AWS Certified AI Practitioner Cheat Sheet

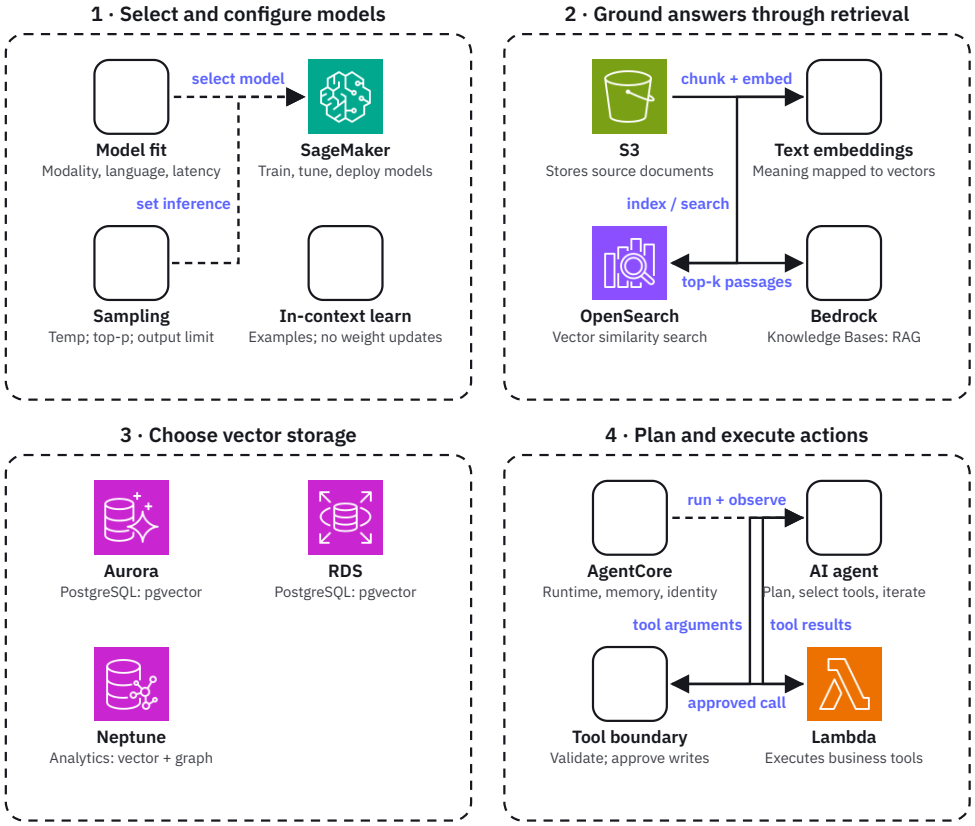
AIF-C01 · Domain 3 of 5

BlowStack

DOMAIN 3 · APPLICATIONS OF FOUNDATION MODELS · 28% OF THE EXAM · TASK STATEMENT 3.1

Describe design considerations for applications that use foundation models (FMs)

Select, ground, store, act



1 Select and configure models

Choose by **workload fit**: cost, model size/complexity, customization support, and input/output limits. Supported prompt caching reuses repeated prefixes to cut input cost and latency; it does not enlarge the context window.

2 Ground answers through retrieval

RAG embeds the query, retrieves relevant passages, and adds them to the prompt without updating weights. Knowledge Bases manages retrieval for enterprise Q&A; semantic search matches meaning rather than exact words.

3 Choose vector storage

Vector storage options include OpenSearch for search workloads, Aurora PostgreSQL and RDS for PostgreSQL with pgvector, and Neptune Analytics for vector similarity combined with graph relationships.

4 Plan and execute actions

Use **agents** for multistep support, such as checking orders and issuing approved refunds. The FM selects tools; the runtime executes them. AgentCore supplies infrastructure, not automatic business authorization.

