DUMPSARENA

AWS Certified AI Practitioner Exam(AI1-C01)

Amazon AWS AIF-C01

Version Demo

Total Demo Questions: 20

Total Premium Questions: 406

Buy Premium PDF

https://dumpsarena.co

sales@dumpsarena.co

sales@dumpsarena.co
dumpsarena.co

QUESTION NO: 1

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

Which steps in the RAG pipeline should the company implement by using offline batch processing to

meet these requirements? (Select TWO.)

A. Generation of content embeddings

B. Generation of embeddings for user queries

C. Creation of the search index

D. Retrieval of relevant content

E. Response generation for the user

ANSWER: A C

Explanation:

In a RAG setup, the parts that make sense to run “offline” are the ones tied to your documents, not the user’s live question. Since new content arrives daily, you can batch-process that new content to create document (content) embeddings ahead of time. That way, when a user asks something, you’re not doing heavy embedding work on the entire library in the moment.

After you generate those embeddings, you also want to batch-update (or rebuild) the vector/search index. Indexing is another compute-heavy step that’s perfect for scheduled jobs, and it keeps retrieval fast at query time. Together, these two offline steps help the system feel near real-time for users because the “slow prep work” is already done.

The remaining steps—embedding the user query, retrieving the most relevant chunks, and generating the final answer—must happen online in real time, because they depend on what the user just asked.

References: https://docs.aws.amazon.com/prescriptive-guidance/latest/retrieval-augmented-generation-options/rag-architecture.html and https://aws.amazon.com/what-is/retrieval-augmented-generation/

QUESTION NO: 2

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

Which AWS services will help the company assess compliance with these requirements? (Select TWO.)

A. AWS Audit Manager

B. AWS Config

C. Amazon Inspector

D. Amazon CloudWatch

E. AWS CloudTrail

ANSWER: A B

Explanation:

The best fit here is **AWS Audit Manager** and **AWS Config**. Audit Manager helps you pull together evidence and map it to common compliance frameworks (and your own controls), which is exactly what a regulated financial company needs when auditors ask “prove it.” It’s built for ongoing compliance checks and reporting, not just one-time reviews.

AWS Config is the service that continuously tracks how your AWS resources are configured over time. That makes it really useful for enforcing internal policies (like “S3 buckets must not be public” or “encryption must be enabled”) and for spotting drift during development and deployment phases.

AWS CloudTrail is also very important for auditing API activity, but since the question says “Select TWO,” Audit Manager + Config most directly cover compliance assessment plus continuous policy monitoring.

References: https://docs.aws.amazon.com/audit-manager/latest/userguide/what-is.html and https://docs.aws.amazon.com/config/latest/developerguide/WhatIsConfig.html

QUESTION NO: 3

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.

ANSWER: A C

Explanation:

The best two actions are **A** and **C**. If you want to reduce bias, you start with the data. Checking for imbalances or disparities helps you spot cases where certain groups are underrepresented or where historical patterns in the data could lead the model to treat some applicants unfairly. Once you know where the data is skewed, you can fix it (for example, by collecting better data, rebalancing, or removing problematic proxy features).

After that, you also need to evaluate how the model behaves in the real world and be able to explain it. Testing outcomes across different groups and documenting what you find is how you create transparency for business owners, risk teams, and compliance. On AWS, a common approach is using Amazon SageMaker Clarify to detect bias and support explainability, which makes it easier to communicate what the model is doing and monitor it over time.

Useful references: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-detect-data-bias.html and https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html

QUESTION NO: 4

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

A. Embeddings

B. Tokens

C. Models

D. Binaries

ANSWER: A

Explanation:

The right term is **embeddings**. An embedding is basically a list of numbers (a vector) that represents something like a word, sentence, or document in a way a model can work with. The big win is that these numbers capture meaning: items that are related in meaning tend to end up “close” to each other in that vector space, which makes tasks like semantic search, clustering, and retrieval much easier.

It’s easy to mix this up with **tokens**, but tokens are just the chunks of text (or token IDs) created during tokenization. They’re more like labels or building blocks. The embedding is the richer numeric representation the model actually uses to understand relationships between those tokens.

Models are the systems that learn from embeddings and produce outputs, and **binaries** is just a generic computing term—not the NLP concept being asked about.

References: https://aws.amazon.com/what-is/embeddings/, https://www.tensorflow.org/tutorials/text/word_embeddings, https://en.wikipedia.org/wiki/Word_embedding

QUESTION NO: 5

A company stores millions of PDF documents in an Amazon S3 bucket. The company needs to extract the text from the PDFs, generate summaries of the text, and index the summaries for fast searching.

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

A. Amazon Translate

B. Amazon Bedrock

C. Amazon Transcribe

D. Amazon Polly

E. Amazon Textract

ANSWER: B E

Explanation:

To pull text out of PDFs at scale, the best fit is **Amazon Textract**. Textract is built for document text extraction (including scanned PDFs) and can read not just plain text but also forms and tables, which is often exactly what you want when you're dealing with real-world PDFs in S3.

Once you have the extracted text, you need a service that can create summaries. **Amazon Bedrock** is the easiest way to do that on AWS because it gives you access to foundation models that can summarize large chunks of text without you managing any model infrastructure.

For the "fast searching" part, you'd typically index those summaries in something like Amazon OpenSearch Service. That said, since the question asks you to pick only TWO services, the core combo is Textract (extract) + Bedrock (summarize). The indexing piece is commonly added as the next step in a real architecture.

