AWS AIF-C01 Exam Questions

Total Questions: 150+
Demo Questions: 29
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

An AI practitioner trained a custom model on Amazon Bedrock by using a training dataset that contains confidential data. The AI practitioner wants to ensure that the custom model does not generate inference responses based on confidential data. How should the AI practitioner prevent responses based on confidential data?

A: Delete the custom model. Remove the confidential data from the training dataset. Retrain the custom model.

B: Mask the confidential data in the inference responses by using dynamic data masking.

C: Encrypt the confidential data in the inference responses by using Amazon SageMaker.

D: Encrypt the confidential data in the custom model by using AWS Key Management Service (AWS KMS).

Correct Answer:

A

Explanation:

When a machine learning model is trained, it learns patterns, relationships, and sometimes specific information from its training dataset. If confidential data is included in this set, the model can "memorize" it and may reproduce it during inference. This is a form of data leakage. The only definitive way to prevent the model from generating responses based on this learned confidential data is to remove the data from the source (the training dataset) and retrain the model from scratch. This ensures the model's internal parameters are never influenced by the confidential information.

Why Incorrect Options are Wrong:

B: Masking is a reactive, post-processing step applied to the model's output. It does not prevent the model from generating the confidential data in the first place and may be imperfect.

C: Encrypting inference responses is a data protection measure for transmission or storage, not a method to prevent the model from generating the sensitive content itself.

D: Encrypting the model artifact with AWS KMS protects the model file at rest. It does not alter the learned knowledge within the model that allows it to generate confidential data.

References:

1. AWS Bedrock Documentation - Custom Models: The process of creating a custom model is through fine-tuning or continued pre-training on a user-provided dataset. The documentation states, "To fine-tune a foundation model, you provide a training dataset... Amazon Bedrock uses your dataset to create a new, fine-tuned version of the foundation model." This implies the model's knowledge is directly derived from the training data, and to change that knowledge, the data must be changed and the process repeated.

Source: AWS Bedrock User Guide, "Custom models," Section: "Fine-tuning".

2. Academic Publication on Data Privacy in ML: Research on "machine unlearning" explores methods to make a model forget specific training data. The most straightforward and guaranteed method, often used as a baseline, is to remove the data and retrain the model from scratch. This is referred to as "retraining from scratch."

Source: Bourtoule, L., et al. (2021). Machine Unlearning. 2021 IEEE Symposium on Security and Privacy (SP), pp. 141-159. IEEE. (This paper establishes that retraining is the gold standard for data removal).

3. AWS Security Documentation - Data Protection in Amazon Bedrock: AWS documentation details how KMS is used to encrypt data at rest, including training data and custom models. This protection applies to the stored artifacts, not the logical content learned by the model during training.

Source: AWS Bedrock User Guide, "Security," Section: "Data protection in Amazon Bedrock".

Question: 2

Which feature of Amazon OpenSearch Service gives companies the ability to build vector database applications?

A: Integration with Amazon S3 for object storage

B: Support for geospatial indexing and queries

C: Scalable index management and nearest neighbor search capability

D: Ability to perform real-time analysis on streaming data

Correct Answer:

C

Explanation:

Vector database applications are built on the capability to store high-dimensional vectors (embeddings) and perform similarity searches. The core function for this is finding the vectors "closest" to a given query vector. Amazon OpenSearch Service provides this through its k-Nearest Neighbor (k-NN) search feature. This allows for efficient and scalable nearest neighbor searches within an index, which is the fundamental operation required for use cases like semantic search, recommendation engines, and image retrieval that rely on vector embeddings.

Why Incorrect Options are Wrong:

A: Integration with Amazon S3 is a data ingestion and backup feature, not the core mechanism that enables vector search functionality itself.

B: Geospatial indexing is a specialized feature for location-based data and queries, which is distinct from the high-dimensional vector search used in AI applications.

D: Real-time analysis on streaming data is a general capability for log analytics and monitoring, not the specific feature that enables vector similarity search.

References:

Amazon OpenSearch Service Documentation: "Vector database capabilities of Amazon OpenSearch Service." This page explicitly states, "OpenSearch Service offers the ability to build vector database solutions with its k-Nearest Neighbor (k-NN) search feature." It details how k-NN search is used for similarity search on vector embeddings.

URL: https://aws.amazon.com/opensearch-service/features/vector-database/

Amazon OpenSearch Service Developer Guide: "k-Nearest Neighbor (k-NN) search in Amazon OpenSearch Service." This guide describes the knn vector type and the k-NN search functionality as the basis for vector search.

URL: https://docs.aws.amazon.com/opensearch-service/latest/developerguide/knn.html

AWS Machine Learning Blog: "Build a powerful semantic search engine with Amazon OpenSearch Service." This article demonstrates building a vector search application and highlights the k-NN index as the central component.

URL: https://aws.amazon.com/blogs/machine-learning/build-a-powerful-semantic-search-engine-with-amazon-opensearch-service/

Question: 3

A company wants to display the total sales for its top-selling products across various retail locations in the past 12 months. Which AWS solution should the company use to automate the generation of graphs?

A: Amazon Q in Amazon EC2

B: Amazon Q Developer

C: Amazon Q in Amazon QuickSight

D: Amazon Q in AWS Chatbot

Correct Answer:

C

Explanation:

Amazon QuickSight is the AWS business intelligence (BI) service specifically designed for creating and publishing interactive dashboards, which include various types of graphs and visualizations. The integration of Amazon Q within QuickSight allows users to ask natural language questions to build visuals automatically. A user can simply type a query like, "What are the total sales for top products by location in the last year?" and Amazon Q will generate the corresponding graph. This directly fulfills the requirement to automate the generation of graphs for business sales data.

Why Incorrect Options are Wrong:

A: Amazon Q in Amazon EC2: This integration assists with troubleshooting, optimizing, and managing applications and resources running on Amazon EC2, not with business data visualization.

B: Amazon Q Developer: This is an AI-powered assistant for software developers, helping with code generation, debugging, and testing within an IDE, not creating BI dashboards.

D: Amazon Q in AWS Chatbot: This allows users to ask questions about their AWS environment and resources via chat applications like Slack, focusing on operational insights, not business analytics.

References:

1. Amazon Web Services (AWS). "Amazon Q in Amazon QuickSight." Amazon QuickSight Documentation. Accessed May 2024. This page details how Amazon Q uses generative AI to allow users to ask questions in natural language to build, discover, and share insights.

URL: https://aws.amazon.com/quicksight/q/

2. Amazon Web Services (AWS). "What is Amazon QuickSight?" Amazon QuickSight User Guide. Accessed May 2024. This document defines QuickSight as a cloud-scale business intelligence (BI) service used to deliver insights, including visualizations and dashboards.

URL: https://docs.aws.amazon.com/quicksight/latest/user/what-is-quicksight.html

3. Amazon Web Services (AWS). "Amazon Q Developer." AWS Documentation. Accessed May 2024. This source describes Amazon Q Developer as an AI assistant for the entire software development life cycle, distinguishing its purpose from BI.

URL: https://aws.amazon.com/q/developer/

Question: 4

A company wants to build an interactive application for children that generates new stories based on classic stories. The company wants to use Amazon Bedrock and needs to ensure that the results and topics are appropriate for children. Which AWS service or feature will meet these requirements?

A: Amazon Rekognition

B: Amazon Bedrock playgrounds

C: Guardrails for Amazon Bedrock

D: Agents for Amazon Bedrock

Correct Answer:

C

Explanation:

Guardrails for Amazon Bedrock is the feature specifically designed to implement safeguards and enforce responsible AI policies in generative AI applications. It allows developers to define denied topics to prevent the model from generating content on subjects deemed inappropriate for children. Additionally, it provides configurable filters to screen for harmful content such as hate speech, violence, and sexual content in both user inputs and model responses. This directly meets the requirement to ensure the generated stories are appropriate for a young audience.

Why Incorrect Options are Wrong:

A: Amazon Rekognition: This service is for image and video analysis. It is not used for moderating or controlling text generation, which is the core of the scenario.

