AIF-C01

AWS Certified AI Practitioner

Exam A

QUESTION 1
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
Section: (none)

Explanation/Reference:
Explanation:
Partial Dependence Plots (PDPs) are a powerful tool for understanding and explaining how the features in a machine learning model impact predictions. They are often used to meet transparency and explainability requirements for stakeholders. Let's go over why this is the correct choice, along with why the other options are less suitable:
Partial Dependence Plots (PDPs)
Purpose: PDPs show the relationship between a feature (or multiple features) and the model's predicted output, which helps to explain the effect of each feature on the model's predictions.
Explainability: By visualizing how each feature influences the prediction, stakeholders can better understand how the model works and why it makes certain predictions. This level of interpretability is essential for gaining trust from non-technical stakeholders.
Transparency: PDPs improve transparency by providing an intuitive way to analyze and present the effects of individual features.

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.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
A summarization chatbot can effectively read legal documents and generate concise versions that highlight key points. This directly addresses the requirement of extracting essential information, unlike the other options, which focus on different tasks.

QUESTION 3
A company wants to classify human genes into 20 categories based on gene characteristics. The company needs an ML algorithm to document how the inner mechanism of the model affects the output.

Which ML algorithm meets these requirements?

A. Decision trees

B. Linear regression
C. Logistic regression
D. Neural networks

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Decision trees provide clear transparency into how the model makes decisions, allowing easy documentation of how the inner mechanism influences the output. The other options do not offer this level of interpretability.

QUESTION 4
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

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Accuracy measures how many images were correctly classified out of the total images, making it the appropriate metric for evaluating the performance of an image classification model. The other metrics are either not suitable for classification tasks or not used for performance evaluation.

QUESTION 5
A company is using a pre-trained large language model (LLM) to build a chatbot for product recommendations. The company needs the LLM outputs to be short and written in a specific language.

Which solution will align the LLM response quality with the company's expectations?

A. Adjust the prompt.
B. Choose an LLM of a different size.
C. Increase the temperature.
D. Increase the Top K value.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Adjusting the prompt allows you to specify the desired length and language of the LLM's responses, making it suitable for tailoring the output to meet the company's needs. The other options do not directly control response length or language.

QUESTION 6
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

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Asynchronous inference is suitable for handling large input data and long processing times while still providing responses without blocking other requests. It allows for near real-time latency, whereas the other options are less suitable given the input size and processing time constraints.

QUESTION 7
A company is using domain-specific models. The company wants to avoid creating new models from the beginning. The company instead wants to adapt pre-trained models to create models for new, related tasks.

Which ML strategy meets these requirements?

A. Increase the number of epochs.
B. Use transfer learning.
C. Decrease the number of epochs.
D. Use unsupervised learning.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Transfer learning allows the company to adapt pre-trained models for new, related tasks, saving time and resources compared to training models from scratch. The other options do not address the goal of reusing existing models.

QUESTION 8
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 Q

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Human-in-the-loop validation ensures high accuracy by involving human reviewers to verify and correct annotations, minimizing the risk of errors in the generated images. The other options are not directly relevant for ensuring annotation accuracy in image generation.

QUESTION 9
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 with the 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.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
The foundation model needs the appropriate permissions to decrypt the encrypted data in the S3 bucket. Ensuring that the role used by Amazon Bedrock has permission to access and decrypt the data will resolve the access failure. The other options are not suitable for addressing the encryption and permission issue.

QUESTION 10
A company wants to use language models to create an application for inference on edge devices. The inference must have the lowest latency possible.

Which solution will meet these requirements?

A. Deploy optimized small language models (SLMs) on edge devices.
B. Deploy optimized large language models (LLMs) on edge devices.
C. Incorporate a centralized small language model (SLM) API for asynchronous communication with edge devices.
D. Incorporate a centralized large language model (LLM) API for asynchronous communication with edge devices.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Deploying optimized small language models (SLMs) directly on edge devices provides low latency for inference since the computation happens locally, avoiding the delays associated with network communication. The other options either increase latency or are less suitable for edge deployment.

QUESTION 11
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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Feature Store is a centralized repository for storing and managing features, allowing multiple teams to share and reuse variables (features) for model development. The other options serve different purposes, such as data preprocessing or model documentation.

QUESTION 12
A company wants to use generative AI to increase developer productivity and software development. The company wants to use Amazon Q Developer.

What can Amazon Q Developer do to help the company meet these requirements?

A. Create software snippets, reference tracking, and open source license tracking.
B. Run an application without provisioning or managing servers.
C. Enable voice commands for coding and providing natural language search.
D. Convert audio files to text documents by using ML models.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon Q Developer is designed to assist developers by generating software snippets, tracking references, and managing open source licenses, which aligns with the company's goal of increasing productivity in software development. The other options do not match the intended use of Amazon Q Developer.

QUESTION 13
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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
AWS PrivateLink allows secure, private connectivity between VPCs and AWS services without needing internet access, making it suitable for meeting regulatory compliance standards. The other options either do not provide private connectivity or require internet access.

QUESTION 14
A company wants to develop an educational game where users answer questions such as the following: "A jar contains six red, four green, and three yellow marbles. What is the probability of choosing a green marble from the jar?"

Which solution meets these requirements with the LEAST operational overhead?

A. Use supervised learning to create a regression model that will predict probability.
B. Use reinforcement learning to train a model to return the probability.
C. Use code that will calculate probability by using simple rules and computations.
D. Use unsupervised learning to create a model that will estimate probability density.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Calculating probability in this scenario involves straightforward arithmetic, making it most efficient to use simple rules and computations. This approach requires the least operational overhead compared to using complex ML models, which are unnecessary for such basic tasks.

QUESTION 15
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

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Average response time measures how quickly an AI model produces an output, which reflects the runtime efficiency of the model. The other options do not directly measure the efficiency of operating AI models.

QUESTION 16
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.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Amazon Transcribe converts audio recordings into text, which allows for further analysis and extraction of key information from customer conversations. The other options do not directly handle audio transcription or extraction of information from audio.

QUESTION 17
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
Section: (none)

Explanation/Reference:
Explanation:
Unsupervised learning is suitable for analyzing unlabeled data and grouping it into clusters or tiers, which aligns with the company’s goal of classifying customers. The other methods require labeled data or are used for different types of problems.

QUESTION 18
An AI practitioner wants to use a foundation model (FM) to design a search application. The search application must handle queries that have text and images.

Which type of FM should the AI practitioner use to power the search application?

A. Multi-modal embedding model
B. Text embedding model
C. Multi-modal generation model
D. Image generation model

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
A multi-modal embedding model can handle both text and image queries by embedding them into a shared space, enabling the search application to process and relate different data types. The other options are not suitable for handling both text and image inputs effectively.

QUESTION 19
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.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Fine-tuning a foundation model involves training it with labeled data that contains both input prompts and corresponding expected completions to adjust the model’s behavior to fit the company’s needs. The other options are not directly related to the fine-tuning process using specific labeled data.

QUESTION 20
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.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
An anomaly detection system can identify suspicious behavior, such as IP addresses that deviate from expected patterns, which helps in protecting the application from threats. The other options are not designed for detecting suspicious IP addresses.

QUESTION 21
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
Section: (none)

Explanation/Reference:
Explanation:
The scalable index management and nearest neighbor search capability in Amazon OpenSearch Service enables companies to build vector database applications, which are crucial for tasks like similarity search in AI models. The other options do not specifically provide the vector search functionality.

QUESTION 22
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
Section: (none)

Explanation/Reference:
Explanation:
Generative AI models are used to create new content, such as photorealistic images from text descriptions, which is useful for digital marketing. The other options involve tasks better suited for analytical or detection systems rather than generative models.

QUESTION 23
A company wants to build a generative AI application by using Amazon Bedrock and needs to choose a foundation model (FM). 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

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
The context window determines how much information can fit into a single prompt. It specifies the number of tokens the foundation model can process at once, affecting the length of input that can be provided. The other options do not directly relate to prompt size.

QUESTION 24
A company wants to make a chatbot to help customers. The chatbot will help solve technical problems without human intervention.

The company chose a foundation model (FM) for the chatbot. The chatbot needs to produce responses that adhere to company tone.

Which solution meets these requirements?

A. Set a low limit on the number of tokens the FM can produce.
B. Use batch inferencing to process detailed responses.
C. Experiment and refine the prompt until the FM produces the desired responses.

D. Define a higher number for the temperature parameter.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Experimenting and refining the prompt allows you to guide the FM to produce responses that align with the company's desired tone. This approach helps to shape the behavior of the chatbot. The other options do not directly ensure adherence to company tone.

QUESTION 25
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.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Providing examples with labels in the prompt helps the LLM understand the context of sentiment analysis, improving its accuracy in classifying the new text passage as positive or negative. The other options do not effectively guide the LLM for sentiment analysis.

QUESTION 26
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
Section: (none)

Explanation/Reference:
Explanation:
AWS CloudTrail records API activity and provides a log of access attempts, which helps identify unauthorized users trying to access Amazon Bedrock. The other services are not specifically used for tracking unauthorized access attempts in this context.

QUESTION 27
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
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Serverless Inference allows the company to deploy the ML model without managing any underlying infrastructure, making it suitable for hosting the model and serving predictions. The other options do not directly provide serverless model deployment capabilities.

QUESTION 28
An AI company periodically evaluates its systems and processes with the help of independent software vendors (ISVs). The company needs to receive email message notifications when an ISV's compliance reports become available.

Which AWS service can the company use to meet this requirement?

A. AWS Audit Manager
B. AWS Artifact
C. AWS Trusted Advisor
D. AWS Data Exchange

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
AWS Artifact provides access to compliance reports, including those from independent software vendors (ISVs). The company can use AWS Artifact to receive notifications when new compliance reports are available. The other services are not used for accessing and notifying about compliance reports.

QUESTION 29
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.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Creating a prompt template that helps the LLM detect common attack patterns can reduce the risk of prompt injection and other undesirable manipulations. The other options do not effectively address the risk of prompt manipulation or unauthorized use.

QUESTION 30
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.

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Building and training a generative AI model from scratch gives the company the most ownership of security responsibilities, as it involves full control over data, training, deployment, and security measures. The other options involve varying levels of dependency on third-party tools and services, which reduces the company's ownership of security.

QUESTION 31
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
Section: (none)

Explanation/Reference:
Explanation:
Object detection is used to automatically identify and categorize objects (in this case, animals) in photos. It can detect the presence of animals and classify them accordingly. The other strategies are not suitable for identifying and categorizing animals in images.

QUESTION 32
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
Section: (none)

Explanation/Reference:
Explanation:
The On-Demand pricing model provides flexibility without requiring a long-term commitment, allowing the company to pay only for the resources used, which fits well with a limited budget. The other options are either not relevant to pricing flexibility or involve specific resource commitments.

QUESTION 33

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
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker JumpStart provides pre-built models, including foundation models, that can be quickly deployed and consumed within a VPC, helping teams get started faster. The other options are not designed for deploying foundation models in this context.

QUESTION 34
How can companies use large language models (LLMs) securely on Amazon Bedrock?

A. 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
Section: (none)

Explanation/Reference:
Explanation:
Designing clear prompts and using IAM roles with least privilege access ensures secure use of LLMs on Amazon Bedrock by minimizing access risks and preventing misuse. The other options do not directly address securing the use of LLMs.

QUESTION 35
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
Section: (none)

Explanation/Reference:
Explanation:
GPT models are well-suited for converting natural language input into structured queries like SQL, making them ideal for building an AI-based application that translates employee-provided text into SQL queries. The other options are not designed for natural language understanding and query generation tasks.

QUESTION 36
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
Section: (none)

Explanation/Reference:
Explanation:
Inference is the process where the model analyzes new data (in this case, a new image) to make predictions or identify objects. The other options are related to different stages of the AI lifecycle, such as building or preparing the model.

QUESTION 37
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
Section: (none)

Explanation/Reference:
Explanation:
Data augmentation for imbalanced classes helps address bias by creating a more balanced dataset, ensuring that different attributes are equally represented. This reduces bias in image generation. The other options do not directly address data bias issues.

QUESTION 38
A company is using an Amazon Titan foundation model (FM) in Amazon Bedrock. The company needs to supplement the model by using relevant data from the company's private data sources.

Which solution will meet this requirement?

A. Use a different FM.
B. Choose a lower temperature value.
C. Create an Amazon Bedrock knowledge base.
D. Enable model invocation logging.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Creating an Amazon Bedrock knowledge base allows the company to supplement the foundation model with relevant data from their private data sources. This ensures that the model has access to the additional, context-specific information needed. The other options do not directly address supplementing the model with private data.

QUESTION 39
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
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Clarify helps with transparency and explainability by generating metrics, reports, and examples that show how the model makes decisions, which is essential for meeting regulatory requirements. The other options are not directly related to improving the model's transparency or explainability.

QUESTION 40
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: BC
Section: (none)

Explanation/Reference:
Explanation:
Threat detection: Ensuring security measures are in place to detect threats is important for compliance with regulatory frameworks.
Data protection: Proper data handling and protection measures are key compliance aspects, especially in applications dealing with sensitive customer information.

The other options (auto scaling, cost optimization, and loosely coupled microservices) are more related to performance and architecture rather than regulatory compliance.

QUESTION 41
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
Section: (none)

Explanation/Reference:
Explanation:
Increasing the number of epochs allows the model to train for more iterations, improving its accuracy until the model reaches an optimal level. The other options are either less effective or unrelated to improving

accuracy.

QUESTION 42
A company is building a large language model (LLM) question answering chatbot. The company wants to decrease the number of actions call center employees need to take to respond to customer questions.

Which business objective should the company use to evaluate the effect of the LLM chatbot?

A. Website engagement rate
B. Average call duration
C. Corporate social responsibility
D. Regulatory compliance

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Reducing the average call duration directly indicates how effectively the LLM chatbot is helping call center employees answer customer questions, thus reducing the number of actions needed. The other options are not directly related to the performance of a call center chatbot.

QUESTION 43
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

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Clarify helps detect potential bias in datasets and models during data preparation, training, and deployment. It also provides tools for explainability. The other options are functionalities that do not directly match SageMaker Clarify's core features.

QUESTION 44
A company is developing a new model to predict the prices of specific items. The model performed well on the training dataset. When the company deployed the model to production, the model's performance decreased significantly.

What should the company do to mitigate this problem?

A. Reduce the volume of data that is used in training.
B. Add hyperparameters to the model.
C. Increase the volume of data that is used in training.
D. Increase the model training time.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Increasing the volume of data used in training helps the model generalize better to new, unseen data, reducing overfitting and improving performance in production. The other options either do not address the issue of model generalization or are unlikely to effectively solve the problem.

QUESTION 45

An ecommerce company wants to build a solution to determine customer sentiments based on written customer reviews of products.

Which AWS services meet these requirements? (Choose two.)

A. Amazon Lex
B. Amazon Comprehend
C. Amazon Polly
D. Amazon Bedrock
E. Amazon Rekognition

Correct Answer: BD
Section: (none)

Explanation/Reference:
Explanation:
Amazon Comprehend: This service is specifically designed for natural language processing (NLP) tasks, including sentiment analysis, making it ideal for analyzing customer reviews.
Amazon Bedrock: Bedrock can be used to leverage foundation models, which can also be employed for sentiment analysis tasks.

The other options are not suitable for sentiment analysis of written customer reviews.

QUESTION 46
A company wants to use large language models (LLMs) with Amazon Bedrock to develop a chat interface for the company's product manuals. The manuals are stored as PDF files.

Which solution meets these requirements MOST cost-effectively?

A. Use prompt engineering to add one PDF file as context to the user prompt when the prompt is submitted to Amazon Bedrock.
B. Use prompt engineering to add all the PDF files as context to the user prompt when the prompt is submitted to Amazon Bedrock.
C. Use all the PDF documents to fine-tune a model with Amazon Bedrock. Use the fine-tuned model to process user prompts.
D. Upload PDF documents to an Amazon Bedrock knowledge base. Use the knowledge base to provide context when users submit prompts to Amazon Bedrock.

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Using an Amazon Bedrock knowledge base allows the model to efficiently access relevant information from the PDF manuals when needed, reducing the cost compared to continuously fine-tuning a model or providing all PDFs as context in each prompt. This approach ensures that only necessary context is provided, making it cost-effective.

QUESTION 47
A social media company wants to use a large language model (LLM) for content moderation. The company wants to evaluate the LLM outputs for bias and potential discrimination against specific groups or individuals.

Which data source should the company use to evaluate the LLM outputs with the LEAST administrative effort?

A. User-generated content
B. Moderation logs
C. Content moderation guidelines
D. Benchmark datasets

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Benchmark datasets are standardized datasets specifically designed for evaluating models for bias and fairness, allowing for efficient assessment with minimal administrative effort. The other options would require more manual processing and might not provide a consistent basis for evaluating bias and discrimination.

QUESTION 48
A company wants to use a pre-trained generative AI model to generate content for its marketing campaigns. The company needs to ensure that the generated content aligns with the company's brand voice and messaging requirements.

Which solution meets these requirements?

A. Optimize the model's architecture and hyperparameters to improve the model's overall performance.
B. Increase the model's complexity by adding more layers to the model's architecture.
C. Create effective prompts that provide clear instructions and context to guide the model's generation.
D. Select a large, diverse dataset to pre-train a new generative model.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Creating effective prompts helps guide the pre-trained generative AI model to produce content that aligns with the company's brand voice and messaging. The other options either involve model architecture changes or require extensive training, which are not necessary for aligning content generation.

QUESTION 49
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? (Choose 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: AC
Section: (none)

Explanation/Reference:
Explanation:
Detect imbalances or disparities in the data: Identifying and addressing data imbalances helps minimize biases that could negatively affect customers.
Evaluate the model's behavior so that the company can provide transparency to stakeholders: Evaluating the model and ensuring transparency is important for responsible AI usage, as it helps stakeholders understand how decisions are made.

The other options are either not directly related to minimizing bias or do not address responsible AI development.

QUESTION 50
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.

Correct Answer: A
Section: (none)

Explanation/Reference:

QUESTION 51
A company needs to choose a model from Amazon Bedrock to use internally. The company must identify a model that generates responses in a style that the company's employees prefer.

What should the company do to meet these requirements?

