CloudNinja.pro

FOUNDATIONAL · AIF-C01

AWS Certified AI Practitioner

AIF-C01 Cheat Sheet

CloudNinja · cloudninja.pro · Free — no signup required

Every model-evaluation metric, prompt pattern, foundation-model concept, SageMaker tool, AWS AI service and responsible-AI guideline the AIF-C01 exam actually tests — organised the way the exam's five domains are weighted, not alphabetically. No prior ML experience assumed.

EXAM SNAPSHOT

- **65 questions** — 50 scored + 15 unscored
- **90 minutes** total time
- **700 / 1000** to pass
- **$100 USD** exam fee

- **AI & ML Fundamentals** — 20%
- **GenAI Fundamentals** — 24%
- **Foundation Models** — 28%
- **Responsible AI** — 14%
- **Security & Governance** — 14%

FREE AIF-C01 practice exam →

1 · AI & ML Fundamentals

Types of Machine Learning

Type	What It Does
Supervised learning	Labelled data (input → target). *Classification* (discrete label) or *regression* (continuous value). Examples: spam detection, house-price prediction.
Unsupervised learning	No labels — model finds structure. Clustering (K-means), dimensionality reduction (PCA, SVD, t-SNE), association rules.
Semi-supervised	Small labelled set + large unlabelled set. Cheaper than full labelling.
Self-supervised	Model generates its own labels from the data (predict next token, masked word). Foundation of LLM pre-training.
Reinforcement learning (RL)	Agent learns from rewards and penalties in an environment. Uses: games, robotics, trading, fine-tuning LLMs (RLHF).
Transfer learning	Start from a pre-trained model and fine-tune on your task. Less data, less compute, better results than training from scratch.

Data Labelling & Data Types

- **Structured data** — rows/columns (tables, spreadsheets). Easy to feed classical ML.
- **Semi-structured** — JSON, XML, logs. Has some schema but not rigid.
- **Unstructured** — free text, images, audio, video. Needs feature extraction or deep learning.
- **Time-series** — ordered observations over time. Forecasting, anomaly detection.
- **Labelling** — Amazon SageMaker Ground Truth (human labellers + active learning) or **Amazon Augmented AI (A2I)** for human review of ML predictions.

Model Fit — Bias vs. Variance

Concept	Definition	Symptom
Bias	Error from oversimplified assumptions — model can't capture the pattern.	Underfitting: high train & test error.
Variance	Error from over-sensitivity to training-set noise.	Overfitting: low train error, high test error.
Bias-variance tradeoff	Total error = bias² + variance + irreducible noise. Raising model complexity cuts bias but raises variance.	—

- **Fixes for underfitting** — more features, more complex model, less regularisation, train longer.
- **Fixes for overfitting** — more data, simpler model, regularisation (L1/L2), dropout, early stopping, ensembles.

Neural Network Architectures

Architecture	Use Case
Feedforward (MLP)	Tabular data, simple classification/regression.
Convolutional Neural Network (CNN)	Image and video. Convolutions learn spatial features. Examples: ResNet, EfficientNet.
Recurrent Neural Network (RNN)	Sequential data — time series, old-style NLP. LSTM / GRU handle long dependencies.

Architecture	Use Case
Transformer	Self-attention across a sequence. Foundation of modern LLMs (GPT, Claude, Llama, Nova, Titan). Enables parallel training at scale.
GAN (Generative Adversarial Network)	Generator vs. discriminator. Produces realistic images, synthetic data.
Diffusion model	Gradually denoise random input to produce images / audio. Backbone of Stable Diffusion, DALL·E.

2 · Generative AI Fundamentals

Foundation Models (FMs) & LLMs

- **Foundation model** — large model pre-trained on broad data, adaptable to many downstream tasks via prompting, fine-tuning, or RAG.
- **LLM** — foundation model for natural language (GPT, Claude, Titan, Llama, Nova).
- **Multimodal** — accepts/emits multiple modalities (text + image + audio + video). Nova, Claude, GPT-4o.
- **Open-source vs. proprietary** — Llama, Mistral (open) vs. Claude, GPT (proprietary API-only). Tradeoff: control + cost vs. capability + speed to ship.