B: Amazon Bedrock playgrounds: These are interactive development environments within the AWS console used for experimenting with models, not for implementing production-level safety controls in a live application.

D: Agents for Amazon Bedrock: Agents are used to orchestrate multi-step tasks and execute API calls. While part of an application's logic, they are not the feature responsible for content safety filtering.

References:

1. Guardrails for Amazon Bedrock: "With Guardrails for Amazon Bedrock, you can create policies to safeguard your generative AI applications... You can define denied topics that are irrelevant to your application's function or that you want to avoid... You can also configure thresholds to filter harmful content across categories like hate, insults, sexual, and violence."

Source: AWS Documentation, "Guardrails for Amazon Bedrock," aws.amazon.com/bedrock/guardrails/.

2. Agents for Amazon Bedrock: "Agents for Amazon Bedrock is a fully managed capability that makes it easier for developers to create generative AI-based applications that can complete complex tasks..."

Source: AWS Documentation, "Agents for Amazon Bedrock," aws.amazon.com/bedrock/agents/.

3. Amazon Rekognition: "Amazon Rekognition makes it easy to add image and video analysis to your applications... Rekognition can identify objects, people, text, scenes, and activities in images and videos..."

Source: AWS Documentation, "What Is Amazon Rekognition?," docs.aws.amazon.com/rekognition/latest/dg/what-is.html.

Question: 5

A company has developed an ML model for image classification. The company wants to deploy the model to production so that a web application can use the model. The company needs to implement a solution to host the model and serve predictions without managing any of the underlying infrastructure. Which solution will meet these requirements?

A: Use Amazon SageMaker Serverless Inference to deploy the model.

B: Use Amazon CloudFront to deploy the model.

C: Use Amazon API Gateway to host the model and serve predictions.

D: Use AWS Batch to host the model and serve predictions.

Correct Answer:

A

Explanation:

Amazon SageMaker Serverless Inference is a purpose-built solution for deploying machine learning models without managing the underlying infrastructure. It automatically provisions, scales, and manages the necessary compute resources based on the volume of inference requests. This aligns perfectly with the requirement to host a model and serve predictions for a web application in a serverless manner, where the company only pays for the compute time used and the amount of data processed, abstracting away all infrastructure concerns.

Why Incorrect Options are Wrong:

B: Use Amazon CloudFront to deploy the model.

C: Use Amazon API Gateway to host the model and serve predictions.

D: Use AWS Batch to host the model and serve predictions.

References:

1. Amazon SageMaker Developer Guide: "With SageMaker Serverless Inference, you can quickly deploy your machine learning models for inference without having to configure or manage the underlying infrastructure... Serverless Inference is ideal for workloads that have intermittent or unpredictable traffic."

URL: https://docs.aws.amazon.com/sagemaker/latest/dg/serverless-inference.html

2. Amazon CloudFront Developer Guide: "Amazon CloudFront is a web service that speeds up distribution of your static and dynamic web content... to your users. You can use CloudFront with any origin server, which is the server where you store the original, definitive version of your content."

URL: https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/Introduction.html

3. AWS Batch User Guide: "AWS Batch enables developers, scientists, and engineers to easily and efficiently run hundreds of thousands of batch computing jobs on AWS."

URL: https://docs.aws.amazon.com/batch/latest/userguide/what-is-batch.html

Question: 6

A company has petabytes of unlabeled customer data to use for an advertisement campaign. The company wants to classify its customers into tiers to advertise and promote the company's products. Which methodology should the company use to meet these requirements?

A: Supervised learning

B: Unsupervised learning

C: Reinforcement learning

D: Reinforcement learning from human feedback (RLHF)

Correct Answer:

B

Explanation:

The core of the question lies in two key details: the data is "unlabeled," and the goal is to "classify its customers into tiers." This process of identifying inherent structures, patterns, or groupings within data that lacks predefined labels is the definition of unsupervised learning. Specifically, the company would use a clustering algorithm (a type of unsupervised learning) to segment customers based on similarities in their data, creating the desired tiers for targeted advertising.

Why Incorrect Options are Wrong:

A: Supervised learning is incorrect because it requires labeled data (i.e., data where the correct output is already known) to train a model, which is not available in this scenario.

C: Reinforcement learning is incorrect as it focuses on training an agent to make sequential decisions in an environment to maximize a reward, which is not applicable to grouping a static dataset.

D: Reinforcement learning from human feedback (RLHF) is a specialized type of reinforcement learning and is irrelevant to the task of clustering unlabeled customer data.

References:

1. Official AWS Documentation: According to AWS, "Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels... A popular unsupervised learning technique is clustering. You can use it for

customer segmentation to group customers with similar purchase attributes." This directly aligns with the scenario.

Source: AWS, "What is Unsupervised Learning?", aws.amazon.com/what-is/unsupervised-learning/

2. University Courseware: Stanford University's machine learning course defines unsupervised learning as the task of finding structure in unlabeled data. It lists clustering as a primary example, where the goal is to group data points into coherent clusters.

Source: Ng, (2023). CS229: Machine Learning - Lecture Notes, Part V: Unsupervised Learning. Stanford University, cs229.stanford.edu/notes2022fall/mainnotes.pdf (See Section on Unsupervised Learning).

3. Peer-Reviewed Academic Publication: In the foundational text The Elements of Statistical Learning, the authors clearly distinguish the learning paradigms. Unsupervised learning is described as a scenario where "we observe only the features... and have no measurements of an outcome... The goal is rather to describe how the data are organized or clustered."

Source: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. (See Chapter 1, Section 1.3).

Question: 7

A company makes forecasts each quarter to decide how to optimize operations to meet expected demand. The company uses ML models to make these forecasts. An AI practitioner is writing a report about the trained ML models to provide transparency and explainability to company stakeholders. What should the AI practitioner include in the report to meet the transparency and explainability requirements?

A: Code for model training

B: Partial dependence plots (PDPs)

C: Sample data for training

D: Model convergence tables

Correct Answer:

B

Explanation:

Partial dependence plots (PDPs) are a primary tool for model explainability. They illustrate the marginal effect of one or two features on the predicted outcome of a machine learning model, holding all other features constant. By visualizing these relationships, PDPs make the model's decision-making process more transparent and understandable to stakeholders, who may not have a technical background. This directly addresses the requirement to explain how the model arrives at its forecasts.

Why Incorrect Options are Wrong:

A: Code for model training: This is too technical for most stakeholders and explains the implementation details, not the model's predictive logic in an interpretable way.

C: Sample data for training: While providing context, sample data does not explain the patterns or relationships the model learned from the data to make its predictions.

D: Model convergence tables: These are technical metrics used by data scientists to verify that the model training process was successful, not to explain the model's behavior to stakeholders.

References:

1. AWS SageMaker Developer Guide: Amazon SageMaker Clarify provides tools for model explainability, including "partial dependence plots (PDPs) to help you visualize the marginal

effect of features on the predicted outcome." This directly links PDPs to the goal of explaining model behavior.

URL: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html

2. Molnar, (2022). Interpretable Machine Learning. This peer-recognized academic resource states, "The partial dependence plot (short PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model." This defines PDPs as a key method for model interpretation.

URL: https://christophm.github.io/interpretable-ml-book/pdp.html (Section 5.1)

3. University of California, Berkeley - STAT 154 Course Materials: Course materials on model interpretation and explainability frequently cover PDPs as a fundamental technique for understanding black-box models.

URL: https://www.stat.berkeley.edu/~breiman/pdptalk.pdf (Page 3, "What is a Partial Dependence Plot?")

Question: 8

Which option is a use case for generative AI models?

A: Improving network security by using intrusion detection systems

B: Creating photorealistic images from text descriptions for digital marketing

C: Enhancing database performance by using optimized indexing

D: Analyzing financial data to forecast stock market trends

Correct Answer:

B

Explanation:

Generative AI is a category of artificial intelligence focused on creating new, original content, such as text, images, audio, and synthetic data. The process of generating photorealistic images from textual descriptions is a hallmark application of generative AI models, often referred to as text-to-image synthesis. These models learn patterns from vast datasets of images and their corresponding text labels to produce novel visual content that aligns with a user's prompt.

