# Exam Questions AIP-C01

AWS Certified Generative AI Developer - Professional

**https://www.2passeasy.com/dumps/AIP-C01/**

**NEW QUESTION 1**
A company runs a Retrieval Augmented Generation (RAG) application that uses Amazon Bedrock Knowledge Bases to perform regulatory compliance queries. The application uses the RetrieveAndGenerateStream API. The application retrieves relevant documents from a knowledge base that contains more than 50,000 regulatory documents, legal precedents, and policy updates.
The RAG application is producing suboptimal responses because the initial retrieval often returns semantically similar but contextually irrelevant documents. The poor responses are causing model hallucinations and incorrect regulatory guidance. The company needs to improve the performance of the RAG application so it returns more relevant documents.
Which solution will meet this requirement with the LEAST operational overhead?

A. Deploy an Amazon SageMaker endpoint to run a fine-tuned ranking mode
B. Use an Amazon API Gateway REST API to route request
C. Configure the application to make requests through the REST API to rerank the results.
D. Use Amazon Comprehend to classify documents and apply relevance score
E. Integrate the RAG application??s reranking process with Amazon Textract to run document analysi
F. Use Amazon Neptune to perform graph-based relevance calculations.
G. Implement a retrieval pipeline that uses the Amazon Bedrock Knowledge Bases Retrieve API to perform initial document retrieva
H. Call the Amazon Bedrock Rerank API to rerank the result
I. Invoke the InvokeModelWithResponseStream operation to generate responses.
J. Use the latest Amazon reranker model through the reranking configuration within Amazon Bedrock Knowledge Base
K. Use the model to improve document relevance scoring and to reorder results based on contextual assessments.

**Answer:** D

**NEW QUESTION 2**
A healthcare company is using Amazon Bedrock to develop a real-time patient care AI assistant to respond to queries for separate departments that handle clinical inquiries, insurance verification, appointment scheduling, and insurance claims. The company wants to use a multi-agent architecture.
The company must ensure that the AI assistant is scalable and can onboard new features for patients. The AI assistant must be able to handle thousands of parallel patient interactions. The company must ensure that patients receive appropriate domain-specific responses to queries.
Which solution will meet these requirements?

A. Isolate data for each agent by using separate knowledge base
B. Use IAM filtering to control access to each knowledge bas
C. Deploy a supervisor agent to perform natural language intent classification on patient inquirie
D. Configure the supervisor agent to route queries to specialized collaborator agents to respond to department-specific querie
E. Configure each specialized collaborator agent to use Retrieval Augmented Generation (RAG) with the agent??s department-specific knowledge base.
F. Create a separate supervisor agent for each departmen
G. Configure individual collaborator agents to perform natural language intent classification for each specialty domain within each departmen
H. Integrate each collaborator agent with department-specific knowledge bases onl
I. Implement manual handoff processes between the supervisor agents.
J. Isolate data for each department in separate knowledge base
K. Use IAM filtering to control access to each knowledge bas
L. Deploy a single general-purpose agen
M. Configure multiple action groups within the general-purpose agent to perform specific department function
N. Implement rule-based routing logic within the general-purpose agent instructions.
O. Implement multiple independent supervisor agents that run in parallel to respond to patient inquiries for each departmen
P. Configure multiple collaborator agents for each supervisor agen
Q. Integrate all agents with the same knowledge bas
R. Use external routing logic to merge responses from multiple supervisor agents.

**Answer:** A

**NEW QUESTION 3**
A company deploys multiple Amazon Bedrock–based generative AI (GenAI) applications across multiple business units for customer service, content generation, and document analysis. Some applications show unpredictable token consumption patterns. The company requires a comprehensive observability solution that provides real-time visibility into token usage patterns across multiple models. The observability solution must support custom dashboards for multiple stakeholder groups and provide alerting capabilities for token consumption across all the foundation models that the company??s applications use.
Which combination of solutions will meet these requirements with the LEAST operational overhead? (Select TWO.)

A. Use Amazon CloudWatch metrics as data sources to create custom Amazon QuickSight dashboards that show token usage trends and usage patterns across FMs.
B. Use CloudWatch Logs Insights to analyze Amazon Bedrock invocation logs for token consumption patterns and usage attribution by applicatio
C. Create custom queries to identify high-usage scenario
D. Add log widgets to dashboards to enable continuous monitoring.
E. Create custom Amazon CloudWatch dashboards that combine native Amazon Bedrock token and invocation CloudWatch metric
F. Set up CloudWatch alarms to monitor token usage thresholds.
G. Create dashboards that show token usage trends and patterns across the company??s FMs by using an Amazon Bedrock zero-ETL integration with Amazon Managed Grafana.
H. Implement Amazon EventBridge rules to capture Amazon Bedrock model invocation event
I. Route token usage data to Amazon OpenSearch Serverless by using Amazon Data Firehos
J. Use OpenSearch dashboards to analyze usage patterns.

**Answer:** CD

**NEW QUESTION 4**
A company is building a legal research AI assistant that uses Amazon Bedrock with an Anthropic Claude foundation model (FM). The AI assistant must retrieve highly relevant case law documents to augment the FM??s responses. The AI assistant must identify semantic relationships between legal concepts, specific legal terminology, and citations. The AI assistant must perform quickly and return precise results.
Which solution will meet these requirements?

A. Configure an Amazon Bedrock knowledge base to use a default vector search configuratio
B. Use Amazon Bedrock to expand queries to improve retrieval for legal documents based on specific terminology and citations.
C. Use Amazon OpenSearch Service to deploy a hybrid search architecture that combines vector search with keyword searc
D. Apply an Amazon Bedrock reranker model to optimize result relevance.
E. Enable the Amazon Kendra query suggestion feature for end user
F. Use Amazon Bedrock to perform post-processing of search results to identify semantic similarity in the documents and to produce precise results.
G. Use Amazon OpenSearch Service with vector search and Amazon Bedrock Titan Embeddings to index and search legal document
H. Use custom AWS Lambda functions to merge results with keyword-based filters that are stored in an Amazon RDS database.

**Answer:** B

**NEW QUESTION 5**
A company uses an AI assistant application to summarize the company??s website content and provide information to customers. The company plans to use Amazon Bedrock to give the application access to a foundation model (FM).
The company needs to deploy the AI assistant application to a development environment and a production environment. The solution must integrate the environments with the FM. The company wants to test the effectiveness of various FMs in each environment. The solution must provide product owners with the ability to easily switch between FMs for testing purposes in each environment.
Which solution will meet these requirements?

A. Create one AWS CDK applicatio
B. Create multiple pipelines in AWS CodePipelin
C. Configure each pipeline to have its own settings for each F
D. Configure the application to invoke the Amazon Bedrock FMs by using the aws_bedrock.ProvisionedModel.fromProvisionedModelArn() method.
E. Create a separate AWS CDK application for each environmen
F. Configure the applications to invoke the Amazon Bedrock FMs by using the aws_bedrock.FoundationModel.fromFoundationModelId() metho
G. Create a separate pipeline in AWS CodePipeline for each environment.
H. Create one AWS CDK applicatio
I. Configure the application to invoke the Amazon Bedrock FMs by using the aws_bedrock.FoundationModel.fromFoundationModelId() metho
J. Create a pipeline in AWS CodePipeline that has a deployment stage for each environment that uses AWS CodeBuild deploy actions.
K. Create one AWS CDK application for the production environmen
L. Configure the application to invoke the Amazon Bedrock FMs by using the aws_bedrock.ProvisionedModel.fromProvisionedModelArn() metho
M. Create a pipeline in AWS CodePipelin
N. Configure the pipeline to deploy to the production environment by using an AWS CodeBuild deploy actio
O. For the development environment, manually recreate the resources by referring to the production application code.

