



# AWS AI Practitioner AIF-C01 Exam Questions

**Total Questions: 150+**

**Demo Questions: 40**

**Version: Updated for 2025**

**Prepared and Verified by Cert Empire – Your Trusted IT Certification Partner**

**For Access to the full set of Updated Questions – Visit:  
[AIF-C01 Exam Dumps](#) by Cert Empire**

## Question: 1

A company is developing a mobile ML app that uses a phone's camera to diagnose and treat insect bites. The company wants to train an image classification model by using a diverse dataset of insect bite photos from different genders, ethnicities, and geographic locations around the world. Which principle of responsible AI does the company demonstrate in this scenario?

- A. Fairness
- B. Explainability
- C. Governance
- D. Transparency

### Answer:

A

### Explanation:

The company is training an image classification model for diagnosing insect bites using a diverse dataset that includes photos from different genders, ethnicities, and geographic locations. This approach demonstrates the principle of fairness in responsible AI, as it aims to reduce bias and ensure the model performs equitably across diverse populations.

Exact Extract from AWS AI Documents:

From the AWS AI Practitioner Learning Path:

"Fairness in AI involves ensuring that models do not exhibit bias against certain groups and perform

equitably across diverse populations. This can be achieved by training models on diverse datasets

that represent various demographics, such as gender, ethnicity, and geographic location."

(Source: AWS AI Practitioner Learning Path, Module on Responsible AI)

Detailed

Option A: Fairness This is the correct answer. By using a diverse dataset, the company ensures the

model is less likely to be biased against specific groups, promoting fairness in its predictions and treatments for insect bites.

Option B: Explainability Explainability refers to making the model's decisions understandable to users, such as by providing insights into how predictions are made. The scenario focuses on dataset

diversity, not explainability.

Option C: Governance Governance involves establishing policies and processes to manage AI systems,

such as compliance and oversight. The scenario does not describe governance mechanisms.

Option D: Transparency Transparency involves disclosing how a model works, its limitations, and its

data sources. While transparency is important, the scenario specifically highlights the diversity of the

dataset, which aligns more directly with fairness.

Reference:

AWS AI Practitioner Learning Path: Module on Responsible AI

AWS Documentation: Responsible AI Principles

(<https://aws.amazon.com/machinelearning/responsible-ai/>)

Amazon SageMaker Developer Guide: Bias and Fairness in ML

(<https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-bias.html>)

CertEmpire

## Question: 2

A company is using a large language model (LLM) on Amazon Bedrock to build a chatbot. The chatbot

processes customer support requests. To resolve a request, the customer and the chatbot must interact a few times.

Which solution gives the LLM the ability to use content from previous customer messages?

- A. Turn on model invocation logging to collect messages.
- B. Add messages to the model prompt.
- C. Use Amazon Personalize to save conversation history.
- D. Use Provisioned Throughput for the LLM.

### Answer:

B

### Explanation:

The company is building a chatbot using an LLM on Amazon Bedrock, and the chatbot needs to use

content from previous customer messages to resolve requests. Adding previous messages to the model prompt (also known as providing conversation history) enables the LLM to maintain context across interactions, allowing it to respond coherently based on the ongoing conversation.

Exact Extract from AWS AI Documents:

From the AWS Bedrock User Guide:

"To enable a large language model (LLM) to maintain context in a conversation, you can include previous messages in the model prompt. This approach, often referred to as providing conversation

history, allows the LLM to generate responses that are contextually relevant to prior interactions."

(Source: AWS Bedrock User Guide, Building Conversational Applications)

Detailed

Option A: Turn on model invocation logging to collect messages. Model invocation logging records interactions for auditing or debugging but does not provide the LLM with access to previous messages during inference to maintain conversation context.

Option B: Add messages to the model prompt. This is the correct answer. Including previous messages in the prompt gives the LLM the conversation history it needs to respond appropriately, a

common practice for chatbots on Amazon Bedrock.

Option C: Use Amazon Personalize to save conversation history. Amazon Personalize is for building recommendation systems, not for managing conversation history in a chatbot. This option is

irrelevant.

Option D: Use Provisioned Throughput for the LLM. Provisioned Throughput in Amazon Bedrock ensures consistent performance for model inference but does not address the need to use previous messages in the conversation.

Reference:

AWS Bedrock User Guide: Building Conversational Applications

(<https://docs.aws.amazon.com/bedrock/latest/userguide/conversational-apps.html>)

AWS AI Practitioner Learning Path: Module on Generative AI and Chatbots

Amazon Bedrock Developer Guide: Managing Conversation Context

(<https://aws.amazon.com/bedrock/>)

CertEmpire

### Question: 3

An ML research team develops custom ML models. The model artifacts are shared with other teams

for integration into products and services. The ML team retains the model training code and data

a. The ML team wants to build a mechanism that the ML team can use to audit models.

Which solution should the ML team use when publishing the custom ML models?

- A. Create documents with the relevant information. Store the documents in Amazon S3.
- B. Use AWS AI Service Cards for transparency and understanding models.
- C. Create Amazon SageMaker Model Cards with Intended uses and training and inference details.
- D. Create model training scripts. Commit the model training scripts to a Git repository.

#### Answer:

C

#### Explanation:

The ML research team needs a mechanism to audit custom ML models while sharing model artifacts

with other teams. Amazon SageMaker Model Cards provide a structured way to document model details, including intended uses, training data, and inference performance, making them ideal for auditing and ensuring transparency when publishing models.

Exact Extract from AWS AI Documents:

From the Amazon SageMaker Developer Guide:

"Amazon SageMaker Model Cards enable you to document critical details about your machine learning models, such as intended uses, training data, evaluation metrics, and inference details. Model Cards support auditing by providing a centralized record that can be reviewed by teams to understand model behavior and limitations."

(Source: Amazon SageMaker Developer Guide, SageMaker Model Cards)

Detailed

Option A: Create documents with the relevant information. Store the documents in Amazon S3. While

storing documents in S3 is feasible, it lacks the structured format and integration with SageMaker that Model Cards provide, making it less suitable for auditing purposes.

Option B: Use AWS AI Service Cards for transparency and understanding models. AWS AI Service

Cards are not a standard feature in AWS documentation. This option appears to be a distractor and is

not a valid solution.

Option C: Create Amazon SageMaker Model Cards with Intended uses and training and inference

details. This is the correct answer. SageMaker Model Cards are specifically designed to document



model details for auditing, transparency, and collaboration, meeting the team's requirements.

Option D: Create model training scripts. Commit the model training scripts to a Git repository. Sharing

training scripts in a Git repository provides access to code but does not offer a structured auditing mechanism for model details like intended uses or inference performance.

Reference:

Amazon SageMaker Developer Guide: SageMaker Model Cards

(<https://docs.aws.amazon.com/sagemaker/latest/dg/model-cards.html>)

AWS AI Practitioner Learning Path: Module on Model Governance and Auditing

AWS Documentation: Responsible AI with SageMaker (<https://aws.amazon.com/sagemaker/>)

CertEmpire

## Question: 4

A manufacturing company uses AI to inspect products and find any damages or defects. Which type of AI application is the company using?

- A. Recommendation system
- B. Natural language processing (NLP)
- C. Computer vision
- D. Image processing

### Answer:

C

### Explanation:

The manufacturing company uses AI to inspect products for damages or defects, which involves analyzing visual data (e.g., images or videos of products). This task falls under computer vision, a type of AI application that enables machines to interpret and understand visual information, such as identifying defects in manufacturing.

Exact Extract from AWS AI Documents:

From the AWS AI Practitioner Learning Path: CertEmpire

"Computer vision enables machines to interpret and understand visual data from the world, such as

images or videos. Common applications include defect detection in manufacturing, where AI models analyze product images to identify damages or anomalies."

(Source: AWS AI Practitioner Learning Path, Module on AI Concepts)

Detailed

Option A: Recommendation system Recommendation systems suggest items or actions based on user

p

### References:

(e.g., product recommendations). They are not relevant for inspecting products for defects.

Option B: Natural language processing (NLP) NLP focuses on processing and understanding text or

speech, not visual data like product images. This option is incorrect.

Option C: Computer vision This is the correct answer. Computer vision is used for tasks like defect detection in manufacturing by analyzing visual data to identify damages or defects.

<https://certempire.com>

Option D: Image processing While image processing involves manipulating images (e.g., filtering, resizing), it is a lower-level technique, not an AI application. Computer vision, which often uses image processing as a component, is the broader AI application here.

Reference:

AWS AI Practitioner Learning Path: Module on AI Concepts

Amazon Rekognition Developer Guide: Image Analysis

(<https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html>)

AWS Documentation: Introduction to Computer Vision (<https://aws.amazon.com/computer-vision/>)

CertEmpire

## Question: 5

A financial institution is building an AI solution to make loan approval decisions by using a foundation model (FM). For security and audit purposes, the company needs the AI solution's decisions to be explainable.