Embeddings & Vector Databases

- **Embedding** — dense vector representing semantic meaning of text/image/audio. Similar items have similar vectors (cosine similarity).
- **Embedding models** — Titan Embeddings, Cohere Embed, Sentence-BERT.
- **Vector database** — stores and searches embeddings. AWS options: OpenSearch Service (with k-NN), Amazon Aurora (pgvector), Amazon Neptune Analytics, Amazon DocumentDB, Amazon MemoryDB, Amazon RDS for PostgreSQL (pgvector), Amazon Bedrock Knowledge Bases (managed).
- **Use cases** — semantic search, recommendation, RAG grounding, clustering.

Tokenisation & Context Windows

- **Token** — chunk of text the model reads/emits. Roughly 0.75 words in English.
- **Context window** — max input+output tokens per call (e.g. 200K for Claude 3.5/4, 1M for some Nova models).
- **Cost** — billed per-input-token + per-output-token. Longer prompts → higher cost.

Dimensionality Reduction

- **PCA (Principal Component Analysis)** — linear method. Projects high-dimensional data to fewer dimensions while preserving variance. Used for visualisation and noise reduction.
- **SVD (Singular Value Decomposition)** — matrix factorisation; foundation of PCA, recommendation systems, latent-semantic analysis.
- **t-SNE / UMAP** — non-linear visualisation of high-dimensional data.

3 · Foundation Models & Prompt Engineering

Customisation Approaches — Lightest to Heaviest

Approach	When to Use	Cost
Prompt engineering	Change the prompt to steer the model. No training. Instant iteration.	Lowest
RAG (Retrieval-Augmented Generation)	Inject retrieved context from your data (via vector DB) so the model grounds its answer. Keeps data fresh.	Low

Approach	When to Use	Cost
Fine-tuning	Continue training on labelled examples so the model adopts your style / format / domain. Use when prompt engineering hits a ceiling.	Medium–High
Continued pre-training	Further train on unlabelled domain corpora. For specialised languages/industries where the base model is weak.	High
Train from scratch	Rarely worth it outside research / hyperscale. Custom architecture or data modality.	Highest

Prompt Patterns — Foundation

- **Zero-shot** — just the task description. "*Classify this review as positive or negative: ...*"
- **Few-shot / in-context learning** — include 2–5 labelled examples in the prompt.
- **Instruction** — explicit, imperative directions. "*Summarise in three bullets. Use plain language.*"
- **Role / persona** — assign identity: "*You are a senior tax advisor answering a client.*"
- **Contextual** — paste relevant docs, then ask the question against them (RAG baseline).
- **Style / tone** — specify voice. "*Reply in the tone of a 19th-century letter.*"

Prompt Patterns — Reasoning

- **Chain-of-Thought (CoT)** — "*Let's think step by step.*" Forces the model to show reasoning. Boosts math/logic accuracy.
- **Self-consistency** — sample multiple CoT completions, take the majority answer. Reduces variance.
- **Tree-of-Thought (ToT)** — explore multiple reasoning branches, back up on dead ends.
- **ReAct** — interleave reasoning with tool/action calls. Foundation of agentic flows.
- **Reflection** — ask the model to critique its own answer and revise.

Prompt Patterns — Structural

- **Output structuring** — demand JSON, XML, markdown tables. Combine with schema validation on output.
- **Template prompting** — reusable skeleton with placeholder variables.
- **Prompt chaining** — split work across multiple prompts: summarise → extract → classify.
- **Active prompting** — model dynamically selects which examples/tools to use. Adaptive.
- **RAG integration** — "*Using the documents below, answer: ...*" Retrieve-then-generate.

Inference Parameters

Parameter	What It Controls
Temperature	Randomness. 0 = deterministic, 1 = creative. Low for code/extraction, high for brainstorming.
Top-p (nucleus)	Sample from the smallest set of tokens whose cumulative probability ≥ p. 0.9 is typical.
Top-k	Consider only the k most-likely next tokens.
Max tokens	Cap on output length. Stops generation regardless of content.
Stop sequences	Strings that end generation when produced. Useful for streaming / structured outputs.
Presence / frequency penalty	Discourage repeating tokens already seen — reduces looping.

