CloudNinja.pro

PROFESSIONAL · AIP-C01

AWS Certified Generative AI Developer – Professional

AIP-C01 Cheat Sheet

CloudNinja · cloudninja.pro · Free — no signup required

Every service in scope, every RAG / agent / fine-tuning pattern, every safety and operations decision the AIP-C01 exam actually tests — organised around the five exam domains. Assumes ≥ 2 years of hands-on AWS development and prior AI/ML experience.

EXAM SNAPSHOT

- **75 questions** — 65 scored + 10 unscored
- **180 minutes** total time
- **750 / 1000** to pass
- **$300 USD** exam fee

- **FM Integration & Data** — 31%
- **Implementation & Integration** — 26%
- **AI Safety, Security & Governance** — 20%
- **Operational Efficiency** — 12%
- **Testing & Validation** — 11%

FREE AIP-C01 practice exam →

Closed services still on the AIP-C01 guide. AWS has closed SageMaker Ground Truth, Clarify, Model Monitor, Amazon Augmented AI, Amazon Kendra, Amazon Q Business and AWS App Runner to new customers, but the AIP-C01 in-scope list still names them, so they stay in this sheet with a CLOSED · ON GUIDE tag. Bedrock Agents Classic is closed and no longer on the guide: DEPRECATED. For a new design, pick the replacement (list on the last page). Amazon Q Developer is on the guide too, but its IDE plugins and paid subscriptions (Pro) reach end of support on 30 April 2027 (Q in the AWS console continues) and its CLI is now Kiro: END OF SUPPORT · ON GUIDE.

1 · GenAI Vocabulary & Customisation Ladder

- **Foundation Model (FM)** — large model pre-trained on broad corpora (Claude, Titan, Llama, Nova, Mistral, Cohere, AI21, Stability). **LLM** = text-only FM. **Multimodal** = text + image + audio.
- **Embeddings** — numeric vectors representing text, image, code; similar meaning ⇒ nearby vectors. Produced by embedding models (Titan Text Embeddings v2, Cohere Embed).
- **Tokens / context window** — unit the model sees; pricing and limits are per-token. Long context (> 200 k tokens) reduces need for chunking but costs more.
- **Inference parameters** — *temperature* (randomness), *top-p*, *top-k*, *max tokens*, *stop sequences*, *frequency/presence penalty*.
- **RAG (Retrieval-Augmented Generation)** — retrieve relevant context from a vector / hybrid store, inject into the prompt, generate.
- **Agent** — FM that calls tools (functions, APIs, retrievers) in a loop to accomplish a goal.
- **MCP (Model Context Protocol)** — open protocol for tool / resource / prompt providers; used by Bedrock AgentCore and Strands Agents.

Customisation Ladder (cheapest → most expensive)

Technique	When to use
Prompt engineering	First try. Zero-shot, few-shot, instruction, chain-of-thought, ReAct, tree-of-thought, self-consistency.
RAG	Domain-specific facts, frequent updates, auditability. No model retraining.
Fine-tuning	Style / format / task patterns not reachable by prompting. Bedrock fine-tuning or SageMaker + LoRA/PEFT.
Continued pre-training	Large domain vocabulary (legal, medical, code). Bedrock CPT with unlabeled corpus.
Train from scratch	Rarely justified outside frontier labs; SageMaker HyperPod / P-instances.

2 · Amazon Bedrock — The Exam's Centre of Gravity

Model Providers

Provider	Model families & strengths
Amazon	*Nova* (Micro, Lite, Pro, Premier) — cost-performance; *Titan* Text / Text Embeddings v2 / Multimodal Embeddings / Image Generator.
Anthropic	Claude (Haiku, Sonnet, Opus) — reasoning, long context, tool use, computer use.
Meta	Llama 3 / 3.1 / 3.2 — open-weights, multimodal variants.
Mistral	Mistral Small / Large — efficient, multilingual.
Cohere	Command R/R+ (RAG-tuned), Embed, Rerank.
AI21 Labs	Jamba — long context.
Stability AI	Stable Diffusion 3.x — image generation.

Inference Modes

- **On-demand** — pay per input+output tokens. No commitment.
- **Provisioned Throughput** — reserved model units for consistent high-rate workloads; required for custom models and some foundation models.

- **Batch Inference** — up to 50% cheaper; process JSONL files in S3 asynchronously.
- **Cross-Region Inference** — automatic routing across a set of Regions to bypass per-Region capacity limits and improve throughput resilience.
- **Latency-Optimized Inference** — lower TTFT for flagship models on supported Regions.

Bedrock Features

- **Bedrock Playgrounds** — interactive testing (chat, text, image).
- **Bedrock Model Evaluation** — automatic (curated datasets, algorithmic metrics), LLM-as-a-judge, or human evaluation with your own team; compare models side-by-side. Replaces SageMaker Clarify FM evaluation.
- **Bedrock Guardrails** — content filters (hate, insults, sexual, violence, misconduct, prompt attacks), denied topics, word filters + managed (profanity) lists, sensitive-information filter (PII redaction + regex), contextual grounding + relevance checks, automated reasoning checks. Applied at request + response.
- **Bedrock Custom Models** — fine-tuning, continued pre-training, model distillation, model import (bring Hugging Face / custom architecture).
- **Bedrock in SageMaker Unified Studio** — web IDE for builders (formerly Bedrock Studio; uses IAM Identity Center).