Which factor relates to the explainability of the AI solution's decisions?

- A. Model complexity
- B. Training time
- C. Number of hyperparameters
- D. Deployment time

### Answer:

A

### Explanation:

The financial institution needs an AI solution for loan approval decisions to be explainable for security and audit purposes. Explainability refers to the ability to understand and interpret how a model makes decisions. Model complexity directly impacts explainability: simpler models (e.g., logistic regression) are more interpretable, while complex models (e.g., deep neural networks) are harder to explain, often behaving like "black boxes."

Exact Extract from AWS AI Documents:

From the Amazon SageMaker Developer Guide:

"Model complexity affects the explainability of AI solutions. Simpler models, such as linear regression, are inherently more interpretable, while complex models, such as deep neural networks,

may require additional tools like SageMaker Clarify to provide insights into their decision-making processes."

(Source: Amazon SageMaker Developer Guide, Explainability with SageMaker Clarify)

Detailed

**Option A: Model complexity** This is the correct answer. The complexity of the model directly influences how easily its decisions can be explained, a critical factor for audit and security purposes in loan approvals.

**Option B: Training time** Training time refers to how long it takes to train the model, which does not directly impact the explainability of its decisions.

**Option C: Number of hyperparameters** While hyperparameters affect model performance, they do not directly relate to explainability. A model with many hyperparameters might still be explainable

if

it is a simple model.

Option D: Deployment timeDeployment time refers to the time taken to deploy the model to production, which is unrelated to the explainability of its decisions.

Reference:

Amazon SageMaker Developer Guide: Explainability with SageMaker Clarify

(<https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-explainability.html>)

AWS AI Practitioner Learning Path: Module on Responsible AI and Explainability

AWS Documentation: Explainable AI (<https://aws.amazon.com/machine-learning/responsible-ai/>)

CertEmpire

## Question: 6

Which phase of the ML lifecycle determines compliance and regulatory requirements?

- A. Feature engineering
- B. Model training
- C. Data collection
- D. Business goal identification

**Answer:**

D

**Explanation:**

The business goal identification phase of the ML lifecycle involves defining the objectives of the project and understanding the requirements, including compliance and regulatory considerations. This phase ensures the ML solution aligns with legal and organizational standards before proceeding to technical stages like data collection or model training.

Exact Extract from AWS AI Documents:

From the AWS AI Practitioner Learning Path:

"The business goal identification phase involves defining the problem to be solved, identifying success metrics, and determining compliance and regulatory requirements to ensure the ML solution adheres to legal and organizational standards."

(Source: AWS AI Practitioner Learning Path, Module on Machine Learning Lifecycle)

Detailed

Option A: Feature engineering Feature engineering involves creating or selecting features for model

training, which occurs after compliance requirements are identified. It does not address regulatory concerns.

Option B: Model training Model training focuses on building the ML model using data, not on determining compliance or regulatory requirements.

Option C: Data collection Data collection involves gathering data for training, but compliance and regulatory requirements (e.g., data privacy laws) are defined earlier in the business goal identification phase.

Option D: Business goal identification This is the correct answer. This phase ensures that compliance

and regulatory requirements are considered at the outset, shaping the entire ML project.

Reference:

AWS AI Practitioner Learning Path: Module on Machine Learning Lifecycle

Amazon SageMaker Developer Guide: ML Workflow

<https://certempire.com>

(<https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-mlconcepts.html>)

AWS Well-Architected Framework: Machine Learning Lens

(<https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/>)

CertEmpire



## Question: 7

A company wants to create a chatbot that answers questions about human resources policies. The company is using a large language model (LLM) and has a large digital documentation base. Which technique should the company use to optimize the generated responses?

- A. Use Retrieval Augmented Generation (RAG).
- B. Use few-shot prompting.
- C. Set the temperature to 1.
- D. Decrease the token size.

### Answer:

A

### Explanation:

The company is building a chatbot using an LLM to answer questions about HR policies, with access to a large digital documentation base. Retrieval Augmented Generation (RAG) optimizes the LLM's responses by retrieving relevant information from the documentation base and using it to generate accurate, contextually grounded answers, reducing hallucinations and improving response quality.

Exact Extract from AWS AI Documents:

From the AWS Bedrock User Guide:

"Retrieval Augmented Generation (RAG) enhances the performance of large language models by retrieving relevant information from external knowledge bases, such as documentation or databases,

and incorporating it into the generation process. This technique ensures responses are accurate and

grounded in the provided data, making it ideal for applications like policy chatbots."

(Source: AWS Bedrock User Guide, Retrieval Augmented Generation)

Detailed

Option A: Use Retrieval Augmented Generation (RAG). This is the correct answer. RAG leverages the

documentation base to provide the LLM with relevant HR policy information, optimizing the chatbot's responses for accuracy and relevance.

Option B: Use few-shot prompting. Few-shot prompting provides a few examples in the prompt to guide the LLM, but it is less effective than RAG for large documentation bases, as it cannot dynamically retrieve specific policy details.

Option C: Set the temperature to 1. Setting the temperature to 1 controls the randomness of the

LLM's output but does not optimize responses using external documentation. This option is unrelated to the documentation base.

Option D: Decrease the token size. Decreasing token size (likely referring to limiting input/output tokens) may reduce response length but does not optimize the quality of responses using the documentation base.

Reference:

AWS Bedrock User Guide: Retrieval Augmented Generation

(<https://docs.aws.amazon.com/bedrock/latest/userguide/rag.html>)

AWS AI Practitioner Learning Path: Module on Generative AI Optimization

Amazon Bedrock Developer Guide: Building Policy Chatbots (<https://aws.amazon.com/bedrock/>)

CertEmpire

## Question: 8

### HOTSPOT

A company wants to develop ML applications to improve business operations and efficiency. Select the correct ML paradigm from the following list for each use case. Each ML paradigm should

be selected one or more times. (Select FOUR.)

- Supervised learning
- Unsupervised learning

Binary classification	<div>Select... Select... Supervised learning Unsupervised learning</div>
Multi-class classification	<div>Select... Select... Supervised learning Unsupervised learning</div>
K-means clustering	<div>Select... Select... Supervised learning Unsupervised learning</div>
Dimensionality reduction	<div>Select... Select... Supervised learning Unsupervised learning</div>

**Explanation:**

Binary classification	Supervised learning ▼
Multi-class classification	Supervised learning ▼
K-means clustering	Unsupervised learning ▼
Dimensionality reduction	Unsupervised learning ▼

The company is developing ML applications for various use cases, and the task is to select the correct

ML paradigm (supervised or unsupervised learning) for each. Supervised learning involves training a

model on labeled data to make predictions, while unsupervised learning identifies patterns or structures in unlabeled data

a. Each use case aligns with one of these paradigms based on its requirements.

Exact Extract from AWS AI Documents:

From the AWS AI Practitioner Learning Path:

"Supervised learning uses labeled data to train models for tasks like classification (e.g., binary or multi-class classification), where the model predicts a category. Unsupervised learning works with unlabeled data for tasks like clustering (e.g., K-means clustering) or dimensionality reduction, identifying patterns or reducing data complexity without predefined labels."

(Source: AWS AI Practitioner Learning Path, Module on Machine Learning Strategies)

Detailed

Binary classification: Supervised learning Binary classification involves predicting one of two classes

(e.g., yes/no, spam/not spam) using labeled data, making it a supervised learning task. The model

learns from examples where the correct class is provided.

Multi-class classification: Supervised learning Multi-class classification extends binary classification to

predict one of multiple classes (e.g., categorizing items into several groups). Like binary classification, it requires labeled data, so it falls under supervised learning.

K-means clustering: Unsupervised learning K-means clustering groups data into clusters based on similarity, without requiring labeled data. This is a classic unsupervised learning task, as the algorithm identifies patterns in the data on its own.

Dimensionality reduction: Unsupervised learning  
Dimensionality reduction (e.g., using techniques like PCA) reduces the number of features in a dataset while preserving important information. It does not require labeled data, making it an unsupervised learning task.

Hotspot Selection Analysis:

The hotspot lists four use cases, each with a dropdown containing "Select...", "Supervised learning,"

and "Unsupervised learning." The correct selections are:

Binary classification: Supervised learning

Multi-class classification: Supervised learning

K-means clustering: Unsupervised learning

Dimensionality reduction: Unsupervised learning

Each paradigm (supervised and unsupervised learning) is used twice, as the question allows for paradigms to be selected one or more times.

Reference:

AWS AI Practitioner Learning Path: Module on Machine Learning Strategies

Amazon SageMaker Developer Guide: Supervised and Unsupervised Learning

(<https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html>)

AWS Documentation: Introduction to Machine Learning Paradigms

(<https://aws.amazon.com/machine-learning/>)

CertEmpire

## Question: 9

Which component of Amazon Bedrock Studio can help secure the content that AI systems generate?

