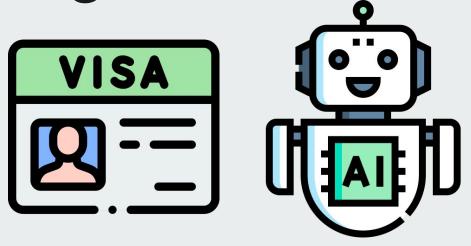
# AgentVisa - Al Assistant for Visa Bulletin Insights

Dinh, Tien Fernandes, Kenneth





### Introduction



**Tech Stack** 



**Local Deployment** 



**Front-End** 



**Cloud Deployment** 



**Challenges and Learnings** 



# Introduction

# **About AgentVisa**

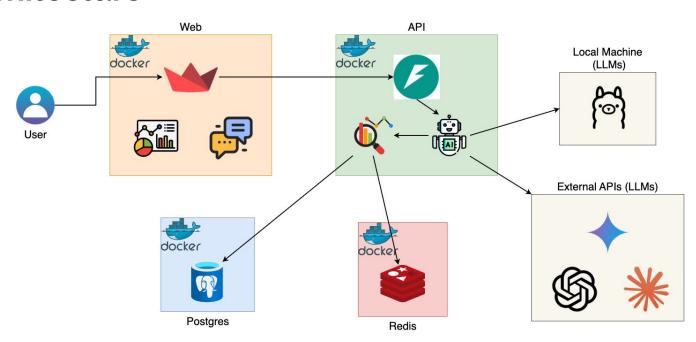
A containerized AI assistant with REST API backend that provides intelligent US visa bulletin analysis and multi-provider LLM chat capabilities

- Key Features
  - O Dual Chat Mode You can chat as a Visa Expert Agent or a normal Chat Bot
  - Multi-Provider LLM Support Google Gemini, OpenAl GPT, Anthropic Claude, Local Ollama
  - Comprehensive Visa Analytics EB & Family-based categories with ML-powered forecasting
  - o Modern Architecture Docker microservices, FastAPI backend, PostgreSQL + Redis
  - Production Ready Full Kubernetes deployment (GKE) with scalable infrastructure backed by Google Cloud
     Platform



# Deployment

# **Architecture**



# **API Endpoints**



- POST /api/agent/chat Send messages to Al agent
- GET
  /api/agent/conversation/{id}
   Get chat history
- GET /api/agent/providers List LLM providers



- POST /api/analytics/trends Analyze visa trends
- POST /api/analytics/predictions  $\ensuremath{\mathsf{ML}}$  forecasting
- GET /api/analytics/historical-Historical data



- GET
  /api/analytics/categories Visa
  categories
- -GET /api/analytics/countries
- Supported countries
- GET /api/analytics/bulletins-Bulletin dates



# **Cloud Deployment**

## **Overview**



Google Cloud Platform as Cloud Provider

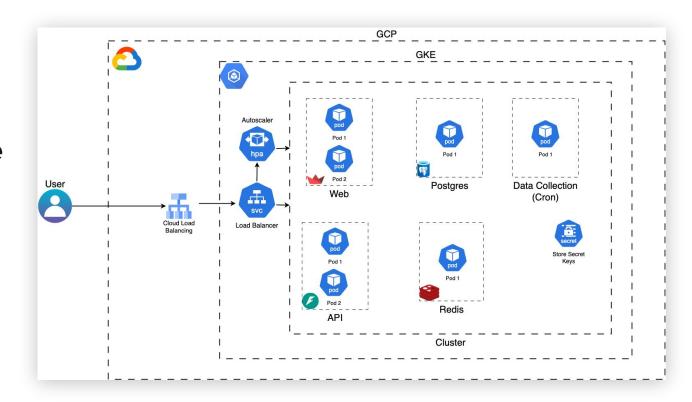
Google Kubernetes Engine as Deployment Engine

Kubernetes Secret is used to store secrets such as API keys

Terraform was used to build the blueprint of GKE



# **Architecture**



# **K9s - GKE Monitoring Dashboard**

Cluster: gke_agent-visa-	-467603_us-centrall-a_agentvisa-clu: -467603_us-centrall-a_agentvisa-clu: -467603_us-centrall-a_agentvisa-clu: 16000	ster <1> visa-app < ster <2> default < <	b> Bench Run/Stop <\> ctrl-d> Delete	Logs Previous  ift-f> Port-Forward  Show PortForward	
NAME†	TYPE	CLUSTER-IP	EXTERNAL-IP	PORTS	AGE
api	ClusterIP	34.118.235.244	EXTERNAL II	8000►0	112m
postgres redis web web-lb web-nodeport-ssl	ClusterIP ClusterIP ClusterIP LoadBalancer NodePort	34.118.226.194 34.118.232.76 34.118.236.0 34.118.230.219 34.118.233.157	34.170.13.186	5432►0 6379►0 8591►0 80−31071 http:8501►30001	112m 112m 112m 74m 38m
<service></service>					