A. Evaluate the models by using built-in prompt datasets.
B. Evaluate the models by using a human workforce and custom prompt datasets.
C. Use public model leaderboards to identify the model.
D. Use the model InvocationLatency runtime metrics in Amazon CloudWatch when trying models.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Evaluating models using a human workforce and custom prompt datasets ensures that the model generates responses in the style that aligns with the company’s preferences. The other options either do not provide direct feedback on style preferences or are not specific enough for determining suitability based on employee preferences.

QUESTION 52
A student at a university is copying content from generative AI to write essays.

Which challenge of responsible generative AI does this scenario represent?

A. Toxicity
B. Hallucinations
C. Plagiarism
D. Privacy

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Copying content from generative AI to write essays without proper attribution constitutes plagiarism, which is a key challenge of responsible generative AI. The other options are unrelated to this specific issue.

QUESTION 53
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

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Amazon EC2 Trn series instances (powered by AWS Trainium chips) are designed to provide efficient and environmentally friendly training of large machine learning models. They are optimized for energy efficiency, which reduces the environmental impact of the training process. The other instance types are not specifically optimized for minimizing environmental effects during training.

QUESTION 54
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
Section: (none)

Explanation/Reference:
Explanation:
Guardrails for Amazon Bedrock can help ensure that the output generated by Amazon Bedrock is appropriate for children. Guardrails are used to apply content moderation, guidelines, and ensure safety by filtering potentially harmful or inappropriate content, which is essential when building an interactive application for children.

QUESTION 55
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

Correct Answer: A
Section: (none)

Explanation/Reference:

QUESTION 56
A digital devices company wants to predict customer demand for memory hardware. The company does not have coding experience or knowledge of ML algorithms and needs to develop a data-driven predictive model. The company needs to perform analysis on internal data and external data.

Which solution will meet these requirements?

A. Store the data in Amazon S3. Create ML models and demand forecast predictions by using Amazon SageMaker built-in algorithms that use the data from Amazon S3.

B. Import the data into Amazon SageMaker Data Wrangler. Create ML models and demand forecast predictions by using SageMaker built-in algorithms.

C. Import the data into Amazon SageMaker Data Wrangler. Build ML models and demand forecast predictions by using an Amazon Personalize Trending-Now recipe.

D. Import the data into Amazon SageMaker Canvas. Build ML models and demand forecast predictions by selecting the values in the data from SageMaker Canvas.

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Canvas is a no-code tool that allows users to build ML models and make predictions without requiring programming knowledge. It is ideal for users with no coding experience, providing an easy interface for importing data and generating predictive models. The other options require more technical expertise or are not designed for no-code model building.

QUESTION 57
A company has installed a security camera. The company uses an ML model to evaluate the security camera footage for potential thefts. The company has discovered that the model disproportionately flags people who are members of a specific ethnic group.

Which type of bias is affecting the model output?

A. Measurement bias

B. Sampling bias

C. Observer bias

D. Confirmation bias

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Sampling bias occurs when the training data is not representative of the overall population, leading to disproportionate flagging of specific groups. In this case, the model may have been trained on biased data that did not adequately represent all ethnic groups, resulting in skewed predictions. The other types of bias do not directly apply to the selection of training data or its representativeness.

QUESTION 58
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

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Reinforcement learning allows the chatbot to learn from interactions by receiving rewards for positive customer feedback, which helps the model self-improve over time. The other options do not directly provide a mechanism for continuous self-improvement based on interactions.

QUESTION 59
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)

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
A confusion matrix provides detailed insights into the performance of a classification model by showing the true positives, false positives, true negatives, and false negatives. This metric helps evaluate how well the model classifies the different types of materials in images. The other metrics are not as suitable for evaluating a classification model.

QUESTION 60
A company has built a chatbot that can respond to natural language questions with images. The company wants to ensure that the chatbot does not return inappropriate or unwanted images.

Which solution will meet these requirements?

A. Implement moderation APIs.
B. Retrain the model with a general public dataset.
C. Perform model validation.
D. Automate user feedback integration.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Implementing moderation APIs can help filter and block inappropriate or unwanted images before they are returned by the chatbot. The other options do not directly address ensuring that the chatbot avoids returning inappropriate images.

QUESTION 61
An AI practitioner is using an Amazon Bedrock base model to summarize session chats from the customer service department. The AI practitioner wants to store invocation logs to monitor model input and output data.

Which strategy should the AI practitioner use?

A. Configure AWS CloudTrail as the logs destination for the model.
B. Enable model invocation logging in Amazon Bedrock.
C. Configure AWS Audit Manager as the logs destination for the model.
D. Configure model invocation logging in Amazon EventBridge.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Enabling invocation logging in Amazon Bedrock allows the AI practitioner to monitor and store the input and output data for model invocations. The other options are not directly used for logging model invocations in Amazon Bedrock.

QUESTION 62

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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Batch transform is ideal for processing large datasets that do not require real-time predictions. It allows the company to perform inference on multiple GBs of data efficiently without needing immediate results. The other options are more suitable for scenarios requiring real-time or near real-time access.

QUESTION 63
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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Embeddings are numerical representations of real-world objects and concepts that help AI and NLP models understand and work with textual information more effectively by capturing relationships and similarities between words or phrases. The other options do not describe this concept.

QUESTION 64
A research company implemented a chatbot by using a foundation model (FM) from Amazon Bedrock. The chatbot searches for answers to questions from a large database of research papers.

After multiple prompt engineering attempts, the company notices that the FM is performing poorly because of the complex scientific terms in the research papers.

How can the company improve the performance of the chatbot?

A. Use few-shot prompting to define how the FM can answer the questions.
B. Use domain adaptation fine-tuning to adapt the FM to complex scientific terms.
C. Change the FM inference parameters.
D. Clean the research paper data to remove complex scientific terms.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Domain adaptation fine-tuning allows the FM to better understand the complex scientific terms by training it with domain-specific data, improving its performance on such specialized content. The other options are either insufficient or not directly related to handling complex terminology effectively.

QUESTION 65
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.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Decreasing the temperature value makes the model's output more deterministic and consistent by reducing randomness in response generation. The other adjustments do not directly ensure consistent responses.

QUESTION 66
A company wants to develop a large language model (LLM) application by using Amazon Bedrock and customer data that is uploaded to Amazon S3. The company's security policy states that each team can access data for only the team's own customers.

Which solution will meet these requirements?

A. Create an Amazon Bedrock custom service role for each team that has access to only the team's customer data.
B. Create a custom service role that has Amazon S3 access. Ask teams to specify the customer name on each Amazon Bedrock request.
C. Redact personal data in Amazon S3. Update the S3 bucket policy to allow team access to customer data.
D. Create one Amazon Bedrock role that has full Amazon S3 access. Create IAM roles for each team that have access to only each team's customer folders.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Creating a custom Amazon Bedrock service role for each team with restricted access to only the team's customer data ensures compliance with the security policy, providing the necessary data segregation and access control. The other options do not effectively enforce team-specific access or may pose risks of broader access than allowed by the policy.

QUESTION 67
A medical company deployed a disease detection model on Amazon Bedrock. To comply with privacy policies, the company wants to prevent the model from including personal patient information in its responses. The company also wants to receive notification when policy violations occur.

Which solution meets these requirements?

A. Use Amazon Macie to scan the model's output for sensitive data and set up alerts for potential violations.
B. Configure AWS CloudTrail to monitor the model's responses and create alerts for any detected personal information.
C. Use Guardrails for Amazon Bedrock to filter content. Set up Amazon CloudWatch alarms for notification of policy violations.
D. Implement Amazon SageMaker Model Monitor to detect data drift and receive alerts when model quality degrades.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Guardrails for Amazon Bedrock can be used to filter content and ensure that personal patient information is not included in model responses. Setting up Amazon CloudWatch alarms allows the company to receive notifications when policy violations occur. The other options are not specifically designed for filtering model output and monitoring policy compliance.

QUESTION 68
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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon Textract can extract text from PDF documents, making it suitable for converting resumes into plain text for further processing. The other services do not provide functionality to extract text from PDFs.

QUESTION 69
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.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Adding a role description to the prompt is the simplest and most effective way to adjust the model's response style based on the user's age range. It requires minimal implementation effort and effectively tailors the output. The other options involve more complex processes, such as fine-tuning or additional reasoning steps.

QUESTION 70
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.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Evaluating a foundation model's accuracy by measuring its performance against a predefined benchmark dataset is the standard approach for assessing accuracy in image classification tasks. The other options do not provide an appropriate measure of classification accuracy.

QUESTION 71
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? (Choose 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.

Correct Answer: AC
Section: (none)

Explanation/Reference:
Explanation:
Include fairness metrics for model evaluation: Fairness metrics help ensure that the LLM is unbiased and treats all cases equitably, which is essential for responsible AI use.
Modify the training data to mitigate bias: Adjusting the training data helps reduce any inherent bias that might exist, contributing to a more fair and responsible LLM.

The other options are related to general model optimization but do not directly address responsible AI practices regarding potential harms like bias and fairness.

QUESTION 72
A company is building an ML model. The company collected new data and analyzed the data by creating a correlation matrix, calculating statistics, and visualizing the data.

Which stage of the ML pipeline is the company currently in?

A. Data pre-processing
B. Feature engineering
C. Exploratory data analysis
D. Hyperparameter tuning

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
The company is currently in the **exploratory data analysis (EDA)** stage, which involves summarizing data through statistics, visualizations, and correlation matrices to understand the dataset before moving on to modeling. The other options are subsequent steps in the ML pipeline.

QUESTION 73
A company has documents that are missing some words because of a database error. The company wants to build an ML model that can suggest potential words to fill in the missing text.

Which type of model meets this requirement?

A. Topic modeling
B. Clustering models
C. Prescriptive ML models
D. BERT-based models

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
BERT-based models are well-suited for natural language understanding tasks, including filling in missing words, because they use contextual information to predict missing tokens in a text. The other types of models are not designed for this type of text completion task.

QUESTION 74
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
Section: (none)

Explanation/Reference:
Explanation:
Amazon Q in Amazon QuickSight allows users to ask questions in natural language and automatically generate graphs and visualizations to display insights, such as total sales for top-selling products. The other options do not provide the same functionality for generating visual analytics.

QUESTION 75
A company is building a chatbot to improve user experience. The company is using a large language model (LLM) from Amazon Bedrock for intent detection. The company wants to use few-shot learning to improve intent detection accuracy.

Which additional data does the company need to meet these requirements?

A. Pairs of chatbot responses and correct user intents
B. Pairs of user messages and correct chatbot responses
C. Pairs of user messages and correct user intents
D. Pairs of user intents and correct chatbot responses

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Few-shot learning involves providing the model with a few examples to help it understand how to perform the task. For intent detection, the company needs pairs of user messages and the correct user intents, which will help the LLM improve its accuracy in detecting user intents. The other options do not provide the necessary pairing for improving intent detection.

QUESTION 76
A company is using few-shot prompting on a base model that is hosted on Amazon Bedrock. The model

currently uses 10 examples in the prompt. The model is invoked once daily and is performing well. The company wants to lower the monthly cost.

Which solution will meet these requirements?

A. Customize the model by using fine-tuning.
B. Decrease the number of tokens in the prompt.
C. Increase the number of tokens in the prompt.
D. Use Provisioned Throughput.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Decreasing the number of tokens in the prompt reduces the amount of data being processed, thereby lowering the cost of using the model. Since the model is performing well, reducing the prompt size is a cost-effective way to maintain performance while lowering expenses. The other options either increase costs or are unrelated to prompt size.

QUESTION 77
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
Section: (none)

Explanation/Reference:
Explanation:
Hallucination occurs when a large language model generates content that appears plausible and factual but is incorrect or fabricated. This is a common issue with LLMs. The other options do not describe this particular behavior.

QUESTION 78
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
Section: (none)

Explanation/Reference:
Explanation:
To ensure that the custom model does not generate responses based on confidential data, the best approach is to retrain the model without including the confidential data. This prevents the model from learning patterns associated with that sensitive information, thereby avoiding its use in inference. The other

options do not address the root cause of the issue—removing confidential data from the training process.

QUESTION 79
A company has built a solution by using generative AI. The solution uses large language models (LLMs) to translate training manuals from English into other languages. The company wants to evaluate the accuracy of the solution by examining the text generated for the manuals.

Which model evaluation strategy meets these requirements?

A. Bilingual Evaluation Understudy (BLEU)
B. Root mean squared error (RMSE)
C. Recall-Oriented Understudy for Gisting Evaluation (ROUGE)
D. F1 score

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
The **BLEU** (Bilingual Evaluation Understudy) score is a common metric used to evaluate the accuracy of machine translation by comparing the generated translation with reference translations. It is specifically designed for translation tasks, whereas the other metrics are not suitable for evaluating translation quality.

QUESTION 80
A large retailer receives thousands of customer support inquiries about products every day. The customer support inquiries need to be processed quickly. The company wants to implement Agents for Amazon Bedrock.

What are the key benefits of using Amazon Bedrock agents that could help this retailer?

A. Generation of custom foundation models (FMs) to predict customer needs
B. Automation of repetitive tasks and orchestration of complex workflows
C. Automatically calling multiple foundation models (FMs) and consolidating the results
D. Selecting the foundation model (FM) based on predefined criteria and metrics

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Amazon Bedrock agents help automate repetitive tasks and orchestrate complex workflows, which is ideal for handling thousands of customer support inquiries efficiently. This helps reduce response times and improves productivity. The other options do not directly address automation and orchestration of tasks for customer support.

QUESTION 81
Which option is a benefit of ongoing pre-training when fine-tuning a foundation model (FM)?

A. Helps decrease the model's complexity
B. Improves model performance over time
C. Decreases the training time requirement
D. Optimizes model inference time

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Ongoing pre-training helps enhance a foundation model's performance by continuously updating it with new data, thereby improving its ability to generalize and perform well on different tasks. The other options do not directly relate to the benefits of ongoing pre-training.

QUESTION 82
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
Section: (none)

Explanation/Reference:
Explanation:
Tokens are the smallest units (e.g., words, subwords, or characters) that generative AI models use to process text. They form the basis of both the input given to and the output generated by the model. The other options do not accurately describe tokens in this context.

QUESTION 83
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
Section: (none)

Explanation/Reference:
Explanation:
Inference costs for large language models are typically driven by the **number of tokens** processed during input and output, as each token incurs computational resources. The other factors (temperature value, training data, and training time) do not directly impact inference costs.

QUESTION 84
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
Section: (none)

Explanation/Reference:
Explanation:
Configuring Amazon SageMaker to use a VPC with an S3 endpoint ensures secure, direct, and managed data flow between Amazon S3 and SageMaker Studio notebooks. This setup avoids public internet exposure and maintains data integrity during transfers. The other options do not provide a solution for managing the data flow in this context.

QUESTION 85
A company has a foundation model (FM) that was customized by using Amazon Bedrock to answer customer queries about products. The company wants to validate the model's responses to new types of queries. The company needs to upload a new dataset that Amazon Bedrock can use for validation.

Which AWS service meets these requirements?

A. Amazon S3
B. Amazon Elastic Block Store (Amazon EBS)
C. Amazon Elastic File System (Amazon EFS)
D. AWS Snowcone

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon S3 is the most suitable AWS service for uploading and storing datasets used for validation purposes. It is highly scalable and integrated with Amazon Bedrock, allowing easy access to data for model validation. The other options do not provide the same level of integration or suitability for managing datasets in this context.

QUESTION 86
Which prompting attack directly exposes the configured behavior of a large language model (LLM)?

A. Prompted persona switches
B. Exploiting friendliness and trust
C. Ignoring the prompt template
D. Extracting the prompt template

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
"Extracting the prompt template" is a type of prompting attack that involves directly exposing the configured behavior or the underlying system prompt of a large language model (LLM). This attack can reveal sensitive details about how the model operates, including its internal instructions or restrictions, which are typically not intended to be disclosed to the user. Such an exposure can compromise the security and reliability of the LLM by making it vulnerable to further exploitation or misuse.

QUESTION 87
A company wants to use Amazon Bedrock. The company needs to review which security aspects the company is responsible for when using Amazon Bedrock.

Which security aspect will the company be responsible for?

A. Patching and updating the versions of Amazon Bedrock
B. Protecting the infrastructure that hosts Amazon Bedrock
C. Securing the company's data in transit and at rest
D. Provisioning Amazon Bedrock within the company network

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
According to the AWS Shared Responsibility Model, AWS manages the security **of** the cloud, including the infrastructure and services like Amazon Bedrock. Customers are responsible for security **in** the cloud, which encompasses protecting their data, managing access controls, and configuring security settings for their applications.

Reference: https://aws.amazon.com/compliance/shared-responsibility-model/?nc1=h_ls

QUESTION 88
A social media company wants to use a large language model (LLM) to summarize messages. The company has chosen a few LLMs that are available on Amazon SageMaker JumpStart. The company wants to compare the generated output toxicity of these models.

Which strategy gives the company the ability to evaluate the LLMs with the LEAST operational overhead?

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

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Automatic model evaluation is the strategy that allows the company to evaluate the LLMs with the least operational overhead. This method leverages automated tools and processes to assess the toxicity or quality of the generated output without the need for manual intervention or crowd-sourced input. By using pre-built evaluation metrics or toxicity detection models, the company can quickly and efficiently evaluate multiple models without the complexity and time required for human evaluations.

QUESTION 89
A company is testing the security of a foundation model (FM). During testing, the company wants to get around the safety features and make harmful content.

Which security technique is this an example of?

A. Fuzzing training data to find vulnerabilities
B. Denial of service (DoS)
C. Penetration testing with authorization
D. Jailbreak

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
"Jailbreaking" refers to attempts to bypass or disable the built-in safety features and restrictions of a system, in this case, a foundation model (FM). This technique involves trying to circumvent the safeguards that prevent the model from generating harmful or unsafe content. Jailbreaking is often performed to exploit vulnerabilities in a model’s filtering or safety protocols, making it a direct attempt to undermine its protections.

QUESTION 90
A company needs to use Amazon SageMaker for model training and inference. The company must comply with regulatory requirements to run SageMaker jobs in an isolated environment without internet access.

Which solution will meet these requirements?

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

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Network isolation in Amazon SageMaker allows you to run training and inference jobs in an environment that does not have access to the internet. This helps ensure that the data and the model do not inadvertently access external resources, meeting regulatory compliance requirements for isolated environments.

QUESTION 91
An ML research team develops custom ML models. The model artifacts are shared with other teams for integration into products and services. The ML team retains the model training code and data. The ML team wants to build a mechanism that the ML team can use to audit models.

