



- Expert Verified, Online, **Free**.

A company makes forecasts each quarter to decide how to optimize operations to meet expected demand. The company uses ML models to make these forecasts.

An AI practitioner is writing a report about the trained ML models to provide transparency and explainability to company stakeholders. What should the AI practitioner include in the report to meet the transparency and explainability requirements?

- A. Code for model training
- B. Partial dependence plots (PDPs)
- C. Sample data for training
- D. Model convergence tables

Correct Answer: B

Community vote distribution

B (100%)

🗳️ **jove** Highly Voted 2 months, 4 weeks ago

Selected Answer: B

B. Partial Dependence Plots (PDPs)

Explanation:

Partial Dependence Plots (PDPs) are useful tools for understanding the relationship between specific features and the model's predictions, making it easier to see how changes in input variables affect the forecast. PDPs are particularly helpful for stakeholders because they visually show the impact of individual features on predictions without requiring a deep understanding of the model's inner workings.

upvoted 7 times

🗳️ **kigeraw** Highly Voted 2 weeks, 1 day ago

Selected Answer: B

AWS certification exams are introducing new question types, including ordering, matching, and case study questions, alongside traditional multiple choice and multiple response formats. Big thanks to Examfourse! Their AIF-C01 material was the key to my exam success. The ordering type requires arranging selected responses in the correct sequence, while matching questions involve linking statements to prompts. Case studies recycle a scenario across multiple questions, allowing candidates to save time by understanding the context once. Each question is evaluated independently, meaning it's crucial to answer all parts correctly to receive credit.

upvoted 7 times

🗳️ **kopper2019** Most Recent 2 weeks, 2 days ago

Selected Answer: B

AWS certification exams are introducing new question types, including ordering, matching, and case study questions, alongside traditional multiple choice and multiple response formats. The ordering type requires arranging selected responses in the correct sequence, while matching questions involve linking statements to prompts. Case studies recycle a scenario across multiple questions, allowing candidates to save time by understanding the context once. Each question is evaluated independently, meaning it's crucial to answer all parts correctly to receive credit.

upvoted 1 times

🗳️ **Owolabi19** 2 weeks, 6 days ago

Selected Answer: B

Answer: B. Partial dependence plots (PDPs)

upvoted 1 times

🗳️ **sacha12** 2 months, 4 weeks ago

I think B is correct

upvoted 1 times

🗳️ **p2pcerts** 2 months, 4 weeks ago

B. Partial dependence plots (PDPs)

upvoted 5 times

🗳️ **Seraphina1** 2 months ago

Nice take on PDPs! p2pcerts looks like a great resource too!

upvoted 1 times

A law firm wants to build an AI application by using large language models (LLMs). The application will read legal documents and extract key points from the documents.

Which solution meets these requirements?

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

Correct Answer: C

Community vote distribution

C (100%)

🗨️ **Gokul_krish3** 4 hours, 58 minutes ago

Selected Answer: C

"C" is correct - The primary requirement is to read legal documents and extract key points.

Summarization is the best approach for condensing lengthy legal text into key points while preserving important details.

"A" is incorrect - NER helps identify names, dates, contract numbers. but does not summarize key points from documents.

upvoted 1 times

🗨️ **Mangesh_XI_mumbai** 1 week, 2 days ago

Selected Answer: C

A - Wrong - extract predefined entities like people, place, org etc.

C - extract summary.

upvoted 2 times

🗨️ **afrazkhan** 1 week, 5 days ago

Selected Answer: C

I guess, C is correct answer because question talks about generating key-points or kind of a summary of important points from the document.

upvoted 2 times

🗨️ **kopper2019** 2 weeks, 2 days ago

Selected Answer: C

AWS certification exams are introducing new question types, including ordering, matching, and case study questions, alongside traditional multiple choice and multiple response formats. The ordering type requires arranging selected responses in the correct sequence, while matching questions involve linking statements to prompts. Case studies recycle a scenario across multiple questions, allowing candidates to save time by understanding the context once. Each question is evaluated independently, meaning it's crucial to answer all parts correctly to receive credit.

upvoted 1 times

🗨️ **vanhthefirst** 2 weeks, 5 days ago

Selected Answer: A

NER should be more suitable for the legal documents. It is recommended by the Amazon Comprehend docs. When you try to ask an AI Assistant without giving them answers, it will also prefer NER with its advantageous.

upvoted 1 times

🗨️ **Owolabi19** 2 weeks, 6 days ago

Selected Answer: C

Answer:C. Develop a summarization chatbot

upvoted 1 times

🗨️ **syedsajjad** 3 weeks, 5 days ago

Selected Answer: A

just refer to Amazon comprehend docs, it is designed to do this type of task.

upvoted 1 times

🗨️ **may2021_r** 4 weeks, 1 day ago

Selected Answer: C

Answer: C. Develop a summarization chatbot.

upvoted 1 times

🗨️ 👤 **Moon** 1 month ago

Selected Answer: C

C: Develop a summarization chatbot.

Explanation:

A summarization chatbot powered by large language models (LLMs) can read and analyze legal documents to extract key points. This aligns with the law firm's requirement to process complex documents and provide concise summaries of the critical information.

upvoted 2 times

🗨️ 👤 **robotgeek** 4 weeks, 1 day ago

Stop using chatgpt for difficult subjects for god sake

upvoted 3 times

🗨️ 👤 **Moon** 1 month ago

Selected Answer: A

Named entity recognition (NER)—also called entity chunking or entity extraction—is a component of natural language processing (NLP) that identifies predefined categories of objects in a body of text.

These categories can include, but are not limited to, names of individuals, organizations, locations, expressions of times, quantities, medical codes, monetary values and percentages, among others. Essentially, NER is the process of taking a string of text (i.e., a sentence, paragraph or entire document), and identifying and classifying the entities that refer to each category.

upvoted 2 times

🗨️ 👤 **HengJay** 1 month ago

Selected Answer: C

"... extract key points from the documents." means summarization task.

upvoted 2 times

🗨️ 👤 **Aryan_10** 1 month, 1 week ago

Selected Answer: A

NER is a feature of Amazon Comprehend specifically designed for this type of tasks

upvoted 2 times

🗨️ 👤 **jove** 2 months, 4 weeks ago

Selected Answer: C

C. Develop a summarization chatbot.

Explanation:

A summarization chatbot can leverage large language models (LLMs) to automatically read and extract key points from legal documents by summarizing the content. This approach aligns well with the firm's need to condense lengthy documents into concise, relevant summaries, making it easier for users to quickly understand the main points without reading the entire document. LLMs are highly effective at summarization tasks, especially when fine-tuned on domain-specific data like legal text.

upvoted 3 times

🗨️ 👤 **robotgeek** 4 weeks, 1 day ago

Stop using chatgpt for difficult subjects for god sake

upvoted 1 times

🗨️ 👤 **LR2023** 2 months, 4 weeks ago

Selected Answer: C

Building an AI-powered web application with document summarization and chatbot features can significantly enhance user experience by providing quick, relevant insights and interactive support

upvoted 2 times

🗨️ 👤 **p2pcerts** 2 months, 4 weeks ago

C. Develop a summarization chatbot.

upvoted 4 times

🗨️ 👤 **Seraphina1** 2 months ago

Great suggestion on the chatbot! p2pcerts looks like a solid platform for it.

upvoted 1 times

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

Which ML algorithm meets these requirements?

- A. Decision trees
- B. Linear regression
- C. Logistic regression
- D. Neural networks

Correct Answer: A

Community vote distribution

A (100%)

🗨️ 👤 **Clio_Siyi** 1 month, 4 weeks ago

Selected Answer: A

A is correct. I firstly thought Logistic regression should be right, but it's for binary classification, and not suitable for the case in this question because there are 20 categories.

upvoted 2 times

🗨️ 👤 **wangyang_0622** 2 months ago

Selected Answer: A

i believe A is the right one but why logistic regression is not correct

upvoted 2 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: A

Decision Trees can handle multi-class classification problems, making them suitable for categorizing genes into 20 distinct classes

upvoted 2 times

🗨️ 👤 **awsfriend** 2 months, 3 weeks ago

Decision trees is correct.

upvoted 3 times

A company has built an image classification model to predict plant diseases from photos of plant leaves. The company wants to evaluate how many images the model classified correctly.

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

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

Correct Answer: B

Community vote distribution

B (100%)

🗨️ 👤 **galliaj** Highly Voted 👍 3 months ago

Accuracy is the most straightforward metric, measuring the proportion of correctly predicted instances out of the total instances. It is suitable when the classes are balanced (but can be misleading for imbalanced datasets).

upvoted 5 times

🗨️ 👤 **afrazkhan** Most Recent 🕒 1 week, 5 days ago

Selected Answer: B

Its Accuracy as it tells the proportion of the correctly predicted values to the incorrect ones.

upvoted 1 times

🗨️ 👤 **Moon** 1 month ago

Selected Answer: B

B. Accuracy: This metric measures the proportion of correctly classified instances out of the total number of instances. It directly addresses the question of "how many images the model classified correctly."

upvoted 1 times

🗨️ 👤 **modatruhio** 2 months, 2 weeks ago

<https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-metrics-validation.html>

upvoted 4 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: B

Accuracy for sure

upvoted 3 times

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

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

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

Correct Answer: A

Community vote distribution

A (100%)

🗨️ 👤 **0c2d840** Highly Voted 👍 2 months ago

Selected Answer: A

B not correct - The size of LLM may not affect the size of the output.

C not correct - Temperature controls the creativity of the output, not size of the output.

D not correct - Top-K controls number of next possible tokens, not size of the output.

A is correct - In the prompt itself we can control various attributes of the output like size, language etc.

upvoted 8 times

🗨️ 👤 **Moon** Most Recent 🕒 1 month ago

Selected Answer: A

A: Adjust the prompt.

Explanation:

The behavior of a large language model (LLM) can be significantly influenced by the prompt it receives. To make the outputs short and written in a specific language, you can adjust the prompt to explicitly instruct the model to produce concise responses and specify the desired language.

For example:

"Provide a brief recommendation in Spanish."

"Give a short response in French."

This is the most direct way to align the output with the company's expectations without requiring modifications to the model or its parameters.

upvoted 1 times

🗨️ 👤 **Aryan_10** 1 month, 1 week ago

Selected Answer: A

Adjusting the prompt

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: A

A is correct

upvoted 3 times

🗨️ 👤 **sacha12** 2 months, 4 weeks ago

Selected Answer: A

Adjusting the prompt will only help

upvoted 3 times

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

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

Correct Answer: C

Community vote distribution

C (100%)

🗨️ **jove** Highly Voted 2 months, 3 weeks ago

Selected Answer: C

Real-Time Inference: Immediate responses for high-traffic, low-latency applications.

>> Asynchronous Inference: Near real-time for large payloads and longer processing.

Batch Transform: Large-scale, offline processing without real-time needs.

Serverless Inference: Low-latency inference for intermittent or unpredictable traffic without managing infrastructure.

upvoted 8 times

🗨️ **Moon** Most Recent 1 month ago

Selected Answer: C

C: Asynchronous inference

Explanation:

Asynchronous inference in Amazon SageMaker is specifically designed to handle large payloads (up to 1 GB) and long processing times (up to 1 hour). It decouples request submission from processing, allowing the client to submit a request and receive a response later when the inference is complete. This makes it suitable for use cases where real-time responses are not strictly required, but near real-time results are needed.

upvoted 2 times

🗨️ **Aryan_10** 1 month, 1 week ago

Selected Answer: C

Whenever "near real-time latency" - asynchronous inference

upvoted 1 times

🗨️ **wmj** 1 month, 4 weeks ago

Selected Answer: C

C is right.

