

Sphera Azure AI — Production Cheat Sheet

PoC → Prod in one page · companion to the [Sphera Azure AI Blueprint Handbook](#)

Decision Quick Calls

Decision	Call
Model access	AI Foundry primary · direct = failover
Gateway	APIM (auth, quotas, analytics)
Compute	Container Apps · Functions for ingestion
Vector search	AI Search serving · Databricks = curation
Embeddings	Jina vs Azure OpenAI — let evals decide
Secrets	Key Vault — non-negotiable
Tracing	Managed tracing (Langfuse / Azure GenAI tracing)

Top 3 P0s first

- **Hardcoded secrets** → Key Vault + scan
- **No retry/rate-limit handling** → gateway wrapper
- **No auth/RBAC + injection risk** → Entra + guardrails

Roadmap at a glance

Phase	Weeks	Exit
Foundation	1-2	IaC landing zone deploys from one command
Harden	3-6	CI green + eval baseline recorded
Migrate	7-10	Parallel-run parity ≥ 95%
GA	11-12	Canary → 100% · AIDLC G1-G6 signed

Anti-patterns

- Notebooks as the runtime · `.env` in repo
- `json.loads()` on raw model output
- Hardcoded model names in 20 files
- Deploying without evals or rollback
- Bypassing APIM "just for internal tools"