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

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

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Model Cards are designed to document the essential details about machine learning models, including their intended uses, training datasets, training parameters, evaluation metrics, and inference environment. This provides a centralized mechanism to store and audit the metadata of the models, which is ideal for the ML team's need to share and audit models effectively.

Key benefits of **Amazon SageMaker Model Cards**:

- Standardized documentation of models' characteristics and intended use cases.
- Transparency and traceability for auditing purposes.
- Integration with other AWS services for lifecycle management.

QUESTION 92
A software company builds tools for customers. The company wants to use AI to increase software development productivity.

Which solution will meet these requirements?

A. Use a binary classification model to generate code reviews.
B. Install code recommendation software in the company's developer tools.
C. Install a code forecasting tool to predict potential code issues.
D. Use a natural language processing (NLP) tool to generate code.

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Natural language processing (NLP) tools can be used to generate code from high-level descriptions or suggestions, which can greatly enhance software development productivity. By leveraging NLP models, developers can automate repetitive coding tasks, generate code snippets, or even complete blocks of code based on natural language inputs. This can speed up development and reduce errors.

QUESTION 93
A retail store wants to predict the demand for a specific product for the next few weeks by using the Amazon SageMaker DeepAR forecasting algorithm.

Which type of data will meet this requirement?

A. Text data
B. Image data
C. Time series data
D. Binary data

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
The Amazon SageMaker DeepAR forecasting algorithm is specifically designed for forecasting scalar (one-dimensional) time series data using recurrent neural networks (RNNs). It excels when trained on datasets containing hundreds of related time series, enabling it to learn patterns across multiple series and provide accurate forecasts.

Reference: **https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html**

QUESTION 94
A large retail bank wants to develop an ML system to help the risk management team decide on loan allocations for different demographics.

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

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

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
In machine learning, class imbalance occurs when certain classes are underrepresented in the training dataset, leading to biased model predictions. To develop an unbiased model, it's crucial to assess the class distribution and adjust the training process to address any imbalances. This can be achieved through techniques such as oversampling the minority class, undersampling the majority class, or applying class weights to ensure the model treats all classes equitably.

QUESTION 95
Which prompting technique can protect against prompt injection attacks?

A. Adversarial prompting
B. Zero-shot prompting
C. Least-to-most prompting
D. Chain-of-thought prompting

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Adversarial prompting is a technique used to defend against prompt injection attacks by crafting inputs that are specifically designed to identify and neutralize malicious prompts. This approach involves generating prompts that can detect and mitigate adversarial inputs, thereby enhancing the robustness of language models against such attacks.

QUESTION 96
A company has fine-tuned a large language model (LLM) to answer questions for a help desk. The company wants to determine if the fine-tuning has enhanced the model's accuracy.

Which metric should the company use for the evaluation?

A. Precision
B. Time to first token
C. F1 score
D. Word error rate

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
The F1 score is a metric that combines precision and recall into a single value, providing a balance between the two. It is particularly useful in evaluating models where there is an uneven class distribution, as it considers both false positives and false negatives. The F1 score is calculated as the harmonic mean of precision and recall:

$$\mathrm{F1} = 2 \times \left(\frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}\right)$$

This metric ranges from 0 to 1, with 1 indicating perfect precision and recall. In the context of evaluating a fine-tuned large language model (LLM) for a help desk application, the F1 score is appropriate because it assesses the model's ability to provide accurate and relevant responses, balancing the trade-off between precision (correctness of responses) and recall (completeness of relevant responses).

Reference: **https://arize.com/blog-course/f1-score/**

QUESTION 97
A company is using Retrieval Augmented Generation (RAG) with Amazon Bedrock and Stable Diffusion to generate product images based on text descriptions. The results are often random and lack specific details. The company wants to increase the specificity of the generated images.

Which solution meets these requirements?

A. Increase the number of generation steps.
B. Use the MASK_IMAGE_BLACK mask source option.
C. Increase the classifier-free guidance (CFG) scale.
D. Increase the prompt strength.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
In Stable Diffusion, the **classifier-free guidance (CFG) scale** parameter controls how closely the generated image adheres to the provided text prompt. By increasing the CFG scale, the model places more emphasis on the prompt, leading to images that more accurately reflect the specified details. However, it's important to balance this setting, as excessively high values can result in less diverse and potentially lower-quality images.

QUESTION 98
A company wants to implement a large language model (LLM) based chatbot to provide customer service agents with real-time contextual responses to customers' inquiries. The company will use the company's policies as the knowledge base.

Which solution will meet these requirements MOST cost-effectively?

A. Retrain the LLM on the company policy data.
B. Fine-tune the LLM on the company policy data.
C. Implement Retrieval Augmented Generation (RAG) for in-context responses.
D. Use pre-training and data augmentation on the company policy data.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Retrieval Augmented Generation (RAG) integrates external data sources with LLMs to produce accurate and contextually relevant outputs without the need for extensive retraining. By connecting the chatbot to the company's policy documents, RAG enables the model to retrieve pertinent information in real-time, ensuring responses are both accurate and up-to-date. This approach is cost-effective as it leverages existing data without the computational expenses associated with retraining or fine-tuning large models.

Reference: **https://aws.amazon.com/what-is/retrieval-augmented-generation/**

QUESTION 99
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
Section: (none)

Explanation/Reference:
Explanation:
Amazon Q Developer is a tool designed to help users with minimal programming experience to work with AWS Glue. It provides a graphical user interface and simplifies the creation of data transformation and extraction workflows, allowing users to perform tasks like querying and working with data without needing deep coding skills.

QUESTION 100
A company is developing a mobile ML app that uses a phone's camera to diagnose and treat insect bites. The company wants to train an image classification model by using a diverse dataset of insect bite photos from different genders, ethnicities, and geographic locations around the world.

Which principle of responsible AI does the company demonstrate in this scenario?

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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
The company is actively seeking to ensure that the image classification model is trained on a diverse dataset that includes insect bite photos from various genders, ethnicities, and geographic locations. This reflects the **fairness** principle of responsible AI, which emphasizes creating models that make unbiased decisions across all demographic groups. By including a diverse range of data, the company is aiming to prevent biases that could lead to inaccurate diagnoses or treatments for certain groups of people.

Fairness ensures that AI systems do not discriminate based on race, gender, geography, or other characteristics.

QUESTION 101
A company is developing an ML model to make loan approvals. The company must implement a solution to detect bias in the model. The company must also be able to explain the model's predictions.

Which solution will meet these requirements?

A. Amazon SageMaker Clarify
B. Amazon SageMaker Data Wrangler
C. Amazon SageMaker Model Cards
D. AWS AI Service Cards

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Clarify is a tool that helps detect bias in machine learning models and provides explainability for model predictions. It allows users to understand the factors influencing a model's predictions and assess whether the model is fair across different demographic groups. This is exactly what the company needs for detecting bias and explaining model decisions, especially for high-stakes applications like loan approvals.

- **Bias detection**: SageMaker Clarify can analyze the training data and the model's predictions to identify and mitigate bias.
- **Explainability**: It provides features for explaining the predictions made by the model, which helps users understand the reasons behind the model's decisions.

QUESTION 102
A company has developed a generative text summarization model by using Amazon Bedrock. The company will use Amazon Bedrock automatic model evaluation capabilities.

Which metric should the company use to evaluate the accuracy of the model?

A. Area Under the ROC Curve (AUC) score
B. F1 score
C. BERTScore
D. Real world knowledge (RWK) score

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
BERTScore is a metric specifically designed for evaluating the quality of text generated by models, particularly in tasks like text summarization and machine translation. It uses contextual embeddings from a pre-trained BERT model to compare generated text with reference text at the word level, making it well-suited for evaluating the accuracy of generative text summarization models.

BERTScore assesses the semantic similarity between the generated text and the reference text, providing a more nuanced evaluation compared to traditional methods that focus on exact matches.

QUESTION 103
An AI practitioner wants to predict the classification of flowers based on petal length, petal width, sepal length, and sepal width.

Which algorithm meets these requirements?

A. K-nearest neighbors (k-NN)

B. K-mean
C. Autoregressive Integrated Moving Average (ARIMA)
D. Linear regression

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
K-nearest neighbors (k-NN) is a supervised learning algorithm used for classification tasks. It works by classifying a data point based on how its features (such as petal length, petal width, sepal length, and sepal width in this case) are similar to the data points in the training set. For this problem, where the task is to classify flowers based on their features, k-NN is an appropriate choice.

k-NN is simple and effective for classification problems where the decision boundary is not necessarily linear, as it looks at the closest neighbors to make predictions.

QUESTION 104
A company is using custom models in Amazon Bedrock for a generative AI application. The company wants to use a company managed encryption key to encrypt the model artifacts that the model customization jobs create.

Which AWS service meets these requirements?

A. AWS Key Management Service (AWS KMS)
B. Amazon Inspector
C. Amazon Macie
D. AWS Secrets Manager

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
AWS Key Management Service (AWS KMS) is a fully managed service that allows you to create and control encryption keys used to encrypt your data. It is ideal for scenarios like this, where a company wants to use a custom, company-managed encryption key to protect model artifacts. AWS KMS allows you to securely manage keys for encrypting and decrypting data, and it integrates with various AWS services, including Amazon Bedrock.

AWS KMS provides centralized key management, and it is designed for use cases that involve encrypting both data and artifacts, such as model customization jobs.

QUESTION 105
A company wants to use large language models (LLMs) to produce code from natural language code comments.

Which LLM feature meets these requirements?

A. Text summarization
B. Text generation
C. Text completion
D. Text classification

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Text generation is the feature of large language models (LLMs) that enables them to produce text based on a given prompt or input. In this scenario, the input would be natural language code comments, and the

LLM would generate code based on those comments. This process involves understanding the context of the comment and generating corresponding code, which is the core functionality of text generation.

QUESTION 106
A company is introducing a mobile app that helps users learn foreign languages. The app makes text more coherent by calling a large language model (LLM). The company collected a diverse dataset of text and supplemented the dataset with examples of more readable versions. The company wants the LLM output to resemble the provided examples.

Which metric should the company use to assess whether the LLM meets these requirements?

A. Value of the loss function
B. Semantic robustness
C. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score
D. Latency of the text generation

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
ROUGE is a set of metrics that evaluates the quality of summaries by comparing the overlap of n-grams, word sequences, and word pairs between the model output and reference examples. Since the company is working on a language model that improves the coherence of text and wants the output to resemble the provided examples (which are more readable versions of the original text), **ROUGE** is the most appropriate metric. It assesses how closely the generated text matches the reference text in terms of content and readability.

ROUGE score is commonly used to evaluate the performance of models in tasks like summarization, where the goal is to ensure the generated text aligns closely with human-provided examples.

QUESTION 107
A company notices that its foundation model (FM) generates images that are unrelated to the prompts. The company wants to modify the prompt techniques to decrease unrelated images.

Which solution meets these requirements?

A. Use zero-shot prompts.
B. Use negative prompts.
C. Use positive prompts.
D. Use ambiguous prompts.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Negative prompts are used to explicitly instruct the model about what to avoid or not generate. By providing the model with specific guidance on what is not desired (e.g., by including terms or concepts that should not appear in the image), the model can better focus on generating relevant and related content. This technique can help reduce unrelated or irrelevant images by constraining the model's creative generation process.

QUESTION 108
A company wants to use a large language model (LLM) to generate concise, feature-specific descriptions for the company’s products.

Which prompt engineering technique meets these requirements?

A. Create one prompt that covers all products. Edit the responses to make the responses more specific, concise, and tailored to each product.

B. Create prompts for each product category that highlight the key features. Include the desired output format and length for each prompt response.

C. Include a diverse range of product features in each prompt to generate creative and unique descriptions.

D. Provide detailed, product-specific prompts to ensure precise and customized descriptions.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
To generate concise, feature-specific descriptions for each product, the company should create prompts tailored to specific product categories. By highlighting the key features of each product category in the prompt, the model can focus on generating descriptions that are relevant and aligned with the unique attributes of each product. Additionally, specifying the desired output format and length ensures that the responses meet the company's requirements for conciseness and clarity.

Tailored prompts help ensure the model generates relevant and accurate descriptions by focusing on the most important features of each product category.

Desired output format and length ensure the responses are consistent and concise, as required.

QUESTION 109
A company is developing an ML model to predict customer churn. The model performs well on the training dataset but does not accurately predict churn for new data.

Which solution will resolve this issue?

A. Decrease the regularization parameter to increase model complexity.

B. Increase the regularization parameter to decrease model complexity.

C. Add more features to the input data.

D. Train the model for more epochs.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
The issue described is a common case of **overfitting**, where the model performs well on the training data but fails to generalize to new, unseen data. This suggests that the model is too complex and has learned to memorize the training data rather than identifying generalizable patterns.

Increasing the **regularization parameter** helps to **reduce model complexity** by penalizing large weights, thereby encouraging simpler models that are less likely to overfit. This will improve the model's ability to generalize to new data, potentially improving performance on unseen customer data.

QUESTION 110
A company is implementing intelligent agents to provide conversational search experiences for its customers. The company needs a database service that will support storage and queries of embeddings from a generative AI model as vectors in the database.

Which AWS service will meet these requirements?

A. Amazon Athena

B. Amazon Aurora PostgreSQL

C. Amazon Redshift

D. Amazon EMR

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Amazon Aurora PostgreSQL is a relational database service that supports a wide range of applications, including those that require vector-based operations. Specifically, Aurora PostgreSQL can be extended with vector search capabilities using extensions like **pgvector**. This extension allows you to store, index, and query vector embeddings from generative AI models, making it well-suited for the company's needs to store and query embeddings as vectors.

pgvector is an extension for PostgreSQL that provides efficient similarity search for vector data, which is ideal for storing and querying embeddings.

QUESTION 111
A financial institution is building an AI solution to make loan approval decisions by using a foundation model (FM). For security and audit purposes, the company needs the AI solution's decisions to be explainable.

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

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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Model complexity plays a significant role in the explainability of an AI solution. Simpler models tend to be more explainable because it is easier to understand how they make decisions. For example, linear regression or decision trees are generally easier to interpret compared to more complex models like deep neural networks.

In contrast, highly complex models, such as deep learning models, may provide very accurate results but are often considered "black boxes," meaning their decision-making process is not easily interpretable. This lack of transparency makes it difficult to explain their decisions in a way that satisfies regulatory requirements or customer understanding.

QUESTION 112
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
Section: (none)

Explanation/Reference:
Explanation:
Amazon Bedrock provides access to a variety of large language models (LLMs), which are well-suited for tasks such as text summarization. By using LLMs, the pharmaceutical company can automatically generate concise and coherent summaries of user reviews for each medication. These models are designed to understand and process large amounts of text, making them ideal for summarizing user reviews in a clear and efficient manner.

Amazon Bedrock allows the company to utilize LLMs that can generate summaries, which is exactly what is needed in this case.

QUESTION 113
A company wants to build a lead prioritization application for its employees to contact potential customers. The application must give employees the ability to view and adjust the weights assigned to different variables in the model based on domain knowledge and expertise.

Which ML model type meets these requirements?

A. Logistic regression model
B. Deep learning model built on principal components
C. K-nearest neighbors (k-NN) model
D. Neural network

Correct Answer: A
Section: (none)

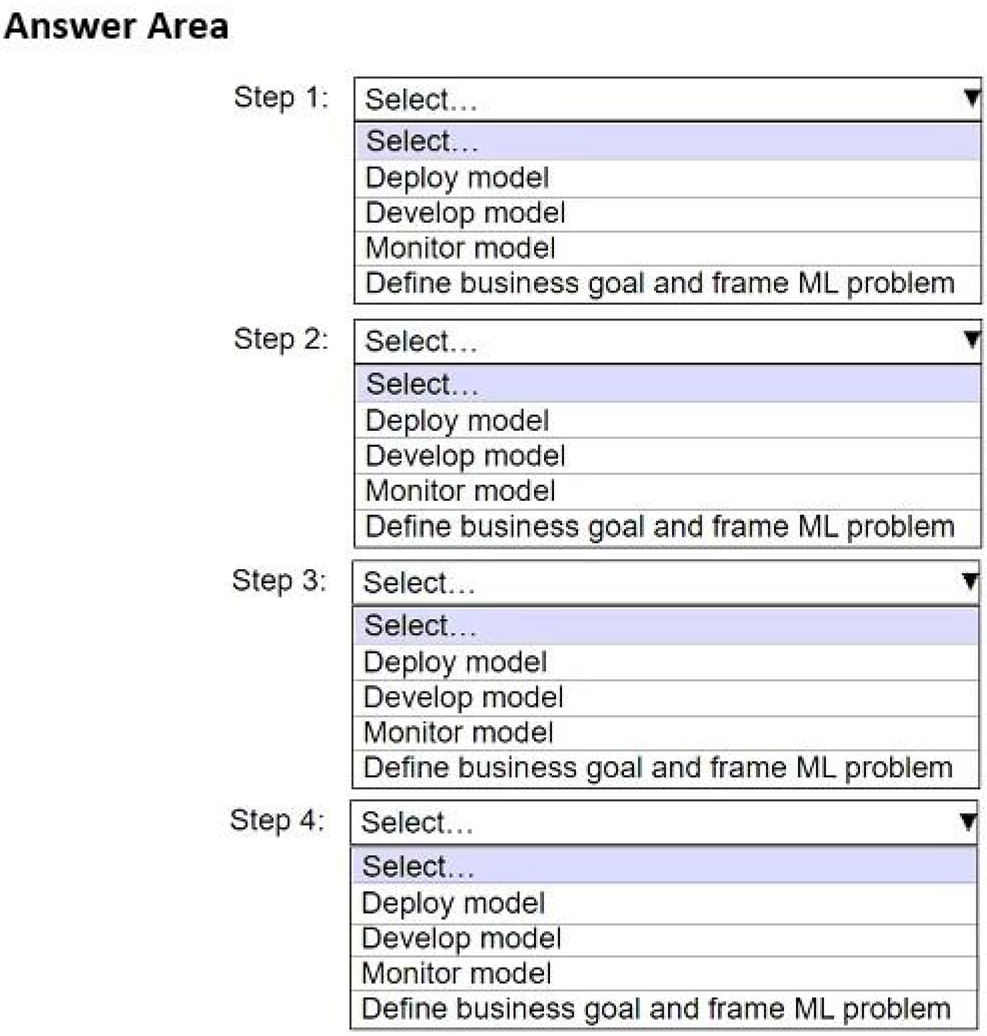
Explanation/Reference:
Explanation:
A **logistic regression model** is a simple, interpretable machine learning model that allows users to adjust the weights of different variables based on domain knowledge and expertise. Since logistic regression is based on a linear relationship between the input features and the predicted outcome, each feature has a corresponding weight (coefficient) that can be easily viewed and adjusted. This makes it a suitable choice for a lead prioritization application where domain experts need the ability to modify the model based on their knowledge.