**Answer:** C

**NEW QUESTION 6**
A GenAI developer is building a Retrieval Augmented Generation (RAG)-based customer support application that uses Amazon Bedrock foundation models (FMs). The application needs to process 50 GB of historical customer conversations that are stored in an Amazon S3 bucket as JSON files. The application must use the processed data as its retrieval corpus. The application??s data processing workflow must extract relevant data from customer support documents, remove customer personally identifiable information (PII), and generate embeddings for vector storage. The processing workflow must be cost- effective and must finish within 4 hours.
Which solution will meet these requirements with the LEAST operational overhead?

A. Use AWS Lambda and Amazon Comprehend to process files in parallel, remove PII, and call Amazon Bedrock APIs to generate vector
B. Configure Lambda concurrency limits and memory settings to optimize throughput.
C. Create an AWS Glue ETL job to run PII detection scripts on the dat
D. Use Amazon SageMaker Processing to run the HuggingFaceProcessor to generate embeddings by using a pre-trained mode
E. Store the embeddings in Amazon OpenSearch Service.
F. Deploy an Amazon EMR cluster that runs Apache Spark with user-defined functions (UDFs) that call Amazon Comprehend to detect PI
G. Use Amazon Bedrock APIs to generate vector
H. Store outputs in Amazon Aurora PostgreSQL with the pgvector extension.
I. Implement a data processing pipeline that uses AWS Step Functions to orchestrate a workload that uses Amazon Comprehend to detect PII and Amazon Bedrock to generate embedding
J. Directly integrate the workflow with Amazon OpenSearch Serverless to store vectors and provide similarity search capabilities.

**Answer:** D

**NEW QUESTION 7**
A company needs a system to automatically generate study materials from multiple content sources. The content sources include document files (PDF files, PowerPoint presentations, and Word documents) and multimedia files (recorded videos). The system must process more than 10,000 content sources daily with peak loads of 500 concurrent uploads. The system must also extract key concepts from document files and multimedia files and create contextually accurate summaries. The generated study materials must support real-time collaboration with version control.
Which solution will meet these requirements?

A. Use Amazon Bedrock Data Automation (BDA) with AWS Lambda functions to orchestrate document file processin
B. Use Amazon Bedrock Knowledge Bases to process all multimedi
C. Store the content in Amazon DocumentDB with replicatio
D. Collaborate by using Amazon SNS topic subscription
E. Track changes by using Amazon Bedrock Agents.
F. Use Amazon Bedrock Data Automation (BDA) with foundation models (FMs) to process document file
G. Integrate BDA with Amazon Textract for PDF extraction and with Amazon Transcribe for multimedia file
H. Store the processed content in Amazon S3 with versioning enable
I. Store the metadata in Amazon DynamoD
J. Collaborate in real time by using AWS AppSync GraphQL subscriptions and DynamoDB.
K. Use Amazon Bedrock Data Automation (BDA) with Amazon SageMaker AI endpoints to host content extraction and summarization model
L. Use Amazon Bedrock Guardrails to extract content from all file type
M. Store document files in Amazon Neptune for time series analysi
N. Collaborate by using Amazon Bedrock Chat for real-time messaging.

O. Use Amazon Bedrock Data Automation (BDA) with AWS Lambda functions to process batches of content file
P. Fine-tune foundation models (FMs) in Amazon Bedrock to classify documents across all content type
Q. Store the processed data in Amazon ElastiCache (Redis OSS) by using Cluster Mode with shardin
R. Use Prompt management in Amazon Bedrock for version control.

**Answer:** B

**NEW QUESTION 8**
A company is designing an API for a generative AI (GenAI) application that uses a foundation model (FM) that is hosted on a managed model service. The API must stream responses to reduce latency, enforce token limits to manage compute resource usage, and implement retry logic to handle model timeouts and partial responses.
Which solution will meet these requirements with the LEAST operational overhead?

A. Integrate an Amazon API Gateway HTTP API with an AWS Lambda function to invoke Amazon Bedroc
B. Use Lambda response streaming to stream response
C. Enforce token limits within the Lambda functio
D. Implement retry logic for model timeouts by using Lambda and API Gateway timeout configurations.
E. Connect an Amazon API Gateway HTTP API directly to Amazon Bedroc
F. Simulate streaming by using client-side pollin
G. Enforce token limits on the fronten
H. Configure retry behavior by using API Gateway integration settings.
I. Connect an Amazon API Gateway WebSocket API to an Amazon ECS service that hosts a containerized inference serve
J. Stream responses by using the WebSocket protoco
K. Enforce token limits within Amazon EC
L. Handle model timeouts by using ECS task lifecycle hooks and restart policies.
M. Integrate an Amazon API Gateway REST API with an AWS Lambda function that invokes Amazon Bedroc
N. Use Lambda response streaming to stream response
O. Enforce token limits within the Lambda functio
P. Implement retry logic by using Lambda and API Gateway timeout configurations.

**Answer:** A

**NEW QUESTION 9**
A healthcare company is developing a document management system that stores medical research papers in an Amazon S3 bucket. The company needs a comprehensive metadata framework to improve search precision for a GenAI application. The metadata must include document timestamps, author information, and research domain classifications.
The solution must maintain a consistent metadata structure across all uploaded documents and allow foundation models (FMs) to understand document context without accessing full content.
Which solution will meet these requirements?

A. Store document timestamps in Amazon S3 system metadat
B. Use S3 object tags for domain classificatio
C. Implement custom user-defined metadata to store author information.
D. Set up S3 Object Lock with legal holds to track document timestamp
E. Use S3 object tags for author informatio
F. Implement S3 access points for domain classification.
G. Use S3 Inventory reports to track timestamp
H. Create S3 access points for domain classificatio
I. Store author information in S3 Storage Lens dashboards.
J. Use custom user-defined metadata to store author informatio
K. Use S3 Object Lock retention periods for timestamp
L. Use S3 Event Notifications for domain classification.

**Answer:** A

**NEW QUESTION 10**
A publishing company is developing a chat assistant that uses a containerized large language model (LLM) that runs on Amazon SageMaker AI. The architecture consists of an Amazon API Gateway REST API that routes user requests to an AWS Lambda function. The Lambda function invokes a SageMaker AI real-time endpoint that hosts the LLM.
Users report uneven response times. Analytics show that a high number of chats are abandoned after 2 seconds of waiting for the first token. The company wants a solution to ensure that p95 latency is under 800 ms for interactive requests to the chat assistant.
Which combination of solutions will meet this requirement? (Select TWO.)

A. Enable model preload upon container startu
B. Implement dynamic batching to process multiple user requests together in a single inference pass.
C. Select a larger GPU instance type for the SageMaker AI endpoin
D. Set the minimum number of instances to 0. Continue to perform per-request processin
E. Lazily load model weights on the first request.
F. Switch to a multi-model endpoin
G. Use lazy loading without request batching.
H. Set the minimum number of instances to greater than 0. Enable response streaming.
I. Switch to Amazon SageMaker Asynchronous Inference for all request
J. Store requests in an Amazon S3 bucke
K. Set the minimum number of instances to 0.

