Amazon-Web-Services

Exam Questions AIF-C01

AWS Certified AI Practitioner

NEW QUESTION 1
A company wants to use a large language model (LLM) on Amazon Bedrock for sentiment analysis. The company wants to know how much information can fit into one prompt.
Which consideration will inform the company's decision?

A. Temperature
B. Context window
C. Batch size
D. Model size

Answer: B

Explanation:
The context window determines how much information can fit into a single prompt when using a large language model (LLM) like those on Amazon Bedrock.
? Context Window:
? Why Option B is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 2
A law firm wants to build an AI application by using large language models (LLMs). The application will read legal documents and extract key points from the documents. Which solution meets these requirements?

A. Build an automatic named entity recognition system.
B. Create a recommendation engine.
C. Develop a summarization chatbot.
D. Develop a multi-language translation system.

Answer: C

Explanation:
A summarization chatbot is ideal for extracting key points from legal documents. Large language models (LLMs) can be used to summarize complex texts, such as legal documents, making them more accessible and understandable.
? Option C (Correct): "Develop a summarization chatbot": This is the correct answer
because a summarization chatbot uses LLMs to condense and extract key information from text, which is precisely the requirement for reading and summarizing legal documents.
? Option A: "Build an automatic named entity recognition system" is incorrect
because it focuses on identifying specific entities, not summarizing documents.
? Option B: "Create a recommendation engine" is incorrect as it is used to suggest products or content, not summarize text.
? Option D: "Develop a multi-language translation system" is incorrect because translation is unrelated to summarizing text.
AWS AI Practitioner References:
? Using LLMs for Text Summarization on AWS: AWS supports developing summarization tools using its AI services, including Amazon Bedrock.

NEW QUESTION 3
An AI practitioner has built a deep learning model to classify the types of materials in images. The AI practitioner now wants to measure the model performance.
Which metric will help the AI practitioner evaluate the performance of the model?

A. Confusion matrix
B. Correlation matrix
C. R2 score
D. Mean squared error (MSE)

Answer: A

Explanation:
A confusion matrix is the correct metric for evaluating the performance of a classification model, such as the deep learning model built to classify types of materials in images.
? Confusion Matrix:
? Why Option A is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 4
A company uses a foundation model (FM) from Amazon Bedrock for an AI search tool. The company wants to fine-tune the model to be more accurate by using the company's data.
Which strategy will successfully fine-tune the model?

A. Provide labeled data with the prompt field and the completion field.
B. Prepare the training dataset by creating a .txt file that contains multiple lines in .csv format.
C. Purchase Provisioned Throughput for Amazon Bedrock.
D. Train the model on journals and textbooks.

Answer: A

Explanation:
Providing labeled data with both a prompt field and a completion field is the correct strategy for fine-tuning a foundation model (FM) on Amazon Bedrock.
? Fine-Tuning Strategy:
? Why Option A is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 5
A company wants to build an ML model by using Amazon SageMaker. The company needs to share and manage variables for model development across multiple teams.
Which SageMaker feature meets these requirements?

A. Amazon SageMaker Feature Store
B. Amazon SageMaker Data Wrangler
C. Amazon SageMaker Clarify
D. Amazon SageMaker Model Cards

Answer: A

Explanation:
Amazon SageMaker Feature Store is the correct solution for sharing and managing variables (features) across multiple teams during model development.
? Amazon SageMaker Feature Store:
? Why Option A is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 6
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

Answer: B

Explanation:
Amazon SageMaker JumpStart is the correct service for quickly deploying and consuming a foundation model (FM) within a team's VPC.
? Amazon SageMaker JumpStart:
? Why Option B is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 7
Which strategy evaluates the accuracy of a foundation model (FM) that is used in image classification tasks?

A. Calculate the total cost of resources used by the model.
B. Measure the model's accuracy against a predefined benchmark dataset.
C. Count the number of layers in the neural network.
D. Assess the color accuracy of images processed by the model.

Answer: B

Explanation:
Measuring the model's accuracy against a predefined benchmark dataset is the correct strategy to evaluate the accuracy of a foundation model (FM) used in image classification tasks.
? Model Accuracy Evaluation:
? Why Option B is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 8
What does an F1 score measure in the context of foundation model (FM) performance?

A. Model precision and recall.
B. Model speed in generating responses.
C. Financial cost of operating the model.
D. Energy efficiency of the model's computations.

Answer: A

Explanation:
The F1 score is the harmonic mean of precision and recall, making it a balanced metric for evaluating model performance when there is an imbalance between false positives and false negatives. Speed, cost, and energy efficiency are unrelated to the F1 score. References: AWS Foundation Models Guide.

NEW QUESTION 9
Which term describes the numerical representations of real-world objects and concepts that AI and natural language processing (NLP) models use to improve understanding of textual information?