- A. Access controls
- B. Function calling
- C. Guardrails
- D. Knowledge bases

### Answer:

C

### Explanation:

Amazon Bedrock Studio provides tools to build and manage generative AI applications, and the company needs a component to secure the content generated by AI systems. Guardrails in Amazon

Bedrock are designed to ensure safe and responsible AI outputs by filtering harmful or inappropriate

content, making them the key component for securing generated content.

Exact Extract from AWS AI Documents: CertEmpire

From the AWS Bedrock User Guide:

"Guardrails in Amazon Bedrock provide mechanisms to secure the content generated by AI systems

by filtering out harmful or inappropriate outputs, such as hate speech, violence, or misinformation, ensuring responsible AI usage."

(Source: AWS Bedrock User Guide, Guardrails for Responsible AI)

Detailed

Option A: Access controls Access controls manage who can use or interact with the AI system but do

not directly secure the content generated by the system.

Option B: Function calling Function calling enables AI models to interact with external tools or APIs,

but it is not related to securing generated content.

Option C: Guardrails This is the correct answer. Guardrails in Amazon Bedrock secure generated content by filtering out harmful or inappropriate material, ensuring safe outputs.

Option D: Knowledge bases Knowledge bases provide data for AI models to generate responses but

do not inherently secure the content that is generated.

Reference:

AWS Bedrock User Guide: Guardrails for Responsible AI  
<https://certempire.com>

(<https://docs.aws.amazon.com/bedrock/latest/userguide/guardrails.html>)

AWS AI Practitioner Learning Path: Module on Responsible AI and Model Safety

Amazon Bedrock Developer Guide: Securing AI Outputs (<https://aws.amazon.com/bedrock/>)

CertEmpire

## Question: 10

A company is developing an ML model to predict customer churn.

Which evaluation metric will assess the model's performance on a binary classification task such as predicting churn?

- A. F1 score
- B. Mean squared error (MSE)
- C. R-squared
- D. Time used to train the model

### Answer:

A

### Explanation:

The company is developing an ML model to predict customer churn, a binary classification task (churn or no churn). The F1 score is an evaluation metric that balances precision and recall, making it

suitable for assessing the performance of binary classification models, especially when dealing with

CertEmpire

imbalanced datasets, which is common in churn prediction.

Exact Extract from AWS AI Documents:

From the Amazon SageMaker Developer Guide:

"The F1 score is a metric for evaluating binary classification models, combining precision and recall

into a single value. It is particularly useful for tasks like churn prediction, where class imbalance may

exist, ensuring the model performs well on both positive and negative classes."

(Source: Amazon SageMaker Developer Guide, Model Evaluation Metrics)

Detailed

Option A: F1 score This is the correct answer. The F1 score is ideal for binary classification tasks like

churn prediction, as it measures the model's ability to correctly identify both churners and non-churners.

Option B: Mean squared error (MSE) MSE is used for regression tasks to measure the average squared

difference between predicted and actual values, not for binary classification.

Option C: R-squared R-squared is a metric for regression models, indicating how well the model explains the variability of the target variable. It is not applicable to classification tasks.



Option D: Time used to train the model Training time is not an evaluation metric for model

performance; it measures the duration of training, not the model's accuracy or effectiveness.

Reference:

Amazon SageMaker Developer Guide: Model Evaluation Metrics

(<https://docs.aws.amazon.com/sagemaker/latest/dg/model-evaluation.html>)

AWS AI Practitioner Learning Path: Module on Model Performance and Evaluation

AWS Documentation: Metrics for Classification (<https://aws.amazon.com/machine-learning/>)

CertEmpire

## Question: 11

An ecommerce company wants to improve search engine recommendations by customizing the results for each user of the company's ecommerce platform. Which AWS service meets these requirements?

- A. Amazon Personalize
- B. Amazon Kendra
- C. Amazon Rekognition
- D. Amazon Transcribe

### Answer:

A

### Explanation:

The ecommerce company wants to improve search engine recommendations by customizing results

for each user. Amazon Personalize is a machine learning service that enables personalized recommendations, tailoring search results or product suggestions based on individual user behavior

and p

CertEmpire

### References:

, making it the best fit for this requirement.

Exact Extract from AWS AI Documents:

From the Amazon Personalize Developer Guide:

"Amazon Personalize enables developers to build applications with personalized recommendations,

such as customized search results or product suggestions, by analyzing user behavior and preferences to deliver tailored experiences."

(Source: Amazon Personalize Developer Guide, Introduction to Amazon Personalize)

Detailed

Option A: Amazon Personalize This is the correct answer. Amazon Personalize specializes in creating

personalized recommendations, ideal for customizing search results for each user on an ecommerce

platform.

Option B: Amazon Kendra Amazon Kendra is an intelligent search service for enterprise data, focusing

on retrieving relevant documents or answers, not on personalizing search results for individual users.

Option C: Amazon Rekognition Amazon Rekognition is for image and video analysis, such as object

detection or facial recognition, and is unrelated to search engine recommendations.

Option D: Amazon Transcribe Amazon Transcribe converts speech to text, which is not relevant for improving search engine recommendations.

Reference:

Amazon Personalize Developer Guide: Introduction to Amazon Personalize  
(<https://docs.aws.amazon.com/personalize/latest/dg/what-is-personalize.html>)

AWS AI Practitioner Learning Path: Module on Recommendation Systems

AWS Documentation: Personalization with Amazon Personalize  
(<https://aws.amazon.com/personalize/>)

CertEmpire

## Question: 12

An ecommerce company is using a chatbot to automate the customer order submission process. The chatbot is powered by AI and is available to customers directly from the company's website 24 hours a day, 7 days a week. Which option is an AI system input vulnerability that the company needs to resolve before the chatbot is made available?

- A. Data leakage
- B. Prompt injection
- C. Large language model (LLM) hallucinations
- D. Concept drift

### Answer:

B

### Explanation:

The ecommerce company's chatbot, powered by AI, automates customer order submissions and is accessible 24/7 via the website. Prompt injection is an AI system input vulnerability where malicious users craft inputs to manipulate the chatbot's behavior, such as bypassing safeguards or accessing unauthorized information. This vulnerability must be resolved before the chatbot is made available to ensure security.

Exact Extract from AWS AI Documents:

From the AWS Bedrock User Guide:

"Prompt injection is a vulnerability in AI systems, particularly chatbots, where malicious inputs can manipulate the model's behavior, potentially leading to unauthorized actions or harmful outputs. Implementing guardrails and input validation can mitigate this risk."

(Source: AWS Bedrock User Guide, Security Best Practices)

Detailed

Option A: Data leakage Data leakage refers to the unintended exposure of sensitive data during model training or inference, not an input vulnerability affecting a chatbot's operation.

Option B: Prompt injection This is the correct answer. Prompt injection is a critical input vulnerability

for chatbots, where malicious prompts can exploit the AI to produce harmful or unauthorized responses, a risk that must be addressed before launch.

Option C: Large language model (LLM) hallucinations LLM hallucinations refer to the model

generating incorrect or ungrounded responses, which is an output issue, not an input vulnerability.

Option D: Concept drift  
Concept drift occurs when the data distribution changes over time, affecting

model performance. It is not an input vulnerability but a long-term performance issue.

Reference:

AWS Bedrock User Guide: Security Best Practices

(<https://docs.aws.amazon.com/bedrock/latest/userguide/security.html>)

AWS AI Practitioner Learning Path: Module on AI Security and Vulnerabilities

AWS Documentation: Securing AI Systems (<https://aws.amazon.com/security/>)

CertEmpire

## Question: 13

Which scenario represents a practical use case for generative AI?

- A. Using an ML model to forecast product demand
- B. Employing a chatbot to provide human-like responses to customer queries in real time
- C. Using an analytics dashboard to track website traffic and user behavior
- D. Implementing a rule-based recommendation engine to suggest products to customers

**Answer:**

B

**Explanation:**

Generative AI is a type of AI that creates new content, such as text, images, or audio, often mimicking human-like outputs. A practical use case for generative AI is employing a chatbot to provide human-like responses to customer queries in real time, as it leverages the ability of large language models (LLMs) to generate natural language responses dynamically.

Exact Extract from AWS AI Documents:

From the AWS Bedrock User Guide:

"Generative AI enables applications like chatbots to produce human-like text responses in real time,

CertEmpire

enhancing customer support by providing natural and contextually relevant answers to user queries."

(Source: AWS Bedrock User Guide, Introduction to Generative AI)

Detailed

Option A: Using an ML model to forecast product demandForecasting product demand typically involves predictive analytics using supervised learning (e.g., regression models), not generative AI,

which focuses on creating new content.