Why Incorrect Options are Wrong:

A: Intrusion detection systems primarily use discriminative AI models for classification and anomaly detection to identify threats, which is an analytical task, not a generative one.

C: Optimizing database indexing is a performance engineering task that relies on algorithms and data structure management, not on creating new content with generative AI.

D: Forecasting stock market trends is a predictive analytics task. It uses historical data to predict future outcomes, which is a form of predictive AI, not generative AI.

References:

1. AWS Documentation on Generative AI: "Generative artificial intelligence (generative AI) is a type of artificial intelligence (AI) that can create new content and ideas, including conversations, stories, images, videos, and music." This source explicitly lists creating images as a core capability.

Source: AWS, "What is Generative AI?", aws.amazon.com/what-is/generative-ai/

2. AWS Documentation on Amazon Forecast: This service is used for predictive tasks like the one described in option "Amazon Forecast is a fully managed service that uses statistical and machine learning (ML) algorithms to deliver highly accurate time-series forecasts." This distinguishes it from generative tasks.

Source: AWS, "What is Amazon Forecast?", docs.aws.amazon.com/forecast/latest/dg/what-is-forecast.html

3. Academic Definition: Generative models are contrasted with discriminative models. Generative models learn the joint probability distribution of the data, allowing them to generate new data points, while discriminative models learn the boundary between classes for classification tasks (like intrusion detection or forecasting).

Source: Goodfellow, I., Bengio, Y., & Courville, (2016). Deep Learning. MIT Press. Chapter 3: Probability and Information Theory.

Question: 9

An AI practitioner is using a large language model (LLM) to create content for marketing campaigns. The generated content sounds plausible and factual but is incorrect. Which problem is the LLM having?

A: Data leakage

B: Hallucination

C: Overfitting

D: Underfitting

Correct Answer:

B

Explanation:

The phenomenon described is known as hallucination. In the context of large language models (LLMs), hallucination occurs when the model generates text that is fluent, plausible, and grammatically correct but is factually incorrect or not grounded in the source data. The model essentially "invents" information, presenting it with confidence, which is a significant challenge when using AI for tasks requiring factual accuracy, such as creating marketing content.

Why Incorrect Options are Wrong:

A: Data leakage: This is a model training error where information from outside the training set is improperly used, leading to inflated performance metrics, not the generation of false content.

C: Overfitting: This occurs when a model learns its training data too well, including noise, and fails to generalize to new data. It is a generalization problem, not the specific act of fabricating plausible facts.

D: Underfitting: This describes a model that is too simple to capture the underlying patterns in the data, leading to poor performance overall, not plausible but incorrect outputs.

References:

1. Official Vendor Documentation: AWS. (2023). Generative AI with Large Language Models. AWS Skill Builder. In Module 3, "Large Language Models," the concept of

hallucination is described as a limitation where models generate responses that are nonsensical or not factually correct.

2. Peer-Reviewed Academic Publication: Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), Article 248, p. 2. The paper defines hallucination as "generated content that is nonsensical or unfaithful to the provided source content."

3. University Courseware: Stanford University Human-Centered Artificial Intelligence (HAI). (2023). What Are AI Hallucinations? Stanford HAI. The article explains that hallucinations are "outputs that are nonsensical or untruthful" and are a result of the model making confident predictions that are not justified by its training data. https://hai.stanford.edu/news/what-are-ai-hallucinations

Question: 10

A loan company is building a generative AI-based solution to offer new applicants discounts based on specific business criteria. The company wants to build and use an AI model responsibly to minimize bias that could negatively affect some customers. Which actions should the company take to meet these requirements? (Select TWO.)

A: Detect imbalances or disparities in the data.

B: Ensure that the model runs frequently.

C: Evaluate the model's behavior so that the company can provide transparency to stakeholders.

D: Use the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) technique to ensure that the model is 100% accurate.

E: Ensure that the model's inference time is within the accepted limits.

Correct Answer:

A, C

Explanation:

Building a responsible AI solution requires focusing on fairness, explainability, and transparency. Detecting imbalances or disparities in the training data (A) is a critical first step to mitigate bias, ensuring the model does not unfairly disadvantage certain customer groups. Evaluating the model's behavior (C) is essential for explainability and transparency. This allows the company to understand and justify the model's decisions to stakeholders, such as customers and regulators, which is a core principle of responsible AI. Together, these actions address the key requirements of minimizing bias and ensuring transparency.

Why Incorrect Options are Wrong:

B: The frequency of model execution is an operational metric related to system performance, not a measure of fairness or transparency.

D: ROUGE is a metric for evaluating text summarization, not a loan application model. Furthermore, 100% accuracy is an unrealistic and often undesirable goal in machine learning.

E: Inference time is a performance metric concerned with speed and efficiency, which is separate from the ethical considerations of bias and transparency.

References:

1. AWS Responsible AI Documentation: "Fairness and explainability are tenets of responsible AI... Amazon SageMaker Clarify provides tools to help you build fairer and more understandable machine learning models. It detects potential bias during data preparation, after model training, and in your deployed model..." This supports detecting imbalances (A) and evaluating behavior for transparency (C).

Source: AWS. (n.d.). Fairness and Explainability with Amazon SageMaker Clarify. AWS Documentation. Retrieved from https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-fairness-and-explainability.html

2. Academic Publication on Algorithmic Fairness: Mehrabi, N., et al. (2021). "A Survey on Bias and Fairness in Machine Learning." ACM Computing Surveys, 54(6), Article 115. This paper extensively discusses that a primary source of bias is skewed or imbalanced data (supporting A) and that transparency and explainability are crucial for auditing and trusting AI systems (supporting C).

Source: Mehrabi, N., et al. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys. https://dl.acm.org/doi/10.1145/3457607 (Section 3 discusses data bias, and Section 6 discusses transparency).

3. AWS AI/ML Well-Architected Framework: The framework's whitepaper emphasizes the importance of the "Fairness" design principle, which includes "Evaluate the model for different types of bias" and "Provide transparency into how the model works." This directly aligns with options A and C.

Source: AWS. (2023). The Machine Learning Lens - AWS Well-Architected Framework. Retrieved from https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/fairness.html

Question: 11

A medical company is customizing a foundation model (FM) for diagnostic purposes. The company needs the model to be transparent and explainable to meet regulatory requirements. Which solution will meet these requirements?

A: Configure the security and compliance by using Amazon Inspector.

B: Generate simple metrics, reports, and examples by using Amazon SageMaker Clarify.

C: Encrypt and secure training data by using Amazon Macie.

D: Gather more data. Use Amazon Rekognition to add custom labels to the data.

Correct Answer:

B

Explanation:

Amazon SageMaker Clarify is the designated AWS service for enhancing model transparency and explainability. It provides tools to detect potential bias and, crucially, to explain how models make predictions. By generating feature attribution reports (e.g., using SHAP), it helps stakeholders understand which data features most influenced a model's outcome. This capability directly fulfills the regulatory requirement for a transparent and explainable AI model, which is essential in sensitive domains like medical diagnostics.

Why Incorrect Options are Wrong:

A: Amazon Inspector is an infrastructure security service that scans for vulnerabilities; it does not analyze or explain machine learning model behavior.

C: Amazon Macie is a data security service for discovering and protecting sensitive data; it does not provide insights into model predictions.

D: While more data can improve a model, gathering and labeling it does not inherently provide the transparency or explainability required by regulations.

References:

AWS Documentation - Amazon SageMaker Clarify: "Amazon SageMaker Clarify helps improve your machine learning (ML) models by detecting potential bias and helping explain how these models make predictions." This directly aligns with the need for transparency and explainability.

URL: https://aws.amazon.com/sagemaker/clarify/

AWS Developer Guide - Amazon SageMaker Clarify: "Amazon SageMaker Clarify provides tools for you to gain a deeper understanding of your machine learning (ML) models and data. It helps you explain the behavior of your models..."