Exam Tip: "*Make output more creative*" → raise temperature or top-p. "*Make output deterministic for evaluation*" → temperature 0. "*Model keeps repeating*" → frequency/presence penalty. "*Model too long-winded*" → lower max tokens + explicit instruction.

4 · Amazon Bedrock

Fully managed service to access foundation models via a single API. No infrastructure to provision. Usage-based pricing (per token) or Provisioned Throughput (hourly commitment). All data stays in your account — AWS never uses your prompts to train base models.

Model Families

- **Amazon Nova** — Amazon's own frontier multimodal FM family: Nova Micro (text-only, fastest), Lite, Pro, Premier, plus Nova Canvas (image) and Nova Reel (video).
- **Amazon Titan** — text generation, embeddings, image generation. Lower cost, good for enterprise use cases.
- **Anthropic Claude** — Claude Haiku / Sonnet / Opus. Strong reasoning and long context.
- **Meta Llama** — open-weight FMs in multiple sizes.
- **Mistral, Cohere, AI21, Stability AI** — additional specialised models.

Bedrock Capabilities

Feature	What It Does
Invoke API	Send a prompt, get a completion (streaming or batch). Same API shape for every model.
Knowledge Bases	Managed RAG. Point at S3; Bedrock chunks, embeds (Titan or Cohere), stores in OpenSearch Serverless / Aurora / Neptune Analytics / Pinecone / Redis, and handles retrieval at query time.
Agents	Multi-step reasoning + tool calling. Define action groups (Lambda functions or OpenAPI schemas) + a knowledge base. Bedrock plans, executes, and returns the final answer.
Guardrails	Content filters (sexual / violent / hate / insults / misconduct), denied topics, word filters, sensitive-info filters (PII redaction), grounding checks, prompt-injection detection.
Custom models	Fine-tune a supported base model on your data (JSONL) or run Continued Pre-training. Access customised weights via Provisioned Throughput.
Model Evaluation	Compare models on your data: automatic metrics (accuracy, BLEU, F1) or human evaluators via Ground Truth / A2I workflows.
Provisioned Throughput	Hourly/monthly commitment that reserves capacity (in *model units*) for predictable latency. Required for custom fine-tuned models.
Prompt Management	Versioned prompt templates and prompt flows. Deploy and A/B test prompts without code changes.

Exam Tip: "*Enterprise wants to chat with its own documents*" → Bedrock Knowledge Bases (RAG). "*Block the model from answering questions about X*" → Guardrails denied topics. "*Model must refuse to output PII*" → Guardrails sensitive-info filter. "*Need predictable latency under a heavy load*" → Provisioned Throughput.

5 · Amazon SageMaker AI

End-to-end managed ML platform: data prep, training, deployment, monitoring, governance. Exam tests the SageMaker *tool names* and what each solves.

Capability	What It Does
SageMaker Studio	Web-based IDE for the full ML lifecycle. Jupyter-based notebooks, experiments, pipelines, debugging.
SageMaker JumpStart	Catalogue of pre-trained foundation models + task-specific models. Deploy in one click. Good starting point for fine-tuning.
SageMaker Canvas	No-code ML for analysts. Build models by pointing at data; get predictions without writing code.