**Answer:** AD

**NEW QUESTION 10**
A medical company is creating a generative AI (GenAI) system by using Amazon Bedrock. The system processes data from various sources and must maintain

end-to-end data lineage. The system must also use real-time personally identifiable information (PII) filtering and audit trails to automatically report compliance. Which solution will meet these requirements?

A. Use AWS Glue Data Catalog to register all data sources and track lineag
B. Use Amazon Bedrock Guardrails PII filter
C. Enable AWS CloudTrail logging for all Amazon Bedrock API calls with Amazon S3 integratio
D. Use Amazon Macie to scan stored data for sensitive information and publish findings to Amazon CloudWatch Log
E. Create CloudWatch dashboards to visualize the findings and generate automated compliance reports.
F. Use AWS Config to track data source configurations and change
G. Use AWS WAF with custom rules to filter PII at the application layer before Amazon Bedrock processes the dat
H. Configure Amazon EventBridge to capture and route audit events to Amazon S3. Use Amazon Comprehend Medical with scheduled AWS Lambda functions to analyze stored outputs for compliance violations.
I. Use AWS DataSync to replicate data sources to track lineag
J. Configure Amazon Macie to scan Amazon Bedrock outputs for sensitive informatio
K. Use AWS Systems Manager Session Manager to log user interaction
L. Deploy Amazon Textract with AWS Step Functions workflows to identify and redact PII from generated reports.
M. Configure Amazon Athena to query data sources to analyze and report on data lineag
N. Use Amazon CloudWatch custom metrics to monitor PII exposure in Amazon Bedrock responses and establish AWS X-Ray tracing to generate an audit trai
O. Use an Amazon Rekognition Custom Labels model to detect sensitive information in the data that Amazon Bedrock processes.

**Answer:** A

**NEW QUESTION 14**
A company is building a generative AI (GenAI) application that produces content based on a variety of internal and external data sources. The company wants to ensure that the generated output is fully traceable. The application must support data source registration and enable metadata tagging to attribute content to its original source. The application must also maintain audit logs of data access and usage throughout the pipeline.
Which solution will meet these requirements?

A. Use AWS Lake Formation to catalog data sources and control acces
B. Apply metadata tags directly in Amazon S3. Use AWS CloudTrail to monitor API activity.
C. Use AWS Glue Data Catalog to register and tag data source
D. Use Amazon CloudWatch Logs to monitor access patterns and application behavior.
E. Store data in Amazon S3 and use object tagging for attributio
F. Use AWS Glue Data Catalog to manage schema informatio
G. Use AWS CloudTrail to log access to S3 buckets.
H. Use AWS Glue Data Catalog to register all data source
I. Apply metadata tags to attribute data source
J. Use AWS CloudTrail to log access and activity across services.

**Answer:** D

**NEW QUESTION 18**
A company has set up Amazon Q Developer Pro licenses for all developers at the company. The company maintains a list of approved resources that developers must use when developing applications. The approved resources include internal libraries, proprietary algorithmic techniques, and sample code with approved styling.
A new team of developers is using Amazon Q Developer to develop a new Java-based application. The company must ensure that the new developer team uses the company??s approved resources. The company does not want to make project-level modifications.
Which solution will meet these requirements?

A. Create a Git repository that contains all of the approved internal libraries, algorithms, and code sample
B. Include this Git repository in the application project locally as part of the workspac
C. Ensure that the developers use the workspace context to retrieve suggestions from the Git repository.
D. In the project root folder, create a folder named amazonq/rule
E. Add the approved internal libraries, algorithms, and code samples to the folder.
F. Create a folder in the application project named rule
G. Store the guidelines and code in the folder for Amazon Q Developer to reference for code suggestions.
H. Create an Amazon Q Developer customization that includes the approved data source
I. Ensure that the developers use the customization to develop the application.

**Answer:** D

**NEW QUESTION 23**
A healthcare company uses Amazon Bedrock to deploy an application that generates summaries of clinical documents. The application experiences inconsistent response quality with occasional factual hallucinations. Monthly costs exceed the company??s projections by 40%. A GenAI developer must implement a near real-time monitoring solution to detect hallucinations, identify abnormal token consumption, and provide early warnings of cost anomalies. The solution must require minimal custom development work and maintenance overhead.
Which solution will meet these requirements?

A. Configure Amazon CloudWatch alarms to monitor InputTokenCount and OutputTokenCount metrics to detect anomalie
B. Store model invocation logs in an Amazon S3 bucke
C. Use AWS Glue and Amazon Athena to identify potential hallucinations.
D. Run Amazon Bedrock evaluation jobs that use LLM-based judgments to detect hallucination
E. Configure Amazon CloudWatch to track token usag
F. Create an AWS Lambda function to process CloudWatch metric
G. Configure the Lambda function to send usage pattern notifications.
H. Configure Amazon Bedrock to store model invocation logs in an Amazon S3 bucke
I. Enable text output loggin
J. Configure Amazon Bedrock guardrails to run contextual grounding checks to detect hallucination
K. Create Amazon CloudWatch anomaly detection alarms for token usage metrics.
L. Use AWS CloudTrail to log all Amazon Bedrock API call

M. Create a custom dashboard in Amazon QuickSight to visualize token usage pattern
N. Use Amazon SageMaker Model Monitor to detect quality drift in generated summaries.

**Answer:** C

**NEW QUESTION 24**
Company configures a landing zone in AWS Control Tower. The company handles sensitive data that must remain within the European Union. The company must use only the eu-central-1 Region. The company uses Service Control Policies (SCPs) to enforce data residency policies. GenAI developers at the company are assigned IAM roles that have full permissions for Amazon Bedrock.
The company must ensure that GenAI developers can use the Amazon Nova Pro model through Amazon Bedrock only by using cross-Region inference (CRI) and only in eu- central-1. The company enables model access for the GenAI developer IAM roles in Amazon Bedrock. However, when a GenAI developer attempts to invoke the model through the Amazon Bedrock Chat/Text playground, the GenAI developer receives the following error:
User arn:aws:sts:123456789012:assumed-role/AssumedDevRole/DevUserName Action: bedrock:InvokeModelWithResponseStream
On resource(s): arn:aws:bedrock:eu-west-3::foundation-model/amazon.nova-pro-v1:0 Context: a service control policy explicitly denies the action
The company needs a solution to resolve the error. The solution must retain the company's existing governance controls and must provide precise access control. The solution must comply with the company's existing data residency policies.
Which combination of solutions will meet these requirements? (Select TWO.)

A. Add an AdministratorAccess policy to the GenAI developer IAM role
B. Extend the existing SCPs to enable CRI for the eu.amazon.nova-pro-v1:0 inference profile
C. Enable Amazon Bedrock model access for Amazon Nova Pro in the eu-west-3 Region
D. Validate that the GenAI developer IAM roles have permissions to invoke Amazon Nova Pro through the eu.amazon.nova-pro-v1:0 inference profile on all European Union AWS Regions that can serve the model
E. Extend the existing SCP to enable CRI for the eu-* inference profile

**Answer:** BE

**NEW QUESTION 25**
A company has a recommendation system. The system's applications run on Amazon EC2 instances. The applications make API calls to Amazon Bedrock foundation models (FMs) to analyze customer behavior and generate personalized product recommendations.
The system is experiencing intermittent issues. Some recommendations do not match customer preferences. The company needs an observability solution to monitor operational metrics and detect patterns of operational performance degradation compared to established baselines. The solution must also generate alerts with correlation data within 10 minutes when FM behavior deviates from expected patterns.
Which solution will meet these requirements?