Exam traps

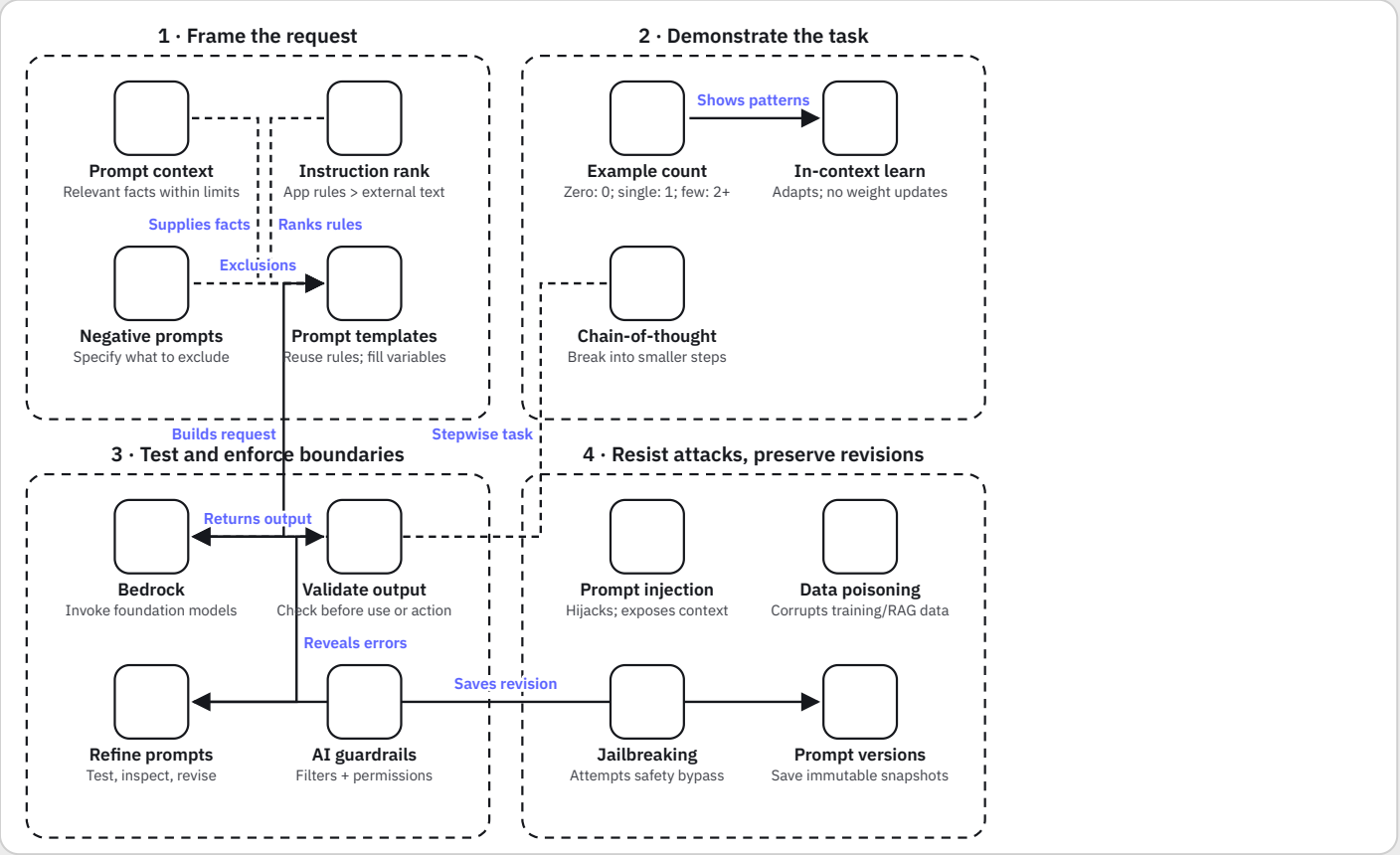
- **Sampling**: low temperature favors likely tokens, not truth; top-p limits choices by cumulative probability. Max output caps generation; input, retrieved context, and output share the context window.
- **Customization cost**: pre-training from scratch costs most; fine-tuning adds training and hosting costs. Distillation pays for teacher data/training to lower student serving cost, subject to quality.
- **RAG versus tuning**: RAG refreshes external knowledge; fine-tuning changes weights for behavior/style. In-context learning needs no training, but longer prompts increase inference cost and latency.

Get practice exams and lessons for AWS Certified AI Practitioner

DOMAIN 3 · APPLICATIONS OF FOUNDATION MODELS · TASK STATEMENT 3.2

Choose effective prompt engineering techniques

Frame, demonstrate, validate, version



1 Frame the request

Use **specific, concise instructions**, clearly separated context, and an explicit output format. Templates standardize repeated requests; negative prompts complement positive directions rather than enforce them.

2 Demonstrate the task

Choose **few-shot prompting** when examples clarify a pattern or format; zero-shot supplies instructions without examples. Chain-of-thought can help multistep tasks. More context can increase inference cost and latency.