QUESTION 114
HOTSPOT

A company wants to build an ML application.

Select and order the correct steps from the following list to develop a well-architected ML workload. Each step should be selected one time.

Hot Area:



Answer Area

Step 1: Select…

- Select…
- Deploy model
- Develop model
- Monitor model
- Define business goal and frame ML problem

Step 2: Select…

- Select…
- Deploy model
- Develop model
- Monitor model
- Define business goal and frame ML problem

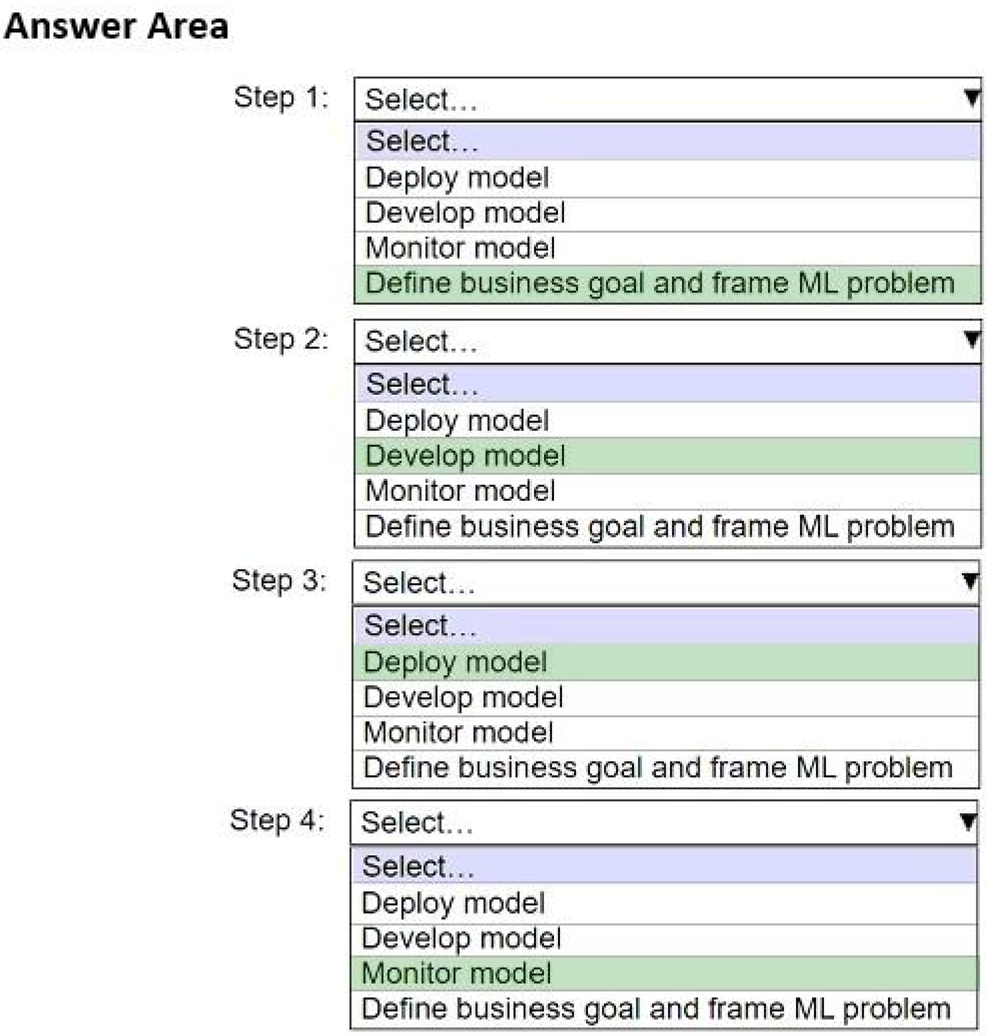
Step 3: Select…

- Select…
- Deploy model
- Develop model
- Monitor model
- Define business goal and frame ML problem

Step 4: Select…

- Select…
- Deploy model
- Develop model
- Monitor model
- Define business goal and frame ML problem

Correct Answer:



Answer Area

Step 1: Select… ▼
- Select…
- Deploy model
- Develop model
- Monitor model
- [✓ Define business goal and frame ML problem]

Step 2: Select… ▼
- Select…
- Deploy model
- [✓ Develop model]
- Monitor model
- Define business goal and frame ML problem

Step 3: Select… ▼
- Select…
- [✓ Deploy model]
- Develop model
- Monitor model
- Define business goal and frame ML problem

Step 4: Select… ▼
- Select…
- Deploy model
- Develop model
- [✓ Monitor model]
- Define business goal and frame ML problem

Section: (none)

Explanation/Reference:
Explanation:

The typical sequence of steps in building an ML application involves:

1. **Defining the business goal and framing the ML problem** - This step involves understanding the business need and determining how ML can address it.
2. **Developing the model** - This involves selecting the appropriate algorithm, training the model, and evaluating it.
3. **Deploying the model** - Once the model is trained, it is deployed to production to serve predictions.
4. **Monitoring the model** - After deployment, the model is monitored to ensure it performs well over time and to detect any issues like model drift.

QUESTION 115
Which strategy will determine if a foundation model (FM) effectively meets business objectives?

A. Evaluate the model's performance on benchmark datasets.
B. Analyze the model's architecture and hyperparameters.
C. Assess the model's alignment with specific use cases.
D. Measure the computational resources required for model deployment.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
To determine if a **foundation model (FM)** effectively meets business objectives, the key is to assess how well the model aligns with the specific use cases or business goals it is intended to address. This involves evaluating whether the model can provide meaningful and accurate outputs relevant to the business's needs, ensuring that it solves the real-world problems the company aims to tackle.

Aligning with specific use cases ensures that the model's capabilities are tailored to the tasks at hand, such as improving customer support, enhancing decision-making, or automating certain processes.

QUESTION 116
A company needs to train an ML model to classify images of different types of animals. The company has a large dataset of labeled images and will not label more data.

Which type of learning should the company use to train the model?

A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning
D. Active learning

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
In **supervised learning**, the model is trained using a labeled dataset, where each image (input) has a corresponding label (the type of animal, in this case). Since the company already has a large dataset of labeled images, supervised learning is the most appropriate approach. The model learns to classify the images based on the features and labels in the training data.

QUESTION 117
Which phase of the ML lifecycle determines compliance and regulatory requirements?

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

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
The **data collection** phase of the ML lifecycle is the most relevant for determining compliance and regulatory requirements. During this phase, organizations must ensure that the data being collected and used for training the model complies with legal and regulatory standards, such as data privacy laws (e.g., GDPR, HIPAA), industry-specific regulations, and ethical considerations. The organization must also verify that they have the proper consent to use the data and that the data does not contain any biases or violate any regulations.

QUESTION 118
A food service company wants to develop an ML model to help decrease daily food waste and increase sales revenue. The company needs to continuously improve the model's accuracy.

Which solution meets these requirements?

A. Use Amazon SageMaker and iterate with newer data.
B. Use Amazon Personalize and iterate with historical data.
C. Use Amazon CloudWatch to analyze customer orders.
D. Use Amazon Rekognition to optimize the model.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
To continuously improve the accuracy of the machine learning model, **Amazon SageMaker** is the appropriate solution. SageMaker allows for efficient model training, deployment, and iteration. The company can use SageMaker to retrain the model regularly with newer data to account for changes in customer behavior, market trends, and other dynamic factors that may affect food waste and sales revenue. Iterating with newer data helps improve the model's performance over time, ensuring it remains accurate and relevant.

QUESTION 119
A company has developed an ML model to predict real estate sale prices. The company wants to deploy the model to make predictions without managing servers or infrastructure.

Which solution meets these requirements?

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

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker provides a fully managed service for deploying machine learning models at scale without needing to manage servers or infrastructure. When you deploy a model using **Amazon SageMaker endpoints**, it handles all the infrastructure, scaling, and maintenance aspects. This allows the company to focus solely on making predictions and integrating the model into their application, without worrying about managing servers or clusters.

Amazon SageMaker simplifies the deployment of ML models by providing scalable, secure, and serverless endpoints, which are ideal for serving predictions in real-time.

QUESTION 120
A company wants to develop an AI application to help its employees check open customer claims, identify details for a specific claim, and access documents for a claim.

Which solution meets these requirements?

A. Use Agents for Amazon Bedrock with Amazon Fraud Detector to build the application.
B. Use Agents for Amazon Bedrock with Amazon Bedrock knowledge bases to build the application.
C. Use Amazon Personalize with Amazon Bedrock knowledge bases to build the application.
D. Use Amazon SageMaker to build the application by training a new ML model.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Amazon Bedrock is a service that allows you to build AI-powered applications using foundation models (FMs) without managing infrastructure. **Agents for Amazon Bedrock** allows you to create intelligent agents that can help users interact with data and perform tasks such as checking open customer claims, identifying claim details, and accessing relevant documents.

By using **Amazon Bedrock knowledge bases**, the AI agent can access relevant information (e.g., details about claims, documents, or customer data) in real time, and provide precise, context-driven answers based on the claim-related data stored in the knowledge base.

- **Agents for Amazon Bedrock** can help automate interactions by understanding and responding to natural language queries, making it a good fit for this use case.
- **Amazon Bedrock knowledge bases** enable the system to access stored information, documents, and details for specific claims efficiently.

QUESTION 121
A manufacturing company uses AI to inspect products and find any damages or defects.

Which type of AI application is the company using?

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

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Computer vision is a field of AI that enables machines to interpret and understand visual information from the world, such as images and videos. In the case of the manufacturing company, the AI is used to inspect products for damages or defects, which involves analyzing visual data (e.g., product images or videos). This is a classic application of computer vision, where the AI system identifies and classifies objects or defects within images.

QUESTION 122
A company wants to create an ML model to predict customer satisfaction. The company needs fully automated model tuning.

Which AWS service meets these requirements?

A. Amazon Personalize
B. Amazon SageMaker
C. Amazon Athena
D. Amazon Comprehend

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker provides a fully managed environment for building, training, and deploying machine learning models, including **automatic model tuning**. Specifically, SageMaker includes a feature called **Automatic Model Tuning** (or **Hyperparameter Optimization**), which automates the process of finding the best hyperparameters for your machine learning model. This is essential when you want to optimize the model’s performance without manual intervention.

Amazon SageMaker allows you to automate the training process and hyperparameter tuning, which aligns perfectly with the company's need for fully automated model tuning.

QUESTION 123
Which technique can a company use to lower bias and toxicity in generative AI applications during the post-processing ML lifecycle?

A. Human-in-the-loop
B. Data augmentation
C. Feature engineering
D. Adversarial training

Correct Answer: A

Section: (none)

Explanation/Reference:
Explanation:
Human-in-the-loop (HITL) is a technique where human oversight is involved in the decision-making process of AI systems. In the context of generative AI applications, HITL can be used during the post-processing phase to identify and mitigate biases or toxic outputs. Humans can review and intervene when the model generates inappropriate or biased content, providing corrections or adjustments that help reduce the likelihood of toxicity and bias. This feedback loop helps refine and improve the model’s outputs over time.

QUESTION 124
A bank has fine-tuned a large language model (LLM) to expedite the loan approval process. During an external audit of the model, the company discovered that the model was approving loans at a faster pace for a specific demographic than for other demographics.

How should the bank fix this issue MOST cost-effectively?

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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
The issue described is a case of **bias** in the model’s decision-making process, where the model is showing a preference for a specific demographic. The most cost-effective way to address this is to **include more diverse training data** that better represents all demographics. By fine-tuning the model with this more diverse data, you can help ensure that the model treats all demographic groups fairly and does not exhibit biased behavior.

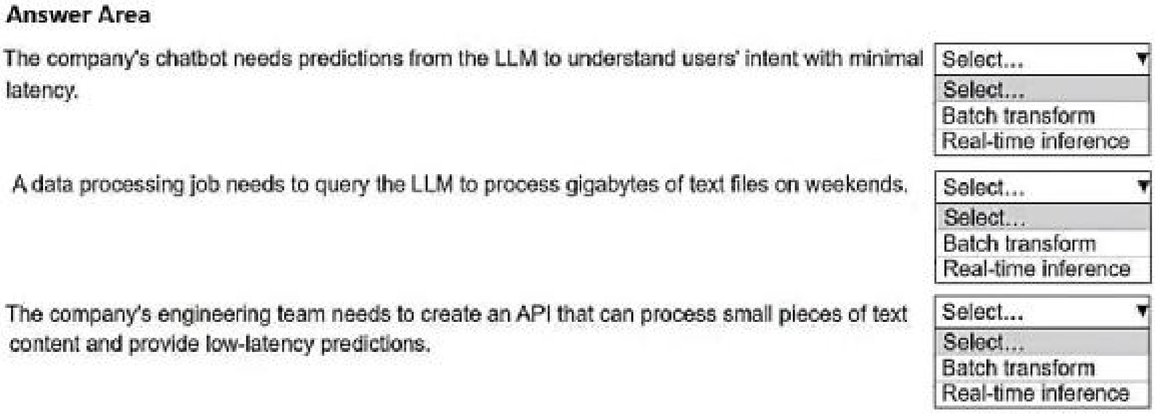
Fine-tuning the model with updated data ensures that the model learns from a more representative sample, improving fairness in its predictions without needing to completely retrain the model from scratch.

QUESTION 125
HOTSPOT

A company has developed a large language model (LLM) and wants to make the LLM available to multiple internal teams. The company needs to select the appropriate inference mode for each team.

Select the correct inference mode from the following list for each use case. Each inference mode should be selected one or more times.

Hot Area:



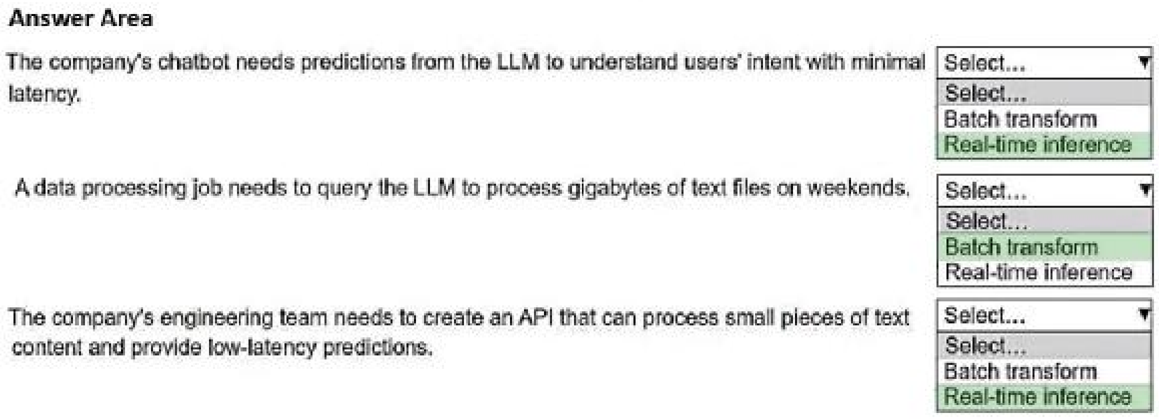
Answer Area

The company's chatbot needs predictions from the LLM to understand users' intent with minimal latency.

A data processing job needs to query the LLM to process gigabytes of text files on weekends.

The company's engineering team needs to create an API that can process small pieces of text content and provide low-latency predictions.

Correct Answer:



Answer Area

The company's chatbot needs predictions from the LLM to understand users' intent with minimal latency.

Select...
Select...
Batch transform
[✓ Real-time inference]

A data processing job needs to query the LLM to process gigabytes of text files on weekends.

Select...
Select...
[✓ Batch transform]
Real-time inference

The company's engineering team needs to create an API that can process small pieces of text content and provide low-latency predictions.

Select...
Select...
Batch transform
[✓ Real-time inference]

Section: (none)

Explanation/Reference:
Explanation:

The company's chatbot needs predictions from the LLM to understand users' intent with minimal latency - Real-time inference
This scenario requires **Real-time inference**, as chatbots typically need immediate responses with low latency to provide an interactive user experience.

A data processing job needs to query the LLM to process gigabytes of text files on weekends - Batch transform
This use case is best suited for **Batch transform**. Since the data processing job involves handling large amounts of data, and it is scheduled for weekends, batch processing can handle these large volumes efficiently.

The company's engineering team needs to create an API that can process small pieces of text content and provide low-latency predictions - Real-time inference
Real-time inference is the correct choice here, as the API needs to process text quickly and provide immediate responses, making real-time inference appropriate.

QUESTION 126
A company needs to log all requests made to its Amazon Bedrock API. The company must retain the logs securely for 5 years at the lowest possible cost.

Which combination of AWS service and storage class meets these requirements? (Choose two.)

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

Correct Answer: AD
Section: (none)

Explanation/Reference:
Explanation:
AWS CloudTrail is the AWS service designed for logging and monitoring API calls made to AWS services, including Amazon Bedrock. CloudTrail records detailed information about the API requests, including the identity of the requester, the time of the request, and the source IP address. This service is ideal for logging all requests made to the Amazon Bedrock API and meets the logging requirement.

Amazon S3 Intelligent-Tiering is a storage class designed for storing data that has unpredictable access patterns. It automatically moves data between two access tiers (frequent and infrequent) based on usage, which helps reduce costs while ensuring data is still available when needed. For retaining logs securely over 5 years at the lowest possible cost, this storage class provides an efficient way to handle long-term storage requirements without incurring unnecessary costs.

QUESTION 127
An ecommerce company wants to improve search engine recommendations by customizing the results for each user of the company's ecommerce platform.

Which AWS service meets these requirements?

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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon Personalize is an AWS service specifically designed to build and deploy personalized recommendations for users. It allows you to create custom machine learning models for personalized user experiences, such as product recommendations, search results, or content suggestions. For an ecommerce company aiming to improve search engine recommendations tailored to each user, Amazon Personalize is the ideal choice as it leverages user behavior data and machine learning to deliver highly relevant recommendations.

