# AIF-C01 / AWS Certified AI Practitioner



**65 questions · 90 minutes · 50 scored + 15 unscored · Pass 700/1000 · No penalty for guessing**

## 1.1 / AI & ML foundations · 20% domain

| | |
|---|---|
| **AI → ML → DL** | AI is the broad field; ML learns from data; deep learning uses multilayer neural networks. |
| **Supervised** | Known labels; classification predicts categories; regression predicts numbers. |
| **Unsupervised** | No target labels; clustering discovers groups; anomaly detection spots unusual behavior. |
| **Reinforcement** | Actions → environment → rewards; optimize cumulative reward. |
| **Training / inference** | Learn model parameters / apply the trained model to new inputs. |
| **Data** | Tabular = rows/columns; time series = ordered observations; text, images and audio often unstructured. |

## 1.1 / Choose an inference mode

| | |
|---|---|
| **Real-time** | Sustained requests; predictable low latency; hosted endpoint. |
| **Serverless** | Intermittent synchronous calls; no instance management; tolerate cold starts. |
| **Asynchronous** | Queue large/long requests; retrieve results later from S3; can scale to zero. |
| **Batch** | Whole dataset available upfront; offline results; no persistent endpoint. |
| **Deciding detail** | Async and serverless both reduce idle costs; output timing and payload distinguish them. |

## 1.2 / Managed service selection

| | |
|---|---|
| **Transcribe / Polly** | Speech → text / text → speech. |
| **Translate** | Text between languages; custom terminology for selected terms. |
| **Comprehend** | Entities, sentiment, key phrases and supported PII in text. |
| **Textract** | Scanned documents → text, forms, tables. |
| **Rekognition** | Images/video → labels, faces, text and moderation. |
| **Lex** | Conversational interface; intents and slots. |
| **Personalize** | Personal recommendations from interaction data. |
| **Rules vs ML** | Exact published rules → deterministic logic; learned predictions → ML; open-ended generation → FM. |

## 1.3 / ML lifecycle & tools

| | |
|---|---|
| **Lifecycle** | Business goal → data → preparation → training → evaluation → deployment → monitoring; iterate. |
| **Ground Truth** | Training labels and human labeling workforce. |
| **Data Wrangler** | Visual data preparation and quality insights. |
| **Feature Store** | Reusable features for training and inference. |
| **Pipelines / Registry** | Workflow orchestration / model versions and approvals. |
| **Model Monitor** | Production data/model quality; supported bias and attribution drift. |
| **JumpStart** | Discover, deploy and customize supported pretrained models. |
| **Split data** | Train learns; validation selects/tunes; holdout test gives final assessment; avoid leakage. |

## 2.1 / GenAI foundations · 24% domain

| | |
|---|---|
| **Foundation model** | Broad pretraining; adaptable to multiple downstream tasks. |
| **Tokens** | Text pieces; input and output token counts affect cost and context limits. |
| **Embeddings** | Numeric representations for semantic similarity and retrieval. |
| **Multimodal** | Supported combinations of text, images, audio or video vary by model. |
| **Hallucination** | Plausible unsupported content; grounding reduces risk, never guarantees truth. |
| **Agentic AI** | Goal-directed planning and tool use with observations and iteration. |
| **Short/long memory** | Within-session context / remembered information across sessions. |

## 2.1 / AgentCore & agent building

| | |
|---|---|
| **Runtime** | Managed hosting and isolated agent sessions. |
| **Memory** | Short-term events and long-term extracted context. |
| **Gateway** | Expose supported APIs/Lambda as tools through MCP. |
| **Identity** | Agent and user credentials for service access. |
| **Policy** | Deterministic authorization of tool calls. |
| **Observability** | Traces and operational metrics. |
| **Code / Browser** | Isolated code execution / browser interaction. |
| **Strands / A2A** | Agent SDK / protocol for agent-to-agent communication. |