3 Test and enforce boundaries

Use **iterative refinement** to improve quality against representative tests. Independently enforced guardrails, output checks, least-privilege tool permissions, and monitoring complement prompts; prompting alone is not enforcement.

4 Resist attacks, preserve revisions

Amazon Bedrock Prompt Management supports drafts, variables, testing, and **immutable versions** for reproducible reuse. Track the deployed version to recover from regressions; protect data provenance against poisoning.

Exam traps

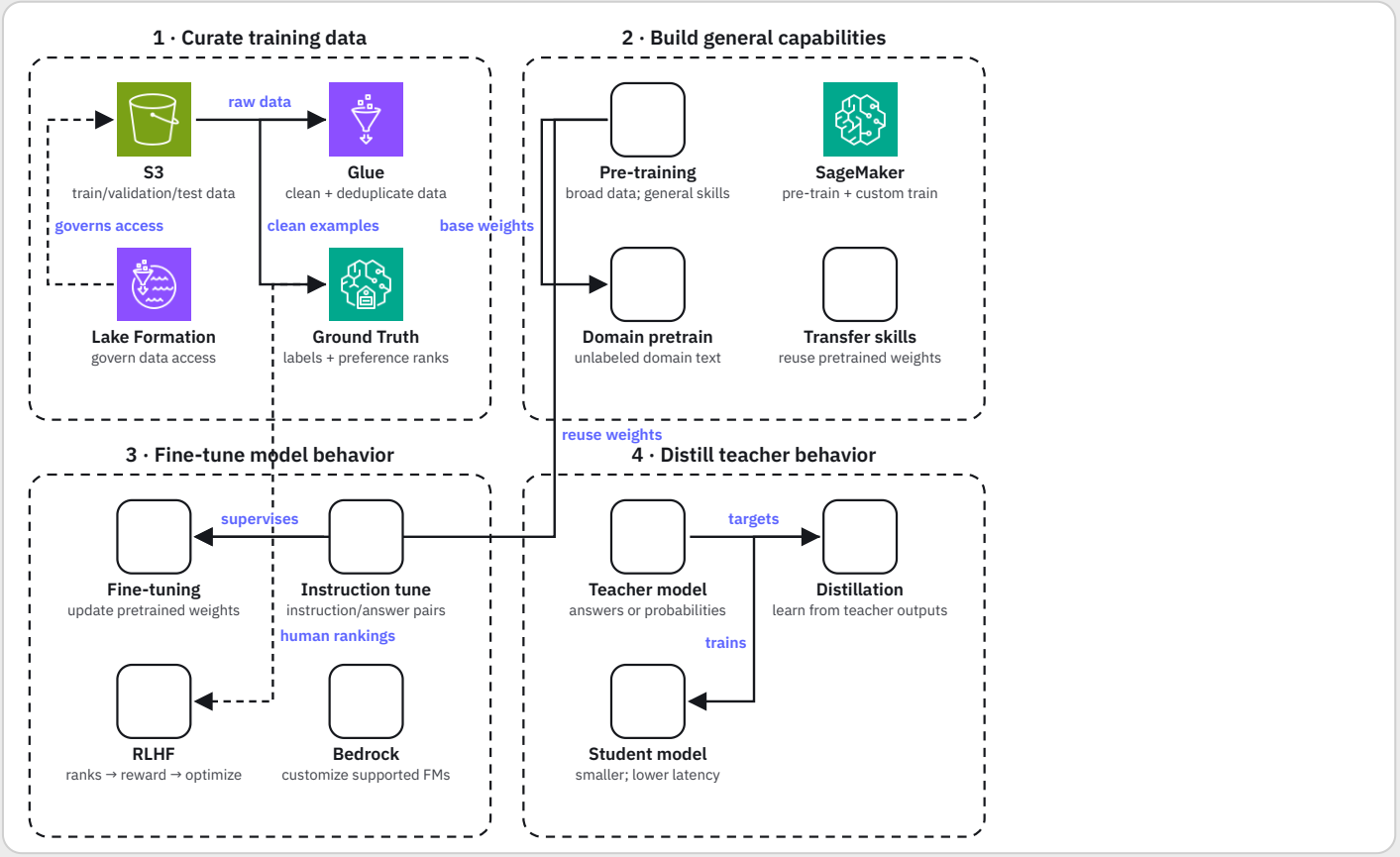
- **Few-shot is not fine-tuning**: examples guide the current inference without updating model weights. Adding examples is not a training job.
- **Instructions are not security boundaries**: hierarchy support varies by model and API. Negative prompts and stepwise explanations guarantee neither safe behavior nor correct answers.
- **Injection versus poisoning**: injection disguises commands in input or retrieved content; poisoning corrupts data sources. Jailbreaking targets safeguards; these attacks can overlap.