Capability	What It Does
SageMaker Data Wrangler	Visual data preparation — import, profile, transform, feature-engineer data for ML training.
SageMaker Feature Store	Central repository of features, online and offline. Share features across teams, avoid skew between train and serve.
SageMaker Ground Truth	Data labelling service. Human labellers (internal, vendor, Mechanical Turk) plus active-learning auto-labelling.
SageMaker Training	Managed distributed training on GPUs/CPUs. Spot training for cost. Built-in algorithms and custom containers.
SageMaker Experiments	Track training runs, hyperparameters, metrics. Compare models side by side.
SageMaker Model Registry	Versioned catalogue of approved models. Promote model versions through dev → staging → prod.
SageMaker Endpoints	Managed inference. Real-time (always on), Serverless (scale to zero), Asynchronous (long inference), Batch transform.
SageMaker Model Monitor	Detects drift — data quality, model quality, bias, feature attribution — on production endpoints. Alerts via CloudWatch.
SageMaker Model Dashboard	Central console of all your models, endpoints, and Model Monitor alerts.
SageMaker Clarify	Detects bias in data and models (pre-training and post-training); explains predictions with Shapley values (SHAP).
SageMaker Pipelines	Native CI/CD for ML. DAG of steps (processing → training → evaluation → register → deploy).
SageMaker MLflow	Managed MLflow tracking server for teams already on open-source MLflow.
SageMaker Role Manager	Predefined permissions templates for data scientists, MLOps engineers, etc.

Exam Tip: "*No-code ML for business analysts*" = Canvas. "*Label 100K images*" = Ground Truth (+ A2I for review). "*Detect feature distribution drift in production*" = Model Monitor. "*Explain why the model said no*" = Clarify (SHAP). "*Share features between training and inference*" = Feature Store.

6 · AWS AI Services (Pre-built APIs)

Service	What It Does
Amazon Comprehend	NLP — sentiment, entity recognition, key-phrase extraction, language detection, topic modelling, PII detection. Comprehend Medical for healthcare text.
Amazon Rekognition	Image & video analysis — object/scene detection, faces, celebrities, text, content moderation, PPE detection.
Amazon Textract	Extract text, tables, and forms from scanned documents. Beyond OCR — understands structure. Queries feature for fielded extraction.
Amazon Transcribe	Speech-to-text. Batch and streaming. Custom vocabulary, speaker diarisation, PII redaction. Transcribe Medical for healthcare.
Amazon Translate	Neural machine translation across 75+ languages. Custom terminology, active custom translation.
Amazon Polly	Text-to-speech with natural voices. Neural and Generative voices. SSML for prosody, breaks, pronunciation.
Amazon Lex	Build chat and voice bots — same tech as Alexa. Intents, slots, fulfilment via Lambda. Integrates with Connect for contact centres.

Service	What It Does
Amazon Kendra	ML-powered intelligent enterprise search. Natural-language queries over S3, RDS, SharePoint, Confluence, etc. Connectors handle auth.
Amazon Personalize	Managed recommendation engine (same tech as Amazon.com). Real-time personalised recommendations, ranking, similar items.
Amazon Fraud Detector	Managed fraud-detection models. Prebuilt detectors for online fraud, account takeover, new-account fraud. Uses Amazon's retail fraud experience.
Amazon Augmented AI (A2I)	Human-in-the-loop review for ML predictions. Works with Rekognition, Textract, SageMaker, or custom. Route low-confidence predictions to human reviewers.

7 · Amazon Q Family

Variant	Audience & Use
Amazon Q Developer	AI assistant for developers and IT. Code generation, inline suggestions in IDEs and the CLI, code review, CLI command generation, troubleshooting. Integrated into Console, IDE, CLI.
Amazon Q Business	Fully managed GenAI assistant over your company's internal data — SharePoint, Confluence, S3, Salesforce, Slack, etc. Grounded answers with citations, no hallucination of facts.
Amazon Q in QuickSight	Natural-language BI. Ask "What were top-5 products by revenue last quarter?" and get a chart. Generative narrative summaries of dashboards.
Amazon Q in Connect	Real-time agent-assist in contact centres. Suggests responses and knowledge articles during customer calls.

8 · Model Evaluation Metrics

Classification

Metric	Formula / Meaning	When to Favour
Accuracy	(TP + TN) / (TP + TN + FP + FN). % predictions correct.	Balanced classes.
Precision	TP / (TP + FP). Of those we said positive, how many really were.	False positives expensive (spam filters, fraud).
Recall (sensitivity)	TP / (TP + FN). Of all actual positives, how many we caught.	False negatives expensive (disease, fraud).
F1 score	Harmonic mean of precision and recall.	Imbalanced classes.
AUC-ROC	Area under ROC curve. Threshold-independent ranking quality.	Need a single score that works across thresholds.
Log loss	Penalises confident wrong predictions.	Probabilistic classifiers.
Confusion matrix	Table of TP/FP/TN/FN per class.	Multi-class diagnosis.

