GenAIOps for Cloud Solution Architects

From Prompt to Production

Purpose

Generative AI is easy to demonstrate and hard to operate. GenAIOps is the operating discipline for moving from prompt experiments to safe, governed, cost-aware production services.

Lifecycle

Stage	CSA focus
Build	Prompts, agents, retrieval, orchestration and tools.
Evaluate	Groundedness, relevance, coherence, safety and task success.
Deploy	Managed endpoints, CI/CD, gateway and release controls.
Monitor	System health, answer health, token use, safety events and feedback.
Govern	Identity, RBAC, audit, responsible AI and project isolation.
Optimise	Cost, model choice, context design, caching and continuous improvement.

Azure mapping

Need	Pattern
Foundation models	Azure OpenAI Service or models through Azure AI Foundry.
Project workspace	Azure AI Foundry.
Retrieval	Azure AI Search and approved data sources.
Safety	Azure AI Content Safety and responsible AI process.
Identity	Microsoft Entra ID and RBAC.
Gateway	Azure API Management as an AI gateway.
Observability	Azure Monitor, Application Insights and OpenTelemetry.
Cost	Azure Cost Management, Azure Monitor logs and gateway analytics.

Discovery questions

- What business outcome would make this use case worth scaling?
- Which data sources are trusted enough to ground responses?
- What happens when the AI is uncertain or wrong?
- How will quality be evaluated before release?
- Who supports the solution after it goes live?
- How will usage and cost be tracked?

Pilot checklist

- Named business owner
- Narrow valuable use case
- Approved data sources
- Representative test set
- Quality and safety criteria
- Access control design
- Monitoring plan
- Cost visibility
- Support and feedback process

Public note

This is personal educational content using public-safe examples only. It does not include customer confidential material.