A. Configure Amazon CloudWatch Container Insights for the application infrastructur
B. Set up CloudWatch alarms for latency threshold
C. Add custom metrics for token counts by using the CloudWatch embedded metric forma
D. Create CloudWatch dashboards to visualize the data.
E. Implement AWS X-Ray to trace requests through the application component
F. Enable CloudWatch Logs Insights for error pattern detectio
G. Set up AWS CloudTrail to monitor all API calls to Amazon Bedroc
H. Create custom dashboards in Amazon QuickSight.
I. Enable Amazon CloudWatch Application Insights for the application resource
J. Create custom metrics for recommendation quality, token usage, and response latency by using the CloudWatch embedded metric format with dimensions for request types and user segment
K. Configure CloudWatch anomaly detection on the model metric
L. Establish log pattern analysis by using CloudWatch Logs Insights.
M. Use Amazon OpenSearch Service with the Observability plugi
N. Ingest model metrics and logs by using Amazon Kinesi
O. Create custom Piped Processing Language (PPL) queries to analyze model behavior pattern
P. Establish operational dashboards to visualizeanomalies in real time.

**Answer:** C

**NEW QUESTION 30**
A company runs a generative AI (GenAI)-powered summarization application in an application AWS account that uses Amazon Bedrock. The application architecture includes an Amazon API Gateway REST API that forwards requests to AWS Lambda functions that are attached to private VPC subnets. The application summarizes sensitive customer records that the company stores in a governed data lake in a centralized data storage account. The company has enabled Amazon S3, Amazon Athena, and AWS Glue in the data storage account.
The company must ensure that calls that the application makes to Amazon Bedrock use only private connectivity between the company's application VPC and Amazon Bedrock.
The company's data lake must provide fine-grained column-level access across the company's AWS accounts.
Which solution will meet these requirements?

A. In the application account, create interface VPC endpoints for Amazon Bedrock runtime
B. Run Lambda functions in private subnet
C. Use IAM conditions on inference and data-plane policies to allow calls only to approved endpoints and role
D. In the data storage account, use AWS Lake Formation LF-tag-based access control to create table-level and column-level cross-account grants.
E. Run Lambda functions in private subnet
F. Configure a NAT gateway to provide access to Amazon Bedrock and the data lak
G. Use S3 bucket policies and ACLs to manage permission
H. Export AWS CloudTrail logs to Amazon S3 to perform weekly reviews.
I. Create a gateway endpoint only for Amazon S3 in the application accoun
J. Invoke Amazon Bedrock through public endpoint
K. Use database-level grants in AWS Lake Formation to manage data acces
L. Stream AWS CloudTrail logs to Amazon CloudWatch Log
M. Do not set up metric filters or alarms.
N. Use VPC endpoints to provide access to Amazon Bedrock and Amazon S3 in the application accoun
O. Use only IAM path-based policies to manage data lake acces

P. Send AWS CloudTrail logs to Amazon CloudWatch Log
Q. Periodically create dashboards and allow public fallback for cross-Region reads to reduce setup time.

**Answer:** B

**NEW QUESTION 35**
A company upgraded its Amazon Bedrock–powered foundation model (FM) that supports a multilingual customer service assistant. After the upgrade, the assistant exhibited inconsistent behavior across languages. The assistant began generating different responses in some languages when presented with identical questions. The company needs a solution to detect and address similar problems for future updates. The evaluation must be completed within 45 minutes for all supported languages. The evaluation must process at least 15,000 test conversations in parallel. The evaluation process must be fully automated and integrated into the CI/CD pipeline. The solution must block deployment if quality thresholds are not met.
Which solution will meet these requirements?

A. Create a distributed traffic simulation framework that sends translation-heavy workloads to the assistant in multiple languages simultaneousl
B. Use Amazon CloudWatch metrics to monitor latency, concurrency, and throughpu
C. Run simulations before production releases to identify infrastructure bottlenecks.
D. Deploy the assistant in multiple AWS Regions with Amazon Route 53 latency-based routing and AWS Global Accelerator to improve global performanc
E. Store multilingual conversation logs in Amazon S3. Perform weekly post-deployment audits to review consistency.
F. Create a pre-processing pipeline that normalizes all incoming messages into a consistent format before sending the messages to the assistan
G. Apply rule-based checks to flag potential hallucinations in the output
H. Focus evaluation on normalized text to simplify testing across languages.
I. Set up standardized multilingual test conversations with identical meanin
J. Run the test conversations in parallel by using Amazon Bedrock model evaluation job
K. Apply similarity and hallucination threshold
L. Integrate the process into the CI/CD pipeline to block releases that fail.

**Answer:** D

**NEW QUESTION 40**
A healthcare company is using Amazon Bedrock to build a system to help practitioners make clinical decisions. The system must provide treatment recommendations to physicians based only on approved medical documentation and must cite specific sources. The system must not hallucinate or produce factually incorrect information.
Which solution will meet these requirements with the LEAST operational overhead?

A. Integrate Amazon Bedrock with Amazon Kendra to retrieve approved document
B. Implement custom post-processing to compare generated responses against source documents and to include citations.
C. Deploy an Amazon Bedrock Knowledge Base and connect it to approved clinical source document
D. Use the Amazon Bedrock RetrieveAndGenerate API to return citations from the knowledge base.
E. Use Amazon Bedrock and Amazon Comprehend Medical to extract medical entitie
F. Implement verification logic against a medical terminology database.
G. Use an Amazon Bedrock knowledge base with Retrieve API calls and InvokeModel API calls to retrieve approved clinical source document
H. Implement verification logic to compare against retrieved sources and to cite sources.

**Answer:** B

**NEW QUESTION 42**
A retail company has a generative AI (GenAI) product recommendation application that uses Amazon Bedrock. The application suggests products to customers based on browsing history and demographics. The company needs to implement fairness evaluation across multiple demographic groups to detect and measure bias in recommendations between two prompt approaches. The company wants to collect and monitor fairness metrics in real time. The company must receive an alert if the fairness metrics show a discrepancy of more than 15% between demographic groups. The company must receive weekly reports that compare the performance of the two prompt approaches.
Which solution will meet these requirements with the LEAST custom development effort?

A. Configure an Amazon CloudWatch dashboard to display default metrics from Amazon Bedrock API call
B. Create custom metrics based on model output
C. Set up Amazon EventBridge rules to invoke AWS Lambda functions that perform post-processing analysis on model responses and publish custom fairness metrics.
D. Create the two prompt variants in Amazon Bedrock Prompt Managemen
E. Use Amazon Bedrock Flows to deploy the prompt variants with defined traffic allocatio
F. Configure Amazon Bedrock guardrails to monitor demographic fairnes
G. Set up Amazon CloudWatch alarms on the GuardrailContentSource dimension by using InvocationsIntervened metrics to detect recommendation discrepancy threshold violations.
H. Set up Amazon SageMaker Clarify to analyze model output
I. Publish fairness metrics to Amazon CloudWatc
J. Create CloudWatch composite alarms that combine SageMaker Clarify bias metrics with Amazon Bedrock latency metrics.
K. Create an Amazon Bedrock model evaluation job to compare fairness between the two prompt variant
L. Enable model invocation logging in Amazon CloudWatc
M. Set up CloudWatch alarms for InvocationsIntervened metrics with a dimension for each demographic group.