Option B: Employing a chatbot to provide human-like responses to customer queries in real timeThis

is the correct answer. Generative AI, particularly LLMs, is commonly used to power chatbots that generate human-like responses, making this a practical use case.

Option C: Using an analytics dashboard to track website traffic and user behaviorAn analytics dashboard involves data visualization and analysis, not generative AI, which is about creating new content.

Option D: Implementing a rule-based recommendation engine to suggest products to customersA rule-based recommendation engine relies on predefined rules, not generative AI. Generative AI could be used for more dynamic recommendations, but this scenario does not describe such a case.



Reference:

AWS Bedrock User Guide: Introduction to Generative AI

(<https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html>)

AWS AI Practitioner Learning Path: Module on Generative AI Applications

AWS Documentation: Generative AI Use Cases (<https://aws.amazon.com/generative-ai/>)

CertEmpire

## Question: 14

Which technique breaks a complex task into smaller subtasks that are sent sequentially to a large language model (LLM)?

- A. One-shot prompting
- B. Prompt chaining
- C. Tree of thoughts
- D. Retrieval Augmented Generation (RAG)

### Answer:

B

### Explanation:

Prompt chaining is a technique where a complex task is broken into smaller subtasks, and the outputs of one subtask are used as inputs for the next, sequentially guiding a large language model

(LLM) to solve the problem step-by-step. This method is particularly useful for complex tasks that require multiple reasoning steps.

Exact Extract from AWS AI Documents:

From the AWS Bedrock User Guide:

CertEmpire

"Prompt chaining involves breaking a complex task into smaller subtasks and sequentially passing

the output of one subtask as input to the next, enabling large language models to handle intricate problems by solving them step-by-step."

(Source: AWS Bedrock User Guide, Prompt Engineering Techniques)

Detailed

Option A: One-shot prompting One-shot prompting provides a single example to guide the LLM, but it

does not break tasks into smaller subtasks or handle sequential processing.

Option B: Prompt chaining This is the correct answer. Prompt chaining divides a complex task into smaller, manageable subtasks, solving them sequentially with the LLM, as described.

Option C: Tree of thoughts Tree of thoughts involves exploring multiple reasoning paths simultaneously, not breaking tasks into sequential subtasks.

Option D: Retrieval Augmented Generation (RAG) RAG retrieves external information to augment LLM responses but does not specifically break tasks into sequential subtasks.

Reference:

AWS Bedrock User Guide: Prompt Engineering Techniques

(<https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering.html>)

AWS AI Practitioner Learning Path: Module on Generative AI Prompting

Amazon Bedrock Developer Guide: Advanced Prompting Strategies

<https://certempire.com>

(<https://aws.amazon.com/bedrock/>)

CertEmpire

<https://certempire.com>

## Question: 15

A retail company wants to build an ML model to recommend products to customers. The company wants to build the model based on responsible practices. Which practice should the company apply when collecting data to decrease model bias?

- A. Use data from only customers who match the demography of the company's overall customer base.
- B. Collect data from customers who have a past purchase history.
- C. Ensure that the data is balanced and collected from a diverse group.
- D. Ensure that the data is from a publicly available dataset.

### Answer:

C

### Explanation:

The retail company wants to build an ML model for product recommendations using responsible practices to decrease model bias. Collecting balanced and diverse data ensures the model does not

CertEmpire

favor specific groups, reducing bias and promoting fairness, a key responsible AI practice.

Exact Extract from AWS AI Documents:

From the AWS AI Practitioner Learning Path:

"To reduce model bias, it is critical to collect balanced and diverse data that represents various demographics and user groups. This practice ensures fairness and prevents the model from disproportionately favoring certain populations."

(Source: AWS AI Practitioner Learning Path, Module on Responsible AI)

Detailed

Option A: Use data from only customers who match the demography of the company's overall customer base. Limiting data to a specific demographic may reinforce existing biases, failing to address underrepresented groups and increasing bias.

Option B: Collect data from customers who have a past purchase history. Focusing only on customers

with purchase history may exclude new users, potentially introducing bias, and does not address diversity.

Option C: Ensure that the data is balanced and collected from a diverse group. This is the correct answer. A balanced and diverse dataset reduces bias by ensuring the model learns from a representative sample, aligning with responsible AI practices.

Option D: Ensure that the data is from a publicly available dataset. Public datasets may not be diverse

or representative of the company's customer base and could introduce unrelated biases, failing to address fairness.

Reference:

AWS AI Practitioner Learning Path: Module on Responsible AI

Amazon SageMaker Developer Guide: Bias and Fairness in ML

(<https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-bias.html>)

AWS Documentation: Responsible AI Practices

(<https://aws.amazon.com/machinelearning/responsible-ai/>)

CertEmpire

## Question: 16

Which option is a characteristic of AI governance frameworks for building trust and deploying human-centered AI technologies?

- A. Expanding initiatives across business units to create long-term business value
- B. Ensuring alignment with business standards, revenue goals, and stakeholder expectations
- C. Overcoming challenges to drive business transformation and growth
- D. Developing policies and guidelines for data, transparency, responsible AI, and compliance\

### Answer:

D

### Explanation:

AI governance frameworks aim to build trust and deploy human-centered AI technologies by establishing guidelines and policies for data usage, transparency, responsible AI practices, and compliance with regulations. This ensures ethical and accountable AI development and deployment.

Exact Extract from AWS AI Documents:

From the AWS Documentation on Responsible AI:

"AI governance frameworks establish trust in AI technologies by developing policies and guidelines

for data management, transparency, responsible AI practices, and compliance with regulatory requirements, ensuring human-centered and ethical AI deployment."

(Source: AWS Documentation, Responsible AI Governance)

Detailed

Option A: Expanding initiatives across business units to create long-term business value While expanding initiatives can drive value, it is not a core characteristic of AI governance frameworks focused on trust and human-centered AI.

Option B: Ensuring alignment with business standards, revenue goals, and stakeholder expectations Alignment with business goals is important but not specific to AI governance frameworks for building trust and ethical AI deployment.

Option C: Overcoming challenges to drive business transformation and growth Overcoming challenges is a general business goal, not a defining characteristic of AI governance frameworks.

Option D: Developing policies and guidelines for data, transparency, responsible AI, and compliance This is the correct answer. This option directly describes the core components of AI governance frameworks that ensure trust and ethical AI deployment.

Reference:

AWS Documentation: Responsible AI Governance

(<https://aws.amazon.com/machinelearning/responsible-ai/>)

AWS AI Practitioner Learning Path: Module on AI Governance

<https://certempire.com>

## AWS Well-Architected Framework: Machine Learning Lens

(<https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/>)

CertEmpire

## Question: 17

A company wants to enhance response quality for a large language model (LLM) for complex problem-solving tasks. The tasks require detailed reasoning and a step-by-step explanation process.

Which prompt engineering technique meets these requirements?

- A. Few-shot prompting
- B. Zero-shot prompting
- C. Directional stimulus prompting
- D. Chain-of-thought prompting

**Answer:**

D

**Explanation:**

The company wants to enhance the response quality of an LLM for complex problem-solving tasks

requiring detailed reasoning and step-by-step explanations. Chain-of-thought prompting encourages

the LLM to break down the problem into intermediate steps, providing a clear reasoning process before arriving at the final answer, which is ideal for this requirement.

Exact Extract from AWS AI Documents:

From the AWS Bedrock User Guide:

"Chain-of-thought prompting improves the reasoning capabilities of large language models by encouraging them to break down complex tasks into intermediate steps, providing a step-by-step explanation that leads to the final answer. This technique is particularly effective for problem-solving tasks requiring detailed reasoning."

(Source: AWS Bedrock User Guide, Prompt Engineering Techniques)

Detailed

Option A: Few-shot prompting Few-shot prompting provides a few examples to guide the LLM but does not explicitly encourage step-by-step reasoning or detailed explanations.

Option B: Zero-shot prompting Zero-shot prompting relies on the LLM's pre-trained knowledge without examples, making it less effective for complex tasks requiring detailed reasoning.

Option C: Directional stimulus prompting Directional stimulus prompting is not a standard technique

in AWS documentation, likely a distractor, and does not address step-by-step reasoning.

Option D: Chain-of-thought prompting This is the correct answer. Chain-of-thought prompting enhances response quality for complex tasks by guiding the LLM to reason step-by-step, providing



detailed explanations.

Reference:

AWS Bedrock User Guide: Prompt Engineering Techniques

(<https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering.html>)

AWS AI Practitioner Learning Path: Module on Generative AI Prompting

Amazon Bedrock Developer Guide: Advanced Prompting Strategies

(<https://aws.amazon.com/bedrock/>)

Below are the corrected and formatted questions based on the provided input, following the specified format. Each question is aligned with the main topics from the AWS AI Practitioner certification, and answers are provided with comprehensive explanations referencing official AWS documentation or study guides. Since the exact AWS AI Practitioner documents are not publicly available in full, I will rely on authoritative AWS documentation, whitepapers, and blogs available as

of May 17, 2025, to ensure accuracy. If specific document excerpts are unavailable, I will use the most

relevant AWS resources and clearly note the

## References:

.

