AWS Certified AI Practitioner (AIF-C01)

Quick Exam Refresher

*This is your **condensed, high-impact review guide** for the AWS Certified AI Practitioner exam. Use it for **quick recall** right before test time — not for deep study. It's structured to help you remember key facts, concepts, and AWS services even if you forget details.*

AI Practitioner (AIF-C01) Domains

Each domain is weighted differently. Foundation models and GenAI make up over half of the test.

- **Domain 1:** Fundamentals of AI and ML (20%)
- **Domain 2:** Fundamentals of Generative AI (24%)
- **Domain 3:** Applications of Foundation Models (28%)
- **Domain 4:** Guidelines for Responsible AI (14%)
- **Domain 5:** Security, Compliance, and Governance (14%)

Quick Reminder: How the Exam Works

- **Number of Questions:** 65
- **Format:** Multiple choice + multiple response
- **Time Limit:** 90 minutes
- **Passing Score:** 700/1000
- **Test Provider:** Pearson VUE (online or onsite)

Remember — You Don't Need to Be Perfect to Pass

The passing score is **700/1000**, meaning you can miss around **15–20 questions** and still pass. Focus on understanding core GenAI concepts, AWS model services, responsible AI practices, and how to match AWS tools to use cases.

Domain 1: Fundamentals of AI and ML (20%)

Key Concepts

- **AI** = Machines mimicking human intelligence (e.g., decision-making, speech recognition)
- **ML** = Subset of AI; uses data to train models that make predictions
- **Deep Learning** = ML using neural networks; good with unstructured data like images or text

Learning Types

- **Supervised Learning**: Labeled data → predictions (classification/regression)
- **Unsupervised Learning**: No labels → find patterns (clustering, dimensionality reduction)
- **Reinforcement Learning**: Agent learns via rewards/punishments (trial-and-error)

ML Lifecycle

1. Problem definition
2. Data collection & preparation
3. Model training (using SageMaker or similar)
4. Evaluation (accuracy, precision, F1, AUC)
5. Deployment (real-time via SageMaker Endpoint or batch)
6. Monitoring (SageMaker Model Monitor)

Core AWS AI/ML Services

- **Amazon SageMaker**: Build, train, deploy custom ML models
- **Amazon Rekognition**: Analyze images/videos (face detection, labels, moderation)
- **Amazon Lex**: Build voice/text chatbots (same tech as Alexa)
- **Amazon Comprehend**: NLP service – sentiment, key phrases, entities
- **Amazon Transcribe**: Convert speech to text
- **Amazon Polly**: Convert text to lifelike speech
- **Amazon Translate**: Language translation

Important Terms

- **Training vs. Inference**: Learn model vs. use model
- **Structured vs. Unstructured Data**
- **Batch vs. Real-Time Inference**
- **Model Bias**: Skewed predictions due to biased data

Domain 2: Fundamentals of Generative AI (24%)

Key Concepts

- **Generative AI**: Produces new content (text, images, audio, code)
- **Foundation Models**: Large pre-trained models for many tasks (e.g., LLMs)
- **Tokens**: Pieces of input text processed by the model
- **Embeddings**: Vector representation of data for similarity searches
- **Transformer**: Neural network architecture behind LLMs (e.g., GPT, BERT)
- **LLM**: Large Language Model – generates human-like text

Generative Use Cases

- Text summarization, translation, image generation (Stable Diffusion)
- Chatbots, Q&A assistants, creative writing, customer service
- Code generation (e.g., CodeWhisperer, Codex)
- Personalized content

Generative AI Lifecycle

- Model selection → data collection → fine-tuning (optional) → evaluation → deployment → feedback

Advantages

- Adaptable to many tasks
- Fast response generation
- Natural interaction (chat-based UI)
- High productivity and content generation

Disadvantages

- **Hallucinations**: Generates false or made-up information
- **Non-determinism**: Outputs vary with each run
- **Interpretability**: Hard to explain how it works
- **Cost**: High resource use, token-based pricing

