AWS Certified

AI Practitioner

AIF-C01 · QUICK REFERENCE CHEAT SHEET

20%	24%	28%	14%	14%
Domain 1	**Domain 2**	**Domain 3**	**Domain 4**	**Domain 5**
Fundamentals of AI & ML	Fundamentals of Gen AI	Applications of Foundation Models	Guidelines for Responsible AI	Security, Compliance & Governance

85 Questions 65 scored + 20 unscored	**90 Minutes** Exam duration	**700 / 1000** Minimum passing score
Multiple Choice + Multiple Response	**No Prerequisites** Open to everyone	**Scaled Score** 100–1000 range

DOMAIN 1 · FUNDAMENTALS OF AI AND ML (20%)

Key AI / ML Concepts

Artificial Intelligence (AI)	Simulation of human intelligence in machines — perceiving, reasoning, learning, and problem-solving.
Machine Learning (ML)	Subset of AI: algorithms that learn patterns from data without being explicitly programmed.
Deep Learning (DL)	Subset of ML using multi-layer neural networks. Powers image recognition, NLP, and speech.
Neural Network	Layers of interconnected nodes (neurons) that transform input data into predictions.
Supervised Learning	Train on labeled data (input→output pairs). Examples: classification, regression.
Unsupervised Learning	Find patterns in unlabeled data. Examples: clustering (K-means), dimensionality reduction (PCA).
Reinforcement Learning	Agent learns by reward/penalty through trial-and-error in an environment.
Semi-supervised Learning	Uses a small amount of labeled data + large unlabeled dataset. Reduces labeling cost.
Self-supervised Learning	Creates labels from the data itself (e.g., predicting masked words). Used to train foundation models.

ML Model Types & Algorithms

Type	Algorithm	Use Case
Classification	Logistic Regression, SVM, Random Forest, XGBoost	Spam detection, sentiment analysis, fraud detection
Regression	Linear Regression, Ridge, Lasso, Gradient Boosting	Price prediction, demand forecasting
Clustering	K-Means, DBSCAN, Hierarchical	Customer segmentation, anomaly detection
Dimensionality Reduction	PCA, t-SNE, UMAP	Visualization, feature compression
Time Series	ARIMA, LSTM, Prophet	Forecasting, anomaly detection in sequences
Recommendation	Collaborative Filtering, Matrix Factorization	Product/content recommendations
NLP	BERT, GPT, Transformers	Text classification, translation, summarization
Computer Vision	CNN, ResNet, YOLO	Image classification, object detection

ML Lifecycle & Key Metrics

ML Lifecycle	Problem framing → Data collection → Data prep → Feature engineering → Model training → Evaluation → Deployment → Monitoring
Training vs. Inference	Training: learning parameters from data (expensive). Inference: applying trained model to new data (production).
Overfitting	Model memorizes training data; performs poorly on new data. Fix: regularization, more data, dropout.
Underfitting	Model too simple; poor on both train and test. Fix: increase model complexity, more features.
Bias-Variance Tradeoff	High bias = underfitting. High variance = overfitting. Goal: balance both for generalization.
Accuracy	(TP+TN)/(Total). Overall correctness — misleading with imbalanced classes.
Precision	TP/(TP+FP). Of predicted positives, how many are correct? Important when false positives are costly.
Recall (Sensitivity)	TP/(TP+FN). Of actual positives, how many did we find? Important when false negatives are costly.
F1 Score	Harmonic mean of Precision and Recall. Best single metric for imbalanced classes.
ROC-AUC	Area under ROC curve. Measures discrimination ability across thresholds (1.0 = perfect).
RMSE/MAE	Root Mean Squared Error / Mean Absolute Error. Used for regression evaluation.

AWS ML Services	Amazon SageMaker: full ML lifecycle platform (train, tune, deploy). SageMaker Autopilot: AutoML — auto feature engineering + model selection. SageMaker Canvas: no-code ML for business users. SageMaker Ground Truth: human data labeling service.

