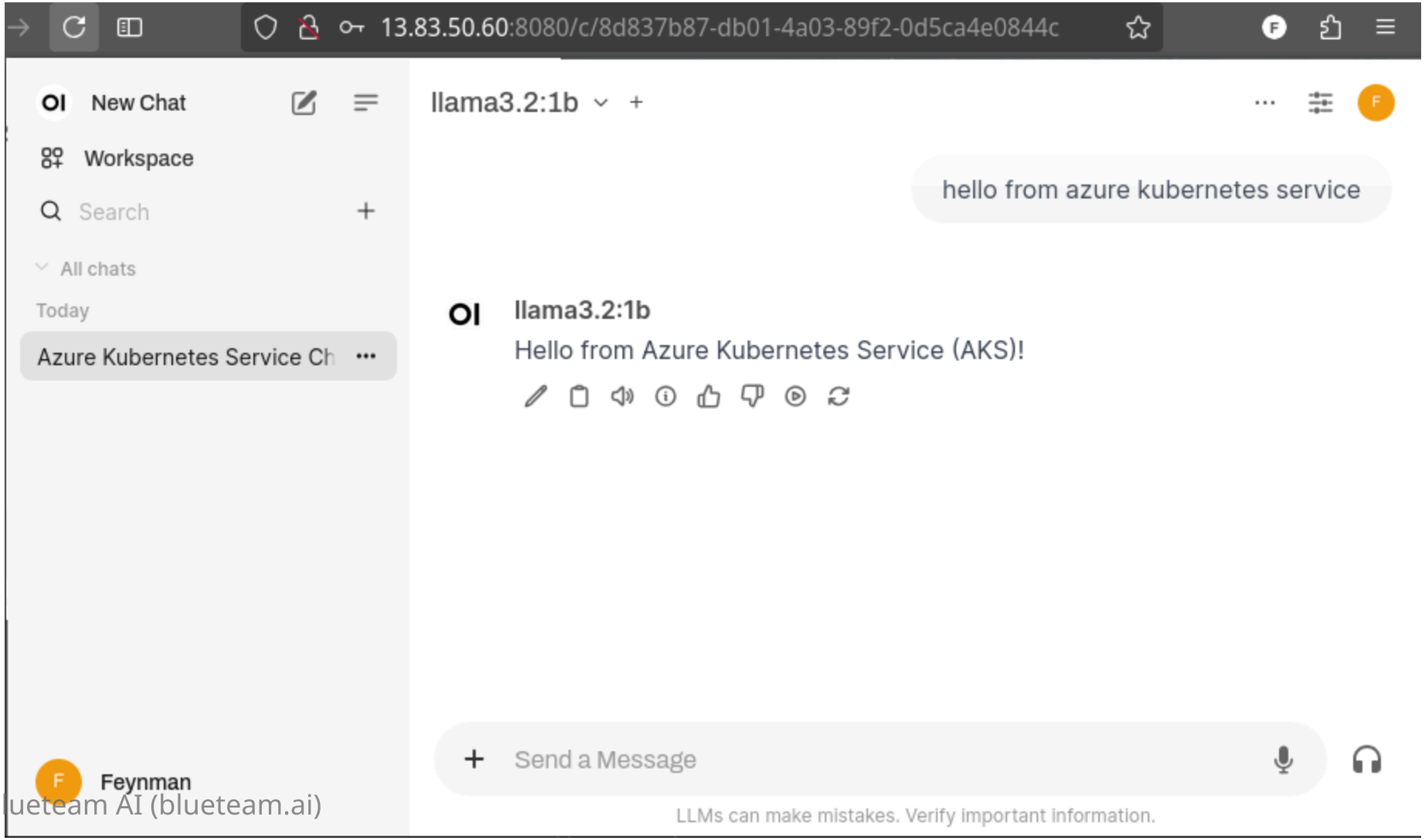
Private AI Infrastucture on Azure

A Practical Tutorial with Ollama & Open WebUI

Feynman Liang (feynman@blueteam.ai)