3 · Knowledge Bases, Prompt Management & Prompt Flows

Amazon Bedrock Knowledge Bases

- Managed RAG — ingest data from S3, Confluence, SharePoint, Salesforce, web crawler, direct documents. Chunks documents (default / fixed / semantic / hierarchical), generates embeddings, writes to a vector store.
- Supported vector stores: **OpenSearch Serverless** (default), OpenSearch managed, **Aurora PostgreSQL pgvector**, Pinecone, Redis Enterprise, MongoDB Atlas, Neptune Analytics.
- Query via *RetrieveAndGenerate* (end-to-end RAG) or *Retrieve* (retrieve only — you generate). Automatic citations + data source filters + reranking.
- Advanced parsing (Bedrock Data Automation) for complex PDFs, images, video.

Amazon Bedrock Prompt Management

- Store, version, compare prompts. Variables, per-model defaults, metadata.
- Approval workflows + IAM + CloudTrail auditing for governance.

Amazon Bedrock Prompt Flows

- Visual DAG — sources (API), nodes (prompt, Lambda, KB, condition, iterator, collector), destinations.
- Deterministic orchestration without building a full Step Function for simple multi-prompt chains.

4 · Agents: Bedrock AgentCore & Frameworks

- **Bedrock AgentCore** — framework-agnostic platform for production agents: *Runtime* (sandboxed execution), *Gateway* (MCP + tool surface with auth), *Memory* (short + long-term), *Identity* (delegated workforce / customer identities, OAuth), *Observability*, *Code Interpreter*, *Browser tool*. Works with Strands Agents, LangGraph, CrewAI, etc.
- **Strands Agents** — AWS-maintained open-source SDK (Python) for building tool-using agents. First-class support for MCP + Bedrock + any OpenAI-compatible endpoint.
- **Agent Squad** — open-source multi-agent orchestrator (supervisor → specialist agents) that pairs naturally with AgentCore.
- **Kiro** — AWS's agentic IDE and CLI: spec-driven development (requirements → design → tasks), then writes and tests the code. Connects to MCP servers.
- **Human-in-the-loop (Amazon Augmented AI — A2I)** CLOSED · ON GUIDE — review workflows for low-confidence model outputs; built-in templates for Textract, Rekognition. New designs: Step Functions task token + your own review UI, or Bedrock human evaluation jobs.

Bedrock Agents Classic DEPRECATED — the original managed single-agent runtime (action groups + Knowledge Bases + Guardrails, Bedrock-run ReAct loop). No longer open to new customers from 30 July 2026 and no longer on the AIP-C01 in-scope list. AWS's migration path is the AgentCore managed harness: action groups become Gateway MCP tools, Knowledge Bases a retrieval tool, return of control becomes inline function tools.

Exam cue: "Agent that calls internal APIs with auth + memory + observability in production" → **Bedrock AgentCore**. "Sequential prompt chain with branches, no custom code" → **Prompt Flows**. "Long-running workflow with fan-out, retries, human approval" → **Step Functions**.

5 · Vector Stores & Retrieval Architectures

Store	When to pick
Amazon OpenSearch Serverless (vector engine)	Default choice. Managed k-NN + BM25 hybrid search; tight Bedrock Knowledge Bases integration (Neural plugin for Bedrock).
Amazon Aurora PostgreSQL pgvector	App already uses Postgres; transactional metadata + vectors in one place.
Amazon Neptune Analytics	Graph-aware retrieval (GraphRAG); entities + relationships + vectors.
Amazon DocumentDB	MongoDB-compatible + vector search on same documents.
Amazon MemoryDB for Redis / ElastiCache for Redis	Ultra-low-latency vector search (< 10 ms); good for real-time recommendations.
Amazon DynamoDB	Metadata + embeddings references; pair with OpenSearch / Aurora for actual similarity search.

Chunking & Embeddings

- **Chunking**: default (Bedrock KB), fixed-size (word count + overlap), semantic (sentence embeddings cluster), hierarchical (parent/child for broader + narrower context). Overlap avoids context loss at boundaries.
- **Embedding model selection**: dimensionality trade-off (Titan Text Embeddings v2 offers 256 / 512 / 1024), domain fit, multilingual support, cost per token.
- **Reranking** — Cohere Rerank or Bedrock rerank models re-score top-k for precision.
- **Hybrid search** — combine BM25 keyword + semantic; normalise + weight.
- **Query transformation** — expand (synonyms), decompose (multi-step), HyDE (hypothetical answer → embed). Chain in Step Functions or Prompt Flows.
- **Maintenance** — incremental updates + scheduled refresh; Bedrock KB handles ingestion jobs.