URL: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-overview.html

AWS Whitepaper - Our Perspective on Responsible AI: This paper discusses explainability as a key pillar of responsible AI and cites SageMaker Clarify as a primary tool. "Amazon SageMaker Clarify helps customers detect bias in data and models and explain model predictions." (Page 6)

URL: https://d1.awsstatic.com/responsible-ai/our-perspective-on-responsible-ai.pdf

Question: 12

A company is building a solution to generate images for protective eyewear. The solution must have high accuracy and must minimize the risk of incorrect annotations. Which solution will meet these requirements?

A: Human-in-the-loop validation by using Amazon SageMaker Ground Truth Plus

B: Data augmentation by using an Amazon Bedrock knowledge base

C: Image recognition by using Amazon Rekognition

D: Data summarization by using Amazon QuickSight

Correct Answer:

A

Explanation:

The primary requirements are high accuracy and minimizing incorrect annotations for an image dataset. Amazon SageMaker Ground Truth Plus is a fully managed data labeling service that uses an expert workforce and multi-step quality assurance workflows, including consensus models and reviews, to deliver high-quality training datasets. This human-in-the-loop approach is specifically designed to achieve high accuracy and validate annotations, directly addressing the problem of minimizing labeling errors for the protective eyewear images.

Why Incorrect Options are Wrong:

B: Data augmentation artificially expands a dataset but does not correct or validate the accuracy of the original annotations. Amazon Bedrock is a generative AI service, not an image annotation tool.

C: Amazon Rekognition is a service for performing image analysis using pre-trained models. It consumes labeled data; it does not create or validate the high-quality annotations required for training.

D: Amazon QuickSight is a business intelligence (BI) service for data visualization and creating dashboards. It is used for analysis, not for creating or validating machine learning training data.

References:

1. Amazon SageMaker Ground Truth Plus Documentation: "Amazon SageMaker Ground Truth Plus is a turnkey data labeling service that enables you to create high-quality training datasets without having to build labeling applications or manage the labeling workforce on your own... To ensure your labels have high accuracy, your dataset is labeled by a workforce trained on ML tasks."

Source: AWS Documentation, "Amazon SageMaker Ground Truth," https://aws.amazon.com/sagemaker/data-labeling/

2. Human-in-the-Loop Machine Learning: The use of human feedback to improve a model's performance or data quality is a core concept. SageMaker Ground Truth Plus institutionalizes this process for data labeling. "Human-in-the-loop computing is a branch of artificial intelligence that leverages both human and machine intelligence to create machine learning models."

Source: Wu, J., et al. (2022). "A Survey on Human-in-the-loop for Machine Learning." ACM Computing Surveys.

3. Amazon Rekognition Documentation: "Amazon Rekognition makes it easy to add image and video analysis to your applications using proven, highly scalable, deep learning technology that requires no machine learning expertise to use." This shows its role is analysis, not annotation.

Source: AWS Documentation, "What Is Amazon Rekognition?," https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html

Question: 13

A security company is using Amazon Bedrock to run foundation models (FMs). The company wants to ensure that only authorized users invoke the models. The company needs to identify any unauthorized access attempts to set appropriate AWS Identity and Access Management (IAM) policies and roles for future iterations of the FMs. Which AWS service should the company use to identify unauthorized users that are trying to access Amazon Bedrock?

A: AWS Audit Manager

B: AWS CloudTrail

C: Amazon Fraud Detector

D: AWS Trusted Advisor

Correct Answer:

B

Explanation:

AWS CloudTrail is the service designed to log all API calls made to AWS services within an account, including Amazon Bedrock. It captures a detailed record of who made a request, when it was made, from what IP address, and what action was requested. Unauthorized access attempts result in "AccessDenied" errors, which are explicitly logged as events in CloudTrail. By analyzing these logs, the security company can identify which users or roles are making unauthorized attempts and use this information to refine their AWS Identity and Access Management (IAM) policies and roles for better security.

Why Incorrect Options are Wrong:

A: AWS Audit Manager: This service automates evidence collection for compliance and audits. It uses CloudTrail logs as a data source but is not the primary tool for identifying individual access attempts.

C: Amazon Fraud Detector: This is a managed service for detecting business-level fraudulent activities like fake account creation or payment fraud, not for monitoring AWS API access permissions.

D: AWS Trusted Advisor: This service provides high-level recommendations on cost optimization, security, and performance based on AWS best practices. It does not provide detailed logs of API access attempts.

References:

AWS CloudTrail Documentation: "AWS CloudTrail is a service that enables governance, compliance, operational auditing, and risk auditing of your AWS account. With CloudTrail, you can log, continuously monitor, and retain account activity related to actions across your AWS infrastructure. CloudTrail provides a complete history of user activity and API calls for your AWS account." (AWS CloudTrail User Guide, "What Is AWS CloudTrail?", Introduction).

Amazon Bedrock and CloudTrail Integration: "Amazon Bedrock is integrated with AWS CloudTrail, a service that provides a record of actions taken by a user, role, or an AWS service in Bedrock. CloudTrail captures all API calls for Bedrock as events." (AWS Bedrock User Guide, "Logging Amazon Bedrock API calls with AWS CloudTrail").

Using CloudTrail for Security: "By using the information collected by CloudTrail, you can determine the request that was made to an AWS service, the IP address from which the request was made, who made the request, when it was made, and additional details. For example, you can identify which users have tried to access Amazon S3 buckets without the correct permissions." (AWS CloudTrail User Guide, "Security in AWS CloudTrail").

Question: 14

A pharmaceutical company wants to analyze user reviews of new medications and provide a concise overview for each medication. Which solution meets these requirements?

A: Create a time-series forecasting model to analyze the medication reviews by using Amazon Personalize.

B: Create medication review summaries by using Amazon Bedrock large language models (LLMs).

C: Create a classification model that categorizes medications into different groups by using Amazon SageMaker.

D: Create medication review summaries by using Amazon Rekognition.

Correct Answer:

B

Explanation:

The core requirement is to analyze user reviews (text data) and generate a "concise overview," which is a text summarization task. Amazon Bedrock provides access to powerful large language models (LLMs) that excel at natural language understanding and generation. These models can process large volumes of text from reviews and create coherent, human-readable summaries, directly fulfilling the company's need for a concise overview of medication feedback.

Why Incorrect Options are Wrong:

A: Amazon Personalize is a service for building recommendation systems, not for text analysis or summarization. Time-series forecasting is also irrelevant to creating a summary.

C: A classification model would categorize reviews (e.g., by sentiment or topic) but would not generate a narrative summary or "concise overview" as requested.

D: Amazon Rekognition is a computer vision service for analyzing images and videos. It is not designed to process or summarize text-based content like medication reviews.

References:

1. Amazon Bedrock for Summarization: The Amazon Bedrock User Guide explicitly lists summarization as a primary use case. It states, "Summarization - Get a summary of long-

form content such as articles, blog posts, books, and documents to get the gist without having to read the full text."

Source: AWS Documentation, "Amazon Bedrock User Guide," Section: "Common use cases for foundation models." (URL: https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html#what-is-bedrock-use-cases)

2. Amazon Personalize Purpose: The official documentation defines Amazon Personalize as a service for creating personalized user experiences and recommendations.

Source: AWS Documentation, "What Is Amazon Personalize?," Amazon Personalize Developer Guide. (URL: https://docs.aws.amazon.com/personalize/latest/dg/what-is-personalize.html)

3. Amazon Rekognition Purpose: The official documentation describes Amazon Rekognition as a service for image and video analysis.

Source: AWS Documentation, "What Is Amazon Rekognition?," Amazon Rekognition Developer Guide. (URL: https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html)

4. Classification vs. Summarization: Academic literature distinguishes text classification (assigning predefined labels) from text summarization (generating new, shorter text). Classification does not produce a summary.

Source: Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Chapter 24 discusses summarization as a sequence-to-sequence task, distinct from classification. (URL: https://web.stanford.edu/~jurafsky/slp3/)

Question: 15

A company wants to assess the costs that are associated with using a large language model (LLM) to generate inferences. The company wants to use Amazon Bedrock to build generative AI applications. Which factor will drive the inference costs?