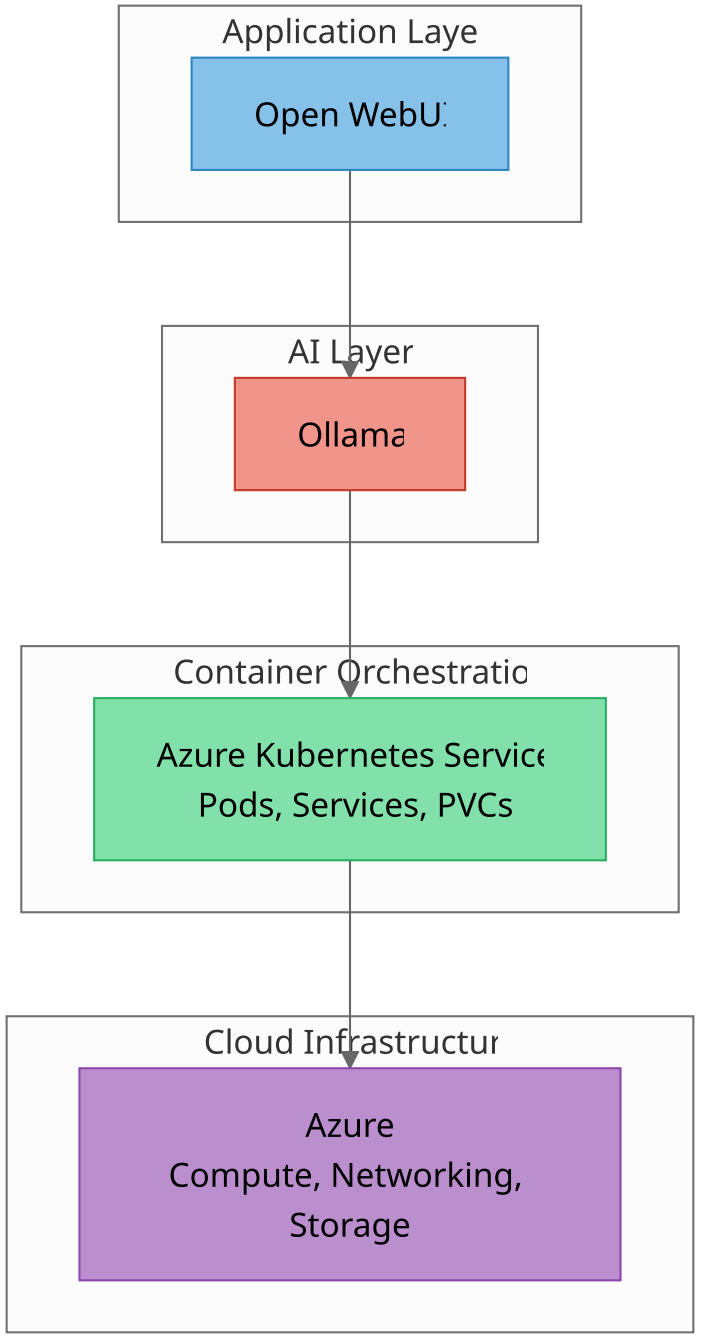
https://linkedin.com/in/feynman

End Result


13.83.50.60:8080/c/8d837b87-db01-4a03-89f2-0d5ca4e0844c
New Chat
Workspace
Search
All chats
Today
Azure Kubernetes Service Ch
Feynman
llama3.2:1b
hello from azure kubernetes service
llama3.2:1b
Hello from Azure Kubernetes Service (AKS)!
Send a Message
LLMs can make mistakes. Verify important information.

Overview

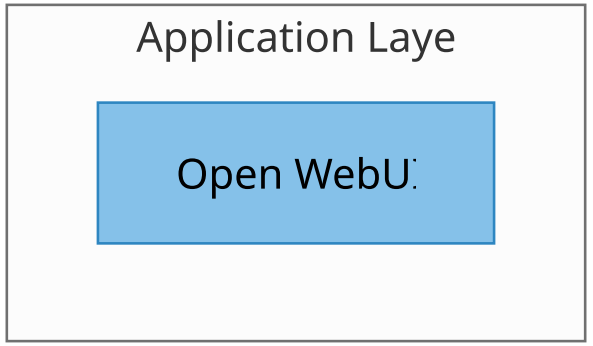
- Open Source AI
  - Open WebUI
  - Ollama
- Declarative Infrastructure
  - OpenTofu
  - Kubernetes
- Walkthrough