# **K9s - GKE Monitoring Dashboard**

```
Context: gke_agent-visa-467603_us-centrall-a_agentvisa-cluster [RW] <0> all
                                                                                          Attach
                                                                                                       <ctrl-k>
                                                                                                                                <o> Show Node
Cluster: gke_agent-visa-467603_us-central1-a_agentvisa-cluster
                                                                  <1> visa-app
                                                                                                                                <f> Show PortForward
                                                                                <ctrl-d> Delete
User: gke agent-visa-467603 us-central1-a agentvisa-cluster
                                                                  <2> default
                                                                                                                Logs Previous
                                                                                                                                <t> Transfer
K9s Rev: v0.50.9
                                                                                                       <shift-f> Port-Forward
                                                                                                                                <V> YAML
K8s Rev: v1.32.6-gke.1096000
        5%
                                                                                <shift-j> Jump Owner
                                                                                                                 Shell
        38%
                                                                               pods(visa-app)[5]
                                                    RESTARTS CPU %CPU/R %CPU/L MEM %MEM/R %MEM/L IP
NAME *
                             PF READY STATUS
                                                                                                                                                                       AGE
api-7d88847487-l8kgn
                                1/1
                                       Runnina
                                                                               0 150
                                                                                           29
                                                                                                   14 10.72.0.11
                                                                                                                  gke-agentvisa-cluste-agentvisa-node-p-9daf58af-f98b
postgres-0
                             • 1/1
                                                                               2 30
                                                                                           23
                                                                                                  11 10.72.2.6
                                                                                                                  gke-agentvisa-cluste-agentvisa-node-p-9daf58af-p1cs
                                       Running
redis-867867f8fd-2x494
                             • 1/1
                                       Running
                                                                                                    1 10.72.2.7
                                                                                                                  gke-agentvisa-cluste-agentvisa-node-p-9daf58af-p1cs
                                       Completed
visa-data-collection-9hcrt
                                                                                                    0 10.72.3.7
                                                                                                                  gke-agentyisa-cluste-agentyisa-node-p-9daf58af-7xpv
web-f8cf565bd-7dmhp
                            • 1/1
                                       Running
                                                                                                  11 10.72.2.14 qke-agentvisa-cluste-agentvisa-node-p-9daf58af-plcs
<namespace>
              <pod>
```



# **Front-End**

### **Settings** Provider google **Expert Mode** General ⊕ Active: General purpose Al assistant Clear General Chat Clear All Chats & API Status **▼** Connected **About**

Al Agent using LangChain with support for multiple providers.

Available Models:

- (Free)
- Ollama (Local)
- OpenAl GPT (Paid)
- Anthropic (Paid)

⊕ Agent Chat

#### AgentVisa Chat - General Mode

(an you tell me when was Al Invented?

pinpointing the exact invention of Al is tricky because it's not a single invention but rather a gradual evolution of ideas and technologies. There's no single "eurekai" moment. However, we can identify key milestones:

- The Dartmouth Workshop (1956): This is widely considered the birth of AI as a field. John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester organized a workshop that brought together leading researchers and formally established the field's goals and research directions. The term "artificial intelligence" was coined there.
- Early Developments (1950s-1970s): The following decades saw significant progress in areas like game playing (e.g., checkers), theorem proving, and early natural language processing. However, progress was slower than initially anticipated, leading to periods of reduced funding and enthusiasm (the "Al winters")
- Expert Systems (1980s): Expert systems, which aimed to mimic the decision-making of human experts in specific domains, became prominent. While successful in niche applications, they lacked the generalizability that researchers initially envisioned.
- Machine Learning Boom (2000s-Present): The availability of vast amounts of data and increased computing power fueled the rise of machine learning, a subfield of Al that focuses on enabling computers to learn from data without explicit programming. This led to breakthroughs in areas like image recognition, natural language processing, and robotics, and continues to drive rapid advancements today.

So, while the Dartmouth Workshop in 1956 is generally considered the birth of Al as a field, its development is an ongoing process with significant milestones spread across several decades. There's no single inventor, but rather a community of researchers who have contributed to its evolution.

Response time: 2.99s | W Mode: General | Provider: google

When was AI chat bot introduced?

The concept of AI chatbots has been around since the 1960s, with ELIZA being a particularly notable early example. However, the technology wasn't sophisticated enough to create truly conversational and helpful bots at that time.

We've seen significant advancements and waves of popularity since then. For example, the rise of rule-based chatbots in the 90s and early 2000s was followed by the emergence of more sophisticated Al-powered chatbots leveraging machine learning, particularly in the last decade. There's no single "introduction" date, but rather a gradual evolution. The current era of highly advanced, large language model (LLM)-based chatbots like myself is relatively recent, really taking off in the last few years.

® Response time: 1.84s | ® Mode: General | ⊕ Provider: google

Type your message here

-

Ollama (Local)

OpenAl GPT (Paid)
Anthropic (Paid)

Type your message here..

#### **Settings**

Provider

ollama

of Expert Mode

Visa Expert

bulletin analysis

V

Quick Actions

✓ Show EB-2 India tre... ✓

Execute Action

Clear Visa Expert Chat

Clear All Chats

Agent Chat

## of AgentVisa Chat - Visa Expert Mode

my priority date is 9/27/2023 EB3, when do I expect it to become current?

