AWS AI Practitioner AIF-C01 Exam Questions

Total Questions: 150+
Demo Questions: 25
Version: Updated for 2025

Prepared and Verified by Cert Empire – Your Trusted IT Certification Partner

For Access to the full set of Updated Questions – Visit:
AIF-C01 Exam Dumps by Cert Empire

Question: 1

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

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

B. Create a custom service role that has Amazon S3 access. Ask teams to specify the customer name
on each Amazon Bedrock request.

C. Redact personal data in Amazon S3. Update the S3 bucket policy to allow team access to customer
data.

D. Create one Amazon Bedrock role that has full Amazon S3 access. Create IAM roles for each team
that have access to only each team's customer folders.

Answer:

A

Explanation:

This solution correctly implements the principle of least privilege using standard AWS identity and access management (IAM) controls. By creating a unique Amazon Bedrock custom service role for each team, you can attach an IAM policy to each role that grants access only to the specific Amazon S3 prefix or folder containing that team's customer data. When a team performs an operation in Bedrock that requires data access (like model fine-tuning or using a Knowledge Base), Bedrock assumes that team's specific role, thereby inheriting its restricted permissions. This enforces strict, infrastructure-level data segregation that aligns perfectly with the security policy.

References:

1. Amazon Bedrock User Guide: In the context of features that access customer data, such as Knowledge Bases, the official documentation mandates the creation of a service role. It states, "Create an IAM role that gives Amazon Bedrock permission to access your vector database and the S3 bucket containing your data sources." The example policies demonstrate scoping

permissions to specific S3 resources. This principle directly supports creating a unique, scoped-down role per team for data segregation.
Source: Amazon Bedrock User Guide, Section: "Create a service role for Knowledge Bases for Amazon Bedrock".
2. AWS IAM User Guide: This guide defines the mechanism used in the correct answer. A service role is an IAM role that a service (like Amazon Bedrock) assumes to perform actions on your behalf. The permissions for the service are determined by the IAM policies attached to that role, not the user who starts the action. This confirms that the service role's permissions are the critical control point.
Source: IAM User Guide, Section: "Role terms and concepts", Subsection: "AWS service role".
3. AWS Well-Architected Framework - Security Pillar: This framework establishes "grant least privilege" as a fundamental design principle. The correct answer is a direct implementation of this principle. "Grant only the permissions required to perform a task. You should grant permissions for specific actions on specific resources under specific conditions." Creating a role per team with access only to that team's data folder is a textbook example.
Source: AWS Well-Architected Framework - Security Pillar, Whitepaper, Section: "SEC 3: How do you manage identities for people and machines?", Principle: "Grant least privilege".

Question: 2

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

A. Temperature

B. Context window

C. Batch size

D. Model size

Answer:

B

Explanation:

The context window, also referred to as context length, is the specific parameter that defines the maximum number of tokens (a combination of the input prompt and the model's generated output) that a model can process at one time. It directly determines how much information can be fed into a single prompt. For a task like sentiment analysis on a large document, a model with a larger context window is necessary to ensure the entire text can be analyzed in one go, preserving the full context for an accurate assessment.

References:

1. Amazon Web Services (AWS) Documentation: The Amazon Bedrock User Guide explicitly lists the "Max context length" or "Max tokens" for each foundation model. This value represents the context window. For example, for Anthropic's Claude models, it states, "Maximum context length (input + output): 200,000 tokens."
Source: Amazon Bedrock User Guide, "Base models in Amazon Bedrock," section on "Model attributes."
2. University Courseware: Stanford University's course CS324: Large Language Models defines context length as a core property of a language model. Lecture slides and notes explain that the "context length" (or window) is the finite number of tokens that a model can take as input.
Source: Stanford University, CS324: Large Language Models, Winter 2023, Lecture 2: "Language Model Properties."
3. Academic Publication: The foundational paper on the Transformer architecture, which underpins most modern LLMs, establishes the concept of processing sequences of a fixed length, which is the basis for the context window.
Source: Vaswani, A., et al. (2017). "Attention Is All You Need." Advances in Neural Information Processing Systems 30 (NIPS 2017). Section 3.1, "Encoder and Decoder Stacks," describes

processing input sequences of symbol representations. (DOI: https://doi.org/10.48550/arXiv.1706.03762)

Question: 3

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

A. Confusion matrix

B. Correlation matrix

C. R2 score

D. Mean squared error (MSE)

Answer:

A

Explanation:

The problem describes a multi-class classification task: classifying images into different material types. A confusion matrix is the standard and most appropriate tool for evaluating the performance of a classification model. It provides a comprehensive summary of prediction results by showing the number of correct and incorrect predictions for each class. This matrix forms the basis for calculating other essential classification metrics such as accuracy, precision, recall, and F1-score, offering a detailed insight into how the model is performing across different categories.

References:

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. In Chapter 4, Section 4.3 "Linear Discriminant Analysis," the concept of misclassification error, which is detailed in a confusion matrix, is presented as a primary measure for classification performance. In contrast, Chapter 3 discusses metrics for regression, such as residual sum of squares (related to MSE and R2).
2. Stanford University. (2023). CS231n: Deep Learning for Computer Vision. Module 1: Neural Networks Part 2, "Setting up the data and the model" section. The course notes state, "To evaluate performance on a classification task it is common to use a confusion matrix."
3. MIT OpenCourseWare. (2020). 6.036 Introduction to Machine Learning, Fall 2020. Lecture 9: Model Selection and Evaluation. The lecture notes introduce the confusion matrix as a fundamental tool to analyze the performance of a classifier by detailing true positives, false positives, true negatives, and false negatives.

