# Runbook: Observability for Azure OpenAI (Key Metrics + KQL Starter Pack)

Last updated: 2026-01-23

## Purpose

Provide a baseline observability approach for Azure OpenAI usage: latency, errors, token usage, throttling, and top callers.

## Golden Signals

| Signal | Goal | Where |
|---|---|---|
| Latency | P50/P95 stable | App Insights / logs |
| Traffic | Requests per minute | App logs / API gateway |
| Errors | 4xx/5xx ratio low | AzureDiagnostics / App logs |
| Saturation | TPM/quota headroom | Azure OpenAI metrics / logs |

## KQL – Errors by Model

```
AzureDiagnostics
| where ResourceProvider == "MICROSOFT.COGNITIVESERVICES"
| summarize Errors=countif(ResultType != "Success"), Total=count() by ModelName_s,
bin(TimeGenerated, 5m)
| extend ErrorRate = todouble(Errors) / todouble(Total)
```

## KQL – Throttling (429) Trend

```
AzureDiagnostics
| where ResultSignature == "429"
| summarize count() by bin(TimeGenerated, 5m), ModelName_s
```

## Operational Guidance

- Alert when 429 count exceeds baseline or when P95 latency exceeds SLO.
- Segment by model, client app, environment, and region.
- Use dashboards to track token burn rate and predict quota exhaustion.