**Answer:** B

**NEW QUESTION 46**
An ecommerce company operates a global product recommendation system that needs to switch between multiple foundation models (FM) in Amazon Bedrock based on regulations,
cost optimization, and performance requirements. The company must apply custom controls based on proprietary business logic, including dynamic cost thresholds, AWS Region-specific compliance rules, and real-time A/B testing across multiple FMs.
The system must be able to switch between FMs without deploying new code. The system must route user requests based on complex rules including user tier, transaction value, regulatory zone, and real-time cost metrics that change hourly and require immediate propagation across thousands of concurrent requests.
Which solution will meet these requirements?

A. Deploy an AWS Lambda function that uses environment variables to store routing rules and Amazon Bedrock FM ID
B. Use the Lambda console to update the environment variables when business requirements chang
C. Configure an Amazon API Gateway REST API to read request parameters to make routing decisions.
D. Deploy Amazon API Gateway REST API request transformation templates to implement routing logic based on request attribute
E. Store Amazon Bedrock FM endpoints as REST API stage variable
F. Update the variables when the system switches between models.
G. Configure an AWS Lambda function to fetch routing configurations from the AWS AppConfig Agent for each user reques
H. Run business logic in the Lambda function to select the appropriate FM for each reques
I. Expose the FM through a single Amazon API Gateway REST API endpoint.
J. Use AWS Lambda authorizers for an Amazon API Gateway REST API to evaluate routing rules that are stored in AWS AppConfi
K. Return authorization contexts based on business logi
L. Route requests to model-specific Lambda functions for each Amazon Bedrock FM.

**Answer:** C

**NEW QUESTION 48**
A company is using Amazon Bedrock to develop a customer support AI assistant. The AI assistant must respond to customer questions about their accounts. The AI assistant must not expose personal information in responses. The company must comply with data residency policies by ensuring that all processing occurs within the same AWS Region where each customer is located.
The company wants to evaluate how effective the AI assistant is at preventing the exposure of personal information before the company makes the AI assistant available to customers.
Which solution will meet these requirements?

A. Configure a cross-Region Amazon Bedrock guardrail to apply sensitive information filter
B. Set the guardrail to detect mode during development and testin
C. Switch to block mode for production deployment.
D. Configure an Amazon Bedrock guardrail to apply sensitive information filter
E. Set theguardrail to mask mode during development and testin
F. Switch to block mode for production deploymen
G. Deploy a copy of the guardrail to each Region where the company operates.
H. Configure an Amazon Bedrock guardrail to apply content and topic filter
I. Set the guardrail to detect mode during development, testing, and productio
J. Disable invocation logging for the Amazon Bedrock model.
K. Configure a cross-Region Amazon Bedrock guardrail to apply a set of content and word filter
L. Set the guardrail to detect mode during development and testin
M. Switch to mask mode for production deployment.

**Answer:** B

**NEW QUESTION 51**
A pharmaceutical company is developing a Retrieval Augmented Generation application that uses an Amazon Bedrock knowledge base. The knowledge base uses Amazon OpenSearch Service as a data source for more than 25 million scientific papers. Users report that the application produces inconsistent answers that cite irrelevant sections of papers when queries span methodology, results, and discussion sections of the papers.
The company needs to improve the knowledge base to preserve semantic context across related paragraphs on the scale of the entire corpus of data.
Which solution will meet these requirements?

A. Configure the knowledge base to use fixed-size chunkin
B. Set a 300-token maximum chunk size and a 10% overlap between chunk
C. Use an appropriate Amazon Bedrock embedding model.
D. Configure the knowledge base to use hierarchical chunkin
E. Use parent chunks that contain 1,000 tokens and child chunks that contain 200 token
F. Set a 50-token overlap between chunks.
G. Configure the knowledge base to use semantic chunkin
H. Use a buffer size of 1 and a breakpoint percentile threshold of 85% to determine chunk boundaries based on content meaning.
I. Configure the knowledge base not to use chunkin
J. Manually split each document into separate files before ingestio
K. Apply post-processing reranking during retrieval.

**Answer:** B

**NEW QUESTION 52**
A financial technology company is using Amazon Bedrock to build an assessment system for the company??s customer service AI assistant. The AI assistant must provide financial recommendations that are factually accurate, compliant with financial regulations, and conversationally appropriate. The company needs to combine automated quality evaluations at scale with targeted human reviews of critical interactions.
What solution will meet these requirements?

A. Configure a pipeline in which financial experts manually score all responses for accuracy, compliance, and conversational qualit
B. Use Amazon SageMaker notebooks to analyze results to identify improvement areas.
C. Configure Amazon Bedrock evaluations that use Anthropic Claude Sonnet as a judge model to assess response accuracy and appropriatenes
D. Configure custom Amazon Bedrock guardrails to check responses for compliance with financial policie
E. Add Amazon Augmented AI (Amazon A2I) human reviews for flagged critical interactions.
F. Create an Amazon Lex bot to manage customer service interaction
G. Configure AWSLambda functions to check responses against a static compliance databas
H. Configure intents that call the Lambda function
I. Add an additional intent to collect end-user reviews.
J. Configure Amazon CloudWatch to monitor response patterns from the AI assistan
K. Configure CloudWatch alerts for potential compliance violation
L. Establish a team of human evaluators to review flagged interactions.

**Answer:** B

**NEW QUESTION 56**
A financial services company uses an AI application to process financial documents by using Amazon Bedrock. During business hours, the application handles approximately 10,000 requests each hour, which requires consistent throughput.
The company uses the CreateProvisionedModelThroughput API to purchase provisioned throughput. Amazon CloudWatch metrics show that the provisioned capacity is unused while on-demand requests are being throttled. The company finds the following code in the application:
response = bedrock_runtime.invoke_model( modelId="anthropic.claude-v2", body=json.dumps(payload)
)
The company needs the application to use the provisioned throughput and to resolve the throttling issues.
Which solution will meet these requirements?

A. Increase the number of model units (MUs) in the provisioned throughput configuration.
B. Replace the model ID parameter with the ARN of the provisioned model that the CreateProvisionedModelThroughput API returns.
C. Add exponential backoff retry logic to handle throttling exceptions during peak hours.
D. Modify the application to use the invokeModelWithResponseStream API instead of the invokeModel API.

**Answer:** B

**NEW QUESTION 59**
A company is creating a workflow to review customer-facing communications before the company sends the communications. The company uses a pre-defined message template to generate the communications and stores the communications in an Amazon S3 bucket. The workflow needs to capture a specific portion from the template and send it to an Amazon Bedrock model. The workflow must store model responses back to the original S3 bucket.
Which solution will meet these requirements?

A. Create a flow in Amazon Bedrock Flow
B. Configure S3 action nodes at the beginning and end of the flow to retrieve and store the communications and the model response
C. In the middle of the flow, configure an expression to parse each communicatio
D. Configure an agent step to send the parsed input to the model for review.
E. Create an AWS Step Functions Express workflow state machin
F. Use an Amazon S3 integration GetObject step to retrieve the original communication
G. Use an intrinsic function Pass step to parse the communications and to pass the results to an Amazon Bedrock InvokeModel ste
H. Configure an Amazon S3 integration PutObject step to store the model responses back to the S3 bucket.
I. Create an Amazon Bedrock agent that has an action grou
J. Configure instructions to define how the agent should parse the communication
K. Configure the action group to retrieve the communications from the S3 bucket, invoke the Amazon Bedrock model, and store the model responses back to the S3 bucket.
L. Create an Amazon Bedrock agent that has a single action grou
M. Configure three AWS Lambda functions in the action grou
N. Configure the functions to retrieve the communications from the S3 bucket, parse the communications and invoke the Amazon Bedrock model, and store the model responses back to the S3 bucket.