6 · Amazon SageMaker AI Suite

Service	Role
Amazon SageMaker AI	End-to-end ML platform: notebooks (Studio), training, hosting (real-time / serverless / async / batch transform / multi-model / shadow), MLOps (Pipelines).
SageMaker Unified Studio	Single pane (preview/GA) combining Studio + Data + Lake Formation + Athena + Redshift + Bedrock; projects + governance.
SageMaker JumpStart	Curated catalog of open-source + partner FMs (Llama, Falcon, Stable Diffusion, Mistral); one-click deploy or fine-tune.
SageMaker Data Wrangler	Low-code data prep; 300+ built-in transforms; exports to Processing jobs + Feature Store.
SageMaker Ground Truth CLOSED · ON GUIDE	Labeling (human, vendor); active learning; RLHF preference data. New designs: Bedrock human evaluation jobs for FM output review.

Service	Role
SageMaker Clarify CLOSED · ON GUIDE	Bias metrics pre-train + post-train, SHAP explanations; FM evaluation (toxicity, factual, adversarial). New designs: Bedrock Model Evaluation, SHAP library.
SageMaker Model Monitor CLOSED · ON GUIDE	Data drift, model quality, bias drift, feature attribution drift on endpoints. New designs: MLflow + CloudWatch drift monitoring.
SageMaker Model Registry	Versioned catalog; approval gates; cross-account deploy via EventBridge + CodePipeline.
SageMaker Processing	Managed batch data-processing containers (SciKit, Spark, PyTorch, bring-your-own).
SageMaker Neo	Compile models for optimised inference on edge / specific instance types.

7 · AWS AI Services & Amazon Q

Service	What it does
Amazon Comprehend	NLP — entities, sentiment, PII redaction, custom classification (topic modelling is closed to new customers). Use for input normalisation / enrichment before FM call.
Amazon Kendra CLOSED · ON GUIDE	Enterprise search with semantic understanding + Kendra Intelligent Ranking. No longer open to new customers; AWS points to Bedrock Knowledge Bases.
Amazon Lex	Build conversational bots — ASR + NLU + DM. Integrates with Bedrock and Connect.
Amazon Rekognition	Images + video: labels, moderation, faces, OCR, PPE, custom labels.
Amazon Textract	Documents — raw text, forms (K/V), tables, signatures, lending + invoice specialised extractors. Pair with Comprehend for entity extraction.
Amazon Transcribe	Speech → text with punctuation, speaker diarisation, custom vocab, Call Analytics, PII redaction, medical.
Amazon Quick	Agentic workspace over company data, successor to Q Business: chat agents, research, flows, Quick Sight BI. The choice for a new enterprise assistant.
Amazon Q Business CLOSED · ON GUIDE	Generative AI assistant for enterprises — plug into data sources (S3, Confluence, Box, ServiceNow, Teams, etc.), IAM Identity Center for user identity, response customisation, guardrails. No longer open to new customers → Amazon Quick.
Amazon Q Business Apps CLOSED · ON GUIDE	Builder-less mini-apps created via natural language; surfaced to users inside Q.
Amazon Q Developer END OF SUPPORT · ON GUIDE	AI coding assistant. IDE plugins and Pro subscriptions (including customizations) reach end of support 30 Apr 2027; Q in the AWS console continues. The CLI is now Kiro. Agent features: */doc*, */test* and */transform* (Java upgrades, .NET porting to Linux); code transformation also ends 30 Apr 2027 → **AWS Transform** (which also covers mainframe).
Amazon Connect	Omnichannel contact centre; embed Lex bots + Bedrock for generative customer interactions; Contact Lens analytics with generative summaries.

8 · Compute for GenAI Workloads

- **AWS Lambda** — glue for orchestration, tool implementation, lightweight inference on Bedrock. 15-min max; memory up to 10 GB; Graviton for 20% cheaper.
- **Lambda@Edge** — run at CloudFront POPs; redact PII at ingress, route by header, A/B prompts.
- **AWS App Runner** CLOSED · ON GUIDE — fully managed HTTP container / source-code service with built-in autoscaling and HTTPS. Closed to new customers but still on the AIP-C01 guide; for a new small GenAI app, AWS points to **Amazon ECS Express Mode** (one call creates an ECS service on Fargate with an ALB and autoscaling).

- **Amazon EC2** — when you need GPU / Neuron accelerators: *p5/p4d* (NVIDIA), *trn2/trn1* (Trainium training), *inf2/inf1* (Inferentia inference), *g5/g6* (general ML). Capacity Blocks for ML give reserved GPU capacity windows.
- **Containers** — ECS + Fargate for app tier; EKS for portable multi-tenant + GPU workloads (Karpenter for GPU node scale-up). ECR for images (Enhanced scanning via Inspector).
- **AWS Outposts / Wavelength** — run inference on-prem or in carrier 5G edges (data-residency or mobile latency).

