AWS Certified Machine Learning Engineer - Associate Exam Guide

# Section 4.1 — Monitor Model Inference (In-Depth)

This section focuses on monitoring deployed ML models in production. It covers drift detection, anomaly monitoring, performance tracking, and AWS tools for observability.

## 1. Monitoring Concepts

• Data Drift → change in input data distribution over time.

• Concept Drift → relationship between input and output changes (model assumptions break).

• Prediction Drift → distribution of predictions changes from expected baseline.

• Model Decay → performance degradation over time due to drift or outdated training data.

## 2. Metrics for Monitoring

• Latency → time per inference request.

• Throughput → number of requests served per second.

• Error rate → failed predictions, timeouts.

• Accuracy metrics → tracked via shadow/A/B testing against labeled data.

## 3. AWS Tools

• SageMaker Model Monitor → detects data quality issues, drift, missing features, schema violations.

• SageMaker Clarify → detects bias and explains model predictions with SHAP values.

• Amazon CloudWatch → collects logs, latency, error rates, and custom metrics.

• SageMaker Debugger (extended) → runtime monitoring of training & inference anomalies.

## 4. Monitoring Strategies

• Schedule monitoring jobs with Model Monitor for real-time or batch analysis.

• Use A/B testing to compare production vs candidate models.

• Shadow testing to monitor new model alongside production.

• Alerts via CloudWatch Alarms when thresholds are breached.

• Retraining triggers when drift/anomalies exceed thresholds.

## 5. Best Practices

• Always baseline data distribution during training for future comparisons.

• Use Clarify for fairness and bias detection in production models.

• Combine CloudWatch logs with Model Monitor metrics for full observability.

• Automate retraining workflows based on monitoring signals.

## ✅ Exam Tips

• Model Monitor = detect drift, bias, and data quality issues.

• Clarify = bias detection and explainability.

• CloudWatch = metrics, logs, alarms.

• Shadow & A/B testing = validate new models before rollout.

• Expect exam scenarios with data drift triggering retraining.