AWS Certified Machine Learning Engineer - Associate Cheat Sheet

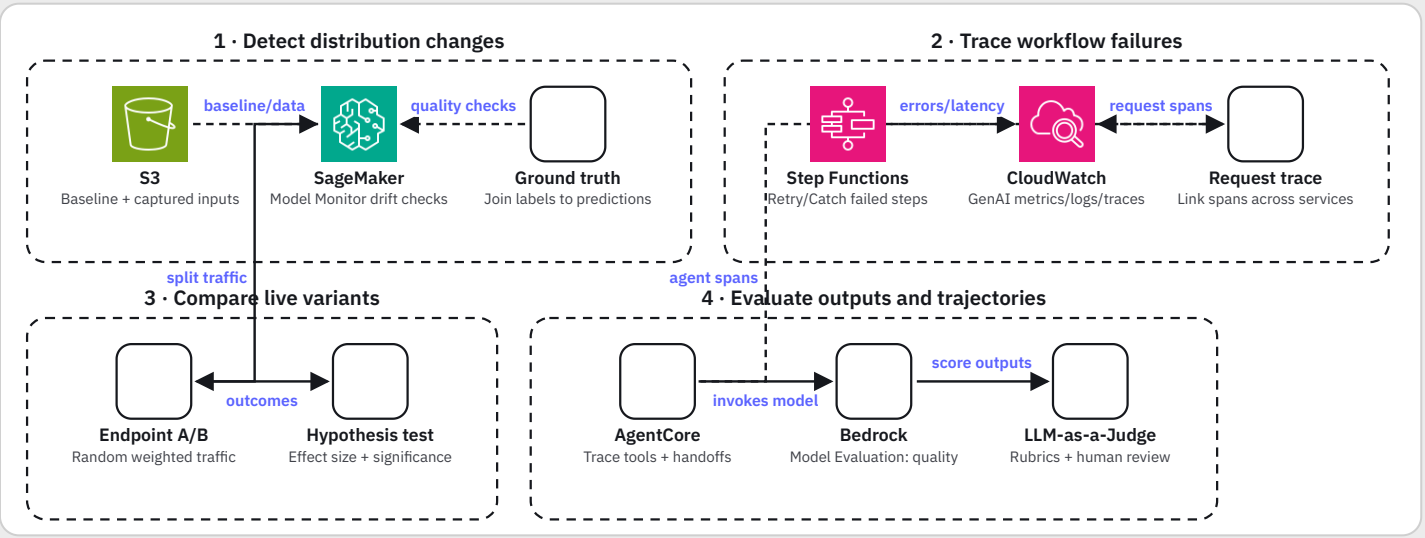
MLA-C02 (beta) · Domain 4 of 4

BlowStack

DOMAIN 4 · OPERATING, MONITORING, AND SECURING ML AND AI SOLUTIONS · 24% OF THE EXAM · TASK STATEMENT 4.1

Monitor ML and AI model inference and performance

Detect drift, trace, compare, evaluate



1 Detect distribution changes

Data drift: schedule Model Monitor comparisons against a baseline in S3. Model-quality monitoring needs ground-truth labels; violations emit metrics, not automatic retraining.

2 Trace workflow failures

Operational anomaly: CloudWatch anomaly detection identifies unusual metric behavior; distributed traces locate failing hops. Step Functions Retry/Catch handles failures only when configured.

3 Compare live variants

A/B testing: use SageMaker production variants with randomized traffic allocation. Compare quality, business outcomes and latency; adequate samples and hypothesis testing distinguish improvement from sampling noise.

4 Evaluate outputs and trajectories

Trajectory evaluation inspects tool calls and handoffs for coordination failures, tool errors and truncated streams. Bedrock evaluations score sampled production outputs; LLM judges need rubrics and human calibration.

Exam traps

- **Data drift is not proof of quality loss**: changed input distributions trigger investigation; labeled outcomes establish model-quality degradation. Stable inputs do not rule out concept drift.
- **Shadow testing is not live A/B testing**: shadow responses do not reach users, so shadow traffic alone cannot measure changes in user behavior or business outcomes.
- **Observability is not remediation**: AgentCore traces expose failures; configured orchestration implements recovery. Low latency and successful tool calls do not prove safe, correct FM answers.