Regression

- **MAE (Mean Absolute Error)** — average absolute difference between predicted and actual. Same units as target.

- **MSE (Mean Squared Error)** — average squared difference. Penalises large errors more.
- **RMSE** — √MSE. Same units as target. More interpretable than MSE.
- **R² (coefficient of determination)** — fraction of variance explained by the model. 1 = perfect, 0 = no better than mean, negative = worse than mean.

Text Generation

Metric	What It Measures	Best For
ROUGE	Recall-based n-gram and sequence overlap with reference.	Summarisation.
BLEU	Precision-based n-gram overlap with reference.	Machine translation.
METEOR	Weighted precision + recall + synonym matching.	Translation, summarisation.
BERTScore	Semantic similarity using BERT embeddings.	Any generation — captures meaning, not just word overlap.
CIDEr	Consensus-based n-gram similarity (TF-IDF weighted).	Image captioning.
Perplexity	How uncertain the model is predicting the next token. Lower is better.	Language modelling.

Exam Tip: ROUGE is recall-weighted (did we capture the reference?). BLEU is precision-weighted (is what we produced in the reference?). BERTScore captures meaning — our output can have zero word overlap yet still score high if it's semantically equivalent. Accuracy lies when classes are imbalanced — switch to F1 or AUC-ROC.

9 · Responsible AI

AWS Responsible AI Dimensions

Dimension	What It Means in Practice
Fairness	Model performs equitably across demographic groups. Measure with SageMaker Clarify bias metrics (DPL, CDD, DI, DAR).
Explainability	Understand *why* the model decided. Clarify SHAP values; partial dependence plots.
Robustness	Resists adversarial input and distribution shift. Monitor with Model Monitor.
Privacy & Security	Protect training and inference data. KMS encryption, VPC endpoints, PII redaction, Bedrock Guardrails.
Governance	Document datasets, models, training runs. SageMaker Model Cards, Model Registry, lineage tracking.
Transparency	Users know they're interacting with AI, model strengths and limitations are published (AWS AI Service Cards).
Veracity	Outputs grounded in reliable sources. RAG with Knowledge Bases, Guardrails grounding checks.

Bias Types & Detection

- **Sampling bias** — training data not representative of the real population (geography, demographics, time).
- **Measurement bias** — features recorded differently across groups.
- **Algorithmic bias** — model amplifies patterns in the data.
- **Confirmation bias** — humans labelling data reflect their assumptions.
- **Clarify** detects bias pre-training (data) and post-training (predictions) with metrics like Class Imbalance, Demographic Parity, Disparate Impact, Conditional Demographic Disparity in Labels.

Explainability Techniques

- **SHAP (Shapley values)** — per-feature contribution to a single prediction. Local explainability. SageMaker Clarify supports this.
- **Partial Dependence Plots (PDP)** — show marginal effect of a feature on model output. Global view.
- **LIME** — approximate a complex model locally with a simpler surrogate. Not AWS-specific.
- **Feature attribution drift** — Model Monitor tracks whether the features driving predictions change over time.

Generative AI Risks & Mitigations

- **Hallucination** — model invents plausible-sounding but wrong facts. Mitigate with RAG, grounding checks, Guardrails.
- **Prompt injection** — adversarial instructions smuggled into user input. Mitigate with Guardrails prompt-injection filter, strict system-prompt boundaries.
- **Toxic / biased outputs** — Guardrails content filters, human-in-the-loop via A2I, red-team evaluation.
- **Training-data leakage** — model reveals sensitive training examples. Use differential privacy, Bedrock (which does not train base models on customer data).
- **Copyright / IP** — AWS offers IP-indemnification for Amazon Titan / Amazon Nova / Claude output under specific conditions. Review service terms.

