

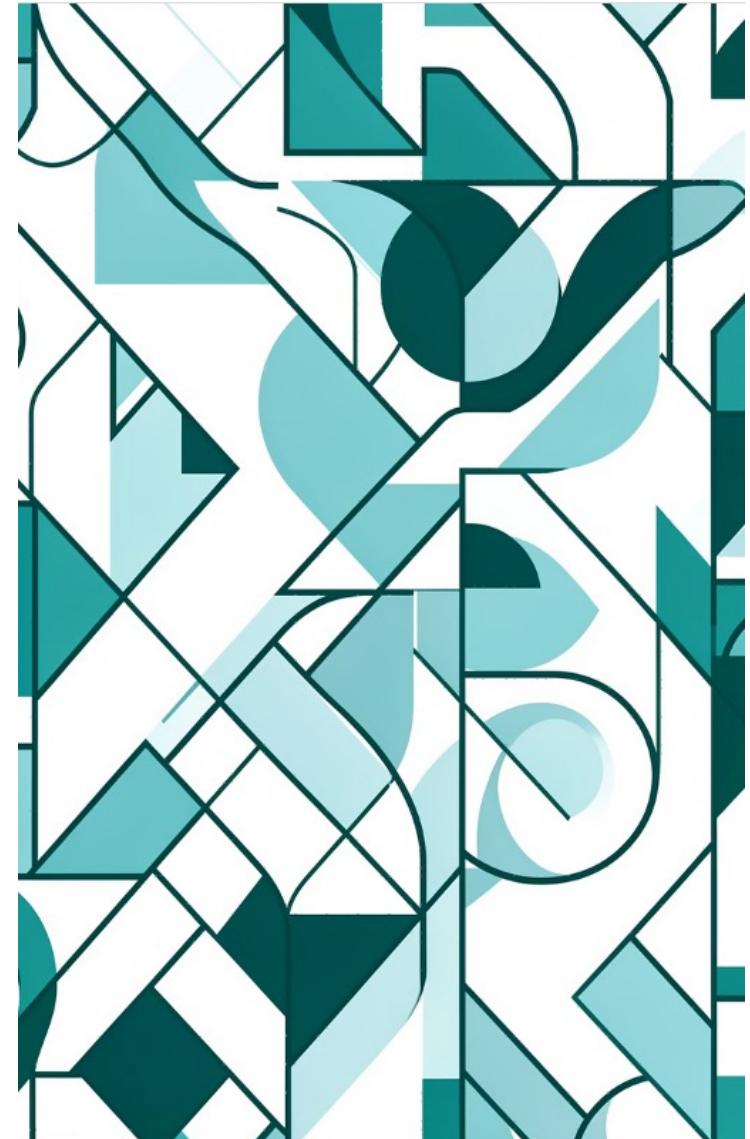


Azure Developer
Community

Own Your AI: Running Open-Source Models on Azure

From Local Experiments to Secure Cloud Deployment

Senthilkumar Srinivasan



About Me

Senthilkumar Srinivasan, Azure Solutions Architect
@ GE Aerospace

- | 18+ years building and running enterprise systems
- | Working on Azure, cloud platforms and applied AI
- | Community speaker



Topics to Cover

 Why Run a Private AI Model on Azure ?

 Local to Cloud Flow

 Architecture Overview

 Live Demo: Private AI Deployment

 Real-World Considerations

 Key Takeaways and Q&A



Why Run a Private AI Model on Azure?

When Managed Services Aren't Enough

Fully managed AI services work well for standard use cases.

Some enterprise scenarios require more control over how AI models are deployed and operated.

This isn't about replacing managed services—it's about owning your infrastructure decisions.



Key Requirements for Private AI



Deployment Control

Control when, where, and how models are deployed



Data Boundaries

Ensure sensitive data stays within your cloud boundary



Version Management

Safely roll back or update model versions



Cost Predictability

Predict and control compute costs at scale

AI models are workloads—deploy and operate them like any other cloud-native service.

