Amazon Bedrock Support

Overview

AI Monitor uses Amazon Bedrock to provide intelligent features that enhance anomaly detection and monitoring. These features are optional – all core detection, alerting, and correlation functionality works without Bedrock.

AI-Powered Features

- **AI Explain** – Generates natural language explanations for detected anomalies, describing what happened and likely root causes.
- **AI Suggest** – Accepts natural language descriptions of what you want to monitor and generates subscription configurations.
- **AI Tuning** – Analyzes anomaly history and recommends threshold/direction adjustments to reduce noise.
- **Report Summaries** – Generates executive summaries in weekly anomaly reports.

Default Model

AI Monitor uses the following model by default:

```
us.anthropic.claude-haiku-5-5
```

This is a cross-region inference profile that automatically routes requests to the nearest US Bedrock endpoint. It provides fast, cost-effective responses suitable for operational monitoring explanations.

Region Availability

AI features require Amazon Bedrock runtime to be available in the deployment region. The stack creates a VPC endpoint for Bedrock in the deployment region.

Fully Supported Regions (Bedrock + Anthropic Claude)

us-east-1	N. Virginia (recommended)
us-west-2	Oregon (recommended)
eu-west-1	Ireland
eu-west-2	London
eu-west-3	Paris
eu-central-1	Frankfurt
ap-northeast-1	Tokyo
ap-southeast-1	Singapore

ap-southeast-2	Sydney
ap-south-1	Mumbai
ca-central-1	Canada
sa-east-1	São Paulo

Regions Without Bedrock Runtime

us-west-1	N. California
af-south-1	Cape Town
ap-east-1	Hong Kong
ap-northeast-2	Seoul
ap-northeast-3	Osaka
ap-southeast-3	Jakarta
eu-north-1	Stockholm
eu-south-1	Milan
me-south-1	Bahrain
me-central-1	UAE

Enabling Model Access

Before AI features will work, you must enable Anthropic Claude model access in the AWS Console:

1. Navigate to **Amazon Bedrock** in the AWS Console
2. Select **Model access** from the left navigation
3. Click **Manage model access**
4. Enable **Anthropic** (Claude models)
5. Submit the request form
6. Wait for activation (~15 minutes)

Graceful Degradation

If Bedrock is unavailable or model access is not enabled:

- Anomaly detection continues normally
- Alerting and SNS notifications work as expected
- Log correlation functions correctly
- The subscription editor is fully operational
- AI Explain returns no explanation (anomaly still recorded)
- AI Suggest returns "AI is busy, please try again later"
- AI Tuning returns a fallback message
- Reports generate without the AI executive summary
- Send us an email, we may be able to possibly accommodate you with a private offer

A circuit breaker protects against repeated Bedrock failures. After 3 consecutive failures, AI calls are paused for 30 minutes before retrying automatically.

Changing the Model

To use a different Bedrock model (e.g., for non-US regions or to use a different model family):

1. Open the AWS Lambda console
2. Update the `BEDROCK_MODEL_ID` environment variable on both:
   `<stack-name>-detector`
   `<stack-name>-editor`
3. Set the value to a valid model ID for your region

Examples:

```
us.anthropic.claude-haiku-5-5 (US cross-region, default)
eu.anthropic.claude-haiku-5-5 (EU cross-region)
anthropic.claude-3-haiku-20240307-v1:0          (single-region, older)
```

Cost

Bedrock usage is pay-per-token with no upfront commitment:

- AI Explain: ~$0.001 per anomaly explanation
- AI Suggest: ~$0.002 per suggestion request
- AI Tuning: ~$0.001 per tuning recommendation
- Report Summary: ~$0.003 per report

Typical monthly cost for AI features: $0.50–$5.00 depending on anomaly volume and usage of Suggest/Tuning features. The caching layer (AI Explain cache) reduces costs by avoiding duplicate explanations for recurring anomaly patterns.

Verification

Run the quota check script to verify Bedrock availability:

```
check-quota.cmd <tier> <region>
check-quota.sh <tier> <region>
```

The script checks:

- Bedrock runtime availability in the target region
- Anthropic Claude model accessibility
- VPC endpoint compatibility