A. Embeddings
B. Tokens
C. Models
D. Binaries

Answer: A

Explanation:
Embeddings are numerical representations of objects (such as words, sentences, or documents) that capture the objects' semantic meanings in a form that AI and

NLP models can easily understand. These representations help models improve their understanding of textual information by representing concepts in a continuous vector space.
? Option A (Correct): "Embeddings": This is the correct term, as embeddings provide
a way for models to learn relationships between different objects in their input space, improving their understanding and processing capabilities.
? Option B: "Tokens" are pieces of text used in processing, but they do not capture
semantic meanings like embeddings do.
? Option C: "Models" are the algorithms that use embeddings and other inputs, not the representations themselves.
? Option D: "Binaries" refer to data represented in binary form, which is unrelated to the concept of embeddings.
AWS AI Practitioner References:
? Understanding Embeddings in AI and NLP: AWS provides resources and tools, like Amazon SageMaker, that utilize embeddings to represent data in formats suitable for machine learning models.

NEW QUESTION 10
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

Answer: A

Explanation:
In generative AI models, such as those built on Amazon Bedrock, inference costs are driven by the number of tokens processed. A token can be as short as one character or as long as one word, and the more tokens consumed during the inference process, the higher the cost.
? Option A (Correct): "Number of tokens consumed": This is the correct answer
because the inference cost is directly related to the number of tokens processed by the model.
? Option B: "Temperature value" is incorrect as it affects the randomness of the
model's output but not the cost directly.
? Option C: "Amount of data used to train the LLM" is incorrect because training data size affects training costs, not inference costs.
? Option D: "Total training time" is incorrect because it relates to the cost of training the model, not the cost of inference.
AWS AI Practitioner References:
? Understanding Inference Costs on AWS: AWS documentation highlights that inference costs for generative models are largely based on the number of tokens processed.

NEW QUESTION 10
An education provider is building a question and answer application that uses a generative AI model to explain complex concepts. The education provider wants to automatically change the style of the model response depending on who is asking the question. The education provider will give the model the age range of the user who has asked the question.
Which solution meets these requirements with the LEAST implementation effort?

A. Fine-tune the model by using additional training data that is representative of the various age ranges that the application will support.
B. Add a role description to the prompt context that instructs the model of the age range that the response should target.
C. Use chain-of-thought reasoning to deduce the correct style and complexity for a response suitable for that user.
D. Summarize the response text depending on the age of the user so that younger users receive shorter responses.

Answer: B

Explanation:
Adding a role description to the prompt context is a straightforward way to instruct the generative AI model to adjust its response style based on the user's age range. This method requires minimal implementation effort as it does not involve additional training or complex logic.
? Option B (Correct): "Add a role description to the prompt context that instructs the model of the age range that the response should target": This is the correct answer because it involves the least implementation effort while effectively guiding the
model to tailor responses according to the age range.
? Option A: "Fine-tune the model by using additional training data" is incorrect because it requires significant effort in gathering data and retraining the model.
? Option C: "Use chain-of-thought reasoning" is incorrect as it involves complex reasoning that may not directly address the need to adjust response style based on age.
? Option D: "Summarize the response text depending on the age of the user" is incorrect because it involves additional processing steps after generating the initial response, increasing complexity.
AWS AI Practitioner References:
? Prompt Engineering Techniques on AWS: AWS recommends using prompt context effectively to guide generative models in providing tailored responses based on specific user attributes.

NEW QUESTION 12
A company is using an Amazon Bedrock base model to summarize documents for an internal use case. The company trained a custom model to improve the summarization quality.
Which action must the company take to use the custom model through Amazon Bedrock?

A. Purchase Provisioned Throughput for the custom model.
B. Deploy the custom model in an Amazon SageMaker endpoint for real-time inference.
C. Register the model with the Amazon SageMaker Model Registry.
D. Grant access to the custom model in Amazon Bedrock.

Answer: B

Explanation:
To use a custom model that has been trained to improve summarization quality, the company must deploy the model on an Amazon SageMaker endpoint. This allows the model to be used for real-time inference through Amazon Bedrock or other AWS services. By deploying the model in SageMaker, the custom model can

be accessed programmatically via API calls, enabling integration with Amazon Bedrock.
? Option B (Correct): "Deploy the custom model in an Amazon SageMaker endpoint
for real-time inference": This is the correct answer because deploying the model on SageMaker enables it to serve real-time predictions and be integrated with Amazon Bedrock.
? Option A: "Purchase Provisioned Throughput for the custom model" is incorrect
because provisioned throughput is related to database or storage services, not model deployment.
? Option C: "Register the model with the Amazon SageMaker Model Registry" is
incorrect because while the model registry helps with model management, it does not make the model accessible for real-time inference.
? Option D: "Grant access to the custom model in Amazon Bedrock" is incorrect
because Bedrock does not directly manage custom model access; it relies on deployed endpoints like those in SageMaker.
AWS AI Practitioner References:
? Amazon SageMaker Endpoints: AWS recommends deploying models to SageMaker endpoints to use them for real-time inference in various applications.

