Category Support:** EB-1 through EB-5, F1 through F4
- **Country Analysis:** India, China, Mexico, Philippines, Worldwide
- **ML-Powered Predictions:** Random Forest & Logistic Regression models
- **Historical Data:** 2020-present with automated monthly updates
- **Interactive Dashboards:** Charts and visualizations

3.3 Data Processing Pipeline

Automated visa bulletin data collection and processing:

- **Web Scraping:** BeautifulSoup4-based HTML parsing from travel.state.gov
- **Data Validation:** Multi-stage quality assurance and verification
- **Historical Analysis:** Trend analysis from 2020-present
- **Predictive Modeling:** ML-powered forecast algorithms

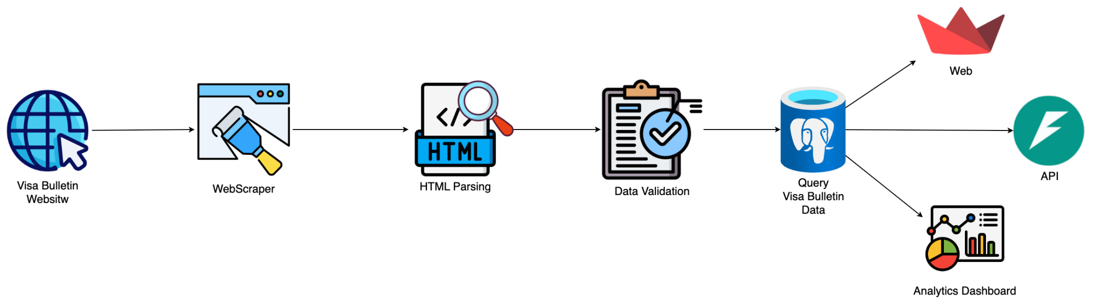

Figure 2: Visa Data Fetching Pipeline - Automated data collection from US State Department through validation to database storage

4 Technical Implementation

4.1 Core Components

4.1.1 AI Agent Core (`src/agent/core.py`)

The central AI agent implementation featuring:

- Multi-provider LLM integration with fallback support
- Conversation memory management
- Tool-calling capabilities for visa analytics
- Context-aware response generation

4.1.2 FastAPI Backend (`src/api/main.py`)

RESTful API service providing:

- RESTful API endpoints for agent interactions
- CORS middleware for cross-origin requests
- Global exception handling
- Health check endpoints

4.1.3 Data Processing Pipeline

Automated data collection and processing:

- Automated visa bulletin scraping from travel.state.gov
- Multi-stage data validation and quality assurance
- Historical data collection and storage
- ML model training and prediction

4.2 Deployment Options

Method	Use Case	Setup Time	Cost	Security
GKE Production	Production deployment	12-20 min	~$143/month	Enterprise
Minikube Local	Development/testing	5-10 min	Free	Basic
Docker Compose	Quick testing	2-5 min	Free	None

Table 1: Deployment Options Comparison

4.3 Infrastructure Features

- **Kubernetes Orchestration:** Full k8s deployment with autoscaling
- **SSL/TLS Security:** Google-managed certificates
- **Load Balancing:** Production-grade traffic distribution
- **Horizontal Pod Autoscaling:** Automatic resource scaling
- **Terraform Automation:** Infrastructure as Code