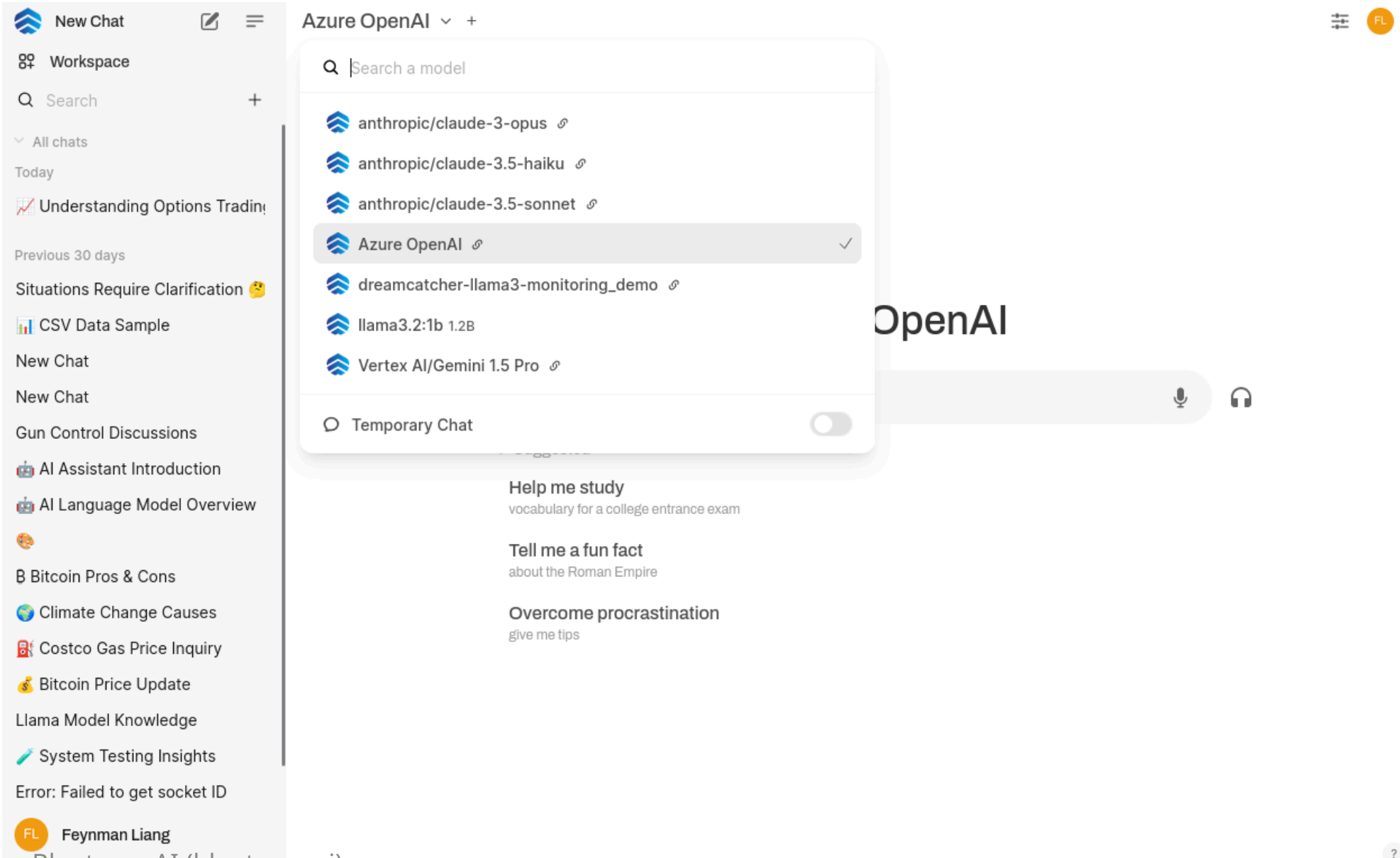


Just show me the code

https://github.com/fmops/azure-private-ai-template

Chat frontend with Open WebUI




New Chat
Workspace
Search
All chats
Today
Understanding Options Tradin
Previous 30 days
Situations Require Clarification
CSV Data Sample
New Chat
New Chat
Gun Control Discussions
AI Assistant Introduction
AI Language Model Overview
Bitcoin Pros & Cons
Climate Change Causes
Costco Gas Price Inquiry
Bitcoin Price Update
Llama Model Knowledge
System Testing Insights
Error: Failed to get socket ID
Feynman Liang
Azure OpenAI
Search a model
anthropic/claude-3-opus
anthropic/claude-3.5-haiku
anthropic/claude-3.5-sonnet
Azure OpenAI
dreamcatcher-llama3-monitoring_demo
llama3.2:1b 1.2B
Vertex AI/Gemini 1.5 Pro
Temporary Chat
OpenAI
Help me study
vocabulary for a college entrance exam
Tell me a fun fact
about the Roman Empire
Overcome procrastination
give me tips

Azure OpenAI + ⋯ FL

anthropic/claude-3.5-sonnet –

llama3.2:1b –

Explain options trading in simple terms if I'm familiar with buying and selling stocks.

Azure OpenAI 6:26 AM

Options trading can be thought of as a way to buy or sell the **right** to trade stocks at a specific price within a certain timeframe, rather than trading the stocks themselves. Here's a simple breakdown:

1. **What is an Option?** : An option is a contract that gives you the right (but not the obligation) to buy or sell a stock at a predetermined price

anthropic/claude-3.5-sonnet 6:26 AM

Here's options trading explained simply:

While regular stock trading involves buying and selling actual shares, options give you the right (but not the obligation) to buy or sell a stock at a specific price (strike price) within a certain time period (expiration date).

There are two basic types:

1. Call Options: The right to BUY stock

llama3.2:1b 6:26 AM

If you're familiar with buying and selling stocks, here's a simplified explanation of options trading:

What is an option?

An option is a contract that gives the buyer (

Azure OpenAI

Price of bitcoin today

Azure OpenAI

Searched 3 sites

The price of Bitcoin (BTC) today is approximately $97,264.18, with a market cap of about $1,924.42 billion and a 24-hour trading volume of $96.84 billion. The price has increased by 3.2% in the last 24 hours.

1 Price of bitcoin today 2 Price of bitcoin today

Web Search

Upload Files

Send a Message

LLMs can make mistakes. Verify important information.

Vertex AI/Gemini 1.5 Pro

deploying-private-oss...
File 5.4 KB

summarize the key points

2/2

Vertex AI/Gemini 1.5 Pro

This document provides a practical tutorial on deploying a private AI infrastructure on Azure using Ollama and Open WebUI. Key points include:

- **Open-source focus:** Leverages open-source AI models for cost benefits and control.
- **Ollama and Open WebUI:** Uses Ollama for LLM management and Open WebUI for a ChatGPT-like interface.
- **OpenTofu for infrastructure:** Employs OpenTofu (similar to Terraform) for Infrastructure as Code.
- **Azure deployment:** Details deploying on Azure, including Kubernetes and resource management.
- **Code availability:** Points to a GitHub repository (https://github.com/fmops/azure-private-ai-template) for the project code.
- **POC-level setup:** The provided setup is proof-of-concept and includes basic networking, single-region deployment, and no advanced security. Future improvements are suggested, such as HA configuration, backup/restore, multi-region deployment, security hardening, and monitoring/alerting.