10 · Security, Compliance & Governance

Identity & Access

- **AWS Identity and Access Management (IAM)** — users, groups, roles, policies (JSON). Always prefer IAM roles for workloads (SageMaker notebooks, Lambda, EC2) over long-lived access keys.
- **Principle of least privilege** — grant only the minimum permissions required. Use managed policies as a starting point, then scope down.
- **Resource-based policies** — attach to S3 buckets, KMS keys, SageMaker endpoints to control cross-account access.

Data Protection

Service	Role in AI Workloads
AWS Key Management Service (KMS)	Encryption keys for S3 training data, EBS volumes, SageMaker artifacts, Bedrock custom-model weights. Customer-managed keys give full control.
AWS Secrets Manager	Store API keys, DB credentials, third-party LLM keys. Automatic rotation.
Amazon Macie	Discover and protect sensitive data (PII, credentials) in S3 — *before* you ship it to a training job.
Amazon Inspector	Vulnerability scanning for EC2, Lambda, container images (ECR). Catches CVEs in your inference stack.

Audit & Compliance

Service	What It Does
AWS CloudTrail	Records every AWS API call. Management events free; data events (S3 object-level, Bedrock Invoke logs) optional. Critical for auditing who invoked which model with which prompt.
Amazon CloudWatch	Metrics, logs, and alarms for model endpoints, training jobs, Bedrock usage.
AWS Config	Tracks resource configuration and compliance. Rules like 'SageMaker notebooks must have root access disabled', 'Bedrock models must have Guardrails attached' (custom rules).
AWS Artifact	Self-service portal for compliance reports (SOC, PCI, ISO, HIPAA BAA). Download what auditors need.

Service	What It Does
AWS Audit Manager	Continuously collect evidence against frameworks (HIPAA, GDPR, PCI DSS, AI RMF). Generates audit-ready reports.
AWS Trusted Advisor	Best-practice checks across cost, performance, security, fault tolerance, service limits. Useful for catching idle SageMaker endpoints.
AWS Well-Architected Tool	Self-assessment against the six Well-Architected pillars and the *ML Lens* / *Generative AI Lens* for workload-specific guidance.

11 · Compute & Containers

Service	AI/ML Relevance
Amazon EC2	Virtual servers. GPU families (P — training, G — inference/graphics), Inferentia (Inf) and Trainium (Trn) for cost-efficient custom silicon. Spot instances for interruptible training.
Amazon ECS	Managed container orchestration. Run custom inference containers with auto-scaling. Fargate launch type = no EC2 to manage.
Amazon EKS	Managed Kubernetes. Pair with the AWS Neuron SDK for Inferentia/Trainium, or NVIDIA device plugin for GPU scheduling.

12 · Storage

- **Amazon S3** — object storage. Primary data lake for ML training sets, model artifacts, Knowledge Base sources. Integrates natively with SageMaker, Bedrock, Glue, Athena, Redshift Spectrum.
- **Amazon S3 Glacier** — archive storage classes (Instant Retrieval, Flexible Retrieval, Deep Archive). Use for training-data retention, audit-log archives, compliance-mandated retention.
- **S3 features for ML** — versioning (roll back bad data), lifecycle rules (tier old experiments), Object Lock (compliance retention), event notifications (trigger retraining on new data).

13 · Database

Service	AI/ML Use
Amazon RDS	Managed relational (MySQL, PostgreSQL, MariaDB, Oracle, SQL Server, Db2). *pgvector* extension on PostgreSQL enables vector search for RAG.
Amazon Aurora	AWS-built relational engine. Aurora PostgreSQL supports pgvector as a cost-effective managed vector DB. Aurora Serverless v2 scales capacity in 0.5 ACU steps.
Amazon DynamoDB	Serverless NoSQL. Session state for chatbots, feature store backing data, metadata for ML experiments.
Amazon DocumentDB	MongoDB-compatible document DB. Supports vector search for RAG over semi-structured corpora.
Amazon Neptune	Graph DB. Knowledge graphs, fraud patterns, recommendation systems. Neptune Analytics supports vector similarity + graph algorithms.
Amazon ElastiCache	In-memory Redis / Memcached. Cache LLM prompts/completions, rate limiters, session state.
Amazon MemoryDB	Redis-compatible *durable* database (Multi-AZ, WAL). Also supports vector search. Use when you need Redis speed + DB durability.

