Course:	AWS Certified AI Practitioner Course
Assessment:	AWS AIF Practice Exam 1 (training mode)
Username:	Charles Dinakaran Kovilpillai Jayapandian
e-mail:	kjcharlesdinakaran@gmail.com

A company has fine-tuned a pre-existing model from Amazon Bedrock to enhance document summarization for an internal project. They now want to use the custom model through Amazon Bedrock for production purposes.

What should the company do to enable the use of the custom model?

○ Allocate Reserved Capacity for the custom model.
○ Host the custom model on an Amazon SageMaker endpoint for real-time predictions.
○ List the custom model in the Amazon SageMaker Model Registry.
○ Enable access to the custom model within Amazon Bedrock.

Correct answer

Enable access to the custom model within Amazon Bedrock.

Feedback

- Explanation:

  Enable access to the custom model within Amazon Bedrock: This is correct because once the model is fine-tuned, it must be made accessible within Amazon Bedrock to be used in production. Enabling access to the custom model ensures the company can use it for inference tasks like document summarization directly through Bedrock’s infrastructure.

Allocate Reserved Capacity for the custom model: This is incorrect because Amazon Bedrock automatically manages the necessary infrastructure for hosting models. There is no need to allocate reserved capacity manually. Bedrock provides the required scaling and infrastructure management.

Host the custom model on an Amazon SageMaker endpoint for real-time predictions: This is incorrect because Amazon Bedrock provides the hosting and inference capabilities needed to use the custom model. There is no need to deploy it on an Amazon SageMaker endpoint separately.

List the custom model in the Amazon SageMaker Model Registry: This is incorrect because while the Amazon SageMaker Model Registry helps manage model versions and deployments, it is not required when using Amazon Bedrock. Bedrock has its own system for managing and deploying models.

References:

https://aws.amazon.com/bedrock/

A company is utilizing the Amazon Titan foundation model (FM) via Amazon Bedrock. To improve the model's output, the company needs to incorporate data from its own private data sources.

What solution should the company implement to meet this requirement?

○ Use a different foundation model.
○ Choose a lower temperature value.
○ Create an Amazon Bedrock knowledge base.
○ Enable model invocation logging.

Correct answer

Create an Amazon Bedrock knowledge base.

Feedback

- Explanation:

  Create an Amazon Bedrock knowledge base is correct because creating a knowledge base in Amazon Bedrock allows the model to integrate private data sources. The knowledge base enables the foundation model to access relevant internal data during inference, which can improve the quality and relevance of the model's responses.

  Use a different foundation model: This is incorrect because the need here is to incorporate private data, not to switch models. Using a different foundation model will not address the requirement to supplement the model with the company's internal data.

  Choose a lower temperature value: This is incorrect because the temperature parameter affects the randomness of the model's output, not its data sources. Lowering the temperature can make responses more focused, but it does not enable the use of private data in the model's processing.

  Enable model invocation logging: This is incorrect because model invocation logging helps in tracking and auditing model usage, but it does not help incorporate private data into the model's output. Logging does not address the requirement for using internal data sources during inference.

  Reference:

  https://aws.amazon.com/bedrock/

A company is building a generative AI application with Amazon Bedrock and wants to understand how much data it can include in a single prompt.

What factor should the company consider?

○ Temperature setting
○ Context window size
○ Maximum inference batch
○ Model architecture

Correct answer

Context window size

Feedback

- Explanation:

  The context window size defines how much input the model can process in a single prompt. It determines the maximum number of tokens (words or characters) that can be included in the input for the model to process.

  Temperature setting: This is incorrect because the temperature setting controls the level of randomness in the model's output, influencing how deterministic or creative the responses are. It doesn't affect the amount of data that can be included in a prompt.

  Maximum inference batch: This is incorrect because batch size refers to the number of inputs processed together during inference, but it does not influence how much information a single prompt can contain. It relates more to processing efficiency rather than input size.

  Model architecture: This is incorrect because model architecture describes the structure and components of the neural network, which affect the model's performance and capabilities, but not how much data can be provided in one prompt.

  References:

  https://aws.amazon.com/bedrock/

A media company is building a news summarization tool using Amazon Bedrock to generate concise summaries of trending news articles. The company wants to ensure that the generated content avoids misinformation and inappropriate topics.

Which AWS service or feature will help meet this requirement?

○ Amazon Rekognition
○ Amazon Bedrock playgrounds
○ Guardrails for Amazon Bedrock
○ Agents for Amazon Bedrock

Correct answer

Guardrails for Amazon Bedrock

Feedback

- Explanation:

  Guardrails can be used to ensure the generated content adheres to certain standards, such as avoiding misinformation, inappropriate topics, or harmful content. In this scenario, guardrails would help ensure that the news summaries are both accurate and appropriate for the audience.

  Amazon Rekognition: This is incorrect because Amazon Rekognition is used for image and video analysis, such as facial recognition or object detection, and is not related to controlling the appropriateness of text content.

  Amazon Bedrock playgrounds: This is incorrect because Amazon Bedrock playgrounds provide an environment for testing and experimenting with foundation models, but they do not offer specific tools to ensure content appropriateness or control over generated outputs.

  Agents for Amazon Bedrock: This is incorrect because Agents for Amazon Bedrock help orchestrate complex multi-step tasks using foundation models, but they do not address the need to filter or restrict the type of content being generated.

  References:

  https://aws.amazon.com/bedrock/guardrails/

Which option is a use case for generative AI models?

○ Improving network security by using intrusion detection systems

○ Creating photorealistic images from text descriptions for digital marketing

○ Enhancing database performance by using optimized indexing

○ Analyzing financial data to forecast stock market trends

Correct answer

Creating photorealistic images from text descriptions for digital marketing

Feedback

- Explanation:

  Generative AI models excel at creating new content, such as images, text, or audio. One of the primary applications of these models is generating visual content from text prompts, which is highly valuable in fields like digital marketing for producing creative, unique visuals.

  Improving network security by using intrusion detection systems: This is incorrect because intrusion detection systems rely on monitoring patterns and anomalies, which is more suited for traditional machine learning models, not generative AI. Generative AI is typically not used for security applications like detecting network intrusions.

  Enhancing database performance by using optimized indexing: This is incorrect because improving database performance involves algorithms and techniques focused on query optimization and indexing, which are not tasks suited for generative AI models. Generative AI is used for creating new data, not for database optimization.

  Analyzing financial data to forecast stock market trends: This is incorrect because predicting stock market trends typically involves statistical models or machine learning algorithms that analyze historical data patterns. Generative AI models are not designed for this kind of data forecasting, but for creating new content like images, text, or even audio.

  References:

  https://aws.amazon.com/ai/generative-ai/

A business is developing a machine learning model to analyze large archived datasets. These datasets are several gigabytes in size, and the company does not require immediate access to the predictions.

What Amazon SageMaker inference option should the company choose?

○ Batch transform

○ Real-time inference

○ Serverless inference

○ Asynchronous inference

Correct answer

Batch transform

Feedback

- Explanation:

  Batch transform is the ideal solution when you need to perform inference on large datasets without needing real-time results. It allows for processing data in bulk and is optimized for situations where immediate model predictions aren't required, making it suitable for analyzing multiple gigabytes of archived data.

  Real-time inference: This is incorrect because real-time inference is designed for low-latency scenarios where predictions are needed almost immediately after a request is made. This option is not necessary when there is no need for instant access to predictions.

  Serverless inference: This is incorrect because serverless inference is best suited for sporadic workloads where the model needs to scale up and down without managing infrastructure. It’s not the best option for processing large, non-urgent datasets in bulk.

  Asynchronous inference: This is incorrect because asynchronous inference is intended for scenarios where you need to handle large payloads with a delay, but it’s more focused on single prediction requests that take time to process, rather than bulk processing like batch transform.

  References:

  https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html
  Sorry, that is not correct.

In the context of generative AI models, what are tokens?

○ Tokens are the smallest units of text, such as words or subwords, that a generative model processes during input and output.

○ Tokens are the numerical encodings of concepts or terms that a generative model uses in its computations.

○ Tokens are the model’s learned parameters that are adjusted during fine-tuning for different tasks.

○ Tokens are the commands or instructions that guide the model to produce a specific response.

Correct answer

Tokens are the smallest units of text, such as words or subwords, that a generative model processes during input and output.

Feedback

- Explanation:

  In generative AI, tokens are the core building blocks that represent segments of text. These units could be full words, subwords, or even characters, depending on the model. The model processes input as sequences of tokens to generate predictions or responses.

  Tokens are the numerical encodings of concepts or terms that a generative model uses in its computations: This is incorrect because tokens are not the numerical encodings themselves. They are converted into numerical representations (embeddings), which the model uses during its computations.

  Tokens are the model’s learned parameters that are adjusted during fine-tuning for different tasks: This is incorrect because tokens are not parameters of the model. Parameters refer to the internal weights of the model, whereas tokens are the pieces of text that the model uses as input.

  Tokens are the commands or instructions that guide the model to produce a specific response: This is incorrect because tokens are not commands or instructions but the basic elements (words or subwords) processed by the model. Prompts, not tokens, represent the specific instructions given to the model.

  References:

  https://aws.amazon.com/ai/generative-ai/

A company is developing an AI application that uses language models for inference on edge devices. The company needs the lowest possible latency for real-time predictions.

Which approach should the company take?

○ Deploy optimized small language models (SLMs) on edge devices.

○ Deploy large-scale language models (LLMs) on edge devices with sufficient processing power.

○ Use cloud-hosted small language models (SLMs) through an API for communication with edge devices.

○ Use cloud-hosted large language models (LLMs) through an API for real-time predictions.

Correct answer

Deploy optimized small language models (SLMs) on edge devices.

Feedback

- Explanation:

  This is correct because deploying small, optimized models directly onto edge devices minimizes latency by avoiding network calls. Small models are specifically designed to run efficiently on devices with limited computational power, ensuring real-time performance.

  Deploy large-scale language models (LLMs) on edge devices with sufficient processing power: This is incorrect because large language models require significantly more memory and compute resources, which are typically not available on edge devices. Even if deployed, they would result in higher latency due to slower inference times.