## 2.2 / Value, limitations & selection

| | |
|---|---|
| **Fit the task** | Modality, latency, context window, quality, cost, license, privacy and Region requirements. |
| **Limitations** | Hallucinations, bias, cutoff, variable output and limited interpretability. |
| **Business value** | Task completion, customer satisfaction, cost per interaction and conversion rate. |
| **Human review** | Use for high-impact or low-confidence outputs; define escalation and accountability. |
| **Small vs large** | Smaller can reduce cost/latency; measure quality on representative tasks before choosing. |

## 2.3 / AWS GenAI technologies

| | |
|---|---|
| **Bedrock** | Managed access to FMs and application building capabilities. |
| **SageMaker AI** | Custom ML training, deployment and lifecycle control. |
| **Nova** | AWS model family; match model capabilities to modality and task. |
| **Kiro / Transform** | Agentic development / modernization of existing workloads. |
| **Amazon Quick** | AI-assisted business analysis and insights. |
| **On-demand** | Variable workload; usage-based inference without provisioned capacity commitment. |
| **Provisioned** | Reserved throughput for suitable steady workloads; model support matters. |
| **Batch / caching** | Delay-tolerant bulk inference / reuse eligible repeated prompt prefixes. |

## EXAM TRAPS / First-pass checks

| | |
|---|---|
| **Service names** | Choose by required capability, not the newest or most powerful name. |
| **No guarantee** | Low temperature, RAG and guardrails do not guarantee correct answers. |
| **Shared duty** | Managed infrastructure does not remove customer access, data and application responsibilities. |
| **Recent changes** | Learn guide objectives; avoid making availability notices or dates the deciding fact. |
| **Reported emphasis** | Responsible AI, service selection and governance lead research signals; anecdotes are not official weights. |

**65 questions · 90 minutes · 50 scored + 15 unscored · Pass 700/1000 · No penalty for guessing**

## 3.1 / FM applications · 28% domain

| | |
|---|---|
| **RAG ingestion** | Prepare/chunk documents → embed → index/store → sync changes. |
| **RAG inference** | Query → retrieve relevant context → augment prompt → generate with citations. |
| **Vector search** | Semantic similarity; hybrid combines semantic and keyword retrieval. |
| **Metadata filters** | Restrict retrieval by document attributes before or during search. |
| **Reranking** | Reorder retrieved candidates for relevance; not model fine-tuning. |
| **Access controls** | Protect both source retrieval and output; citations do not grant access permission. |
| **Agents** | Use tools for actions; knowledge retrieval alone is not autonomous execution. |

## 3.2 / Prompt engineering

| | |
|---|---|
| **Prompt parts** | Instruction, context, input, output format; delimit trusted instructions from untrusted data. |
| **Zero / one / few** | No examples / one example / several examples in context. |
| **Chain of thought** | Structured reasoning technique; do not confuse prompt instructions with retraining. |
| **Chaining** | Sequence prompts and pass intermediate results. |
| **Temperature** | Lower usually more predictable; higher more variable. |
| **Top K / top P** | Top K candidates / cumulative probability nucleus. |
| **Max tokens / stop** | Output length limit / sequences that end generation. |
| **Injection** | Direct user instruction or indirect retrieved content tries to override intended behavior. |

## 3.3 / Customization choices

| | |
|---|---|
| **Prompting** | Fast task instructions; no weight updates. |
| **RAG** | Current/private factual context at inference; no weight updates from retrieval. |
| **Fine-tuning** | Adapt behavior with training examples; changes model weights. |
| **Continued pretrain** | Additional domain corpus; adapt domain language and patterns. |
| **Distillation** | Large teacher → smaller student; trade quality against cost/latency. |
| **LoRA** | Parameter-efficient adaptation; train a smaller set of parameters. |
| **RLHF / DPO** | Preference alignment from human or preference data. |
| **Data quality** | Representative, permitted, clean data; separate train/validation/test and evaluate regressions. |