References: https://docs.aws.amazon.com/textract/latest/dg/what-is.html and https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html

QUESTION NO: 6

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:

The cleanest way to enforce "each team can only see its own customers" is to lock it down at the permission layer, not in the app logic. If you create a separate Amazon Bedrock service role per team and scope that role to only that team's S3 prefix (for example, `s3://bucket/teamA/*`), then Bedrock simply can't read anyone else's data—even if someone makes a bad request or the app has a bug.

Option B is risky because it depends on teams always passing the right customer name in each request. That's not a real security control—someone could just request a different customer name if the underlying role can read broadly.

Option C (redaction) is about hiding sensitive fields, not preventing cross-team access. It doesn't guarantee that Team A can't fetch Team B's objects.

Option D creates a single Bedrock role with full S3 access, which breaks least privilege and becomes a big blast-radius problem if anything goes wrong.

References: https://docs.aws.amazon.com/IAM/latest/UserGuide/best-practices.html ,
https://docs.aws.amazon.com/AmazonS3/latest/userguide/security-iam.html ,
https://docs.aws.amazon.com/bedrock/latest/userguide/security-iam.html

QUESTION NO: 7

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
Auto scaling inference endpoints: While auto-scaling can improve performance and availability, it's primarily related to operational efficiency, not direct regulatory compliance. Though high availability can be
indirectly linked to certain business continuity compliance requirements, it's not a primary compliance aspect for chatbots concerning regulated data.

B. Threat detection
Threat Detection: Deploying a chatbot application, especially one dealing with potentially sensitive customer data, necessitates robust threat detection mechanisms. Compliance frameworks often mandate security measures to identify and mitigate potential threats, such as unauthorized access, data breaches, and malicious attacks. Utilizing services like Amazon GuardDuty or AWS Security Hub, which can monitor API calls, network activity, and identify potential vulnerabilities, demonstrably supports compliance regarding threat detection. SageMaker JumpStart models themselves don't inherently provide threat detection; rather, the deployment environment must be secured.

C. Data protection
Data Protection: Data protection is a cornerstone of most regulatory frameworks. A conversational chatbot inevitably processes and potentially stores customer data. Compliance mandates implementing appropriate data protection measures, including encryption (both in transit and at rest), access control, and adherence to data residency requirements. Fine-tuning a SageMaker JumpStart model does not automatically guarantee data protection. Instead, data protection relies on how the company handles the input data to the chatbot, the data processed during conversation, and the model's outputs. Using AWS KMS (Key Management Service) for encryption, IAM (Identity and Access Management) for granular access control, and adhering to data residency requirements using appropriate AWS regions are crucial for demonstrating data protection compliance.
Why the other options are less appropriate:

D. Cost optimization
Cost optimization: Cost optimization is an important aspect of cloud management, but it's not a core requirement for demonstrating compliance with regulatory frameworks pertaining to data handling and security.

E. Loosely coupled microservices

Loosely coupled microservices: Microservices architecture offers benefits such as scalability and maintainability, but it doesn't directly address the core requirements of regulatory compliance in the context of a chatbot application. The compliance focus remains on data protection and threat detection, regardless of the application's architecture.
Supporting Links:

AWS Security: https://aws.amazon.com/security/ Amazon GuardDuty: https://aws.amazon.com/guardduty/ AWS IAM: https://aws.amazon.com/iam/
AWS KMS: https://aws.amazon.com/kms/
AWS Compliance: https://aws.amazon.com/compliance/

ANSWER: B C

Explanation:

For a regulated, customer-facing chatbot, the easiest compliance story to tell is usually around protecting data and spotting suspicious activity. That's why **data protection** and **threat detection** are the best fits here.

Data protection maps directly to common requirements in many frameworks (privacy, confidentiality, auditability). On AWS you can show encryption at rest and in transit, strong access controls, and auditable key usage—often using services like AWS KMS, IAM, and CloudTrail. Those are the kinds of controls auditors expect when customer prompts or chat logs might include sensitive info.

Threat detection is also a common requirement because regulators want proof you can monitor and respond to suspicious behavior. Services like Amazon GuardDuty can continuously analyze logs and network activity for indicators of compromise, and AWS Security Hub can centralize findings to help with reporting and evidence.

Auto scaling, cost optimization, and microservices are useful design choices, but they don't directly prove regulatory compliance the way security monitoring and data safeguards do.

References: https://aws.amazon.com/guardduty/ https://aws.amazon.com/security-hub/ https://aws.amazon.com/kms/ https://aws.amazon.com/compliance/

QUESTION NO: 8

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

A. To encrypt text data

B. To compress text files

C. To break text into smaller units for processing

D. To translate text between languages

ANSWER: C

Explanation:

In NLP, tokenization is basically the "chopping up" step. You take raw text (like a sentence or paragraph) and split it into smaller pieces called tokens—often words, subwords, or sometimes characters. Models can't work directly with free-form text, so tokenization turns it into structured units that can be converted into numbers and fed into an ML model.

This is why tokenization shows up everywhere in language tasks: sentiment analysis, chatbots, summarization, and more. It's not about securing the text (encryption), shrinking file size (compression), or directly converting between languages (translation). Translation might use tokenization as part of its pipeline, but tokenization itself just prepares the text for processing.