9 · Storage & Data for GenAI

- **Amazon S3** — primary datalake + KB source. Versioning, **S3 Intelligent-Tiering**, **S3 Lifecycle policies** → IA / Glacier / Deep Archive, **S3 Cross-Region Replication** (RTC for 15-min SLA). Event Notifications trigger Lambda for new-doc ingest.
- **Amazon EBS** — gp3 default; io2 Block Express for high-IOPS model hosts.
- **Amazon EFS** — shared NFS for training workstations / Studio users.
- **AWS DataSync** — sync on-prem / other clouds to S3 for ingestion.
- **AWS Transfer Family** — SFTP / FTPS / FTP / AS2 / WebApp landing for partner data uploads.
- **AWS Glue** — serverless ETL with Data Catalog + Data Quality; build tabular features and document corpora.
- **Amazon EMR** — Spark / Hive / Iceberg at scale for large-corpus prep; EMR Serverless when spiky.
- **Amazon Kinesis / Amazon MSK** — streaming ingestion; use with Lambda or Flink / Kafka Connect to build real-time RAG update pipelines.
- **Amazon Athena / Amazon Quick Sight** — SQL over S3 + dashboards (e.g., monitor prompt usage + cost).

10 · Database Options for GenAI Apps

Service	Role
Amazon Aurora	MySQL / PostgreSQL compatible; *pgvector* extension for embeddings; Aurora Serverless v2; Global Database for cross-Region; zero-ETL to Redshift.
Amazon RDS	Managed MySQL / PostgreSQL / MariaDB / Oracle / SQL Server / Db2. Use PostgreSQL + pgvector when you need full RDS features (e.g., RDS Proxy).
Amazon DynamoDB · DynamoDB Streams	Single-digit-ms NoSQL for session + chat-history store. DynamoDB Streams trigger Lambda for async tasks (index new docs, send moderation job, sync to OpenSearch).
Amazon DocumentDB	MongoDB-compatible; vector search available on same collection.
Amazon ElastiCache	Redis OSS / Valkey / Memcached — prompt / response cache; semantic cache layer in front of Bedrock.
Amazon Neptune	Graph for GraphRAG; analytics variant for embeddings + graph in one store.
Amazon OpenSearch Service	Full-text + vector; SAML / IAM; serverless or managed. BM25 + k-NN + Neural plugin.

11 · Application Integration & Orchestration

Service	Purpose
Amazon EventBridge	Event bus (default, custom, partner), Pipes, Scheduler. Fire workflows on new S3 doc / finished ingestion / Bedrock finish.
AWS Step Functions	Standard (durable, < 1 y) + Express (streaming, < 5 min). Retry/Catch/parallel/Map; great for circuit-breaker and ReAct patterns around Bedrock calls.
Amazon SNS / SQS	Decouple producers from workers (moderation queue, async long prompts with DLQ).
Amazon AppFlow	Managed connectors for SaaS → AWS (Salesforce, Zendesk, Slack, Marketo, SAP) → feed KB / datalake.

Service	Purpose
AWS AppConfig	Runtime feature-flag and config deployment with validators + CloudWatch alarm rollback. Swap FM / prompts / guardrails without redeploy.
AWS Chatbot	Slack / Microsoft Teams / Amazon Chime integration for ops events — great for human-in-the-loop approvals.

12 · API, Front-End & Edge

- **Amazon API Gateway** — REST / HTTP / WebSocket in front of Lambda; IAM, Cognito, Lambda authorizers, mTLS (REST), usage plans for API keys.
- **AWS AppSync** — managed GraphQL (and *AppSync Events* for subscriptions); good fit for chat / streaming LLM token-by-token delivery.
- **Amazon CloudFront** — edge delivery + Lambda@Edge/CloudFront Functions for request shaping; signed URLs for private model artefacts; WAF integration.
- **AWS Global Accelerator** — static anycast IPs + AWS backbone; ideal for non-HTTP / low-jitter voice agents.
- **AWS PrivateLink** — expose internal GenAI service privately to partner VPCs; VPC interface endpoints for Bedrock / SageMaker.
- **Elastic Load Balancing** — ALB (routing, weighted blue/green for new model versions), NLB (static IPs, TLS passthrough), GWLB (inline security appliances).
- **Amazon Route 53** — weighted / latency / failover records for multi-Region FM deployments.
- **AWS Amplify** — hosting + auth + APIs for front-end GenAI apps (React / Next.js) with built-in Cognito + AppSync + Lambda wiring.

13 · Security, Identity & Compliance for GenAI

Identity

- **AWS IAM** — least-privilege roles for Lambda / SageMaker / EC2 / Bedrock. Conditions: `aws:PrincipalOrgID`, `aws:SourceArn`, `aws:SecureTransport`, `aws:MultiFactorAuthPresent`.
- **IAM Access Analyzer** — external access, unused access, custom policy checks, policy generation from CloudTrail.
- **AWS IAM Identity Center** — workforce SSO; used by Amazon Quick, Amazon Q Business and SageMaker Unified Studio.
- **Amazon Cognito** — end-user auth for GenAI apps (User Pools + Identity Pools), federation with social/OIDC/SAML.

Data Protection

- **AWS KMS** — customer-managed keys for S3 buckets holding training data, Bedrock custom models, SageMaker artefacts; `kms:ViaService` condition to scope.
- **AWS Encryption SDK** — client-side envelope encryption; when data must be encrypted before S3 / Kinesis.
- **AWS Secrets Manager** — rotate API keys (third-party APIs, vendor SaaS, DB creds) referenced by Lambda at runtime.
- **Amazon Macie** — scan S3 corpora for PII / regulated data before ingestion into a KB.
- **AWS WAF** — on API Gateway / ALB / CloudFront / AppSync; rate-based rules on chat endpoints protect from prompt-injection scraping and token-flood attacks.