  Use cloud-hosted small language models (SLMs) through an API for communication with edge devices: This is incorrect because even though small models require fewer resources, using a cloud API introduces network latency, which can delay predictions compared to running models locally on the device.

Use cloud-hosted large language models (LLMs) through an API for real-time predictions: This is incorrect because large models are computationally intensive, and network latency from communicating with a cloud-hosted LLM would increase the time required to get predictions, failing to meet the requirement for low latency.

References:

https://aws.amazon.com/edge/

Which feature of Amazon OpenSearch Service enables companies to create applications that use vector-based search functionality?

○ Seamless integration with Amazon S3 for data storage.

○ Capabilities for handling location-based queries and geospatial data.

○ Advanced vector indexing and similarity search for high-dimensional data.

○ Real-time processing of incoming data streams for immediate insights.

Correct answer

Advanced vector indexing and similarity search for high-dimensional data.

Feedback

- Explanation:

  Amazon OpenSearch Service provides vector search capabilities, which are essential for building vector databases. These features allow companies to search for similar items based on vector representations, often used in applications like recommendation engines and AI-driven search.

  Seamless integration with Amazon S3 for data storage: This is incorrect because integration with Amazon S3 is primarily used for storing large volumes of data or backups, not for managing or searching vector data.

  Capabilities for handling location-based queries and geospatial data: This is incorrect because geospatial queries are used to manage and query geographic data, not vectors. These features help with map-based applications but don't support vector search.

Real-time processing of incoming data streams for immediate insights: This is incorrect because real-time analysis is related to processing and analyzing streaming data as it arrives. While useful for certain analytics applications, it doesn’t enable vector-based search functionality.

References:

https://aws.amazon.com/opensearch-service/

A company needs to visualize the total sales of its best-performing products across different retail outlets over the last year. The company wants an automated solution to create these graphs based on its data.

Which AWS service should the company use?

○ Amazon QuickSight Q for automated insights
○ Amazon SageMaker for custom model deployment
○ Amazon EC2 for data processing and reporting
○ AWS Glue for data visualization

Correct answer

Amazon QuickSight Q for automated insights

Feedback

- Explanation:

  Amazon QuickSight Q is a natural language query tool that allows users to generate graphs and reports automatically by asking questions in plain language. It’s specifically designed to create visualizations like sales reports without manual effort.

  Amazon SageMaker for custom model deployment: This is incorrect because Amazon SageMaker is used to build, train, and deploy machine learning models. It’s not intended for generating sales graphs or creating automated visualizations based on business data.

Amazon EC2 for data processing and reporting: This is incorrect because Amazon EC2 provides compute resources for running virtual machines, but it doesn’t offer tools for data visualization or graph generation. The company would still need a separate service like QuickSight to generate graphs.

AWS Glue for data visualization: This is incorrect because AWS Glue is a data integration service designed for ETL (extract, transform, load) processes. It is not a visualization tool, and it cannot directly generate graphs or reports from data.

References:

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

A company is developing a machine learning model using Amazon SageMaker and needs a solution to store and share feature sets across different teams for collaborative model building.

Which Amazon SageMaker feature should the company use?

○ Amazon SageMaker Feature Store
○ Amazon SageMaker Data Wrangler
○ Amazon SageMaker Clarify
○ Amazon SageMaker Model Registry

Correct answer

Amazon SageMaker Feature Store

Feedback

- Explanation:

  SageMaker Feature Store is designed to allow teams to store, manage, and share features (attributes or variables) in a central repository. This ensures consistency across models and helps teams collaborate by reusing the same features across multiple projects.

Amazon SageMaker Data Wrangler: This is incorrect because Data Wrangler is used for data transformation and preparation, helping users to clean and structure data before training models. It does not provide a mechanism to store and share features across teams.

Amazon SageMaker Clarify: This is incorrect because Clarify is focused on detecting bias in machine learning models and ensuring explainability. It is not used for storing or sharing features between teams.

Amazon SageMaker Model Registry: This is incorrect because the Model Registry is designed for managing and versioning machine learning models, not for storing or sharing feature sets used during model development.

References:

https://docs.aws.amazon.com/sagemaker/latest/dg/feature-store.html

A business uses Amazon SageMaker to run its machine learning pipeline in a production environment. The company processes large datasets, sometimes reaching 1 GB in size, with processing times that can take up to an hour. To support its operations, the company requires low-latency predictions.

Which Amazon SageMaker inference option should the company choose?

○ Real-time inference
○ Serverless inference
○ Asynchronous inference
○ Batch transform

Correct answer

Real-time inference

Feedback

- Explanation:

  Real-time inference is designed for scenarios where low-latency responses are needed. It is ideal when predictions need to be generated immediately upon receiving input data, making it suitable for use cases requiring near real-time results, even with large datasets.

  Serverless inference: This is incorrect because serverless inference is optimized for intermittent workloads that don’t require low-latency predictions. While it’s cost-effective for occasional requests, it doesn't meet the requirement for near real-time performance when processing large datasets.

  Asynchronous inference: This is incorrect because asynchronous inference is intended for situations where the input size is large or the processing time is lengthy, but real-time predictions are not required. It’s useful when results can be delayed, but it doesn't support the company's need for near real-time latency.

  Batch transform: This is incorrect because batch transform is used for processing large datasets in batches without a focus on real-time results. It is better suited for use cases where predictions are processed in bulk and not required immediately after input.

  References:

  https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html

Which method is used to evaluate the accuracy of a foundation model (FM) applied to image classification tasks?

○ Calculate the total cost of resources consumed by the model.

○ Measure the model's accuracy using a benchmark dataset specifically designed for image classification.

○ Count the total number of neural network layers in the model architecture.

○ Assess the color accuracy of the images that the model processes.

Correct answer

Measure the model's accuracy using a benchmark dataset specifically designed for image classification.

Feedback

- Explanation:

  Evaluating the accuracy of a foundation model involves comparing its predictions to known labels from a predefined dataset. Benchmark datasets are specifically curated for tasks like image classification to assess how well a model performs against a standardized set of images.

  Calculate the total cost of resources consumed by the model: This is incorrect because resource consumption measures efficiency, not model accuracy. Evaluating costs is important for financial considerations, but it doesn’t reflect how well the model classifies images.

  Count the total number of neural network layers in the model architecture: This is incorrect because the number of layers in the neural network doesn't directly measure the model’s performance or accuracy. While deeper networks can often be more powerful, accuracy is evaluated through performance on actual data.

  Assess the color accuracy of the images that the model processes: This is incorrect because color accuracy pertains to the visual quality or fidelity of an image. It has no direct relationship to the model's ability to classify images correctly.

  References:

  https://docs.aws.amazon.com/machine-learning/

An AI researcher is developing a model to generate synthetic faces for a facial recognition application. During training, they realize that the dataset contains significantly fewer samples of certain ethnic groups, leading to biased model outputs.

Which technique can the researcher use to address this bias?

○ Data augmentation for imbalanced classes

○ Model monitoring for accuracy drift

○ Retrieval Augmented Generation (RAG)

○ Edge detection for image processing

Correct answer

Data augmentation for imbalanced classes

Feedback

- Explanation:

  Data augmentation helps balance the dataset by creating new samples from underrepresented groups. By applying transformations such as rotation, cropping, or flipping to the existing images, the researcher can reduce the bias in the dataset and improve the fairness of the model's outputs.

  Model monitoring for accuracy drift: This is incorrect because monitoring for accuracy drift helps track model performance over time but does not correct bias in the training data. It identifies performance issues but doesn't resolve them.

  Retrieval Augmented Generation (RAG): This is incorrect because RAG is a method for improving generative models by retrieving relevant information from external sources, which doesn't help balance the data or correct the underlying bias in this scenario.

  Edge detection for image processing: This is incorrect because edge detection is a technique used to identify boundaries in images, which is unrelated to addressing dataset bias. It focuses on feature extraction, not on balancing data or mitigating bias.

  References:

  https://docs.aws.amazon.com/machine-learning/

A company is utilizing machine learning models for specialized tasks in a specific domain. To save time and resources, the company prefers to modify existing pre-trained models instead of building new ones from scratch.

Which machine learning approach should the company use?

○ Increase the number of training iterations.

○ Apply transfer learning.

○ Reduce the number of training iterations.

○ Implement unsupervised learning.

Correct answer

Apply transfer learning.

Feedback

- Explanation:

  Transfer learning allows a company to leverage pre-trained models and adapt them to new, related tasks. Instead of training a model from scratch, the pre-trained model's knowledge is fine-tuned for the new task, significantly reducing the training time and required data.

  Increase the number of training iterations: This is incorrect because increasing the number of epochs (iterations) only affects the training of the model, but it doesn't allow you to reuse knowledge from pre-trained models. This would require starting with a model from scratch, which the company wants to avoid.

  Reduce the number of training iterations: This is incorrect because reducing the number of epochs does not address the need for reusing pre-trained models. While this may shorten training time, it doesn't leverage pre-existing model knowledge, which is the main objective.

  Implement unsupervised learning: This is incorrect because unsupervised learning involves training a model without labeled data. The company's goal is to adapt pre-trained models for new tasks, which typically involves supervised or fine-tuned learning rather than unsupervised techniques.

  References:

  https://aws.amazon.com/ai/machine-learning/

A company has customized a foundation model (FM) with Amazon Bedrock to handle customer support queries. The company now wants to test the model's ability to respond accurately to new types of queries. They need to upload and store a new dataset that Amazon Bedrock can access for the validation process.

Which AWS service should they use for storing this dataset?

- ○ Amazon S3
- ○ Amazon FSx
- ○ Amazon RDS
- ○ AWS Snowball Edge

Correct answer

Amazon S3

Feedback

- Explanation:

  Amazon S3 is a scalable, cost-effective object storage service that is widely used for storing large datasets. It is commonly integrated with machine learning services, including Amazon Bedrock, for tasks like training and validating models. S3 allows easy access to datasets needed for model validation.

  Amazon FSx: This is incorrect because Amazon FSx is a fully managed file system service designed for specific high-performance workloads like Windows file storage or Lustre for HPC (high-performance computing). It is not typically used for general-purpose dataset storage for model validation.