Get practice exams and lessons for AWS Certified Machine Learning Engineer - Associate

AWS Certified Machine Learning Engineer - Associate Cheat Sheet

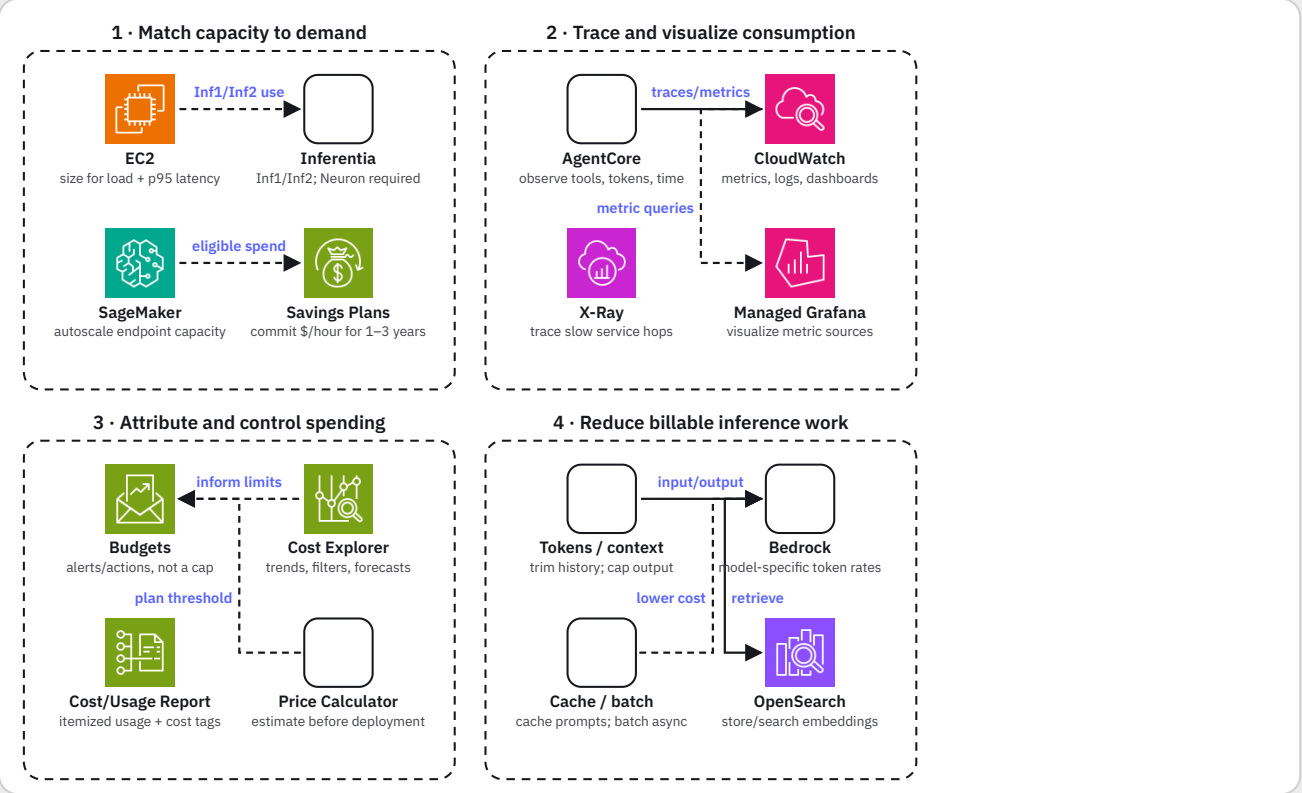
MLA-C02 (beta) · Domain 4 of 4

BlowStack

DOMAIN 4 · OPERATING, MONITORING, AND SECURING ML AND AI SOLUTIONS · TASK STATEMENT 4.2

Optimize and manage ML and AI infrastructure costs and performance

Right-size, trace, allocate, economize



1 Match capacity to demand

Use **right-sizing** for measured latency and utilization; autoscaling changes capacity, not hardware speed. Multi-AZ redundancy supports failover. Commit predictable baseline usage; keep bursts flexible.