Responsible-AI Controls

- **Bedrock Guardrails** — see §2; enforce denied topics, PII redaction, contextual grounding, prompt-injection detection.
- **Input validation** — Comprehend PII detection + custom Lambda filters before sending to FM; strip / redact / refuse.
- **Output filtering** — re-run through Guardrails on response; log decisions to CloudTrail + S3.
- **Data residency** — choose Region with Bedrock access + cross-Region inference off if required.
- **Customer data NOT used for training** by Bedrock, Q, SageMaker — document in data-handling addendum.

14 · Observability & Operations

Service	Use
Amazon CloudWatch	Bedrock metrics (*InputTokenCount*, *OutputTokenCount*, *InvocationLatency*, *InvocationClientErrors*), custom app metrics (hallucination rate, grounding score), dashboards, alarms (standard / composite / anomaly).
Amazon CloudWatch Logs	Lambda + Bedrock Model Invocation Logging (S3 + CW Logs). Logs Insights for slow-latency / refusal analysis. Subscription filters → Firehose → OpenSearch.
Amazon CloudWatch Synthetics	Canary for chat endpoints; scheduled prompt regressions to detect model drift after version swaps.
AWS X-Ray	Distributed tracing across API GW → Lambda → Bedrock + KB + tools; pinpoint RAG vs inference latency.
AWS CloudTrail	Full API audit: who called which Bedrock model, which Guardrail, which KB. Org trail to Log Archive.
Amazon Managed Grafana	Dashboards merging CloudWatch + Prometheus + X-Ray + OpenSearch; SSO via IAM Identity Center.
AWS Systems Manager	Parameter Store for prompt templates, Patch Manager for GPU node hygiene, OpsCenter for prod incidents (Incident Manager is closed to new customers), AppConfig for feature flags.
AWS Service Catalog	Curated CFN products — approved Bedrock app stacks, approved KB wiring, Guardrail defaults.
AWS Auto Scaling	Target tracking for SageMaker endpoint instances; predictive for known traffic shapes.
AWS Well-Architected Tool (*Generative AI Lens*)	Review against prescriptive pillar for GenAI workloads.
AWS Cost Anomaly Detection / Cost Explorer	Flag runaway token spend (classic — agent loop burning Opus tokens). Alert via SNS / Slack.

15 · DevOps for GenAI Applications

- **AWS CDK / AWS CloudFormation** — IaC for AgentCore agents, KBs, Guardrails, Prompt Flows, SageMaker endpoints, Cognito, API GW.
- **AWS CodePipeline + CodeBuild + CodeDeploy + CodeArtifact** — pipeline building + artefact management. Promotion from dev → staging → prod with model-eval gates.
- **AWS Amplify** — front-end hosting + CI/CD for the React / Next.js / Vue app.
- **AWS CLI + AWS Tools and SDKs** — CLI / Python / JS / Java / .NET; LangChain + LlamaIndex + Strands Agents interoperate through them.
- **Model promotion** — SageMaker Model Registry approval → EventBridge → pipeline deploy to endpoints. For Bedrock custom models, promote new Provisioned Throughput alias.
- **Blue/green on endpoints** — SageMaker production variants with shadow + A/B weighting; for Bedrock, use alias-like config via AppConfig + two Provisioned-Throughput model ARNs.

16 · Testing, Validation & Metrics

Evaluation Types

- **Automatic (Bedrock Model Evaluation)** — curated benchmark datasets + metrics: BERTScore (semantic similarity), F1, Exact Match, Real-world Knowledge, Robustness, Toxicity, Accuracy.
- **Human evaluation** — your own team via Bedrock Model Evaluation human jobs; Ground Truth CLOSED · ON GUIDE for labelled comparisons.

- **SageMaker Clarify FM evaluation** CLOSED · ON GUIDE — toxicity, factual knowledge, accuracy, semantic robustness. New designs: Bedrock Model Evaluation.
- **RAG-specific** (RAGAS-style): context precision / recall, answer relevance, faithfulness (grounding). Implement via Step Functions calling an evaluator model.
- **Generation metrics** — BLEU / ROUGE / METEOR (reference-based), BERTScore (semantic), perplexity (for LMs).

Troubleshooting Checklist

- **Hallucinations** → add RAG with reranking; enable Bedrock Guardrails contextual-grounding check; reduce temperature; include "If unsure, say so" instruction; fine-tune for style.
- **Prompt injection** → Bedrock Guardrails prompt-attack filter; separate system + user turns; strip/flag injected patterns; rate-limit + WAF.
- **Latency too high** → switch to Nova Micro / Haiku / smaller model; latency-optimized inference; parallelise retrieval + prompt; caching with ElastiCache; async via Step Functions Express.
- **Throttling / limits** → cross-Region inference; provisioned throughput; exponential back-off with jitter; queue into SQS + worker.
- **Agent loops / runaway cost** → Step Functions with max-iteration state; Lambda timeouts; AppConfig kill-switch; Cost Anomaly Detection alarm.
- **Drift in retrieval quality** → scheduled RAGAS eval via CloudWatch Synthetics; re-embed with updated model; re-chunk; update KB.
- **Answer quality drop after model swap** → A/B via AppConfig weights + eval harness; rollback via AppConfig; keep Bedrock model ID and inference params in Parameter Store.