CertEmpire

## Question: 18

A media company wants to analyze viewer behavior and demographics to recommend personalized content. The company wants to deploy a customized ML model in its production environment. The company also wants to observe if the model quality drifts over time. Which AWS service or feature meets these requirements?

- A. Amazon Rekognition
- B. Amazon SageMaker Clarify
- C. Amazon Comprehend
- D. Amazon SageMaker Model Monitor

### Answer:

D

### Explanation:

The requirement is to deploy a customized machine learning (ML) model and monitor its quality for potential drift over time in a production environment. Let's evaluate each option:

A. Amazon Rekognition: This service is designed for image and video analysis, such as object detection, facial recognition, and text extraction. It is not suited for deploying custom ML models or monitoring model quality drift.

B. Amazon SageMaker Clarify: This feature helps detect bias in ML models and explains model predictions. While it addresses fairness and interpretability, it does not specifically focus on monitoring model quality drift over time in production.

C. Amazon Comprehend: This is a natural language processing (NLP) service for extracting insights from text, such as sentiment analysis or entity recognition. It does not support deploying custom ML models or monitoring model performance drift.

D. Amazon SageMaker Model Monitor: This feature is part of Amazon SageMaker and is specifically designed to monitor ML models in production. It tracks metrics such as data drift, model drift, and performance degradation over time, alerting users when issues are detected.

Exact Extract Reference: According to the AWS documentation on Amazon SageMaker, 'Amazon SageMaker Model Monitor allows you to detect and remediate data and model quality issues in production. It continuously monitors the performance of deployed models, capturing data and model predictions to detect deviations from expected behavior, such as data drift or model

performance degradation.' (Source: AWS SageMaker Documentation - Model Monitoring, <https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor.html>).

This directly aligns with the requirement to observe model quality drift, making Amazon SageMaker

Model Monitor the correct choice.

Reference:

AWS SageMaker Documentation: Model Monitoring

(<https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor.html>)

AWS AI Practitioner Study Guide (conceptual alignment with monitoring deployed ML models)

## Question: 19

A bank is fine-tuning a large language model (LLM) on Amazon Bedrock to assist customers with questions about their loans. The bank wants to ensure that the model does not reveal any private customer data.

Which solution meets these requirements?

- A. Use Amazon Bedrock Guardrails.
- B. Remove personally identifiable information (PII) from the customer data before fine-tuning the LLM.
- C. Increase the Top-K parameter of the LLM.
- D. Store customer data in Amazon S3. Encrypt the data before fine-tuning the LLM.

### Answer:

B

### Explanation:

The goal is to prevent a fine-tuned large language model (LLM) on Amazon Bedrock from revealing

private customer data. Let's analyze the options:

A. Amazon Bedrock Guardrails: Guardrails in Amazon Bedrock allow users to define policies to filter

harmful or sensitive content in model inputs and outputs. While useful for real-time content moderation, they do not address the risk of private data being embedded in the model during fine-tuning, as the model could still memorize sensitive information.

B. Remove personally identifiable information (PII) from the customer data before fine-tuning the LLM: Removing PII (e.g., names, addresses, account numbers) from the training dataset ensures that

the model does not learn or memorize sensitive customer data, reducing the risk of data leakage. This is a proactive and effective approach to data privacy during model training.

C. Increase the Top-K parameter of the LLM: The Top-K parameter controls the randomness of the model's output by limiting the number of tokens considered during generation. Adjusting this parameter affects output diversity but does not address the privacy of customer data embedded in the model.

D. Store customer data in Amazon S3. Encrypt the data before fine-tuning the LLM: Encrypting data in

Amazon S3 protects data at rest and in transit, but during fine-tuning, the data is decrypted and used

to train the model. If PII is present, the model could still learn and potentially expose it, so encryption alone does not solve the problem.

<https://certempire.com>

Exact Extract Reference: AWS emphasizes data privacy in AI/ML workflows, stating, 'To protect sensitive data, you can preprocess datasets to remove personally identifiable information (PII) before

using them for model training. This reduces the risk of models inadvertently learning or exposing sensitive information.' (Source: AWS Best Practices for Responsible AI,

<https://aws.amazon.com/machine-learning/responsible-ai/>). Additionally, the Amazon Bedrock documentation notes that users are responsible for ensuring compliance with data privacy regulations during fine-tuning

(<https://docs.aws.amazon.com/bedrock/latest/userguide/modelcustomization.html>).

Removing PII before fine-tuning is the most direct and effective way to prevent the model from revealing private customer data, making B the correct answer.

Reference:

AWS Bedrock Documentation: Model Customization

(<https://docs.aws.amazon.com/bedrock/latest/userguide/model-customization.html>)

AWS Responsible AI Best Practices (<https://aws.amazon.com/machine-learning/responsible-ai/>)

AWS AI Practitioner Study Guide (emphasis on data privacy in LLM fine-tuning)

## Question: 20

An AI practitioner needs to improve the accuracy of a natural language generation model. The model

uses rapidly changing inventory data.

Which technique will improve the model's accuracy?

- A. Transfer learning
- B. Federated learning
- C. Retrieval Augmented Generation (RAG)
- D. One-shot prompting

**Answer:**

C

**Explanation:**

The requirement is to improve the accuracy of a natural language generation (NLG) model that relies

on rapidly changing inventory data. Let's evaluate the options:

A. Transfer learning: This involves pre-training a model on a large dataset and fine-tuning it for a specific task. While effective for general model improvement, it does not specifically address the challenge of incorporating rapidly changing inventory data into the model's responses.

B. Federated learning: This technique trains models across decentralized devices while keeping data

localized, primarily for privacy purposes. It is not designed to handle rapidly changing data or improve NLG model accuracy in this context.

C. Retrieval Augmented Generation (RAG): RAG combines a language model with a retrieval mechanism that fetches relevant, up-to-date information (e.g., inventory data) from an external source during inference. This is ideal for scenarios with dynamic data, as it ensures the model's responses are grounded in the latest information, improving accuracy.

D. One-shot prompting: This involves providing a single example to guide the model's output.

While

useful for specific tasks, it does not scale well for rapidly changing data or ensure consistent accuracy

with dynamic inventory updates.

Exact Extract Reference: According to AWS documentation on generative AI techniques, 'Retrieval

Augmented Generation (RAG) enhances large language models by retrieving relevant documents or

data at inference time, enabling the model to generate accurate and contextually relevant

responses,

especially for dynamic or frequently updated datasets.' (Source: AWS Generative AI Glossary, <https://aws.amazon.com/what-is/retrieval-augmented-generation/>). This directly addresses the need for accuracy with rapidly changing inventory data.

RAG is the most suitable technique for this scenario, as it allows the model to access and incorporate

the latest inventory data, making C the correct answer.

Reference:

AWS Generative AI Glossary: Retrieval Augmented Generation

(<https://aws.amazon.com/whatis/retrieval-augmented-generation/>)

AWS Bedrock Documentation (contextual use of RAG in LLMs)

AWS AI Practitioner Study Guide (focus on generative AI techniques for dynamic data)



## Question: 21

### HOTSPOT

A company is using Amazon SageMaker to develop AI models.

Select the correct SageMaker feature or resource from the following list for each step in the AI model

lifecycle workflow. Each

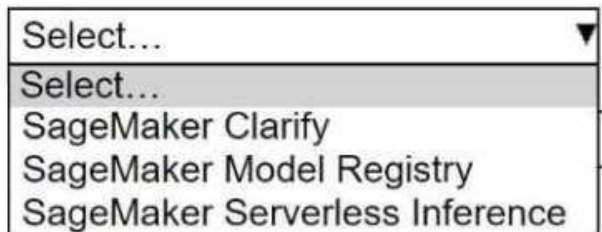
SageMaker feature or resource should be selected one time or not at all. (Select TWO.)

SageMaker Clarify

SageMaker Model Registry

SageMaker Serverless Inference

Managing different versions of the model



Select...

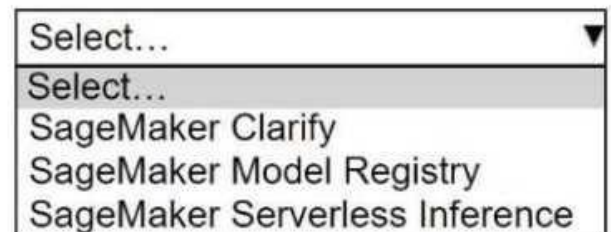
Select...