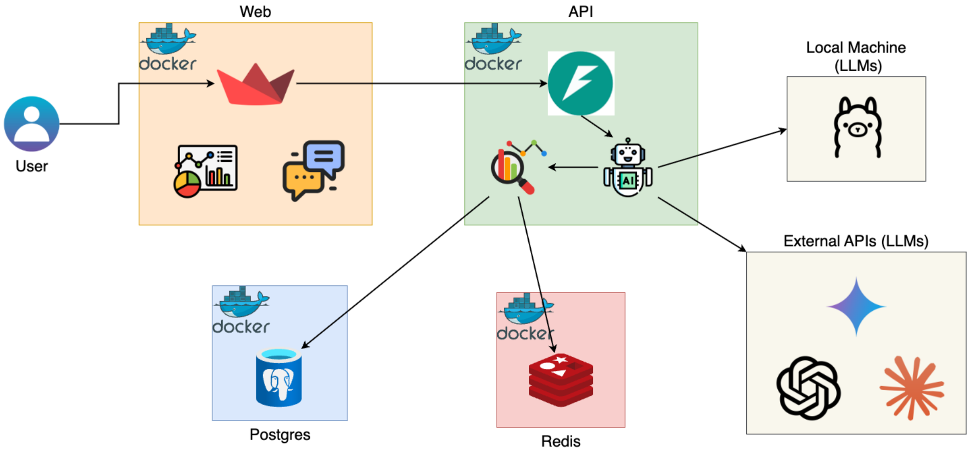

Figure 3: Local Docker Deployment - Microservices architecture for development

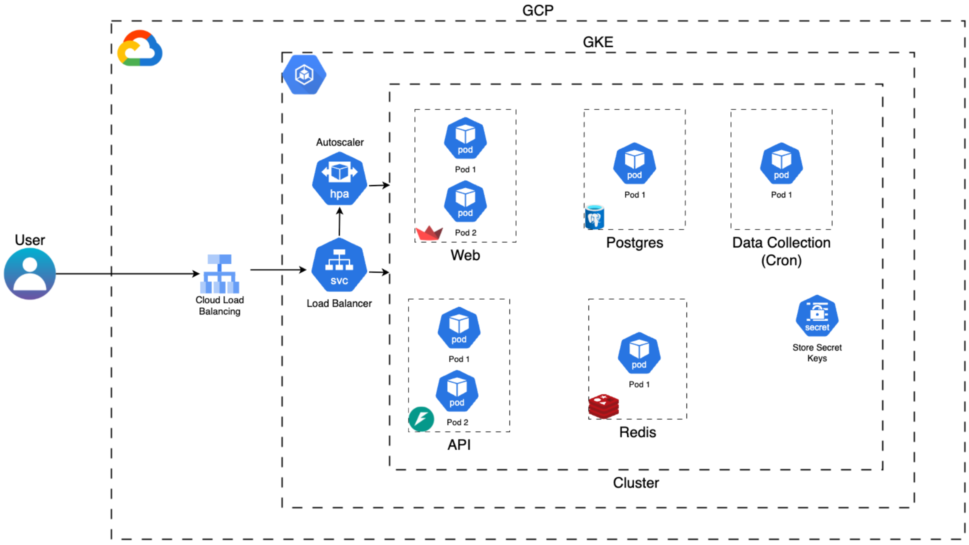

Figure 4: GKE Cloud Architecture - Production deployment with scaling

5 Key Technologies & Dependencies

5.1 AI/ML Stack

- **LangChain Framework:** Agent orchestration and tool integration
- **Scikit-learn:** ML models for visa prediction
- **Pandas/NumPy:** Data processing and analysis

5.2 Web Framework

- **FastAPI:** Modern async web framework
- **Streamlit:** Interactive frontend interface
- **Uvicorn:** ASGI server implementation

5.3 Database & Caching

- **PostgreSQL:** Primary database with JSONB support

- **Redis:** High-performance caching layer

5.4 DevOps & Testing

- **Docker/Docker Compose:** Containerization
- **Kubernetes:** Container orchestration
- **Pytest:** Comprehensive testing suite
- **GitHub Actions:** CI/CD pipeline

6 Deployment & Operations

6.1 Deployment Architecture Overview

AgentVisa supports multiple deployment strategies optimized for different use cases:

Method	Use Case	Setup Time	Cost	Security
GKE Production	Production deployment	12-20 min	~$143/month	Enterprise
Minikube Local	Development/testing	5-10 min	Free	Basic
Docker Compose	Quick testing	2-5 min	Free	None