17 · Reference Architectures

A · Document-grounded chatbot (RAG)

- S3 corpus → Bedrock Knowledge Base (OpenSearch Serverless or Aurora pgvector).
- Frontend (Amplify + Cognito) → API Gateway / AppSync → Lambda → Bedrock *RetrieveAndGenerate*.
- Guardrails for PII + denied topics; contextual grounding check; prompt stored in Prompt Management.
- Conversation history in DynamoDB; semantic cache in ElastiCache Redis.
- Observability: CloudWatch metrics, X-Ray traces, Bedrock Model Invocation Logging to S3 for audit + eval.

B · Agentic assistant with internal tools

- Bedrock AgentCore Runtime + Gateway (MCP tool surface) + Memory + Identity; Gateway tools call internal APIs via Lambda.
- Knowledge Base attached; Guardrails; Bedrock Model Evaluation periodically against scripted scenarios.
- Step Functions around agent for circuit-breaker; CloudWatch alarm stops execution on cost-per-request spike.

C · Fine-tuned domain model on SageMaker

- Data curation in Data Wrangler; preference data from your own reviewers (Ground Truth on older exam answers, closed to new customers); Processing jobs for tokenisation.
- SageMaker Training with LoRA (parameter-efficient); JumpStart base model.
- Evaluation with Bedrock Model Evaluation (Clarify FM evaluation on older exam answers); register in Model Registry with approval gate.
- Deploy to real-time endpoint with Auto Scaling + CloudWatch drift alarms (Model Monitor on older exam answers); expose via API Gateway + Cognito.

D · Multimodal intake pipeline

- Ingest: S3 (docs/images), Transcribe (audio), Rekognition (video / images) → normalise → Comprehend PII scrub → Bedrock multimodal FM.
- Output fanned-out via SNS; Step Functions orchestrates human review of low-confidence outputs (task token + review UI; on the exam, Augmented AI (A2I), closed to new customers).

E · Enterprise assistant (Amazon Q Business on the exam, Amazon Quick for new builds)

- IAM Identity Center as identity source; data sources = S3, Confluence, ServiceNow, SharePoint.
- Guardrails: allowed topics, blocked phrases, response customisation; admin controls per group.
- Q Business Apps created by builders for self-service workflows (e.g., HR questions, IT tickets).
- Q Business is closed to new customers; AWS points new deployments to Amazon Quick.

18 · Scenario → Answer Patterns

Scenario	Answer
Answer questions from constantly-updated docs without re-training	Bedrock Knowledge Bases + RetrieveAndGenerate (managed RAG).
Expose internal APIs to an FM with auth + memory + observability	Bedrock AgentCore (Runtime + Gateway + Memory + Identity + Observability).
Block the FM from talking about competitors	Bedrock Guardrails denied topics + prompt-injection filter.
Redact PII in incoming chat + outgoing responses	Comprehend PII detection + Bedrock Guardrails sensitive-info filter.
Fine-tune a domain FM with limited labelled data	SageMaker JumpStart + LoRA / PEFT; register in Model Registry; deploy to real-time endpoint.
Avoid regional capacity throttling on flagship models	Bedrock Cross-Region Inference.
Hard rollback if a new model version regresses	AppConfig feature flag with CloudWatch alarm auto-rollback; keep model ID in Parameter Store.
Cheap overnight batch classification on FMs	Bedrock Batch Inference (JSONL in / out via S3), ~50% cheaper.
Real-time low-latency semantic cache before Bedrock	ElastiCache for Redis / MemoryDB + Titan embeddings for key; exact-or-near-match shortcut.
Vector store inside existing Postgres app	Aurora PostgreSQL + pgvector.
GraphRAG over entities + relationships	Neptune Analytics (graph + vector in one store).
Human review for low-confidence outputs	Amazon Augmented AI (A2I) + Step Functions (A2I is closed to new customers but still on the guide).
Chat UI streaming tokens at scale	AppSync (GraphQL subscriptions) or API Gateway WebSocket in front of Lambda + Bedrock streaming.
Enterprise GenAI assistant with workforce SSO + connectors	Amazon Q Business + IAM Identity Center on the exam; Amazon Quick for a new build.
Identify runaway Bedrock spend	AWS Cost Anomaly Detection + Budgets alert via SNS; Cost Explorer filtered by service = Bedrock.
Track which user ran which prompt	Bedrock Model Invocation Logging to S3 + CloudTrail; Athena for audit queries.
Compare two models on your own dataset	Bedrock Model Evaluation (automatic, LLM-as-a-judge or human).
Reduce hallucinations with contextual grounding	Bedrock Guardrails contextual-grounding + relevance checks + reranking top-k.
Deploy a GenAI web app with auth	Amplify Hosting + Cognito + AppSync + Lambda + Bedrock.