2 Trace and visualize consumption

Use **distributed tracing** to isolate slow model/tool hops; correlate latency, errors, utilization, and tokens per task. Instrument telemetry first. Amazon Managed Service for Prometheus stores container metrics.

3 Attribute and control spending

Activate **cost allocation tags** for owner, application, and environment reporting. Forecast spending, then configure budget thresholds and actions; tags alone neither enforce quotas nor stop resources.

4 Reduce billable inference work

Repeated prompts favor caching; offline jobs favor batch. Compare token charges with hosted-instance TCO. Reuse embeddings; vector quantization saves memory/storage but can reduce recall.

Exam traps

- **Budget ≠ hard spending cap**: alerts notify; configured actions affect supported resources or permissions. Neither guarantees that all charges stop at the threshold.
- **Discount ≠ reserved capacity**: Savings Plans discount eligible usage, not capacity guarantees. Compute Savings Plans do not cover SageMaker; SageMaker Savings Plans are separate.
- **One task ≠ one model call**: agent loops, retries, and tool results can multiply token usage. Tokens are not words; include embeddings, vector storage, and idle hosting in TCO.

Get practice exams and lessons for AWS Certified Machine Learning Engineer - Associate

AWS Certified Machine Learning Engineer - Associate Cheat Sheet

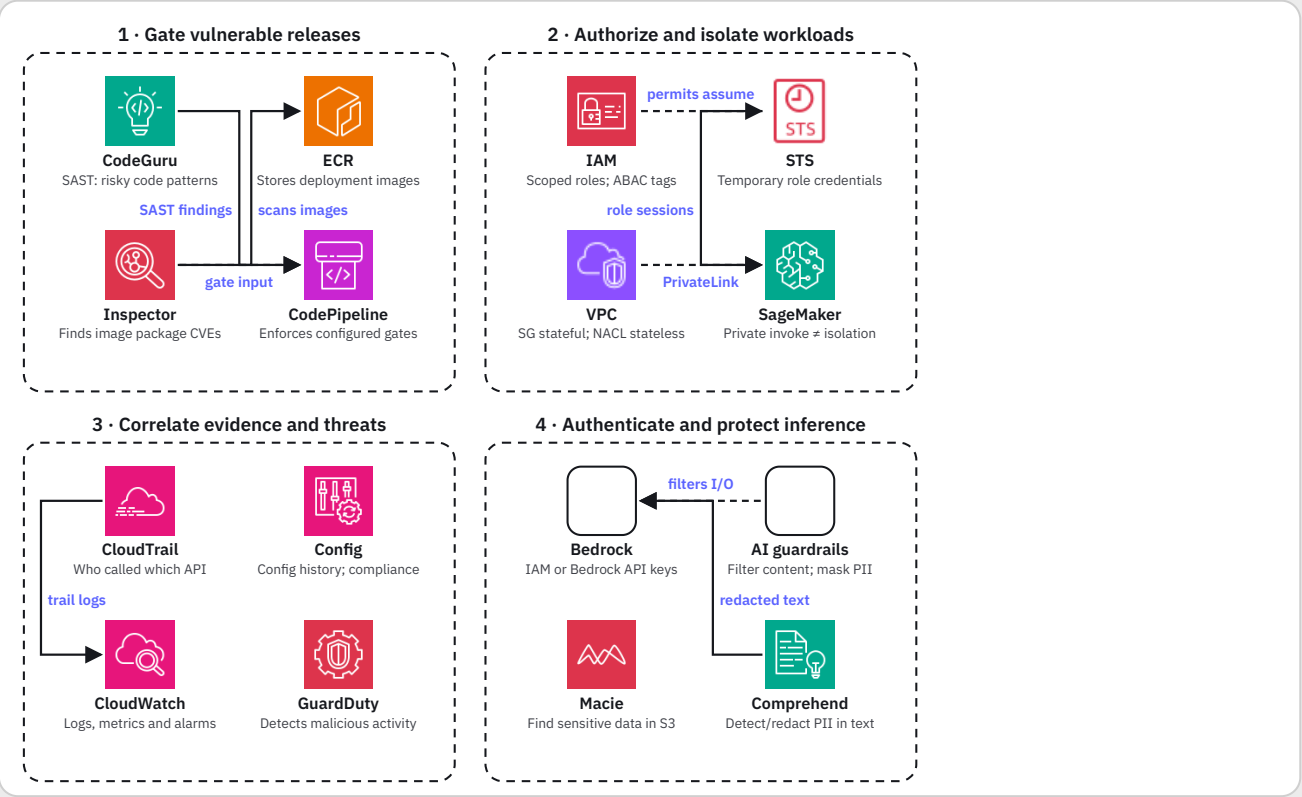
MLA-C02 (beta) · Domain 4 of 4

BlowStack

DOMAIN 4 · OPERATING, MONITORING, AND SECURING ML AND AI SOLUTIONS · TASK STATEMENT 4.3

Secure ML and AI workloads and model endpoints

Scan, authorize, isolate, audit, safeguard



1 Gate vulnerable releases

Security quality gates turn findings into release decisions. Configure pipeline checks to fail or require approval at the chosen severity; Inspector scanning alone does not block deployment.

2 Authorize and isolate workloads

Least privilege scopes workload roles to S3 artifacts/actions; ABAC uses trusted tags. Cross-account role assumption needs trust plus caller permission. Encrypted artifacts also require KMS authorization.

3 Correlate evidence and threats

AccessDenied investigation correlates CloudTrail identity/API errors with CloudWatch logs, then checks effective permissions, VPC paths and KMS policies. Set log retention; Config rules detect configuration drift.

4 Authenticate and protect inference

