CDW Documentation

Azure API Monitoring

Azure API Monitoring

Monitoring Azure AI APIs is critical for performance, usage tracking, quota management, and troubleshooting. Azure provides multiple built-in and extensible options to monitor its AI services (like Azure OpenAI, Cognitive Services, and Azure Machine Learning). Here's a breakdown of the available monitoring options:

---

1. Azure Monitor (Primary Platform Monitoring Tool)

Azure Monitor provides a centralized platform for collecting, analyzing, and acting on telemetry from Azure resources.

Key Features:

- **Metrics**: Track request volume, latency, and error rates.
- **Logs (via Log Analytics)**: Ingest and query detailed activity and diagnostics logs.
- **Alerts**: Set up rules to get notified on API failures, high latency, quota breaches, etc.
- **Dashboards**: Build visual dashboards to monitor trends.

---

2. Metrics for Azure AI Services

Each AI service exposes its own set of metrics in Azure Monitor:

Common Metrics:

Metric	Description
Total Calls	Total number of API calls
Successful Calls	Count of HTTP 200 responses
Failed Calls	Count of 4xx/5xx errors
Latency	Response time percentiles (P50, P90, P95, etc.)
Throttled Calls	Requests blocked due to quota limits

You can find these under:

> **Azure Portal → Monitor → Metrics → Select your AI resource**

---

3. Diagnostic Settings

You can configure **Diagnostic Settings** on each Azure AI resource to send logs and metrics to:

- **Log Analytics Workspace**
- **Event Hubs**
- **Azure Storage**

Logs may include:

- Request logs (time, endpoint, status)
- Quota usage
- Custom logs depending on the service

Enable via:

> **Resource → Monitoring → Diagnostic settings**

---

4. Application Insights (Optional for Custom Apps)

If you're calling Azure AI APIs from your own application, you can use **Application Insights** to:

- Track dependency calls to Azure AI APIs
- Monitor end-to-end latency and failures
- View distributed traces and performance bottlenecks

> **Integrates well with web apps, functions, and APIs**

---

5. Quota and Usage Tracking

For services like Azure OpenAI and Cognitive Services:

- **Azure Quotas API**: Track how close you are to usage limits
- **Azure Portal → Usage + Quotas** under the AI service

You can set up alerts when usage approaches or exceeds thresholds.

⚙ 6. Azure Machine Learning (if used)

If you're deploying models via Azure ML:

- **Azure ML Studio** has its own monitoring for:
  - Endpoint request count
  - Response times
  - Status codes
  - Resource utilization (CPU/Memory)

> **Studio → Endpoints → Monitoring tab**

7. Custom Monitoring via API Wrappers

You can build wrappers or proxies around API calls to:

- Log request/response time
- Capture payloads for analysis
- Push logs to Azure Log Analytics or external tools (e.g., Splunk)

---

8. Security and Compliance Monitoring

Use:

- **Microsoft Defender for Cloud**: To monitor security posture
- **Azure Policy**: To enforce tagging, logging, and retention policies
- **Sentinel (SIEM)**: For advanced threat detection and analytics

---

Best Practices

- Enable **Diagnostic Logs** and send to **Log Analytics**
- Configure **Alerts** for latency, error rate, and quota limits
- Use **Workbooks** for visual monitoring dashboards
- Use **App Insights** for client-side monitoring if you build apps that call the APIs