QUESTION 128
A hospital is developing an AI system to assist doctors in diagnosing diseases based on patient records and medical images. To comply with regulations, the sensitive patient data must not leave the country the data is located in.

Which data governance strategy will ensure compliance and protect patient privacy?

A. Data residency
B. Data quality
C. Data discoverability
D. Data enrichment

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Data residency refers to the practice of ensuring that data is stored and processed within a specific geographical location to comply with regulations or policies. In this scenario, where sensitive patient data must not leave the country it originates from, implementing a data residency strategy ensures compliance with legal and regulatory requirements while protecting patient privacy. This approach is crucial for organizations like hospitals operating under strict data governance frameworks such as HIPAA or GDPR.

QUESTION 129
A company needs to monitor the performance of its ML systems by using a highly scalable AWS service.

Which AWS service meets these requirements?

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

Correct Answer: A

Section: (none)

Explanation/Reference:
Explanation:
Amazon CloudWatch is a highly scalable AWS service designed for monitoring and observability. It provides real-time monitoring of system metrics, including performance data for ML systems, such as resource utilization (CPU, memory, etc.), model inference latency, request counts, and errors. CloudWatch enables users to set up alarms, visualize metrics, and automate actions based on performance thresholds, making it ideal for monitoring the performance of ML systems.

QUESTION 130
An AI practitioner is developing a prompt for an Amazon Titan model. The model is hosted on Amazon Bedrock. The AI practitioner is using the model to solve numerical reasoning challenges. The AI practitioner adds the following phrase to the end of the prompt: “Ask the model to show its work by explaining its reasoning step by step.”

Which prompt engineering technique is the AI practitioner using?

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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Chain-of-thought prompting is a technique in prompt engineering where the AI model is encouraged to break down its reasoning process step by step to solve a problem, such as numerical reasoning challenges. By explicitly instructing the model to "show its work by explaining its reasoning step by step," the practitioner ensures the model provides a logical sequence of intermediate steps leading to the solution. This improves the accuracy and transparency of the model's outputs, particularly for complex reasoning tasks.

QUESTION 131
Which AWS service makes foundation models (FMs) available to help users build and scale generative AI applications?

A. Amazon Q Developer
B. Amazon Bedrock
C. Amazon Kendra
D. Amazon Comprehend

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Amazon Bedrock is an AWS service that allows users to build and scale generative AI applications using foundation models (FMs) without the need to manage infrastructure. It provides access to pre-trained models from AWS and third-party providers, enabling developers to integrate generative AI capabilities like text generation, summarization, and question-answering into their applications. This makes Amazon Bedrock the ideal choice for building and scaling generative AI solutions.

QUESTION 132
A company is building a mobile app for users who have a visual impairment. The app must be able to hear what users say and provide voice responses.

Which solution will meet these requirements?

A. Use a deep learning neural network to perform speech recognition.

B. Build ML models to search for patterns in numeric data.
C. Use generative AI summarization to generate human-like text.
D. Build custom models for image classification and recognition.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
To meet the requirements of enabling the app to "hear what users say and provide voice responses," the solution must include speech recognition and text-to-speech capabilities. Using a deep learning neural network for speech recognition allows the app to convert spoken words into text. Once the input is understood, text-to-speech systems can provide voice responses back to users. This approach is fundamental in applications that assist users with visual impairments by enabling interaction through spoken language.

QUESTION 133
A company wants to enhance response quality for a large language model (LLM) for complex problem-solving tasks. The tasks require detailed reasoning and a step-by-step explanation process.

Which prompt engineering technique meets these requirements?

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

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Chain-of-thought prompting is specifically designed to enhance the reasoning capabilities of a large language model (LLM) by encouraging it to provide step-by-step explanations for complex problem-solving tasks. This approach helps the model break down a problem into smaller, logical steps, ensuring that the response is detailed, accurate, and easy to follow.

QUESTION 134
A company wants to keep its foundation model (FM) relevant by using the most recent data. The company wants to implement a model training strategy that includes regular updates to the FM.

Which solution meets these requirements?

A. Batch learning
B. Continuous pre-training
C. Static training
D. Latent training

Correct Answer: C
Section: (none)

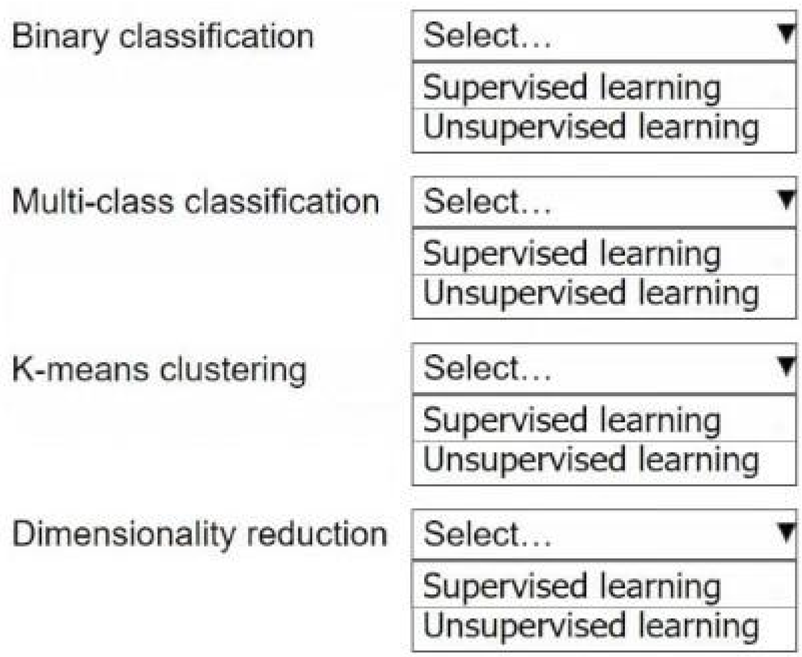
Explanation/Reference:
Explanation:
Continuous pre-training involves regularly updating a foundation model (FM) by training it on new data as it becomes available. This approach ensures that the model remains relevant and accurate by incorporating the most recent information. It is especially important for applications where up-to-date knowledge is crucial, such as news aggregation, customer behavior analysis, or real-time market trends.

QUESTION 135
HOTSPOT

A company wants to develop ML applications to improve business operations and efficiency.

Select the correct ML paradigm from the following list for each use case. Each ML paradigm should be selected one or more times.

Hot Area:



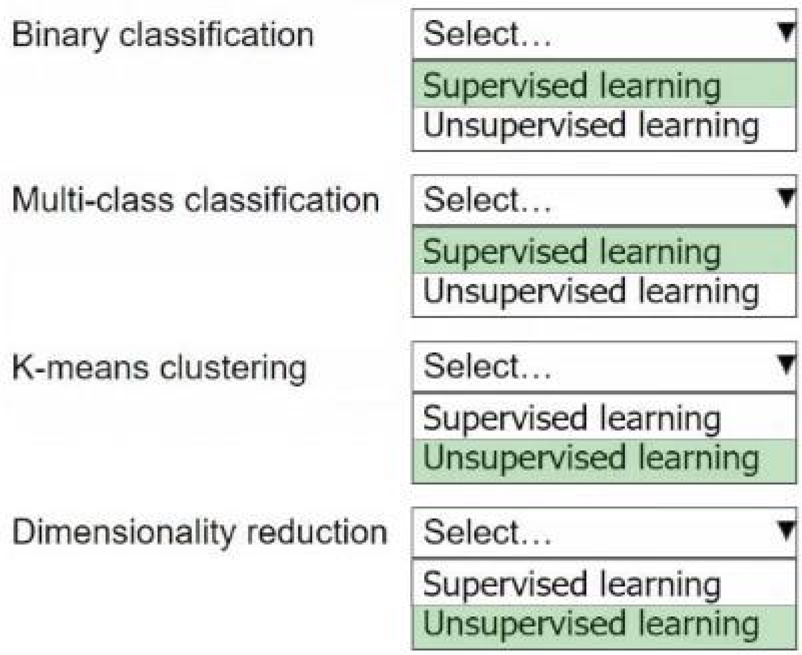
Binary classification
Select... ▼
Supervised learning
Unsupervised learning

Multi-class classification
Select... ▼
Supervised learning
Unsupervised learning

K-means clustering
Select... ▼
Supervised learning
Unsupervised learning

Dimensionality reduction
Select... ▼
Supervised learning
Unsupervised learning

Correct Answer:



Binary classification
Select... ▼
[✓ Supervised learning]
Unsupervised learning

Multi-class classification
Select... ▼
[✓ Supervised learning]
Unsupervised learning

K-means clustering
Select... ▼
Supervised learning
[✓ Unsupervised learning]

Dimensionality reduction
Select... ▼
Supervised learning
[✓ Unsupervised learning]

Section: (none)

Explanation/Reference:
Explanation:

Binary classification: **Supervised learning**
Binary classification involves predicting one of two possible outcomes, requiring labeled training data.

Multi-class classification: **Supervised learning**
Multi-class classification extends binary classification to predict one of several classes, also requiring labeled data.

K-means clustering: **Unsupervised learning**
K-means clustering is used to group data points into clusters without requiring labeled data.

Dimensionality reduction: **Unsupervised learning**
Dimensionality reduction techniques, like PCA, reduce the number of features in the dataset without needing labeled data.

QUESTION 136
Which option is a characteristic of AI governance frameworks for building trust and deploying human-centered AI technologies?

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

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
AI governance frameworks focus on building trust and ensuring the responsible deployment of AI technologies. They establish clear policies and guidelines to address critical aspects such as data management, transparency, ethical considerations, responsible AI practices, and regulatory compliance. These frameworks help organizations mitigate risks, promote fairness, and foster public trust in AI systems, making them essential for creating human-centered AI technologies.

QUESTION 137
An ecommerce company is using a generative AI chatbot to respond to customer inquiries. The company wants to measure the financial effect of the chatbot on the company’s operations.

Which metric should the company use?

A. Number of customer inquiries handled
B. Cost of training AI models
C. Cost for each customer conversation
D. Average handled time (AHT)

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
To measure the financial effect of a generative AI chatbot on the company's operations, the **cost for each customer conversation** is the most relevant metric. It directly quantifies the operational expense associated with using the chatbot to handle customer inquiries. By analyzing this metric, the company can evaluate how much it spends per conversation and compare it to the cost of alternative methods (e.g., human agents), providing insight into the chatbot's financial efficiency and ROI.

QUESTION 138
A company wants to find groups for its customers based on the customers’ demographics and buying patterns.

Which algorithm should the company use to meet this requirement?

A. K-nearest neighbors (k-NN)
B. K-means
C. Decision tree
D. Support vector machine

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
K-means is a clustering algorithm used in **unsupervised learning** to group data points into clusters based on their similarities. It is well-suited for finding patterns in customer demographics and buying behavior. The algorithm identifies groups (clusters) of customers with similar characteristics, which can then be used for targeted marketing, personalized recommendations, or segmentation.

QUESTION 139
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
Section: (none)

Explanation/Reference:
Explanation:
In the context of large language models (LLMs), the **temperature inference parameter** controls the randomness of the model's output. Lowering the temperature reduces randomness and makes the model's responses more deterministic and focused on the most likely predictions. By decreasing the temperature, the likelihood of hallucinations (when the model generates incorrect or nonsensical information) is reduced, as the model relies more on high-probability outputs rather than exploring less likely possibilities.

QUESTION 140
A company is using a large language model (LLM) on Amazon Bedrock to build a chatbot. The chatbot processes customer support requests. To resolve a request, the customer and the chatbot must interact a few times.

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

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

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
To give a large language model (LLM) the ability to use content from previous customer messages in a conversation, the previous messages must be included in the model prompt. This technique is known as **prompt engineering** and allows the LLM to retain context by incorporating a history of the interaction within the prompt. By appending prior exchanges to the prompt, the model can generate contextually relevant and coherent responses throughout the multi-turn conversation.

QUESTION 141
A company's employees provide product descriptions and recommendations to customers when customers call the customer service center. These recommendations are based on where the customers are located. The company wants to use foundation models (FMs) to automate this process.

Which AWS service meets these requirements?

A. Amazon Macie
B. Amazon Transcribe
C. Amazon Bedrock
D. Amazon Textract

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Amazon Bedrock enables companies to use foundation models (FMs) to build and automate tasks like generating product descriptions and recommendations. It allows the integration of pre-trained FMs into applications without managing infrastructure, making it an ideal choice for automating customer service tasks. With Amazon Bedrock, the company can leverage FMs to generate tailored recommendations based on customer locations, enabling dynamic and efficient customer interactions.

QUESTION 142
A company wants to upload customer service email messages to Amazon S3 to develop a business analysis application. The messages sometimes contain sensitive data. The company wants to receive an alert every time sensitive information is found.

Which solution fully automates the sensitive information detection process with the LEAST development effort?

A. Configure Amazon Macie to detect sensitive information in the documents that are uploaded to Amazon S3.
B. Use Amazon SageMaker endpoints to deploy a large language model (LLM) to redact sensitive data.
C. Develop multiple regex patterns to detect sensitive data. Expose the regex patterns on an Amazon SageMaker notebook.
D. Ask the customers to avoid sharing sensitive information in their email messages.

Correct Answer: A
Section: (none)

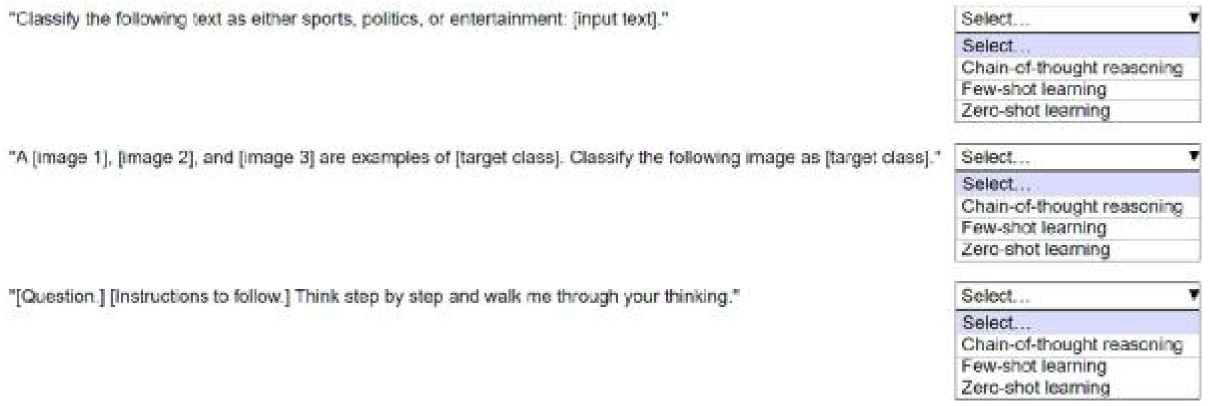
Explanation/Reference:
Explanation:
Amazon Macie is a fully managed data security and privacy service that uses machine learning to discover and protect sensitive data in Amazon S3. It can automatically detect sensitive information, such as personally identifiable information (PII) or financial data, and send alerts when such data is found. This approach minimizes development effort, as it does not require custom regex patterns or model development, and it is specifically designed to handle the scenario described.

QUESTION 143
HOTSPOT

A company is training its employees on how to structure prompts for foundation models.

Select the correct prompt engineering technique from the following list for each prompt template. Each prompt engineering technique should be selected one time.

Hot Area:

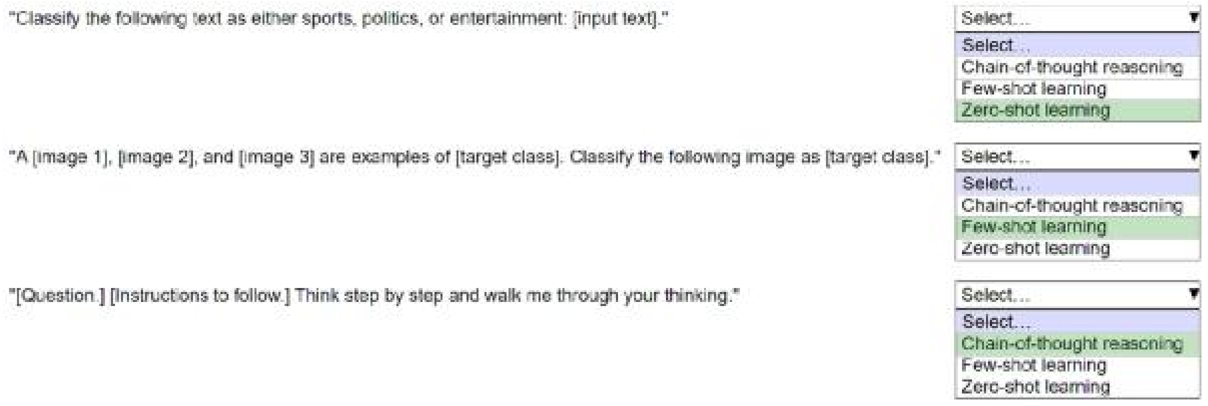


"Classify the following text as either sports, politics, or entertainment: [input text]."

"A [image 1], [image 2], and [image 3] are examples of [target class]. Classify the following image as [target class]."

"[Question.] [Instructions to follow.] Think step by step and walk me through your thinking."

Correct Answer:



"Classify the following text as either sports, politics, or entertainment: [input text]."

"A [image 1], [image 2], and [image 3] are examples of [target class]. Classify the following image as [target class]."

"[Question.] [Instructions to follow.] Think step by step and walk me through your thinking."

Section: (none)

Explanation/Reference:
Explanation:

"Classify the following text as either sports, politics, or entertainment: [input text]."
Correct Answer: Zero-shot learning
Zero-shot learning involves providing the model with a task and no prior examples, relying entirely on the model's pre-trained knowledge to perform the classification.

"A [image 1], [image 2], and [image 3] are examples of [target class]. Classify the following image as [target class]."
Correct Answer: Few-shot learning
Few-shot learning provides the model with a few examples (image 1, image 2, and image 3) before asking it to classify a new instance, helping it generalize to the task with minimal examples.