Question: 4

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

A. Data augmentation for imbalanced classes

B. Model monitoring for class distribution

C. Retrieval Augmented Generation (RAG)

D. Watermark detection for images

Answer:

A

Explanation:

The problem described is a biased image generation model resulting from imbalanced input data. Data augmentation is a pre-processing technique used to artificially increase the size and diversity of a training dataset. When specific classes or attributes are underrepresented, augmentation can be applied to create more synthetic examples for those classes. This process helps to balance the class distribution in the training data, forcing the model to learn from a more equitable representation and thereby mitigating the generation of biased or stereotypical images. It directly addresses the root cause of the problem-the biased training data.

References:

1. Academic Publication: Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1-35. In Section 4.1, "Pre-processing techniques," the authors state, "These techniques aim to mitigate bias by modifying the training data... A common approach to rebalance a dataset is oversampling the minority class... Data augmentation is another approach to generate new data points." This directly supports using augmentation to fix biased data. (DOI: https://doi.org/10.1145/3457607)
2. University Courseware: Stanford University, CS231n: Convolutional Neural Networks for Visual Recognition, Spring 2017, Lecture 5 notes. The section on "Data Augmentation" describes techniques like flipping, cropping, and color jittering to increase the size of the training set. The notes explain, "In practice, data augmentation can significantly improve the performance of ConvNets." While focused on performance, this mechanism is the foundation for creating additional samples for underrepresented classes to achieve balance.
3. Official Vendor Documentation: Amazon Web Services (AWS). (2023). Amazon SageMaker

Developer Guide. Section: "Mitigate Bias". The guide discusses pre-training mitigation strategies, stating, "You can attempt to rebalance the dataset so that the labels are more equitably distributed across the values of a facet." Data augmentation is a primary method for achieving this rebalancing through synthetic oversampling.

Question: 5

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

A. Batch transform

B. Real-time inference

C. Serverless inference

D. Asynchronous inference

Answer:

A

Explanation:

The scenario requires performing inference on large, multi-GB archived datasets where immediate predictions are not necessary. Amazon SageMaker Batch Transform is the ideal service for this use case. It is specifically designed for offline, asynchronous processing of entire datasets. SageMaker manages the provisioning of compute resources for the duration of the job, runs inference on the data, and saves the output predictions to a specified Amazon S3 location, making it efficient and cost-effective for large-scale batch processing.

References:

1. Amazon SageMaker Developer Guide. "Choose an inference option." This official documentation provides a comparison table. For Batch Transform, the "When to use it" column states, "Get inferences for an entire dataset... You don't need subsecond latency for your inferences." This directly aligns with the question's requirements.
2. Amazon SageMaker Developer Guide. "Get Inferences for an Entire Dataset with Batch Transform." This section explicitly states, "Use batch transform when you have a large dataset and don't need subsecond latency... With batch transform, you create a batch transform job. The job reads the input data from a specified S3 location and outputs the predictions to another S3 location."
3. AWS Training and Certification. "AWS Certified Machine Learning - Specialty, Module 5: Modeling." In the section discussing deployment, the courseware distinguishes between real-time inference for low-latency needs and batch transform for offline processing of large datasets, reinforcing that batch transform is the correct choice for the described scenario.

Question: 6

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.

Answer:

B

Explanation:

The core requirement is to select a model based on a subjective criterion: the "style" that employees "prefer." This cannot be measured by automated metrics like accuracy or latency. A human workforce, specifically the company's employees, is necessary to provide this qualitative feedback. Using custom prompt datasets ensures the evaluation is relevant to the company's specific internal use cases and communication style. Amazon Bedrock's model evaluation feature directly supports using a human work team to assess models on subjective metrics like style and brand alignment, making this the most appropriate method.

References:

1. Amazon Bedrock User Guide: In the section on Model evaluation, the guide states, "For subjective or custom metrics, such as friendliness, style, and brand alignment, you can set up a human evaluation work team... Your human workers can then start a task to review the model's responses and provide feedback." This directly supports using a human workforce for style-based evaluation. (Source: AWS Documentation, Amazon Bedrock User Guide, "Model evaluation," "Set up a human evaluation work team").
2. Liang, P., et al. (2022). Holistic Evaluation of Language Models. Stanford University. This foundational paper on LLM evaluation emphasizes the limitations of automated metrics. Section 2.3, "Human evaluation," details scenarios where human judgment is indispensable, stating, "For more open-ended generation tasks... human evaluation is generally required." Evaluating a preferred "style" is a prime example of such a task. (DOI: https://doi.org/10.48550/arXiv.2211.09110).

Question: 7

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

A. Using a third-party enterprise application that has embedded generative AI features.

B. Building an application by using an existing third-party generative AI foundation model (FM).

C. Refining an existing third-party generative AI foundation model (FM) by fine-tuning the model by
using data specific to the business.

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

Answer:

D

Explanation:

Building and training a generative AI model from scratch (Scope 4 in the matrix) assigns the maximum security ownership to the company. This scope requires the company to be responsible for the entire lifecycle, including the security of the training data, the integrity of the data processing and model training infrastructure, the security of the model architecture itself, and all post-training deployment and operational security. This contrasts with other scopes where a third-party vendor manages the security of the foundational model and its underlying infrastructure, creating a shared responsibility model.