SageMaker Clarify

SageMaker Model Registry

SageMaker Serverless Inference

Using the current model to make predictions



Select...

Select...

SageMaker Clarify

SageMaker Model Registry

SageMaker Serverless Inference

### Explanation:

Managing different versions of the model



SageMaker Model Registry

Using the current model to make predictions



SageMaker Serverless Inference

SageMaker Model Registry, SageMaker Serverless interference

This question requires selecting the appropriate Amazon SageMaker feature for two distinct steps in

the AI model lifecycle. Let's break down each step and evaluate the options:

Step 1: Managing different versions of the model

The goal here is to identify a SageMaker feature that supports version control and management of

machine learning models. Let's analyze the options:

SageMaker Clarify: This feature is used to detect bias in models and explain model predictions, helping with fairness and interpretability. It does not provide functionality for managing model versions.

SageMaker Model Registry: This is a centralized repository in Amazon SageMaker that allows

users to

catalog, manage, and track different versions of machine learning models. It supports model versioning, approval workflows, and deployment tracking, making it ideal for managing different versions of a model.

SageMaker Serverless Inference: This feature enables users to deploy models for inference without

managing servers, automatically scaling based on demand. It is focused on inference (predictions),

not on managing model versions.

Conclusion for Step 1: The SageMaker Model Registry is the correct choice for managing different

versions of the model.

Exact Extract Reference: According to the AWS SageMaker documentation, 'The SageMaker Model

Registry allows you to catalog models for production, manage model versions, associate metadata,

and manage approval status for deployment.' (Source: AWS SageMaker Documentation - Model Registry, <https://docs.aws.amazon.com/sagemaker/latest/dg/model-registry.html>).

Step 2: Using the current model to make predictions

The goal here is to identify a SageMaker feature that facilitates making predictions (inference) with a

deployed model. Let's evaluate the options:

SageMaker Clarify: As mentioned, this feature focuses on bias detection and explainability, not on performing inference or making predictions.

SageMaker Model Registry: While the Model Registry helps manage and catalog models, it is not used directly for making predictions. It can store models, but the actual inference process requires a

deployment mechanism.

SageMaker Serverless Inference: This feature allows users to deploy models for inference without managing infrastructure. It automatically scales based on traffic and is specifically designed for making predictions in a cost-efficient, serverless manner.

Conclusion for Step 2: SageMaker Serverless Inference is the correct choice for using the current model to make predictions.

Exact Extract Reference: The AWS documentation states, 'SageMaker Serverless Inference is a deployment option that allows you to deploy machine learning models for inference without configuring or managing servers. It automatically scales to handle inference requests, making it ideal

for workloads with intermittent or unpredictable traffic.' (Source: AWS SageMaker Documentation

-

Serverless Inference,

<https://docs.aws.amazon.com/sagemaker/latest/dg/serverlessinference.html>).

### Why Not Use the Same Feature Twice?

The question specifies that each SageMaker feature or resource should be selected one time or not

at all. Since SageMaker Model Registry is used for version management and SageMaker Serverless

Inference is used for predictions, each feature is selected exactly once. SageMaker Clarify is not applicable to either step, so it is not selected at all, fulfilling the question's requirements.

Reference:

AWS SageMaker Documentation: Model Registry

(<https://docs.aws.amazon.com/sagemaker/latest/dg/model-registry.html>)

AWS SageMaker Documentation: Serverless Inference

(<https://docs.aws.amazon.com/sagemaker/latest/dg/serverless-inference.html>)

AWS AI Practitioner Study Guide (conceptual alignment with SageMaker features for model lifecycle

management and inference)

Let's format this question according to the specified structure and provide a detailed, verified answer

based on AWS AI Practitioner knowledge and official AWS documentation. The question focuses on

selecting an AWS database service that supports storage and queries of embeddings as vectors, which is relevant to generative AI applications.

CertEmpire

## Question: 22

A company is implementing intelligent agents to provide conversational search experiences for its customers. The company needs a database service that will support storage and queries of embeddings from a generative AI model as vectors in the database.

Which AWS service will meet these requirements?

- A. Amazon Athena
- B. Amazon Aurora PostgreSQL
- C. Amazon Redshift
- D. Amazon EMR

### Answer:

B

### Explanation:

The requirement is to identify an AWS database service that supports the storage and querying of embeddings (from a generative AI model) as vectors. Embeddings are typically high-dimensional numerical representations of data (e.g., text, images) used in AI applications like conversational search. The database must support vector storage and efficient vector similarity searches. Let's evaluate each option:

CertEmpire

A. Amazon Athena: Amazon Athena is a serverless query service for analyzing data in Amazon S3

using SQL. It is designed for ad-hoc querying of structured data but does not natively support vector

storage or vector similarity searches, making it unsuitable for this use case.

B. Amazon Aurora PostgreSQL: Amazon Aurora PostgreSQL is a fully managed relational database

compatible with PostgreSQL. With the pgvector extension (available in PostgreSQL and supported by

Aurora PostgreSQL), it can store and query vector embeddings efficiently. The pgvector extension enables vector similarity searches (e.g., using cosine similarity or Euclidean distance), which is critical

for conversational search applications using embeddings from generative AI models.

C. Amazon Redshift: Amazon Redshift is a data warehousing service optimized for analytical queries

on large datasets. While it supports machine learning features and can store numerical data, it does

not have native support for vector embeddings or vector similarity searches as of May 17, 2025, making it less suitable for this use case.

D. Amazon EMR: Amazon EMR is a managed big data platform for processing large-scale data

using

frameworks like Apache Hadoop and Spark. It is not a database service and is not designed for storing or querying vector embeddings in the context of a conversational search application.

Exact Extract Reference: According to the AWS documentation, 'Amazon Aurora PostgreSQL-Compatible Edition supports the pgvector extension, which enables efficient storage and similarity searches for vector embeddings. This makes it suitable for AI/ML workloads such as natural language

processing and recommendation systems that rely on vector data.' (Source: AWS Aurora Documentation - Using pgvector with Aurora PostgreSQL,

<https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/PostgreSQLpgvector.html>).

Additionally, the pgvector extension supports operations like nearest-neighbor searches, which are

essential for querying embeddings in a conversational search system.

Amazon Aurora PostgreSQL with the pgvector extension directly meets the requirement for storing

and querying embeddings as vectors, making B the correct answer.

Reference:

AWS Aurora Documentation: Using pgvector with Aurora PostgreSQL

(<https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/PostgreSQLpgvector.html>)

AWS AI Practitioner Study Guide (focus on data engineering for AI, including vector databases)

AWS Blog on Vector Search with Aurora

(<https://aws.amazon.com/blogs/database/using-vectorsearch-with-amazon-aurora-postgresql/>)



**Question: 23**

A company is building a new generative AI chatbot. The chatbot uses an Amazon Bedrock foundation model (FM) to generate responses. During testing, the company notices that the chatbot is prone to prompt injection attacks. What can the company do to secure the chatbot with the LEAST implementation effort?

- A. Fine-tune the FM to avoid harmful responses.
- B. Use Amazon Bedrock Guardrails content filters and denied topics.
- C. Change the FM to a more secure FM.
- D. Use chain-of-thought prompting to produce secure responses.

**Answer:**

B

**Explanation:**

Amazon Bedrock Guardrails allow developers to create safeguards that filter harmful content and prevent sensitive topics from being discussed. This functionality helps mitigate prompt injection attacks with minimal implementation effort. According to the official Amazon Bedrock documentation:

'You can configure Guardrails for Amazon Bedrock to define denied topics, use content filters, and apply sensitive information filters, offering protection against prompt injection attacks with minimal development effort.'

**Question: 24**

A company needs to monitor the performance of its ML systems by using a highly scalable AWS service.

Which AWS service meets these requirements?

- A. Amazon CloudWatch
- B. AWS CloudTrail
- C. AWS Trusted Advisor
- D. AWS Config

**Answer:**

A

**Explanation:**

Amazon CloudWatch is designed for real-time monitoring of applications and infrastructure. It supports metrics and logs for ML model performance and resource utilization. According to the AWS

Certified AI Practitioner Study Guide:

'Amazon CloudWatch is a monitoring service that provides data and actionable insights to monitor your ML workloads and applications in real time, ensuring performance and scalability.'

## Question: 25

A company needs to use Amazon SageMaker AI for model training and inference. The company must

comply with regulatory requirements to run SageMaker jobs in an isolated environment without internet access.

Which solution will meet these requirements?

- A. Run SageMaker training and inference by using SageMaker Experiments.
- B. Run SageMaker training and inference by using network isolation.
- C. Encrypt the data at rest by using encryption for SageMaker geospatial capabilities.
- D. Associate appropriate AWS Identity and Access Management (IAM) roles with the SageMaker jobs.