NEW QUESTION 17
A company is using the Generative AI Security Scoping Matrix to assess security responsibilities for its solutions. The company has identified four different solution scopes based on the matrix.
Which solution scope gives the company the MOST ownership of security responsibilities?

A. Using a third-party enterprise application that has embedded generative AI features.
B. Building an application by using an existing third-party generative AI foundation model (FM).
C. Refining an existing third-party generative AI foundation model (FM) by fine-tuning the model by using data specific to the business.
D. Building and training a generative AI model from scratch by using specific data that a customer owns.

Answer: D

Explanation:
Building and training a generative AI model from scratch provides the company with the most ownership and control over security responsibilities. In this scenario, the company is responsible for all aspects of the security of the data, the model, and the infrastructure.
? Option D (Correct): "Building and training a generative AI model from scratch by
using specific data that a customer owns": This is the correct answer because it involves complete ownership of the model, data, and infrastructure, giving the company the highest level of responsibility for security.
? Option A: "Using a third-party enterprise application that has embedded generative
AI features" is incorrect as the company has minimal control over the security of the AI features embedded within a third-party application.
? Option B: "Building an application using an existing third-party generative AI
foundation model (FM)" is incorrect because security responsibilities are shared with the third-party model provider.
? Option C: "Refining an existing third-party generative AI FM by fine-tuning the
model with business-specific data" is incorrect as the foundation model and part of
the security responsibilities are still managed by the third party.
AWS AI Practitioner References:
? Generative AI Security Scoping Matrix on AWS: AWS provides a security responsibility matrix that outlines varying levels of control and responsibility depending on the approach to developing and using AI models.

NEW QUESTION 19
Which AWS feature records details about ML instance data for governance and reporting?

A. Amazon SageMaker Model Cards
B. Amazon SageMaker Debugger
C. Amazon SageMaker Model Monitor
D. Amazon SageMaker JumpStart

Answer: A

Explanation:
Amazon SageMaker Model Cards provide a centralized and standardized repository for documenting machine learning models. They capture key details such as the model's intended use, training and evaluation datasets, performance metrics, ethical considerations, and other relevant information. This documentation facilitates governance and reporting by ensuring that all stakeholders have access to consistent and comprehensive information about each model. While Amazon SageMaker Debugger is used for real-time debugging and monitoring during training, and Amazon SageMaker Model Monitor tracks deployed models for data and prediction quality, neither offers the comprehensive documentation capabilities of Model Cards. Amazon SageMaker JumpStart provides pre-built models and solutions but does not focus on governance documentation.
Reference: Amazon SageMaker Model Cards

NEW QUESTION 21
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

Answer: B

Explanation:
Inference is the correct answer because it is the AI process that occurs when a deployed model analyzes new data (such as an image) to make predictions or identify objects.
? Inference:
? Why Option B is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 26
A company deployed an AI/ML solution to help customer service agents respond to frequently asked questions. The questions can change over time. The company wants to give customer service agents the ability to ask questions and receive automatically generated answers to common customer questions. Which strategy will meet these requirements MOST cost-effectively?

A. Fine-tune the model regularly.
B. Train the model by using context data.
C. Pre-train and benchmark the model by using context data.
D. Use Retrieval Augmented Generation (RAG) with prompt engineering techniques.

Answer: D

Explanation:
RAG combines large pre-trained models with retrieval mechanisms to fetch relevant context from a knowledge base. This approach is cost-effective as it eliminates the need for frequent model retraining while ensuring responses are contextually accurate and up to date. References: AWS RAG Techniques.

NEW QUESTION 29
A company has thousands of customer support interactions per day and wants to analyze these interactions to identify frequently asked questions and develop insights.
Which AWS service can the company use to meet this requirement?

A. Amazon Lex
B. Amazon Comprehend
C. Amazon Transcribe
D. Amazon Translate

Answer: B

Explanation:
Amazon Comprehend is the correct service to analyze customer support interactions and identify frequently asked questions and insights.
? Amazon Comprehend:
? Why Option B is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 33
A company wants to use a large language model (LLM) to develop a conversational agent. The company needs to prevent the LLM from being manipulated with common prompt engineering techniques to perform undesirable actions or expose sensitive information.
Which action will reduce these risks?

A. Create a prompt template that teaches the LLM to detect attack patterns.
B. Increase the temperature parameter on invocation requests to the LLM.
C. Avoid using LLMs that are not listed in Amazon SageMaker.
D. Decrease the number of input tokens on invocations of the LLM.

Answer: A

Explanation:
Creating a prompt template that teaches the LLM to detect attack patterns is the most effective way to reduce the risk of the model being manipulated through prompt engineering.
? Prompt Templates for Security:
? Why Option A is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 36
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.