  Amazon RDS: This is incorrect because Amazon RDS is a managed relational database service used for structured data and transactional workloads. While RDS is useful for database-driven applications, it is not suitable for storing large datasets required for model validation in machine learning.

  AWS Snowball Edge: This is incorrect because AWS Snowball Edge is a physical device used for transferring large volumes of data in or out of AWS when network transfer is impractical. It is used for data migration, not for directly storing and validating datasets in cloud-based machine learning services.

  References:

  https://aws.amazon.com/s3/

A law firm is deploying a large language model (LLM) to automate the drafting of legal documents. The firm wants to ensure the model is developed responsibly to minimize risks, such as biased outputs.

Which two actions should the firm take? (Select TWO.)

- ○ Conduct fairness evaluations on the model's outputs.
- ○ Adjust the model’s temperature to increase output variety.
- ○ Retrain the model with diverse datasets to reduce bias.
- ○ Limit training epochs to prevent overfitting.
- ○ Apply regularization techniques to tune hyperparameters.

Correct answers

- Conduct fairness evaluations on the model's outputs.
- Retrain the model with diverse datasets to reduce bias.

Feedback

- Explanation:

  Ensuring the model generates unbiased outputs is crucial for responsible AI deployment, especially in sensitive tasks like legal drafting. Using diverse data helps the model generalize better and reduces bias in its predictions.

  Adjust the model’s temperature to increase output variety: Adjusting the temperature controls output randomness but does not address bias or fairness.

  Limit training epochs to prevent overfitting: While useful for preventing overfitting, this doesn’t address bias or fairness concerns in model outputs.

  Apply regularization techniques to tune hyperparameters: Regularization can improve model performance but does not directly address fairness or bias.

  References:

  https://aws.amazon.com/ai/responsible-ai/

A company has developed a chatbot that responds to user queries with images. The company needs to ensure that the chatbot avoids displaying inappropriate or offensive images.

Which approach should the company take to achieve this?

- ○ Use content moderation tools to filter image responses.
- ○ Retrain the model using larger, publicly available datasets.
- ○ Implement regular performance checks on the chatbot.
- ○ Enable automatic updates from user feedback.

Correct answer

Use content moderation tools to filter image responses.

Feedback

- Explanation:

  Content moderation tools can scan and block images that are inappropriate, ensuring that only safe and relevant images are returned by the chatbot.

  Retrain the model using larger, publicly available datasets: This is incorrect because retraining with a public dataset will not directly address the issue of inappropriate images unless the dataset itself is highly curated.

  Implement regular performance checks on the chatbot: This is incorrect because performance checks assess the chatbot's overall functioning but won't specifically prevent inappropriate images from being displayed.

  Enable automatic updates from user feedback: This is incorrect because relying on user feedback is reactive. The company needs a proactive solution like moderation to prevent inappropriate content from being displayed in the first place.

  References:

  https://aws.amazon.com/rekognition/content-moderation/

A company is developing an application using Amazon Bedrock. With a limited budget, the company seeks a flexible pricing model that does not require long-term commitments.

Which Amazon Bedrock pricing model is most suitable for this requirement?

○ On-Demand
○ Reserved Instances
○ Pay-per-request
○ Subscription-based

Correct answer

On-Demand

Feedback

- Explanation:

  The On-Demand pricing model allows the company to pay only for the resources it uses, with no long-term commitments. This offers flexibility and is ideal for companies with a limited budget who want to avoid upfront costs.

  Reserved Instances: This is incorrect because reserved instances require a commitment to a long-term contract, which contradicts the company's need for flexibility and a limited budget.

  Pay-per-request: This is incorrect because while pay-per-request may exist for some services, it is not the standard pricing model for Amazon Bedrock. On-Demand is the appropriate model for flexible usage.

  Subscription-based: This is incorrect because subscription-based pricing often involves recurring charges, which may not provide the cost control and flexibility the company is looking for.

  References:

  https://aws.amazon.com/bedrock/pricing/

Which functionality is provided by Amazon SageMaker Clarify?

○ Implements a Retrieval Augmented Generation (RAG) pipeline

○ Tracks the performance of machine learning models in production

○ Captures key metadata about machine learning models

○ Detects possible bias during the data preparation phase

Correct answer

Detects possible bias during the data preparation phase

Feedback

- Explanation:

  Amazon SageMaker Clarify helps identify and mitigate bias in machine learning datasets and models, ensuring fairness during both data preparation and model training.

  Implements a Retrieval Augmented Generation (RAG) pipeline: This is incorrect because RAG is a method used to improve generative models by retrieving relevant information, and SageMaker Clarify does not handle RAG workflows.

  Tracks the performance of machine learning models in production: This is incorrect because tracking model performance in production is handled by tools like SageMaker Model Monitor, not SageMaker Clarify.

  Captures key metadata about machine learning models: This is incorrect because documenting metadata about models is a feature of SageMaker Model Cards, not SageMaker Clarify.

  References:

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

A company wants to train a large language model (LLM) using only its private data. In addition to performance, the company is focused on minimizing the environmental footprint during training.

Which Amazon EC2 instance type should the company choose to achieve this?

- ○ Amazon EC2 M series
- ○ Amazon EC2 Inf series
- ○ Amazon EC2 P series
- ○ Amazon EC2 Trn series

Correct answer

Amazon EC2 Trn series

Feedback

- Explanation:

  The EC2 Trn series (using AWS Trainium processors) is designed to optimize the energy efficiency of large-scale machine learning training tasks. It consumes less power while delivering high-performance training for LLMs, reducing the environmental impact.

  Amazon EC2 M series: This is incorrect because the M series is a general-purpose instance type and is not optimized for energy efficiency or machine learning model training, making it less suitable for minimizing environmental impact.

  Amazon EC2 Inf series: This is incorrect because the Inf series is optimized for inference tasks rather than training. While it can provide efficient inference, it is not designed for the training phase, especially for LLMs.

  Amazon EC2 P series: This is incorrect because the P series is optimized for high-performance GPU workloads, but it is less efficient in terms of energy consumption compared to the Trn series, which is built specifically for efficient ML training.

  References:

  https://aws.amazon.com/machine-learning/trainium/

A company is developing a machine learning model. The company has gathered new data and is analyzing it by generating correlation matrices, calculating statistics, and visualizing patterns in the dataset.

What stage of the machine learning pipeline is the company in?

- ○ Data cleansing
- ○ Feature extraction
- ○ Exploratory data analysis
- ○ Model evaluation

Correct answer

Exploratory data analysis

Feedback

- Explanation:

  The company is focused on understanding the data by visualizing relationships and calculating statistical measures. These activities are key components of exploratory data analysis (EDA), which helps identify patterns and guide further steps in the pipeline.

  Data cleansing: This is incorrect because data cleansing focuses on correcting or removing inaccurate records from the dataset. The scenario describes analysis and visualization, not the cleaning of data.

  Feature extraction: This is incorrect because feature extraction refers to creating or selecting specific attributes (features) to improve the model's performance. The company is still in the data exploration phase and hasn’t started engineering or extracting features yet.

  Model evaluation: This is incorrect because model evaluation is the process of assessing the performance of a trained model. The company has not yet built or trained a model, so evaluation is not relevant at this stage.

  References:

  https://aws.amazon.com/ai/machine-learning/

A company is using a large language model (LLM) on Amazon Bedrock for sentiment analysis. The company wants to classify text passages as positive or negative.

Which prompt engineering strategy should the company use?

○ Provide examples of text passages with their corresponding positive or negative labels, followed by the new passage to classify.

○ Include a thorough explanation of sentiment analysis techniques and how LLMs work in the prompt.

○ Input the new text passage without any examples or context and ask the model to classify it.

○ Include the new text passage along with examples of other tasks, like text summarization or translation, in the prompt.

Correct answer

Provide examples of text passages with their corresponding positive or negative labels, followed by the new passage to classify.

Feedback

- Explanation:

  Including labeled examples in the prompt helps guide the LLM by demonstrating how to classify sentiments, improving the model's accuracy in identifying whether a new passage is positive or negative.

  Include a thorough explanation of sentiment analysis techniques and how LLMs work in the prompt: This is incorrect because providing an explanation of the underlying techniques does not help the LLM in performing the specific task of sentiment classification. The model responds better to concrete examples rather than theoretical explanations.

  Input the new text passage without any examples or context and ask the model to classify it: This is incorrect because not providing examples or context decreases the model's ability to accurately perform sentiment classification, especially if it lacks prior instructions on how to approach the task.

Include the new text passage along with examples of other tasks, like text summarization or translation, in the prompt: This is incorrect because mixing different tasks like summarization or translation dilutes the model’s focus on sentiment analysis and can lead to incorrect results.

References:

https://aws.amazon.com/bedrock/

An e-commerce company is using Amazon Bedrock to power a product recommendation system. The company wants to ensure that the system does not generate recommendations based on customers' sensitive personal information, such as payment details or personal addresses. Additionally, the company requires notifications when any policy violations occur.

Which solution meets these requirements?

○ Use Amazon Macie to scan the recommendation system’s output for sensitive data and configure alerts for policy violations.

○ Set up AWS CloudTrail to monitor the system’s output and notify the company when sensitive data is detected.

○ Implement Guardrails for Amazon Bedrock to prevent sensitive content from being included in recommendations. Configure Amazon CloudWatch alarms for policy violation notifications.

○ Enable Amazon SageMaker Model Monitor to track data quality and notify the company if sensitive data is found in the training data.

Correct answer
Implement Guardrails for Amazon Bedrock to prevent sensitive content from being included in recommendations. Configure Amazon CloudWatch alarms for policy violation notifications.

Feedback

- Explanation:

Guardrails can help filter out sensitive data from being used or included in the system's responses. Combined with CloudWatch alarms, the company can receive alerts whenever violations occur.

Use Amazon Macie to scan the recommendation system's output for sensitive data and configure alerts for policy violations: This is incorrect because Amazon Macie is primarily used for scanning data in S3 buckets for sensitive information, not for monitoring model outputs in real-time.