1 deploying-private-oss-llms-on-azure.md

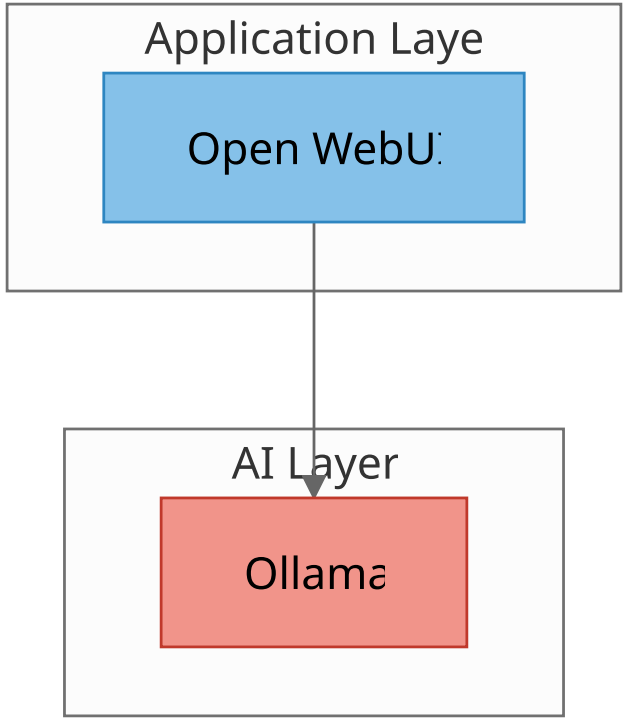
Web Search

Upload Files

Send a Message

LLMs can make mistakes. Verify important information.

Open Source LLMs with Ollama

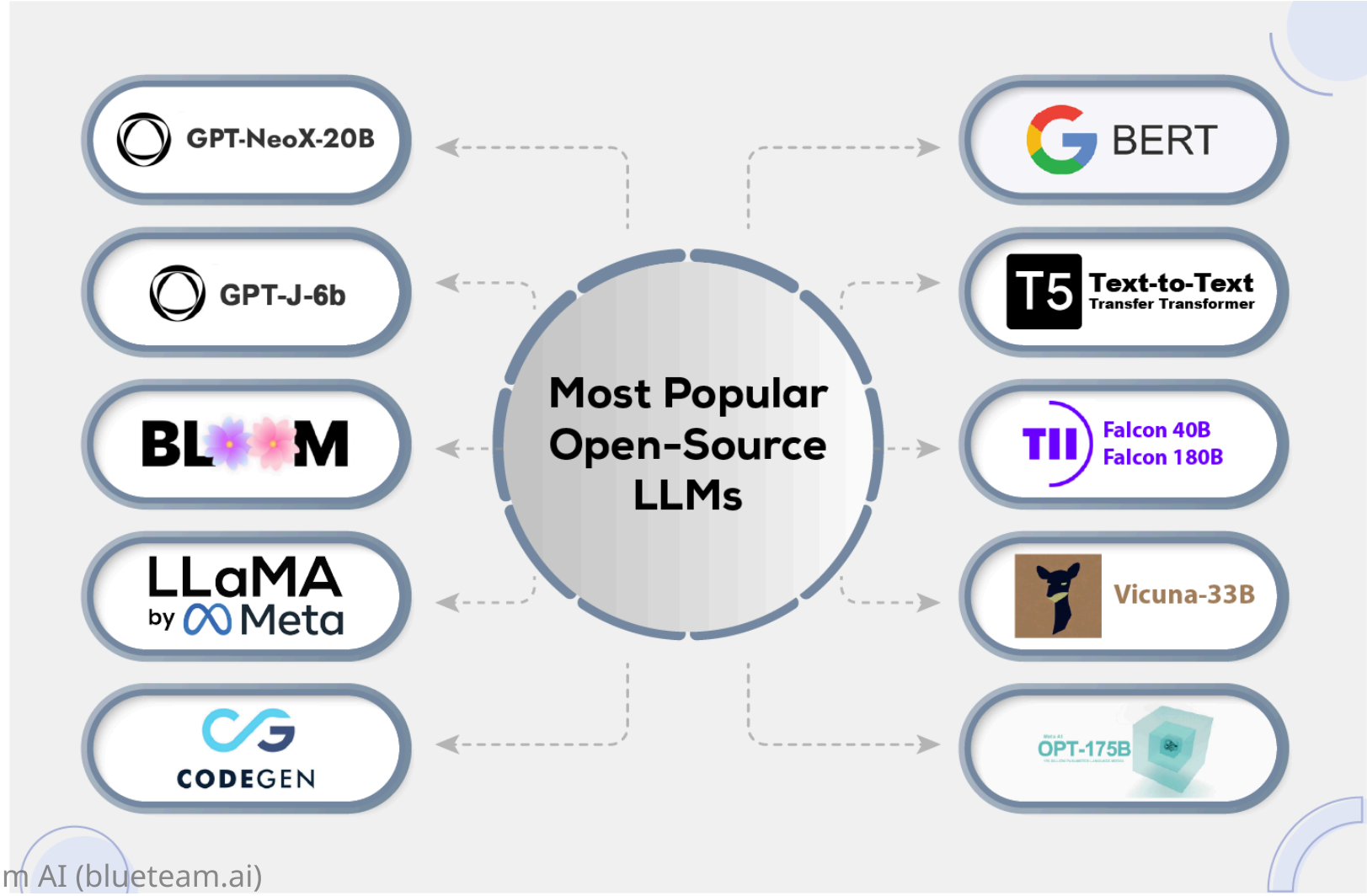


Why Open Source AI?