Model & Runtime used

For this demo, we use an open-source LLM running via Ollama — a lightweight inference runtime that simplifies local and container-based execution.

Model

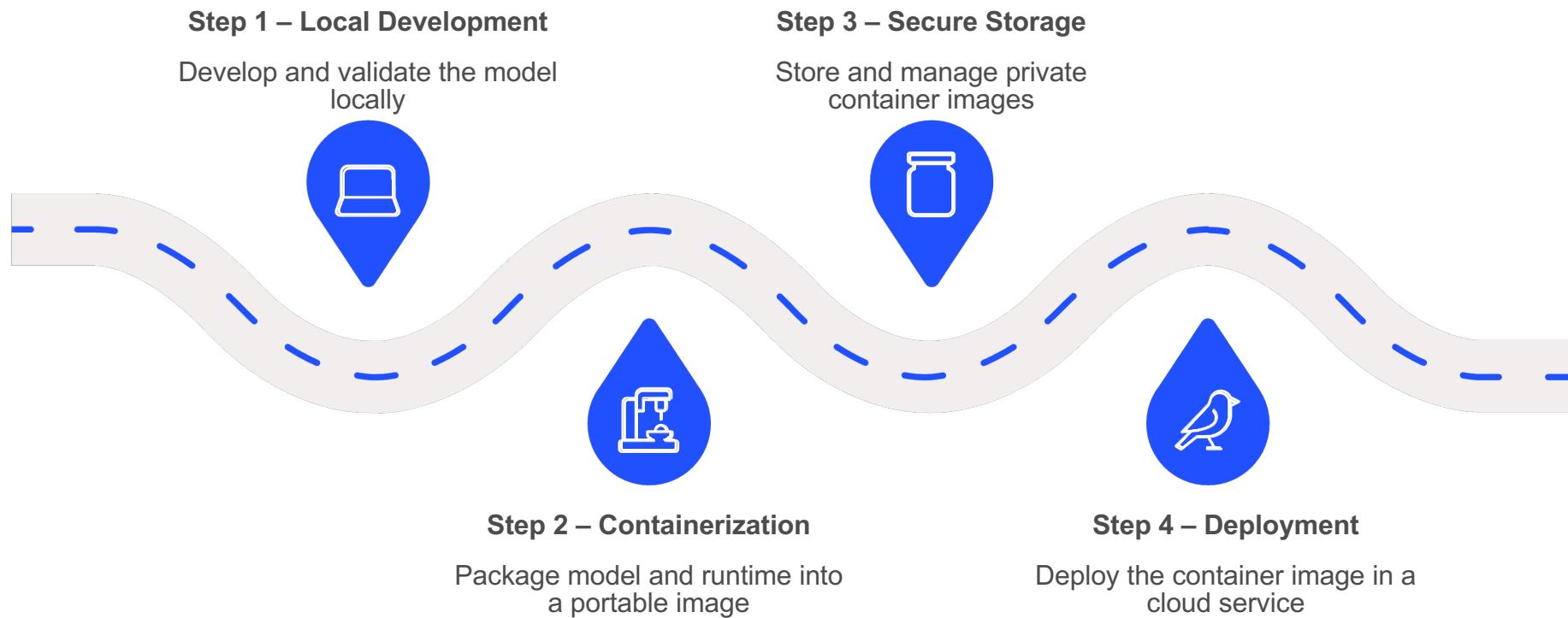
- Llama 3.2 (3B variant)
- ~ 2 GB model size
- Chosen for lightweight deployment