Set up AWS CloudTrail to monitor the system's output and notify the company when sensitive data is detected: This is incorrect because CloudTrail is used for logging API calls and activities in AWS, not for detecting sensitive data in real-time model outputs.

Enable Amazon SageMaker Model Monitor to track data quality and notify the company if sensitive data is found in the training data: This is incorrect because SageMaker Model Monitor is designed to track model performance and data drift, not for filtering sensitive information in real-time outputs or ensuring compliance.

References:

https://aws.amazon.com/bedrock/guardrails/

A media company is using a large language model (LLM) on Amazon Bedrock to summarize movie reviews. The company wants the model to generate concise summaries that capture the overall sentiment of the review.

Which prompt engineering strategy should the company use?

○ Provide detailed instructions on how LLMs generate summaries and sentiment analysis.

○ Include a few examples of movie reviews with their corresponding summaries before providing the new review to summarize.

○ Provide the new review without any context or examples and ask the model to summarize it.

○ Include instructions for other tasks, such as generating a product description or categorizing the review, along with the summarization task.

Correct answer

Include a few examples of movie reviews with their corresponding summaries before providing the new review to summarize.

Feedback

- Explanation:

  Providing a few examples of reviews with their corresponding summaries helps guide the model by showing it how to perform the task. This is known as few-shot learning and helps the LLM generate more accurate summaries based on the patterns in the provided examples.

  Provide detailed instructions on how LLMs generate summaries and sentiment analysis: This is incorrect because giving the model detailed technical explanations doesn’t guide it in generating the specific type of summary required.

  Provide the new review without any context or examples and ask the model to summarize it: This is incorrect because without context or examples, the model may not generate consistent or accurate summaries.

  Include instructions for other tasks, such as generating a product description or categorizing the review, along with the summarization task: This is incorrect because mixing different tasks in a single prompt can confuse the model and lead to poor-quality outputs for the specific task of summarization.

  References:

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

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?

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

Correct answer

Use domain adaptation fine-tuning to adapt the FM to complex scientific terms.

Feedback

- Explanation:

  Lowering the temperature value makes the model’s output more deterministic and consistent. A lower temperature reduces randomness in the generated responses, ensuring the same input yields more similar outputs.

  Raise the temperature value: This is incorrect because increasing the temperature introduces more randomness, making the model's responses less predictable and more varied.

  Shorten the output token limit: This is incorrect because reducing the token limit affects the length of the output, not the consistency of the responses. It limits how much text the model can generate but does not control randomness.

  Extend the maximum sequence length: This is incorrect because increasing the sequence length allows the model to produce longer outputs but does not directly affect how consistent or varied the responses are.

  References:

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

A company wants to improve developer productivity and streamline software development using generative AI. The company plans to use Amazon Q Developer.

What functionality does Amazon Q Developer offer to help meet these goals?

○ Generate code snippets, track references, and manage open source licenses.

○ Run applications without needing to provision or manage servers.

○ Enable voice-activated coding and natural language search capabilities.

○ Convert audio files into text documents using machine learning models.

Correct answer

Generate code snippets, track references, and manage open source licenses.

Feedback

- Explanation:

  Amazon Q Developer is designed to assist developers by automating tasks like creating code snippets, tracking dependencies, and ensuring compliance with open-source licenses. This helps improve developer productivity by reducing manual effort.

  Run applications without needing to provision or manage servers: This is incorrect because running applications without server management refers to AWS Lambda or serverless computing, not Amazon Q Developer.

  Enable voice-activated coding and natural language search capabilities: This is incorrect because Amazon Q Developer does not focus on voice commands or natural language search features, which are more associated with tools like Amazon Alexa.

  Convert audio files into text documents using machine learning models: This is incorrect because converting audio to text is typically done by Amazon Transcribe, not Amazon Q Developer, which is focused on developer tools and productivity.

  References:

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

A company is using Amazon SageMaker Studio notebooks to build and train machine learning models. The data is stored in an Amazon S3 bucket, and the company needs to manage the data flow between Amazon S3 and SageMaker Studio notebooks.

Which solution will meet this requirement?

- ○ Use Amazon Inspector to monitor SageMaker Studio.
- ○ Use Amazon Macie to track data flow in SageMaker Studio.
- ○ Configure SageMaker to use a VPC with an S3 VPC endpoint.
- ○ Configure SageMaker to use S3 Glacier Deep Archive for data access.

Correct answer

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

Feedback

- Explanation:

  Setting up a VPC with an S3 VPC endpoint allows secure and efficient access to data stored in Amazon S3 without using the public internet, ensuring smooth data flow between SageMaker and S3.

  Use Amazon Inspector to monitor SageMaker Studio: This is incorrect because Amazon Inspector is a security assessment service, not a service for managing data flow between S3 and SageMaker.

  Use Amazon Macie to track data flow in SageMaker Studio: This is incorrect because Amazon Macie is used for data security and privacy, specifically for identifying sensitive data in S3. It doesn't manage the flow of data between SageMaker and S3.

  Configure SageMaker to use S3 Glacier Deep Archive for data access: This is incorrect because S3 Glacier Deep Archive is used for long-term, low-cost storage of infrequently accessed data. It's not suitable for active data access and flow management between SageMaker and S3.

  References:

  https://docs.aws.amazon.com/sagemaker/

A media company has implemented a generative AI solution that uses large language models (LLMs) to automatically generate subtitles for video content in different languages. The company wants to assess the quality of the translations generated by the model.

Which model evaluation strategy should the company use?

○ Root mean squared error (RMSE)
○ Recall-Oriented Understudy for Gisting Evaluation (ROUGE)
○ Bilingual Evaluation Understudy (BLEU)
○ F1 score

Correct answer

Bilingual Evaluation Understudy (BLEU)

Feedback

- Explanation:

  BLEU is a widely used metric for evaluating the quality of machine-generated translations by comparing them to reference translations. It measures the accuracy of translation models, making it ideal for assessing subtitle generation.

  Root mean squared error (RMSE): This is incorrect because RMSE is used to measure the differences between predicted and observed values in regression models, not for evaluating text translation quality.

  Recall-Oriented Understudy for Gisting Evaluation (ROUGE): This is incorrect because ROUGE is primarily used for evaluating text summarization, not translation accuracy.

  F1 score: This is incorrect because the F1 score measures the balance between precision and recall in classification tasks, not for assessing translations in natural language processing.

  References:

  https://aws.amazon.com/what-is/nlp/

A financial institution has trained a large language model (LLM) on Amazon Bedrock using a dataset that contains sensitive financial records. The institution needs to ensure the model does not generate responses that reveal or are influenced by the confidential financial data.

What action should the institution take to prevent this?

○ Apply dynamic data masking to hide sensitive data in inference responses.

○ Encrypt the sensitive data in inference outputs using Amazon SageMaker.

○ Use AWS Key Management Service (AWS KMS) to encrypt the sensitive data within the model.

○ Delete the trained model, remove the sensitive financial data from the dataset, and retrain the model.

Correct answer

Delete the trained model, remove the sensitive financial data from the dataset, and retrain the model.

Feedback

- Explanation:

  Once a model is trained on sensitive data, the only way to ensure it doesn’t generate responses based on that data is to remove the sensitive data from the training set and retrain the model. This guarantees the model won't use confidential information in its responses.

  Apply dynamic data masking to hide sensitive data in inference responses: This is incorrect because masking hides specific fields in outputs but doesn't prevent the model from generating content influenced by the sensitive data.

  Encrypt the sensitive data in inference outputs using Amazon SageMaker: This is incorrect because encryption protects data in transit or at rest but doesn't stop the model from referencing the sensitive data during inference.

Use AWS Key Management Service (AWS KMS) to encrypt the sensitive data within the model: This is incorrect because encryption ensures data protection but doesn’t prevent the model from generating outputs that may still be influenced by the sensitive data.

References:

https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html

A logistics company has thousands of warehouse images and wants to automatically identify and classify different types of items stored in the images without manual effort.

Which strategy will help the company achieve this?

○ Anomaly detection
○ Object detection
○ Named entity recognition
○ Semantic segmentation

Correct answer

Object detection

Feedback

- Explanation:

  Object detection is a computer vision technique used to identify and classify multiple objects within an image. In this case, it can automatically identify and categorize different items stored in the warehouse.

  Anomaly detection: This is incorrect because anomaly detection is used to identify unusual patterns or outliers in data, not to recognize or classify objects in images.

  Named entity recognition: This is incorrect because named entity recognition (NER) is a natural language processing technique used to identify entities in text, not for identifying objects in images.

Semantic segmentation: This is incorrect because while semantic segmentation labels each pixel in an image to classify different parts of the image, it is more detailed than what is needed for simply identifying and categorizing items.

References:

https://docs.aws.amazon.com/sagemaker/latest/dg/algo-object-detection-tech-notes.html

An AI researcher is using an Amazon Bedrock base model to generate product descriptions for an e-commerce platform. The researcher needs to store logs of each model invocation, including input and output data, for later review and analysis.

What is the best strategy to meet this requirement?

○ Configure AWS CloudTrail to log the model’s input and output data.
○ Enable invocation logging in Amazon Bedrock to track inputs and outputs.
○ Use AWS Audit Manager to log the input and output data of the model.
○ Set up logging through Amazon EventBridge for capturing model responses.

Correct answer
Enable invocation logging in Amazon Bedrock to track inputs and outputs.

Feedback

- Explanation:

  Amazon Bedrock provides native invocation logging, which allows users to store input and output data for each invocation. This logging is essential for tracking model performance and ensuring data integrity during operations.

  Configure AWS CloudTrail to log the model’s input and output data: This is incorrect because AWS CloudTrail logs API activity and service-level actions, but it doesn’t log the specific inputs and outputs of model invocations.

Use AWS Audit Manager to log the input and output data of the model: This is incorrect because AWS Audit Manager is used for auditing and compliance checks, not for capturing real-time invocation logs of AI models.

Set up logging through Amazon EventBridge for capturing model responses: This is incorrect because Amazon EventBridge routes events between AWS services but is not intended for storing or managing detailed model invocation logs.