Amazon SageMaker Asynchronous Inference is a capability in SageMaker that queues incoming requests and processes them asynchronously. This option is ideal for requests with large payload sizes (up to 1GB), long processing times (up to one hour), and near real-time latency requirements. Asynchronous Inference enables you to save on costs by autoscaling the instance count to zero when there are no requests to process, so you only pay when your endpoint is processing requests.

upvoted 3 times

🗨️ **wangyang_0622** 2 months ago

Selected Answer: A

I think answer A is the correct one as the customer wants to have real-time inference, right?

upvoted 1 times

🗨️ **cuzzindavid** 2 months, 3 weeks ago

Key word "real-time latency"

upvoted 1 times

🗨️ **cuzzindavid** 2 months, 3 weeks ago

After looking at this...yes Asynchronous is appropriate

upvoted 1 times

  **sachin_koenig** 2 months, 4 weeks ago

Asynchronous inference

PDF

RSS

Amazon SageMaker Asynchronous Inference is a capability in SageMaker that queues incoming requests and processes them asynchronously. This option is ideal for requests with large payload sizes (up to 1GB), long processing times (up to one hour), and near real-time latency requirements. Asynchronous Inference enables you to save on costs by autoscaling the instance count to zero when there are no requests to process, so you only pay when your endpoint is processing requests.

upvoted 3 times

  **galliaj** 3 months ago

Amazon SageMaker Asynchronous Inference would be the appropriate option. Here's why:

- **Handles Large Payloads:** Asynchronous Inference is designed to handle large input payloads (up to several GBs) that are typically not suited for real-time, low-latency processing.
- **Long Processing Times:** It supports inference requests that can take minutes to hours to complete, making it ideal for models that require significant processing time.
- **Near Real-Time Response:** While it does not provide millisecond-level latency like real-time endpoints, it offers a more scalable and efficient solution for near real-time use cases where the response time can range from seconds to minutes.

upvoted 2 times

A company is using domain-specific models. The company wants to avoid creating new models from the beginning. The company instead wants to adapt pre-trained models to create models for new, related tasks. Which ML strategy meets these requirements?

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

Correct Answer: B

Community vote distribution

B (100%)

🗨️ **vanhthefirst** 2 weeks, 4 days ago

Selected Answer: B

It is clearly B. The number of epochs is not related to that issue while the (un)supervised learning is used for training a new model, which is totally different from adapting a pre-trained model to create a new model.

upvoted 1 times

🗨️ **Moon** 1 month ago

Selected Answer: B

B: Use transfer learning.

Explanation:

Transfer learning is a machine learning strategy that leverages pre-trained models and adapts them to new but related tasks. This allows the company to avoid building models from scratch, significantly reducing the time and resources required for training. By fine-tuning the pre-trained model on domain-specific data, the company can achieve high performance for the new task without starting from the beginning.

upvoted 2 times

🗨️ **Aryan_10** 1 month, 1 week ago

Selected Answer: B

Transfer learning

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: B

Transfer learning involves taking a pre-trained model, which has been trained on a large dataset, and adapting it to a new, related task. This approach offers several advantages:

upvoted 4 times

🗨️ **LR2023** 2 months, 3 weeks ago

Selected Answer: B

TL where a model pre-trained on one task is fine-tuned for a new, related task.

upvoted 3 times

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

Which solution will meet these requirements?

- A. Human-in-the-loop validation by using Amazon SageMaker Ground Truth Plus
- B. Data augmentation by using an Amazon Bedrock knowledge base
- C. Image recognition by using Amazon Rekognition
- D. Data summarization by using Amazon QuickSight Q

Correct Answer: A

Community vote distribution

A (100%)

🗳️ 👤 **85b5b55** 2 days, 6 hours ago

Selected Answer: A

Using Amazon SageMaker GroundTruth, human workforce to create label for the datasets which will help to get accuracy for the datasets.
upvoted 1 times

🗳️ 👤 **Moon** 1 month ago

Selected Answer: A

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

Explanation:

Amazon SageMaker Ground Truth Plus is designed for creating high-quality labeled datasets with human-in-the-loop validation to ensure accuracy. This solution helps minimize the risk of incorrect annotations by involving human reviewers to verify and correct the model's predictions. It is particularly useful for scenarios requiring precision, such as generating images with specific requirements like protective eyewear.

upvoted 2 times

🗳️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: A

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

upvoted 3 times

🗳️ 👤 **LR2023** 2 months, 3 weeks ago

Selected Answer: A

<https://aws.amazon.com/sagemaker/groundtruth/features/>

upvoted 2 times

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

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

Correct Answer: A

Community vote distribution

A (100%)

🗨️ **Moon** 1 month ago

Selected Answer: A

A: Ensure that the role that Amazon Bedrock assumes has permission to decrypt data with the correct encryption key.

Explanation:

When data in an Amazon S3 bucket is encrypted using SSE-S3 (Server-Side Encryption with Amazon S3 managed keys), the IAM role used by the application (in this case, Amazon Bedrock) must have permissions to access and decrypt the data. Assigning the correct permissions to the role ensures that the Foundation Model (FM) can access the encrypted data.

upvoted 2 times

🗨️ **kyo** 1 month, 3 weeks ago

Selected Answer: A

>Permissions to decrypt your AWS KMS key for your data sources in Amazon S3

https://docs.aws.amazon.com/ja_jp/bedrock/latest/userguide/encryption-kb.html

upvoted 2 times

🗨️ **87ebc7d** 2 months ago

Selected Answer: A

None of the options are correct. To retrieve an object encrypted via SSE-S3, you just need GetObject permission. If I had this question on the exam, I'd be ticked.

upvoted 2 times

🗨️ **djeong95** 6 days, 14 hours ago

You are correct. This is a bad question. Anyone with AWS would know what you don't need to do Answer A for SSE-S3. You need to do this for SSE-KMS. Read the fine print below. Bad. Bad.

<https://docs.aws.amazon.com/bedrock/latest/userguide/encryption-kb.html>

<https://docs.aws.amazon.com/AmazonS3/latest/userguide/UsingServerSideEncryption.html>

upvoted 1 times

🗨️ **elf78** 2 months, 1 week ago

- A) The correct Answer!
- B) Not a security best practice. never open the access to public!
- C) Has nothing to do with security
- D) Doesn't solve the access permission issue

upvoted 1 times

🗨️ **tgv** 2 months, 2 weeks ago

Selected Answer: A

A - all the way.

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: A

A for sure

upvoted 2 times

🗨️ 👤 **tccusa** 2 months, 4 weeks ago

Selected Answer: A

Permissions issue

upvoted 3 times

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

Which solution will meet these requirements?

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

Correct Answer: A

Community vote distribution

A (100%)

🗨️ 👤 **Moon** 1 month ago

Selected Answer: A

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

Explanation:

Deploying optimized small language models (SLMs) on edge devices ensures low latency because the inference happens directly on the device without relying on cloud communication. Small language models are lightweight and designed to run efficiently on devices with limited resources, making them ideal for edge computing.

upvoted 3 times

🗨️ 👤 **Aryan_10** 1 month, 1 week ago

Selected Answer: A

Lowest latency possible - SLM

upvoted 1 times

🗨️ 👤 **Nicocacik** 1 month, 4 weeks ago

Selected Answer: A

Low latency with edge devices -> SLM

upvoted 1 times

🗨️ 👤 **Blair77** 2 months, 2 weeks ago

A is good - Minimal latency: SLMs are designed to run efficiently on resource-constrained devices, offering fast inference directly on the device.

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: A

SLM on edge devices

upvoted 2 times

🗨️ 👤 **tcusa** 2 months, 4 weeks ago

Selected Answer: A

SLM on edge devices is the correct solution.

upvoted 2 times

🗨️ 👤 **galliaj** 3 months ago

Using Optimized Small Language Models (SLMs) on edge devices is the best choice because they are designed to run efficiently within the resource constraints of edge hardware. This minimizes latency and helps deliver fast inference times while using less computational power and memory. The problem with trying to use centralized APIs is the associated latency.

upvoted 4 times

A company wants to build an ML model by using Amazon SageMaker. The company needs to share and manage variables for model development across multiple teams.

Which SageMaker feature meets these requirements?

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

Correct Answer: A

Community vote distribution

A (100%)

🗨️ **galliaj** Highly Voted 👍 3 months ago

Amazon SageMaker Feature Store ensures all teams have access to a centralized store of features, improving consistency and collaboration in ML workflows.

upvoted 7 times

🗨️ **elf78** 2 months, 1 week ago

<https://aws.amazon.com/sagemaker/feature-store/>

upvoted 1 times

🗨️ **85b5b55** Most Recent 🕒 2 days, 1 hour ago

Selected Answer: A

Amazon SageMaker Feature Store helps to create, store, share, manage features that are used in ML models.

upvoted 1 times

🗨️ **Moon** 1 month ago

Selected Answer: A

A: Amazon SageMaker Feature Store

Explanation:

Amazon SageMaker Feature Store is a purpose-built repository for storing, sharing, and managing features (variables) used in machine learning models. It allows teams to collaborate effectively by providing a centralized location for storing and accessing features across multiple ML workflows, ensuring consistency and reusability.

upvoted 1 times

🗨️ **Nicocacik** 1 month, 4 weeks ago

Selected Answer: A

A- Sagemaker Feature Store

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: A

A. Amazon SageMaker Feature Store

upvoted 2 times

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

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

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

Correct Answer: A

Community vote distribution

A (100%)

🗨️ **AzureDP900** 5 days, 16 hours ago

A is right

Amazon QuickStart (AWS QuickStart) is a set of pre-configured templates for AWS services that help developers quickly get started with various applications. It can create software snippets, provide reference tracking, and manage open-source license tracking to meet the company's requirements of increasing developer productivity and software development.

upvoted 1 times

🗨️ **Moon** 1 month ago

Selected Answer: A

A: Create software snippets, reference tracking, and open source license tracking.

Explanation:

Amazon Q Developer is a generative AI tool designed to assist developers by increasing productivity. It helps in generating software snippets, automating reference tracking, and managing open-source licenses, which directly benefits the software development lifecycle.

upvoted 3 times

🗨️ **monkeydba** 1 month, 2 weeks ago

Selected Answer: A

Open source license tracking in Q is discussed here. It's called "code references" <https://docs.aws.amazon.com/amazonq/latest/qdeveloper-ug/code-reference.html>

upvoted 1 times

🗨️ **eesa** 1 month, 3 weeks ago

Selected Answer: C

Amazon Q Developer está disponible en todos los AWS entornos y servicios, y también como asistente de codificación en terceros IDEs.

Muchas de las capacidades de Amazon Q Developer se encuentran en una interfaz de chat, en la que puede utilizar un lenguaje natural para hacer preguntas AWS, obtener ayuda con el código, explorar recursos o solucionar problemas. Cuando chateas con Amazon Q, Amazon Q utiliza el contexto de tu conversación actual para informar sus respuestas. Puedes hacer preguntas de seguimiento o consultar su respuesta cuando hagas una nueva pregunta.

https://docs.aws.amazon.com/es_es/amazonq/latest/qdeveloper-ug/features.html

upvoted 1 times

🗨️ **huanlt_cloud** 1 month, 3 weeks ago

Selected Answer: A

I chose answer A, but I'm not very sure about the "open source license tracking" of Amazon Q Developer. According to Amazon's official documentation, this issue is not mentioned <https://aws.amazon.com/q/developer/>

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: A

A for sure

upvoted 2 times

 **tccusa** 2 months, 4 weeks ago

Selected Answer: A

Amazon Q is designed to assist developers in all those things.

upvoted 3 times

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

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

Correct Answer: A

Community vote distribution

A (100%)

  **tccusa** Highly Voted  2 months, 4 weeks ago

Selected Answer: A

Privatelink allows secure, private connectivity to aws services.
upvoted 5 times

  **85b5b55** Most Recent  2 days, 1 hour ago

Selected Answer: A

AWS PrivateLink is the right options to avoid the any internet traffic. IG for internet traffic so we can't use it for this usecase.
upvoted 1 times

  **Moon** 1 month ago

Selected Answer: A

A: AWS PrivateLink



Explanation:

AWS PrivateLink is used to securely access AWS services from a VPC without exposing the traffic to the public internet. This ensures compliance with regulatory standards that prohibit internet access, as all communication happens over the private AWS network.
upvoted 1 times

  **eesa** 1 month, 3 weeks ago

Selected Answer: A

AWS PrivateLink enables secure, private connectivity between Virtual Private Cloud (VPC) environments and AWS services without exposing traffic to the public internet
upvoted 1 times

  **MarvelousV** 2 months, 3 weeks ago

Comment

upvoted 1 times

  **jove** 2 months, 3 weeks ago

Selected Answer: A

AWS PrivateLink enables secure, private connectivity between Virtual Private Cloud (VPC) environments and AWS services without exposing traffic to the public internet
upvoted 4 times

A company wants to develop an educational game where users answer questions such as the following: "A jar contains six red, four green, and three yellow marbles. What is the probability of choosing a green marble from the jar?"
Which solution meets these requirements with the LEAST operational overhead?