- Cost Benefits
- Freedom & Control
- Comparing to Closed Models

Image credits https://www.inspirisys.com/blog-details/Top-10-Open-Source-LLMs-of-2024-A-Complete-Guide/175

Open Source LLM Landscape



Ollama

Blog Discord GitHub

Models Sign in Download

Search models

All Embedding Vision Tools Popular

qwen2.5-coder
The latest series of Code-Specific Qwen models, with significant improvements in code generation, code reasoning, and code fixing.
tools 0.5b 1.5b 3b 7b 14b 32b
475.1K Pulls 196 Tags Updated 9 days ago

llama3.2
Meta's Llama 3.2 goes small with 1B and 3B models.
tools 1b 3b
3.2M Pulls 63 Tags Updated 8 weeks ago

llama3.1
Llama 3.1 is a new state-of-the-art model from Meta available in 8B, 70B and 405B parameter sizes.
tools 8b 70b 405b
10.5M Pulls 93 Tags Updated 2 months ago

mistral
The 7B model released by Mistral AI, updated to version 0.3.
tools 7b
5.6M Pulls 84 Tags Updated 4 months ago

qwen2
Qwen2 is a new series of large language models from Alibaba group
tools 0.5b 1.5b 7b 72b
3.9M Pulls 97 Tags Updated 2 months ago

Image credits https://www.jeremymorgan.com/blog/generative-ai/ollama-windows/

```
C:\Users\jerem>ollama help
Large language model runner

Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve       Start ollama
  create      Create a model from a Modelfile
  show        Show information for a model
  run         Run a model
  pull        Pull a model from a registry
  push        Push a model to a registry
  list        List models
  cp          Copy a model
  rm          Remove a model
  help        Help about any command

Flags:
  -h, --help      help for ollama
  -v, --version   Show version information

Use "ollama [command] --help" for more information about a command.

C:\Users\jerem>
```

Image credits https://www.jeremymorgan.com/blog/generative-ai/ollama-windows/

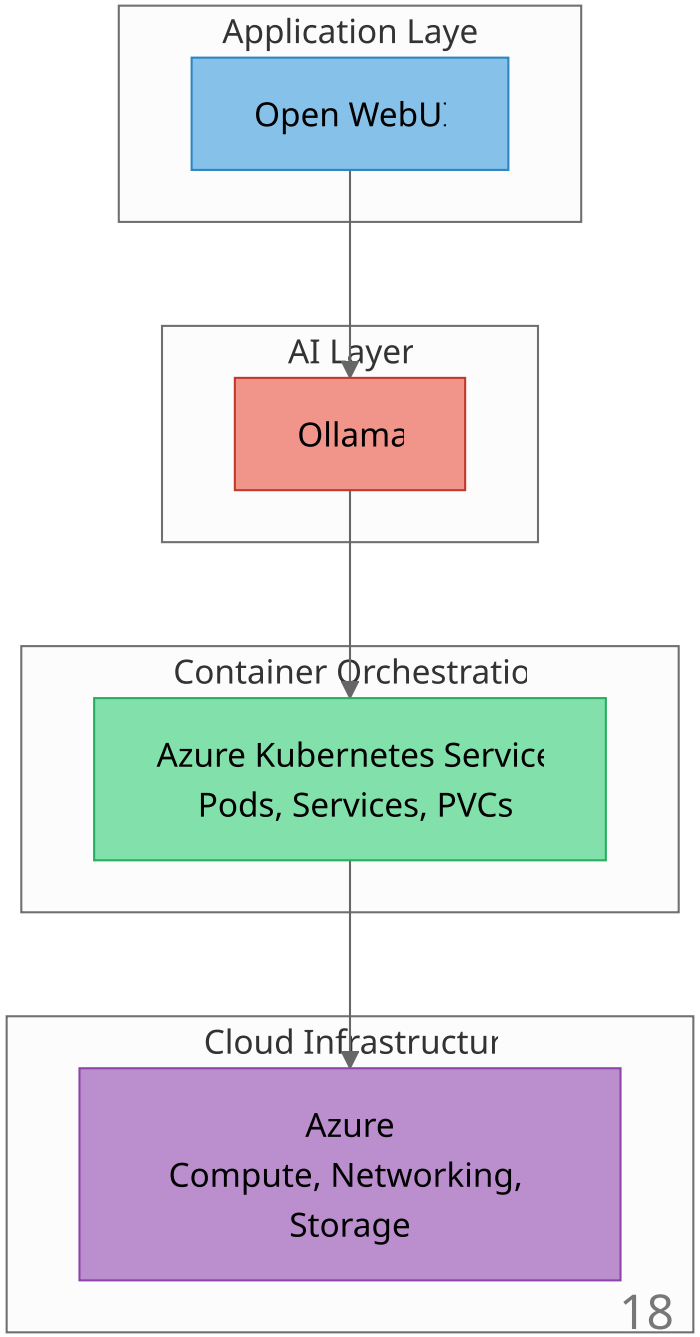
```
C:\Users\jerem>ollama run llava
pulling manifest
pulling 170370233dd5... 100%  4.1 GB
pulling 72d6f08a42f6... 100%  624 MB
pulling 43070e2d4e53... 100%  11 KB
pulling c43332387573... 100%  67 B
pulling ed11eda7790d... 100%  30 B
pulling 7c658f9561e5... 100%  564 B
verifying sha256 digest
writing manifest
removing any unused layers
success
>>> tell me a funny joke about Python?
 Sure! Here's a joke for you:

Why was the Python programmer always cold?

Because he was used to working in low-level languages like C and assembly!

I hope that made you smile. Do you have any other questions I can help with?
```

