CDW Documentation

Potential Monitoring Plan

Potential Monitoring Plan

Overview

This Azure AI Monitoring and Reporting Plan provides enterprise-grade visibility, accountability, and operational assurance for AI workloads running in Microsoft Azure. Designed for applications leveraging services like Azure OpenAI, Speech, Language, and Cognitive APIs, this plan ensures your AI solutions remain secure, reliable, and performance-optimized.

Goals

- Proactively monitor AI system health, latency, and errors
- Ensure SLA compliance and minimize downtime
- Provide actionable performance and usage insights
- Enable cost transparency and anomaly detection
- Secure sensitive data and track access

Scope of Monitoring

Area	Monitored Item
AI Services	OpenAI (GPT), Speech, Language, Vision, etc.
Performance	Latency, token usage, call frequency, timeouts
Reliability	HTTP error rates, dependency failures
Security	Key Vault access, secret usage, auth events
Cost Tracking	Resource consumption and budget alerts
Custom Metrics	AI response accuracy, sentiment scores, etc.

Implementation Approach

1. **Azure Monitor Integration**
    - Real-time tracking of latency, usage, errors
    - Application Insights or custom telemetry
2. **Diagnostic Settings**
    - Export logs to Log Analytics & Storage
    - Enable audit trails and long-term retention
3. **Alerts and Notifications**
    - Latency thresholds (e.g., > 500ms)
    - Error alerts for failed API calls
    - Custom thresholds for model usage
4. **Dashboards and Visualization**
    - Azure Workbooks with real-time and historical metrics
    - KPI charts for token usage, sentiment trend, etc.
5. **Weekly/Monthly Reports**
    - Performance summaries
    - Anomaly detection insights
    - Recommendations for scaling or tuning

Security & Compliance

- Access monitoring for Key Vault and secrets
- RBAC enforcement on logs and diagnostics
- Optional integration with Microsoft Defender for Cloud

Deliverables

- Baseline monitoring template for all AI services
- Custom alerts and dashboards tailored to client KPIs
- Weekly AI health check report (PDF or Power BI)
- Optional cost optimization and usage analysis

Optional Add-ons

- GPT behavior audits (accuracy, bias, hallucinations)
- Compliance logging for regulated industries (HIPAA, GDPR)
- Integration with ServiceNow, PagerDuty, or Slack for alerting