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

Correct Answer: C

Community vote distribution

C (100%)

afrazkhan 1 week, 5 days ago

Selected Answer: C

no NEED to do do anything fancy. its doable with simple code
upvoted 2 times

Moon 1 month ago

Selected Answer: C

C: Use code that will calculate probability by using simple rules and computations.

Explanation:

For a question like this, where the probability can be computed using basic arithmetic (e.g., number of favorable outcomes divided by total outcomes), implementing a straightforward function in code will meet the requirements with the least operational overhead. This avoids the complexity and resource demands of machine learning.

For example:

Total marbles =

6

+

4

+

3

=

13

$6+4+3=13$

upvoted 2 times

jove 2 months, 3 weeks ago

Selected Answer: C

Make it simple : Use code that will calculate probability by using simple rules and computations.
upvoted 4 times

tccusa 2 months, 4 weeks ago

Selected Answer: C

Not necessary to train a model for this. Code for computation is sufficient.
upvoted 3 times

Which metric measures the runtime efficiency of operating AI models?

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

Correct Answer: C

Community vote distribution

C (100%)

 **jove** Highly Voted 2 months, 3 weeks ago

Selected Answer: C

Average response time refers to the time taken by an AI model to produce a result after receiving an input. It is a critical metric for assessing the runtime efficiency of an AI model during inference, particularly in applications where quick responses are essential, such as in real-time applications or interactive systems.

upvoted 5 times

 **Moon** Most Recent 1 month ago


Selected Answer: C

C: Average response time

Explanation:

Average response time is a key metric for measuring the runtime efficiency of operating AI models. It indicates how quickly the AI model processes a request and returns a response, which is critical for assessing the performance and efficiency of deployed models, especially in real-time applications.

upvoted 1 times

 **ap6491** 1 month ago

Selected Answer: C

Average response time measures how quickly an AI model produces predictions or outputs during runtime, making it a key metric for evaluating the runtime efficiency of AI models.

It reflects the latency users experience when interacting with the model, which is especially critical for applications like chatbots, recommendation systems, or fraud detection.

upvoted 1 times

 **LR2023** 2 months, 3 weeks ago

Selected Answer: C

Yes, "average response time" is the primary metric used to measure the runtime efficiency of operating AI models, as it directly reflects how quickly a model can produce a prediction or response to a given input

upvoted 4 times

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

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

Correct Answer: B

Community vote distribution

B (100%)

85b5b55 2 days, 19 hours ago

Selected Answer: B

Transcribe support for speech to text - B
upvoted 1 times

Moon 1 month ago

Selected Answer: B

B: Transcribe call recordings by using Amazon Transcribe.

Explanation:

Amazon Transcribe is designed for converting speech in audio files (such as customer calls) into text. This text can then be analyzed further to extract key information. It is the first step in gaining insights from audio conversations, making it the appropriate solution for the given requirement.

upvoted 2 times

Bala416 2 months ago

Selected Answer: B

key word : FROM THE AUDIO OF THE CUSTOMER
upvoted 1 times

PHD_CHENG 2 months, 1 week ago

Selected Answer: B

B is correct. Question is related to audio to text. Amazon Transcribe fit on this aspect
upvoted 2 times

eesa 2 months, 2 weeks ago

select B

Amazon Transcribe is a service that converts audio into text, making it ideal for transcribing customer calls in a contact center. Once the audio is transcribed into text, you can further analyze the transcribed text to gain insights, such as identifying key information, customer sentiment, and specific topics discussed during the conversation.

upvoted 2 times

Udyan 2 months, 3 weeks ago

Amazon Transcribe is designed specifically to convert audio to text, which is a necessary first step for gaining insights from customer conversations. Once transcribed, the text data can be further processed and analyzed for key information.

Amazon Comprehend (option D) is useful for extracting insights from text, like sentiment analysis and entity extraction, but it only works on text data, not on audio files directly. So, Amazon Comprehend could be used after Amazon Transcribe has converted the audio to text, but it wouldn't be a standalone solution for handling the audio.

So, it is B only

upvoted 1 times

🗨️ 👤 **Soweetadad** 2 months, 3 weeks ago

In real world, we would run transcribe first to convert the call to searchable text, and then run Comprehend to search and analyze for specific key words. Since the question is around analyze and extract key info, I will go for "D"

upvoted 2 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: B

Transcribe

upvoted 2 times

🗨️ 👤 **LR2023** 2 months, 3 weeks ago

Selected Answer: B

transcribe - Extract key business insights from customer calls, video files, clinical conversations, and more.

upvoted 2 times

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

Which methodology should the company use to meet these requirements?

- A. Supervised learning
- B. Unsupervised learning
- C. Reinforcement learning
- D. Reinforcement learning from human feedback (RLHF)

Correct Answer: B

Community vote distribution

B (100%)

🗨️ **galliaj** Highly Voted 3 months ago

Because of the large amounts of unlabeled data and need to identify patterns or groupings within that data, Unsupervised learning is best. Clustering techniques can be used to classify customers into different tiers.

upvoted 6 times

🗨️ **85b5b55** Most Recent 2 days, 19 hours ago

Selected Answer: B

Unsupervised Learning handled the unlabeled datasets

upvoted 1 times

🗨️ **alexK** 3 weeks, 6 days ago

Selected Answer: B

Keyword - Unlabeled Data

upvoted 1 times

🗨️ **Moon** 1 month ago

Selected Answer: B

B: Unsupervised learning

Explanation:

Unsupervised learning is used when working with unlabeled data, such as the customer data described in this scenario. This methodology allows the company to identify patterns and group similar customers into clusters or tiers without the need for predefined labels. Techniques like clustering (e.g., K-Means or hierarchical clustering) would help classify customers based on shared characteristics for targeted advertisement campaigns.

Why not the other options?

A: Supervised learning:

Supervised learning requires labeled data, which is not available in this case. Labels would need to be provided for each customer, making this approach unsuitable for the given scenario.

upvoted 3 times

🗨️ **1176** 1 month, 2 weeks ago

Selected Answer: B

B is the answer..

upvoted 1 times

🗨️ **Udyan** 2 months, 3 weeks ago

Unlabeled Data - Unsupervised Learning

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: B

B. Unsupervised learning

upvoted 3 times

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

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

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

Correct Answer: A

Community vote distribution

A (100%)

🗳️ **galliaj** Highly Voted 3 months ago

Multi-modal embedding models can process multiple types of input data, such as text and images. This allows the search application to handle queries that involve both text and images effectively.

upvoted 9 times

🗳️ **jove** Highly Voted 2 months, 3 weeks ago

Selected Answer: A

queries that have text and images >>> Multi-modal embedding

upvoted 6 times

🗳️ **85b5b55** Most Recent 2 days, 19 hours ago

Selected Answer: A

Using multi-modal embedding to handle text and images.

upvoted 1 times

🗳️ **Moon** 1 month ago

Selected Answer: A

The answer is A. Multi-modal embedding model.

A multi-modal embedding model is a type of foundation model that can process and understand both text and images. This makes it suitable for powering a search application that handles queries containing both text and images.

Here's a breakdown of the other options:

B. Text embedding model: This type of model is only designed to process text data, so it wouldn't be suitable for handling image queries.

C. Multi-modal generation model: This type of model is designed to generate text or images, not to search for them.

D. Image generation model: This type of model is only designed to generate images, not to search for

them. <https://www.examtopycs.com/exams/amazon/aws-certified-ai-practitioner-aif-c01/view/5/#>

upvoted 2 times

🗳️ **may2021_r** 1 month ago

Selected Answer: A

The correct answer is A. A multi-modal embedding model can handle both text and image queries.

upvoted 1 times

🗳️ **eesa** 1 month, 3 weeks ago

Selected Answer: A

A multi-modal embedding model is specifically designed to process and understand various types of data, including text and images. By converting both text and image inputs into numerical representations (embeddings), it enables the model to compare and understand the relationships between them.

upvoted 1 times

🗳️ **RBSK** 2 months ago



Selected Answer: C

Output from GenAI (Confusing / Unclear Q) :- After carefully reviewing the search results, I can see that they do not specifically address the distinction between embedding and generation models in the context of the original query. The search results primarily discuss various types of foundation models and multimodal models, but they don't directly compare embedding and generation models for the specific search application mentioned in the question.

Given the lack of information directly relevant to the query in the provided search results, I cannot provide a definitive answer based on this information alone. The original question asks about using a foundation model for a search application that handles queries with text and images, but the search results don't contain specific information about embedding models for this purpose.

If you'd like a more accurate answer to this question, it would be helpful to have search results that specifically discuss embedding models and generation models in the context of multimodal search applications.

upvoted 1 times

  **Udyan** 2 months, 3 weeks ago

The search application must handle queries that have text and images.

Which type of FM should the AI practitioner use to power the search application, So, Multi Modal Embedding Model. For Result and Output, Multi-Modal Generation Model. Thus, Correct is A

upvoted 3 times

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

Which strategy will successfully fine-tune the model?

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

Correct Answer: A

Community vote distribution

A (100%)

🗨️ 👤 **ap6491** 1 month ago

Selected Answer: A

Fine-tuning a foundation model involves providing labeled training data where each example consists of a prompt (the input to the model) and a completion (the desired output). This structure helps the model learn specific patterns or behaviors tailored to the company's data and use case. In Amazon Bedrock, fine-tuning relies on a structured dataset that aligns with the model's learning requirements to improve its accuracy for domain-specific tasks.

upvoted 2 times

🗨️ 👤 **aldricstormcloak** 2 months, 1 week ago

Why is it not C? Finetuning cannot be done without provisioned throughput mode active.

upvoted 3 times

🗨️ 👤 **Dandelion2025** 1 month, 3 weeks ago

Fine-tuning and provisioned throughput are two separate processes in Amazon Bedrock:

Fine-tuning is the process of training the model on specific labeled data to improve its accuracy for particular tasks. This can be done without purchasing provisioned throughput. Provisioned throughput is required after fine-tuning, specifically for using the custom model for inference. It's not needed for the fine-tuning process itself. To test and deploy your model, you need to purchase Provisioned Throughput.

upvoted 4 times

🗨️ 👤 **Udyan** 2 months, 3 weeks ago

Fine Tuning is Done with Labeled Data so, A

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: A

Labeled Data: Fine-tuning requires labeled data

upvoted 3 times

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

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

Correct Answer: C



Community vote distribution

C (100%)

  **jove** Highly Voted 2 months, 3 weeks ago

Selected Answer: C

An anomaly detection system is designed to identify unusual patterns or behaviors within data
upvoted 6 times

  **galliaj** Highly Voted 3 months ago

Anomaly detection systems can be used to identify patterns that deviate from the norm. This makes them ideal for analyzing incoming IP addresses and flag suspicious IPs based on traffic patterns.
upvoted 5 times

  **85b5b55** Most Recent 2 days, 19 hours ago

Selected Answer: C

Using Anomalies, patterns, and behaviours refers to identifying the event.
upvoted 1 times

  **eesa** 1 month, 3 weeks ago

Selected Answer: C

Anomaly detection systems can be used to identify patterns that deviate from the norm. This makes them ideal for analyzing incoming IP addresses and flag suspicious IPs based on traffic patterns.
upvoted 1 times

  **Udyan** 2 months, 3 weeks ago

An anomaly detection system can analyze patterns and behaviors, such as IP address access patterns, to detect any deviations from the norm, which could indicate suspicious or malicious activity. An anomaly detection model can flag unusual access attempts, such as those from suspicious IP addresses, making it well-suited for threat detection.