DOMAIN 2 · FUNDAMENTALS OF GENERATIVE AI (24%)

Generative AI Concepts

Generative AI	AI that creates new content (text, images, code, audio) by learning patterns from training data.
Foundation Model (FM)	Large model trained on broad data at scale; adaptable to many tasks via fine-tuning or prompting.
Large Language Model (LLM)	Type of FM trained on text corpora to understand and generate human language.
Transformer Architecture	Attention-based neural network architecture underlying most modern LLMs. "Attention is All You Need" (2017).
Attention Mechanism	Allows model to focus on relevant parts of input when generating each output token.
Token	Smallest unit of text processed by an LLM (word, sub-word, or character). Models have context window limits.
Context Window	Maximum number of tokens an LLM can process at once (input + output). Larger = more context.
Embeddings	Dense numerical vector representations of text/images capturing semantic meaning.
Vector Database	Stores embeddings for fast similarity search. Used in RAG architectures (e.g., Amazon OpenSearch).
Diffusion Models	Generative models for images: add noise then learn to reverse (denoise). e.g., Stable Diffusion.
GAN (Generative Adversarial Network)	Generator vs. Discriminator: adversarial training to create realistic synthetic data/images.
Multimodal Model	Handles multiple input/output types: text, image, audio, video (e.g., Claude, GPT-4o).

Prompt Engineering Techniques

Technique	Description	Example Use Case
Zero-shot	No examples provided; model uses learned knowledge	General Q&A;, summarization
One-shot	One example provided in the prompt	Specific format output
Few-shot	2–5 examples in the prompt	Consistent classification or extraction
Chain-of-Thought (CoT)	Ask model to reason step-by-step before answering	Math, logic, multi-step reasoning
ReAct	Combine reasoning + action (tool use)	Agents that call APIs or search
System Prompt	Instructions that set model behavior/persona	Chatbot persona, safety rules
Temperature	Controls randomness (0=deterministic, 1+=creative)	Creative writing vs. factual answers
Top-p / Top-k	Nucleus/top-k sampling: controls token selection diversity	Fine-tune creativity vs. accuracy

Model Customization Approaches

Prompt Engineering	Modify input prompt only — no training required. Fastest and cheapest. Best for simple adaptations.
RAG (Retrieval-Augmented Generation)	Retrieve relevant documents at inference time; inject into prompt. Keeps knowledge current, no retraining.
Fine-tuning	Update model weights on domain-specific data. Better task performance but requires labeled data and compute.
RLHF	Reinforcement Learning from Human Feedback: align model outputs with human preferences. Used for ChatGPT, Claude.

Continued Pre-training	Train on large domain corpus to add domain knowledge before fine-tuning. Most expensive approach.
Distillation	Transfer knowledge from a large model (teacher) to a smaller model (student). Reduces deployment cost.

Amazon Bedrock Fully managed service to access foundation models from leading AI companies (Anthropic Claude, Meta Llama, Mistral, AI21, Cohere, Amazon Titan) via a single API. No infrastructure to manage. Supports fine-tuning, RAG with Knowledge Bases, and Agents for multi-step reasoning.

DOMAIN 3 · APPLICATIONS OF FOUNDATION MODELS (28%)

AWS AI Services Quick Reference

Service	Category	Key Capability
Amazon Bedrock	FM access & GenAI	Access Claude, Llama, Titan models; Agents; Knowledge Bases (RAG)
Amazon Q Business	Enterprise GenAI assistant	Answer questions from company docs/data; connects to 40+ data sources
Amazon Q Developer	AI coding assistant	Code generation, debugging, security scanning in IDEs; CLI support
Amazon SageMaker	ML platform	Full ML lifecycle: data prep, train, tune, deploy, monitor
Amazon Comprehend	NLP / Text analysis	Sentiment, entities, key phrases, PII detection, topic modeling
Amazon Rekognition	Computer Vision	Object/face detection, text in images, content moderation, video analysis
Amazon Textract	Document analysis	Extract text, tables, forms from PDFs and images (OCR+)
Amazon Transcribe	Speech-to-text	ASR with speaker diarization, custom vocabulary, PII redaction
Amazon Polly	Text-to-speech	Lifelike speech synthesis; SSML support; 60+ voices
Amazon Translate	Neural MT	Real-time translation, custom terminology, batch translation
Amazon Lex	Conversational AI	Build chatbots/voice bots; NLU + ASR; integrates with Lambda
Amazon Personalize	Recommendations	Real-time personalized recommendations using ML; no ML expertise needed
Amazon Forecast	Time series	ML-based demand forecasting; auto model selection
Amazon Kendra	Intelligent search	ML-powered enterprise search; 40+ connectors; NLP understanding
Amazon Titan	AWS foundation models	Titan Text, Titan Embeddings, Titan Image Generator via Bedrock