Table 2: Deployment Options Comparison

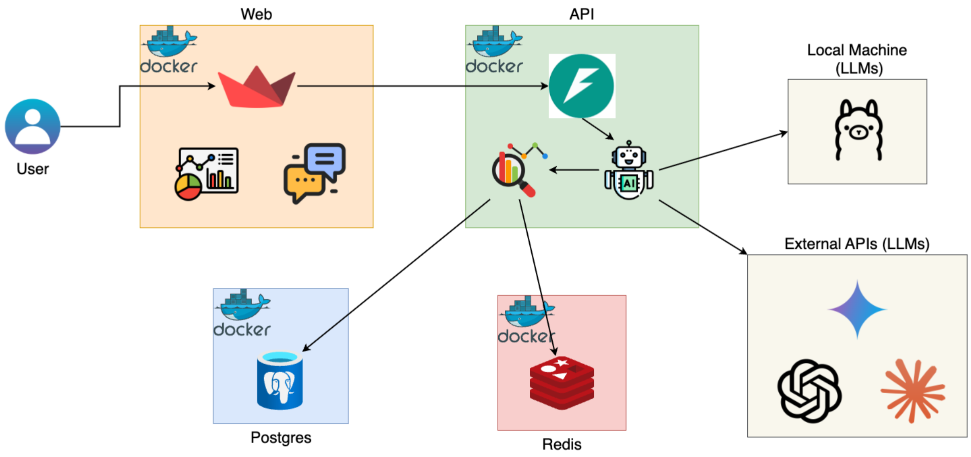

Figure 5: Local Docker Deployment - Microservices architecture for development

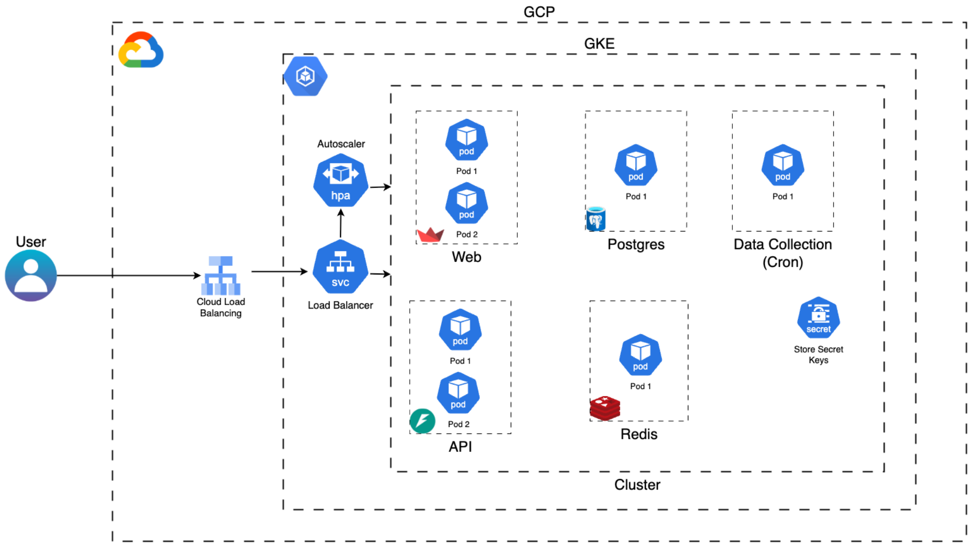

Figure 6: GKE Cloud Architecture - Production deployment with scaling

6.2 Cloud Infrastructure (GKE)

Production deployment on Google Kubernetes Engine features:

- **Google Cloud Platform:** Primary cloud provider with integrated services
- **Kubernetes Engine:** Managed container orchestration with autoscaling
- **Terraform Automation:** Infrastructure as Code for consistent deployments
- **Kubernetes Secrets:** Secure API key and credential management
- **Load Balancing:** Production-grade traffic distribution with health checks
- **SSL/TLS Security:** Google-managed certificates with automatic renewal

6.3 Local Development Environment

Developer-friendly local setup options:

- **Minikube:** Local Kubernetes environment for testing
- **Docker Compose:** Simplified multi-container orchestration
- **Hot Reload:** Development-friendly configuration with live updates
- **Resource Optimization:** Efficient resource usage for local development

7 Development Highlights

7.1 Advanced System Features

1. **Intelligent Tool Integration:** Automatic tool selection based on query context and user intent
2. **Multi-Provider Fallback:** Seamless switching between LLM providers with graceful degradation
3. **Production Monitoring:** Comprehensive health checks, logging, and observability with K9s dashboard
4. **Cost Optimization:** Efficient resource utilization and automated scaling (~$143/month for cloud)

8 Quality Assurance

8.1 Testing Strategy

The project implements comprehensive testing across multiple dimensions:

- **Unit Tests:** Individual component testing
- **Integration Tests:** End-to-end workflow validation
- **API Tests:** REST endpoint functionality
- **Cache Tests:** Redis performance and fallback
- **Testing Focus:** Core APIs and critical functionality covered