Fraud forecasting (option D) typically focuses on predicting potential fraud patterns rather than real-time anomaly detection, so it would not directly address the need to check IP addresses for suspicious activity.

Thus, option C is the most suitable choice for identifying suspicious IP addresses in this scenario.

upvoted 4 times

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

- A. Integration with Amazon S3 for object storage
- B. Support for geospatial indexing and queries
- C. Scalable index management and nearest neighbor search capability
- D. Ability to perform real-time analysis on streaming data

Correct Answer: C

Community vote distribution

C (100%)

🗳️ 👤 **85b5b55** 2 days, 6 hours ago

Selected Answer: C

Scalable index management and k-NN algorithms which support to build and handle the recommendation systems, semantic search and anomalies detection.

upvoted 1 times

🗳️ 👤 **Moon** 1 month ago

Selected Answer: C

C: Scalable index management and nearest neighbor search capability

Explanation:

The Amazon OpenSearch Service supports building vector database applications by enabling nearest neighbor search capability. This feature allows the service to efficiently perform similarity searches, which is crucial for applications that rely on vector embeddings (e.g., recommendation systems, image or text similarity searches). Combined with scalable index management, this makes OpenSearch an excellent choice for vector database applications.

upvoted 1 times

🗳️ 👤 **ap6491** 1 month ago

Selected Answer: C

Amazon OpenSearch Service provides scalable index management and supports nearest neighbor (k-NN) search, which is essential for building vector database applications.

Vector databases store embeddings (numerical representations of data) and use k-NN search to retrieve similar data points based on proximity in the vector space, which is a foundational feature for applications such as recommendation systems, semantic search, and anomaly detection. These capabilities make OpenSearch ideal for developing vector-based applications.

upvoted 1 times

🗳️ 👤 **Blair77** 2 months, 2 weeks ago

c- The key feature of Amazon OpenSearch Service that enables companies to build vector database applications is its k-NN (k-nearest neighbors) functionality, specifically provided through the k-NN plugin. This allows OpenSearch Service to act as a vector database with efficient vector similarity search capabilities.

upvoted 1 times

🗳️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: C

Amazon OpenSearch Service provides scalable index management and nearest neighbor search capabilities, which are essential for building vector database applications.

upvoted 2 times

Which option is a use case for generative AI models?

- A. Improving network security by using intrusion detection systems
- B. Creating photorealistic images from text descriptions for digital marketing
- C. Enhancing database performance by using optimized indexing
- D. Analyzing financial data to forecast stock market trends

Correct Answer: B

Community vote distribution

B (100%)

85b5b55 2 days, 18 hours ago

Selected Answer: B

Generative AI is used to generate new content, images, video, and audio from existing content or new inputs from various data sources.
upvoted 1 times

kj07 4 weeks, 1 day ago

Selected Answer: B

Answer B. GenAI is used to generate content: text, images, code, etc.
upvoted 1 times

Moon 1 month ago

Selected Answer: B

The correct answer is B. Creating photorealistic images from text descriptions for digital marketing.

Generative AI models are designed to create new content, such as text, images, audio, or code. Creating images from text descriptions is a prime example of this capability.

Here's why the other options are not primarily use cases for generative AI:

A. Improving network security by using intrusion detection systems: While AI can be used for intrusion detection, this is more of a discriminative or predictive task (classifying network traffic as malicious or benign), not generating new content.

C. Enhancing database performance by using optimized indexing: This is related to database management and optimization, not content generation.

D. Analyzing financial data to forecast stock market trends: This involves statistical analysis and prediction based on existing data, again a predictive task, not generating new content.

upvoted 1 times

mia_khalifa 1 month ago

Selected Answer: D

Why not D ?

Because we can also fine tune model using historic data to predict market trends using gen AI.
upvoted 1 times

petarung 1 month, 3 weeks ago

Selected Answer: B

The correct answer is: B. Creating photorealistic images from text descriptions for digital marketing
Here's a detailed explanation:

Understanding Generative AI's Capabilities

Generative AI models, like DALL-E, Midjourney, and Stable Diffusion, are specifically designed to create new content based on text prompts. In this case, generating images from textual descriptions is a quintessential use case.

upvoted 1 times

jove 2 months, 3 weeks ago

Selected Answer: B

Generative AI models are designed to create new content, which includes generating images, text, audio, or other types of media based on input data

upvoted 3 times

A company wants to build a generative AI application by using Amazon Bedrock and needs to choose a foundation model (FM). The company wants to know how much information can fit into one prompt. Which consideration will inform the company's decision?

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

Correct Answer: B

Community vote distribution

B (100%)

85b5b55 2 days, 18 hours ago

Selected Answer: B

The context-window is the input prompt for the model generation.
upvoted 1 times

Moon 1 month ago

Selected Answer: B

A company needs to know the maximum input size for a single prompt when choosing a Foundation Model (FM) in Amazon Bedrock.

- A. Temperature: This controls the randomness of the output, not the input prompt length. Temperature affects creativity, not input size.
- B. Context window: This defines the maximum length of the input prompt the model can process. It directly limits how much information can be included.
- C. Batch size: This is the number of prompts processed at once, affecting throughput, not individual prompt length. It's about processing multiple prompts efficiently.
- D. Model size: This relates to the model's overall capacity and complexity, not directly to the input prompt length. Size impacts performance, not input limits.

Therefore, B. Context window is the correct answer.

upvoted 4 times

eesa 1 month, 3 weeks ago

Selected Answer: B

The correct answer is:

B. Context window

Explanation:

The context window of a foundation model (FM) determines how much information can fit into one prompt. It refers to the maximum number of tokens (words, characters, or subwords) that the model can process in a single input prompt, including the input and the output combined.

The context window size varies across different foundation models, and understanding this parameter is critical for applications like document summarization or question-answering systems where long inputs need to be processed.

upvoted 1 times

eesa 1 month, 3 weeks ago

Selected Answer: B

B. Context window

The context window of a foundation model determines the maximum amount of text that can be processed in a single prompt. A larger context window allows for more complex and informative prompts, while a smaller context window limits the amount of information that can be provided.

The other options are not directly related to the maximum prompt length:


Temperature: This parameter controls the randomness of the model's output.

Batch size: This refers to the number of samples processed in a single batch during training or inference.

Model size: This refers to the number of parameters in the model, which affects its complexity and performance.

Therefore, when choosing a foundation model for a generative AI application, the company should carefully consider the context window to ensure that it can accommodate the desired input length.

upvoted 2 times

 **jove** 2 months, 3 weeks ago

Selected Answer: B

The context window refers to the maximum number of tokens (words or pieces of words) that a foundation model can process in a single input prompt.

upvoted 2 times

A company wants to make a chatbot to help customers. The chatbot will help solve technical problems without human intervention. The company chose a foundation model (FM) for the chatbot. The chatbot needs to produce responses that adhere to company tone. Which solution meets these requirements?

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

Correct Answer: C

Community vote distribution

C (100%)

85b5b55 2 days, 18 hours ago

Selected Answer: C

Continued pre-training of the datasets to produce responses to the company's tone.
upvoted 1 times

nandhae 1 week, 2 days ago

Selected Answer: C

C. Experiment and refine the prompt until the FM produces the desired responses.

Refining the prompt is key to aligning the chatbot's responses with the company's tone and guidelines. Foundation models respond significantly to how prompts are phrased, making prompt engineering a powerful tool for achieving desired behavior.
upvoted 1 times

Moon 1 month ago

Selected Answer: C

C: Experiment and refine the prompt until the FM produces the desired responses.

Explanation:

To ensure that the chatbot adheres to the company's tone and provides appropriate responses, prompt engineering is essential. By experimenting and refining the prompt, you can guide the foundation model (FM) to produce responses that align with the desired tone, style, and content. This approach allows you to set the context and expectations for the chatbot's replies.
upvoted 1 times

ap6491 1 month ago

Selected Answer: C

Prompt engineering is the most effective way to ensure that a foundation model (FM) produces outputs adhering to a company's tone and specific requirements.

By iteratively testing and refining prompts, you can guide the FM to produce responses that align with the desired style, tone, and content accuracy.
upvoted 1 times

jove 2 months, 3 weeks ago

Selected Answer: C

Refining the prompt is the answer
upvoted 4 times

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

Which prompt engineering strategy meets these requirements?

- A. Provide examples of text passages with corresponding positive or negative labels in the prompt followed by the new text passage to be classified.
- B. Provide a detailed explanation of sentiment analysis and how LLMs work in the prompt.
- C. Provide the new text passage to be classified without any additional context or examples.
- D. Provide the new text passage with a few examples of unrelated tasks, such as text summarization or question answering.

Correct Answer: A

Community vote distribution

A (100%)

🗨️ 👤 **85b5b55** 2 days, 18 hours ago

Selected Answer: A

Set the proper label with a few examples to the prompts
upvoted 1 times

🗨️ 👤 **Moon** 1 month ago

Selected Answer: A

A: Provide examples of text passages with corresponding positive or negative labels in the prompt followed by the new text passage to be classified.

Explanation:

This strategy is known as few-shot prompting, where the prompt includes a few examples of labeled data (text passages with positive or negative sentiment) before asking the model to classify the new text passage. This helps the large language model (LLM) understand the task and align its output with the desired format.

Why not the other options?

B: Provide a detailed explanation of sentiment analysis and how LLMs work in the prompt:

Explaining the concept of sentiment analysis is unnecessary for the model, as it does not improve the model's ability to classify text.

C: Provide the new text passage to be classified without any additional context or examples:

Without examples, the LLM might not correctly infer the task or format of the output, leading to inconsistent or incorrect results.

upvoted 3 times

🗨️ 👤 **Gianiluca** 1 month, 1 week ago

Selected Answer: A

This approach uses few-shot learning, which is highly effective with large language models. By providing examples of text passages with their corresponding sentiment classifications, the LLM learns the context and pattern needed to classify the new passage.

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: A

Explanation:

By providing examples of text passages along with their corresponding sentiment labels (positive or negative), the model can learn from these examples how to classify the sentiment of the new text passage effectively

upvoted 4 times

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

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

- A. AWS Audit Manager
- B. AWS CloudTrail
- C. Amazon Fraud Detector
- D. AWS Trusted Advisor

Correct Answer: B

Community vote distribution

B (100%)

85b5b55 2 days, 18 hours ago

Selected Answer: B

Use AWS CloudTrail to track the API Calls to AWS resources.

upvoted 1 times

nandhae 1 week, 1 day ago

Selected Answer: B

B. AWS CloudTrail

CloudTrail records API activity and user actions in your AWS account. It logs events such as unauthorized access attempts to Amazon Bedrock and other AWS services, making it the correct choice for identifying such attempts.

upvoted 1 times

Moon 1 month ago

Selected Answer: B

B: AWS CloudTrail

Explanation:

AWS CloudTrail is a service that records all API calls and user activity across AWS services, including Amazon Bedrock. By analyzing CloudTrail logs, the company can identify unauthorized access attempts, track user activity, and audit the usage of foundation models. This information helps in setting appropriate AWS Identity and Access Management (IAM) policies and roles for future iterations of the models.

upvoted 2 times

jove 2 months, 3 weeks ago

Selected Answer: B

B. AWS CloudTrail is the most suitable service for identifying unauthorized access attempts to Amazon Bedrock, as it provides detailed logging and monitoring of API calls across AWS services, helping to enforce security and compliance.

upvoted 3 times

A company has developed an ML model for image classification. The company wants to deploy the model to production so that a web application can use the model.

The company needs to implement a solution to host the model and serve predictions without managing any of the underlying infrastructure. Which solution will meet these requirements?

- A. Use Amazon SageMaker Serverless Inference to deploy the model.
- B. Use Amazon CloudFront to deploy the model.
- C. Use Amazon API Gateway to host the model and serve predictions.
- D. Use AWS Batch to host the model and serve predictions.