References: https://en.wikipedia.org/wiki/Lexical_analysis#Tokenization and https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models.html

QUESTION NO: 9

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

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

A. AWS CloudTrail
AWS CloudTrail: CloudTrail is an AWS service that enables governance, compliance, operational auditing, and risk auditing of your AWS account. By enabling CloudTrail, the company can log all API calls made to Amazon Bedrock. This provides a detailed audit trail of who made which requests, when they were made, and from where. CloudTrail is necessary for capturing the request logs themselves. [https://aws.amazon.com/cloudtrail/]

B. Amazon CloudWatch
Amazon CloudWatch: CloudWatch is primarily used for monitoring and observability. While it can collect logs, it's not designed for long-term, cost-optimized archival of audit logs like CloudTrail. CloudWatch logs can become expensive for long-term storage.

C. AWS Audit Manager
AWS Audit Manager: Audit Manager automates audit preparation by collecting evidence from your AWS environment. It leverages services like CloudTrail but is not a replacement for it. It's a tool for audit readiness, not the primary log collection service.

D. Amazon S3 Intelligent-Tiering
Amazon S3 Intelligent-Tiering: Amazon S3 Intelligent-Tiering is a storage class designed to optimize storage costs by automatically moving data to the most cost-effective access tier based on access patterns, without operational overhead. Since the logs need to be retained for 5 years, and the access frequency is likely to decrease over time, Intelligent-Tiering helps to move the logs to less expensive tiers like infrequent access or glacier, providing cost savings. S3 offers durability and meets the security requirements due to its access control policies and encryption options.
[https://aws.amazon.com/s3/storage-classes/intelligent- tiering/]
Why other options are incorrect:

E. Amazon S3 Standard

Amazon S3 Standard: While S3 Standard provides high durability and availability, it's the most expensive S3 storage class. It's suitable for frequently accessed data, which isn't the expected use case for long-term audit logs, making Intelligent-Tiering a more cost-effective choice.
Therefore, CloudTrail captures the Amazon Bedrock API request logs, and S3 Intelligent-Tiering provides a secure, cost-optimized storage solution for retaining those logs for the required 5-year period.

ANSWER: A D

Explanation:

To capture a record of every call made to the Amazon Bedrock API, the right tool is **AWS CloudTrail**. CloudTrail is built specifically for tracking AWS API activity (who called what, when, from where, and with which parameters). That’s exactly what “log all requests made to its API” usually means in an audit/compliance sense. See https://aws.amazon.com/cloudtrail/ and https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html.

For keeping those logs for 5 years at the lowest cost, you typically deliver CloudTrail logs to **Amazon S3** and use a cost-optimized storage class. **S3 Intelligent-Tiering** is a solid fit because it automatically moves objects into cheaper tiers as they become rarely accessed, without you having to guess access patterns up front. You still get S3’s strong security controls (encryption, bucket policies, IAM) for “retain securely.” References: https://aws.amazon.com/s3/storage-classes/intelligent-tiering/ and https://docs.aws.amazon.com/AmazonS3/latest/userguide/storage-class-intro.html.

QUESTION NO: 10

A company wants to use AWS services to build an AI assistant for internal company use. The AI assistant's responses must reference internal documentation. The company stores internal documentation as PDF, CSV, and image files.

Which solution will meet these requirements with the LEAST operational overhead?

A. Use Amazon SageMaker AI to fine-tune a model.

B. Use Amazon Bedrock Knowledge Bases to create a knowledge base.

C. Configure a guardrail in Amazon Bedrock Guardrails.

D. Select a pre-trained model from Amazon SageMaker JumpStart.

ANSWER: B

Explanation:

Amazon Bedrock Knowledge Bases is the simplest fit here because it’s built specifically for “answer using my company docs” use cases. You can point it at your content (often stored in Amazon S3), it handles chunking and creating embeddings, and it connects retrieval directly to a foundation model so the assistant can pull the right passages and cite that context in its responses.

This is a classic retrieval-augmented generation (RAG) setup, and it avoids a lot of work you’d otherwise have to do yourself (building a pipeline, managing vector databases, updating indexes, and so on). It also works well with mixed document types like PDFs and CSVs, and you can include text extracted from images as part of the knowledge base content.

The other choices don’t meet the “least overhead + must reference internal docs” requirement as cleanly. Fine-tuning or JumpStart still leaves you needing a retrieval layer for up-to-date internal documents, and Guardrails is about safety controls, not grounding answers in your files.

References: https://docs.aws.amazon.com/bedrock/latest/userguide/knowledge-base.html and https://docs.aws.amazon.com/bedrock/latest/userguide/agents-rag.html

QUESTION NO: 11

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

Which solution meets these requirements?

A. Batch learning

B. Continuous pre-training
Continuous pre-trainingTo keep a foundation model (FM) updated with the most recent data on a regular basis, you need a training approach that continually integrates new information. Continuous pre-training fits this requirement because it periodically (or even continuously) retrains or fine-tunes the model with the latest data, ensuring relevance and improved performance.Here's why the other options are less suitable:A. Batch learning: Trains in large, discrete batches and may introduce significant delays between training cycles, potentially causing the model to become stale.C. Static training: Trains the model once and does not update it with new data, leading to outdated predictions.D. Latent training: Not a standard industry term or recognized strategy for regularly updating foundation models.

C. Static training

D. Latent training

ANSWER: B

Explanation:

To keep a foundation model up to date, you need a training approach where the model is periodically trained again using newly collected data. That's exactly what **continuous pre-training** is about: you keep running new training cycles over time so the model's weights actually learn from the latest information, instead of getting "stale."

