Runbook: Azure OpenAI Service – Incident Triage

Purpose

This runbook provides step-by-step guidance for triaging incidents related to Azure OpenAI Service usage, availability, latency, and quota exhaustion.

Symptoms

- HTTP 429 or quota exceeded errors
- Increased latency for completions or embeddings
- Timeouts from downstream applications
- Sudden drop in token throughput

Immediate Checks

- Verify Azure region and resource availability
- Check recent deployments or configuration changes
- Confirm API key or managed identity validity

KQL – Token Usage Investigation

```
AzureDiagnostics
| where ResourceProvider == "MICROSOFT.COGNITIVESERVICES"
| where OperationName contains "Tokens"
| summarize sum(TotalTokens) by bin(TimeGenerated, 5m)
```

Resolution Steps

- Increase TPM quota if sustained load is expected
- Enable client-side retry with exponential backoff
- Throttle non-critical workloads

Escalation

If the issue persists beyond 15 minutes or impacts production clients, escalate to Azure Support with correlation IDs and timestamps.