8.2 Development Best Practices

- **Clean Architecture:** Layered design with separation of concerns
- **Factory Patterns:** Flexible agent creation and configuration
- **Environment Configuration:** Secure secret management
- **Documentation:** Comprehensive markdown documentation

9 Project Success Metrics

9.1 Technical Achievements

✓ Multi-provider AI agent successfully implemented

✓ Production-ready microservices architecture

✓ Comprehensive visa analytics with ML predictions

✓ Full containerization and k8s deployment

✓ Automated CI/CD pipeline with testing

9.2 Learning Objectives Met

✓ Modern AI framework integration (LangChain)

✓ Multiple LLM provider support and flexibility

✓ Modular and extensible architecture design

✓ Practical AI integration skills demonstration

10 Future Enhancements & Recommendations

1. **Enhanced ML Models:** Implement deep learning for improved predictions
2. **Real-time Updates:** WebSocket integration for live data streaming
3. **Advanced Analytics:** More sophisticated trend analysis and reporting
4. **Mobile Support:** React Native or Progressive Web App
5. **API Rate Limiting:** Enhanced security and resource protection

11 API Architecture & Endpoints

The AgentVisa system exposes a comprehensive REST API organized into three main categories:

11.1 Core Agent Endpoints

- `POST /api/agent/chat` - Send messages to AI agent with dual mode support
- `GET /api/agent/conversation/{id}` - Retrieve chat history and conversation state
- `GET /api/agent/providers` - List available LLM providers and their status

11.2 Visa Analytics Endpoints

- `POST /api/analytics/trends` - Analyze historical visa bulletin trends
- `POST /api/analytics/predictions` - Generate ML-powered forecasting predictions
- `GET /api/analytics/historical` - Access historical visa bulletin data

11.3 Data Access Endpoints

- `GET /api/analytics/categories` - Retrieve supported visa categories (EB-1 to EB-5, F1 to F4)
- `GET /api/analytics/countries` - List supported countries for analysis
- `GET /api/analytics/bulletins` - Access available bulletin dates and metadata

12 Challenges & Technical Learnings

The development process provided valuable insights into modern AI system deployment and integration challenges:

12.1 Communication Architecture Decisions

- **WebSocket Integration Issues:** Initial WebSocket implementation with Streamlit proved unstable, leading to adoption of reliable HTTP-based communication for better compatibility and stability
- **Solution:** Implemented robust REST API communication patterns with proper error handling and timeout management

12.2 Cloud Deployment Challenges

- **Ollama on GKE Limitations:** CPU constraints and lack of GPU support made llama3 1B/3B models unstable without custom high-resource nodes
- **Autoscaling Delays:** Resource constraints occurred when autoscaler delayed node provisioning during experimental pod additions
- **Data Population:** Initial visa data insertion issues resolved through Kubernetes cron job implementation for automated data population

12.3 Local Development Optimizations

- **M2 Air Performance:** Docker performance limitations due to resource constraints and lack of GPU support necessitated use of compact models like LLaMA 3.2 1B
- **Solution:** Optimized local development workflow with efficient model selection and resource management

12.4 Development Tools & AI Integration

- **AI-Enhanced Development:** Extensive use of VS Code with GitHub Copilot and Claude for AI-enhanced coding, auto-suggestions, and intelligent code reviews
- **Collaboration:** Project developed through intensive brainstorming sessions with Claude AI and ChatGPT, demonstrating effective human-AI collaboration

13 Conclusion

AgentVisa successfully demonstrates the practical application of modern AI agent development concepts, combining multiple cutting-edge technologies into a cohesive, production-ready system. The project showcases technical proficiency in AI/ML, cloud infrastructure, and software engineering best practices while delivering genuine value through intelligent visa bulletin analysis.

The comprehensive architecture, from local development to cloud deployment, reflects industry-standard practices and positions the project as a strong foundation for further AI agent development endeavors. The successful integration of multiple LLM providers, sophisticated data processing pipelines, and enterprise-grade deployment infrastructure validates the learning objectives of the CISC 691 course.

As demonstrated through the visual architecture diagrams (Figures 1, 5, 6, and 2), the system provides a complete end-to-end solution for AI-powered visa bulletin analysis, from data collection through user interaction and intelligent response generation.

Generated on: August 4, 2025
Project: AgentVisa - AI Assistant for Visa Bulletin Insights
Course: CISC 691 - Foundations of Next-Gen AI