References:

1. AWS Documentation, Overview of Generative AI Security. This document introduces the Generative AI Security Scoping Matrix. It explicitly defines four scopes of use. Scope 4, "Build and train your own generative AI foundation model," is described as the scenario where "you own all the security responsibilities." The accompanying diagram visually places this scope as having the highest level of customer responsibility.
2. Stanford University, CS329S: Machine Learning Systems Design, Lecture 15, MLOps: Tooling and Best Practices. This courseware discusses the end-to-end ML lifecycle. The lecture covers the extensive responsibilities involved in building a model from scratch, including data collection, infrastructure management, and deployment, which inherently include all associated security tasks, reinforcing that this path entails maximum ownership. (Reference to the principles of MLOps and full lifecycle ownership).

Question: 8

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

A. Real-time inference

B. Serverless inference

C. Asynchronous inference

D. Batch transform

Answer:

C

Explanation:

The scenario requires an inference solution that can handle large input payloads (up to 1 GB) and long processing times (up to 1 hour). Amazon SageMaker Asynchronous Inference is specifically designed for these requirements. It queues incoming requests and processes them one by one, which is ideal for workloads with large data sizes or long-running models. While not instantaneous, it provides a "near real-time" experience by accepting the request immediately and notifying the client upon completion, which is the most suitable pattern given the constraints.

References:

1. Amazon SageMaker Developer Guide, "Asynchronous inference": "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 1 GB) and/or long processing times (up to one hour)." This directly supports the choice of Asynchronous Inference for the specified requirements.
2. Amazon SageMaker Developer Guide, "Real-time inference": "For inferences with a payload up to 6 MB and a processing time of 60 seconds or less, use a SageMaker real-time endpoint." This documentation explicitly states the limitations that make real-time inference unsuitable for the scenario.
3. Amazon SageMaker Developer Guide, "Serverless Inference": Under the "Quotas" section, the documentation specifies a "Payload size up to 4 MB" and a "Timeout up to 60 seconds," confirming it cannot meet the question's requirements.
4. Amazon SageMaker Developer Guide, "Use batch transform": "Use batch transform when you need to get inferences for an entire dataset... Batch transform is ideal for scenarios where you

have large batches of data, need to preprocess and post-process data, and don't need sub-second latency." This clarifies its purpose for offline, non-interactive jobs.

Question: 9

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.

Answer:

A

Explanation:

To achieve the lowest possible inference latency, processing must occur locally on the edge device, which eliminates network round-trip time inherent in API-based solutions. Small language models (SLMs) are specifically designed and optimized for resource-constrained environments like edge devices. Their smaller size and lower computational requirements result in significantly faster processing (lower computational latency) and a smaller memory footprint compared to large language models (LLMs). Therefore, deploying an optimized SLM directly on the edge device is the most effective strategy to meet the requirement for the lowest possible latency.

References:

1. Academic Publication: Ghasemzadeh, P., et al. (2024). A Survey on Large Language Models for Edge Intelligence: A New Horizon. arXiv preprint arXiv:2402.18231. In Section III.A, "Model Compression for On-Device LLMs," the paper discusses that deploying large models on edge devices is challenging due to "high latency and energy consumption." It highlights the necessity of using smaller, compressed models for efficient on-device execution. (DOI: https://doi.org/10.48550/arXiv.2402.18231)
2. University Courseware: Stanford University, CS 329T: "On-device AI". Lecture 11, "On-device Inference," details the primary motivations for on-device AI, listing "Latency: no network roundtrip" as a key advantage. The lecture materials emphasize that models must be small and computationally efficient to run effectively on mobile and edge hardware, directly supporting the use of SLMs over LLMs for this purpose.

3. Academic Publication: Zafrir, O., et al. (2021). Prune Once for All: Sparse Pre-Trained Language Models. In Advances in Neural Information Processing Systems 34 (NeurIPS 2021). This paper and similar research on model optimization demonstrate that creating smaller, sparser models (a characteristic of SLMs) is a critical technique for enabling "fast inference on resource-constrained hardware," such as edge devices. (Available via NeurIPS proceedings).

Question: 10

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

A. Build a conversational chatbot by using Amazon Lex.

B. Transcribe call recordings by using Amazon Transcribe.

C. Extract information from call recordings by using Amazon SageMaker Model Monitor.

D. Create classification labels by using Amazon Comprehend.

Answer:

B

Explanation:

The primary requirement is to analyze and extract information from the audio of customer calls. The foundational step for any analysis of spoken language is to convert the audio (speech) into a machine-readable format, which is text. Amazon Transcribe is an automatic speech recognition (ASR) service designed specifically for this purpose. It accurately transcribes audio files, such as call recordings, into text. This transcribed text can then be used by other services for further analysis to gain the desired insights. Therefore, transcribing the call recordings is the essential first step that directly addresses the problem of processing the audio data.