AWS Generative AI Services

- **Amazon Bedrock**: Access foundation models via API (Titan, Claude, Jurassic, Stable Diffusion)
- **SageMaker JumpStart**: Pre-built models and templates
- **Amazon PartyRock**: No-code generative AI app builder (based on Bedrock)
- **Amazon Q**: Generative AI assistant (chat, business knowledge)

Domain 3: Applications of Foundation Models (28%)

Key Concepts

- **Prompt Engineering**: Crafting effective prompts for better responses
- **Few-shot Prompting**: Giving examples in the prompt
- **Chain-of-Thought Prompting**: Asking model to think step-by-step
- **System vs. User Prompt**: Role definition to shape response behavior
- **Agents**: LLMs coordinating multi-step tasks (e.g., “Agents for Bedrock”)

RAG (Retrieval-Augmented Generation)

- Combines LLMs + external data sources (docs, KBs)
- Steps: Store embeddings → search vector DB → feed context to LLM
- Reduces hallucinations, supports custom knowledge

Vector Databases on AWS

- **Amazon OpenSearch (k-NN)**: Primary vector store
- **RDS/Aurora PostgreSQL (pgvector)**: Store embeddings
- **Amazon Neptune**: Graph database with embedding support
- **Amazon DocumentDB**: Basic vector use possible

Model Customization Options

- **Prompting**: Fast, no training
- **RAG**: Augment with external info
- **Fine-tuning**: Adjust model weights on custom data
- **Pre-training**: Full model training (rare, costly)

Fine-Tuning Techniques

- **Instruction tuning**: Makes model better at following commands
- **Domain adaptation**: Train model on industry-specific data
- **RLHF**: Human preference feedback to refine outputs

Evaluation Metrics

- **ROUGE**: For summarization
- **BLEU**: For translation
- **BERTScore**: Semantic similarity
- **Human Eval**: For output quality

Domain 4: Guidelines for Responsible AI (14%)

Responsible AI Features

- **Fairness & Bias Mitigation**: Avoid discrimination
- **Inclusivity**: Serve diverse users
- **Transparency & Explainability**: Understand and justify model decisions
- **Veracity**: Ensure truthful outputs
- **Safety & Robustness**: Avoid harmful or broken outputs

AWS Tools

- **SageMaker Clarify**: Detect bias, explain predictions
- **Model Cards**: Document model use, training, limitations
- **Bedrock Guardrails**: Filter unsafe/inappropriate content
- **Amazon A2I**: Human review workflow for sensitive predictions

Legal & Ethical Risks

- Copyright violations, privacy breaches
- Bias in outcomes (e.g., hiring, lending)
- Lack of user consent, hallucinations

Key Practices

- Use diverse, balanced, curated datasets
- Track data sources, permissions, consent
- Use explainable models where possible
- Allow human-in-the-loop for sensitive cases

Domain 5: Security, Compliance, and Governance (14%)

AWS Security Basics

- **IAM**: Role-based access control (least privilege)
- **Encryption**: In transit (TLS) + at rest (KMS, S3 encryption)
- **PrivateLink**: Keep traffic within AWS VPC
- **Macie**: Scan S3 for sensitive data (PII)

Secure Model Access

- SageMaker/Bedrock endpoints must use IAM controls
- No public endpoints for sensitive data
- Validate inputs to avoid prompt injection

Compliance & Monitoring Tools

- **AWS Config**: Detect non-compliant configurations
- **Amazon Inspector**: Scan EC2, containers for vulnerabilities
- **AWS Audit Manager**: Track audit readiness and controls
- **AWS Artifact**: Download AWS compliance certifications
- **CloudTrail**: Log all actions (who accessed what)

Governance Practices

- **Model Cards**: Document model behavior and limits
- **Data lifecycle policies**: Retain, delete, or archive data
- **Scoping Matrix**: Classify risk level of generative AI use
- **Transparent design**: Show users when AI is in use

Privacy Enhancing Techniques

- Data anonymization
- Differential privacy
- Federated learning
- Limit model access to PII