Answer: C

Explanation:
To manage the flow of data from Amazon S3 to SageMaker Studio notebooks securely, using a VPC with an S3 endpoint is the best solution.
? Amazon SageMaker and S3 Integration:
? Why Option C is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 40
Which metric measures the runtime efficiency of operating AI models?

A. Customer satisfaction score (CSAT)
B. Training time for each epoch
C. Average response time
D. Number of training instances

Answer: C

Explanation:
The average response time is the correct metric for measuring the runtime efficiency of operating AI models.
? Average Response Time:
? Why Option C is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 42
An accounting firm wants to implement a large language model (LLM) to automate document processing. The firm must proceed responsibly to avoid potential harms.
What should the firm do when developing and deploying the LLM? (Select TWO.)

A. Include fairness metrics for model evaluation.
B. Adjust the temperature parameter of the model.
C. Modify the training data to mitigate bias.
D. Avoid overfitting on the training data.
E. Apply prompt engineering techniques.

Answer: AC

Explanation:
To implement a large language model (LLM) responsibly, the firm should focus on fairness and mitigating bias, which are critical for ethical AI deployment.
? A. Include Fairness Metrics for Model Evaluation:
? C. Modify the Training Data to Mitigate Bias:
? Why Other Options are Incorrect:

NEW QUESTION 44
A company is building a customer service chatbot. The company wants the chatbot to improve its responses by learning from past interactions and online resources.
Which AI learning strategy provides this self-improvement capability?

A. Supervised learning with a manually curated dataset of good responses and bad responses
B. Reinforcement learning with rewards for positive customer feedback
C. Unsupervised learning to find clusters of similar customer inquiries
D. Supervised learning with a continuously updated FAQ database

Answer: B

Explanation:
Reinforcement learning allows a model to learn and improve over time based on feedback from its environment. In this case, the chatbot can improve its responses by being rewarded for positive customer feedback, which aligns well with the goal of self- improvement based on past interactions and new information.
? Option B (Correct): "Reinforcement learning with rewards for positive customer
feedback": This is the correct answer as reinforcement learning enables the chatbot to learn from feedback and adapt its behavior accordingly, providing self-improvement capabilities.
? Option A: "Supervised learning with a manually curated dataset" is incorrect
because it does not support continuous learning from new interactions.
? Option C: "Unsupervised learning to find clusters of similar customer inquiries" is incorrect because unsupervised learning does not provide a mechanism for improving responses based on feedback.
? Option D: "Supervised learning with a continuously updated FAQ database" is incorrect because it still relies on manually curated data rather than self-improvement from feedback.
AWS AI Practitioner References:
? Reinforcement Learning on AWS: AWS provides reinforcement learning
frameworks that can be used to train models to improve their performance based on feedback.

NEW QUESTION 46
A company has built an image classification model to predict plant diseases from photos of plant leaves. The company wants to evaluate how many images the model classified correctly.
Which evaluation metric should the company use to measure the model's performance?

A. R-squared score
B. Accuracy
C. Root mean squared error (RMSE)
D. Learning rate

Answer: B

Explanation:
Accuracy is the most appropriate metric to measure the performance of an image classification model. It indicates the percentage of correctly classified images out of the total number of images. In the context of classifying plant diseases from images, accuracy will help the company determine how well the model is performing by showing how many images were correctly classified.
? Option B (Correct): "Accuracy": This is the correct answer because accuracy
measures the proportion of correct predictions made by the model, which is suitable for evaluating the performance of a classification model.
? Option A: "R-squared score" is incorrect as it is used for regression analysis, not
classification tasks.
? Option C: "Root mean squared error (RMSE)" is incorrect because it is also used for regression tasks to measure prediction errors, not for classification accuracy.
? Option D: "Learning rate" is incorrect as it is a hyperparameter for training, not a performance metric.
AWS AI Practitioner References:
? Evaluating Machine Learning Models on AWS: AWS documentation emphasizes the use of appropriate metrics, like accuracy, for classification tasks.

NEW QUESTION 48
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.

Answer: A

Explanation:
Tokens in generative AI models are the smallest units that the model processes, typically representing words, subwords, or characters. They are essential for the model to understand and generate language, breaking down text into manageable parts for processing.
? Option A (Correct): "Tokens are the basic units of input and output that a
generative AI model operates on, representing words, subwords, or other linguistic units": This is the correct definition of tokens in the context of generative AI models.
? Option B: "Mathematical representations of words" describes embeddings, not
tokens.
? Option C: "Pre-trained weights of a model" refers to the parameters of a model, not tokens.
? Option D: "Prompts or instructions given to a model" refers to the queries or commands provided to a model, not tokens.
AWS AI Practitioner References:
? Understanding Tokens in NLP: AWS provides detailed explanations of how tokens are used in natural language processing tasks by AI models, such as in Amazon Comprehend and other AWS AI services.