Batch learning can mean retraining in chunks, but it's a broader term and doesn't clearly communicate the idea of an ongoing, planned strategy to refresh a foundation model. Static training is the opposite of what you want (train once and stop), and "latent training" isn't a standard, commonly accepted strategy for maintaining foundation models in production.

If you were implementing this on AWS, you'd typically orchestrate repeated training jobs and version the resulting models so you can evaluate and roll out updated versions safely over time. For more background, see https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html and https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html.

QUESTION NO: 12

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

A. Prompted persona switches

B. Exploiting friendliness and trust

C. Ignoring the prompt template

D. Extracting the prompt template

ANSWER: D

Explanation:

The attack that most directly exposes how an LLM is configured to behave is **extracting the prompt template**. That "template" (often the hidden system/developer instructions) is basically the rulebook that tells the model what tone to use, what to refuse, what to prioritize, and how it should respond. If an attacker can get the model to reveal that hidden instruction text, they've essentially uncovered the configuration that drives the model's behavior.

The other choices are more about steering or bypassing behavior, not revealing the underlying setup. A persona switch ("pretend you are…") changes the role the model plays, but doesn't automatically show the original rules. Exploiting friendliness is just social engineering to get compliance. Ignoring the prompt template is about overriding guardrails, which may change outputs, but it still doesn't inherently disclose the hidden instructions themselves.

For more background, see OWASP's LLM Top 10 (prompt injection and related disclosure risks): https://owasp.org/www-project-top-10-for-large-language-model-applications/ and AWS guidance on guardrails and controlling model behavior: https://docs.aws.amazon.com/bedrock/latest/userguide/guardrails.html

QUESTION NO: 13

A company is developing an AI solution to help make hiring decisions. Which strategy complies with AWS guidance for responsible AI?

A. Use the AI solution to make final hiring decisions without human review.

B. Train the AI solution exclusively on data from previous successful hires.

C. Test the AI solution to ensure that it does not discriminate against any protected groups.

D. Keep the AI decision-making process confidential to maintain a competitive advantage.

ANSWER: C

Explanation:

The strategy that best matches AWS responsible AI guidance is to actively test for bias and discrimination, especially in a high-impact area like hiring. That means checking whether the model's outcomes differ unfairly across protected groups (for example, based on gender, age, or race), and then fixing issues through data improvements, model changes, or adding guardrails.

AWS specifically calls out fairness and bias detection as a key part of building trustworthy AI systems. In practice, teams often use tools like Amazon SageMaker Clarify to detect bias in training data and in model predictions, and to improve explainability so decisions can be reviewed and challenged when needed.

The other options don't align well with responsible AI: fully automated hiring without human review raises serious accountability risks; training only on "successful hires" can bake historical bias into the model; and keeping the process secret goes against transparency expectations for systems that affect people's lives.

References: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-bias-detect-data-pre-training.html and https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-bias-detect-model-post-training.html

QUESTION NO: 14

A documentary filmmaker wants to reach more viewers. The filmmaker wants to automatically add subtitles and voice-overs in multiple languages to their films.

Which combination of steps will meet these requirements? (Select TWO.)

A. Use Amazon Transcribe and Amazon Translate to generate subtitles in other languages

B. Use Amazon Textract and Amazon Translate to generate subtitles in other languages

C. Use Amazon Polly to generate voice-overs in other languages

D. Use Amazon Translate to generate voice-overs in other languages E. Use Amazon Textract to generate voice-overs in other languages

ANSWER: A C

Explanation:

To create subtitles automatically, you first need to turn the spoken audio in the film into text. That's exactly what Amazon Transcribe is for—it listens to the dialogue and produces a time-stamped transcript. Once you have the text, Amazon

Translate can convert those subtitles into multiple languages so you can reach viewers who don't speak the original language. See https://docs.aws.amazon.com/transcribe/ and https://docs.aws.amazon.com/translate/.

For voice-overs, you need text-to-speech. After you translate the script (or the transcript) into the target language, Amazon Polly can read that text out loud using natural-sounding voices in many languages. That gives you audio tracks you can line up with the video as dubbed narration. Reference: https://docs.aws.amazon.com/polly/.

The other options don't fit because Amazon Textract is meant for pulling text out of images and documents (like scanned PDFs), not for transcribing movie audio or generating speech. Also, Amazon Translate only translates text—it doesn't produce audio by itself.