14 · Analytics & Data Preparation

Service	Role in AI/ML
AWS Glue	Serverless ETL. Crawlers auto-detect schema; Glue Data Catalog is the central metadata store used by Athena, Redshift Spectrum, EMR, SageMaker.
AWS Glue DataBrew	Visual data preparation — no-code cleaning, normalising, deduplicating. Profile datasets before training. Outputs cleaned data to S3/Redshift/RDS.
AWS Lake Formation	Build data lakes on S3 with centralised permissions, fine-grained (row/column/cell) access control, cross-account sharing.
Amazon EMR	Managed Hadoop/Spark for distributed data processing. Spark MLlib for classical ML at scale; common for feature engineering on billions of rows.
Amazon OpenSearch Service	Search + analytics. *k-NN / vector engine* makes it a common RAG backend. Also log analytics / ELK-stack replacement.
Amazon QuickSight	Serverless BI. SPICE in-memory engine. Q in QuickSight = natural-language BI.
Amazon Redshift	Columnar data warehouse. Redshift ML trains SageMaker models on Redshift data via SQL. Zero-ETL integration with Aurora and DynamoDB.
AWS Data Exchange	Find, subscribe to, and use third-party datasets (weather, finance, healthcare, demographics) directly in S3. Accelerates training-data acquisition.

15 · Networking & Content Delivery

- **Amazon VPC** — isolated virtual network. SageMaker training jobs, endpoints, and notebooks can run inside a VPC with no internet egress. Bedrock supports VPC Endpoints (PrivateLink) so prompts never traverse the public internet.
- **Amazon CloudFront** — CDN with 700+ edge locations. Cache inference results for public models, serve a chat-bot web front-end with low latency, protect origins with AWS WAF.
- **VPC Endpoints** — Gateway (S3, DynamoDB — free) and Interface / PrivateLink (SageMaker, Bedrock, most services) keep traffic on the AWS backbone.

16 · Management & Governance

- **Amazon CloudWatch** — metrics (endpoint invocations, latency, errors), logs (SageMaker training output, Bedrock Invoke logs), alarms (on model drift or cost spikes), dashboards.
- **AWS CloudTrail** — who invoked which Bedrock model, who edited which SageMaker endpoint. Management events free; data events (e.g. Bedrock Invoke) optional + charged.
- **AWS Config** — track resource configuration over time. Write a custom rule like 'all Bedrock agents must have a Guardrail attached'.
- **AWS Trusted Advisor** — best-practice checks. Flags idle SageMaker endpoints, underutilised RIs, unassociated Elastic IPs.
- **AWS Well-Architected Tool** — review workloads against the six pillars plus the *Machine Learning Lens* and *Generative AI Lens* for AI-specific guidance.

17 · Cost Management

- **AWS Cost Explorer** — visualise and forecast AI spend by service, tag (e.g. `project:rag`), account.
- **AWS Budgets** — thresholds on cost or usage. Alert when a Bedrock-heavy project trends over budget; can trigger SCPs to stop new spend.
- **Bedrock pricing model** — per-token on-demand (default) vs. Provisioned Throughput (hourly commitment, required for custom fine-tuned models).

- **SageMaker cost levers** — Spot training (up to 90% off), right-size inference instances, Serverless Inference for bursty traffic, shut down notebooks when idle, Savings Plans for long-running endpoints.
- **Free Tier** — SageMaker Studio: 250 hrs/month ml.t3.medium for 2 months; limited Bedrock and Comprehend free usage.

> **Exam Tip:** 'Cut training cost on fault-tolerant jobs' = SageMaker Spot training. 'Predictable inference latency at scale' = Provisioned Throughput (Bedrock) or real-time endpoints with Auto Scaling. 'Pay nothing when no traffic' = SageMaker Serverless Inference.

cloudninja.pro — Free AWS certification practice exams, study guides, and cheat sheets.
Pair this cheat sheet with our free AIF-C01 practice exam for the best results.