NEW QUESTION 53
A company wants to use a large language model (LLM) on Amazon Bedrock for sentiment analysis. The company needs the LLM to produce more consistent responses to the same input prompt.
Which adjustment to an inference parameter should the company make to meet these requirements?

A. Decrease the temperature value
B. Increase the temperature value
C. Decrease the length of output tokens
D. Increase the maximum generation length

Answer: A

Explanation:
The temperature parameter in a large language model (LLM) controls the randomness of the model's output. A lower temperature value makes the output more deterministic and consistent, meaning that the model is less likely to produce different results for the same input prompt.
? Option A (Correct): "Decrease the temperature value": This is the correct answer
because lowering the temperature reduces the randomness of the responses, leading to more consistent outputs for the same input.
? Option B: "Increase the temperature value" is incorrect because it would make the
output more random and less consistent.
? Option C: "Decrease the length of output tokens" is incorrect as it does not directly affect the consistency of the responses.
? Option D: "Increase the maximum generation length" is incorrect because this adjustment affects the output length, not the consistency of the model??s responses.
AWS AI Practitioner References:
? Understanding Temperature in Generative AI Models: AWS documentation explains that adjusting the temperature parameter affects the model??s output randomness, with lower values providing more consistent outputs.

NEW QUESTION 58
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 dat
E. Use Amazon Rekognition to add custom labels to the data.

Answer: B

Explanation:
Amazon SageMaker Clarify provides transparency and explainability for machine learning models by generating metrics, reports, and examples that help to understand model predictions. For a medical company that needs a foundation model to be transparent and explainable to meet regulatory requirements, SageMaker Clarify is the most suitable solution.
? Amazon SageMaker Clarify:
? Why Option B is Correct:
? Why Other Options are Incorrect:
Thus, B is the correct answer for meeting transparency and explainability requirements for the foundation model

NEW QUESTION 59
A company wants to use a large language model (LLM) on Amazon Bedrock for sentiment analysis. The company wants to classify the sentiment of text passages as positive or negative.
Which prompt engineering strategy meets these requirements?

A. Provide examples of text passages with corresponding positive or negative labels in the prompt followed by the new text passage to be classified.
B. Provide a detailed explanation of sentiment analysis and how LLMs work in the prompt.

C. Provide the new text passage to be classified without any additional context or examples.
D. Provide the new text passage with a few examples of unrelated tasks, such as text summarization or question answering.

Answer: A

Explanation:
Providing examples of text passages with corresponding positive or negative labels in the prompt followed by the new text passage to be classified is the correct prompt engineering strategy for using a large language model (LLM) on Amazon Bedrock for sentiment analysis.
? Example-Driven Prompts:
? Why Option A is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 63
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)

Answer: B

Explanation:
Unsupervised learning is the correct methodology for classifying customers into tiers when the data is unlabeled, as it does not require predefined labels or outputs.
? Unsupervised Learning:
? Why Option B is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 64
A company manually reviews all submitted resumes in PDF format. As the company grows, the company expects the volume of resumes to exceed the company's review capacity. The company needs an automated system to convert the PDF resumes into plain text format for additional processing.
Which AWS service meets this requirement?

A. Amazon Textract
B. Amazon Personalize
C. Amazon Lex
D. Amazon Transcribe

Answer: A

Explanation:
Amazon Textract is a service that automatically extracts text and data from scanned documents, including PDFs. It is the best choice for converting resumes from PDF format to plain text for further processing.
? Amazon Textract:
? Why Option A is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 68
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

Answer: B

Explanation:
In the context of AI, "hallucination" refers to the phenomenon where a model generates outputs that are plausible-sounding but are not grounded in reality or the training data. This problem often occurs with large language models (LLMs) when they create information that sounds correct but is actually incorrect or fabricated.
? Option B (Correct): "Hallucination": This is the correct answer because the
problem described involves generating content that sounds factual but is incorrect, which is characteristic of hallucination in generative AI models.
? Option A: "Data leakage" is incorrect as it involves the model accidentally learning
from data it shouldn't have access to, which does not match the problem of generating incorrect content.
? Option C: "Overfitting" is incorrect because overfitting refers to a model that has
learned the training data too well, including noise, and performs poorly on new data.
? Option D: "Underfitting" is incorrect because underfitting occurs when a model is
too simple to capture the underlying patterns in the data, which is not the issue here.
AWS AI Practitioner References:
? Large Language Models on AWS: AWS discusses the challenge of hallucination in large language models and emphasizes techniques to mitigate it, such as using guardrails and fine-tuning.

NEW QUESTION 72
A company has a database of petabytes of unstructured data from internal sources. The company wants to transform this data into a structured format so that its

data scientists can perform machine learning (ML) tasks.
Which service will meet these requirements?