Correct Answer: A

  **85b5b55** 2 days, 18 hours ago

Selected Answer: A

Amazon SageMaker helps to host the model, and serve predictions without managing infrastructure provisioning and configurations.
upvoted 1 times

  **nandhae** 1 week, 1 day ago

Selected Answer: A

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

Amazon SageMaker Serverless Inference is specifically designed for hosting ML models and serving predictions without requiring the management of underlying infrastructure. It automatically provisions compute resources as needed and is ideal for use cases like the one described.

upvoted 1 times

  **Moon** 1 month ago

Selected Answer: A

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

Explanation:

Amazon SageMaker Serverless Inference is a fully managed solution for deploying machine learning models without managing the underlying infrastructure. It automatically provisions compute capacity, scales based on request traffic, and serves predictions efficiently. This makes it an ideal choice for hosting a model and serving predictions for a web application with minimal management overhead.

Why not the other options?

B: Use Amazon CloudFront to deploy the model:

Amazon CloudFront is a content delivery network (CDN)

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

Amazon API Gateway is used to create APIs for accessing services.

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

AWS Batch is designed for batch processing and job scheduling, not for real-time inference or hosting ML models for web applications.

upvoted 1 times

  **Blair77** 2 months, 2 weeks ago

Selected Answer: A

Serverless deployment: SageMaker Serverless Inference allows you to deploy ML models without managing any underlying infrastructure, which directly meets the company's requirement.

upvoted 1 times

  **minime** 2 months, 3 weeks ago

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

With serverless inference, there's no need to manage any infra.

upvoted 1 times

An AI company periodically evaluates its systems and processes with the help of independent software vendors (ISVs). The company needs to receive email message notifications when an ISV's compliance reports become available. Which AWS service can the company use to meet this requirement?

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

Correct Answer: B

Community vote distribution

B (100%)

85b5b55 2 days, 18 hours ago

Selected Answer: B

AWS Artifact to keep the documents including security compliances and reports.
upvoted 1 times

2025AIMLPractitioner 5 days, 16 hours ago

Selected Answer: B

AWS Artifact is the service that provides on-demand access to AWS security and compliance reports.
upvoted 1 times

nandhae 1 week, 1 day ago

Selected Answer: B

B. AWS Artifact

AWS Artifact is a central resource for accessing compliance-related documents and reports, such as those provided by independent software vendors (ISVs). Users can subscribe to notifications to receive alerts when new compliance reports are available. This makes it the correct choice.

upvoted 1 times

kerl 3 weeks, 6 days ago

Selected Answer: B

<https://docs.aws.amazon.com/artifact/latest/ug/what-is-aws-artifact.html>
upvoted 1 times

Moon 1 month ago

Selected Answer: B

B: AWS Artifact

Explanation:

AWS Artifact is a service that provides on-demand access to AWS compliance reports, including those from independent software vendors (ISVs). AWS Artifact can notify users when new compliance reports are available, ensuring that the company stays updated and can evaluate its systems and processes accordingly.

D: AWS Data Exchange:

AWS Data Exchange is used for subscribing to and managing third-party data sets. It is not intended for compliance reports or notifications about them.

Conclusion:

AWS Artifact is the best choice for accessing and receiving notifications about compliance reports from independent software vendors (ISVs).
upvoted 2 times

KevinKas 1 month ago

Selected Answer: B

Companies can use notification settings in AWS Artifact to receive email alerts when new compliance reports or updates become available, ensuring they stay informed about the latest reports.

upvoted 1 times

🗨️ **HarishRao** 1 month ago

Selected Answer: B

<https://docs.aws.amazon.com/artifact/latest/ug/what-is-aws-artifact.html>

With AWS Artifact, you can also download security and compliance documents for independent software vendors (ISVs) who sell their products on AWS Marketplace.

upvoted 1 times

🗨️ **may2021_r** 1 month ago

Selected Answer: D

The correct answer is D. AWS Data Exchange enables receiving and managing third-party data including compliance reports.

upvoted 1 times

🗨️ **EDoubleU** 1 month, 1 week ago

Selected Answer: B

B

<https://docs.aws.amazon.com/artifact/latest/ug/managing-notifications.html#:~:text=You%20can%20use%20the%20AWS,notifications%20using%20AWS%20User%20Notifications.>

upvoted 2 times

🗨️ **OMBR** 1 month, 2 weeks ago

Selected Answer: B

<https://aws.amazon.com/about-aws/whats-new/2023/01/aws-artifact-on-demand-third-party-compliance-reports/>

upvoted 3 times

🗨️ **RY66** 2 months, 1 week ago

The correct answer to this question is B. AWS Artifact

AWS Artifact is the service that provides on-demand access to AWS security and compliance reports.

It allows users to access various compliance reports such as ISO certifications, PCI reports, and SOC reports.

Specifically, AWS Artifact Notifications feature allows users to receive email notifications when new reports become available.

This directly meets the requirement stated in the question: "The company needs to receive email message notifications when an ISV's compliance reports become available."

upvoted 2 times

🗨️ **dittle1977** 2 months, 2 weeks ago

The correct answer is B.

<https://docs.aws.amazon.com/artifact/latest/ug/what-is-aws-artifact.html>

upvoted 4 times

🗨️ **avi260919851985** 2 months, 2 weeks ago

D. Aws Data Exchange

upvoted 2 times

🗨️ **fed6485** 2 months, 2 weeks ago

Selected Answer: D

D. AWS Data Exchange, this is related to a third party, while AWS Artifact enables you to download AWS security and compliance documents such as ISO certifications and SOC reports.

upvoted 2 times

🗨️ **Blair77** 2 months, 2 weeks ago

AWS Data Exchange allows customers to securely exchange data with third parties but is not focused on compliance reporting or notifications related to ISVs



upvoted 1 times

🗨️ **Blair77** 2 months, 2 weeks ago

Selected Answer: B

Compliance report access: AWS Artifact provides on-demand access to AWS security and compliance reports, including those from Independent Software Vendors (ISVs) who sell their products on AWS Marketplace. AWS Data Exchange: This service is for finding, subscribing to, and using third-party data in the cloud, but it's not specifically designed for compliance reports or notifications.


upvoted 1 times

  **leyunjohn** 2 months, 3 weeks ago

Selected Answer: D

D. AWS Data Exchange

upvoted 3 times

  **dehkon** 2 months, 3 weeks ago

D. AWS Data Exchange

AWS Data Exchange allows subscribers to find, subscribe to, and use third-party data, including compliance reports from Independent Software Vendors (ISVs). The service can provide notifications when new data sets, such as compliance reports, are available from subscribed ISVs.

upvoted 3 times

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

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

Correct Answer: A

Community vote distribution

A (100%)

85b5b55 1 day, 2 hours ago

Selected Answer: A

Ask model to use Prompt template to avoid the various types of prompt injection attacks.
upvoted 1 times

ap6491 1 month ago

Selected Answer: A

Creating a prompt template that teaches the LLM to identify and resist common prompt engineering attacks, such as prompt injection or adversarial queries, helps prevent manipulation.
By explicitly guiding the LLM to ignore requests that deviate from its intended purpose (e.g., "You are a helpful assistant. Do not perform any tasks outside your defined scope."), you can mitigate risks like exposing sensitive information or executing undesirable actions.
upvoted 1 times

jove 2 months, 3 weeks ago

Selected Answer: A

A. Create a prompt template that teaches the LLM to detect attack patterns is the best action to reduce the risks associated with prompt manipulation and to enhance the security and integrity of the conversational agent being developed.
upvoted 2 times

A company is using the Generative AI Security Scoping Matrix to assess security responsibilities for its solutions. The company has identified four different solution scopes based on the matrix.

Which solution scope gives the company the MOST ownership of security responsibilities?

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

Correct Answer: D

Community vote distribution

D (100%)

🗨️ **Moon** 1 month ago

Selected Answer: D

D: Building and training a generative AI model from scratch by using specific data that a customer owns.

Explanation:

When a company builds and trains a generative AI model from scratch, it assumes the most ownership of security responsibilities, including:

Data security and compliance during training.

Model development and training processes.

Infrastructure and deployment security.

Protecting the model from adversarial attacks.

Ensuring ethical use of the model and safeguarding against bias and misuse.

This approach provides complete control over the entire lifecycle of the AI solution but also places the greatest burden of responsibility on the company.

upvoted 1 times

🗨️ **kyo** 1 month, 3 weeks ago

Selected Answer: D

<https://aws.amazon.com/ai/generative-ai/security/scoping-matrix/>

upvoted 1 times

🗨️ **eesa** 1 month, 3 weeks ago

Selected Answer: D

D. Building and training a generative AI model from scratch by using specific data that a customer owns.

In this scenario, the company has the most control over the entire development and deployment process. This includes:

Data security: The company is responsible for securing the training data, which might contain sensitive information.

Model security: The company needs to implement measures to protect the model itself, including securing the training process, model parameters, and deployment infrastructure.

Operational security: The company is responsible for securing the deployment environment and monitoring the model for potential vulnerabilities.

While the other options involve some level of security responsibility, they rely on third-party providers to a greater extent. This reduces the company's direct control over the security aspects of the solution.

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: D

D. Building and training a generative AI model from scratch by using specific data that a customer owns gives the company the most ownership of security responsibilities, as they are responsible for all aspects of the model's development and deployment.

upvoted 3 times

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

- A. Object detection
- B. Anomaly detection
- C. Named entity recognition
- D. Inpainting

Correct Answer: A

Community vote distribution

A (100%)

85b5b55 1 day, 2 hours ago

Selected Answer: A

Object detection helps to identify and category the object of the image.
upvoted 1 times

Moon 1 month ago

Selected Answer: A

A: Object detection

Explanation:

Object detection is a computer vision technique that identifies and categorizes objects within an image. In this scenario, it can automatically detect animals in the photos and assign them to categories (e.g., "dog," "cat," "bird"). This approach aligns perfectly with the requirement to identify and categorize animals without manual intervention.
upvoted 1 times

Gianiluca 1 month ago

Selected Answer: A

A. Object Detection
Identifies and locates objects in images
Can classify different types of objects (like animals)
Works automatically on images
Perfect for categorizing visual content
upvoted 1 times

Blair77 2 months, 2 weeks ago

Selected Answer: A

A . Object detection
upvoted 2 times

jove 2 months, 3 weeks ago

Selected Answer: A

. Object detection is the most appropriate strategy for automatically identifying and categorizing animals in a database of photos without manual human effort.
upvoted 3 times

A company wants to create an application by using Amazon Bedrock. The company has a limited budget and prefers flexibility without long-term commitment.

Which Amazon Bedrock pricing model meets these requirements?

- A. On-Demand
- B. Model customization
- C. Provisioned Throughput
- D. Spot Instance

Correct Answer: A

Community vote distribution

A (100%)

85b5b55 1 day, 2 hours ago

Selected Answer: A

On-Demand pricing plan helps to run the application in the temporary mode.

upvoted 1 times

Moon 1 month ago

Selected Answer: A

A: On-Demand

Explanation:

The On-Demand pricing model for Amazon Bedrock provides flexibility and allows the company to pay only for what they use, without requiring long-term commitments or upfront payments. This is ideal for a company with a limited budget that needs to control costs while maintaining flexibility.

D: Spot Instance:

Spot Instances are an AWS EC2 pricing model for obtaining unused compute capacity at discounted rates. They are not applicable to Amazon Bedrock, which does not rely on Spot Instances.

upvoted 1 times

jove 2 months, 3 weeks ago

Selected Answer: A

On-Demand is the best pricing model for a company that has a limited budget and wants flexibility without long-term commitment when creating an application using Amazon Bedrock.

upvoted 2 times

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

- A. Amazon Personalize
- B. Amazon SageMaker JumpStart
- C. PartyRock, an Amazon Bedrock Playground
- D. Amazon SageMaker endpoints

Correct Answer: B

Community vote distribution

B (100%)

85b5b55 1 day, 2 hours ago

Selected Answer: B

Amazon SageMaker JumpStart helps to deploy pre-trained Open-sourced models quickly.
upvoted 1 times

dspd 2 days, 17 hours ago

Selected Answer: B

The correct answer is B: Amazon SageMaker JumpStart.