A: Number of tokens consumed

B: Temperature value

C: Amount of data used to train the LLM

D: Total training time

Correct Answer:

A

Explanation:

The primary factor driving inference costs for large language models (LLMs) on Amazon Bedrock is the number of tokens processed. The on-demand pricing model bills based on the volume of text data handled during an inference request. This includes both the tokens in the input prompt provided by the user and the tokens in the output generated by the model. Therefore, longer inputs and more extensive generated responses directly increase the overall cost of using the service for inference.

Why Incorrect Options are Wrong:

B: Temperature value: The temperature is a hyperparameter that controls the randomness of the model's output. It does not directly influence the computational workload or the billing metric for an inference call.

C: Amount of data used to train the LLM: This is a cost associated with the initial training or subsequent fine-tuning of a model, not the cost of running inference on an existing, pre-trained model.

D: Total training time: Similar to the amount of training data, the time spent training a model is a cost factor for model creation or customization, which is separate from inference costs.

References:

Amazon Web Services (AWS). (n.d.). Amazon Bedrock Pricing. Retrieved from https://aws.amazon.com/bedrock/pricing/. The official pricing page explicitly states, "With the on-demand mode, you pay for model inference based on the number of input tokens

and output tokens." This directly supports that token consumption is the key cost driver for inference.

Amazon Web Services (AWS). (2023). Generative AI on AWS. AWS Whitepaper. Page 18 discusses the operational cost of generative AI, highlighting that inference costs are a significant component and are often tied to usage metrics like the number of API calls or tokens processed. (A general reference to the concept within an official AWS whitepaper).

Question: 16

A company wants to create a new solution by using AWS Glue. The company has minimal programming experience with AWS Glue. Which AWS service can help the company use AWS Glue?

A: Amazon Q Developer

B: AWS Config

C: Amazon Personalize

D: Amazon Comprehend

Correct Answer:

A

Explanation:

Amazon Q Developer is a generative AI-powered assistant designed to help users build, secure, and operate applications on AWS. For a company with minimal programming experience, Amazon Q can interpret natural language prompts to generate code, such as AWS Glue scripts for data integration tasks. This capability directly addresses the challenge of using AWS Glue without extensive programming knowledge by automating script creation and providing expert guidance, thereby accelerating development and lowering the technical barrier to entry.

Why Incorrect Options are Wrong:

B: AWS Config: This service is for auditing and evaluating the configurations of AWS resources for compliance and governance, not for assisting with code development.

C: Amazon Personalize: This is a managed machine learning service used to create recommendation engines and user personalization, not a tool to help program other AWS services.

D: Amazon Comprehend: This is a natural language processing (NLP) service for analyzing text to find insights; it does not generate code or assist with developing on AWS Glue.

References:

Amazon Q Developer: AWS. (2024). Amazon Q Developer. "Amazon Q Developer is your AI assistant for the entire software development life cycle (SDLC) on AWS... Get expert

guidance... generate code from natural language prompts." Retrieved from https://aws.amazon.com/q/developer/

AWS Glue: AWS. (2024). What is AWS Glue?. AWS Documentation. "AWS Glue is a serverless data integration service that makes it easier to discover, prepare, move, and integrate data from multiple sources..." Retrieved from https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html

AWS Config: AWS. (2024). What Is AWS Config?. AWS Documentation. "AWS Config is a service that enables you to assess, audit, and evaluate the configurations of your AWS resources." Retrieved from https://docs.aws.amazon.com/config/latest/developerguide/what-is-config.html

Question: 17

A company is using Amazon SageMaker Studio notebooks to build and train ML models. The company stores the data in an Amazon S3 bucket. The company needs to manage the flow of data from Amazon S3 to SageMaker Studio notebooks. Which solution will meet this requirement?

A: Use Amazon Inspector to monitor SageMaker Studio.

B: Use Amazon Macie to monitor SageMaker Studio.

C: Configure SageMaker to use a VPC with an S3 endpoint.

D: Configure SageMaker to use S3 Glacier Deep Archive.

Correct Answer:

C

Explanation:

To manage and secure the flow of data between Amazon SageMaker Studio and Amazon S3, the best practice is to configure SageMaker Studio to operate within a Virtual Private Cloud (VPC). By creating a VPC endpoint for Amazon S3, you ensure that the data traffic between your SageMaker Studio notebooks and your S3 bucket travels over the private AWS network, not the public internet. This provides a secure, controlled, and managed path for data access, directly addressing the requirement to manage the data flow.

Why Incorrect Options are Wrong:

A: Amazon Inspector is a vulnerability management service that scans for software vulnerabilities and unintended network exposure; it does not manage data flow between services.

B: Amazon Macie is a data security service that discovers and protects sensitive data within Amazon S3; it does not control the network path for data access.

D: S3 Glacier Deep Archive is a storage class for long-term data archiving. Its slow retrieval times (hours) make it unsuitable for active ML model training.

References:

1. AWS SageMaker Documentation: "Connect SageMaker Studio Notebooks in a VPC to External Resources". This document explicitly states, "To control access to your data and model artifacts, we recommend that you create a private Amazon S3 bucket... Create a

VPC endpoint for Amazon S3. This allows Studio to access the buckets that store your data and model artifacts, without the traffic going over the internet."

URL: https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html

2. AWS PrivateLink Documentation: "Access an AWS service using an interface VPC endpoint". This explains how VPC endpoints provide private connectivity to AWS services like S3.

URL: https://docs.aws.amazon.com/vpc/latest/privatelink/privatelink-access-aws-services.html

3. AWS S3 Documentation: "Amazon S3 storage classes". This page details the use cases for each storage class, highlighting that S3 Glacier Deep Archive is for long-term archiving with retrieval times of 12 hours or more.

URL: https://aws.amazon.com/s3/storage-classes/

Question: 18

What are tokens in the context of generative AI models?

A: Tokens are the basic units of input and output that a generative AI model operates on, representing words, subwords, or other linguistic units.

B: Tokens are the mathematical representations of words or concepts used in generative AI models.

C: Tokens are the pre-trained weights of a generative AI model that are fine-tuned for specific tasks.

D: Tokens are the specific prompts or instructions given to a generative AI model to generate output.

Correct Answer:

A

Explanation:

In generative AI, particularly in Large Language Models (LLMs), text is not processed as a raw string of characters. Instead, it is first broken down into smaller, manageable pieces called tokens through a process called tokenization. A token is the fundamental unit of data that the model processes for both input and output. These units can be whole words (e.g., "hello"), parts of words or subwords (e.g., "token" and "ization" from "tokenization"), punctuation, or even individual characters, depending on the tokenization strategy used. The model's vocabulary consists of all possible tokens it can understand and generate.

Why Incorrect Options are Wrong:

B: This describes an embedding. An embedding is the dense vector (mathematical representation) that corresponds to a token, not the token itself.

C: This describes model weights or parameters. These are the numerical values within the neural network that are learned during training and are distinct from the input/output text units.

D: This describes a prompt. A prompt is the complete input text given to a model, which is itself composed of a sequence of tokens.

References:

1. Official Vendor Documentation (AWS): Amazon Bedrock User Guide, Glossary. It defines a token as "a sequence of characters that are grouped together for a common meaning. For example, a word, part of a word, or punctuation." This directly supports that tokens are the basic units of text.

URL: https://docs.aws.amazon.com/bedrock/latest/userguide/glossary.html

2. University Courseware (Stanford): Stanford University's CS224N: NLP with Deep Learning course materials explain tokenization as the first step in the NLP pipeline, which involves "breaking text into sentences and words or 'tokens'." The course further details subword tokenization methods like Byte-Pair Encoding (BPE), reinforcing that tokens can be words or subwords.

URL: https://web.stanford.edu/class/cs224n/ (See lectures on "Language Models" and "Subword Models").

3. Peer-reviewed Academic Publication (ACL): Sennrich, R., Haddow, B., & Birch, (2016). "Neural Machine Translation of Rare Words with Subword Units." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. This foundational paper on subword tokenization describes the method of representing text as a sequence of subword units (tokens) to handle vocabulary complexity.

URL: https://aclanthology.org/P16-1162/ (Section 2 describes the subword unit segmentation).