References:

https://docs.aws.amazon.com/bedrock/latest/userguide/model-invocation-logging.html

A healthtech startup has created a machine learning model that analyzes X-ray images to detect potential signs of illness. The company wants to deploy the model to production so that doctors can upload X-rays via a web application and receive predictions in real-time. The company prefers a solution that does not require managing underlying infrastructure.

Which solution should the company use?

○ Use Amazon SageMaker Serverless Inference to deploy the model.

○ Use Amazon CloudFront to serve the model for real-time predictions.

○ Use Amazon API Gateway to deploy the model and serve predictions.

○ Use AWS Batch to deploy the model for processing X-ray images.

Correct answer

Use Amazon SageMaker Serverless Inference to deploy the model.

Feedback

- Explanation:

  SageMaker Serverless Inference provides a fully managed, serverless environment for hosting and serving machine learning models. It allows the company to focus on deploying the model without managing any underlying infrastructure, which meets the company's needs.

Use Amazon CloudFront to serve the model for real-time predictions: This is incorrect because CloudFront is a content delivery network (CDN) used to deliver data with low latency but is not designed for hosting and serving machine learning models.

Use Amazon API Gateway to deploy the model and serve predictions: This is incorrect because API Gateway is used to expose APIs, but it doesn't host or serve machine learning models. It can be used alongside SageMaker but isn't sufficient on its own for model hosting.

Use AWS Batch to deploy the model for processing X-ray images: This is incorrect because AWS Batch is designed for batch processing large jobs, not for real-time inference or serving predictions for a web application.

References:

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

An education platform is developing a chatbot to help students with homework questions. The company has selected a foundation model (FM) but wants the chatbot’s responses to maintain an encouraging and educational tone.

What should the company do to achieve this?

○ Limit the token output to control the length of responses.

○ Refine the prompt to ensure the FM produces responses in the desired tone.

○ Use batch inference to process multiple student queries at once.

○ Increase the temperature to make responses more dynamic.

Correct answer

Refine the prompt to ensure the FM produces responses in the desired tone.

Feedback

- Explanation:

  Adjusting and refining the prompt can help guide the model to respond in a consistent, encouraging tone, which aligns with the company's goals for the chatbot.

  Limit the token output to control the length of responses: This is incorrect because limiting tokens affects response length, not the tone or style.

  Use batch inference to process multiple student queries at once: This is incorrect because batch inference is used for handling large-scale data and doesn't control the chatbot's tone.

  Increase the temperature to make responses more dynamic: This is incorrect because increasing temperature makes responses more random and creative, which could make them inconsistent with the desired tone.

  References:

  https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering-guidelines.html

A company wants to develop a large language model (LLM) application using Amazon Bedrock with customer data stored in Amazon S3. The company's security policy mandates that each team can only access data for their own customers.

Which solution will meet these requirements?

○ Create an Amazon Bedrock custom service role for each team that has access to only the team's customer data.

○ Assign one service role to all teams and use Amazon S3 event triggers to restrict data access based on customer information.

○ Use AWS Secrets Manager to encrypt customer data and allow each team to decrypt only their specific customer information.

○ Set up a shared Bedrock role and log data access with Amazon CloudWatch to monitor unauthorized access attempts.

Correct answer

Create an Amazon Bedrock custom service role for each team that has access to only the team's customer data.

Feedback

- Explanation:

  Creating a separate custom service role for each team ensures that access to customer data is restricted in accordance with the company's security policies. By using custom roles, each team only has permissions to access the specific data associated with their customers in Amazon S3.

  This strategy aligns with the principle of least privilege, providing granular control over data access and ensuring that teams do not have unauthorized access to another team's data. It also simplifies management, as each team can only perform operations within the bounds of their assigned role, preventing accidental or malicious access to restricted information.

  Assign one service role to all teams and use Amazon S3 event triggers to restrict data access based on customer information: This is incorrect because event triggers are not designed to control real-time access to specific data. A single role for all teams does not provide the necessary access control, and event triggers are more suited for initiating workflows, not for restricting data access dynamically.

  Use AWS Secrets Manager to encrypt customer data and allow each team to decrypt only their specific customer information: This is incorrect because AWS Secrets Manager is used for managing secrets, such as API keys or database credentials, rather than controlling access to large datasets. While encryption ensures data protection, it does not solve the problem of restricting access to customer-specific data in S3.

  Set up a shared Bedrock role and log data access with Amazon CloudWatch to monitor unauthorized access attempts: This is incorrect because logging access with CloudWatch provides visibility into unauthorized actions but does not prevent unauthorized access from happening in the first place. A shared role would still allow teams to access data they shouldn't, violating the security policy.

References:

https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies.html

A retail company has collected terabytes of customer purchase data but the data is not labeled. The company wants to segment its customers into groups for a targeted marketing campaign based on their purchasing patterns.

Which machine learning approach should the company use to achieve this?

○ Data clustering
○ Unsupervised learning
○ Semi-supervised learning
○ Deep reinforcement learning

Correct answer

Unsupervised learning

Feedback

- Explanation:

  Unsupervised learning is ideal for this task because it works with unlabeled data and can identify patterns in the data to group customers based on their purchasing behaviors. This method will allow the company to classify its customers into segments for targeted marketing.

  Data clustering: This is incorrect because clustering is a technique used in unsupervised learning but is not a learning methodology itself. The broader approach should be unsupervised learning.

  Semi-supervised learning: This is incorrect because semi-supervised learning works with a mix of labeled and unlabeled data. Since the company's data is entirely unlabeled, this method would not be the best fit.

Deep reinforcement learning: This is incorrect because deep reinforcement learning is focused on learning through interactions in an environment with rewards, not on data classification or segmentation of large, unlabeled datasets.

References:

https://aws.amazon.com/compare/the-difference-between-machine-learning-supervised-and-unsupervised/

A fintech company is building a generative AI application using Amazon Bedrock. The company wants to assess the costs related to generating inferences with a large language model (LLM).

Which factor will influence the inference costs?

○ Number of tokens processed

○ Model accuracy rate

○ Size of training dataset

○ Total memory used for training

Correct answer

Number of tokens processed

Feedback

- Explanation:

  Inference costs for large language models on Amazon Bedrock are driven by the number of tokens processed. Tokens represent pieces of the input or output text, and the more tokens involved in a single inference request, the higher the cost. Monitoring token usage helps control and manage costs effectively.

  Model accuracy rate: This is incorrect because the accuracy of the model does not directly influence the cost of inference, which is based on the number of tokens processed during requests.

Size of training dataset: This is incorrect because the cost of inference is unrelated to the size of the training dataset. It’s the inference stage that affects token usage and costs.

Total memory used for training: This is incorrect because memory usage during training impacts training costs, not inference costs. Inference costs are driven by token processing during real-time requests.

References:

https://aws.amazon.com/bedrock/pricing/

A healthcare organization is handling a large number of patient records in PDF format. As the volume of records continues to grow, the organization needs an automated system to convert these PDF documents into plain text format for integration into their electronic health record (EHR) system.

Which AWS service meets this requirement?

○ Amazon Personalize
○ Amazon Lex
○ Amazon Textract
○ Amazon Transcribe

Correct answer

Amazon Textract

Feedback

- Explanation:

  Amazon Textract is the ideal solution because it can automatically extract text, tables, and forms from PDFs and other scanned documents. This allows the healthcare organization to convert patient records into plain text, making it easier to integrate the data into their EHR system.

Amazon Personalize: This is incorrect because Amazon Personalize is used to provide personalized recommendations, not for extracting text from documents.

Amazon Lex: This is incorrect because Amazon Lex is focused on building conversational interfaces and chatbots, not on converting documents into text.

Amazon Transcribe: This is incorrect because Amazon Transcribe is designed for converting speech to text from audio files, not from PDF documents.

References:

https://aws.amazon.com/textract/

A financial services company is deploying a chatbot using a fine-tuned Amazon SageMaker JumpStart model to handle customer queries about loans. The company must ensure that the chatbot complies with various financial regulatory frameworks for secure data handling.

Which two capabilities can the company demonstrate to meet these compliance requirements?

○ Automated scaling of model inference

○ Intrusion detection and monitoring

○ Encryption of sensitive data

○ Optimizing server uptime

○ Using containerized microservices for scaling

Correct answers

- Intrusion detection and monitoring
- Encryption of sensitive data

Feedback

- Explanation:

  Intrusion detection and monitoring: This is correct because compliance in the financial sector often requires monitoring for suspicious or unauthorized activity, ensuring the system is protected from potential breaches.

  Encryption of sensitive data: This is correct because encrypting customer and financial data is critical to comply with regulatory standards such as PCI-DSS and GDPR. Proper encryption ensures that sensitive information is protected both in transit and at rest.

  Automated scaling of model inference: This is incorrect because auto-scaling improves resource efficiency but does not directly address regulatory compliance.

  Optimizing server uptime: This is incorrect because ensuring uptime is important for reliability but does not fulfill compliance requirements related to security and data protection.

  Using containerized microservices for scaling: This is incorrect because microservices are an architectural choice for building scalable applications but do not directly relate to compliance with security and data protection regulations.

  References:

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

A retail company wants to predict customer demand for seasonal products. The company lacks coding experience and knowledge of machine learning algorithms but needs to build a predictive model using internal sales data and external market data.

Which solution will meet these requirements?

○ Import the data into Amazon SageMaker Studio. Build ML models and predict demand using built-in SageMaker algorithms.

○ Import the data into Amazon SageMaker Data Wrangler and build a demand forecasting model with SageMaker JumpStart.

○ Import the data into Amazon SageMaker Canvas. Build ML models and predict demand by selecting values in the data from SageMaker Canvas.

○ Use Amazon Lex to analyze the data and automatically generate predictions for product demand.

Correct answer

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

Feedback

- Explanation:

  SageMaker Canvas allows users with no coding or machine learning experience to create models by simply interacting with the data via a point-and-click interface. This makes it the best choice for the retail company to generate demand forecasts without technical expertise.

  Import the data into Amazon SageMaker Studio. Build ML models and predict demand using built-in SageMaker algorithms: This is incorrect because using SageMaker Studio and its algorithms requires coding and machine learning knowledge, which the company does not have.