References:

1. AWS Documentation for Amazon Transcribe: "Amazon Transcribe is an automatic speech recognition (ASR) service that makes it easy for you to add speech-to-text capabilities to your applications... For example, you can use Amazon Transcribe to transcribe customer service calls and analyze the audio for customer sentiment."
Source: AWS Developer Guide, "What is Amazon Transcribe?", Introduction.
2. AWS Solutions Library for Contact Center Intelligence (CCI): The official AWS architecture for this exact use case shows that audio from calls is first processed by Amazon Transcribe to convert speech to text, which is then sent to services like Amazon Comprehend for analysis.
Source: AWS Solutions Library, "Contact Center Intelligence on AWS", Architecture Diagram and description.
3. AWS Documentation for Amazon Comprehend: "Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text."
Source: AWS Developer Guide, "What is Amazon Comprehend?", Introduction.
4. AWS Documentation for Amazon Lex: "Amazon Lex is an AWS service for building conversational interfaces into any application using voice and text."

Source: AWS Developer Guide, "What Is Amazon Lex?", Introduction.

Question: 11

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

A. Amazon SageMaker Feature Store

B. Amazon SageMaker Data Wrangler

C. Amazon SageMaker Clarify

D. Amazon SageMaker Model Cards

Answer:

A

Explanation:

Amazon SageMaker Feature Store is a fully managed, purpose-built repository to store, update, retrieve, and share machine learning (ML) features. It serves as a single source of truth for features, which are the "variables" used in model development. By centralizing features, different teams can discover, share, and reuse them across various models, ensuring consistency, reducing redundant data processing efforts, and accelerating the ML development lifecycle. This directly addresses the company's requirement to share and manage variables across multiple teams.

References:

1. Amazon SageMaker Developer Guide: "Amazon SageMaker Feature Store is a fully managed repository where you can store and access features so they can be more easily discovered and reused across your organization. SageMaker Feature Store provides a unified store for features during training and real-time inference without the need to write additional code or create and maintain complex pipelines."
Source: AWS Documentation, What Is Amazon SageMaker Feature Store?, Section: "Amazon SageMaker Feature Store".
2. Amazon SageMaker Developer Guide: "Amazon SageMaker Data Wrangler reduces the time it takes to aggregate and prepare data for machine learning (ML) from weeks to minutes... You can use the transformations and analyses to create a data processing workflow..."
Source: AWS Documentation, Prepare ML Data with Amazon SageMaker Data Wrangler, Section: "Amazon SageMaker Data Wrangler".
3. Amazon SageMaker Developer Guide: "Amazon SageMaker Clarify helps improve your machine learning models by detecting potential bias and helping explain how these models make predictions."

Source: AWS Documentation, Amazon SageMaker Clarify, Section: "Fairness and Explainability with Amazon SageMaker Clarify".
4. Amazon SageMaker Developer Guide: "Amazon SageMaker Model Cards provide a single location to store critical model information. Model cards can be used to document a model's intended use cases, performance objectives, and to report on the evaluation results of a model against these objectives."
Source: AWS Documentation, Amazon SageMaker Model Cards, Section: "Create an Amazon SageMaker Model Card".

Question: 12

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.

Answer:

A

Explanation:

Prompt engineering is the most direct and effective method for controlling the output of a large language model (LLM). By adjusting the prompt, a user can provide explicit instructions to the model regarding the desired length, language, tone, and format of the response. For instance, including phrases like "Respond in Spanish" and "Keep the answer under 30 words" within the prompt directly guides the model to generate an output that aligns with these specific constraints. This technique leverages the model's instruction-following capabilities to tailor its behavior without altering its core architecture or generation randomness parameters.

References:

1. Google Cloud. (2024). Introduction to prompt design. Google Cloud AI Documentation. Retrieved from https://cloud.google.com/vertex-ai/docs/generative-ai/learn/introduction-prompt-design.
Reference Details: The section "Prompt components" explains that an "Instruction" is a key part of a prompt, telling the model what to do. The document states, "You can use prompts to guide the model to generate text that is a summary, a translation, or an answer to a question," directly supporting the use of prompts to control language and format.
2. Amazon Web Services. (2024). Prompt engineering guidelines. Amazon Bedrock User Guide. Retrieved from https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering-guidelines.html.
Reference Details: The guide explicitly states, "Be specific and detailed in your prompt... To get a specific response, provide instructions, context, or examples of the output that you want." This confirms that instructions for length and language should be placed in the prompt.
3. Stanford University. (2023). Lecture 10: Prompting, Instruction Tuning. CS224N: Natural Language Processing with Deep Learning. Retrieved from

https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture10-prompting.pdf.
Reference Details: Slide 11 ("Prompting") describes how a prompt is composed of a task description and examples. The task description is precisely where one would specify constraints like output language and length to guide the LLM's generation.
4. Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., & Sui, Z. (2022). A Survey on In-context Learning. arXiv. https://doi.org/10.48550/arXiv.2301.00234.
Reference Details: Section 2, "Preliminaries," formally defines how a prompt, consisting of a demonstration or instruction, is used to condition the LLM to produce a desired output. This foundational concept underpins the use of prompts to control model behavior.

Question: 13

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

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

B. Prepare the training dataset by creating a .txt file that contains multiple lines in .csv format.

C. Purchase Provisioned Throughput for Amazon Bedrock.

D. Train the model on journals and textbooks.

Answer:

A

Explanation:

The most effective strategy for fine-tuning a foundation model in Amazon Bedrock is to use supervised fine-tuning with a custom, labeled dataset. This process requires preparing a training dataset where each entry is a structured example consisting of an input (prompt) and the desired output (completion). By providing these prompt-completion pairs derived from the company's own data, the model learns to generate more accurate and contextually relevant responses for the specific domain of the AI search tool. This directly addresses the goal of improving accuracy using proprietary data.