Get practice exams and lessons for AWS Certified AI Practitioner

DOMAIN 3 · APPLICATIONS OF FOUNDATION MODELS · TASK STATEMENT 3.3

Describe the training and fine-tuning process for FMs

Curate, pretrain, specialize, distill



1 Curate training data

Data curation requires clean, representative, correctly labeled data and sensitive-data controls. Dataset size depends on model/task; separate training and evaluation data to prevent leakage.

2 Build general capabilities

Continued pre-training uses unlabeled domain text; pre-training learns general capabilities from broad data. Transfer learning reuses learned representations or weights, reducing training time and data needs.

3 Fine-tune model behavior

Instruction tuning learns from task/answer examples. Reinforcement Learning from Human Feedback commonly learns a reward model from human rankings, then optimizes behavior. Customization support varies by model.

4 Distill teacher behavior

Knowledge distillation trains a smaller student using teacher outputs to reduce inference cost, memory, and latency. Capability retention is not guaranteed; the student requires evaluation.

Exam traps

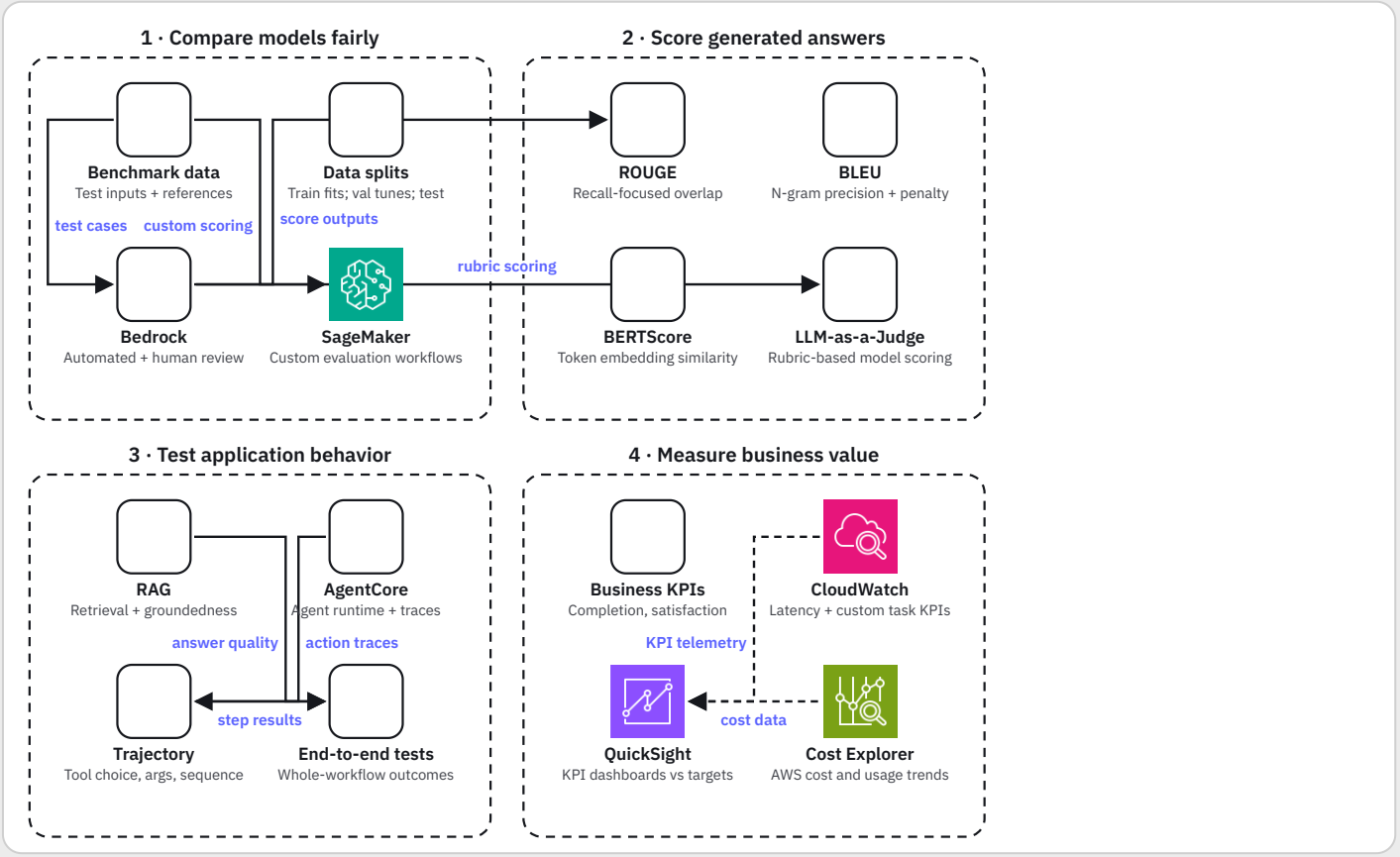
- **Prompting and RAG do not update model weights**; fine-tuning does. Retrieving changing business facts does not automatically require retraining.
- **Continued pre-training is not instruction tuning**: unlabeled domain documents teach domain familiarity; instruction/answer examples teach desired responses.
- **Human labeling alone is not RLHF**. Ground Truth can collect labels or preferences; RLHF additionally uses human feedback to optimize model behavior.