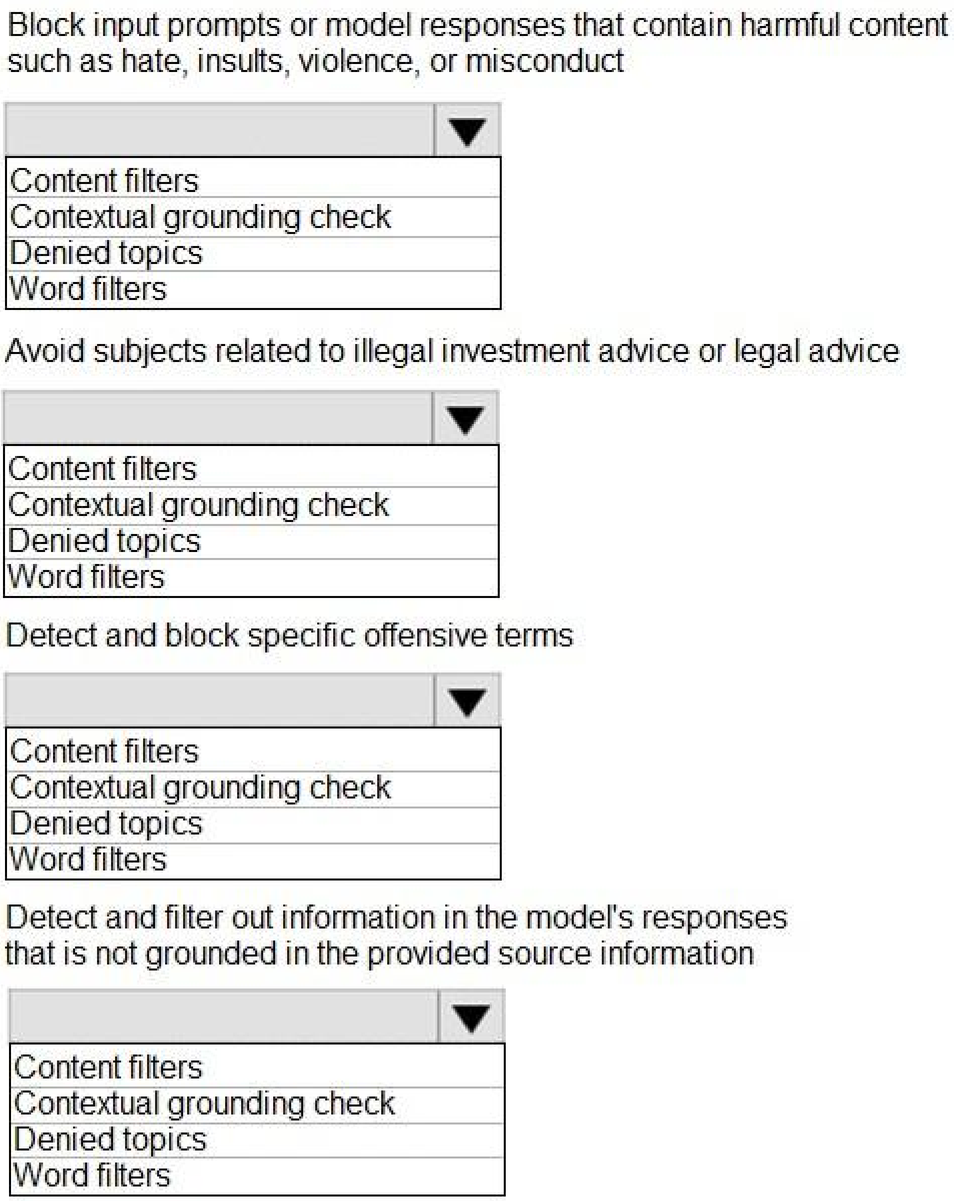
"[Question.] [Instructions to follow.] Think step by step and walk me through your thinking."
Correct Answer: Chain-of-thought reasoning
Chain-of-thought reasoning encourages the model to break down its reasoning process step by step, enhancing its ability to solve complex tasks logically.

QUESTION 144
HOTSPOT

A company is using a generative AI model to develop a digital assistant. The model's responses occasionally include undesirable and potentially harmful content.

Select the correct Amazon Bedrock filter policy from the following list for each mitigation action. Each filter policy should be selected one time.

Hot Area:



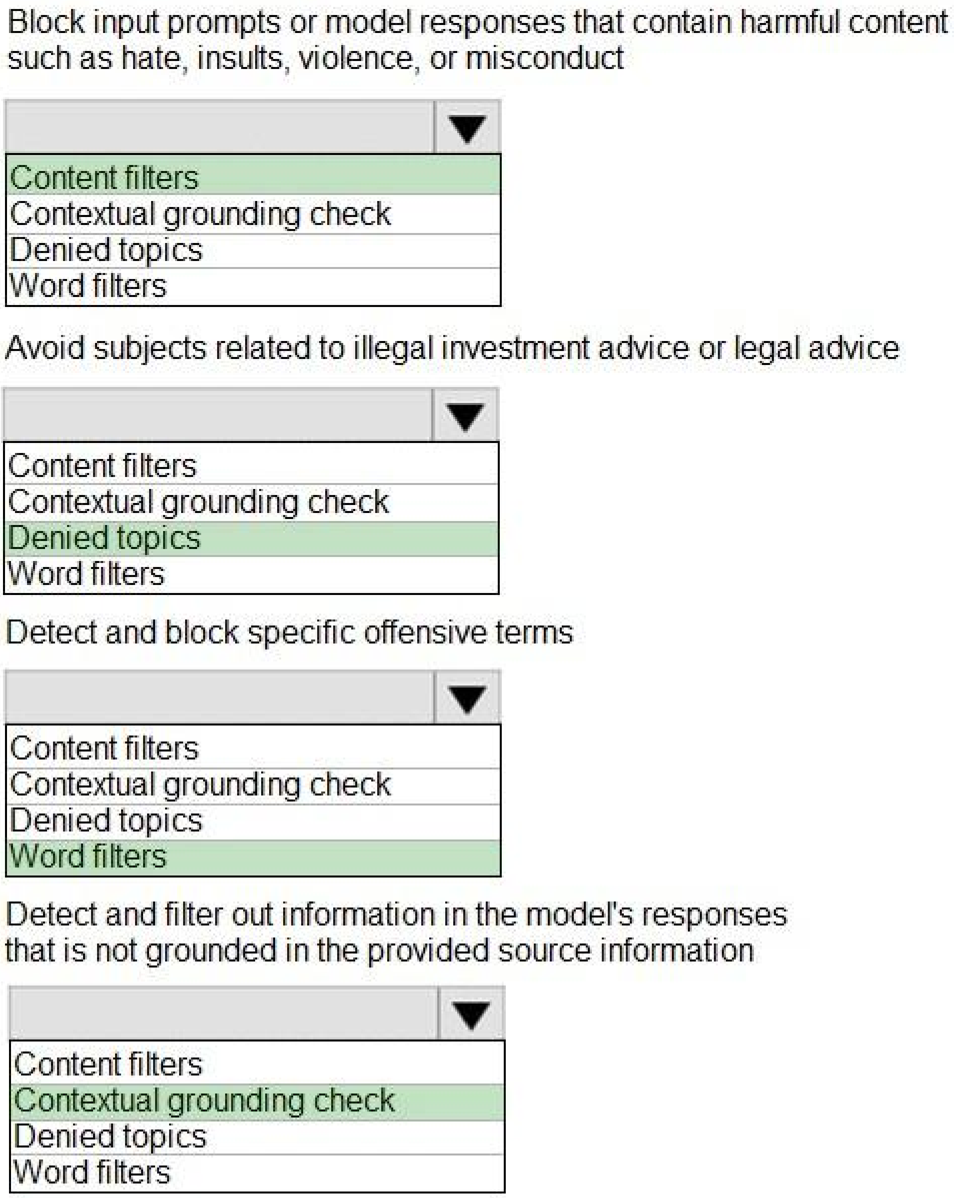
Block input prompts or model responses that contain harmful content such as hate, insults, violence, or misconduct

Avoid subjects related to illegal investment advice or legal advice

Detect and block specific offensive terms

Detect and filter out information in the model's responses that is not grounded in the provided source information

Correct Answer:



Block input prompts or model responses that contain harmful content such as hate, insults, violence, or misconduct

▼
- [✓ Content filters]
- Contextual grounding check
- Denied topics
- Word filters

Avoid subjects related to illegal investment advice or legal advice

▼
- Content filters
- Contextual grounding check
- [✓ Denied topics]
- Word filters

Detect and block specific offensive terms

▼
- Content filters
- Contextual grounding check
- Denied topics
- [✓ Word filters]

Detect and filter out information in the model's responses that is not grounded in the provided source information

▼
- Content filters
- [✓ Contextual grounding check]
- Denied topics
- Word filters

Section: (none)

Explanation/Reference:
Explanation:

Block input prompts or model responses that contain harmful content such as hate, insults, violence, or misconduct - Content filters
Avoid subjects related to illegal investment advice or legal advice - Denied topics
Detect and block specific offensive terms - Word filters
Detect and filter out information in the model's responses that is not grounded in the provided source information - Contextual grounding check

QUESTION 145
Which option is a benefit of using Amazon SageMaker Model Cards to document AI models?

A. Providing a visually appealing summary of a mode’s capabilities.
B. Standardizing information about a model’s purpose, performance, and limitations.
C. Reducing the overall computational requirements of a model.
D. Physically storing models for archival purposes.

Correct Answer: B

Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Model Cards provide a structured way to document key details about AI models, including their intended use, performance metrics, and limitations. This helps organizations maintain transparency, compliance, and governance in AI model development, making it easier to track and manage models over time.

QUESTION 146
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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
The **F1 score** is a metric used to evaluate the performance of a classification model by considering both **precision** (the proportion of correctly predicted positive cases out of all predicted positives) and **recall** (the proportion of correctly predicted positive cases out of all actual positives). It is the **harmonic mean** of precision and recall, ensuring a balance between them, especially when dealing with imbalanced datasets.

QUESTION 147
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.

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Retrieval Augmented Generation (RAG) is a cost-effective approach that enhances AI-generated responses by retrieving relevant information from external knowledge sources. Instead of fine-tuning or re-training a model, RAG dynamically pulls the most recent and relevant data at query time. This is particularly useful in scenarios where questions change over time, ensuring that the AI/ML solution provides accurate and up-to-date responses without requiring expensive and time-consuming model retraining. Prompt engineering techniques further optimize how the model processes and generates responses, improving accuracy and relevance.

QUESTION 148
A company built an AI-powered resume screening system. The company used a large dataset to train the model. The dataset contained resumes that were not representative of all demographics.

Which core dimension of responsible AI does this scenario present?

A. Fairness
B. Explainability
C. Privacy and security

D. Transparency

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
The scenario describes a dataset that is **not representative of all demographics**, which can lead to biased model predictions. This directly relates to **fairness**, a core dimension of responsible AI that ensures AI systems make unbiased and equitable decisions across different demographic groups. Addressing fairness involves techniques such as balanced dataset curation, bias detection, and mitigation strategies to ensure that the AI system does not discriminate against any group.

QUESTION 149
A global financial company has developed an ML application to analyze stock market data and provide stock market trends. The company wants to continuously monitor the application development phases and to ensure that company policies and industry regulations are followed.

Which AWS services will help the company assess compliance requirements? (Choose two.)

A. AWS Audit Manager
B. AWS Config
C. Amazon Inspector
D. Amazon CloudWatch
E. AWS CloudTrail

Correct Answer: AB
Section: (none)

Explanation/Reference:
Explanation:
AWS Audit Manager helps organizations **continuously assess and audit compliance** with industry regulations and internal policies by automating evidence collection and generating audit reports. This is essential for ensuring that the ML application meets regulatory requirements.

AWS Config enables **continuous monitoring and compliance checks** by tracking configuration changes in AWS resources. It helps the company ensure that infrastructure settings align with security policies and industry standards.

QUESTION 150
A company wants to improve the accuracy of the responses from a generative AI application. The application uses a foundation model (FM) on Amazon Bedrock.

Which solution meets these requirements MOST cost-effectively?

A. Fine-tune the FM.
B. Retrain the FM.
C. Train a new FM.
D. Use prompt engineering.

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Using **prompt engineering** is the most cost-effective way to improve the accuracy of responses from a generative AI application without retraining or fine-tuning the foundation model (FM). Prompt engineering involves carefully designing the input prompts to guide the model toward producing better responses, improving relevance and accuracy.

QUESTION 151

A company wants to identify harmful language in the comments section of social media posts by using an ML model. The company will not use labeled data to train the model.

Which strategy should the company use to identify harmful language?

A. Use Amazon Rekognition moderation.
B. Use Amazon Comprehend toxicity detection.
C. Use Amazon SageMaker built-in algorithms to train the model.
D. Use Amazon Polly to monitor comments.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Amazon Comprehend provides **pre-trained NLP models**, including **toxicity detection**, to analyze text for harmful language. Since the company does not plan to use labeled data for training, Amazon Comprehend is a suitable choice because it does not require custom training and can automatically detect toxic or harmful content in comments.

QUESTION 152
A media company wants to analyze viewer behavior and demographics to recommend personalized content. The company wants to deploy a customized ML model in its production environment. The company also wants to observe if the model quality drifts over time.

Which AWS service or feature meets these requirements?

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

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Model Monitor **continuously tracks deployed ML models in production** to detect data drift, model drift, and quality degradation over time. This is essential for ensuring that the recommendation model remains accurate as viewer behavior and demographics change. Model Monitor helps detect anomalies and provides alerts when model performance deviates from expected trends, allowing the company to take corrective action.

QUESTION 153
A company is deploying AI/ML models by using AWS services. The company wants to offer transparency into the models' decision-making processes and provide explanations for the model outputs.

Which AWS service or feature meets these requirements?

A. Amazon SageMaker Model Cards
B. Amazon Rekognition
C. Amazon Comprehend
D. Amazon Lex

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Model Cards help provide **transparency into AI/ML models** by documenting key details such as model purpose, training data, performance metrics, and limitations. This documentation

enables organizations to **explain model outputs and decision-making processes**, ensuring accountability and compliance with responsible AI principles.

QUESTION 154
A manufacturing company wants to create product descriptions in multiple languages.

Which AWS service will automate this task?

A. Amazon Translate
B. Amazon Transcribe
C. Amazon Kendra
D. Amazon Polly

Correct Answer: A
Section: (none)

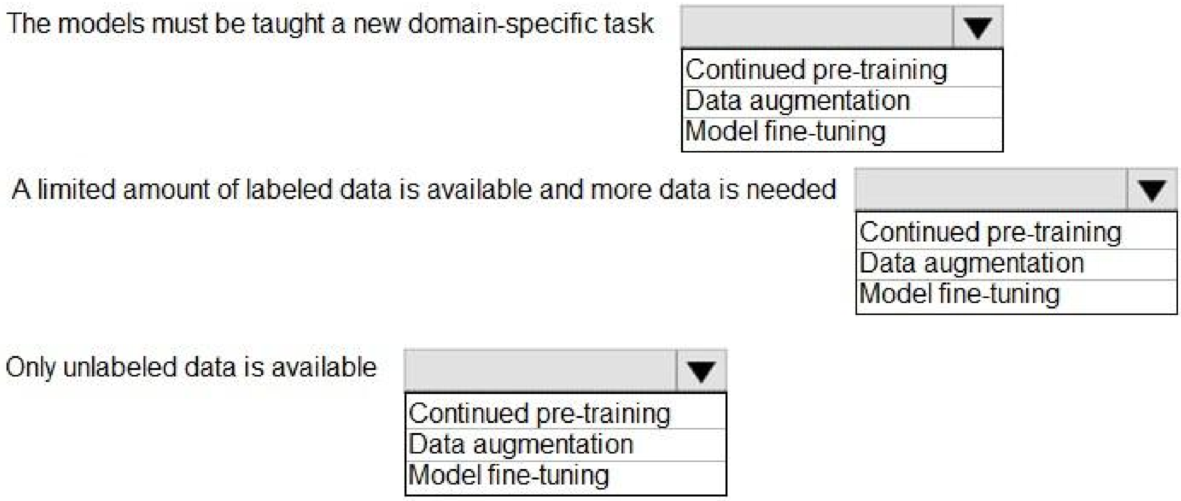
Explanation/Reference:
Explanation:
Amazon Translate is an AWS service that **automates language translation** using neural machine translation (NMT). It enables businesses to **generate product descriptions in multiple languages** quickly and accurately, making it the best choice for this task.

QUESTION 155
HOTSPOT

A company wants more customized responses to its generative AI models' prompts.

Select the correct customization methodology from the following list for each use case. Each use case should be selected one time.

Hot Area:

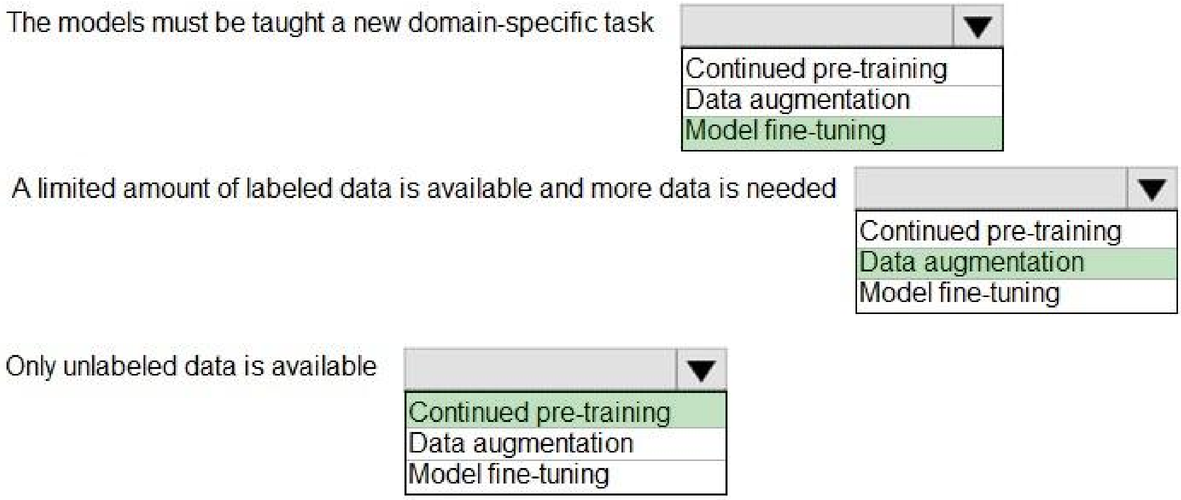


The models must be taught a new domain-specific task ▼
- Continued pre-training
- Data augmentation
- Model fine-tuning

A limited amount of labeled data is available and more data is needed ▼
- Continued pre-training
- Data augmentation
- Model fine-tuning

Only unlabeled data is available ▼
- Continued pre-training
- Data augmentation
- Model fine-tuning

Correct Answer:



The models must be taught a new domain-specific task ▼

- Continued pre-training
- Data augmentation
- [✓ Model fine-tuning]

A limited amount of labeled data is available and more data is needed ▼

- Continued pre-training
- [✓ Data augmentation]
- Model fine-tuning

Only unlabeled data is available ▼

- [✓ Continued pre-training]
- Data augmentation
- Model fine-tuning

Section: (none)

Explanation/Reference:
Explanation:

The models must be taught a new domain-specific task - Model fine-tuning
Fine-tuning is the best approach when a model needs to learn a specific domain-related task. It involves adjusting a pre-trained model with a smaller, task-specific dataset to improve performance in that domain.