References:

1. AWS Bedrock User Guide: In the section on custom models, the guide specifies the data preparation process for fine-tuning: "Create a training dataset and (optional) validation dataset in JSON Lines format. Each JSON object in the file is an example consisting of a prompt and completion field."
Source: AWS Documentation, Amazon Bedrock User Guide, "Custom models," "Prepare the training data" section.
2. AWS Bedrock User Guide on Provisioned Throughput: The documentation defines the purpose of this feature: "With Provisioned Throughput, you can purchase model units for a specific base or custom model. You can use the purchased model units to perform inference on that model at a guaranteed throughput." This confirms it is for inference, not training.
Source: AWS Documentation, Amazon Bedrock User Guide, "Provisioned Throughput" section.
3. Stanford University Courseware (CS224N): Lecture materials on adapting language models discuss instruction tuning, a form of fine-tuning. This method trains a model on examples formatted as (instruction, output) pairs, which is conceptually identical to the prompt-completion structure used by Bedrock. This academic source validates the fundamental approach described

in the correct answer.
Source: Stanford University, CS224N: NLP with Deep Learning, Winter 2023, Lecture 11: "Prompting, Instruction Tuning, and RLHF," Slide 45.

Question: 14

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

Answer:

A

Explanation:

Object detection is a computer vision technology that deals with identifying and locating objects within an image or video. Specifically, it draws bounding boxes around detected objects (identification/localization) and assigns a class label to them (categorization). This directly addresses the requirement to automatically find where the animals are in the photos and determine what kind of animals they are. The process is automated, fulfilling the "without manual human effort" constraint once the model is trained.

References:

1. Stanford University CS231n Course Notes. In the "Object Detection" section, it is defined as the task of classification and localization of objects. The notes state, "we want to classify each object and localize it using a bounding box."
Source: Stanford CS231n: Convolutional Neural Networks for Visual Recognition, Spring 2022, Lecture 11: Detection and Segmentation, Slides 4-6.
2. Official Vendor Documentation (Amazon Web Services). The Amazon Rekognition documentation describes its "Object and Scene Detection" feature as being able to "identify thousands of objects such as cars, furniture, or animals, and scenes such as a beach or city."
Source: AWS Documentation, Amazon Rekognition Developer Guide, "Detecting objects in images," Section: "Objects and scenes detected in images."
3. Peer-Reviewed Academic Publication. A comprehensive survey on object detection defines the task as determining "where objects are located in a given image (object localization) and which category each object belongs to (object classification)."
Source: Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object Detection in 20 Years: A Survey. arXiv preprint arXiv:1905.05055, Section 1, Paragraph 1. (Note: While an arXiv preprint, this survey is widely cited and foundational in the field, reflecting academic consensus).

Question: 15

A research company implemented a chatbot by using a foundation model (FM) from Amazon Bedrock. The chatbot searches for answers to questions from a large database of research papers.
After multiple prompt engineering attempts, the company notices that the FM is performing poorly because of the complex scientific terms in the research papers.
How can the company improve the performance of the chatbot?

A. Use few-shot prompting to define how the FM can answer the questions.

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

C. Change the FM inference parameters.

D. Clean the research paper data to remove complex scientific terms.

Answer:

B

Explanation:

The foundation model (FM) is underperforming because it lacks familiarity with the specialized vocabulary ("complex scientific terms") of the research domain. Domain adaptation fine-tuning is the process of further training a pre-trained model on a specific, unlabeled dataset-in this case, the research papers. This process adjusts the model's internal parameters (weights) to learn the nuances, terminology, and statistical patterns of the scientific domain. This directly addresses the root cause of the problem by enhancing the model's core comprehension of the specialized content, leading to significantly improved performance.

References:

1. AWS Documentation: The Amazon Bedrock User Guide explains that fine-tuning adapts a model for specific tasks or domains. It states, "Fine-tuning is the process of taking a pre-trained foundation model (FM) and further training it on your own dataset... to make it more specialized for your specific application."
Source: Amazon Bedrock User Guide, "Custom models," section on "Fine-tuning."
2. University Courseware: Stanford University's course on Large Language Models distinguishes between in-context learning (prompting) and fine-tuning. Fine-tuning modifies the model's weights to specialize it, which is necessary when the task requires deep domain knowledge that cannot be conveyed in a few examples.
Source: Stanford University, CS324: Large Language Models, Winter 2022, Lecture 3: "Capabilities," section on "Adaptation."
3. Academic Publication: A foundational paper on language models explains that fine-tuning is a critical step for adapting large pre-trained models to specific downstream tasks or domains, which

significantly improves performance over using the base model alone.
Source: Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186. Section 4: "Experiments." (https://doi.org/10.18653/v1/N19-1423)

Question: 16

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

A. Use Amazon Macie to scan the model's output for sensitive data and set up alerts for potential violations.

B. Configure AWS CloudTrail to monitor the model's responses and create alerts for any detected personal information.

C. Use Guardrails for Amazon Bedrock to filter content. Set up Amazon CloudWatch alarms for notification of policy violations.

D. Implement Amazon SageMaker Model Monitor to detect data drift and receive alerts when model
quality degrades.

Answer:

C

Explanation:

Guardrails for Amazon Bedrock is a purpose-built feature designed to implement safeguards in generative AI applications. It allows organizations to define policies to control conversation topics and filter content based on specific criteria, including the removal of Personally Identifiable Information (PII). This directly addresses the requirement to prevent the model from including personal patient data.
When a guardrail policy is violated, Amazon Bedrock integrates with Amazon EventBridge (formerly CloudWatch Events), which can capture these violation events. These events can then be used to trigger Amazon CloudWatch alarms or send notifications via services like Amazon Simple Notification Service (SNS), fulfilling the requirement for notification.