Image credits https://www.jeremymorgan.com/blog/generative-ai/ollama-windows/

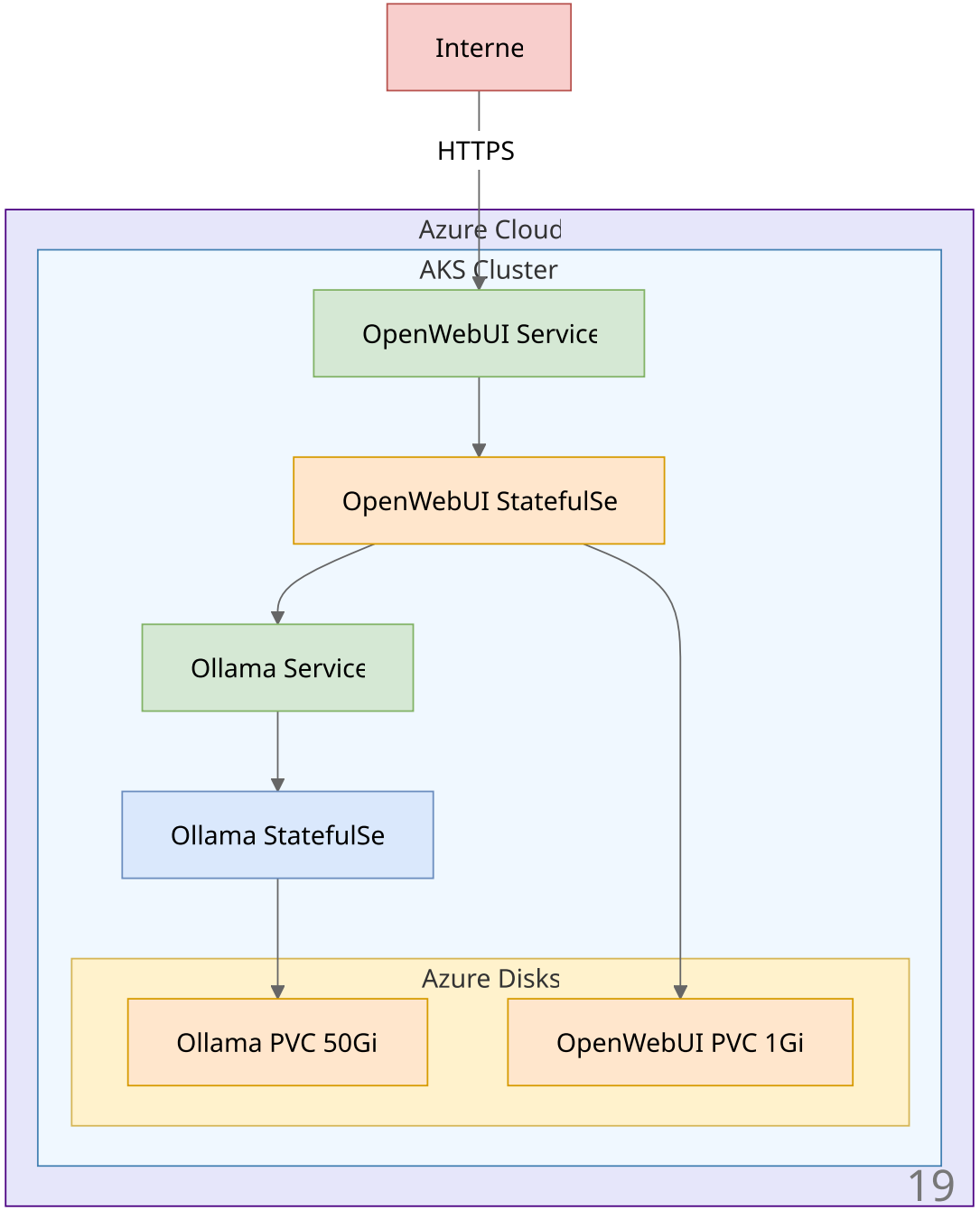
Administrator: Windows PowerShell

```
PS C:\Windows\system32> (Invoke-WebRequest -method POST -Body '{"model":"llava", "prompt":"Tell me a funny joke about
WIndows", "stream": false}' -uri http://localhost:11434/api/generate ).Content | ConvertFrom-json


model                : llava
created_at           : 2024-02-15T19:44:02.1328476Z
response             :  Why did the Windows programmer get arrested?

                       Because he was caught trying to steal some code!
done                 : True
context              : {733, 16289, 28793, 15259...}
total_duration       : 3225812900
load_duration        : 2564599000
prompt_eval_count    : 17
prompt_eval_duration : 122861000
eval_count           : 23
eval_duration        : 537955000



PS C:\Windows\system32>
```