  Import the data into Amazon SageMaker Data Wrangler and build a demand forecasting model with SageMaker JumpStart: This is incorrect because, although SageMaker JumpStart simplifies access to pre-built models, using Data Wrangler and tuning these models still requires some machine learning understanding.

  Use Amazon Lex to analyze the data and automatically generate predictions for product demand: This is incorrect because Amazon Lex is designed for building conversational AI, like chatbots, and is not suitable for forecasting product demand.

  References:

  https://aws.amazon.com/sagemaker/canvas/

A cybersecurity company regularly assesses its internal processes with assistance from independent software vendors (ISVs). The company requires email notifications when compliance reports from the ISVs are available for review.

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

- ○ AWS Audit Manager
- ○ AWS Artifact
- ○ AWS Security Hub
- ○ Amazon SNS (Simple Notification Service)

Correct answer

AWS Artifact

Feedback

- Explanation:

  AWS Artifact provides access to compliance-related documents, such as security and compliance reports from third-party ISVs. The company can use AWS Artifact to download and monitor these reports and configure notifications when new reports are available.

  AWS Audit Manager: This is incorrect because AWS Audit Manager helps automate the process of auditing by collecting evidence, but it does not provide access to compliance reports from ISVs.

  AWS Security Hub: This is incorrect because AWS Security Hub is focused on centralizing security findings from AWS services, not on handling compliance reports from ISVs.

  Amazon SNS (Simple Notification Service): This is incorrect because while SNS handles notifications, it is not a service that directly manages or provides compliance reports.

  References:

  https://aws.amazon.com/artifact/

A food processing company has built an AI model to classify different types of fruits based on images. The company wants to evaluate how many images the model has correctly classified into the right fruit categories.

Which evaluation metric should the company use to measure the model's performance?

○ F1 score
○ Accuracy
○ Mean Absolute Error (MAE)
○ Dropout rate

Correct answer

Accuracy

Feedback

- Explanation:

  Accuracy is the appropriate metric because it measures the proportion of images that the model classified correctly out of the total number of images. This is the most straightforward metric for evaluating how well a classification model is performing.

  F1 score: This is incorrect because while the F1 score balances precision and recall, it is more commonly used when dealing with imbalanced datasets. In this case, accuracy is a simpler and more direct measure for evaluating overall classification performance.

  Mean Absolute Error (MAE): This is incorrect because MAE is used to measure errors in regression tasks, not classification. It calculates the difference between predicted and actual continuous values, not category labels.

  Dropout rate: This is incorrect because dropout rate is a parameter used during the training of neural networks to prevent overfitting. It is not an evaluation metric for measuring model performance.

  References:

  https://aws.amazon.com/ai/machine-learning/

A marketing agency needs to select a model from Amazon Bedrock that will be used internally to generate campaign slogans and advertisements. The agency must find a model that produces content in a tone and style that aligns with the agency's creative standards.

What should the agency do to meet these requirements?

○ Evaluate the models using built-in datasets for prompt evaluation.

○ Evaluate the models by testing custom prompts and collecting feedback from the agency's creative team.

○ Use popular model benchmarks and rankings to identify the best model.

○ Analyze the model InvocationLatency runtime metrics in Amazon CloudWatch to assess response times.

Correct answer

Evaluate the models by testing custom prompts and collecting feedback from the agency's creative team.

Feedback

- Explanation:

  The best way to find a model that fits the agency's style and tone is to use custom prompts that reflect real-world use cases and collect feedback from employees who are familiar with the company's preferences. This ensures that the chosen model aligns with internal creative standards.

  Evaluate the models using built-in datasets for prompt evaluation: This is incorrect because built-in datasets may not reflect the company's specific needs or style preferences. Custom prompts are more effective for evaluating the model in a real-world context.

  Use popular model benchmarks and rankings to identify the best model: This is incorrect because public leaderboards often rank models based on general performance metrics, which may not reflect how well a model fits the company's specific style requirements.

Analyze the model InvocationLatency runtime metrics in Amazon CloudWatch to assess response times: This is incorrect because InvocationLatency measures the speed of a model’s responses, not the quality or style of its outputs.

References:

https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering-guidelines.html

A biotechnology company needs to classify human genes into 20 categories based on various gene characteristics. The company also requires a machine learning algorithm that can clearly document how the inner workings of the model influence its decisions and outputs.

Which machine learning algorithm should the company use?

○ K-means clustering
○ Support vector machines (SVM)
○ Decision trees
○ Neural networks

Correct answer

Decision trees

Feedback

- Explanation:

  Decision trees are well-suited for this task because they provide transparency in their decision-making process. The structure of a decision tree allows the company to trace how the input characteristics lead to specific classifications, making it easy to document the inner mechanism of the model and its impact on the output.

K-means clustering: This is incorrect because K-means is used for unsupervised learning and doesn’t inherently document how specific features influence its clustering results.

Support vector machines (SVM): This is incorrect because while SVMs are effective for classification, they are less interpretable and harder to document in terms of how inputs lead to decisions compared to decision trees.

Neural networks: This is incorrect because neural networks, while powerful, are often seen as "black box" models, making it difficult to document and explain the inner workings and how inputs affect outputs.

References:

https://aws.amazon.com/ai/machine-learning/

A travel company is using a pre-trained large language model (LLM) to create a chatbot that provides vacation suggestions. The company needs the chatbot’s responses to be concise and delivered in a specific language.

Which solution will help ensure the LLM produces responses that meet the company's requirements?

○ Adjust the prompt.
○ Select an LLM with different architecture.
○ Raise the temperature value.
○ Increase the maximum number of tokens.

Correct answer

Adjust the prompt.

Feedback

- Explanation:

Adjusting the prompt allows the company to control the style, length, and language of the responses generated by the LLM. Providing clear instructions within the prompt helps guide the model to produce shorter responses and ensures they are written in the specified language.

Select an LLM with different architecture: This is incorrect because changing the architecture or size of the model does not directly control the length or language of its outputs. Prompt adjustments are a more effective solution for this need.

Raise the temperature value: This is incorrect because increasing the temperature makes the model's responses more creative and diverse, which could lead to longer and less predictable outputs.

Increase the maximum number of tokens: This is incorrect because increasing the token limit allows for longer responses, which contradicts the company's need for concise answers.

References:

https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering-guidelines.html

A financial institution is using Amazon Bedrock to build an AI application hosted in a VPC. Due to regulatory compliance standards, the VPC must not have any internet access.

Which AWS service or feature will help meet these requirements?

- ○ Amazon Macie
- ○ Amazon Route 53
- ○ NAT gateway
- ○ AWS PrivateLink

Correct answer

AWS PrivateLink

Feedback

- Explanation:

  AWS PrivateLink enables the financial institution to securely access Amazon Bedrock services from within a VPC without exposing traffic to the public internet. This ensures compliance with regulations that restrict internet access.

  Amazon Macie: This is incorrect because Amazon Macie is used for identifying sensitive data and monitoring security risks, not for managing VPC internet access.

  Amazon Route 53: This is incorrect because Amazon Route 53 is a DNS service, and it doesn't provide the private connectivity needed to avoid internet exposure.

  NAT gateway: This is incorrect because a NAT gateway allows private instances to access the internet, which contradicts the requirement of preventing internet access.

  References:

  https://aws.amazon.com/privatelink/

A healthcare company is training a foundation model (FM) to analyze medical records. The company wants to improve the model's accuracy until it reaches a specific threshold for acceptable performance.

Which solution will help the company achieve this?

○ Decrease the batch size.
○ Increase the epochs.
○ Lower the learning rate.
○ Increase the dropout rate.

Correct answer

Increase the epochs.

Feedback

- Explanation:

  Increasing the epochs allows the model to continue learning by passing through the dataset multiple times, which can improve the model’s accuracy. The company can adjust the number of epochs to achieve the desired performance, ensuring the model trains until it meets the accuracy threshold.

  Decrease the batch size: This is incorrect because reducing the batch size affects the training dynamics but does not directly improve accuracy to a specific level. It may instead slow down training.

  Lower the learning rate: This is incorrect because decreasing the learning rate helps in fine-tuning but may result in longer training times without necessarily improving accuracy beyond a certain point.

  Increase the dropout rate: This is incorrect because increasing the dropout rate prevents overfitting, but an overly high dropout rate can reduce the model’s ability to learn, negatively impacting accuracy.

  References:

  https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-llms-finetuning-hyperparameters.html

A company is developing an educational app where users solve basic math problems such as: "A bag contains 8 blue balls, 5 red balls, and 2 yellow balls. What is the probability of picking a red ball?" The company needs a solution that minimizes operational overhead.

Which solution will meet these requirements with the least operational complexity?

○ Use supervised learning to create a classification model for probability prediction.
○ Use reinforcement learning to teach a model to compute probabilities.
○ Use a simple algorithm that calculates probability using basic rules and formulas.
○ Use unsupervised learning to generate a model for probability estimation.

Correct answer

Use a simple algorithm that calculates probability using basic rules and formulas.

Feedback

- Explanation:

  Using a simple algorithm is the best solution because probability problems can be solved using basic math formulas, without the need for complex machine learning models. This approach requires minimal operational overhead and ensures accurate results through straightforward computations.

  Use supervised learning to create a classification model for probability prediction: This is incorrect because supervised learning is unnecessary for basic probability calculations, which are deterministic and do not require a model.

  Use reinforcement learning to teach a model to compute probabilities: This is incorrect because reinforcement learning is used for decision-making tasks where an agent interacts with an environment, which is excessive for simple probability problems.

  Use unsupervised learning to generate a model for probability estimation: This is incorrect because unsupervised learning is meant for discovering patterns in unlabeled data, not for calculating explicit probabilities based on known values.

  References:

  https://aws.amazon.com/ai/machine-learning/

A cybersecurity firm wants to use AI to enhance the protection of its web application from potential threats. The AI solution must be able to identify whether an IP address originates from a suspicious source.

Which solution will meet these requirements?

○ Build a chatbot for user support.
○ Develop an anomaly detection system.
○ Create a system for translating IP logs into multiple languages.
○ Implement a text summarization system.