**Answer:** A

**NEW QUESTION 63**
A company is planning to deploy multiple generative AI (GenAI) applications to five independent business units that operate in multiple countries in Europe and the Americas.
Each application uses Amazon Bedrock Retrieval Augmented Generation (RAG) patterns with business unit-specific knowledge bases that store terabytes of unstructured data.
The company must establish well-architected, standardized components for security controls, observability practices, and deployment patterns across all the GenAI applications. The components must be reusable, versioned, and governed consistently.
Which solution will meet these requirements?

A. Configure Amazon API Gateway REST API endpoints for the GenAI application
B. Deploy common security, observability, and RAG patterns based on the AWS Well- Architected Generative AI Lens in standardized AWS CloudFormation template
C. Use CloudFormation Guard after deployment to validate policy compliance in each business unit.
D. Create standardized AWS CloudFormation templates to implement security, observability, and RAG patterns based on the AWS Well-Architected Generative AI Len
E. Establish a centralized repository for version contro
F. Integrate a CI/CD pipeline with CloudFormation Guard to enforce consistent and repeatable deployments across business units.
G. Use AWS Service Catalog to define standardized portfolios and versioned products for each business uni
H. Use the portfolios to enforce security, observability, and RAG patterns based on the AWS Well-Architected Generative AI Len
I. Require business units to use the Service Catalog console to deploy resources.
J. Document security controls, observability requirements, and RAG patterns based on the AWS Well-Architected Generative AI Lens in a shared design documen
K. Use Amazon Macie to enforce deploymen
L. Delegate implementation responsibility to each business unit.

**Answer:** B

**NEW QUESTION 65**
A finance company is developing an AI assistant to help clients plan investments and manage their portfolios. The company identifies several high-risk conversation patterns such as requests for specific stock recommendations or guaranteed returns. High-risk conversation patterns could lead to regulatory violations if the company cannot implement appropriate controls.
The company must ensure that the AI assistant does not provide inappropriate financial advice, generate content about competitors, or make claims that are not factually grounded in the company's approved financial guidance. The company wants to use Amazon Bedrock Guardrails to implement a solution.
Which combination of steps will meet these requirements? (Select THREE)

A. Add the high-risk conversation patterns to a denied topics guardrail.
B. Configure a content filter guardrail to filter prompts that contain the high-risk conversation patterns.
C. Configure a content filter guardrail to filter prompts that contain competitor names.

D. Add the names of competitors as custom word filter
E. Set the input and output actions to block.
F. Set a low grounding score threshold.
G. Set a high grounding score threshold.

**Answer:** ADF

**NEW QUESTION 67**
A company wants to select a new FM for its AI assistant. A GenAI developer needs to generate evaluation reports to help a data scientist assess the quality and safety of various foundation models FMs. The data scientist provides the GenAI developer with sample prompts for evaluation. The GenAI developer wants to use Amazon Bedrock to automate report generation and evaluation.
Which solution will meet this requirement?

A. Combine the sample prompts into a single JSON documen
B. Create an Amazon Bedrock knowledge base with the documen
C. Write a prompt that asks the FM to generate a response to each sample promp
D. Use the RetrieveAndGenerate API to generate a report for each model.
E. Combine the sample prompts into a single JSONL documen
F. Store the document in an Amazon S3 bucke
G. Create an Amazon Bedrock evaluation job that uses a judge mode
H. Specify the S3 location as input and a different S3 location as outpu
I. Run an evaluation job for each FM and select the FM as the generator.
J. Combine the sample prompts into a single JSONL documen
K. Store the document in an Amazon S3 bucke
L. Create an Amazon Bedrock evaluation job that uses a judge mode
M. Specify the S3 location as input and Amazon QuickSight as outpu
N. Run an evaluation job for each FM and select the FM as the evaluator.
O. Combine the sample prompts into a single JSON documen
P. Create an Amazon Bedrock knowledge base from the documen
Q. Create an Amazon Bedrock evaluation job that uses the retrieval and response generation evaluation typ
R. Specify an Amazon S3 bucket as the outpu
S. Run an evaluation job for each FM.

**Answer:** B

**NEW QUESTION 68**
A financial services company is developing a real-time generative AI (GenAI) assistant to support human call center agents. The GenAI assistant must transcribe live customer speech, analyze context, and provide incremental suggestions to call center agents while a customer is still speaking. To preserve responsiveness, the GenAI assistant must maintain end-to-end latency under 1 second from speech to initial response display. The architecture must use only managed AWS services and must support bidirectional streaming to ensure that call center agents receive updates in real time.
Which solution will meet these requirements?

A. Use Amazon Transcribe streaming to transcribe call
B. Pass the text to Amazon Comprehend for sentiment analysi
C. Feed the results to Anthropic Claude on Amazon Bedrock by using the InvokeModel AP
D. Store results in Amazon DynamoD
E. Use a WebSocket API to display the results.
F. Use Amazon Transcribe streaming with partial results enabled to deliver fragments oftranscribed text before customers finish speakin
G. Forward text fragments to Amazon Bedrock by using the InvokeModelWithResponseStream AP
H. Stream responses to call center agents through an Amazon API Gateway WebSocket API.
I. Use Amazon Transcribe batch processing to convert calls to tex
J. Pass complete transcripts to Anthropic Claude on Amazon Bedrock by using the ConverseStream AP
K. Return responses through an Amazon Lex chatbot interface.
L. Use the Amazon Transcribe streaming API with an AWS Lambda function to transcribe each audio segmen
M. Call the Amazon Titan Embeddings model on Amazon Bedrock by using the InvokeModel AP
N. Publish results to Amazon SNS.

**Answer:** B

**NEW QUESTION 72**
A financial services company is developing a customer service AI assistant by using Amazon Bedrock. The AI assistant must not discuss investment advice with users. The AI assistant must block harmful content, mask personally identifiable information (PII), and maintain audit trails for compliance reporting. The AI assistant must apply content filtering to both user inputs and model responses based on content sensitivity.
The company requires an Amazon Bedrock guardrail configuration that will effectively enforce policies with minimal false positives. The solution must provide multiple handling strategies for multiple types of sensitive content.
Which solution will meet these requirements?

A. Configure a single guardrail and set content filters to high for all categorie
B. Set up denied topics for investment advice and include sample phrases to bloc
C. Set up sensitive information filters that apply the block action for all PII entitie
D. Apply the guardrail to all model inference calls.
E. Configure multiple guardrails by using tiered policie
F. Create one guardrail and set content filters to hig
G. Configure the guardrail to block PII for public interaction
H. Configure a second guardrail and set content filters to mediu
I. Configure the second guardrail to mask PII for internal us
J. Configure multiple topic-specific guardrails to block investment advice and set up contextual grounding checks.
K. Configure a guardrail and set content filters to medium for harmful conten
L. Set updenied topics for investment advice and include clear definitions and sample phrases to bloc
M. Configure sensitive information filters to mask PII in responses and to block financial information in input

N. Enable both input and output evaluations that use custom blocked messages for audits.
O. Create a separate guardrail for each use cas
P. Create one guardrail that applies a harmful content filte
Q. Create a guardrail to apply topic filters for investment advic
R. Create a guardrail to apply sensitive information filters to block PI
S. Use AWS Step Functions to chain the guardrails sequentially.