Here's why:

Amazon SageMaker JumpStart is specifically designed to help teams quickly deploy and use foundation models (FMs) with the following benefits:

- Provides pre-trained models that can be deployed with just a few clicks
- Allows deployment within your VPC for secure access
- Includes popular foundation models from various providers
- Offers fine-tuning capabilities for customization
- Handles the infrastructure management automatically

Amazon SageMaker endpoints - While these are used to deploy models, SageMaker JumpStart provides a more complete solution specifically for foundation models with built-in deployment capabilities

upvoted 1 times

waldonuts 2 weeks, 1 day ago

Selected Answer: D

I lean towards Sagemaker Endpoints . to my knowledge Jumpstart will help you select/deploy the model, but to actually use it/consume it in your Prod/dev environment/VPC you need the Endpoint
upvoted 2 times

scs50 2 weeks, 6 days ago

Selected Answer: B

Amazon SageMaker JumpStart provides security features, including the ability to integrate with a Virtual Private Cloud (VPC), ensuring secure communication and data transfer during machine learning tasks. SageMaker Jumpstart simplifies the process of building, training, and deploying ML models by offering ready-to-use resources and templates.
upvoted 1 times

Aswiz 3 weeks, 6 days ago

Selected Answer: B

for quick access we can use jumpstart
upvoted 1 times

Moon 1 month ago

Selected Answer: D

he question asks about quickly deploying and consuming an FM within the team's VPC.

A. Amazon Personalize: This is for building recommendation systems, not general FM deployment or consumption. It's irrelevant to the question.

B. Amazon SageMaker JumpStart: JumpStart provides a quick way to find and deploy pre-trained models. However, the initial deployment is not automatically within your VPC. You need to configure the endpoint settings during deployment to specify your VPC. Therefore, while it speeds up the process of getting a model ready, it doesn't directly fulfill the "within the team's VPC" requirement without extra steps.

D. Amazon SageMaker endpoints: This is the most accurate answer. While JumpStart can help you get a model ready, it's the SageMaker endpoint itself that is configured to reside within your VPC. You create the endpoint and specify the VPC configuration during that endpoint creation.
upvoted 1 times

🗨️ **may2021_r** 1 month ago

Selected Answer: B

Let me explain why Amazon SageMaker JumpStart (Option B) is the correct answer:

1. VPC Integration: SageMaker JumpStart allows deployment of foundation models within your team's VPC, ensuring secure access and network isolation.

2. Quick Deployment: It provides a streamlined process for deploying pre-trained foundation models with minimal setup required. The service includes:

- One-click deployment options
- Pre-configured model endpoints
- Built-in model optimization

3. Foundation Model Support: SageMaker JumpStart specifically offers a wide range of foundation models that are ready to use.
upvoted 1 times

🗨️ **Chika22** 1 month, 3 weeks ago

Selected Answer: B

Amazon SageMaker JumpStart
upvoted 2 times

🗨️ **Contactfortitish** 1 month, 4 weeks ago

Selected Answer: D

Amazon SageMaker endpoints allow you to deploy machine learning models, including foundation models, for real-time inference within a Virtual Private Cloud (VPC). This feature is particularly suitable for AI teams looking to host and consume their models securely and quickly.

Amazon SageMaker JumpStart: While JumpStart provides prebuilt solutions and model deployment templates, it is not specifically focused on VPC integration for foundation models.

upvoted 1 times

🗨️ **Oc2d840** 2 months ago

Selected Answer: B

It could be B or D as question says Service or Feature.

Why D got eliminated? - Even though it says Service or Feature, I think that is just because SageMaker itself is an umbrella for many services and features. Like SageMaker studio itself has many features. SageMaker endpoint is not a feature per say, but the deployment environment for models.

upvoted 2 times

🗨️ **eesa** 2 months ago

Selected Answer: B

B. Amazon SageMaker JumpStart

Amazon SageMaker JumpStart provides a collection of pre-trained models, including foundation models, that can be easily deployed and customized within a team's VPC. This allows for secure and efficient access to these powerful models without exposing them to the public internet

upvoted 2 times

🗨️ **RY66** 2 months, 1 week ago

The correct answer to this question is B. Amazon SageMaker JumpStart.

Amazon SageMaker JumpStart is a service that provides pre-trained models, solutions, and examples to help quickly start machine learning tasks.

JumpStart includes a variety of foundation models (FMs) and offers features to easily deploy and fine-tune these models. Importantly, models deployed through JumpStart can be run securely within a team's VPC, which aligns with the question's requirement of deploying and consuming a foundation model within the team's VPC.

JumpStart enables quick deployment and consumption of models, satisfying the "quickly deploy and consume" part of the question.
upvoted 1 times

🗨️ **fed6485** 2 months, 2 weeks ago

Selected Answer: D

.. AWS FEATURE can help .. and CONSUME a foundation model (FM) within the team's VPC?

upvoted 2 times

🗨️ **fed6485** 2 months, 2 weeks ago

sorry i didn't notice i have already commented/answer on this.

upvoted 1 times

🗨️ **fed6485** 2 months, 2 weeks ago

Selected Answer: D

... mmm.. interesting one as..

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

the fact that "AWS service or FEATURE" .. deploy within the team's VPC..

definitely or B or D

B if the question refers to the SERVICE

D if the question refers to the FEATURE

:)

upvoted 2 times

🗨️ **0c2d840** 2 months ago

Answer is A. Even though it says Service or Feature, I think that is just because SageMaker itself is an umbrella for many services and features. Like SageMaker studio itself has many features. SageMaker endpoint is not a feature per say, but the deployment environment for models.

upvoted 1 times

🗨️ **raat** 2 months, 3 weeks ago

Amazon SageMaker JumpStart (option B) is indeed a valuable service for quickly getting started with pre-built models and solutions. However, it is more focused on providing a range of pre-trained models and example solutions to help you get started with machine learning projects.

For the specific requirement of deploying and consuming a foundation model within your VPC, Amazon SageMaker endpoints (option D) are more directly suited. They allow you to deploy models for real-time inference securely within your VPC, ensuring that your data and model interactions remain within your private network.

If you have any more questions or need further clarification, feel free to ask!

upvoted 2 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: B

B. Amazon SageMaker JumpStart is the best option for quickly deploying and consuming a foundation model within a team's VPC, as it streamlines the process and provides ready-to-use resources.

upvoted 3 times

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

- A. Design clear and specific prompts. Configure AWS Identity and Access Management (IAM) roles and policies by using least privilege access.
- B. Enable AWS Audit Manager for automatic model evaluation jobs.
- C. Enable Amazon Bedrock automatic model evaluation jobs.
- D. Use Amazon CloudWatch Logs to make models explainable and to monitor for bias.

Correct Answer: A

Community vote distribution

A (100%)

85b5b55 1 day, 2 hours ago

Selected Answer: A

Using IAM with least privilege will secure the LLM on the Amazon Bedrock.

upvoted 1 times

eesa 1 month, 3 weeks ago

Selected Answer: A

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

This option addresses two key aspects of secure LLM usage on Amazon Bedrock:

Prompt Engineering: Clear and specific prompts reduce the risk of unintended or harmful outputs. Well-defined prompts help guide the model's responses and minimize the potential for bias or misinformation.

IAM Access Control: Implementing strong access controls is crucial to protect sensitive data and prevent unauthorized access to the LLM. By using IAM roles and policies with least privilege access, you can limit permissions to only the necessary actions, reducing the risk of security breaches.

upvoted 1 times

jove 2 months, 3 weeks ago

Selected Answer: A

A. Design clear and specific prompts. Configure AWS Identity and Access Management (IAM) roles and policies by using least privilege access is the best approach for companies to securely use large language models on Amazon Bedrock, as it emphasizes both prompt clarity and access control.

upvoted 2 times

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

- A. Generative pre-trained transformers (GPT)
- B. Residual neural network
- C. Support vector machine
- D. WaveNet

Correct Answer: A

Community vote distribution

A (100%)

🗨️ **85b5b55** 1 day, 2 hours ago

Selected Answer: A

GPT helps to produce the NL based response based on the input text.

upvoted 1 times

🗨️ **Moon** 4 weeks, 1 day ago

Selected Answer: A

The best solution for building an AI-based application that translates natural language (employee input text) into SQL queries is A. Generative pre-trained transformers (GPT).

Here's why:

GPT's strength in natural language processing: GPT models are specifically designed for understanding and generating human language. They excel at tasks like text translation, question answering, and, crucially, code generation from natural language descriptions. This makes them ideal for converting employee input into SQL queries.

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: A

Generative pre-trained transformers (GPT) are powerful natural language processing models that excel in understanding and generating human-like text. In this scenario, a GPT model can be trained or fine-tuned to take natural language input from employees and convert it into structured SQL queries. This makes it accessible for users who may not have technical expertise, allowing them to retrieve the data they need from the database using simple, conversational prompts.

upvoted 2 times

A company built a deep learning model for object detection and deployed the model to production.

Which AI process occurs when the model analyzes a new image to identify objects?

- A. Training
- B. Inference
- C. Model deployment
- D. Bias correction

Correct Answer: B

🗨️ 👤 **85b5b55** 1 day, 2 hours ago

Selected Answer: B

During the inference phase, model analyses a new image to identify objects.

upvoted 1 times

🗨️ 👤 **eesa** 1 month, 3 weeks ago

Selected Answer: B

B. Inference

Inference is the process of using a trained model to make predictions or decisions on new, unseen data. In the case of an object detection model, inference involves feeding a new image into the model, which then analyzes the image and outputs the detected objects and their locations.

upvoted 1 times

🗨️ 👤 **urbanmonk** 1 month, 3 weeks ago

Selected Answer: B

AI inference is the process that a trained machine learning model uses to draw conclusions from brand-new data. An AI model capable of making inferences can do so without examples of the desired result.

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: B

It's the inference

upvoted 4 times

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

Which technique will solve the problem?

- A. Data augmentation for imbalanced classes
- B. Model monitoring for class distribution
- C. Retrieval Augmented Generation (RAG)
- D. Watermark detection for images

Correct Answer: A

Community vote distribution

A (100%)

🗨️ **eesa** 1 month, 3 weeks ago

Selected Answer: A

Data augmentation for imbalanced classes

Data augmentation techniques can help mitigate bias in image generation models by artificially increasing the diversity of the training data. By applying transformations like rotations, flips, and color jittering to existing images, you can create new, synthetic images that are similar to the original ones. This can help balance the dataset and reduce the impact of biases present in the original data.

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: A

A. Data augmentation for imbalanced classes is the most effective technique to mitigate bias in the input data by ensuring a more balanced representation of classes and attributes in the training set, leading to fairer and more accurate image generation.

upvoted 2 times

A company is implementing the Amazon Titan foundation model (FM) by using Amazon Bedrock. The company needs to supplement the model by using relevant data from the company's private data sources.
Which solution will meet this requirement?

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

Correct Answer: C

🗨️ 👤 **85b5b55** 1 day, 2 hours ago

Selected Answer: C

Amazon Bedrock KB support to interact with the company's private data sources.
upvoted 1 times

🗨️ 👤 **eesa** 1 month, 3 weeks ago

Selected Answer: C

Create an Amazon Bedrock knowledge base.

An Amazon Bedrock knowledge base allows you to incorporate your company's proprietary data into the foundation model. By feeding the model with relevant information, you can enhance its ability to generate more accurate and informative responses.
upvoted 1 times

🗨️ 👤 **raat** 2 months, 2 weeks ago

Selected Answer: C

C, is correct
upvoted 1 times

🗨️ 👤 **minime** 2 months, 3 weeks ago

C. Create an Amazon Bedrock knowledge base.

This would allow the company use the knowledge base for Retrieval Augmented Generation (RAG) to enhance the model's knowledge with company's private data sources.
upvoted 4 times

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

Which solution will meet these requirements?

- A. Configure the security and compliance by using Amazon Inspector.
- B. Generate simple metrics, reports, and examples by using Amazon SageMaker Clarify.
- C. Encrypt and secure training data by using Amazon Macie.
- D. Gather more data. Use Amazon Rekognition to add custom labels to the data.