References:

1. Guardrails for Amazon Bedrock: The official AWS documentation states, "With Guardrails for Amazon Bedrock, you can... configure a set of policies to safeguard your generative AI applications... You can create multiple guardrails, each with a different combination of policies. The policies in a guardrail include... Content filters to filter harmful content... and Denied topics to avoid unwanted topics."
Source: AWS Documentation, "Guardrails for Amazon Bedrock," Introduction.

2. Monitoring Guardrails with CloudWatch: The documentation further explains the notification mechanism: "Amazon Bedrock integrates with Amazon CloudWatch Events to notify you of interventions by a guardrail... You can create rules in CloudWatch Events that trigger programmatic actions in response to an event."
Source: AWS Documentation, "Monitor Guardrails for Amazon Bedrock."
3. Amazon Macie Functionality: "Amazon Macie is a data security service that discovers sensitive data by using machine learning and pattern matching... Macie automatically detects a large and growing list of sensitive data types, including personally identifiable information (PII)... in your Amazon S3 buckets."
Source: AWS Documentation, "What is Amazon Macie?"
4. AWS CloudTrail Functionality: "AWS CloudTrail is an AWS service that helps you enable operational and risk auditing, governance, and compliance of your AWS account. Actions taken by a user, role, or an AWS service are recorded as events in CloudTrail."
Source: AWS Documentation, "What Is AWS CloudTrail?"

Question: 17

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

A. Fine-tune the model by using additional training data that is representative of the various age ranges that the application will support.

B. Add a role description to the prompt context that instructs the model of the age range that the response should target.

C. Use chain-of-thought reasoning to deduce the correct style and complexity for a response suitable
for that user.

D. Summarize the response text depending on the age of the user so that younger users receive shorter responses.

Answer:

B

Explanation:

This solution uses prompt engineering, a technique that guides the model's output by providing specific instructions within the prompt itself. Adding a role description (e.g., "Explain this to a 10-year-old") leverages the model's in-context learning capabilities to adopt the appropriate tone, vocabulary, and style for the specified age range. This method is highly effective and requires only a minor modification to the input text, making it the solution with the least implementation effort compared to model retraining or multi-step processing.

References:

1. Vendor Documentation: Google Cloud. (2024). Introduction to prompt design. Vertex AI Documentation. In the section "Prompt types," the "Persona prompt" is described as a way to assign a role to the model (e.g., "You are an expert in...") to tailor its response style, which aligns directly with the proposed solution.
2. Academic Publication: Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems 35. This paper defines Chain-of-Thought (CoT) as a method for solving arithmetic, commonsense, and symbolic reasoning problems (Section 2), distinguishing its purpose from stylistic control.

3. University Courseware: Jurafsky, D., & Manning, C. (2023). Lecture 10: Prompting, Instruction-Tuning, and RLHF. Stanford University, CS224N: Natural Language Processing with Deep Learning. The lecture discusses prompting as a low-effort way to guide model behavior without updating model weights, contrasting it with the higher effort of fine-tuning (instruction-tuning). Assigning a persona is a fundamental prompting technique.

Question: 18

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

Answer:

D

Explanation:

Benchmark datasets are specifically created and curated by the research community for evaluating model performance on specific tasks, including fairness and bias detection. These datasets are pre-labeled, cleaned, and structured, allowing for standardized and repeatable evaluations. Using an existing, relevant benchmark dataset (e.g., Civil Comments, Jigsaw Toxicity Datasets) eliminates the need for data collection, annotation, and structuring, which are time-consuming and resource-intensive tasks. This makes it the most efficient option with the least administrative overhead for the company to systematically assess the LLM's outputs for bias.

References:

1. Dinan, E., et al. (2020). Multi-dimensional Gender Bias Classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). This paper introduces a benchmark dataset specifically for evaluating gender bias, illustrating the role of such datasets. (See Section 3: "A New Dataset for Gender Bias Classification", pp. 2-4). DOI: https://doi.org/10.18653/v1/2020.emnlp-main.391
2. Mehrabi, N., et al. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1-35. This survey discusses evaluation methodologies, highlighting the use of benchmark datasets as a primary tool for auditing and quantifying bias in models. (See Section 4: "BIAS MITIGATION"). DOI: https://doi.org/10.1145/3457607
3. Stanford University. (2023). CS224N: Natural Language Processing with Deep Learning. Course materials frequently emphasize the use of standardized benchmark datasets (e.g., GLUE, SQuAD) for model evaluation to ensure comparability and reproducibility, a principle that extends to fairness and bias evaluation. (See Lecture on "Model Evaluation").

Question: 19

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

A. Calculate the total cost of resources used by the model.

B. Measure the model's accuracy against a predefined benchmark dataset.

C. Count the number of layers in the neural network.

D. Assess the color accuracy of images processed by the model.

Answer:

B

Explanation:

The most direct and standard strategy for evaluating the accuracy of any classification model, including a foundation model (FM) applied to image classification, is to test it against a benchmark dataset. A benchmark dataset (e.g., ImageNet, CIFAR-10) contains a large set of images with pre-verified, correct labels, known as the "ground truth." The model's predictions on this dataset are compared to the ground truth labels to calculate performance metrics, with accuracy being the most common. This process quantifies how well the model generalizes to new, unseen data, which is the primary goal of model evaluation.