**Answer:** C

**NEW QUESTION 76**
A company uses Amazon Bedrock to generate technical content for customers. The company has recently experienced a surge in hallucinated outputs when the company??s model generates summaries of long technical documents. The model outputs include inaccurate or fabricated details. The company??s current solution uses a large foundation model (FM) with a basic one-shot prompt that includes the full document in a single input.
The company needs a solution that will reduce hallucinations and meet factual accuracy goals. The solution must process more than 1,000 documents each hour and deliver summaries within 3 seconds for each document.
Which combination of solutions will meet these requirements? (Select TWO.)

A. Implement zero-shot chain-of-thought (CoT) instructions that require step-by-step reasoning with explicit fact verification before the model generates each summary.
B. Use Retrieval Augmented Generation (RAG) with an Amazon Bedrock knowledge bas
C. Apply semantic chunking and tuned embeddings to ground summaries in source content.
D. Configure Amazon Bedrock guardrails to block any generated output that matches patterns that are associated with hallucinated content.
E. Increase the temperature parameter in Amazon Bedrock.
F. Prompt the Amazon Bedrock model to summarize each full document in one pass.

**Answer:** BC

**NEW QUESTION 80**
A financial services company uses an AI application to process financial documents by using Amazon Bedrock. During business hours, the application handles approximately 10,000 requests each hour, which requires consistent throughput.
The company uses the CreateProvisionedModelThroughput API to purchase provisioned throughput. Amazon CloudWatch metrics show that the provisioned capacity is unused while on-demand requests are being throttled. The company finds the following code in the application:
python
response = bedrock_runtime.invoke_model(modelId="anthropic.claude-v2", body=json.dumps(payload))
The company needs the application to use the provisioned throughput and to resolve the throttling issues.
Which solution will meet these requirements?

A. Increase the number of model units (MUs) in the provisioned throughput configuration.
B. Replace the model ID parameter with the ARN of the provisioned model that the CreateProvisionedModelThroughput API returns.
C. Add exponential backoff retry logic to handle throttling exceptions during peak hours.
D. Modify the application to use the InvokeModelWithResponseStream API instead of the InvokeModel API.

**Answer:** B

**NEW QUESTION 83**
A software company is using Amazon Q Business to build an AI assistant that allows employees to access company information and personal information by using natural language prompts. The company stores this information in an Amazon S3 bucket.
Each department in the company has a dedicated prefix in the S3 bucket. Each object name includes the S3 prefix of the department that it belongs to. Each department can belong to only a single group in AWS IAM Identity Center. Each employee belongs to a single department.
The company configures Amazon Q Business to access data stored in an S3 bucket as a data source. The company needs to ensure that the AI assistant respects access controls based on the user's IAM Identity Center group membership.
Which solution will meet this requirement with the LEAST operational overhead?

A. Create a JSON file named acl.json in each department folde
B. In each file, create access control entries that specify the IAM Identity Center group that should have access to that department's dat
C. Indicate the location of the JSON file in the Access Control section of the data source settings.
D. Create a single JSON file named acl.json at the top level of the S3 bucke
E. Add access control entries that map each department's S3 prefix to its corresponding IAM Identity Center grou
F. Indicate the location of the JSON file in the Access Control section of the data source settings.
G. For each IAM Identity Center group, create a separate permissions set that denies access to all prefixes in the S3 bucke
H. Add a StringNotEquals condition key to the permissions set for each group that specifies the department each group is associated wit
I. Attach the permissions sets to the Identity Center groups.
J. Create a metadata file named metadata.json at the top level of the S3 bucke
K. Add anAccessControlList object to the file that specifies the S3 path of each department's prefi
L. Specify the IAM Identity Center group that should have access to each department's prefi
M. Reference the file location in the data source metadata settings.

**Answer:** B

**NEW QUESTION 87**
A company provides a service that helps users from around the world discover new restaurants. The service has 50 million monthly active users. The company wants to implement a semantic search solution across a database that contains 20 million restaurants and 200 million reviews. The company currently stores the data in PostgreSQL.
The solution must support complex natural language queries and return results for at least 95% of queries within 500 ms. The solution must maintain data freshness for restaurant details that update hourly. The solution must also scale cost-effectively during peak usage periods.
Which solution will meet these requirements with the LEAST development effort?

A. Migrate the restaurant data to Amazon OpenSearch Servic
B. Implement keyword-based search rules that use custom analyzers and relevance tuning to find restaurants based on attributes such as cuisine type, features,

and locatio
C. Create Amazon API Gateway HTTP API endpoints to transform user queries into structured search parameters.
D. Migrate the restaurant data to Amazon OpenSearch Servic
E. Use a foundation model (FM) in Amazon Bedrock to generate vector embeddings from restaurant descriptions, reviews, and menu item
F. When users submit natural language queries, convert the queries to embeddings by using the same F
G. Perform k-nearest neighbors (k-NN) searches to find semantically similar results.
H. Keep the restaurant data in PostgreSQL and implement a pgvector extensio
I. Use a foundation model (FM) in Amazon Bedrock to generate vector embeddings from restaurant dat
J. Store the vector embeddings directly in PostgreSQ
K. Create an AWS Lambda function to convert natural language queries to vector representations by using the same F
L. Configure the Lambda function to perform similarity searches within the database.
M. Migrate restaurant data to an Amazon Bedrock knowledge base by using a custom ingestion pipelin
N. Configure the knowledge base to automatically generate embeddings from restaurant informatio
O. Use the Amazon Bedrock Retrieve API with built-in vector search capabilities to query the knowledge base directly by using natural language input.

**Answer:** B

**NEW QUESTION 91**
A healthcare company is using Amazon Bedrock to build a Retrieval Augmented Generation (RAG) application that helps practitioners make clinical decisions. The application must achieve high accuracy for patient information retrievals, identify hallucinations in generated content, and reduce human review costs.
Which solution will meet these requirements?

A. Use Amazon Comprehend to analyze and classify RAG responses and to extract medical entities and relationship
B. Use AWS Step Functions to orchestrate automated evaluation
C. Configure Amazon CloudWatch metrics to track entity recognition confidence score
D. Configure CloudWatch to send an alert when accuracy falls below specified thresholds.
E. Implement automated large language model (LLM)-based evaluations that use a specialized model that is fine-tuned for medical content to assess all response
F. Deploy AWS Lambda functions to parallelize evaluation
G. Publish results to Amazon CloudWatch metrics that track relevance and factual accuracy.
H. Configure Amazon CloudWatch Synthetics to generate test queries that have known answers on a regular schedule, and track model success rate
I. Set up dashboards that compare synthetic test results against expected outcomes.
J. Deploy a hybrid evaluation system that uses an automated LLM-as-a-judge evaluation to initially screen responses and targeted human reviews for edge case
K. Use a built-in Amazon Bedrock evaluation to track retrieval precision and hallucination rates.

**Answer:** D

**NEW QUESTION 96**
A company is designing a canary deployment strategy for a payment processing API. The system must support automated gradual traffic shifting between multiple Amazon Bedrock models based on real-time inference metrics, historical traffic patterns, and service health. The solution must be able to gradually increase traffic to new model versions. The system must increase traffic if metrics remain healthy and decrease traffic if the performance degrades below acceptable thresholds. The company needs to comprehensively monitor inference latency and error rates during the deployment phase. The company must also be able to halt deployments and revert to a previous model version without any manual intervention.
Which solution will meet these requirements?