RAG Architecture & Agentic AI

RAG Pipeline	1. Query → 2. Retrieve relevant docs from vector DB → 3. Augment prompt with docs → 4. Generate grounded response
Amazon Bedrock Knowledge Bases	Managed RAG: connect S3 data, auto-embed, store in vector DB. Retrieves context at inference automatically.
Bedrock Agents	Orchestrate multi-step tasks: break goal into steps, call APIs/Lambda, reason with FM, return final answer.
Agentic AI	AI system that autonomously takes actions (tool calls, web search, code execution) to achieve a goal.
Hallucination	Model confidently generates factually incorrect information. RAG and grounding help reduce this.
Grounding	Connecting model output to verifiable external sources to improve factual accuracy.
Guardrails (Bedrock)	Content filtering, PII redaction, topic denial, grounding checks — applied to FM inputs and outputs.
Model Evaluation (Bedrock)	Automatic (BERTScore, ROUGE) and human evaluation of FM responses for quality and safety.

Amazon Q vs. Amazon Bedrock Bedrock: developer tool — access/customize FMs, build GenAI apps with fine-tuning, RAG, and agents. Amazon Q Business: pre-built enterprise assistant — connect to company data sources for employee Q&A;, no ML expertise. Amazon Q Developer: AI coding assistant integrated into IDEs and CLI.

DOMAIN 4 · GUIDELINES FOR RESPONSIBLE AI (14%)

Responsible AI Principles

Fairness	AI should treat all individuals equitably. Avoid biases based on race, gender, age, etc.
Transparency	Stakeholders should understand how AI decisions are made. Explainability is key.
Privacy	Protect personal data; comply with GDPR, CCPA; apply data minimization and anonymization.
Safety	Prevent AI from causing harm. Test for edge cases, adversarial inputs, and unintended behaviors.
Truthfulness	AI should only assert things it believes to be true. Reduce hallucinations and misinformation.
Robustness	AI should perform reliably across diverse inputs and resist manipulation or adversarial attacks.
Human Oversight	Keep humans in the loop for high-stakes decisions. AI should support, not replace, human judgment.
Accountability	Clear ownership of AI outcomes. Document model cards, data lineage, and decision processes.

AI Bias, Fairness & Explainability

Data Bias	Training data does not represent real-world distribution. Fix: diverse, balanced datasets.
Algorithmic Bias	Model amplifies biases present in training data. Fix: bias auditing, fairness constraints.
Societal Bias	Historical discrimination encoded in data (e.g., hiring algorithms). Fix: debiasing techniques.
Explainability (XAI)	Methods to interpret why a model made a decision. SHAP, LIME, attention visualization.
Model Cards	Documentation summarizing model purpose, training data, limitations, and fairness evaluations.
Disparate Impact	When model outcomes disproportionately affect a protected group. Measure with fairness metrics.
Amazon SageMaker Clarify	Detects bias in data and model predictions; provides feature importance (SHAP) explanations.

DOMAIN 5 · SECURITY, COMPLIANCE & GOVERNANCE FOR AI (14%)

AI Security & Governance

AI Security Risks	Prompt injection, data poisoning, model inversion, adversarial attacks, sensitive data exposure.
Prompt Injection	Malicious input overrides system prompt to manipulate model behavior. Mitigate with Bedrock Guardrails.
Data Poisoning	Attacker corrupts training data to manipulate model predictions. Mitigate with data validation.
Model Inversion	Extract training data from model outputs. Mitigate with differential privacy.
PII Protection	Use Amazon Comprehend PII detection, Macie (S3), Transcribe PII redaction, Bedrock Guardrails.
Bedrock Guardrails	Input/output filtering: harmful content, PII, topic denial, grounding, word filters.
AWS AI Service Cards	Responsible AI documentation for each AWS AI service: intended use, limitations, design choices.
SageMaker Model Monitor	Detect data drift and model quality degradation in production. Alert on threshold violations.
AWS Config + CloudTrail	Track configuration changes and API calls for AI resources. Essential for compliance auditing.
VPC Endpoints for Bedrock	Keep Bedrock API calls within AWS network; no public internet. Enhances data privacy.