A. Amazon Lex
B. Amazon Rekognition
C. Amazon Kinesis Data Streams
D. AWS Glue

Answer: D

Explanation:
AWS Glue is the correct service for transforming petabytes of unstructured data into a structured format suitable for machine learning tasks.
? AWS Glue:
? Why Option D is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 75
A company needs to build its own large language model (LLM) based on only the company's private data. The company is concerned about the environmental effect of the training process.
Which Amazon EC2 instance type has the LEAST environmental effect when training LLMs?

A. Amazon EC2 C series
B. Amazon EC2 G series
C. Amazon EC2 P series
D. Amazon EC2 Trn series

Answer: D

Explanation:
The Amazon EC2 Trn series (Trainium) instances are designed for high-performance, cost- effective machine learning training while being energy-efficient. AWS Trainium-powered instances are optimized for deep learning models and have been developed to minimize environmental impact by maximizing energy efficiency.
? Option D (Correct): "Amazon EC2 Trn series": This is the correct answer because the Trn series is purpose-built for training deep learning models with lower energy consumption, which aligns with the company's concern about environmental effects.
? Option A: "Amazon EC2 C series" is incorrect because it is intended for compute-
intensive tasks but not specifically optimized for ML training with environmental considerations.
? Option B: "Amazon EC2 G series" (Graphics Processing Unit instances) is
optimized for graphics-intensive applications but does not focus on minimizing environmental impact for training.
? Option C: "Amazon EC2 P series" is designed for ML training but does not offer
the same level of energy efficiency as the Trn series.
AWS AI Practitioner References:
? AWS Trainium Overview: AWS promotes Trainium instances as their most energy- efficient and cost-effective solution for ML model training.

NEW QUESTION 78
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

Answer: C

Explanation:
Amazon OpenSearch Service (formerly Amazon Elasticsearch Service) has introduced capabilities to support vector search, which allows companies to build vector database applications. This is particularly useful in machine learning, where vector representations (embeddings) of data are often used to capture semantic meaning.
Scalable index management and nearest neighbor search capability are the core features enabling vector database functionalities in OpenSearch. The service allows users to index high-dimensional vectors and perform efficient nearest neighbor searches, which are crucial for tasks such as recommendation systems, anomaly detection, and semantic search.
Here is why option C is the correct Answer:
? Scalable Index Management: OpenSearch Service supports scalable indexing of vector data. This means you can index a large volume of high-dimensional vectors
and manage these indexes in a cost-effective and performance-optimized way. The service leverages underlying AWS infrastructure to ensure that indexing scales seamlessly with data size.
? Nearest Neighbor Search Capability: OpenSearch Service's nearest neighbor
search capability allows for fast and efficient searches over vector data. This is essential for applications like product recommendation engines, where the system needs to quickly find the most similar items based on a user's query or behavior.
? AWS AI Practitioner References:
The other options do not directly relate to building vector database applications:
? A. Integration with Amazon S3 for object storage is about storing data objects, not vector-based searching or indexing.
? B. Support for geospatial indexing and queries is related to location-based data, not vectors used in machine learning.
? D. Ability to perform real-time analysis on streaming data relates to analyzing incoming data streams, which is different from the vector search capabilities.

NEW QUESTION 82
Which functionality does Amazon SageMaker Clarify provide?

A. Integrates a Retrieval Augmented Generation (RAG) workflow
B. Monitors the quality of ML models in production
C. Documents critical details about ML models
D. Identifies potential bias during data preparation

Answer: D

Explanation:
Exploratory data analysis (EDA) involves understanding the data by visualizing it, calculating statistics, and creating correlation matrices. This stage helps identify patterns, relationships, and anomalies in the data, which can guide further steps in the ML pipeline.
? Option C (Correct): "Exploratory data analysis": This is the correct answer as the
tasks described (correlation matrix, calculating statistics, visualizing data) are all part of the EDA process.
? Option A: "Data pre-processing" is incorrect because it involves cleaning and
transforming data, not initial analysis.
? Option B: "Feature engineering" is incorrect because it involves creating new features from raw data, not analyzing the data's existing structure.
? Option D: "Hyperparameter tuning" is incorrect because it refers to optimizing model parameters, not analyzing the data.
AWS AI Practitioner References:
? Stages of the Machine Learning Pipeline: AWS outlines EDA as the initial phase of understanding and exploring data before moving to more specific preprocessing, feature engineering, and model training stages.

NEW QUESTION 87
A company uses Amazon SageMaker for its ML pipeline in a production environment. The company has large input data sizes up to 1 GB and processing times up to 1 hour. The company needs near real-time latency.
Which SageMaker inference option meets these requirements?