A limited amount of labeled data is available and more data is needed - Data augmentation
Data augmentation helps in scenarios where labeled data is scarce by artificially increasing the size and variability of the dataset, improving model generalization.

Only unlabeled data is available - Continued pre-training
Continued pre-training allows an existing foundation model to be further trained on domain-specific, unlabeled data to adapt to new contexts without requiring labeled data.

QUESTION 156
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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon SageMaker Model Cards record key details about machine learning models, including intended use, training data, evaluation metrics, and compliance information. They support governance and reporting by providing a standardized way to document model information throughout its lifecycle.

QUESTION 157
A financial company is using ML to help with some of the company’s tasks.

Which option is a use of generative AI models?

A. Summarizing customer complaints
B. Classifying customers based on product usage
C. Segmenting customers based on type of investments
D. Forecasting revenue for certain products

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Generative AI models are designed to generate new content such as text, images, or audio. Summarizing customer complaints involves generating concise versions of longer texts, which is a task well-suited for generative AI models like large language models.

QUESTION 158
A medical company wants to develop an AI application that can access structured patient records, extract relevant information, and generate concise summaries.

Which solution will meet these requirements?

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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Amazon Comprehend Medical is specifically designed to extract structured medical information (such as medication, condition, test results) from unstructured text. By applying rule-based logic afterward, relevant data can be formatted into concise summaries, meeting both the extraction and summarization needs.

QUESTION 159
Which option describes embeddings in the context of AI?

A. A method for compressing large datasets
B. An encryption method for securing sensitive data
C. A method for visualizing high-dimensional data
D. A numerical method for data representation in a reduced dimensionality space

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Embeddings are numerical representations of data, such as words, images, or items, in a lower-dimensional vector space. They capture semantic relationships and patterns in the data, enabling models to process and compare inputs efficiently.

QUESTION 160
A company is building an AI application to summarize books of varying lengths. During testing, the application fails to summarize some books.

Why does the application fail to summarize some books?

A. The temperature is set too high.
B. The selected model does not support fine-tuning.
C. The Top P value is too high.
D. The input tokens exceed the model’s context size.

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Language models have a maximum context size, which limits the number of input tokens they can process at once. If a book's length exceeds this limit, the model cannot handle the full input, leading to failure in summarization.

QUESTION 161
An airline company wants to build a conversational AI assistant to answer customer questions about flight schedules, booking, and payments. The company wants to use large language models (LLMs) and a knowledge base to create a text-based chatbot interface.

Which solution will meet these requirements with the LEAST development effort?

A. Train models on Amazon SageMaker Autopilot.
B. Develop a Retrieval Augmented Generation (RAG) agent by using Amazon Bedrock.
C. Create a Python application by using Amazon Q Developer.
D. Fine-tune models on Amazon SageMaker Jumpstart.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Amazon Bedrock allows quick integration of large language models with a knowledge base using RAG architecture, enabling accurate and dynamic responses based on up-to-date information. This approach requires minimal development effort while leveraging powerful generative capabilities.

QUESTION 162
What is tokenization used for in natural language processing (NLP)?

A. To encrypt text data
B. To compress text files
C. To break text into smaller units for processing
D. To translate text between languages

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Tokenization is the process of dividing text into smaller units, such as words, subwords, or characters, which can then be analyzed or processed by NLP models. It is a foundational step in preparing text data for machine learning.

QUESTION 163
Which option is a characteristic of transformer-based language models?

A. Transformer-based language models use convolutional layers to apply filters across an input to capture local patterns through filtered views.
B. Transformer-based language models can process only text data.
C. Transformer-based language models use self-attention mechanisms to capture contextual relationships.
D. Transformer-based language models process data sequences one element at a time in cyclic iterations.

Correct Answer: C
Section: (none)

Explanation/Reference:

Explanation:
A key characteristic of transformer models is their use of self-attention mechanisms, which allow the model to weigh the importance of different words in a sequence relative to each other, enabling a deep understanding of context and meaning across the entire input.

QUESTION 164
A financial company is using AI systems to obtain customer credit scores as part of the loan application process. The company wants to expand to a new market in a different geographic area. The company must ensure that it can operate in that geographic area.

Which compliance laws should the company review?

A. Local health data protection laws
B. Local payment card data protection laws
C. Local education privacy laws
D. Local algorithm accountability laws

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
When deploying AI systems that affect individuals, such as credit scoring, companies must comply with local algorithm accountability laws. These laws regulate the fairness, transparency, and impact of automated decision-making, ensuring ethical use of AI in new geographic regions.

QUESTION 165
A company uses Amazon Bedrock for its generative AI application. The company wants to use Amazon Bedrock Guardrails to detect and filter harmful user inputs and model-generated outputs.

Which content categories can the guardrails filter? (Choose two.)

A. Hate
B. Politics
C. Violence
D. Gambling
E. Religion

Correct Answer: AC
Section: (none)

Explanation/Reference:
Explanation:
Amazon Bedrock Guardrails provide configurable content filters to detect and block harmful content in generative AI applications. Specifically, they include filters for categories such as Hate and Violence, among others. These filters can be applied to both user inputs and model-generated outputs to ensure that the AI application adheres to responsible AI practices and organizational policies.

QUESTION 166
Which scenario describes a potential risk and limitation of prompt engineering in the context of a generative AI model?

A. Prompt engineering does not ensure that the model always produces consistent and deterministic outputs, eliminating the need for validation.
B. Prompt engineering could expose the model to vulnerabilities such as prompt injection attacks.
C. Properly designed prompts reduce but do not eliminate the risk of data poisoning or model hijacking.
D. Prompt engineering does not ensure that the model will consistently generate highly reliable outputs when working with real-world data.

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
A key risk in prompt engineering is prompt injection, where attackers manipulate input prompts to alter a model's behavior or produce unintended outputs. This vulnerability arises from the model's sensitivity to input structure and content, making it a critical limitation in secure prompt design.

QUESTION 167
A publishing company built a Retrieval Augmented Generation (RAG) based solution to give its users the ability to interact with published content. New content is published daily. The company wants to provide a near real-time experience to users.

Which steps in the RAG pipeline should the company implement by using offline batch processing to meet these requirements? (Choose two.)

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

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
In a RAG pipeline, generating content embeddings and creating the search index can be done offline in batch processes because they involve static published content. This enables the system to be updated periodically without affecting real-time user interactions. User queries and response generation, on the other hand, must occur in real time.

QUESTION 168
Which technique breaks a complex task into smaller subtasks that are sent sequentially to a large language model (LLM)?

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

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Prompt chaining involves breaking a complex task into smaller, manageable subtasks and passing them sequentially to a large language model. Each step builds upon the previous one, enabling more structured reasoning and improved task execution.

QUESTION 169
An AI practitioner needs to improve the accuracy of a natural language generation model. The model uses rapidly changing inventory data.

Which technique will improve the model's accuracy?

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

Correct Answer: C

Section: (none)

Explanation/Reference:
Explanation:
RAG enhances a language model by fetching up-to-date, domain-specific data (e.g., current inventory) at inference time and conditioning the generation on those facts, ensuring the output reflects the latest information and improving accuracy without retraining the core model.

QUESTION 170
A company wants to collaborate with several research institutes to develop an AI model. The company needs standardized documentation of model version tracking and a record of model development.

Which solution meets these requirements?

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

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
SageMaker Model Cards provide a built-in framework for capturing and versioning key metadata, such as training data details, performance metrics, lineage, and intended use, for each model iteration. This standardized documentation and history of development fulfills requirements for model version tracking and auditable records.

QUESTION 171
A company that uses multiple ML models wants to identify changes in original model quality so that the company can resolve any issues.

Which AWS service or feature meets these requirements?

A. Amazon SageMaker JumpStart
B. Amazon SageMaker HyperPod
C. Amazon SageMaker Data Wrangler
D. Amazon SageMaker Model Monitor

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
SageMaker Model Monitor continuously measures data and prediction quality in production, detects deviations from your model’s baseline (such as data drift or accuracy degradation), and alerts you so you can investigate and resolve issues promptly.

QUESTION 172
What is the purpose of chunking in Retrieval Augmented Generation (RAG)?

A. To avoid database storage limitations for large text documents by storing parts or chunks of the text
B. To improve efficiency by avoiding the need to convert large text into vector embeddings
C. To improve the contextual relevancy of results retrieved from the vector index
D. To decrease the cost of storage by storing parts or chunks of the text

Correct Answer: C
Section: (none)

Explanation/Reference:

Explanation:
By splitting large documents into smaller, semantically coherent chunks, the retrieval system can match and return only the most context-relevant segments for a given query, enhancing the precision and accuracy of the generated responses.

QUESTION 173
A company is developing an editorial assistant application that uses generative AI. During the pilot phase, usage is low and application performance is not a concern. The company cannot predict application usage after the application is fully deployed and wants to minimize application costs.

Which solution will meet these requirements?

A. Use GPU-powered Amazon EC2 instances.
B. Use Amazon Bedrock with Provisioned Throughput.
C. Use Amazon Bedrock with On-Demand Throughput.
D. Use Amazon SageMaker JumpStart.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
On-Demand Throughput in Bedrock charges only for the inference calls you actually make, with no upfront capacity commitment. This lets you start with minimal pilot usage and elastically handle unpredictable future load while keeping costs as low as possible.

QUESTION 174
A company deployed a Retrieval Augmented Generation (RAG) application on Amazon Bedrock that gathers financial news to distribute in daily newsletters. Users have recently reported politically influenced ideas in the newsletters.

Which Amazon Bedrock guardrail can identify and filter this content?

A. Word filters
B. Denied topics
C. Sensitive information filters
D. Content filters

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
The Denied topics guardrail lets you define high-level categories (such as politics) that the model must avoid. Bedrock enforces these rules during generation, filtering out any content related to politically influenced ideas.

QUESTION 175
A financial company is developing a fraud detection system that flags potential fraud cases in credit card transactions. Employees will evaluate the flagged fraud cases. The company wants to minimize the amount of time the employees spend reviewing flagged fraud cases that are not actually fraudulent.

Which evaluation metric meets these requirements?

A. Recall
B. Accuracy
C. Precision
D. Lift chart

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Precision measures the proportion of flagged cases that are truly fraudulent (TP / [TP + FP]). Maximizing precision reduces the number of false positives employees must review, cutting down wasted effort on non-fraudulent cases.

QUESTION 176
A company designed an AI-powered agent to answer customer inquiries based on product manuals.

Which strategy can improve customer confidence levels in the AI-powered agent's responses?

A. Writing the confidence level in the response
B. Including referenced product manual links in the response
C. Designing an agent avatar that looks like a computer
D. Training the agent to respond in the company's language style

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Providing direct links to the exact sections of the product manual that support each answer lets customers verify and trust the information, boosting confidence in the AI agent’s responses.

QUESTION 177
A hospital developed an AI system to provide personalized treatment recommendations for patients. The AI system must provide the rationale behind the recommendations and make the insights accessible to doctors and patients.

Which human-centered design principle does this scenario present?

A. Explainability
B. Privacy and security
C. Fairness
D. Data governance

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Explainability ensures that the AI system reveals the reasoning behind its recommendations in an understandable way, making insights transparent and accessible to both doctors and patients.

QUESTION 178
Which statement presents an advantage of using Retrieval Augmented Generation (RAG) for natural language processing (NLP) tasks?

A. RAG can use external knowledge sources to generate more accurate and informative responses.
B. RAG is designed to improve the speed of language model training.
C. RAG is primarily used for speech recognition tasks.
D. RAG is a technique for data augmentation in computer vision tasks.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
By retrieving relevant documents or data at inference time and conditioning the generation on that external knowledge, RAG enriches model outputs with up-to-date, domain-specific information, boosting accuracy

and informativeness without retraining the core model.

QUESTION 179
A company has created a custom model by fine-tuning an existing large language model (LLM) from Amazon Bedrock. The company wants to deploy the model to production and use the model to handle a steady rate of requests each minute.

Which solution meets these requirements MOST cost-effectively?

A. Deploy the model by using an Amazon EC2 compute optimized instance.
B. Use the model with on-demand throughput on Amazon Bedrock.
C. Store the model in Amazon S3 and host the model by using AWS Lambda.
D. Purchase Provisioned Throughput for the model on Amazon Bedrock.

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
Provisioned Throughput is priced lower per request when you have a predictable, steady volume of calls. By committing to a fixed throughput level, you secure the necessary capacity at a reduced unit cost compared to on-demand, making it the most cost-effective choice for steady-minute usage.

QUESTION 180
Which technique involves training AI models on labeled datasets to adapt the models to specific industry terminology and requirements?

A. Data augmentation
B. Fine-tuning
C. Model quantization
D. Continuous pre-training

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Fine-tuning takes a pre-trained model and continues training it on a labeled, domain-specific dataset so the model learns industry terminology and task nuances directly from that specialized data.

QUESTION 181
A company is creating an agent for its application by using Amazon Bedrock Agents. The agent is performing well, but the company wants to improve the agent's accuracy by providing some specific examples.

Which solution meets these requirements?

A. Modify the advanced prompts for the agent to include the examples.
B. Create a guardrail for the agent that includes the examples.
C. Use Amazon SageMaker Ground Truth to label the examples.
D. Run a script in AWS Lambda that adds the examples to the training dataset.

Correct Answer: A
Section: (none)

Explanation/Reference:
Explanation:
Embedding specific input–output examples directly into the agent's advanced prompt (few-shot prompting) guides the model toward more accurate behavior without retraining or additional tooling.

QUESTION 182

Which option is a benefit of using infrastructure as code (IaC) in machine learning operations (MLOps)?

A. IaC eliminates the need for hyperparameter tuning.
B. IaC always provisions powerful compute instances, contributing to the training of more accurate models.
C. IaC streamlines the deployment of scalable and consistent ML workloads in cloud environments.
D. IaC minimizes overall expenses by deploying only low-cost instances.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
By describing infrastructure in code, teams can version, automate, and repeatably provision resources, ensuring that ML training and inference environments are consistent, scalable, and easy to reproduce across development, testing, and production.

QUESTION 183
A company wants to fine-tune a foundation model (FM) to answer questions for a specific domain. The company wants to use instruction-based fine-tuning.

How should the company prepare the training data?

A. Gather company internal documents and industry-specific materials. Merge the documents and materials into a single file.
B. Collect external company reviews from various online sources. Manually label each review as either positive or negative.
C. Create pairs of questions and answers that specifically address topics related to the company's industry domain.
D. Create few-shot prompts to instruct the model to answer only domain knowledge.

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Instruction-based fine-tuning requires a dataset of instruction–response examples. By curating question–answer pairs focused on the company’s domain, you teach the model exactly how to interpret domain-specific queries and generate the correct responses during inference.

QUESTION 184
Which ML technique ensures data compliance and privacy when training AI models on AWS?

A. Reinforcement learning
B. Transfer learning
C. Federated learning
D. Unsupervised learning

Correct Answer: C
Section: (none)

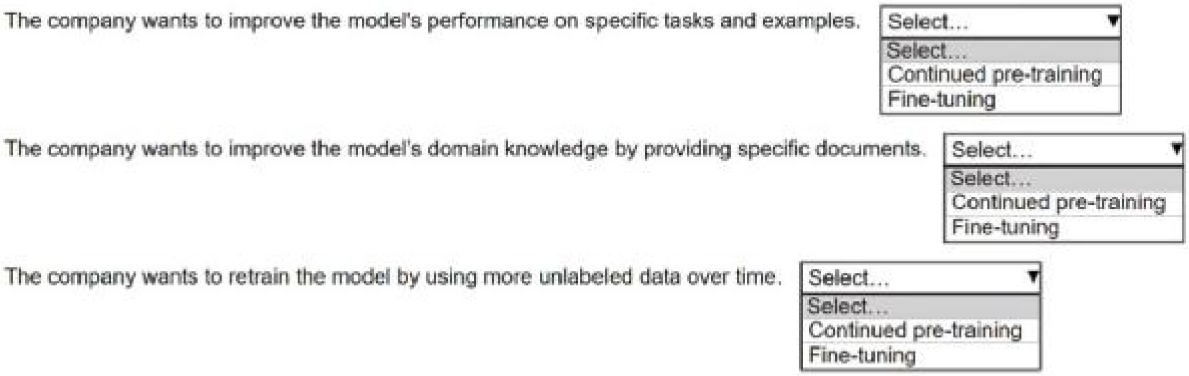
Explanation/Reference:
Explanation:
Federated learning lets you train a global model across multiple data holders (for example, different AWS accounts or edge devices) without moving raw data to a central location. Each participant trains locally on its own private dataset and only shares model updates, preserving data privacy and ensuring compliance.

QUESTION 185
HOTSPOT

A company needs to customize a base model that is hosted on Amazon Bedrock.

Select the correct model customization method from the following list of company requirements. Each model customization method should be selected one or more times.