Correct answer

Develop an anomaly detection system.

Feedback

- Explanation:

  Anomaly detection systems are designed to identify unusual patterns or deviations from normal behavior. In this case, it can be used to detect suspicious IP addresses by identifying traffic patterns or access behaviors that do not match normal activity, providing an extra layer of security.

  Build a chatbot for user support: This is incorrect because chatbots are designed for customer interaction, not for identifying or analyzing security threats.

  Create a system for translating IP logs into multiple languages: This is incorrect because translating logs does not help in detecting suspicious IP addresses or enhancing security.

  Implement a text summarization system: This is incorrect because text summarization condenses large pieces of text and is unrelated to identifying security threats from IP addresses.

  References:

  https://aws.amazon.com/what-is/anomaly-detection/

A security company is using Amazon Bedrock to run foundation models (FMs). The company wants to ensure that only authorized users can invoke the models and needs to detect any unauthorized access attempts to refine AWS Identity and Access Management (IAM) policies.

Which AWS service should the company use to identify unauthorized users trying to access Amazon Bedrock?

○ AWS Shield

○ AWS CloudTrail

○ Amazon Macie

○ Amazon Inspector

Correct answer

AWS CloudTrail

Feedback

- Explanation:

  AWS CloudTrail records all API calls and actions across AWS services, including attempts to invoke Amazon Bedrock models. By reviewing these logs, the company can identify unauthorized access attempts and set appropriate IAM policies for future model use.

  AWS Shield: This is incorrect because AWS Shield provides protection against distributed denial of service (DDoS) attacks, not for identifying unauthorized access attempts within AWS services.

  Amazon Macie: This is incorrect because Macie is designed to detect sensitive data in Amazon S3, not for monitoring access attempts to AWS services.

  Amazon Inspector: This is incorrect because Amazon Inspector is used to assess security vulnerabilities in EC2 instances, not to track or identify unauthorized users attempting to access AWS services.

  References:

  https://aws.amazon.com/cloudtrail/

A university student is copying content from a generative AI system to write essays without proper attribution.

Which challenge of responsible generative AI does this scenario represent?

○ Misinformation

○ Hallucinations

○ Plagiarism

○ Bias

Correct answer

Plagiarism

Feedback

- Explanation:

  Plagiarism occurs when content generated by AI is copied without proper attribution, leading to ethical and academic concerns. In this scenario, the student is using AI-generated text without crediting the source, which constitutes plagiarism.

  Misinformation: This is incorrect because misinformation refers to false or inaccurate information being generated, not copying content without proper credit.

  Hallucinations: This is incorrect because hallucinations refer to AI generating false or misleading information that wasn't part of the original input.

  Bias: This is incorrect because bias refers to unfair or unbalanced outputs from AI systems, not the unauthorized copying of content.

  References:

  https://aws.amazon.com/ai/responsible-ai/

An AI-driven marketing agency uses machine learning models to predict consumer trends each season. The company's AI practitioner is preparing a report to explain the models' behavior and predictions to stakeholders, ensuring transparency and trust in the process.

What should the AI practitioner include in the report to meet these requirements?

○ The raw data that was used to test the model

○ A summary of the model's memory usage during training

○ Partial dependence plots (PDPs) to show how features affect predictions

○ The dataset schema

Correct answer

Partial dependence plots (PDPs) to show how features affect predictions

Feedback

- Explanation:

  Partial dependence plots (PDPs) help stakeholders understand how specific features influence the model's predictions. By showing the relationship between inputs and outputs, PDPs provide insights into the model's decision-making process, enhancing transparency and explainability.

  The raw data that was used to test the model: This is incorrect because providing raw data doesn't explain how the model makes predictions. Raw data alone does not enhance the interpretability of the model.

  A summary of the model's memory usage during training: This is incorrect because memory usage is related to model efficiency, not its transparency or explainability.

  The dataset schema: This is incorrect because while the schema shows the structure of the dataset, it does not help explain how the model makes decisions.

  References:

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

A company wants to develop a chat interface using large language models (LLMs) with Amazon Bedrock to help users navigate the company's product manuals, which are stored as PDF files. The company needs a cost-effective solution to provide relevant answers from the manuals.

Which solution meets these requirements most cost-effectively?

○ Use prompt engineering to add one relevant PDF file as context to the user prompt when submitted to Amazon Bedrock.

○ Use prompt engineering to add all PDF files as context to every user prompt submitted to Amazon Bedrock.

○ Fine-tune a model with Amazon Bedrock using all the PDF files and process user prompts with the fine-tuned model.

○ Upload PDF documents to an Amazon Bedrock knowledge base and use the knowledge base to provide context when users submit prompts.

Correct answer

Upload PDF documents to an Amazon Bedrock knowledge base and use the knowledge base to provide context when users submit prompts.

Feedback

- Explanation:

  Uploading PDF documents to an Amazon Bedrock knowledge base allows the company to store the manuals and provide relevant context dynamically based on user prompts. This solution is cost-effective because the model does not need to be fine-tuned or retrained, and it leverages an existing knowledge base to serve the necessary information without repeatedly submitting large amounts of data.

  Use prompt engineering to add one relevant PDF file as context to the user prompt when submitted to Amazon Bedrock: This is incorrect because adding a single document as context may not provide sufficient information for accurate responses when the user's query spans multiple manuals.

  Use prompt engineering to add all PDF files as context to every user prompt submitted to Amazon Bedrock: This is incorrect because adding all the PDF files as context to each prompt would be expensive in terms of token usage and processing, making it a less cost-effective solution.

  Fine-tune a model with Amazon Bedrock using all the PDF files and process user prompts with the fine-tuned model: This is incorrect because fine-tuning a model with a large dataset like all the product manuals would be costly and unnecessary when a knowledge base can serve the same purpose more efficiently.

  References:

  https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering-guidelines.html

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?

○ Training

○ Inference

○ Model deployment

○ Bias correction

Correct answer

Inference

Feedback

- Explanation:

  Inference is the process where a trained model is used to analyze new data and make predictions. In this case, when the deployed object detection model analyzes a new image to identify objects, it is performing inference. The model is applying what it learned during the training phase to make predictions on unseen data.

  Training: This is incorrect because training refers to the process of teaching the model using a labeled dataset. The model has already been trained and is now being used to make predictions.

  Model deployment: This is incorrect because model deployment refers to the process of moving the trained model into a production environment. The question focuses on what happens *after* the model has been deployed.

  Bias correction: This is incorrect because bias correction involves adjusting the model to prevent biased outputs. The question does not mention any bias-related issues, and it focuses on object detection in a production environment.

  References:

  https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html

A healthcare company’s AI development team wants to quickly deploy and consume a foundation model (FM) within their VPC to process medical data securely.

Which AWS service or feature will help them achieve this?

○ Amazon Comprehend Medical
○ AWS Fargate
○ Amazon SageMaker endpoints
○ AWS Lambda

Correct answer

Amazon SageMaker endpoints

Feedback

- Explanation:

  Amazon SageMaker endpoints allow the team to securely deploy machine learning models, including foundation models, within their VPC. This ensures the data stays within the VPC while enabling real-time access to the deployed models for secure processing of medical data.

  Amazon Comprehend Medical: This is incorrect because Amazon Comprehend Medical is a service designed to extract and analyze medical information from text, not to deploy and host foundation models within a VPC.

  AWS Fargate: This is incorrect because AWS Fargate is a serverless compute engine for containerized applications and is not specifically designed for deploying or hosting machine learning models.

  AWS Lambda: This is incorrect because AWS Lambda is a serverless compute service for running functions in response to events. It’s not designed for hosting and deploying foundation models within a VPC.

  References:

  https://docs.aws.amazon.com/sagemaker/latest/dg/realtime-endpoints.html

An AI researcher has developed a deep learning model to identify different types of textures in images. The researcher now wants to assess how well the model performs in classifying these textures.

Which metric will help the researcher evaluate the model’s performance?

- ○ Confusion matrix
- ○ Correlation matrix
- ○ R2 score
- ○ Mean absolute error (MAE)

Correct answer

Confusion matrix

Feedback

- Explanation:

  A confusion matrix is a tool used to evaluate the performance of a classification model by showing the true positives, true negatives, false positives, and false negatives. This helps the AI researcher understand how well the model is classifying the textures in the images and where it may be making mistakes.

  Correlation matrix: This is incorrect because a correlation matrix shows relationships between variables in a dataset but does not help evaluate classification performance.

  R2 score: This is incorrect because R2 score is used to measure the goodness-of-fit for regression models, not classification models.

  Mean absolute error (MAE): This is incorrect because MAE is used to measure error in regression models, not classification tasks like image classification.

  References:

  https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html

A retail company has terabytes of data stored in its database, which can be used for business analysis. The company wants to develop an AI application that can generate SQL queries from simple text inputs provided by employees with minimal technical experience.

Which solution meets these requirements?

- ○ Generative pre-trained transformers (GPT)
- ○ Convolutional neural network (CNN)
- ○ Random forest
- ○ Recurrent neural network (RNN)

Correct answer

Generative pre-trained transformers (GPT)

Feedback

- Explanation:

  Generative pre-trained transformers (GPT) are ideal for this task because they are designed for natural language processing (NLP) tasks, such as converting human language into structured queries like SQL. GPT models can understand and interpret employee text inputs and generate the appropriate SQL queries, even for users with minimal technical skills.

  Convolutional neural network (CNN): This is incorrect because CNNs are typically used for image processing tasks, not for generating SQL queries from text inputs.

  Random forest: This is incorrect because random forests are used for classification and regression tasks, not for natural language understanding or generating queries from text.

  Recurrent neural network (RNN): This is incorrect because while RNNs are used for sequential data like time series or language modeling, GPT is a more advanced architecture for NLP tasks such as converting text to SQL.

References:

https://aws.amazon.com/what-is/nlp/

An AI practitioner is using a large language model (LLM) to generate content for marketing campaigns. While the content sounds plausible and fact-based, some of the information is actually incorrect.

Which problem is the LLM experiencing?