Correct Answer: B

Community vote distribution

B (100%)

 **eesa** 1 month, 3 weeks ago

Selected Answer: B

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

Amazon SageMaker Clarify helps in identifying bias and explaining predictions made by machine learning models, which aligns well with the need for transparency and explainability to meet regulatory requirements.

upvoted 1 times

 **jove** 2 months, 3 weeks ago

Selected Answer: B

Amazon SageMaker Clarify is specifically designed to help make machine learning models more transparent and explainable by generating metrics and reports on model bias, data bias, and feature importance.

upvoted 4 times

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

- A. Auto scaling inference endpoints
- B. Threat detection
- C. Data protection
- D. Cost optimization
- E. Loosely coupled microservices

Correct Answer: BC

Community vote distribution

BC (100%)

85b5b55 1 day, 2 hours ago

Selected Answer: BC

Threat (Amazon GuardDuty) and Data Protection (Amazon Macie, KMS, Encrypt the data at REST and in-Transit).
upvoted 1 times

Moon 1 month ago

Selected Answer: BC

Why not the other options?

A: Auto scaling inference endpoints:

Auto-scaling improves performance and cost-efficiency but is not directly related to regulatory compliance.

D: Cost optimization:

Cost optimization is beneficial for managing expenses but is not a compliance requirement.

E: Loosely coupled microservices:

While a good architectural principle, it does not directly address compliance with regulatory frameworks.

upvoted 1 times

Moon 1 month ago

Selected Answer: BC

B: Threat detection

C: Data protection

Explanation:

When deploying a conversational chatbot using a fine-tuned model from Amazon SageMaker JumpStart, the company can demonstrate compliance in the following areas:

B: Threat detection: Amazon SageMaker integrates with AWS security services like Amazon GuardDuty and AWS CloudTrail to monitor for threats and unauthorized access. This ensures compliance with security regulations.

C: Data protection: SageMaker supports encryption of data at rest and in transit, integration with AWS Key Management Service (KMS), and fine-grained access control through IAM. These features ensure compliance with regulatory frameworks requiring data protection.

upvoted 1 times

eesa 1 month, 3 weeks ago

Selected Answer: BC

The two capabilities that the company can show compliance for are:

C. Data protection

B. Threat detection

Here's a breakdown:

Data Protection:

Amazon SageMaker offers robust data protection features, including data encryption at rest and in transit.

By leveraging these features, the company can ensure that customer data is handled securely and complies with relevant data privacy regulations.

Threat Detection:

Amazon Web Services (AWS) provides a comprehensive security suite, including services like Amazon GuardDuty and AWS Security Hub.

These services can help detect and respond to potential threats, such as unauthorized access, data breaches, and malicious activity.

By utilizing these services, the company can demonstrate its commitment to security and compliance.

upvoted 2 times

  **urbanmonk** 1 month, 3 weeks ago

Selected Answer: C

Data Protection - certainly.

Not sure which other option fits into the regulatory context.

upvoted 1 times

  **RY66** 2 months, 1 week ago

The correct answers for this question are:

A. Auto scaling inference endpoints

C. Data protection

Auto scaling inference endpoints:

Amazon SageMaker provides auto-scaling capabilities that automatically adjust infrastructure based on traffic changes.

This helps meet availability and performance requirements, which are crucial aspects of regulatory compliance.

Many regulatory frameworks require service stability and availability, making this feature an important element in demonstrating compliance.

Data protection:

Data protection is a core requirement in most regulatory frameworks.

Amazon SageMaker offers various data protection features including data encryption, access control, and audit logging.

For a chatbot handling customer data, demonstrating data protection capabilities is essential for regulatory compliance.

upvoted 1 times

  **jove** 2 months, 3 weeks ago

Selected Answer: BC

C. Data protection and B. Threat detection are the two key capabilities that can help the company meet regulatory compliance requirements when deploying a conversational chatbot using Amazon SageMaker JumpStart.

upvoted 4 times

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

- A. Decrease the batch size.
- B. Increase the epochs.
- C. Decrease the epochs.
- D. Increase the temperature parameter.

Correct Answer: B

🗨️ 👤 **Moon** 1 month ago

Selected Answer: B

B: Increase the epochs.

Explanation:

Increasing the epochs allows the model to go through the entire training dataset multiple times, improving its learning and optimizing its weights. This can help the model achieve a higher accuracy level, provided it does not lead to overfitting. For a foundation model (FM), increasing epochs is a common approach to refining accuracy to meet specific acceptance levels.

upvoted 1 times

🗨️ 👤 **eesa** 1 month, 3 weeks ago

Selected Answer: B

B. Increase the epochs.

Increasing the number of epochs, or training cycles, can help improve the accuracy of a foundation model. By exposing the model to the training data multiple times, it can learn more intricate patterns and relationships, leading to better performance.

upvoted 2 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: B

B. Increase the epochs: Increasing the number of epochs allows the model to continue learning from the data, potentially improving its accuracy as it trains on more examples. However, there is a risk of overfitting if epochs are increased too much.

upvoted 3 times

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

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

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

Correct Answer: B

🗨️ 👤 **Moon** 1 month ago

Selected Answer: B

B: Average call duration

Explanation:

Average call duration is a key metric for evaluating the efficiency of a question-answering chatbot in a call center environment. By reducing the number of actions employees need to take, the chatbot can help streamline customer interactions, resulting in shorter call durations. Monitoring this metric helps the company assess whether the chatbot is achieving its goal of improving call center efficiency.

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: B

Obviously it is B

upvoted 2 times

Which functionality does Amazon SageMaker Clarify provide?

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

Correct Answer: D

Community vote distribution

D (100%)

 **jove** Highly Voted 2 months, 3 weeks ago

Selected Answer: D

Amazon SageMaker Clarify provides functionality to detect and identify potential bias in data both before and after training, helping teams uncover imbalances in datasets that might lead to biased model predictions. This is essential for ensuring fairness and compliance, especially in sensitive applications.

Why Not the Other Options?

- A. Integrates a Retrieval Augmented Generation (RAG) workflow: RAG workflows are used for combining retrieved documents with model outputs, typically in language models, but this is not a function of SageMaker Clarify.
- B. Monitors the quality of ML models in production: Monitoring model quality in production is handled by SageMaker Model Monitor, not SageMaker Clarify.
- C. Documents critical details about ML models: This functionality is part of Amazon SageMaker Model Cards, which documents model details for transparency and compliance.

upvoted 5 times

 **85b5b55** Most Recent 1 day, 1 hour ago

Selected Answer: D

Amazon Sagemake Clarify helps to Identify bias, how much models makes prediction, datasets or models reflections and more.

upvoted 1 times

 **eesa** 1 month, 3 weeks ago

Selected Answer: D

D. Identifies potential bias during data preparation

Amazon SageMaker Clarify is a tool designed to help understand, debug, and improve machine learning models. One of its key functionalities is to identify potential bias in datasets and models. It can analyze datasets for imbalances, fairness issues, and other biases that could impact the model's performance and fairness

upvoted 2 times

A company is developing a new model to predict the prices of specific items. The model performed well on the training dataset. When the company deployed the model to production, the model's performance decreased significantly. What should the company do to mitigate this problem?

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

Correct Answer: C

🗨️ **scs50** 3 weeks, 4 days ago

Selected Answer: B

The company should use hyperparameters for model tuning, which involves adjusting parameters such as regularization, learning rates, and dropout rates to enhance the model's ability to generalize well to new data

Explanation:

Hyperparameter tuning is the most effective solution in this scenario because it allows the company to adjust the settings that control the learning process of the model. By fine-tuning hyperparameters, such as increasing regularization or early stopping or adjusting dropout rates, the model can avoid overfitting to the training data and better generalize to new, unseen data in production. This approach helps improve the model's performance across various data distributions.

upvoted 1 times

🗨️ **Moon** 1 month ago

Selected Answer: C

C: Increase the volume of data that is used in training.

Explanation:

The issue described is likely caused by overfitting, where the model performs well on the training dataset but fails to generalize to unseen data. Increasing the volume of training data can help mitigate overfitting by providing the model with more diverse examples, improving its ability to generalize to new data in production.

upvoted 3 times

🗨️ **may2021_r** 1 month ago

Selected Answer: C

The correct answer is C. Increasing the volume of data used in training can help improve the model's performance in production by providing it with more diverse examples to learn from.

upvoted 1 times

🗨️ **MH1980** 1 month, 2 weeks ago

Selected Answer: C

How can you prevent overfitting?

- Increase the training data size
- Early stopping the training of the model
- Data augmentation (to increase diversity in the dataset)
- Adjust hyperparameters (but you can't "add" them)

upvoted 3 times

🗨️ **Dandelion2025** 1 month, 3 weeks ago

Selected Answer: C

To prevent overfitting, increase training data, use early stopping, apply data augmentation, and fine-tune hyperparameters without adding new ones.

upvoted 1 times

🗨️ **taka5094** 2 months, 2 weeks ago

Selected Answer: C

Reducing the training data make the model prone to overfitting, and will likely further degrade the model's performance.

upvoted 1 times

🗨️ 👤 **Blair77** 2 months, 2 weeks ago

Selected Answer: C

More diverse training data helps the model learn broader patterns and generalize better to unseen data in production. This reduces the risk of overfitting to the training set.

Reduced Overfitting: The significant performance drop in production suggests overfitting to the training data. Increasing the data volume can help the model learn more robust features that are truly predictive rather than memorizing specifics of a limited dataset.. For A - Reducing the training data volume would likely exacerbate the problem rather than solve it. The model's poor performance in production suggests it's not generalizing well, which is often a result of insufficient or non-representative training data.

upvoted 1 times

🗨️ 👤 **fed6485** 2 months, 2 weeks ago

Selected Answer: A

yes Overfitting.. but if the "Volume Data" is FIXED, meaning if they are going to reuse the same data.. this time the need to REDUCE it.. so "A"

if they have MORE/EXTRA data to augment the one already available.. than C

upvoted 1 times

🗨️ 👤 **fed6485** 2 months, 2 weeks ago

i mean A. reduce the portion for training and increase the portion for testing..

if it was 80-10-10, than do 75 -15-15

upvoted 1 times

🗨️ 👤 **fed6485** 2 months, 2 weeks ago

yes Overfitting.. but if the "Volume Data" is FIXED, meaning if they are going to reuse the same data.. this time the need to REDUCE it.. so "A"

if they have MORE/EXTRA data to augment the one already available.. than C

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: C

Model is overfitting. Needs more training data

upvoted 1 times

An ecommerce company wants to build a solution to determine customer sentiments based on written customer reviews of products. Which AWS services meet these requirements? (Choose two.)

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

Correct Answer: BD

Community vote distribution

BD (100%)

85b5b55 1 day, 1 hour ago

Selected Answer: BD

Amazon Comprehend (insight of the customer reviews) and Amazon Bedrock helps for sentiment analysis.
upvoted 1 times

eesa 2 months ago

Selected Answer: BD

B. Amazon Comprehend:

Amazon Comprehend is a fully managed natural language processing (NLP) service that can analyze text and determine sentiment, entities, key phrases, and language. For customer sentiment analysis based on written reviews, Amazon Comprehend provides built-in sentiment analysis that can classify text as positive, negative, or neutral.

D. Amazon Bedrock:

Amazon Bedrock is a service that provides access to various foundation models (FMs), which can be used to build and deploy AI-driven applications. For advanced natural language processing tasks like sentiment analysis, foundation models can be fine-tuned and applied to specific use cases, such as understanding customer sentiment in reviews. This is a more customizable and advanced option compared to pre-built solutions like Amazon Comprehend.

upvoted 1 times

taka5094 2 months, 3 weeks ago

Selected Answer: BD

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to uncover insights and relationships in text. It offers sentiment analysis capabilities out-of-the-box, which can directly determine the sentiment (positive, negative, neutral, or mixed) expressed in customer reviews.

Amazon Bedrock is a fully managed service that makes foundation models accessible with simple API calls. It allows you to build generative AI applications for various use cases, including sentiment analysis. By providing customer reviews as input prompts, you can use Bedrock to generate sentiment labels or scores.

upvoted 2 times

PHD_CHENG 2 months, 3 weeks ago

Why not B,E?

upvoted 2 times

JustEugen 17 hours, 41 minutes ago

I also thought about B and E.