Compliance Frameworks for AI	NIST AI RMF (Risk Management Framework) · EU AI Act (risk-based tiers) · ISO 42001 (AI Management System) · GDPR/CCPA (data privacy) · SOC 2 Type II · AWS Shared Responsibility Model applies to AI services too — AWS secures the infrastructure; you secure your data, prompts, and model configurations.

MASTER QUICK REFERENCE — MOST FREQUENTLY TESTED

Know These Cold

Service / Concept	Category	Remember This
Amazon Bedrock	GenAI platform	Access FMs (Claude, Llama, Titan) via API; fine-tuning, RAG, Agents, Guardrails
Amazon Q Business	Enterprise assistant	Pre-built GenAI for employee Q&A; connects to 40+ company data sources
Amazon Q Developer	AI coding tool	Code gen, debugging, security in IDE; CLI; code reviews
SageMaker	ML lifecycle	Full ML platform: label, train, tune, deploy, monitor
SageMaker Clarify	Bias & explainability	Detect bias in data/model; SHAP feature importance
SageMaker Canvas	No-code ML	Business users build ML models without writing code
SageMaker Ground Truth	Data labeling	Human + automated labeling; active learning
Amazon Comprehend	NLP	Sentiment, entities, PII detection, key phrases
Amazon Rekognition	Computer Vision	Object/face detection, text in images, video analysis
Amazon Textract	Document OCR+	Extract text, tables, forms from PDFs/images
Amazon Transcribe	Speech-to-text	ASR; speaker diarization; PII redaction
Amazon Polly	Text-to-speech	Lifelike voice synthesis; 60+ voices
Amazon Translate	Translation	Neural machine translation; custom terminology
Amazon Lex	Chatbots	Build conversational bots; NLU + ASR
Amazon Personalize	Recommendations	Real-time personalized recs; no ML expertise
Amazon Forecast	Time series	Demand forecasting; automatic model selection
Amazon Kendra	Enterprise search	ML-powered search; 40+ connectors; NLP
RAG	Architecture pattern	Retrieval-Augmented Generation: ground LLM in external docs
Bedrock Knowledge Bases	Managed RAG	Connect S3/data to FM; auto-embedding; vector search
Bedrock Agents	Agentic AI	Orchestrate multi-step tasks; call APIs; reason to goal
Bedrock Guardrails	Safety controls	Content filtering, PII, topic denial, grounding checks
Foundation Model	Large pre-trained model	Broad training; adaptable to many tasks via prompts/fine-tuning
LLM	Language model	Understands + generates text; powers Q&A;, summarization, code
Hallucination	AI inaccuracy risk	Model generates false info confidently. Fix: RAG, grounding
Fine-tuning	Model customization	Update weights on domain data; requires labeled examples
RLHF	Alignment technique	Reinforcement Learning from Human Feedback; aligns with human prefs
Embeddings	Vector representations	Dense vectors capturing semantic meaning; used in RAG/search
Prompt Engineering	Input optimization	Zero-shot / few-shot / CoT / ReAct; no model retraining
Responsible AI	Ethical AI principles	Fairness, transparency, privacy, safety, accountability
SageMaker Clarify	Bias & explainability	Bias detection + SHAP explanations in one service

EXAM DAY REMINDERS

700/1000 to pass | 85 questions (65 scored) | 90 minutes | Answer every question — no penalty for guessing! | Bedrock = FM access platform | Q Business = employee assistant | Q Developer = coding AI | RAG = grounding via retrieval | Guardrails = content safety | Clarify = bias + explainability | Fine-tuning > Prompt Engineering for task-specific accuracy | Shared Responsibility: AWS secures infrastructure; YOU secure data, prompts, model configs