Scenario	Answer
Expose Bedrock privately to a partner VPC	PrivateLink interface endpoints for Bedrock + resource-based policies.
Orchestrate multi-agent specialists	Bedrock AgentCore + Agent Squad or Strands Agents with supervisor agent.
Swap prompt templates without redeploy	Bedrock Prompt Management + AppConfig for runtime prompt ID.
Data residency — keep training data in a specific Region	S3 in that Region only; Bedrock model in same Region; disable cross-Region inference; Control Tower SCP pinning Region.

19 · Common AIP-C01 Pitfalls

- **"Use prompt engineering first"** — every exam tree bottoms out on prompt/RAG before fine-tuning. Don't jump to fine-tune.
- **Prompt Flows ≠ Step Functions** — Prompt Flows is prompt-focused visual orchestration (no human approvals, no long-running state). Use Step Functions when you need durable state + complex orchestration.
- **AgentCore, not Bedrock Agents** — AgentCore is the agent platform on the AIP-C01 guide (framework-agnostic: Runtime, Gateway, Memory, Identity, Observability). Bedrock Agents Classic is closed to new customers DEPRECATED; if an older question offers it, AgentCore is the current answer.
- **Model Context Protocol (MCP)** is the canonical tool-surface protocol now — expect it in questions about connecting agents to existing tools.
- **Bedrock Model Invocation Logging** is *opt-in*. Without it, you have no body-level audit trail; CloudTrail records only the API call metadata.
- **Knowledge Base vs Kendra** — KB is RAG-optimised with vector store; Kendra is enterprise search with managed connectors + ranking. Kendra is closed to new customers (still on the guide): for a new build, pick Knowledge Bases.
- **Guardrails apply to both input + output** — don't split detection from response filtering.
- **Provisioned Throughput** is per-model-unit commitment; required for custom models; gives deterministic latency. Not the same as "just reserve capacity globally".
- **SageMaker endpoints scale independently** of Bedrock. Right-size with CloudWatch metrics + Auto Scaling.
- **Tokens dominate cost** — cache, compress, summarise chat history, use smaller models for retrieval-only steps.

AIP-C01 default answers: "Answer from private docs" → Bedrock KB + Guardrails. "Agent with auth + memory + tools" → AgentCore. "Safety controls" → Guardrails (denied topics, PII, prompt attack, contextual grounding). "Evaluate model" → Bedrock Model Evaluation. "Cut latency / cost" → smaller model, Batch Inference, caching, Cross-Region Inference. "Swap behaviour without redeploy" → AppConfig + Prompt Management.

20 · Prompt Engineering Deep Dive

Foundational patterns

- **Zero-shot** — instructions only, no examples. Cheapest.
- **Few-shot** — include 2–8 input/output demonstrations. Use when task format matters.
- **Instruction** — direct command with constraints ("Respond in JSON with keys X, Y").
- **Role / persona** — system prompt sets expertise + voice ("You are a compliance analyst…").
- **Contextual** — inject retrieved documents between system + user turns; label as "Context" and instruct to use it only.
- **Style / template** — constrain with regex-like grammar or output-format spec; pair with Bedrock tool-use for guaranteed JSON.

Reasoning patterns

- **Chain-of-Thought (CoT)** — "Think step by step." Useful for math, logic, multi-hop.

- **Self-Consistency** — sample N CoT paths and majority-vote the answer; more reliable for hard tasks.
- **Tree-of-Thoughts** — explore branches + backtrack; orchestrate with Step Functions or Prompt Flows.
- **ReAct (Reasoning + Acting)** — alternate Thought → Action (tool call) → Observation → Thought. Native to AgentCore and Strands Agents.
- **Reflection** — model critiques its own answer then revises.
- **Plan-and-Solve** — plan steps first, then execute; reduces omissions on complex tasks.

Structural patterns

- **Structured output** — request specific JSON schema; validate with Lambda; reprompt on parse failure. Bedrock tool-use guarantees schema-conformant outputs.
- **Chaining** — split complex ask into sequential sub-prompts (summarise → classify → draft → review).
- **RAG** — retrieve before generate; always include "Cite the source" instruction.
- **Programmatic guardrails** — pre-filter (Comprehend PII, allowlist intents), post-filter (Guardrails, regex, denylists).

Inference parameters — practical settings

Goal	Settings
Deterministic, factual Q&A	temperature 0 – 0.2, top-p 0.9, max_tokens tight.
Creative writing	temperature 0.7 – 0.9, top-p 0.95, higher max_tokens.
Structured output / code	temperature 0 – 0.3, stop sequences on fences, use tool-use.
Brainstorm / variety	temperature 0.8 – 1.0, top-k 50+, increase frequency penalty.

21 · Responsible AI — Practical Checklist

- **Fairness & bias** — pre-train + post-train bias metrics (SageMaker Clarify on the guide, closed to new customers); run on evaluation sets representative of protected groups.
- **Explainability** — SHAP values (SHAP library or Clarify); citations via Bedrock KB; record reasoning traces for agent decisions.
- **Privacy** — Macie on ingestion buckets; Comprehend PII + Bedrock Guardrails sensitive-info filter; KMS at rest; TLS in transit; least-privilege IAM.
- **Safety** — Guardrails content filters + denied topics; test with red-team prompts; rate-limit + WAF.
- **Accountability** — CloudTrail + Bedrock Model Invocation Logging; immutable S3 log with Object Lock; tag every request with user + session + prompt version.
- **Transparency** — disclose model provider + capabilities + limitations to end users; publish data-handling policy.
- **Veracity** — contextual-grounding + relevance checks in Guardrails; RAGAS or Bedrock Model Evaluation scheduled; human review for high-impact outputs.
- **Governance** — Bedrock Prompt Management with versioning + approvals; Service Catalog only for approved FM stacks; SCP denying un-allowlisted model IDs.