Declarative Infrastructure with OpenTofu and Kubernetes

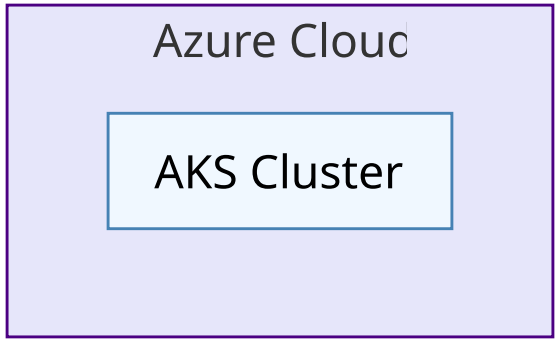


Architecture Overview

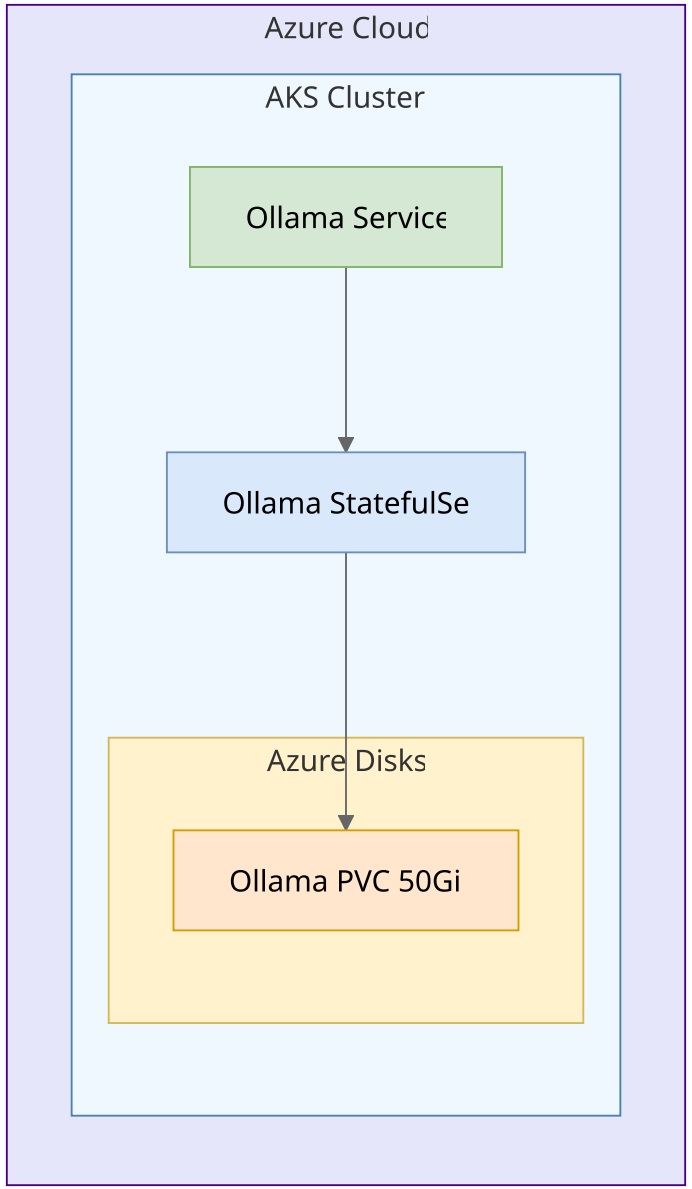


```
resource "azurerm_resource_group" "rg" {
  location = var.resource_group_location
   ...
}

resource "azurerm_kubernetes_cluster" "k8s" {
  resource_group_name = azurerm_resource_group.rg.name

  default_node_pool {
    vm_size             = "Standard_D1_v2"
    node_count          = var.node_count
  }
   ...
}
```