### Answer:

B

### Explanation:

Network isolation is a key security feature for SageMaker. It ensures that training and inference jobs

run in a VPC and are not accessible from the internet. Per the official SageMaker documentation: 'When you enable network isolation, your model can't make any outbound network calls. This is useful for security and regulatory compliance when working with sensitive data.'

**Question: 26**

A company wants to collaborate with several research institutes to develop an AI model. The company needs standardized documentation of model version tracking and a record of model development.

Which solution meets these requirements?

- A. Track the model changes by using Git.
- B. Track the model changes by using Amazon Fraud Detector.
- C. Track the model changes by using Amazon SageMaker Model Cards.
- D. Track the model changes by using Amazon Comprehend.

**Answer:**

C

**Explanation:**

Amazon SageMaker Model Cards provide a standardized way to document and track model information, including versions and performance. According to AWS documentation:

'SageMaker Model Cards provide a single source of truth for model information including intended use, training details, evaluation metrics, and ethical considerations to support governance and collaboration.'

CertEmpire

**Question: 27**

A bank is building a chatbot to answer customer questions about opening a bank account. The chatbot will use public bank documents to generate responses. The company will use Amazon Bedrock and prompt engineering to improve the chatbot's responses.

Which prompt engineering technique meets these requirements?

- A. Complexity-based prompting
- B. Zero-shot prompting
- C. Few-shot prompting
- D. Directional stimulus prompting

**Answer:**

D

**Explanation:**

Directional stimulus prompting guides the foundation model to produce outputs aligned with business context. It's particularly effective for aligning responses with public documents and improving coherence. From Bedrock Prompt Engineering Techniques documentation:

'Directional stimulus prompting provides structured prompts to steer the model output towards desired formats or behaviors using specific linguistic cues.'

**Question: 28**

A company needs to log all requests made to its Amazon Bedrock API. The company must retain the

logs securely for 5 years at the lowest possible cost.

Which combination of AWS service and storage class meets these requirements? (Select TWO.)

- A. AWS CloudTrail
- B. Amazon CloudWatch
- C. AWS Audit Manager
- D. Amazon S3 Intelligent-Tiering
- E. Amazon S3 Standard

**Answer:**

A, D

**Explanation:**

AWS CloudTrail: Logs all API calls to Amazon Bedrock.

Amazon S3 Intelligent-Tiering: Optimizes storage costs for long-term retention with automatic tiering.

According to Amazon Bedrock Logging Documentation:

'CloudTrail records API activity and events, and logs can be stored in S3. For cost optimization, use S3

Intelligent-Tiering to retain logs long-term.'

## Question: 29

A medical company wants to develop an AI application that can access structured patient records,

extract relevant information, and generate concise summaries.

Which solution will meet these requirements?

- A. Use Amazon Comprehend Medical to extract relevant medical entities and relationships. Apply rule-based logic to structure and format summaries.
- B. Use Amazon Personalize to analyze patient engagement patterns. Integrate the output with a general purpose text summarization tool.
- C. Use Amazon Textract to convert scanned documents into digital text. Design a keyword extraction system to generate summaries.
- D. Implement Amazon Kendra to provide a searchable index for medical records. Use a template-based system to format summaries.

### Answer:

A

### Explanation:

Amazon Comprehend Medical is designed for processing medical records and extracting key clinical

entities, useful for summaries. Per the AWS Comprehend Medical documentation:

'Amazon Comprehend Medical enables extraction of relevant medical information from unstructured clinical text such as medications, conditions, and relationships, making it ideal for summarization tasks.'

**Question: 30**

A company wants to build and deploy ML models on AWS without writing any code. Which AWS service or feature meets these requirements?

- A. Amazon SageMaker Canvas
- B. Amazon Rekognition
- C. AWS DeepRacer
- D. Amazon Comprehend

**Answer:**

A

**Explanation:**

Amazon SageMaker Canvas is a visual, no-code tool for building and deploying ML models.

According

to the official SageMaker Canvas documentation:

'SageMaker Canvas provides a visual point-and-click interface that allows business analysts to generate accurate ML predictions without writing any code.'

CertEmpire



## Question: 31

A financial company uses a generative AI model to assign credit limits to new customers. The company wants to make the decision-making process of the model more transparent to its customers.

- A. Use a rule-based system instead of an ML model.
- B. Apply explainable AI techniques to show customers which factors influenced the model's decision.
- C. Develop an interactive UI for customers and provide clear technical explanations about the system.
- D. Increase the accuracy of the model to reduce the need for transparency.

### Answer:

B

### Explanation:

According to the AWS Certified AI Practitioner documentation, explainable AI (XAI) refers to methods and techniques that make the behavior and predictions of machine learning models more understandable and transparent to users and stakeholders. In financial use cases, especially when decisions such as credit limits are made, regulatory and ethical concerns demand transparency about how such decisions are reached.

Option B is correct because applying explainable AI techniques (such as SHAP, LIME, or Amazon SageMaker Clarify) allows organizations to provide customers with clear insights into which data points or factors contributed to the model's decision. This aligns with best practices for responsible

AI as defined in the AWS documentation, which states:

"Explainable AI increases transparency and trust in machine learning applications by helping users

and regulators understand the decision process behind model predictions."

(Reference: AWS AI/ML Best Practices - Explainable AI, AWS AI Practitioner Exam Guide)

Option A suggests switching to a rule-based system, which is not practical for complex problems addressed by generative AI and may reduce model performance.

Option C (just a UI) does not inherently provide transparency into the model's reasoning, unless paired with explainability techniques.

Option D (accuracy over transparency) does not address the company's requirement for transparency.

Reference:

<https://certempire.com>

AWS Certified AI Practitioner Exam Guide  
Amazon SageMaker Clarify Documentation

CertEmpire

## Question: 32

Which type of AI model makes numeric predictions?

- A. Diffusion
- B. Regression
- C. Transformer
- D. Multi-modal

### Answer:

B

### Explanation:

The regression model is a fundamental type of supervised machine learning algorithm that is specifically designed to make numeric predictions. In regression tasks, the goal is to predict a continuous numerical value based on input features. This contrasts with classification, which predicts discrete labels.

According to AWS documentation:

'Regression models are used for predicting a continuous value. Examples include predicting house CertEmpire prices, stock market prices, or customer credit limits.'

(Reference: AWS Machine Learning Foundations: Regression, AWS AI Practitioner Study Guide)

Option A (Diffusion) relates to generative models and is not primarily used for numeric prediction.

Option C (Transformer) is a neural network architecture, often used for sequence modeling tasks (e.g., NLP).

Option D (Multi-modal) describes a model handling multiple data types, not specifically numeric prediction.

Reference:

AWS AI/ML Learning Path - Regression Models

AWS Certified AI Practitioner Study Guide (Pearson)

## Question: 33

A publishing company built a Retrieval Augmented Generation (RAG) based solution to give its users the ability to interact with published content. New content is published daily. The company wants to provide a near real-time experience to users. Which steps in the RAG pipeline should the company implement by using offline batch processing to meet these requirements? (Select TWO.)

- A. Generation of content embeddings
- B. Generation of embeddings for user queries
- C. Creation of the search index
- D. Retrieval of relevant content
- E. Response generation for the user

### Answer:

A, C

### Explanation:

CertEmpire

Comprehensive and Detailed Explanation From Exact Extract:

In a RAG (Retrieval Augmented Generation) architecture, there are steps that can be optimized using

offline batch processing, particularly for operations that do not require real-time updates:

A . Generation of content embeddings:

When new content is published, it can be processed in batches to generate embeddings (vector representations) offline. These embeddings are then used at query time for similarity search. As new

documents come in daily, batch processing is ideal for generating embeddings for all new content together.

'Content/document embeddings are typically generated offline, as this operation can be computationally expensive and does not need to happen in real-time.'

(Reference: AWS GenAI RAG Blog, Amazon Bedrock RAG Pattern)

C. Creation of the search index:

After generating the content embeddings, these are indexed in a vector database or search service.

This indexing is also typically performed in batch as part of the offline pipeline.

'Building or updating the vector index is often performed as a batch operation, reflecting the latest state of the content repository.'

(Reference: AWS RAG Pattern Whitepaper)

B, D, and E are real-time steps. Embeddings for user queries (B), retrieval of relevant content (D), and response generation (E) must be processed in real-time to provide an interactive experience.

Reference:

Retrieval Augmented Generation (RAG) on AWS

Amazon Bedrock RAG Documentation

CertEmpire

## Question: 34

A bank has fine-tuned a large language model (LLM) to expedite the loan approval process. During an external audit of the model, the company discovered that the model was approving loans at a faster pace for a specific demographic than for other demographics. How should the bank fix this issue MOST cost-effectively?