22 · Cost Control for GenAI Workloads

- **Model choice** — pick the smallest model that meets quality. Nova Micro → Lite → Pro → Premier ladder; Haiku → Sonnet → Opus for Claude.
- **Token minimisation** — summarise chat history; compress retrieved context; cap max_tokens; strip prompts of boilerplate.
- **Caching** — ElastiCache / MemoryDB semantic cache; Bedrock prompt caching (new) reuses repeated context at reduced cost.
- **Batch Inference** — ~50% off for async workloads (summaries, classification, embeddings backfill).
- **Provisioned Throughput** — reserve units for steady high-rate; otherwise pay On-Demand.
- **Cross-Region Inference** — can reduce effective cost by load-spreading for capacity-limited flagship models.

- **Right-size endpoints** — SageMaker Auto Scaling target tracking; shut down dev endpoints overnight via EventBridge Scheduler.
- **Observability** — CloudWatch metric *OutputTokenCount* + custom metric per prompt ID; Cost Anomaly Detection monitors + alerts.

23 · Multi-Region & Resilience for GenAI

- **Bedrock Cross-Region Inference** — route across a profile of supported Regions; first-line resilience for flagship model availability.
- **Multi-Region KB** — replicate S3 source via CRR; duplicate KB + vector store in target Region.
- **SageMaker endpoint** — deploy to 2+ Regions, Route 53 failover or latency records.
- **Stateful app tier** — Aurora Global Database; DynamoDB Global Tables; ElastiCache Global Datastore.
- **Graceful degradation** — circuit breaker in Step Functions; fallback to smaller model; cached "safe" response; static doc hand-off.
- **Chaos** — inject Bedrock throttles via AppConfig feature flag + kill-switch; verify fallback path.

24 · Operational Runbooks

Model latency regression

1. Check CloudWatch Bedrock *InvocationLatency* p95/p99 per model; correlate with deploy event via EventBridge.
2. X-Ray trace — RAG vs inference vs guardrails breakdown.
3. Flip AppConfig weight back to last-known-good model alias.
4. File Bedrock Model Evaluation job on last 1% sample to measure quality delta.

Prompt injection incident

1. Identify via Guardrails prompt-attack blocked event in CloudTrail + Model Invocation Logging.
2. Quarantine user session via API Gateway usage plan / Cognito disable.
3. Add WAF rate-based rule; tune Guardrails word filters; escalate to security.
4. Replay session through Athena on log bucket for forensic scope.

KB data drift (answers stale)

1. Check KB ingestion job status; schedule incremental re-ingest via EventBridge Scheduler.
2. Re-embed with updated embedding model; run RAGAS eval harness via Synthetics.
3. Update chunking strategy if recall < target.

Agent cost runaway

1. Cost Anomaly Detection alarm fires; pinpoint by Bedrock usage tag (user / app).
2. AppConfig kill-switch disables agent; smaller model fallback.
3. Add Step Functions *max-iteration* state + Lambda timeouts; reduce tool call budget; enable prompt caching.

25 · Closed to New Customers or End of Support

Know these for the exam (most are still on the AIP-C01 in-scope list); for a new design, pick the replacement.

Service	Status → What to use instead
Bedrock Agents Classic DEPRECATED	Closed from 30 Jul 2026, not on the guide → Bedrock AgentCore.
SageMaker Ground Truth · Amazon A2I CLOSED · ON GUIDE	Closed from 30 Jul 2026 → Bedrock human evaluation jobs; Step Functions review steps.
SageMaker Clarify CLOSED · ON GUIDE	Closed from 30 Jul 2026 → Bedrock Model Evaluation (FMs), SHAP library.
SageMaker Model Monitor CLOSED · ON GUIDE	Closed from 30 Jul 2026 → MLflow + CloudWatch drift monitoring.
Amazon Kendra CLOSED · ON GUIDE	No longer open to new customers → Bedrock Knowledge Bases.
Amazon Q Business · Q Business Apps CLOSED · ON GUIDE	No longer open to new customers → Amazon Quick.
AWS App Runner CLOSED · ON GUIDE	No longer open to new customers → Amazon ECS Express Mode.
Amazon Q Developer END OF SUPPORT · ON GUIDE	IDE plugins + Pro (incl. /transform) end of support 30 Apr 2027 (console unaffected); the CLI is now Kiro → Kiro; /transform → AWS Transform.
Systems Manager Incident Manager DEPRECATED	Closed to new customers since 7 Nov 2025, not on the guide → Systems Manager OpsCenter, or partner paging tools.

Built by **CloudNinja** · **cloudninja.pro** · Free AWS practice exams & cheat sheets. No signup, no paywall.