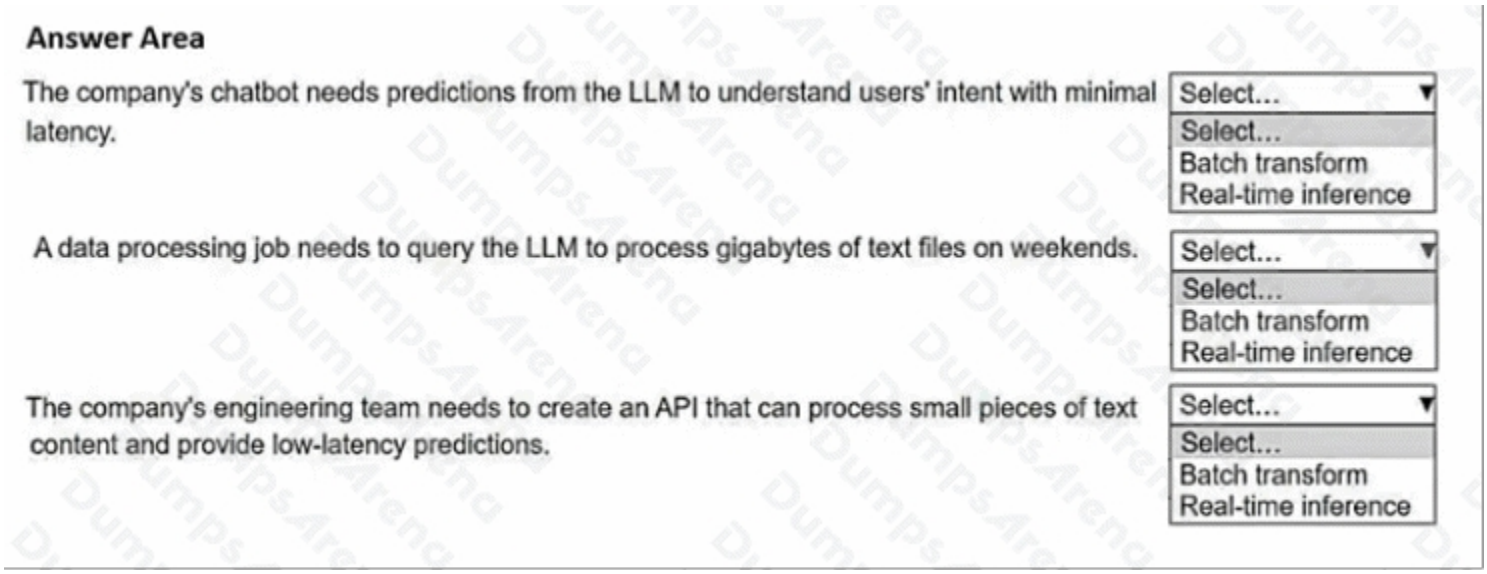
QUESTION NO: 15 - (HOTSPOT)

HOTSPOT

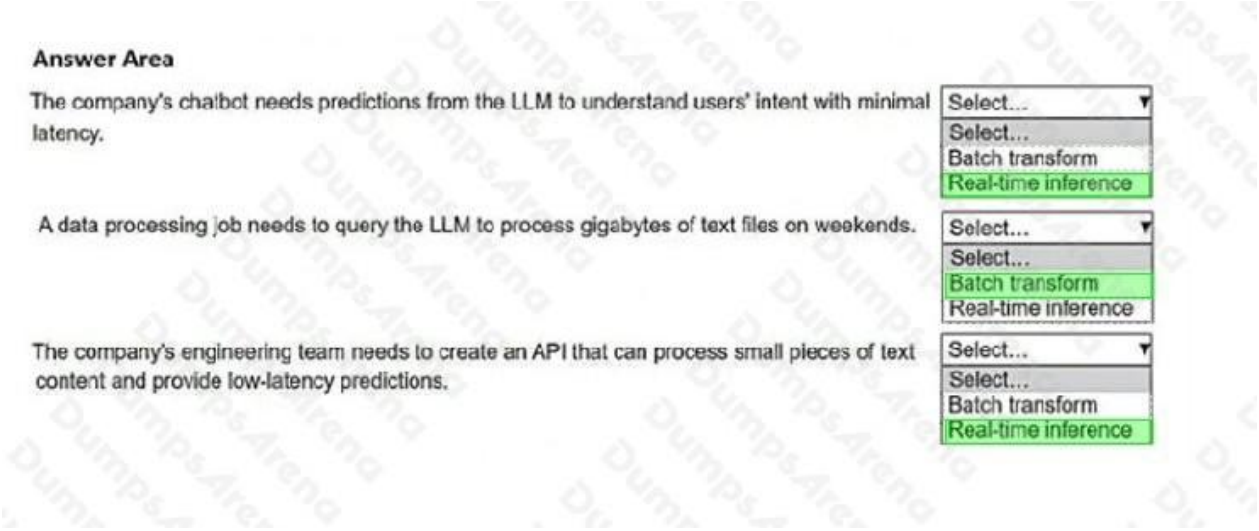
-

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

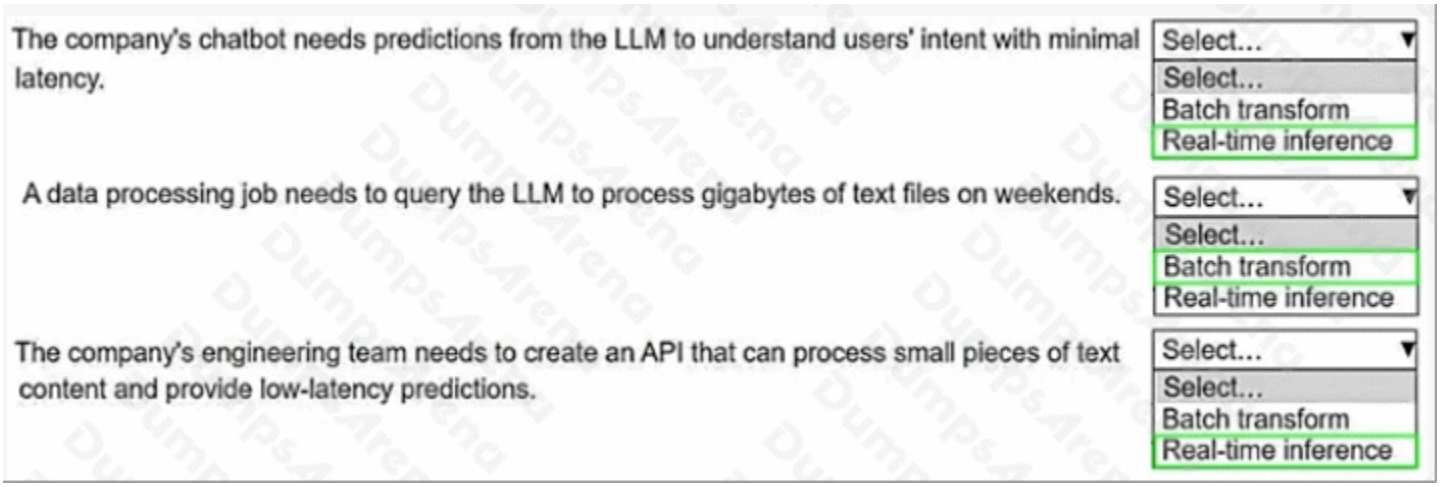
Select the correct inference mode from the following list for each use case. Each inference mode should be selected one or more times.