## 3.4 / Evaluation metrics

| | |
|---|---|
| **Precision** | TP / (TP + FP); minimize false alarms. |
| **Recall** | TP / (TP + FN); minimize missed positives. |
| **F1 / accuracy** | Balance precision + recall / overall correct fraction; imbalance can mislead. |
| **MAE / RMSE** | Absolute numeric error / stronger penalty for large errors. |
| **BLEU / ROUGE** | Translation-oriented n-gram precision / summary-reference overlap. |
| **BERTScore** | Contextual embedding similarity; not factuality proof. |
| **Perplexity** | Next-token predictive uncertainty; compare under compatible evaluation conditions. |
| **Human / LLM judge** | Human expertise / scalable rubric scoring; calibrate and check judge bias. |

## 3.4 / Evaluate the entire application

| | |
|---|---|
| **Context relevance** | Are retrieved passages relevant to the query? |
| **Context coverage** | Do passages contain the required reference information? |
| **Faithfulness** | Are answer claims supported by retrieved passages? |
| **Correctness** | Is the answer accurate? A faithful answer can repeat an incorrect source. |
| **Completeness** | Does it address all parts of the request? |
| **Agents** | Tool selection, argument accuracy, task/goal success, harmfulness. |
| **Business** | Measure cost, latency, satisfaction and completed user tasks alongside model quality. |

## 4.1 / Responsible AI · 14% domain

| | |
|---|---|
| **Fairness** | Avoid harmful group disparities; evaluate representative subgroups. |
| **Explainability** | Understand why the system produced an outcome. |
| **Privacy & security** | Protect personal information and system/data access. |
| **Safety** | Reduce harmful outputs and consequences. |
| **Controllability** | Monitor, guide and intervene in system behavior. |
| **Veracity / robustness** | Truthfulness and reliable behavior under variation or adversity. |
| **Governance** | Ownership, policies, review and accountability across the lifecycle. |
| **Transparency** | Disclose intended use, limitations and AI involvement. |
| **Clarify / A2I** | Bias/feature attribution / human review of supported predictions. |

## 4.1 / Guardrails distinctions

| | |
|---|---|
| **Content filters** | Harmful content categories and supported prompt attacks. |
| **Denied topics** | Block topic categories, not merely literal words. |
| **Word filters** | Block configured terms and supported profanity. |
| **Sensitive info** | Detect/block/mask supported PII and configured patterns. |
| **Grounding** | Check answer grounding and relevance to supplied context. |
| **Automated Reasoning** | Validate against formal policy rules; findings need application handling. |
| **Layer defenses** | Guardrails complement IAM, tool authorization, retrieval permissions and evaluation. |

## 4.2 / Explainability & documentation

| | |
|---|---|
| **Interpretability** | Understand internal workings directly. |
| **Local / global** | Explain one prediction / aggregate model behavior. |
| **SHAP** | Feature attributions; association is not proof of causation. |
| **Model Cards** | Your model: intended use, risk, training, evaluation. |
| **AI Service Cards** | AWS-published service uses, limitations and responsible design. |
| **Tradeoffs** | Explainability vs privacy; transparency vs security/IP; model complexity vs interpretability. |
| **Human-centered** | Present useful explanations to the actual decision-maker; keep escalation paths. |

## 5.1 & 5.2 / Security & governance · 14%

| | |
|---|---|
| **IAM / KMS** | Least-privilege resource access / encryption key control. |
| **PrivateLink** | Private service connectivity; separate from encryption and authorization. |
| **Macie / Inspector** | Sensitive data in S3 / supported software vulnerabilities. |
| **CloudTrail** | Who called which AWS API and when. |
| **CloudWatch** | Operational logs, metrics and alarms. |
| **Config / Artifact** | Resource compliance rules / AWS compliance reports. |
| **Pseudonymize** | Re-identification mapping remains; protect separately. |
| **Anonymize / DP** | Irreversible de-identification / calibrated privacy noise. |
| **Governance** | Data lineage, approvals, retention, residency, access and incident response. |
| **ISO 42001 / NIST** | AI management-system standard / voluntary AI risk framework. |
| **Security scopes** | Consumer app → enterprise app → pretrained → fine-tuned → self-trained; increasing customer ownership. |
| **FINAL TRAP** | 700 scaled ≠ 70% raw. Review all unanswered items; no guessing penalty. |