Core AWS Services You Must Know

(High Priority – Frequently Tested)

Service	Key Info You Must Know
Amazon SageMaker	Full ML lifecycle: build, train, tune, deploy models. Used for custom ML. Supports JumpStart & fine-tuning.
Amazon Bedrock	Serverless access to foundation models (LLMs, diffusion). Supports customization, RAG, and agents.
Amazon Rekognition	Computer vision: label detection, face comparison, celebrity ID, unsafe content detection.
Amazon Lex	Build chatbots/voice bots with speech-to-text and intent recognition.
Amazon Comprehend	NLP: sentiment analysis, key phrases, entities, language detection, topic modeling.
Amazon Transcribe	Automatic speech-to-text. Supports streaming and custom vocabulary.
Amazon Polly	Text-to-speech. Produces lifelike voices. Multiple languages and formats (MP3, OGG).
Amazon Translate	Neural machine translation for many language pairs. Real-time or batch.
SageMaker Clarify	Bias detection, explainability (e.g., SHAP), fairness reporting, pre- and post-training analysis.
SageMaker Model Monitor	Tracks model drift, quality, and bias in production.
SageMaker JumpStart	Deploy pre-trained models and solution templates easily (e.g., Stable Diffusion, text classifiers).
Amazon PartyRock	No-code generative AI app builder based on Bedrock. Great for experimentation and prototyping.
Amazon Q	Generative AI assistant for business data Q&A and internal app creation.

Other AWS Services You Should Know

(Moderate Priority – Occasionally or Partially Tested)

Service	Key Info You Must Know
Amazon Kendra	Intelligent search engine for enterprise data. Used in RAG to retrieve relevant info from documents.
Amazon OpenSearch Service	Supports vector search (k-NN). Used as a vector DB in RAG for semantic similarity.
Amazon Aurora (PostgreSQL)	Supports pgvector for vector search. Can store embeddings.
Amazon RDS for PostgreSQL	Also supports pgvector extension. Vector DB alternative for RAG.
Amazon Neptune	Graph DB with vector and feature similarity support. Used in advanced semantic use cases.
Amazon DocumentDB	MongoDB-compatible doc DB. Can store vectors; not purpose-built for vector search.
Amazon S3	Storage for datasets, models, and logs. Supports encryption, lifecycle policies.
AWS Glue DataBrew	No-code data preparation for ML workflows. Visual transformations.
SageMaker Data Wrangler	Visual tool to prepare, transform, and analyze data for ML models.
SageMaker Feature Store	Central repository for storing and retrieving ML features consistently across training/inference.
AWS Macie	Scans S3 for sensitive data (e.g., PII). Ensures privacy for AI datasets.

AWS IAM	Control access to AI services, endpoints, and data. Use least privilege and roles.
AWS KMS	Manage encryption keys for data at rest (e.g., S3, EBS, SageMaker volumes).
AWS PrivateLink	Keeps traffic between AI services and apps inside VPC (no internet exposure).
AWS CloudTrail	Logs API activity for auditing AI workflows.
AWS Config	Monitors AWS resource compliance (e.g., S3 not public, encryption enabled).
Amazon Inspector	Scans EC2 and containers for vulnerabilities. Relevant for self-managed AI infra.
AWS Audit Manager	Tracks and maps AWS compliance controls (e.g., ISO, SOC2) – useful for AI governance.
AWS Artifact	Download AWS compliance reports and certifications.
Amazon A2I (Augmented AI)	Adds human review step to ML inference workflows. Useful for high-risk or subjective outputs.
Amazon DataZone	Catalog and govern data assets across teams. Helps track data sources used in AI training.
SageMaker Model Cards	Documentation for models (use, data, metrics, risks). Enhances transparency and compliance.
AWS Trusted Advisor	Checks AWS account for security, performance, and cost issues. Helpful in governance and deployment best practices.
AWS VPC	Used to isolate SageMaker/Bedrock endpoints and restrict access.