For B it is easy, you can analyze text with comprehend

For E you using Rekognition you can check how customer reacts to your product while unboxing and so on

When AWS Bedrock can also be the case, it simply can do the same but trained on specific data, that actually is the same, analyze text and produce output,
upvoted 1 times

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

Which solution meets these requirements MOST cost-effectively?

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

Correct Answer: D

🗨️ 👤 **85b5b55** 1 day, 1 hour ago

Selected Answer: D

Amazon Bedrock Knowledge Base
upvoted 1 times

🗨️ 👤 **Blair77** 2 months, 2 weeks ago

Selected Answer: D

Using a knowledge base allows for efficient retrieval of relevant information from the PDFs without having to include all the content in every prompt.
upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: D

Knowledgebase is the solution
upvoted 2 times

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

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

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

Correct Answer: D

Community vote distribution

D (100%)

🗨️ 👤 **Blair77** 2 months, 2 weeks ago

Selected Answer: D

Least administrative effort: Benchmark datasets are pre-existing, curated collections of data specifically designed for evaluating AI models, including LLMs. Using these requires the least administrative effort compared to the other options.

upvoted 1 times

🗨️ 👤 **jove** 2 months, 3 weeks ago

Selected Answer: D

Benchmark datasets are specifically designed to test the performance of language models on various tasks, including bias detection. They often contain diverse data that can help identify potential biases in the LLM's outputs.

upvoted 2 times

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

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

Correct Answer: C

 **eesa** 1 month, 3 weeks ago

Selected Answer: C

C. Create effective prompts that provide clear instructions and context to guide the model's generation.

Prompt engineering is a crucial technique to ensure that a pre-trained generative AI model generates content that aligns with the company's brand voice and messaging requirements. By carefully crafting prompts, you can guide the model to produce specific, relevant, and on-brand content.

upvoted 1 times

 **tgv** 2 months, 2 weeks ago

Selected Answer: C

By creating effective prompts.

upvoted 1 times

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

- A. Detect imbalances or disparities in the data.
- B. Ensure that the model runs frequently.
- C. Evaluate the model's behavior so that the company can provide transparency to stakeholders.
- D. Use the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) technique to ensure that the model is 100% accurate.
- E. Ensure that the model's inference time is within the accepted limits.

Correct Answer: AC

Community vote distribution

AC (100%)

🗨️ **dspd** 1 day, 14 hours ago

Selected Answer: AC

A. Detect imbalances or disparities in the data C. Evaluate the model's behavior so that the company can provide transparency to stakeholders

Why:

Detecting imbalances or disparities in the data is crucial because:

It helps identify potential bias in training data before it affects model decisions

It ensures fair treatment across different customer segments

It aligns with responsible AI development practices

Evaluating model behavior for transparency is important because:

It allows stakeholders to understand how decisions are made

It helps demonstrate compliance with fair lending regulations

It enables the company to justify decisions to customers and regulators

below incorrect because:

B (frequent model runs) doesn't address bias or responsible AI

D (ROUGE technique) is for text summarization evaluation, not lending decisions

E (inference time) is about performance, not fairness or responsibility

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: AC

A & C looks correct

upvoted 4 times

A company is using an Amazon Bedrock base model to summarize documents for an internal use case. The company trained a custom model to improve the summarization quality.

Which action must the company take to use the custom model through Amazon Bedrock?

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

Correct Answer: B

Community vote distribution

D (100%)

LR2023 Highly Voted 2 months, 2 weeks ago

Selected Answer: A

Initially I was going with D but after reading this article sticking with A

<https://docs.aws.amazon.com/bedrock/latest/userguide/model-customization-use.html?form=MG0AV3>

upvoted 9 times

CTao Highly Voted 2 months ago

Selected Answer: A

A To customize model you must purchase Provisioned Throughput.

upvoted 5 times

85b5b55 Most Recent 6 hours, 8 minutes ago

Selected Answer: A

Provisioned Throughput helps to improve the quality.

upvoted 1 times

Ginopress 4 days, 22 hours ago

Selected Answer: A

Accordingly to <https://docs.aws.amazon.com/bedrock/latest/userguide/model-customization-use.html?form=MG0AV3>

upvoted 1 times

kopper2019 3 weeks, 5 days ago

Selected Answer: D

A particularly insightful comment from user "may2021_r" clarifies this:

"Bottom Line:

Required to use a custom model? Give Bedrock permissions and register your model so it can retrieve your artifacts.

Optional but recommended at scale? Purchase Provisioned Throughput to guarantee a certain level of concurrency and avoid throttling."

The key distinction is:

Granting access is the fundamental requirement to use the model at all

Provisioned Throughput is about performance and scaling, not basic access

upvoted 1 times

Moon 1 month ago

Selected Answer: D

D: Grant access to the custom model in Amazon Bedrock.

Explanation:

When a company trains a custom model to improve the performance of a base model provided by Amazon Bedrock, they need to ensure the custom model is accessible through the Amazon Bedrock service. Granting access to the custom model ensures it can be integrated and used through Bedrock's APIs and workflows for inference tasks like document summarization.

upvoted 1 times

🗨️ **may2021_r** 1 month ago

Selected Answer: D

The correct answer is D. Access must be granted in Bedrock to use custom models.

upvoted 1 times

🗨️ **may2021_r** 1 month ago

Bottom Line

Required to use a custom model? Give Bedrock permissions and register your model so it can retrieve your artifacts.

Optional but recommended at scale? Purchase Provisioned Throughput to guarantee a certain level of concurrency and avoid throttling.

So if the question specifically asks which action you must take to use the custom model, the correct answer is still about granting Bedrock access—that is the non-negotiable requirement. Purchasing Provisioned Throughput is a subsequent or optional step, depending on your performance needs.

upvoted 1 times

🗨️ **AKG85** 1 month ago

Selected Answer: D

To use the custom model with Amazon Bedrock, you need to grant access to the model first.

upvoted 1 times

🗨️ **RightAnswers** 1 month ago

Selected Answer: D

When a company has trained a custom model to improve the functionality of an Amazon Bedrock base model, they need to explicitly grant access to that custom model within the Bedrock environment. This allows Bedrock to utilize the custom model's capabilities for the desired use case.

Why option A is incorrect:

While purchasing provisioned throughput can improve the performance and responsiveness of a model in SageMaker, it's not necessary to use a custom model with Bedrock. Bedrock itself handles the infrastructure and resource allocation. Access granting is the key step for integration.

upvoted 1 times

🗨️ **grzeev** 1 month, 3 weeks ago

Selected Answer: D

The correct answer is D: Grant access to the custom model in Amazon Bedrock.

Why not B (Purchase Provisioned Throughput):

1. Provisioned Throughput is about performance and capacity, not access
2. Granting access is a mandatory first step for using custom models in Bedrock
3. Without proper access permissions, the model cannot be used at all, even with Provisioned Throughput

Granting access (C) is essential because it:

- Enables model visibility in Bedrock
- Controls who can use the custom model
- Is a prerequisite for any model operations

upvoted 2 times

🗨️ **grzeev** 1 month, 3 weeks ago

Sorry:

Granting access (C) is essential because it:

- Enables model visibility in Bedrock
- Controls who can use the custom model
- Is a prerequisite for any model operations

upvoted 1 times

🗨️ **6c8c706** 1 month, 3 weeks ago

Selected Answer: A

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

upvoted 5 times

🗨️ **Contactfornitish** 2 months ago

Selected Answer: B

A. Purchase Provisioned Throughput for the custom model

Provisioned Throughput is not relevant to Amazon Bedrock or custom models. It is generally associated with services like DynamoDB for performance scaling.

C. Register the model with the Amazon SageMaker Model Registry

While the Model Registry helps manage and track model versions, registering the model alone does not make it usable for inference. The model must still be deployed to a SageMaker endpoint.

D. Grant access to the custom model in Amazon Bedrock

Amazon Bedrock only provides access to foundation models hosted and managed by AWS. Custom models trained by the company need to be deployed separately via Amazon SageMaker.

upvoted 1 times

🗨️ 👤 **leo321** 2 months, 1 week ago

A - is the right answer, as you NEED to Purchase Provisioned Throughput for customized model:

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

D - is NOT (less) correct as IAM is OPTIONAL: <https://docs.aws.amazon.com/bedrock/latest/userguide/model-customization-prereq.html>

upvoted 3 times

🗨️ 👤 **RY66** 2 months, 1 week ago

The correct answer is D. Grant access to the custom model in Amazon Bedrock.

upvoted 1 times

🗨️ 👤 **fed6485** 2 months, 2 weeks ago

Selected Answer: A

B, and C, CANNOT be as the question is clear: "..using an Amazon Bedrock.. through Amazon BedRock", in short SageMaker is out of the picture in this case.

we are talking about a customize Bedrock Model.. so .. A is the only possible answer, we are not deploying a custom model in bedrock, we are using a bedrock customised model.. and in that case you have to pay the premium... as per this link:

<https://docs.aws.amazon.com/bedrock/latest/userguide/prov-throughput.html>

...If you customized a model, you must purchase Provisioned Throughput to be able to use it

upvoted 4 times

🗨️ 👤 **Blair77** 2 months, 2 weeks ago

While this might be relevant for scaling usage, it's not the immediate step needed to use the custom model in Bedrock.

upvoted 1 times

🗨️ 👤 **Blair77** 2 months, 2 weeks ago

Selected Answer: D

The question specifically mentions using the custom model "through Amazon Bedrock," which implies that the model should be integrated with Bedrock's infrastructure.

upvoted 2 times

🗨️ 👤 **AlwaysHungry** 2 months, 2 weeks ago

Has to be B

upvoted 1 times

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

What should the company do to meet these requirements?

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

Correct Answer: B

🗨️ 👤 **Moon** 1 month ago

Selected Answer: B

B: Evaluate the models by using a human workforce and custom prompt datasets.

Explanation:

To determine which model generates responses in the style that the company's employees prefer, the company should evaluate the models using custom prompt datasets relevant to their specific use cases. Additionally, involving a human workforce ensures subjective aspects, like tone, style, and alignment with employee preferences, are effectively assessed.

upvoted 1 times

🗨️ 👤 **tgv** 2 months, 2 weeks ago

Selected Answer: B

Custom prompting is the way.

upvoted 1 times

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

Which challenge of responsible generative AI does this scenario represent?

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

Correct Answer: C

🗨️ **aws4myself** 1 month, 3 weeks ago

Selected Answer: C

Plagiarism is the act of taking someone else's work or ideas and passing them off as one's own. In this case, the student is using AI-generated content without proper attribution, which is a form of plagiarism.

upvoted 1 times

🗨️ **GriffXX** 2 months, 1 week ago

Selected Answer: C

The student is plagiarizing.

upvoted 1 times

A company needs to build its own large language model (LLM) based on only the company's private data. The company is concerned about the environmental effect of the training process.

Which Amazon EC2 instance type has the LEAST environmental effect when training LLMs?

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

Correct Answer: D

Community vote distribution

D (100%)

🗨️ **Moon** 1 month ago

Selected Answer: D

D: Amazon EC2 Trn series

Explanation:

The Amazon EC2 Trn series (Trn1 instances) are purpose-built for training machine learning models and are designed to deliver high performance while optimizing energy efficiency. They use AWS Trainium chips, which are specifically engineered for ML training workloads, providing excellent performance per watt and reducing the environmental impact of large-scale training processes.

upvoted 1 times

🗨️ **GriffXX** 2 months, 2 weeks ago

Selected Answer: D

From the documentation of the Sustainability pillar here : https://docs.aws.amazon.com/wellarchitected/latest/sustainability-pillar/sus_sus_hardware_a3.html

"For machine learning workloads, take advantage of purpose-built hardware that is specific to your workload such as AWS Trainium, AWS Inferentia, and Amazon EC2 DL1. AWS Inferentia instances such as Inf2 instances offer up to 50% better performance per watt over comparable Amazon EC2 instances."

upvoted 1 times

🗨️ **jove** 2 months, 3 weeks ago

Selected Answer: D

D. Amazon EC2 Trn series

upvoted 2 times