Question: 19

A retail company is tagging its product inventory. A tag is automatically assigned to each product based on the product description. The company created one product category by using a large language model (LLM) on Amazon Bedrock in few-shot learning mode. The company collected a labeled dataset and wants to scale the solution to all product categories. Which solution meets these requirements?

A: Use prompt engineering with zero-shot learning.

B: Use prompt engineering with prompt templates.

C: Customize the model with continued pre-training.

D: Customize the model with fine-tuning.

Correct Answer:

D

Explanation:

The company aims to scale its product categorization task from a single category (tested with few-shot learning) to its entire inventory, for which it has prepared a labeled dataset. Fine-tuning is the appropriate method for this scenario. It involves taking a pre-trained foundation model and further training it on a specific, labeled dataset to adapt its parameters for a particular task. This creates a custom model that is highly optimized for the company's unique product descriptions and categories, offering higher accuracy and performance at scale compared to prompt engineering alone.

Why Incorrect Options are Wrong:

A: Zero-shot learning provides no examples and would be a step backward from the few-shot approach already used, failing to leverage the valuable labeled dataset.

B: Prompt templates improve consistency but are still a form of prompt engineering. They do not fundamentally adapt the model's internal knowledge using the large labeled dataset.

C: Continued pre-training is used to adapt a model to a new domain using a large corpus of unlabeled data, not the labeled dataset the company has prepared.

References:

1. AWS Documentation on Fine-Tuning: "Fine-tuning customizes a foundation model by training it with a labeled dataset to improve performance on a specific task... For example,

you can fine-tune a model to improve its performance on classifying text for your specific business needs."

Source: Amazon Web Services, "Custom models in Amazon Bedrock," Section: "Fine-tuning." Retrieved from https://docs.aws.amazon.com/bedrock/latest/userguide/custom-models.html#custom-models-fine-tuning

2. AWS Documentation on Continued Pre-training: "Continued Pre-training customizes a foundation model by training it with a large corpus of unlabeled data... Use Continued Pre-training to adapt a model to a specific domain or industry."

Source: Amazon Web Services, "Custom models in Amazon Bedrock," Section: "Continued Pre-training." Retrieved from https://docs.aws.amazon.com/bedrock/latest/userguide/custom-models.html#custom-models-continued-pre-training

3. Academic Distinction between Fine-Tuning and Prompting: "While prompting can steer a model to a task, fine-tuning modifies the model's weights to specialize it for that task, typically yielding higher performance when a sufficient labeled dataset is available."

Source: Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Chapter 10, Section 10.4. Retrieved from https://web.stanford.edu/~jurafsky/slp3/10.pdf

Question: 20

A company's large language model (LLM) is experiencing hallucinations. How can the company decrease hallucinations?

A: Set up Agents for Amazon Bedrock to supervise the model training.

B: Use data pre-processing and remove any data that causes hallucinations.

C: Decrease the temperature inference parameter for the model.

D: Use a foundation model (FM) that is trained to not hallucinate.

Correct Answer:

C

Explanation:

The temperature is an inference parameter that controls the randomness of a large language model's (LLM) output. A higher temperature value increases randomness, encouraging more creative but potentially less accurate or nonsensical responses. Conversely, decreasing the temperature makes the model's output more deterministic and focused on the most probable tokens based on its training data. This reduces the likelihood of the model generating factually incorrect or fabricated information, thereby directly decreasing hallucinations. This is a standard technique for tuning model behavior at inference time to favor factuality over creativity.

Why Incorrect Options are Wrong:

A: Agents for Amazon Bedrock are used to execute tasks and orchestrate API calls at inference time; they do not supervise the training of foundation models.

B: Hallucinations are an emergent property of the model's generation process, not typically caused by specific, identifiable data points that can be simply removed during pre-processing.

D: No foundation model is completely free from hallucinations; it is an inherent challenge in current LLM technology. While some models are better, a model "trained to not hallucinate" does not exist.

References:

AWS Documentation - Inference parameters for foundation models: "Temperature: Modulates the probability distribution for the next token. ... Use a lower value to decrease

randomness in the response." This directly links lower temperature to less random (and thus less hallucinatory) output.

Source: AWS Documentation, Inference parameters for foundation models in Amazon Bedrock, Section: "Common inference parameters".

URL: https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters.html

AWS Documentation - Amazon Titan Text Models: "To encourage more factual and concise responses, decrease the Temperature value." This explicitly states the relationship between decreasing temperature and increasing factuality.

Source: AWS Documentation, Amazon Titan Text generation models, Section: "Inference parameters".

URL: https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-titan-text.html

Academic Publication - A Survey of Hallucination in Large Language Models: This survey discusses various mitigation strategies. Under "Inference-Time Mitigation," it notes that decoding strategies like adjusting temperature are used to control output quality. Lowering temperature is a form of greedy decoding which reduces the chance of generating low-probability, often nonsensical, sequences.

Source: Zhang, Y., et al. (2023). "A Survey of Hallucination in Large Language Models." ACM Computing Surveys.

URL: https://dl.acm.org/doi/10.1145/3626428 (See Section 4.2 on Decoding Strategies)

Question: 21

An AI practitioner has a database of animal photos. The AI practitioner wants to automatically identify and categorize the animals in the photos without manual human effort. Which strategy meets these requirements?

A: Object detection

B: Anomaly detection

C: Named entity recognition

D: Inpainting

Correct Answer:

A

Explanation:

Object detection is a computer vision technique that deals with identifying and locating objects within an image or video. This process involves two steps: localizing the object (e.g., by drawing a bounding box around it) and classifying the object into a specific category (e.g., "cat," "dog," "bird"). This directly fulfills the requirement to automatically identify and categorize animals in photos. AWS services like Amazon Rekognition use object detection models to perform this exact task, analyzing images to return labels for detected objects and their locations.

Why Incorrect Options are Wrong:

B: Anomaly detection: This technique identifies rare events or outliers that deviate from a norm. It would not categorize known animals but rather flag unusual images.

C: Named entity recognition: This is a Natural Language Processing (NLP) task for extracting entities like names and locations from text, not for analyzing images.

D: Inpainting: This is an image restoration technique used to fill in missing or corrupted parts of an image, not to identify the content within it.

References:

1. Amazon Web Services (AWS) Documentation: "Amazon Rekognition can identify thousands of objects, such as vehicles, pets, or furniture. It also detects scenes within an image, such as a sunset or a beach... When you analyze an image, Amazon Rekognition returns a list of objects and scenes that it detects."

Source: AWS Documentation, Amazon Rekognition, "Detecting objects and scenes in images."

2. Stanford University Courseware: "Object Detection: In this task, we want to classify and localize one or more objects in the image. We need to output a class label and a bounding box for each object in the image."

Source: Stanford University, CS231n: Deep Learning for Computer Vision, "Object Detection."

3. Amazon Web Services (AWS) Documentation: "Named entity recognition (NER) is a natural language processing (NLP) task that identifies named entities in a text and classifies them into predefined categories..."

Source: AWS Documentation, Amazon Comprehend, "Named entity recognition."

Question: 22

A company wants to create an application by using Amazon Bedrock. The company has a limited budget and prefers flexibility without long-term commitment. Which Amazon Bedrock pricing model meets these requirements?

A: On-Demand

B: Model customization

C: Provisioned Throughput

D: Spot Instance

Correct Answer:

A

Explanation:

The On-Demand pricing model for Amazon Bedrock is a pay-as-you-go approach. This model allows users to pay only for the inference they consume, measured in input and output tokens, without any upfront costs or time-based commitments. This structure provides maximum flexibility and is ideal for applications with unpredictable workloads or for companies that wish to avoid long-term financial commitments, directly addressing the requirements of a limited budget and the need for flexibility.

Why Incorrect Options are Wrong:

B: Model customization: This is a feature for fine-tuning models, not a primary pricing model for inference. It has associated costs but does not represent a plan for application usage.

C: Provisioned Throughput: This model requires a commitment to a specific term (one or six months) in exchange for a guaranteed throughput rate, which contradicts the requirement for no long-term commitment.

D: Spot Instance: Spot Instances are a pricing option for Amazon EC2 compute capacity and are not a pricing model available for Amazon Bedrock services.