FM credentials: prefer workload roles with temporary credentials; manage API-key expiry, rotation and revocation. Bedrock Guardrails must be applied to inference; content filtering does not grant or restrict IAM access.

Exam traps

- **Explicit deny wins** over an allow. Adding IAM allows cannot override an applicable explicit deny in an SCP, permissions boundary, session policy, endpoint policy or resource policy.
- **PrivateLink is not network isolation**: private endpoint access does not replace IAM. SageMaker network isolation blocks container network calls; required dependencies must be staged.
- **Guardrails do not prevent data poisoning**. Threat modeling must cover training and retrieval sources; restrict writes, verify provenance and validate data before ingestion.

Get practice exams and lessons for AWS Certified Machine Learning Engineer - Associate

AWS Certified Machine Learning Engineer - Associate Cheat Sheet

MLA-C02 (beta) · Domain 4 of 4

BlowStack

DOMAIN 4 · OPERATING, MONITORING, AND SECURING ML AND AI SOLUTIONS · WHAT TO REMEMBER

Concepts and services of Operating, Monitoring, and Securing ML and AI Solutions

Concepts to know

Model Drift

Production behavior can diverge from validation results. **Data drift alone does not prove quality loss**; compare input distributions with baselines, then use labeled outcomes to assess performance before deciding to retrain.

Right-Sizing

Right-sizing matches resources to measured demand and latency requirements. **Select capacity before purchasing discounts**; scale for variable traffic and commit only to predictable usage.

Least privilege

Least privilege limits actions, resources, and access conditions to workload needs. **An explicit deny overrides an allow**; investigate identity, resource, endpoint, and key policies rather than adding broader permissions.

Core services



Amazon CloudWatch

Use metrics, logs, dashboards, and alarms for **operational health and automated responses**. Latency and error alarms detect serving failures; model quality requires evaluation signals.



Amazon SageMaker

Use Model Monitor for **baseline-based drift monitoring**, production variants for live A/B comparisons, and endpoint auto scaling to match serving capacity to demand.



Amazon Bedrock

Use evaluations to compare FM quality and Guardrails to filter configured content risks. **Optimize token usage, caching, and batch inference** according to latency and workload requirements.

Get practice exams and lessons for AWS Certified Machine Learning Engineer - Associate