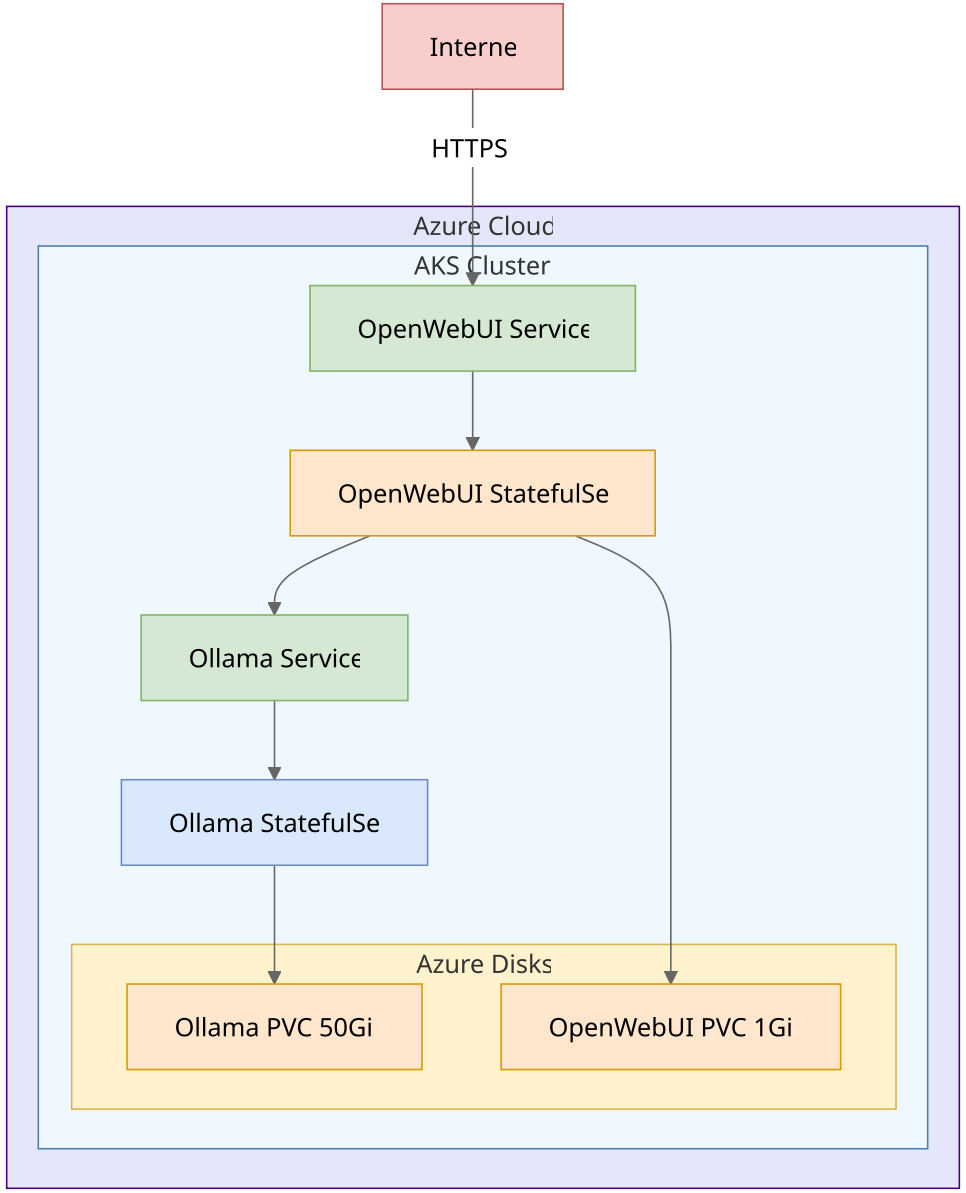
```
❯ tofu apply

OpenTofu will perform the following actions:

  # azurerm_kubernetes_cluster.k7s will be created
  + resource "azurerm_kubernetes_cluster" "k7s" {
      + api_server_authorized_ip_ranges     = (known after apply)
      + current_kubernetes_version          = (known after apply)
      + dns_prefix                          = (known after apply)
       ...
    }

  # azurerm_resource_group.rg will be created
  + resource "azurerm_resource_group" "rg" {
      + id       = (known after apply)
      + location = "westus"
      + name     = (known after apply)
    }
     ...

Plan: 7 to add, 0 to change, 0 to destroy.
```



```yaml
apiVersion: apps/v1
kind: StatefulSet
spec:
  template:
    spec:
      containers:
      - name: ollama
        image: ollama/ollama:latest
        ports:
        - name: http
          containerPort: 11434
  volumeClaimTemplates:
  - spec:
      resources:
        requests:
          storage: 50Gi
---
apiVersion: v1
kind: Service
spec:
  ports:
  - port: 80
    targetPort: http
```



```yaml
apiVersion: apps/v1
kind: StatefulSet
spec:
  template:
    spec:
      containers:
      - name: open-webui
        image: ghcr.io/open-webui/open-webui:main
        env:
          - name: OLLAMA_BASE_URLS
            value: http://ollama:80
        ports:
        - name: http
          containerPort: 8080
  volumeClaimTemplates:
  - spec:
      resources:
        requests:
          storage: 1Gi
---
apiVersion: v1
kind: Service
spec:
  ports:
  - port: 8080
    targetPort: http
```



Step by Step Walkthrough

Infrastructure Setup

```
git clone https://github.com/fmops/azure-private-ai-template
cd tofu && tofu init && tofu apply
```

Application Deployment

```
kubectl apply -f k8s/
```

Done

Just kidding...

Day 2 Operations

- Logging: `kubectl logs`
- Monitoring: Azure Monitor
- Patching: `kubectl edit`
- Scaling: `kubectl scale`
- Disaster Recovery: VolumeSnapshots

Thank you!

Questions?