ANSWER:



Explanation:



1.Chatbot needing real-time predictions with minimal latency → Real-time inference .

A chatbot requires immediate responses to user queries.

Real-time inference is designed for low-latency predictions, making it the ideal choice.

This ensures that the chatbot can quickly understand and respond to user inputs without delays. 2.Processing gigabytes of text files on weekends → Batch transform.

A batch processing job does not require real-time responses.

Since this job processes large amounts of text on a schedule (weekends), batch inference is more efficient. Batch transform allows processing large datasets at once, reducing costs and optimizing performance.

3.Engineering team creating an API for small text inputs with low-latency predictions → Real-time inference. An API that processes small text inputs requires real-time predictions to maintain a smooth user experience. Real-time inference ensures quick response times for users making API requests.

QUESTION NO: 16

A company wants to improve a large language model (LLM) for content moderation within 3 months.

The company wants the model to moderate content according to the company's values and ethics. The LLM must also be able to handle emerging trends and new types of problematic content. Which solution will meet these requirements?

A. Conduct continuous pre-training on a large amount of text-based internet content.

B. Create a high-quality dataset of historical moderation decisions.

C. Fine-tune the LLM on a diverse set of general ethical guidelines from various sources.

D. Conduct reinforcement learning from human feedback (RLHF) by using real-time input from skilled moderators.

ANSWER: D

Explanation:

RLHF is the best fit here because it directly teaches the model what your company considers “good” and “bad” moderation decisions. Instead of hoping the model picks up your values from generic data, you use skilled human moderators to review outputs and provide feedback, and the model learns to prefer the responses that match your policies.

It also works well in a tight 3-month window. You don’t need to retrain on massive internet data (slow, expensive, and risky for introducing misaligned content). You can start improving behavior quickly by focusing feedback on the exact edge cases your moderators see in the real world.

Finally, RLHF is naturally adaptable. When new trends or new abuse patterns show up, moderators can label and reward/penalize those cases, and the model can be updated to handle them without waiting for a big new static dataset cycle.

References: https://docs.aws.amazon.com/sagemaker/latest/dg/reinforcement-learning.html and https://aws.amazon.com/what-is/reinforcement-learning/

QUESTION NO: 17

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.

ANSWER: A

Explanation:

With SSE-S3, Amazon S3 handles encryption and decryption automatically using S3-managed keys. So if the model (or the Bedrock feature you're using, like a knowledge base ingestion job) can't read the objects, it's almost always a permissions problem—not an encryption-key problem you need to manually manage.

The fix is to make sure the IAM role being used has the right S3 permissions (typically `s3:GetObject`, and sometimes `s3:ListBucket`), and that the bucket policy doesn't block that role. You don't need to make the bucket public, and prompt engineering can't override IAM/S3 access controls.

Helpful references: https://docs.aws.amazon.com/AmazonS3/latest/userguide/UsingServerSideEncryption.html and https://docs.aws.amazon.com/AmazonS3/latest/userguide/sse-s3.html and https://docs.aws.amazon.com/AmazonS3/latest/userguide/example-bucket-policies.html

QUESTION NO: 18

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

ANSWER: B D

Explanation:

For sentiment analysis on written product reviews, **Amazon Comprehend** is the most straightforward fit. It's built for natural language processing tasks like detecting sentiment (positive, negative, neutral, mixed) directly from text, without you having to train a model. You can run it in real time for new reviews or in batch for a big backlog, which is perfect for ecommerce review pipelines. See https://aws.amazon.com/comprehend/ and https://docs.aws.amazon.com/comprehend/latest/dg/how-sentiment.html.

Amazon Bedrock also works well here because foundation models can classify sentiment from text via prompting (and can go further, like giving aspect-based sentiment such as "delivery: negative, quality: positive"). It's a good choice when you want more flexible or richer outputs than a standard sentiment API. More details at https://aws.amazon.com/bedrock/ and https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html.

The other options don't really match: **Lex** is for chatbots, **Polly** turns text into speech, and **Rekognition** analyzes images/video—not written reviews.