A. Real-time inference
B. Serverless inference
C. Asynchronous inference
D. Batch transform

Answer: A

Explanation:
Real-time inference is designed to provide immediate, low-latency predictions, which is necessary when the company requires near real-time latency for its ML models. This option is optimal when there is a need for fast responses, even with large input data sizes and substantial processing times.
? Option A (Correct): "Real-time inference": This is the correct answer because it
supports low-latency requirements, which are essential for real-time applications where quick response times are needed.
? Option B: "Serverless inference" is incorrect because it is more suited for intermittent, small-scale inference workloads, not for continuous, large-scale, low-latency needs.
? Option C: "Asynchronous inference" is incorrect because it is used for workloads that do not require immediate responses.
? Option D: "Batch transform" is incorrect as it is intended for offline, large-batch processing where immediate response is not necessary.
AWS AI Practitioner References:
? Amazon SageMaker Inference Options: AWS documentation describes real-time inference as the best solution for applications that require immediate prediction results with low latency.

NEW QUESTION 88
A financial institution is using Amazon Bedrock to develop an AI application. The application is hosted in a VPC. To meet regulatory compliance standards, the VPC is not allowed access to any internet traffic.
Which AWS service or feature will meet these requirements?

A. AWS PrivateLink
B. Amazon Macie
C. Amazon CloudFront
D. Internet gateway

Answer: A

Explanation:
AWS PrivateLink enables private connectivity between VPCs and AWS services without exposing traffic to the public internet. This feature is critical for meeting regulatory compliance standards that require isolation from public internet traffic.
? Option A (Correct): "AWS PrivateLink": This is the correct answer because it
allows secure access to Amazon Bedrock and other AWS services from a VPC without internet access, ensuring compliance with regulatory standards.
? Option B: "Amazon Macie" is incorrect because it is a security service for data
classification and protection, not for managing private network traffic.
? Option C: "Amazon CloudFront" is incorrect because it is a content delivery network service and does not provide private network connectivity.
? Option D: "Internet gateway" is incorrect as it enables internet access, which violates the VPC's no-internet-traffic policy.
AWS AI Practitioner References:
? AWS PrivateLink Documentation: AWS highlights PrivateLink as a solution for connecting VPCs to AWS services privately, which is essential for organizations with strict regulatory requirements.

NEW QUESTION 93
A company is building an ML model to analyze archived data. The company must perform inference on large datasets that are multiple GBs in size. The company does not need to access the model predictions immediately.
Which Amazon SageMaker inference option will meet these requirements?

A. Batch transform
B. Real-time inference
C. Serverless inference
D. Asynchronous inference

Answer: A

Explanation:
Batch transform in Amazon SageMaker is designed for offline processing of large datasets. It is ideal for scenarios where immediate predictions are not required, and the inference can be done on large datasets that are multiple gigabytes in size. This method processes data in batches, making it suitable for analyzing

archived data without the need for real- time access to predictions.
? Option A (Correct): "Batch transform": This is the correct answer because batch
transform is optimized for handling large datasets and is suitable when immediate access to predictions is not required.
? Option B: "Real-time inference" is incorrect because it is used for low-latency, real-
time prediction needs, which is not required in this case.
? Option C: "Serverless inference" is incorrect because it is designed for small-scale, intermittent inference requests, not for large batch processing.
? Option D: "Asynchronous inference" is incorrect because it is used when immediate predictions are required, but with high throughput, whereas batch transform is more suitable for very large datasets.
AWS AI Practitioner References:
? Batch Transform on AWS SageMaker: AWS recommends using batch transform for large datasets when real-time processing is not needed, ensuring cost-effectiveness and scalability.

NEW QUESTION 95
A company wants to use AI to protect its application from threats. The AI solution needs to check if an IP address is from a suspicious source.
Which solution meets these requirements?

A. Build a speech recognition system.
B. Create a natural language processing (NLP) named entity recognition system.
C. Develop an anomaly detection system.
D. Create a fraud forecasting system.

Answer: C

Explanation:
An anomaly detection system is suitable for identifying unusual patterns or behaviors, such as suspicious IP addresses, which might indicate a potential threat.
? Anomaly Detection:
? Why Option C is Correct:
? Why Other Options are Incorrect:
Thus, C is the correct answer for detecting suspicious IP addresses.

NEW QUESTION 99
A company is building a contact center application and wants to gain insights from customer conversations. The company wants to analyze and extract key information from the audio of the customer calls.
Which solution meets these requirements?

A. Build a conversational chatbot by using Amazon Lex.
B. Transcribe call recordings by using Amazon Transcribe.
C. Extract information from call recordings by using Amazon SageMaker Model Monitor.
D. Create classification labels by using Amazon Comprehend.

Answer: B

Explanation:
Amazon Transcribe is the correct solution for converting audio from customer calls into text, allowing the company to analyze and extract key information from the conversations.
? Amazon Transcribe:
? Why Option B is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 101
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.