References:

Amazon Web Services (AWS). (n.d.). Amazon Bedrock Pricing. Retrieved from https://aws.amazon.com/bedrock/pricing/

Section: Inference pricing: This page explicitly details the two pricing modes. It states, "With the On-Demand mode, you pay for what you use with no time-based term commitments." It describes Provisioned Throughput as requiring a "1-month or 6-month commitment term."

Amazon Web Services (AWS). (n.d.). What Is Amazon Bedrock?. AWS Documentation. Retrieved from https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html

Section: Provisioned Throughput: This section of the user guide explains that Provisioned Throughput is for "large, consistent inference workloads" and involves purchasing model units for a commitment term, reinforcing why it is unsuitable for the scenario.

Question: 23

Which AWS service or feature can help an AI development team quickly deploy and consume a foundation model (FM) within the team's VPC?

A: Amazon Personalize

B: Amazon SageMaker JumpStart

C: PartyRock, an Amazon Bedrock Playground

D: Amazon SageMaker endpoints

Correct Answer:

B

Explanation:

Amazon SageMaker JumpStart is a machine learning hub that accelerates the ML journey by providing access to a wide range of pre-trained models, including hundreds of foundation models (FMs). It allows development teams to select an FM and deploy it with just a few clicks onto a secure, scalable Amazon SageMaker endpoint. These endpoints can be configured to be private and accessible only from within the team's Amazon Virtual Private Cloud (VPC), directly addressing the need for quick deployment and secure consumption of FMs.

Why Incorrect Options are Wrong:

A: Amazon Personalize: This is a managed service specifically for building and deploying real-time recommendation systems, not for deploying general-purpose foundation models.

C: PartyRock, an Amazon Bedrock Playground: PartyRock is a public, web-based playground for experimenting with generative AI. It is not designed for deploying models into a private, enterprise VPC environment.

D: Amazon SageMaker endpoints: While an endpoint is the mechanism used to consume the deployed model, it is the result of a deployment. SageMaker JumpStart is the feature that facilitates the quick deployment of the foundation model to an endpoint.

References:

Amazon SageMaker JumpStart: AWS Documentation. "Amazon SageMaker JumpStart helps you quickly and easily get started with machine learning... You can access pre-trained

models, including foundation models, to solve your use cases." Retrieved from: https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart.html

Deploying Models with SageMaker JumpStart: AWS Documentation. "With SageMaker JumpStart, you can choose from a list of curated models and deploy them. When you deploy a model, SageMaker hosts the model in a secure and scalable environment and creates an endpoint that you can use to get inferences." Retrieved from: https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-deploy.html

SageMaker Endpoints and VPC: AWS Documentation. "To control access to your models, we recommend that you configure your SageMaker session to use a private VPC endpoint." This confirms that endpoints created (e.g., via JumpStart) can be secured within a VPRetrieved from: https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-vpc-endpoint.html

Question: 24

How can companies use large language models (LLMs) securely on Amazon Bedrock?

A: Design clear and specific prompts. Configure AWS Identity and Access Management (IAM) roles and policies by using least privilege access.

B: Enable AWS Audit Manager for automatic model evaluation jobs.

C: Enable Amazon Bedrock automatic model evaluation jobs.

D: Use Amazon CloudWatch Logs to make models explainable and to monitor for bias.

Correct Answer:

A

Explanation:

Securing large language models (LLMs) on Amazon Bedrock involves a multi-layered approach. The most critical and foundational layer is controlling access to the service itself, which is achieved using AWS Identity and Access Management (IAM). By configuring IAM roles and policies with the principle of least privilege, companies ensure that only authorized entities can invoke models or manage Bedrock resources. Additionally, designing clear and specific prompts (prompt engineering) is a key security practice at the application level to mitigate risks like prompt injection, where malicious inputs could cause the model to behave in unintended and insecure ways.

Why Incorrect Options are Wrong:

B: AWS Audit Manager is a compliance and audit service. It helps verify that your AWS usage meets certain standards but does not directly secure the use of LLMs.

C: Amazon Bedrock's automatic model evaluation is for assessing model performance (e.g., accuracy, toxicity). While related to responsible AI, it is not a primary security mechanism for access control.

D: Amazon CloudWatch Logs are for monitoring and logging API calls, which is useful for auditing and forensics. However, it is a detective control, not a preventative security measure like IAM.

References:

1. AWS Documentation - Security in Amazon Bedrock: This official guide explicitly states, "AWS Identity and Access Management (IAM) is a central component of security in Amazon

Bedrock. You use IAM to control who can access your Bedrock resources and what they can do with them." It details the use of identity-based policies for managing permissions.

URL: https://docs.aws.amazon.com/bedrock/latest/userguide/security.html

2. AWS Documentation - Guardrails for Amazon Bedrock: This documentation discusses how to implement safeguards, which relates to controlling model responses. This aligns with the concept of using well-designed prompts to guide model behavior securely. While not the same as prompt engineering, it highlights the importance of controlling model output for safety and security.

URL: https://docs.aws.amazon.com/bedrock/latest/userguide/guardrails.html

3. AWS Documentation - Model evaluation in Amazon Bedrock: This page describes the feature for evaluating, comparing, and selecting foundation models based on metrics. This confirms that its purpose is performance assessment, not security enforcement.

URL: https://docs.aws.amazon.com/bedrock/latest/userguide/model-evaluation.html

Question: 25

A company has terabytes of data in a database that the company can use for business analysis. The company wants to build an AI-based application that can build a SQL query from input text that employees provide. The employees have minimal experience with technology. Which solution meets these requirements?

A: Generative pre-trained transformers (GPT)

B: Residual neural network

C: Support vector machine

D: WaveNet

Correct Answer:

A

Explanation:

The requirement is to build an application that translates natural language text into SQL queries. This is a sequence-to-sequence task that falls under the domain of natural language generation and understanding. Generative Pre-trained Transformers (GPTs) are a class of large language models based on the transformer architecture, which excels at such tasks. They are pre-trained on vast datasets of text and code, enabling them to understand the semantics of a natural language request and generate syntactically correct SQL code as output. This makes them the most suitable solution for the described problem.

Why Incorrect Options are Wrong:

B: Residual neural network (ResNet): This architecture is primarily designed for computer vision tasks, such as image classification, and is not suited for natural language processing or code generation.

C: Support vector machine (SVM): SVMs are supervised learning models used for classification and regression. They cannot perform generative tasks like creating structured text or code from a prompt.

D: WaveNet: WaveNet is a deep generative model specifically designed for producing raw audio waveforms, most notably for text-to-speech synthesis, not for text-to-SQL translation.

References:

1. Vaswani, A., et al. (2017). "Attention Is All You Need." Advances in Neural Information Processing Systems 30. The paper introduces the Transformer architecture, the foundation for GPT models, demonstrating its effectiveness in machine translation, a task analogous to text-to-SQL. (URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

2. AWS Documentation. "What is generative AI?". This official AWS page explains that generative AI can create various types of content, including "code," which directly aligns with the question's requirement to generate SQL queries. (URL: https://aws.amazon.com/what-is/generative-ai/)

3. Stanford University. CS224N: NLP with Deep Learning, Winter 2023, Lecture 11: "Pretraining and Transformers". This course material details how transformer-based models like GPT are pre-trained and then fine-tuned for various downstream tasks, including generation. (URL: https://web.stanford.edu/class/cs224n/)

4. He, K., et al. (2016). "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. This paper establishes ResNets as a solution for image recognition, a different domain. (URL: https://ieeexplore.ieee.org/document/7780459)

Question: 26

A company built a deep learning model for object detection and deployed the model to production. Which AI process occurs when the model analyzes a new image to identify objects?

A: Training

B: Inference

C: Model deployment

D: Bias correction

Correct Answer:

B

Explanation:

Inference is the phase in the machine learning lifecycle where a trained model is used to make predictions on new, previously unseen data. The question describes a scenario where a deployed object detection model is actively analyzing a new image to identify objects. This act of generating a prediction (i.e., identifying objects) from new input is precisely the definition of inference. The training and deployment stages have already been completed.