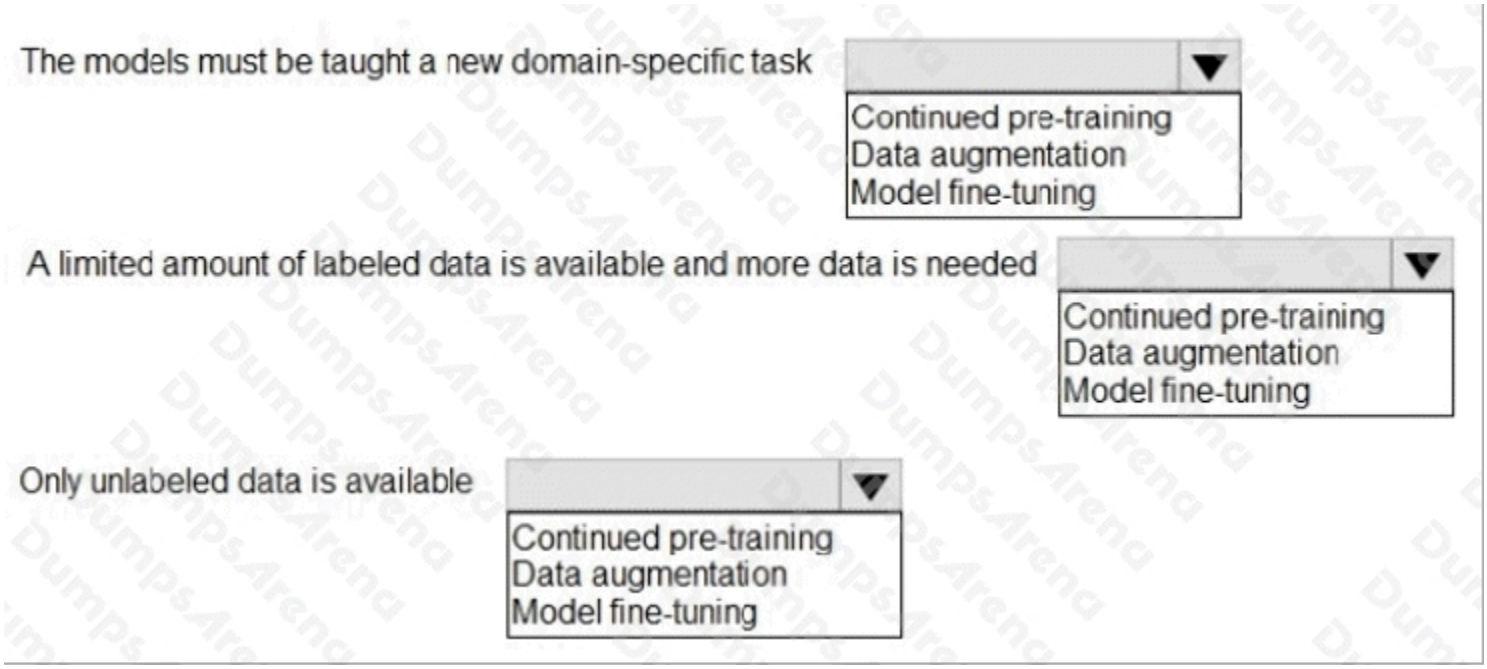
QUESTION NO: 19 - (HOTSPOT)