Hot Area:



The company wants to improve the model's performance on specific tasks and examples. Select... ▼

Select...
Continued pre-training
Fine-tuning

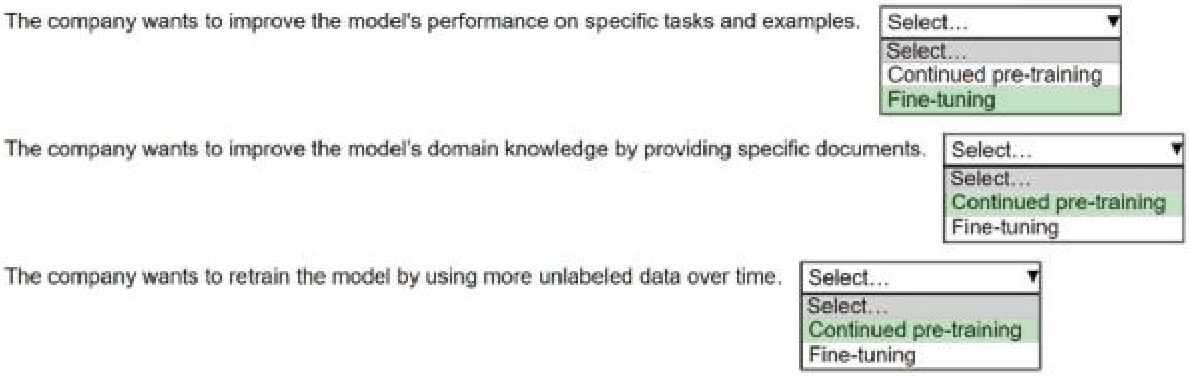
The company wants to improve the model's domain knowledge by providing specific documents. Select... ▼

Select...
Continued pre-training
Fine-tuning

The company wants to retrain the model by using more unlabeled data over time. Select... ▼

Select...
Continued pre-training
Fine-tuning

Correct Answer:



The company wants to improve the model's performance on specific tasks and examples. Select... ▼

Select...
Continued pre-training
[✓ Fine-tuning]

The company wants to improve the model's domain knowledge by providing specific documents. Select... ▼

Select...
[✓ Continued pre-training]
Fine-tuning

The company wants to retrain the model by using more unlabeled data over time. Select... ▼

Select...
[✓ Continued pre-training]
Fine-tuning

Section: (none)

Explanation/Reference:

QUESTION 186
A manufacturing company has an application that ingests consumer complaints from publicly available sources. The application uses complex hard-coded logic to process the complaints. The company wants to scale this logic across markets and product lines.

Which advantage do generative AI models offer for this scenario?

A. Predictability of outputs
B. Adaptability
C. Less sensitivity to changes in inputs
D. Explainability

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
Generative AI models can generalize learned patterns to new domains and formats with minimal manual reconfiguration. This adaptability lets you extend complaint-processing logic across different markets and product lines without rewriting complex hard-coded rules.

QUESTION 187
A financial company wants to flag all credit card activity as possibly fraudulent or non-fraudulent based on transaction data.

Which type of ML model meets these requirements?

A. Regression
B. Diffusion
C. Binary classification
D. Multi-class classification

Correct Answer: C
Section: (none)

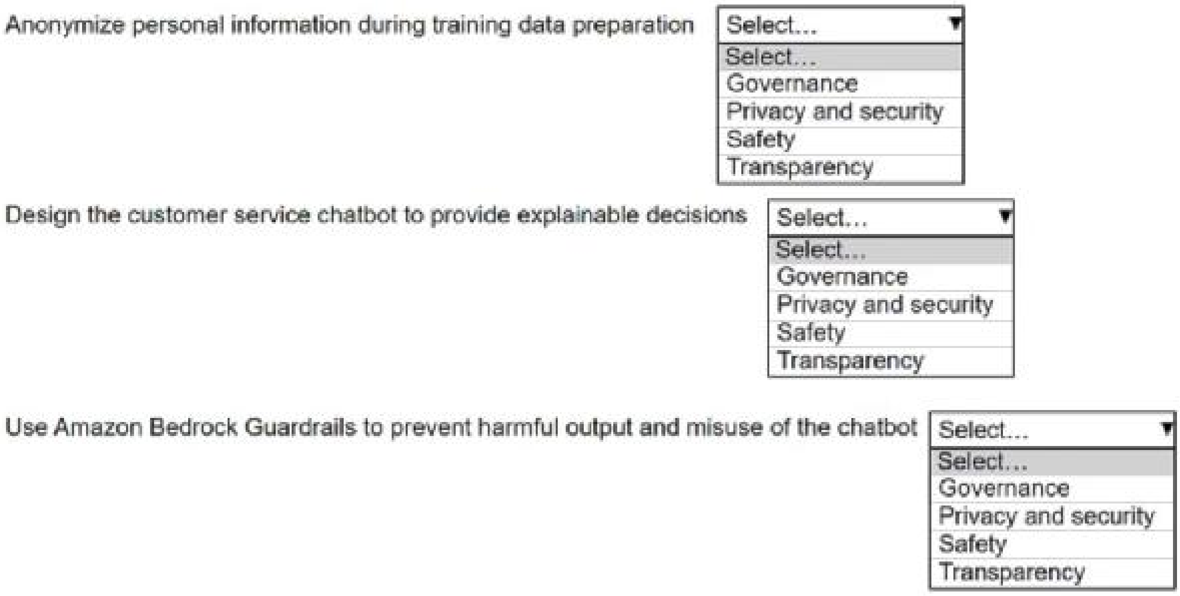
Explanation/Reference:
Explanation:
Binary classification models are designed to distinguish between two classes - fraudulent versus non-fraudulent transactions - making them the appropriate choice for this use case.

QUESTION 188
HOTSPOT

A company is designing a customer service chatbot by using a fine-tuned large language model (LLM). The company wants to ensure that the chatbot uses responsible AI characteristics.

Select the correct responsible AI characteristic from the following list for each application design action. Each responsible AI characteristic should be selected one time or not at all.

Hot Area:

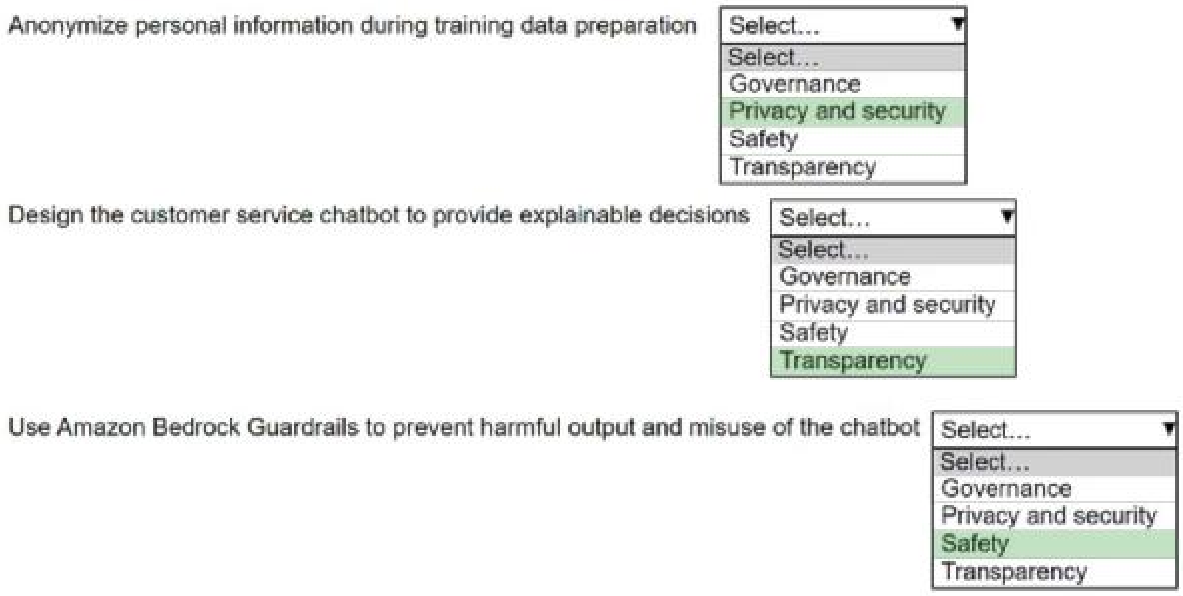


Anonymize personal information during training data preparation

Design the customer service chatbot to provide explainable decisions

Use Amazon Bedrock Guardrails to prevent harmful output and misuse of the chatbot

Correct Answer:



Anonymize personal information during training data preparation

Select... ▼

Select...
Governance
[✓ Privacy and security]
Safety
Transparency

Design the customer service chatbot to provide explainable decisions

Select... ▼

Select...
Governance
Privacy and security
Safety
[✓ Transparency]

Use Amazon Bedrock Guardrails to prevent harmful output and misuse of the chatbot

Select... ▼

Select...
Governance
Privacy and security
[✓ Safety]
Transparency

Section: (none)

Explanation/Reference:

QUESTION 189
A hospital wants to use a generative AI solution with speech-to-text functionality to help improve employee skills in dictating clinical notes.

Which AWS service meets these requirements?

A. Amazon Q Developer
B. Amazon Polly
C. Amazon Rekognition
D. AWS HealthScribe

Correct Answer: D
Section: (none)

Explanation/Reference:
Explanation:
AWS HealthScribe is specifically designed for healthcare workflows, providing accurate speech-to-text transcription of clinical conversations and generating structured clinical notes, meeting the hospital’s need to improve employee dictation of clinical documentation.

QUESTION 190
Which type of AI model makes numeric predictions?

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

Correct Answer: B
Section: (none)

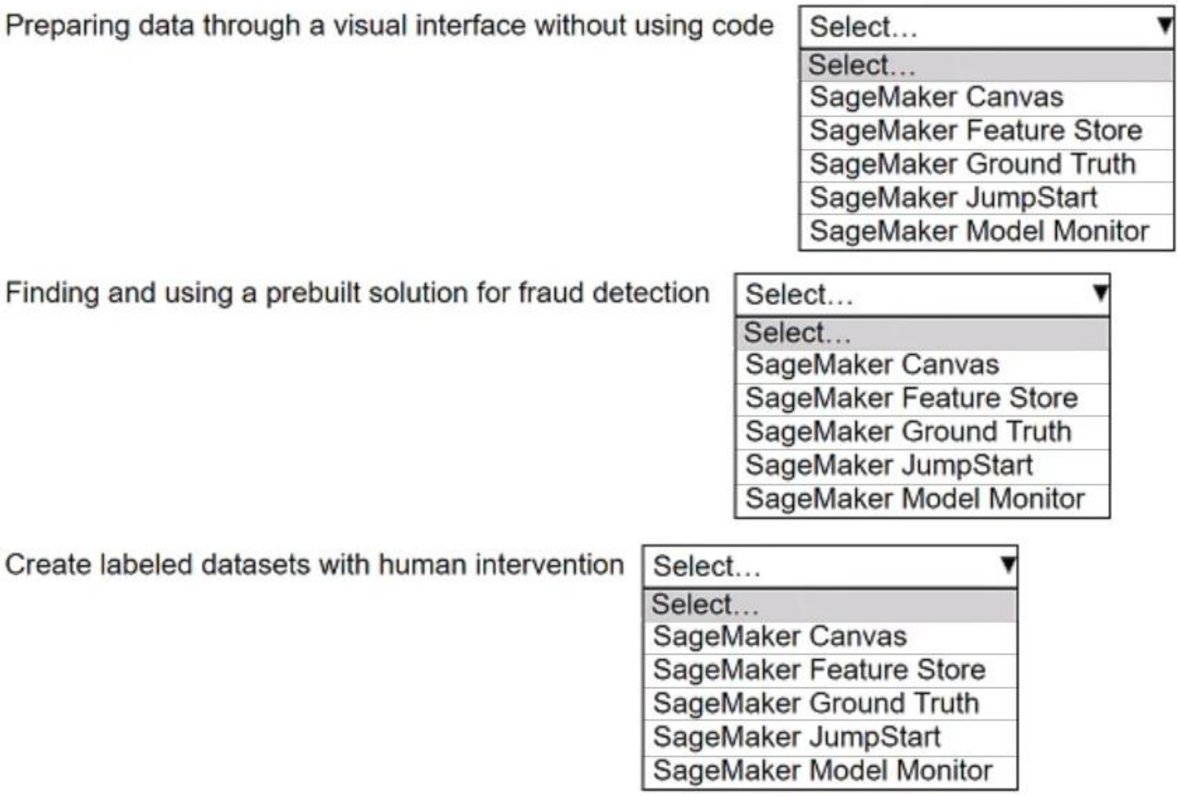
Explanation/Reference:
Explanation:
Regression models are designed to predict continuous numerical values (for example, forecasting sales figures or estimating house prices), making them the appropriate choice for numeric predictions.

QUESTION 191
HOTSPOT

A company wants to use Amazon SageMaker features for various use cases.

Select the correct SageMaker feature from the following list for each use case. Each SageMaker feature should be selected one time or not at all.

Hot Area:



Preparing data through a visual interface without using code

Select…
- Select…
- SageMaker Canvas
- SageMaker Feature Store
- SageMaker Ground Truth
- SageMaker JumpStart
- SageMaker Model Monitor

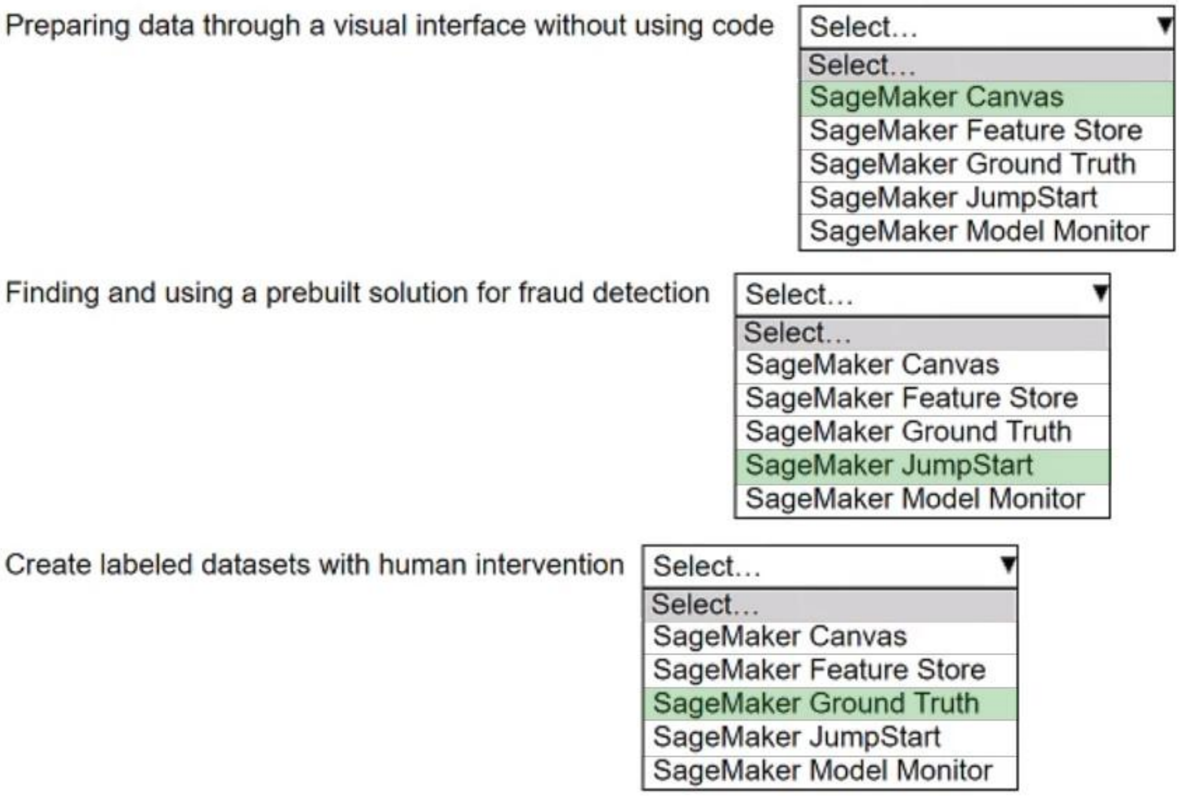
Finding and using a prebuilt solution for fraud detection

Select…
- Select…
- SageMaker Canvas
- SageMaker Feature Store
- SageMaker Ground Truth
- SageMaker JumpStart
- SageMaker Model Monitor

Create labeled datasets with human intervention

Select…
- Select…
- SageMaker Canvas
- SageMaker Feature Store
- SageMaker Ground Truth
- SageMaker JumpStart
- SageMaker Model Monitor

Correct Answer:



Preparing data through a visual interface without using code

Select…

- Select…
- [✓ SageMaker Canvas]
- SageMaker Feature Store
- SageMaker Ground Truth
- SageMaker JumpStart
- SageMaker Model Monitor

Finding and using a prebuilt solution for fraud detection

Select…

- Select…
- SageMaker Canvas
- SageMaker Feature Store
- SageMaker Ground Truth
- [✓ SageMaker JumpStart]
- SageMaker Model Monitor

Create labeled datasets with human intervention

Select…

- Select…
- SageMaker Canvas
- SageMaker Feature Store
- [✓ SageMaker Ground Truth]
- SageMaker JumpStart
- SageMaker Model Monitor

Section: (none)

Explanation/Reference:

QUESTION 192
What is the purpose of vector embeddings in a large language model (LLM)?

A. Splitting text into manageable pieces of data
B. Grouping a set of characters to be treated as a single unit
C. Providing the ability to mathematically compare texts
D. Providing the count of every word in the input

Correct Answer: C
Section: (none)

Explanation/Reference:
Explanation:
Embeddings convert text into high-dimensional vectors that capture semantic relationships, allowing you to compute distances or similarities between pieces of text for tasks like retrieval or clustering.

QUESTION 193
A company wants to fine-tune a foundation model (FM) by using AWS services. The company needs to ensure that its data stays private, safe, and secure in the source AWS Region where the data is stored.

Which combination of steps will meet these requirements MOST cost-effectively? (Choose two.)

A. Host the model on premises by using AWS Outposts.
B. Use the Amazon Bedrock API.
C. Use AWS PrivateLink and a VPC.
D. Host the Amazon Bedrock API on premises.
E. Use Amazon CloudWatch logs and metrics.

Correct Answer: BC
Section: (none)

Explanation/Reference:
Explanation:
Use the Amazon Bedrock API: You'll call Bedrock's managed fine-tuning endpoints directly in your AWS Region, so your data never leaves that region.

Use AWS PrivateLink and a VPC: Front Bedrock API traffic through a VPC endpoint via PrivateLink to keep all network traffic on the AWS backbone and within your account's private network.

QUESTION 194
A financial company uses AWS to host its generative AI models. The company must generate reports to show adherence to international regulations for handling sensitive customer data.

Which AWS service meets these requirements?

A. Amazon Macie
B. AWS Artifact
C. AWS Secrets Manager
D. AWS Config

Correct Answer: B
Section: (none)

Explanation/Reference:
Explanation:
AWS Artifact provides on-demand access to AWS's compliance reports and certifications (for example, ISO, SOC, GDPR), enabling the company to demonstrate its generative AI workloads and data handling practices adhere to international regulatory requirements.