Get practice exams and lessons for AWS Certified AI Practitioner

DOMAIN 3 · APPLICATIONS OF FOUNDATION MODELS · TASK STATEMENT 3.4

Describe methods to evaluate FM performance

Compare, score, test, measure



1 Compare models fairly

Representative benchmarks enable repeatable comparisons. **Human review** assesses usefulness, factual correctness, safety, and appropriateness. Training fits, validation guides choices, and held-out testing estimates generalization.

2 Score generated answers

ROUGE commonly scores summaries; BLEU scores translation. BERTScore tolerates paraphrases. **LLM-as-a-Judge** scales rubric-based assessment but requires calibration against human judgment and checks for bias.

3 Test application behavior

RAG evaluation checks retrieval relevance and answer grounding. **Trajectory evaluation** checks agent actions and tool use; end-to-end testing checks realistic outcomes across retrieval, prompts, tools, and safeguards.

4 Measure business value

Define **business targets** and review periods for completion rate, satisfaction, productivity, and engagement. Combine application counts with attributable spend to calculate cost per interaction; technical scores alone do not prove value.

Exam traps

- **Test-set tuning** leaks evaluation information into development. Validation guides model and prompt choices; an untouched test set provides the independent final assessment.
- **High overlap is not factual accuracy**: BLEU and ROUGE reward shared wording. BERTScore measures semantic similarity, not proof of correctness, safety, or business value.
- **A plausible final answer can hide failure**: agent trajectories may contain unsafe tool calls or unnecessary steps. Task completion, satisfaction, latency, and cost remain separate checks.

Get practice exams and lessons for AWS Certified AI Practitioner

DOMAIN 3 · APPLICATIONS OF FOUNDATION MODELS · WHAT TO REMEMBER

Concepts and services of Applications of Foundation Models

Concepts to know

Retrieval-Augmented Generation

RAG retrieves external information and adds it to the model's context. Choose it for **current or organization-specific knowledge without retraining**; answer quality depends on retrieval relevance and grounding.

Fine-Tuning

Fine-tuning updates a pretrained model's weights using curated examples. Choose it to **specialize behavior, style, or task performance**, accepting training and maintenance costs; it is not a substitute for retrieving frequently changing facts.

In-Context Learning

In-context learning guides a model through instructions and examples without updating weights. Choose **zero-shot, single-shot, or few-shot prompting** according to the examples needed; longer prompts consume context and increase inference cost.

Core services

Amazon Bedrock

Choose **managed foundation model access** with capabilities for Knowledge Bases, agents, guardrails, prompt management, customization, and model evaluation without managing model-serving infrastructure.

Amazon OpenSearch Service

Choose **vector similarity search** over embeddings to retrieve relevant content for RAG; hybrid search combines lexical matching with semantic retrieval.

Amazon SageMaker

Choose **greater control over training and deployment** for custom ML workflows, including foundation model fine-tuning, evaluation, and hosting.

Get practice exams and lessons for AWS Certified AI Practitioner