Runtime

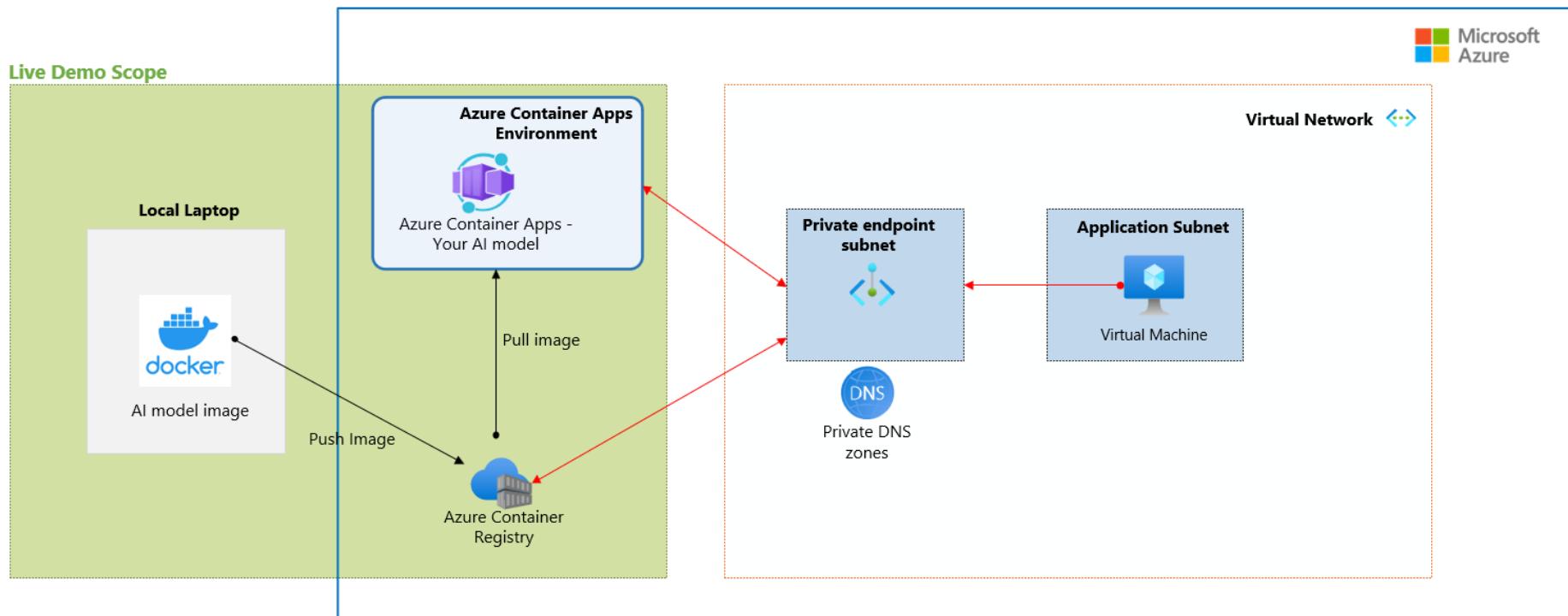
- Ollama packages model + runtime together
- Containerized and deployed to Azure

The deployment pattern works for any open-source or custom model packaged as a container image

From Local Model to Cloud Service



Architecture Overview



Real-World Considerations



Image Size Matters

Large container images increase pull and deployment time



Startup Time Is Real

Cold starts (30–60 seconds) impact first-request latency



Dedicated Profile Behavior

Consumption and Dedicated profiles behave differently under load



Operational Ownership

You're responsible for monitoring, updates, and scaling



Resource Sizing Affects Performance

Memory and CPU allocation directly impact inference latency

Key Takeaways

Private AI on Azure is practical

Open-source models can be securely deployed using Azure-native services like Container Apps and Private Endpoints.

AI models are compute-heavy workloads

Image size, cold starts, CPU, and memory directly affect performance.

Containerization enables consistency from local to cloud

The same image moves from development to Azure production.

Control requires operational ownership

You manage scaling, updates, monitoring, and cost.



Q&A – Own Your AI

❑ Questions? Let's discuss

Connect with me



Senthilkumar Srinivasan

Azure Solutions Architect | Azure 10x Certified | Cloud & AI Platform Leader