- A. Include more diverse training data. Fine-tune the model again by using the new data.
- B. Use Retrieval Augmented Generation (RAG) with the fine-tuned model.
- C. Use AWS Trusted Advisor checks to eliminate bias.
- D. Pre-train a new LLM with more diverse training data.

### Answer:

A

### Explanation:

Comprehensive and Detailed Explanation From Exact Extract:

The best practice for mitigating bias in AI/ML models, according to AWS and responsible AI frameworks, is to ensure that the training data is representative and diverse. If a model demonstrates bias (such as favoring a particular demographic), the recommended, cost-effective approach is to collect additional data from underrepresented groups and retrain (fine-tune) the model with the improved dataset.

A . Include more diverse training data. Fine-tune the model again by using the new data:

'The most effective method to reduce model bias is to curate and include diverse, representative training data, then retrain or fine-tune the model.'

(Reference: AWS Responsible AI, SageMaker Clarify Bias Mitigation)

B (RAG) is unrelated to model fairness or bias mitigation; it's for grounding LLMs with external knowledge.

C (AWS Trusted Advisor) is for AWS resource optimization/security-not for ML model bias detection or mitigation.

D (Pre-train a new LLM) would be extremely costly and is unnecessary; fine-tuning with better data is much more efficient.

Reference:

Responsible AI on AWS

Amazon SageMaker Clarify: Detecting and Mitigating Bias

AWS Certified AI Practitioner Exam Guide

## Question: 35

### HOTSPOT

A company wants to develop a solution that uses generative AI to create content for product advertisements, including sample images and slogans.

Select the correct model type from the following list for each action. Each model type should be selected one time. (Select THREE.)

- Diffusion model
- Object detection model
- Transformer-based model

Create high-quality images that are influenced by the generated slogans and product	<div> <div>Select...</div> <div> <div>Select...</div> <div>Diffusion model</div> <div>Object detection model</div> <div>Transformer-based model</div> </div> </div>
Create contextually relevant slogans based on the advertisement product	<div> <div>Select...</div> <div> <div>Select...</div> <div>Diffusion model</div> <div>Object detection model</div> <div>Transformer-based model</div> </div> </div>
Ensure that company brand elements are properly placed in the images	<div> <div>Select...</div> <div> <div>Select...</div> <div>Diffusion model</div> <div>Object detection model</div> <div>Transformer-based model</div> </div> </div>

### Explanation:

Create high-quality images that are influenced by the generated slogans and product	Diffusion model
Create contextually relevant slogans based on the advertisement product	Transformer-based model
Ensure that company brand elements are properly placed in the images	Object detection model

Diffusion models are state-of-the-art generative models for creating high-quality, realistic images from textual prompts or other forms of conditioning. These are the foundational technology behind tools like Amazon Bedrock Titan Image Generator and other generative image models.

Reference: AWS Generative AI Overview, Diffusion Models Explained - AWS Blog

Transformer-based models (such as GPT or Amazon Titan Text) are designed for generating and understanding natural language. These models can generate coherent, contextually relevant slogans

based on product information.

Reference: AWS Generative AI on Bedrock, Transformers Explained - AWS

Object detection models are designed to identify and locate objects within images, which makes them suitable for verifying that specific brand elements (like logos or products) are correctly



positioned in the generated content.

Reference: AWS Rekognition Object Detection, Object Detection Overview - AWS

CertEmpire

**Question: 36**

Which option is an example of unsupervised learning?

- A. A model that groups customers based on their purchase history
- B. A model that classifies images as dogs or cats
- C. A model that predicts a house's price based on various features
- D. A model that learns to play chess by using trial and error

**Answer:**

A

**Explanation:**

Unsupervised learning involves training a model on unlabeled data, letting it find patterns or groupings on its own, without explicit outputs provided. Clustering is a primary unsupervised learning technique.

Option A is correct: Grouping customers based on purchase history (without predefined categories) is

clustering, a classic unsupervised task.

B and C are supervised learning (classification and regression, respectively).

D is reinforcement learning, not unsupervised learning.

"Unsupervised learning involves training on data without labels and is often used for clustering or dimensionality reduction."

(Reference: AWS Certified AI Practitioner Official Study Guide, AWS ML Concepts)

**Question: 37**

A social media company wants to use a large language model (LLM) to summarize messages. The company has chosen a few LLMs that are available on Amazon SageMaker JumpStart. The company wants to compare the generated output toxicity of these models. Which strategy gives the company the ability to evaluate the LLMs with the LEAST operational overhead?

- A. Crowd-sourced evaluation
- B. Automatic model evaluation
- C. Model evaluation with human workers
- D. Reinforcement learning from human feedback (RLHF)

**Answer:**

B

**Explanation:**

The least operational overhead comes from automated tools that can scan and evaluate LLM outputs for toxicity. AWS and SageMaker JumpStart support integrations with automatic evaluation tools and

APIs (such as Amazon Comprehend or third-party toxicity classifiers).

B is correct: Automated evaluation provides quick, scalable, and repeatable analysis, requiring minimal human intervention.

A and C require manual effort, increasing operational overhead.

D (RLHF) is resource-intensive and not designed for rapid, automated model comparison.

"Automated evaluation can quickly assess generated text for specific attributes like toxicity, sentiment, or compliance using pre-trained classifiers, reducing human involvement and operational complexity."

(Reference: AWS SageMaker JumpStart Evaluation, AWS AI Practitioner Guide)

## Question: 38

An AI practitioner is developing a prompt for an Amazon Titan model. The model is hosted on Amazon Bedrock. The AI practitioner is using the model to solve numerical reasoning challenges.

The

AI practitioner adds the following phrase to the end of the prompt: "Ask the model to show its work

by explaining its reasoning step by step."

Which prompt engineering technique is the AI practitioner using?

- A. Chain-of-thought prompting
- B. Prompt injection
- C. Few-shot prompting
- D. Prompt templating

### Answer:

A

### Explanation:

Chain-of-thought prompting is a prompt engineering technique where you instruct the model to explain its reasoning step by step, which is particularly useful for tasks involving logic, math, or reasoning.

A is correct: Asking the model to "explain its reasoning step by step" directly invokes chain-of-thought prompting, as documented in AWS and generative AI literature.

B is unrelated (prompt injection is a security concern).

C (few-shot) provides examples, but doesn't specifically require step-by-step reasoning.

D (templating) is about structuring the prompt format.

"Chain-of-thought prompting elicits step-by-step explanations from LLMs, which improves performance on complex reasoning tasks."

(Reference: Amazon Bedrock Prompt Engineering Guide, AWS Certified AI Practitioner Study Guide)

**Question: 39**

A large retail bank wants to develop an ML system to help the risk management team decide on loan

allocations for different demographics.

What must the bank do to develop an unbiased ML model?

- A. Reduce the size of the training dataset.
- B. Ensure that the ML model predictions are consistent with historical results.
- C. Create a different ML model for each demographic group.
- D. Measure class imbalance on the training dataset. Adapt the training process accordingly.

**Answer:**

D

**Explanation:**

Class imbalance in a training dataset can cause ML models to favor overrepresented groups, leading to biased predictions. The AWS AI Practitioner guide and SageMaker Clarify documentation emphasize the need to identify and mitigate class imbalance to ensure fairness and unbiased model outcomes.

CertEmpire

D is correct: By measuring class imbalance and adapting the training process (e.g., through oversampling, undersampling, or using class weights), organizations can improve fairness and reduce bias across demographic groups.

A (reducing data size) could worsen bias by removing potentially useful diverse data.

B (consistency with historical results) might reinforce existing biases.

C (separate models) is not scalable and can introduce other fairness issues.

'To reduce bias, examine class imbalance in your training data and use techniques to ensure all groups are fairly represented.'

(Reference: AWS SageMaker Clarify: Mitigating Bias, AWS Responsible AI)

**Question: 40**

A company has developed an ML model to predict real estate sale prices. The company wants to deploy the model to make predictions without managing servers or infrastructure.

Which solution meets these requirements?

- A. Deploy the model on an Amazon EC2 instance.
- B. Deploy the model on an Amazon Elastic Kubernetes Service (Amazon EKS) cluster.
- C. Deploy the model by using Amazon CloudFront with an Amazon S3 integration.
- D. Deploy the model by using an Amazon SageMaker AI endpoint.

**Answer:**

D

**Explanation:**

Amazon SageMaker endpoints provide fully managed, serverless model deployment for real-time and batch predictions, allowing companies to deploy ML models without handling any servers or infrastructure management.

D is correct: SageMaker endpoints let you deploy, scale, and monitor ML models with no infrastructure overhead.

A and B require infrastructure management. CertEmpire

C (CloudFront/S3) is not for model deployment, but for static content delivery.

'Amazon SageMaker endpoints allow you to deploy machine learning models for inference without the need to manage underlying infrastructure.'

(Reference: AWS SageMaker Model Deployment, AWS Certified AI Practitioner Study Guide)