Runbook: AWS Lambda API Failure – LLM SRE Site

Purpose

This runbook helps diagnose and resolve failures in the LLM SRE Site Lambda-based API.

Common Symptoms

- 5XX errors from API Gateway
- Cold start latency spikes
- Lambda timeout errors

Log Analysis

- Check CloudWatch Logs for recent invocation errors
- Look for timeout or memory limit exceeded messages

CloudWatch Logs Insights Query

```
fields @timestamp, @message
| filter @message like /ERROR/
| sort @timestamp desc
| limit 20
```

Mitigation

- Increase Lambda memory (also increases CPU)
- Review S3 permissions for access denied errors
- Enable provisioned concurrency if needed