Based on your priority date of September 27th, 2023, under EB-3 category, it appears that this category is experiencing an average advancement rate of about 20 days per month.

Given this information, we can expect that your priority date will likely become current within the next few months. However, predicting an exact timeline is challenging due to various factors such as changes in visa policy and new immigrant visa availability.

To put this into perspective, if we apply the average advancement rate of 20 days per month to your priority date, it would suggest that your priority date will likely become current around late 2025 or early 2026. However, please note that this is a rough estimate and actual times may vary.

It's essential to keep in mind that the visa bulletin is updated quarterly, and the advancement rates can fluctuate significantly between updates. Therefore, it's crucial to continue monitoring the bulletin for any changes and adjustments to your priority date.

I recommend regularly checking the official U.S. Department of State website or reputable visa bulletin sources for the latest updates on your category's status.

® Response time: 8.59s | ® Mode: Visa Expert | ® Provider: ollama



# **Tech Stack**

### **Tech Stack**



**Streamlit** – Interactive web UI for real-time dashboards and analytics



Python based FastAPI -

High-performance asynchronous REST API with auto-generated docs



**LangChain** – Agent-based architecture supporting multi-provider LLM integration



**PostgreSQL** – ACID-compliant database for structured visa bulletin data



**Redis** – API response caching and session storage for improved performance



GitHub - Source control and CI/CD pipeline



**Docker** – Multi-stage builds for packaging and deploying web and API services



**Kubernetes** – Scalable, production-grade deployment with load balancing



VS Code with GitHub Copilot and Claude – AI-enhanced coding, auto-suggestions, and smart code reviews.

### Al and LLMs



#### **Currently Supported**



Google Gemini (gemini-1.5-flash)

Free tier, fast responses, and cost-effective



#### **Future Release**



OpenAl GPT (gpt-4o)

Excellent for complex reasoning, nuanced immigration logic, and high reliability in critical decisions



Ollama (Ilama3.2) - (on local)

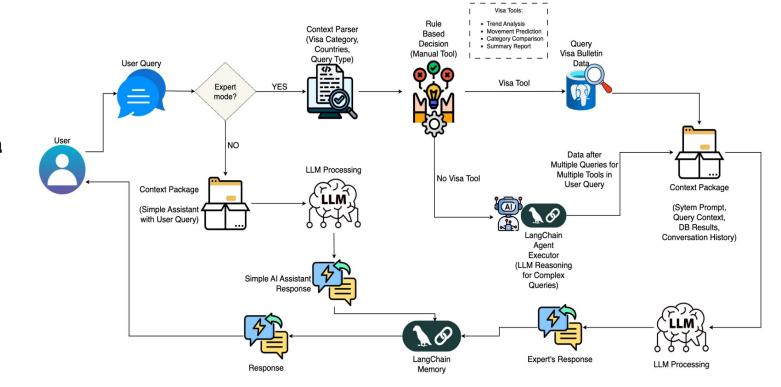
Offline processing, no API costs, and complete data privacy



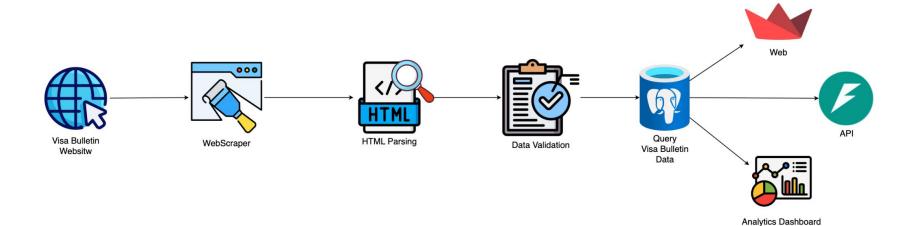
Anthropic Claude (claude-3-5-sonnet)

Long-context processing, trend analysis, and strong safety guarantees

Our AgentVisa app in a nutshell



# Visa Bulletin Data Fetching





# Challenges and Learnings

# **Challenges and Learnings**

- **Websockets**: Faced WebSocket integration issues with Streamlit, so switched to reliable HTTP-based communication for better compatibility and stability.
- Ollama on GKE: Faced CPU limitations and lacked GPU support, making llama 3 1B/3B models unstable without custom high-resource nodes, so used local deployments.
- Autoscaling on GKE: Adding new pods for experiments led to resource constraints as autoscaler delayed node provisioning.
- **Populate Visa Data for LLMs on GKE:** Initially faced issues with data insertion, but implemented a workaround by spinning up a Kubernetes pod to populate data via a cron job.
- Ollama Local Dev on M2 Air: Docker performance was poor due to limited resources and lack of GPU support, so compact models like LLaMA 3.2 1B were used.

## References

- This project was proudly pieced together through an intense brainstorming session with Claude, endless existential debates with ChatGPT, and a sprinkle of manual debugging tears.
- Basically, it took two Als and one tired human to make it work.