Answer: B

Explanation:
Amazon Bedrock provides large language models (LLMs) that are optimized for natural language understanding and text summarization tasks, making it the best choice for creating concise summaries of user reviews. Time-series forecasting, classification, and image analysis (Rekognition) are not suitable for summarizing textual data. References: AWS Bedrock Documentation.

NEW QUESTION 106
A company is building an application that needs to generate synthetic data that is based on existing data.
Which type of model can the company use to meet this requirement?

A. Generative adversarial network (GAN)
B. XGBoost
C. Residual neural network
D. WaveNet

Answer: A

Explanation:
Generative adversarial networks (GANs) are a type of deep learning model used for generating synthetic data based on existing datasets. GANs consist of two neural networks (a generator and a discriminator) that work together to create realistic data.

? Option A (Correct): "Generative adversarial network (GAN)": This is the correct answer because GANs are specifically designed for generating synthetic data that closely resembles the real data they are trained on.
? Option B: "XGBoost" is a gradient boosting algorithm for classification and regression tasks, not for generating synthetic data.
? Option C: "Residual neural network" is primarily used for improving the performance of deep networks, not for generating synthetic data.
? Option D: "WaveNet" is a model architecture designed for generating raw audio waveforms, not synthetic data in general.
AWS AI Practitioner References:
? GANs on AWS for Synthetic Data Generation: AWS supports the use of GANs for creating synthetic datasets, which can be crucial for applications like training machine learning models in environments where real data is scarce or sensitive.

NEW QUESTION 111
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

Answer: A

Explanation:
Data augmentation for imbalanced classes is the correct technique to address bias in input data affecting image generation.
? Data Augmentation for Imbalanced Classes:
? Why Option A is Correct:
? Why Other Options are Incorrect:

NEW QUESTION 113
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

Answer: B

Explanation:
Partial dependence plots (PDPs) are visual tools used to show the relationship between a feature (or a set of features) in the data and the predicted outcome of a machine learning model. They are highly effective for providing transparency and explainability of the model's behavior to stakeholders by illustrating how different input variables impact the model's predictions.
? Option B (Correct): "Partial dependence plots (PDPs)": This is the correct answer because PDPs help to interpret how the model's predictions change with varying values of input features, providing stakeholders with a clearer understanding of the model's decision-making process.
? Option A: "Code for model training" is incorrect because providing the raw code for model training may not offer transparency or explainability to non-technical stakeholders.
? Option C: "Sample data for training" is incorrect as sample data alone does not explain how the model works or its decision-making process.
? Option D: "Model convergence tables" is incorrect. While convergence tables can show the training process, they do not provide insights into how input features affect the model's predictions.
AWS AI Practitioner References:
? Explainability in AWS Machine Learning: AWS provides various tools for model explainability, such as Amazon SageMaker Clarify, which includes PDPs to help explain the impact of different features on the model??s predictions.

NEW QUESTION 115
A company wants to create a chatbot by using a foundation model (FM) on Amazon Bedrock. The FM needs to access encrypted data that is stored in an Amazon S3 bucket.
The data is encrypted with Amazon S3 managed keys (SSE-S3).
The FM encounters a failure when attempting to access the S3 bucket data. Which solution will meet these requirements?

A. Ensure that the role that Amazon Bedrock assumes has permission to decrypt data withthe correct encryption key.
B. Set the access permissions for the S3 buckets to allow public access to enable access over the internet.
C. Use prompt engineering techniques to tell the model to look for information in Amazon S3.
D. Ensure that the S3 data does not contain sensitive information.

Answer: A

Explanation:
Amazon Bedrock needs the appropriate IAM role with permission to access and decrypt data stored in Amazon S3. If the data is encrypted with Amazon S3 managed keys (SSE- S3), the role that Amazon Bedrock assumes must have the required permissions to access and decrypt the encrypted data.
? Option A (Correct): "Ensure that the role that Amazon Bedrock assumes has permission to decrypt data with the correct encryption key": This is the correct solution as it ensures that the AI model can access the encrypted data securely without changing the encryption settings or compromising data security.
? Option B: "Set the access permissions for the S3 buckets to allow public access" is incorrect because it violates security best practices by exposing sensitive data to the public.
? Option C: "Use prompt engineering techniques to tell the model to look for information in Amazon S3" is incorrect as it does not address the encryption and permission issue.
? Option D: "Ensure that the S3 data does not contain sensitive information" is incorrect because it does not solve the access problem related to encryption.
AWS AI Practitioner References:

? Managing Access to Encrypted Data in AWS: AWS recommends using proper IAM roles and policies to control access to encrypted data stored in S3.

NEW QUESTION 118

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

AIF-C01 Practice Exam Features:

* AIF-C01 Questions and Answers Updated Frequently

* AIF-C01 Practice Questions Verified by Expert Senior Certified Staff

* AIF-C01 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* AIF-C01 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year