References:

1. University Courseware:
Stanford University, CS231n: Convolutional Neural Networks for Visual Recognition, Module 1, "Setting up the data and the model". The course notes explain the necessity of a test set: "Finally, after the best hyperparameters are found, we evaluate the best model on the test set to get a measurement of how well the model is expected to perform on new data." This test set is a form of benchmark dataset.
2. Official Vendor Documentation:
Amazon Web Services (AWS), Amazon SageMaker Developer Guide, "Evaluate a Model". The documentation states: "After you have trained a model, you need to evaluate it to get an estimation of its quality on new data... by comparing the predictions that the model makes with the ground truth labels from a labeled test dataset." This directly supports using a benchmark dataset to measure accuracy.
3. Peer-reviewed Academic Publications:
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248-255. This paper introduced the ImageNet dataset, which became a fundamental

benchmark for evaluating the performance of image classification models. The entire premise of the work is to provide a standardized, labeled dataset for robust model evaluation. (DOI: https://doi.org/10.1109/CVPR.2009.5206848)

Question: 20

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

Answer:

A

Explanation:

The core requirement is to translate natural language text into structured SQL queries. This is a natural language processing (NLP) task that involves both understanding the user's intent and generating syntactically correct code. Generative Pre-trained Transformers (GPT) are a class of large language models (LLMs) based on the transformer architecture. They excel at understanding context and generating coherent, structured text, including programming code. A GPT-based model can be trained or fine-tuned to specifically handle "text-to-SQL" tasks, providing an intuitive interface for non-technical users to query complex databases.

References:

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., & Polosukhin, I. (2017). Attention Is All You Need. In Advances in Neural Information Processing Systems 30 (NIPS 2017). Section 1, "Introduction," describes the transformer model's suitability for transduction tasks, which includes translation between languages (like English to SQL). Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
2. Stanford University. (2023). CS224N: NLP with Deep Learning, Lecture 11: Transformers and Pretraining. This lecture material discusses how transformer-based models like GPT are pre-trained on vast text and code corpora, enabling them to perform tasks like code generation from natural language prompts.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. The abstract and introduction clearly state that ResNets were developed to address challenges in training very deep networks for image recognition. DOI: 10.1109/CVPR.2016.90

4. Stanford University. (2022). CS229: Machine Learning, Course Notes: Support Vector Machines. Section 1, "Margins, Intuition," describes SVMs as a method for finding an optimal separating hyperplane for classification tasks.
5. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. The abstract explicitly states, "This paper introduces WaveNet, a deep neural network for generating raw audio waveforms." Available from: https://arxiv.org/pdf/1609.03499.pdf

Question: 21

Which metric measures the runtime efficiency of operating AI models?

A. Customer satisfaction score (CSAT)

B. Training time for each epoch

C. Average response time

D. Number of training instances

Answer:

C

Explanation:

Runtime efficiency for an operating AI model refers to its performance during the inference phase, which is when the model is actively making predictions on new data. The "average response time," also known as latency, is a direct measure of this efficiency. It quantifies the time elapsed between receiving a request and returning a prediction. A lower average response time indicates higher runtime efficiency, which is critical for real-time applications and a positive user experience. This metric specifically evaluates the model's speed in its deployed, operational state.

References:

1. Google Cloud AI Platform Documentation: In the documentation for Model Monitoring, "latency" (response time) is listed as a primary performance metric for prediction nodes. It is defined as the "distribution of the amount of time, in seconds, that it takes for AI Platform Prediction to return a prediction."
Source: Google Cloud. "Understanding model monitoring." Vertex AI Documentation. Accessed October 2023. (Specifically, see the table of metrics under the "Drift detection" or "Performance monitoring" sections).
2. AWS SageMaker Documentation: The official documentation for monitoring SageMaker endpoints lists ModelLatency as a key invocation metric. This metric is defined as "the time elapsed, in microseconds, from when a request enters the container until the container is ready to return a response." This directly corresponds to response time.
Source: Amazon Web Services. "Monitor Amazon SageMaker with Amazon CloudWatch." Amazon SageMaker Developer Guide, section on "SageMaker Endpoint Invocation Metrics."
3. University Courseware (Stanford): In Stanford's course on Machine Learning Systems Design (CS 329S), lecture materials on "Model Serving" emphasize latency (response time) and throughput as the two main performance metrics for a deployed model. Latency is critical for user-facing applications.
Source: Chip Huyen. "CS 329S: Machine Learning Systems Design, Lecture 7: Model Serving." Stanford University, Winter 2021, Slides 11-14.

Question: 22

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

A. Helps decrease the model's complexity

B. Improves model performance over time

C. Decreases the training time requirement

D. Optimizes model inference time

Answer:

B

Explanation:

Ongoing pre-training, also known as continued pre-training or domain-adaptive pre-training, involves further training a general foundation model on a large corpus of domain-specific, unlabeled data. This process adapts the model's internal representations, vocabulary, and understanding to the nuances of the target domain (e.g., legal, medical, or financial text). By aligning the model with the specific data distribution of the target domain before task-specific fine-tuning, it achieves a better starting point, which consistently leads to improved performance and accuracy on downstream tasks within that domain.