Why Incorrect Options are Wrong:

A: Training: This is the process of teaching the model using a labeled dataset. The model in the scenario is already built, so this phase is complete.

C: Model deployment: This is the process of making the trained model available in a production environment. The scenario states this has already occurred.

D: Bias correction: This is a specific procedure to mitigate systematic errors in a model, not the general term for making a prediction on new data.

References:

1. Official AWS Documentation: Amazon SageMaker's documentation defines inference as the process of using a trained model to make predictions. It states, "After you deploy a model, you can use it to get inferences, which are predictions that the model makes."

Source: AWS Documentation, "What Is Amazon SageMaker?", Section: "Get inferences".

URL: https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

2. Academic Publication: In the foundational paper on deep learning, the authors distinguish between the learning (training) phase and the subsequent use of the model to classify or predict, which is termed inference.

Source: LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. (See section on "Supervised learning").

URL: https://www.nature.com/articles/nature14539

3. University Courseware: Stanford University's course on Convolutional Neural Networks for Visual Recognition explains that after a network is trained, it can be used for inference to classify new images.

Source: Stanford University, CS231n: Deep Learning for Computer Vision, "Module 1: Neural Networks".

URL: https://cs231n.github.io/

Question: 27

An AI practitioner is building a model to generate images of humans in various professions. The AI practitioner discovered that the input data is biased and that specific attributes affect the image generation and create bias in the model. Which technique will solve the problem?

A: Data augmentation for imbalanced classes

B: Model monitoring for class distribution

C: Retrieval Augmented Generation (RAG)

D: Watermark detection for images

Correct Answer:

A

Explanation:

The core problem described is biased input data leading to a biased generative model. This is often caused by an imbalance in the training dataset, where certain attributes (e.g., gender) are underrepresented within specific classes (e.g., professions). Data augmentation is a pre-processing technique used to solve this by creating new, synthetic data points for the underrepresented classes. By generating more varied examples (e.g., more images of women as doctors or men as nurses), data augmentation helps to balance the dataset, which in turn mitigates the bias in the final trained model.

Why Incorrect Options are Wrong:

B: Model monitoring for class distribution: This is a post-deployment technique to detect bias or data drift in a live model, but it does not solve the underlying problem in the training data itself.

C: Retrieval Augmented Generation (RAG): RAG is a technique primarily for language models to enhance text generation with external knowledge; it is not designed to fix class imbalance in image datasets.

D: Watermark detection for images: This technique is used to identify if an image was generated by an AI or to protect copyright. It is unrelated to training a model or correcting data bias.

References:

1. AWS SageMaker Developer Guide: Discusses handling imbalanced data, a common cause of bias. Techniques like data augmentation are standard pre-processing steps to address this. "If your training dataset is imbalanced, the trained model is likely to be biased... You can use data augmentation techniques to generate new data for the minority class." (Paraphrased from concepts in data preparation sections).

Source: Amazon SageMaker Developer Guide, "Handle Imbalanced Data", AWS Documentation. https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-mitigation-handle-imbalanced-data.html

2. Peer-Reviewed Publication (IEEE): Academic literature confirms data augmentation as a primary method for bias mitigation. "Data augmentation is a widely used strategy to mitigate data scarcity and imbalance issues... by generating synthetic data for minority groups, it can help debias the model."

Source: Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1-35. (This concept is widely covered in Section 4.1, "Pre-processing techniques").

3. University Courseware (Stanford): Machine learning courses frequently teach data augmentation as a solution for imbalanced datasets.

Source: Ng, (n.d.). CS229 Machine Learning Course Notes. Stanford University. The course materials cover data augmentation as a practical technique to improve model performance and robustness, especially when data is skewed or imbalanced.

Question: 28

A company wants to deploy a conversational chatbot to answer customer questions. The chatbot is based on a fine-tuned Amazon SageMaker JumpStart model. The application must comply with multiple regulatory frameworks. Which capabilities can the company show compliance for? (Choose two.)

A: Auto scaling inference endpoints

B: Threat detection

C: Data protection

D: Cost optimization

E: Loosely coupled microservices

Correct Answer:

B, C

Explanation:

Regulatory frameworks (e.g., GDPR, HIPAA, PCI DSS) mandate stringent controls for security and data privacy. Data protection (C), including encryption of data at rest and in transit and managing access controls, is a core requirement. Amazon SageMaker integrates with AWS Key Management Service (KMS) to meet these needs. Threat detection (B) is another critical compliance capability, involving monitoring for and responding to security threats. AWS provides services like Amazon GuardDuty and AWS CloudTrail logging for SageMaker API calls, which helps companies demonstrate that they have systems in place to detect and investigate potential security incidents, a common requirement in regulated industries.

Why Incorrect Options are Wrong:

A: Auto scaling inference endpoints: This is a performance and cost-management feature, not a primary capability for demonstrating regulatory compliance, which focuses on security and data governance.

D: Cost optimization: This is a financial and operational objective. Regulatory bodies are concerned with data security and privacy, not the cost-efficiency of the infrastructure.

E: Loosely coupled microservices: This is an architectural design pattern. While it can contribute to a secure posture, it is not a direct compliance capability reported to regulators.

References:

1. Data Protection: AWS Documentation, "Data Protection in Amazon SageMaker." This document details how SageMaker provides encryption at rest and in transit, stating, "Data protection refers to protecting data while in transit... and at rest... Amazon SageMaker provides the mechanisms for you to encrypt your data." This is a fundamental compliance requirement.

URL: https://docs.aws.amazon.com/sagemaker/latest/dg/data-protection.html

2. Threat Detection & Monitoring: AWS Documentation, "Security in Amazon SageMaker." This page explains how to use AWS CloudTrail to monitor API calls, which is essential for security analysis and threat detection. It states, "You can use the information collected by CloudTrail to determine the request that was made to SageMaker... and who made the request."

URL: https://docs.aws.amazon.com/sagemaker/latest/dg/security.html

3. Compliance Overview: AWS Documentation, "AWS Compliance Programs." This resource lists the various regulatory frameworks AWS complies with, which inherently rely on underlying capabilities like data protection and threat detection/monitoring to meet their standards.

URL: https://aws.amazon.com/compliance/programs/

Question: 29

A company is training a foundation model (FM). The company wants to increase the accuracy of the model up to a specific acceptance level. Which solution will meet these requirements?

A: Decrease the batch size.

B: Increase the epochs.

C: Decrease the epochs.

D: Increase the temperature parameter.

Correct Answer:

B

Explanation:

An epoch represents one complete pass of the entire training dataset through the learning algorithm. By increasing the number of epochs, the model is exposed to the training data more times, allowing it to learn the underlying patterns more thoroughly and adjust its internal parameters (weights) to minimize error. This iterative process is fundamental to improving a model's accuracy. Training is typically continued for more epochs until the desired accuracy on a validation set is achieved or until the model begins to overfit.

Why Incorrect Options are Wrong:

A: Decrease the batch size. While batch size is a critical hyperparameter, decreasing it primarily affects the stability and speed of convergence, not directly guaranteeing an increase in final accuracy.

C: Decrease the epochs. This would reduce the amount of training the model receives, likely resulting in underfitting and lower accuracy, which is the opposite of the stated goal.

D: Increase the temperature parameter. Temperature is a parameter used during the model's inference (generation) phase to control the randomness of its predictions, not during the training phase to improve its accuracy.

References:

1. Goodfellow, I., Bengio, Y., & Courville, (2016). Deep Learning. MIT Press. In Chapter 8, "Optimization for Training Deep Models," the text explains that training involves iterating

through the dataset multiple times (epochs) to minimize a cost function, which is directly related to improving model accuracy. (See Section 8.1.3, "Batches and Mini-Batches").

2. Stanford University CS231n Course Notes. The glossary defines an epoch as "a full pass over the entire dataset." The notes on the training process explain that iterating for multiple epochs is how the model learns.

URL: https://cs231n.github.io/neural-networks-3/#update

3. AWS SageMaker Developer Guide. The documentation on automatic model tuning explains that the number of epochs is a common hyperparameter to tune. The goal of tuning is to find the value that maximizes the objective metric (e.g., accuracy).

URL: https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-define-ranges.html