○ Data leakage
○ Hallucination
○ Feature selection
○ Gradient explosion

Correct answer

Hallucination

Feedback

- Explanation:

  Hallucination refers to the problem where a model generates content that appears factual but is not based on accurate or relevant data. In this case, the LLM is generating plausible-sounding but incorrect marketing content, which is a hallmark of hallucination in language models.

  Data leakage: This is incorrect because data leakage happens when the model has access to information during training that it wouldn’t have during real-world use, which leads to overly optimistic results. It does not explain why the generated content is incorrect.

  Feature selection: This is incorrect because feature selection refers to choosing the most relevant input variables in traditional machine learning models. It does not apply to the generation of incorrect text by LLMs.

Gradient explosion: This is incorrect because gradient explosion refers to issues in the training process of deep learning models, where gradients grow uncontrollably, but it is unrelated to generating incorrect but plausible text.

References:

https://aws.amazon.com/blogs/machine-learning/improve-llm-performance-with-human-and-ai-feedback-on-amazon-sagemaker-for-amazon-engineering/

A global transportation company receives thousands of requests daily from customers seeking updates on package deliveries. To manage the volume, the company wants to deploy Agents for Amazon Bedrock to streamline responses and automate workflows.

What are the key benefits of using Amazon Bedrock agents that could assist the transportation company?

○ Simplification of data visualization for customer tracking insights
○ Automation of routine inquiries and coordination of multi-step processes
○ Generation of marketing analytics to predict future package deliveries
○ Enhancement of email campaigns for customer engagement based on delivery data

Correct answer

Automation of routine inquiries and coordination of multi-step processes

Feedback

- Explanation:

  Amazon Bedrock agents automate repetitive tasks like responding to common customer inquiries, such as package status, and orchestrate complex workflows, including escalating unresolved issues or initiating specific processes across different systems. This improves efficiency in handling large volumes of customer service requests.

Simplification of data visualization for customer tracking insights: This is incorrect because Bedrock agents focus on automating tasks and workflows, not on creating visualizations or dashboards for tracking insights.

Generation of marketing analytics to predict future package deliveries: This is incorrect because Bedrock agents are not used for marketing analytics or predictive delivery models. They are designed for workflow automation and task handling.

Enhancement of email campaigns for customer engagement based on delivery data: This is incorrect because Bedrock agents are not designed for email campaign management or customer engagement tools.

References:

https://aws.amazon.com/bedrock/agents/

A financial services company is using few-shot prompting on a base model hosted on Amazon Bedrock to generate daily reports. The model currently uses 10 examples in each prompt and performs well. However, the company wants to reduce monthly operational costs.

Which solution will meet these requirements?

○ Customize the model through fine-tuning.
○ Decrease the number of tokens in the prompt.
○ Increase the number of examples used in the prompt.
○ Deploy the model on a dedicated server instance.

Correct answer

Decrease the number of tokens in the prompt.

Feedback

- Explanation:

Decreasing the number of tokens in the prompt will directly reduce the amount of input the model processes during each invocation, leading to lower inference costs. Fewer tokens in each prompt mean the company can maintain performance while cutting costs.

Customize the model through fine-tuning: This is incorrect because fine-tuning adds complexity and cost. It does not directly help in reducing inference costs based on token usage.

Increase the number of examples used in the prompt: This is incorrect because increasing the number of examples would result in more tokens being processed, which would raise costs rather than lowering them.

Deploy the model on a dedicated server instance: This is incorrect because running the model on a dedicated server may not necessarily reduce costs. It can, in fact, increase operational expenses depending on usage patterns.

References:

https://docs.aws.amazon.com/bedrock/latest/userguide/design-a-prompt.html

A healthcare company is developing an application that needs to generate synthetic medical data based on patterns observed in existing patient datasets.

Which type of model should the company use to meet this requirement?

○ Generative adversarial network (GAN)
○ Support vector machine (SVM)
○ Convolutional neural network (CNN)
○ Decision tree

Correct answer

Generative adversarial network (GAN)

Feedback

- Explanation:

  A Generative adversarial network (GAN) is ideal for generating synthetic data that mimics real data by learning the patterns from existing datasets. GANs are commonly used for creating synthetic images, text, and other types of data, making them well-suited for this task.

  Support vector machine (SVM): This is incorrect because SVM is used for classification and regression tasks, not for generating synthetic data.

  Convolutional neural network (CNN): This is incorrect because CNNs are typically used for image recognition and processing tasks, not for generating synthetic data.

  Decision tree: This is incorrect because decision trees are used for decision-making and classification tasks, but they do not generate synthetic data.

  References:

  https://aws.amazon.com/what-is/gan/

A research company has historical transcripts of interviews, but some portions of the text are missing due to errors in data collection. The company needs to build a machine learning model that can predict and fill in the missing words based on the surrounding context.

Which type of model meets this requirement?

○ LDA (Latent Dirichlet Allocation) models
○ K-means clustering models
○ BERT-based models
○ Time series models

Correct answer

BERT-based models

Feedback

- Explanation:

  BERT-based models are designed for natural language processing tasks, including predicting missing words in a sentence. BERT’s ability to understand context by looking at the words before and after the missing sections makes it highly effective for tasks such as filling in gaps in transcripts.

  LDA (Latent Dirichlet Allocation) models: This is incorrect because LDA is used for topic modeling, which helps in identifying themes in a text but is not suited for predicting missing words.

  K-means clustering models: This is incorrect because K-means is a clustering algorithm that groups similar data points together, not for completing text or predicting missing words.

  Time series models: This is incorrect because time series models are used for predicting future values based on historical data trends, not for completing sentences or predicting words in text.

  References:

  https://aws.amazon.com/blogs/machine-learning/fine-tune-and-host-hugging-face-bert-models-on-amazon-sagemaker/

A museum is developing an AI-powered virtual tour guide to explain historical artifacts to visitors. The AI needs to adjust its language and tone depending on the visitor’s background, such as children, history enthusiasts, or academic researchers. The visitor’s background will be provided to the model when they ask questions.

Which solution meets these requirements with the least implementation effort?

○ Fine-tune the model by adding specialized datasets for each visitor group, such as children and researchers.

○ Include a description of the visitor's background in the prompt to instruct the model on how to adjust its response.

○ Use a separate machine learning model to analyze and transform the model’s output for different user types.

○ Implement a multi-step dialogue process where the model asks follow-up questions to adjust its response style based on the visitor's feedback.

Correct answer

Include a description of the visitor's background in the prompt to instruct the model on how to adjust its response.

Feedback

- Explanation:

  Including a description of the visitor's background in the prompt is the most efficient solution, requiring minimal implementation effort. By simply adjusting the prompt, the model can tailor its responses to match the visitor's background, whether it’s a child or a researcher, without the need for additional training or complex workflows.

  Fine-tune the model by adding specialized datasets for each visitor group, such as children and researchers: This is incorrect because fine-tuning the model with additional datasets requires significant effort and is more resource-intensive than prompt engineering.

  Use a separate machine learning model to analyze and transform the model’s output for different user types: This is incorrect because using an additional machine learning model introduces unnecessary complexity when prompt adjustments can achieve the desired result more simply.

  Implement a multi-step dialogue process where the model asks follow-up questions to adjust its response style based on the visitor's feedback: This is incorrect because a multi-step dialogue process increases complexity and implementation effort, which is not necessary for simply adjusting the response style based on the provided background.

  References:

  https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering-guidelines.html

A legal firm is using a foundation model (FM) from Amazon Bedrock to power its AI legal search tool. The firm wants to fine-tune the model using its own proprietary legal documents to improve the tool's accuracy.

Which strategy will successfully fine-tune the model?

- ○ Organize the dataset into a JSON file and upload it directly to the model for fine-tuning.
- ○ Purchase additional storage on Amazon S3 to store the training data.
- ○ Integrate AWS CloudWatch for real-time model monitoring and adjustments.
- ○ Provide labeled data with the prompt field and the completion field.

Correct answer

Provide labeled data with the prompt field and the completion field.

Feedback

- Explanation:

  Providing labeled data with specific inputs (prompts) and expected outputs (completions) allows the foundation model to learn from the company's proprietary documents. Fine-tuning requires structured data with clear prompts and corresponding completions to improve the model's performance in generating accurate responses for legal searches.

  Organize the dataset into a JSON file and upload it directly to the model for fine-tuning: This is incorrect because while data format is important, uploading a simple JSON file without labeled prompts and completions does not ensure effective fine-tuning.

  Purchase additional storage on Amazon S3 to store the training data: This is incorrect because buying storage does not directly help in fine-tuning the model. The focus should be on preparing the data for model training, not storage.

  Integrate AWS CloudWatch for real-time model monitoring and adjustments: This is incorrect because AWS CloudWatch is used for monitoring and logging, but it does not contribute to the actual fine-tuning process of the model.

References:

https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering-guidelines.html

An ecommerce company wants to develop a solution that analyzes customer reviews of products to determine customer sentiments based on the text.

Which AWS services meet these requirements? (Select TWO.)

○ Amazon Lex
○ Amazon Comprehend
○ Amazon Polly
○ Amazon Transcribe
○ Amazon Bedrock

Correct answers

- Amazon Comprehend
- Amazon Bedrock

Feedback

- Explanation:

  Amazon Comprehend is a natural language processing (NLP) service that can analyze customer reviews to detect sentiment, including positive, negative, neutral, or mixed. It's ideal for automatically analyzing written text to determine customer opinions.
  Amazon Bedrock offers foundation models for various AI tasks, including text analysis and sentiment detection. It provides pre-trained models that can be fine-tuned for tasks like sentiment analysis based on customer reviews.

  Amazon Lex: This is incorrect because Amazon Lex is designed for building conversational interfaces like chatbots, not for analyzing sentiment in written reviews.

Amazon Polly: This is incorrect because Amazon Polly converts text to speech, which is unrelated to sentiment analysis from written customer reviews.

Amazon Transcribe: This is incorrect because Amazon Transcribe converts speech to text, which is not relevant for analyzing already written customer reviews.

References:

https://aws.amazon.com/comprehend/

https://aws.amazon.com/bedrock/