References:

1. Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8342-8356). The paper's central thesis, summarized in the abstract and demonstrated in Section 4 ("Results"), is that "pretraining on data from the target domain (domain-adaptive pretraining) leads to performance gains." (DOI: https://doi.org/10.18653/v1/2020.acl-main.740)
2. Bommasani, R., et al. (2021). On the Opportunities and Risks of Foundation Models. Stanford University Center for Research on Foundation Models (CRFM). In Section 4.2.1 ("Adaptation"), the report discusses methods for adapting FMs, stating, "The goal of adaptation is to steer the behavior of a foundation model to better perform a desired downstream task." Continued pre-training is a key method to achieve this performance improvement. (Page 63).
3. Stanford University. (2023). CS224N: NLP with Deep Learning, Winter 2023 Lecture 12: Pretraining and Transfer Learning. The lecture notes explain that continued pretraining on a domain-specific corpus before fine-tuning helps the model learn the specific statistics and vocabulary of that domain, which improves final task performance.

Question: 23

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

Answer:

A

Explanation:

A search application that handles both text and image queries requires a model capable of understanding and processing both data types (modalities) simultaneously. A multi-modal embedding model is specifically designed for this purpose. It converts the combined text and image input into a single, dense vector representation, known as an embedding, within a shared semantic space. This embedding can then be used to efficiently find and rank relevant items in a database by comparing vector similarity, which is the core mechanism of modern semantic search applications.

References:

1. Official Vendor Documentation: Amazon Web Services (AWS) documentation for "Amazon Titan Multimodal Embeddings" explicitly states its primary use case: "By converting images and short text into numerical representations (known as embeddings), the model supports a wide variety of multimodal search, recommendation, and ranking tasks." This directly aligns with the question's scenario. (Source: AWS Documentation, Amazon Bedrock, "Amazon Titan models").
2. Peer-Reviewed Academic Publication: The foundational paper on CLIP, a model that creates a joint embedding space for images and text, describes its utility for retrieval tasks. The model learns a "multi-modal embedding space" to perform tasks like zero-shot image retrieval from text queries. (Source: Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8748-8763. Section 3.1.3).
3. University Courseware: Stanford University's course CS231n discusses models that create joint embeddings for vision and language. These models are designed to map images and text to a shared vector space, enabling tasks like retrieving images based on text descriptions (and vice-versa), which is a form of multi-modal search. (Source: Stanford University, CS231n:

Convolutional Neural Networks for Visual Recognition, Spring 2023, Lecture 11: "Vision + Language").

Question: 24

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

A. Purchase Provisioned Throughput for the custom model.

B. Deploy the custom model in an Amazon SageMaker endpoint for real-time inference.

C. Register the model with the Amazon SageMaker Model Registry.

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

Answer:

A

Explanation:

To use a custom model (a model that has been fine-tuned) for inference within the Amazon Bedrock service, it is mandatory to purchase Provisioned Throughput. This action allocates dedicated, managed inference capacity for the custom model, ensuring consistent throughput and performance. Once Provisioned Throughput is purchased, the custom model becomes available for real-time inference calls via the Amazon Bedrock API, using the specific Amazon Resource Name (ARN) of the provisioned capacity.

References:

1. Amazon Bedrock User Guide, "Provisioned Throughput": The documentation explicitly states, "To use your custom models for inference, you must purchase Provisioned Throughput for them. You can't use custom models with the On-Demand throughput mode." This confirms that purchasing Provisioned Throughput is a required action. (Source: AWS Official Documentation).
2. Amazon Bedrock User Guide, "Custom models": The section on using custom models details the workflow, which involves fine-tuning or importing a model, followed by purchasing Provisioned Throughput to make it available for inference. The guide does not mention deploying to a SageMaker endpoint as a step for using the model within Bedrock. (Source: AWS Official Documentation).

Question: 25

A company has built a solution by using generative AI. The solution uses large language models (LLMs) to translate training manuals from English into other languages. The company wants to evaluate the accuracy of the solution by examining the text generated for the manuals.
Which model evaluation strategy meets these requirements?

A. Bilingual Evaluation Understudy (BLEU)

B. Root mean squared error (RMSE)

C. Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

D. F1 score

Answer:

A

Explanation:

The scenario describes evaluating a machine translation system that translates training manuals. The Bilingual Evaluation Understudy (BLEU) score is the industry-standard metric for this exact purpose. BLEU evaluates the quality of machine-generated text by comparing it to one or more high-quality human reference translations. It measures the correspondence of n-grams (contiguous sequences of n items) between the machine's output and the reference translations, adding a brevity penalty for translations that are too short. This directly assesses the accuracy and fluency of the generated text as required by the company.

References:

1. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311-318. In Section 2, "The Bleu Metric," the paper states, "The closer a machine translation is to a professional human translation, the better it is. This is the central idea behind our work." DOI: https://doi.org/10.3115/1073083.1073135
2. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Stanford University. In Chapter 9, "Machine Translation and Encoder-Decoder Models," Section 9.5, "Evaluation of Machine Translation," the text introduces BLEU as the "dominant metric" for MT evaluation.
3. Manning, C., & Jurafsky, D. (2021). CS224N: Natural Language Processing with Deep Learning, Lecture 8: Machine Translation, Seq2seq, and Attention. Stanford University. The lecture notes state, "BLEU (Bilingual Evaluation Understudy) is a popular metric for MT Machine Translation evaluation." (Slide 10).