HOTSPOT

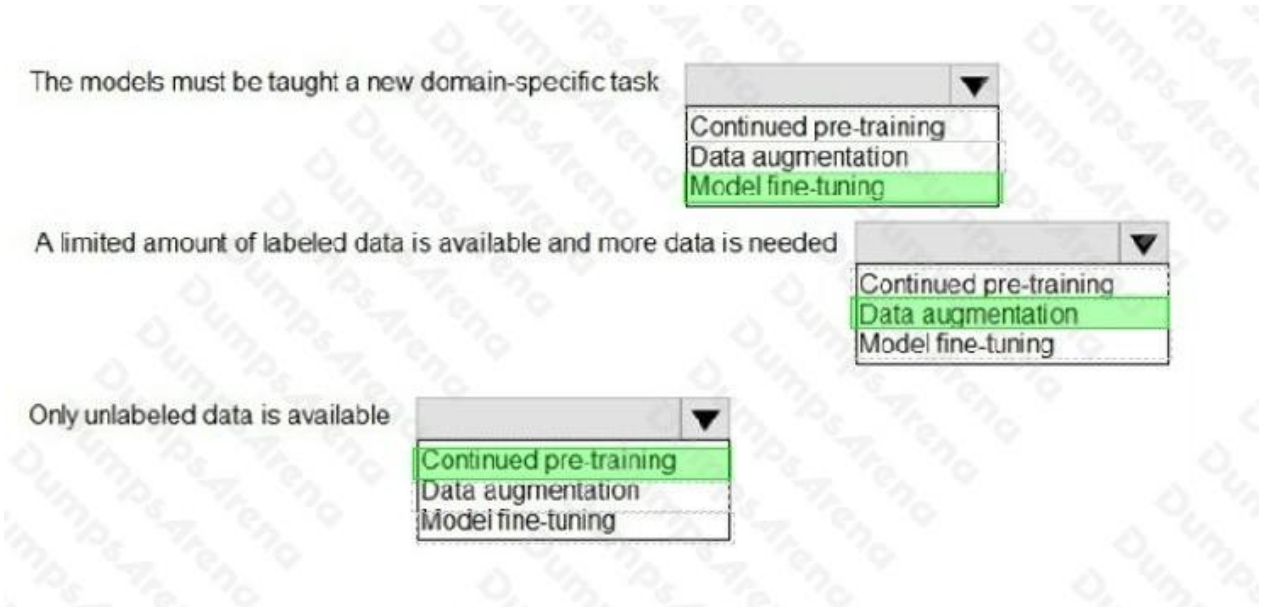
-

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

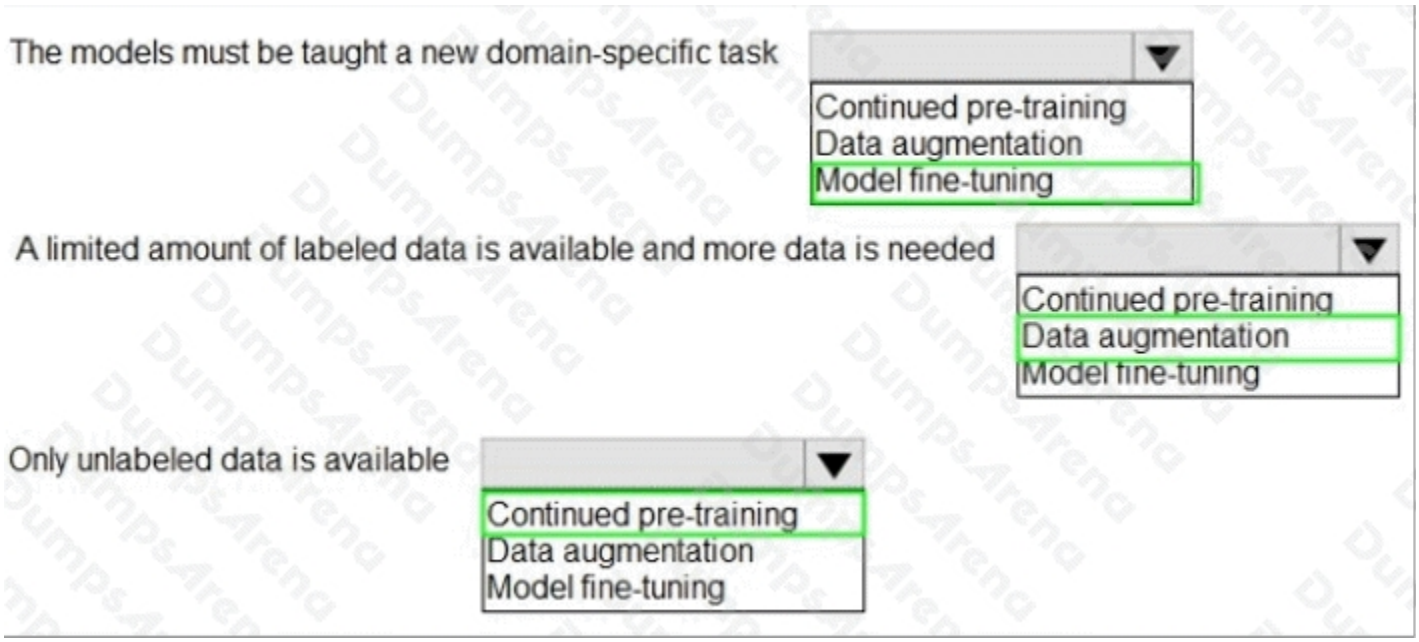
Select the correct customization methodology from the following list for each use case. Each use case should be selected one time.



ANSWER:



Explanation:



1. The models must be taught a new domain-specific task

Model fine-tuning.

When a model needs to adapt to a new task, especially within a specific domain (like legal, medical, or finance), you typically:

Start with a pre-trained model.

Use labeled data from the new domain to fine-tune it.

This adapts the model's behavior to the specific objectives or outputs of the domain-specific task. Example: Fine-tuning a language model trained on general English to classify legal documents.

2. A limited amount of labeled data is available and more data is needed.

Data augmentation.

Data augmentation is used to:

Artificially expand your dataset by generating modified versions of existing samples. Improve model performance and generalization without requiring new labeled data. Common in NLP, CV, and other domains — for instance:

Synonym replacement in NLP.

Image rotation or flipping in computer vision.

It's especially useful when labeled data is scarce but you want to improve model robustness.

3. Only unlabeled data is available.

Continued pre-training.

When you have unlabeled data, you cannot fine-tune or augment it directly. Instead, you use the data to:

Continue pre-training a foundational model using self-supervised learning techniques (like masked language modeling in NLP).

This helps the model adapt better to the target domain, even without labels.

Example: Using a corpus of technical manuals to further pre-train a language model before any fine-tuning.

QUESTION NO: 20

A company wants to use Amazon Q Business for its data. The company needs to ensure the security and privacy of the data. Which combination of steps will meet these requirements? (Select TWO.)

A. Enable AWS Key Management Service (AWS KMS) keys for the Amazon Q Business Enterprise index.

B. Set up cross-account access to the Amazon Q index.

C. Configure Amazon Inspector for authentication.

D. Allow public access to the Amazon Q index.

E. Configure AWS Identity and Access Management (IAM) for authentication.

ANSWER: A E

Explanation:

To keep Amazon Q Business data secure and private, you generally want two things: strong access control and strong encryption. Configuring IAM (and typically IAM Identity Center behind the scenes) makes sure only the right users and roles can sign in and query the Q Business app, following least-privilege access.

On top of that, enabling AWS KMS keys for the Amazon Q Business index helps protect data at rest. Using customer-managed KMS keys is a common requirement for compliance and gives you more control over key policies, rotation, and auditing.

The other options don’t really match the goal. Cross-account access is only needed if you’re intentionally sharing resources across AWS accounts, not as a core security requirement. Amazon Inspector is for finding vulnerabilities on compute workloads, not for authenticating users to Q Business. And public access would obviously work against privacy.

References: https://docs.aws.amazon.com/amazonq/latest/qbusiness-ug/security.html and https://docs.aws.amazon.com/kms/latest/developerguide/overview.html