A. Use Amazon Bedrock with provisioned throughput to host model version
B. Configure an Amazon EventBridge rule to invoke an AWS Step Functions workflow when a new model version is release
C. Configure the workflow to shift traffic in stages, wait for a specified time period, and invoke an AWS Lambda function to check Amazon CloudWatch performance metric
D. Configure the workflow to increase traffic if metrics meet thresholds and to trigger a traffic rollback if performance metrics fall below thresholds.
E. Use AWS Lambda functions to invoke various Amazon Bedrock model version
F. Use an Amazon API Gateway HTTP API with stage variables and weighted routing to shift traffic graduall
G. Use Amazon CloudWatch to monitor performanc
H. Use external logic to adjust traffic and roll back if performance falls below thresholds.
I. Use Amazon SageMaker AI endpoint variants to represent multiple Amazon Bedrock model version
J. Use variant weights to shift traffi
K. Use Amazon CloudWatch and SageMaker Model Monitor to trigger rollback
L. Use EventBridge to roll back deployments if an anomaly is detected.
M. Use Amazon OpenSearch Service to track inference log
N. Configure OpenSearch Service to invoke an AWS Systems Manager Automation runbook to update Amazon Bedrock model endpoints to shift traffic based on inference logs.

**Answer:** A

**NEW QUESTION 101**
A financial services company needs to pre-process unstructured data such as customer transcripts, financial reports, and documentation. The company stores the unstructured data in Amazon S3 to support an Amazon Bedrock application.
The company must validate data quality, create auditable metadata, monitor data metrics, and customize text chunking to optimize foundation model (FM) performance.
Which solution will meet these requirements with the LEAST development effort?

A. Use Amazon SageMaker Data Wrangler to create a data flo
B. Configure Amazon CloudWatch metrics and alarms to monitor data qualit
C. Use a custom AWS Lambda function to pre-process the dat
D. Load processed data into Amazon Bedrock.
E. Set up an AWS Glue crawler to catalog data source
F. Create AWS Glue ETL jobs to run custom transformation script
G. Use AWS Glue Data Quality to validate and monitor data qualit
H. Load processed data into Amazon Bedrock.
I. Use Amazon Comprehend to extract entitie
J. Create an AWS Lambda function to chunk tex

K. Run Amazon Athena to query and validate data qualit
L. Load processed data into Amazon Bedrock.
M. Create an AWS Step Functions workflow to orchestrate data pre-processing task
N. Run custom code on Amazon EC2 instance
O. Use Amazon SageMaker Model Monitor to monitor data qualit
P. Load processed data into Amazon Bedrock.

**Answer:** B

**NEW QUESTION 105**
A financial services company is developing a Retrieval Augmented Generation (RAG) application to help investment analysts query complex financial relationships across multiple investment vehicles, market sectors, and regulatory environments. The dataset contains highly interconnected entities that have multi-hop relationships. Analysts must examine relationships holistically to provide accurate investment guidance. The application must deliver comprehensive answers that capture indirect relationships between financial entities and must respond in less than 3 seconds.
Which solution will meet these requirements with the LEAST operational overhead?

A. Use Amazon Bedrock Knowledge Bases with GraphRAG and Amazon Neptune Analytics to store financial dat
B. Analyze multi-hop relationships between entities and automatically identify related information across documents.
C. Use Amazon Bedrock Knowledge Bases and an Amazon OpenSearch Service vector store to implement custom relationship identification logic that uses AWS Lambda to query multiple vector embeddings in sequence.
D. Use Amazon OpenSearch Serverless vector search with k-nearest neighbor (k-NN). Implement manual relationship mapping in an application layer that runs on Amazon EC2 Auto Scaling.
E. Use Amazon DynamoDB to store financial data in a custom indexing syste
F. Use AWS Lambda to query relevant record
G. Use Amazon SageMaker to generate responses.

**Answer:** A

**NEW QUESTION 106**
A bank is building a generative AI (GenAI) application that uses Amazon Bedrock to assess loan applications by using scanned financial documents. The application must extract structured data from the documents. The application must redact personally identifiable information (PII) before inference. The application must use foundation models (FMs) to generate approvals. The application must route low-confidence document extraction results to human reviewers who are within the same AWS Region as the loan applicant.
The company must ensure that the application complies with strict Regional data residency and auditability requirements. The application must be able to scale to handle 25,000 applications each day and provide 99.9% availability.
Which combination of solutions will meet these requirements? (Select THREE.)

A. Deploy Amazon Textract and Amazon Augmented AI within the same Region to extract relevant data from the scanned document
B. Route low-confidence pages to human reviewers.
C. Use AWS Lambda functions to detect and redact PII from submitted documents before inferenc
D. Apply Amazon Bedrock guardrails to prevent inappropriate or unauthorized content in model output
E. Configure Region-specific IAM roles to enforce data residency requirements and to control access to the extracted data.
F. Use Amazon Kendra and Amazon OpenSearch Service to extract field-level values semantically from the uploaded documents before inference.
G. Store uploaded documents in Amazon S3 and apply object metadat
H. Configure IAM policies to store original documents within the same Region as each applican
I. Enable object tagging for future audits.
J. Use AWS Glue Data Quality to validate the structured document dat
K. Use AWS Step Functions to orchestrate a review workflow that includes a prompt engineering step that transforms validated data into optimized prompts before invoking Amazon Bedrock to assess loan applications.
L. Use Amazon SageMaker Clarify to generate fairness and bias reports based on model scoring decisions that Amazon Bedrock makes.

**Answer:** ABD

**NEW QUESTION 111**
An enterprise application uses an Amazon Bedrock foundation model (FM) to process and analyze 50 to 200 pages of technical documents. Users are experiencing inconsistent responses and receiving truncated outputs when processing documents that exceed the FM's context window limits.
Which solution will resolve this problem?

A. Configure fixed-size chunking at 4,000 tokens for each chunk with 20% overla
B. Use application-level logic to link multiple chunks sequentially until the FM's maximum context window of 200,000 tokens is reached before making inference calls.
C. Use hierarchical chunking with parent chunks of 8,000 tokens and child chunks of 2,000 token
D. Use Amazon Bedrock Knowledge Bases built-in retrieval to automatically select relevant parent chunks based on query contex
E. Configure overlap tokens to maintain semantic continuity.
F. Use semantic chunking with a breakpoint percentile threshold of 95% and a buffer size of 3 sentence
G. Use the RetrieveAndGenerate API to dynamically select the most relevant chunks based on embedding similarity scores.
H. Create a pre-processing AWS Lambda function that analyzes document token count by using the FM's tokenize
I. Configure the Lambda function to split documents into equal segments that fit within 80% of the context windo
J. Configure the Lambda function to process each segment independently before aggregating the results.

**Answer:** C

**NEW QUESTION 114**
An insurance company uses existing Amazon SageMaker AI infrastructure to support a web-based application that allows customers to predict what their insurance premiums will be. The company stores customer data that is used to train the SageMaker AI model in an Amazon S3 bucket. The dataset is growing rapidly. The company wants a solution to continuously re-train the model. The solution must automatically re-train and re-deploy the model to the application when an employee uploads a new customer data file to the S3 bucket.
Which solution will meet these requirements?

A. Use AWS Glue to run an ETL job on each uploaded fil
B. Configure the ETL job to use the AWS SDK to invoke the SageMaker AI model endpoin
C. Use real-time inference with the endpoint to re-deploy the model after it is re-trained on the updated customer dataset.
D. Create an AWS Lambda function and webhook handlers to generate an event when an employee uploads a new fil
E. Configure SageMaker Pipelines to re-deploy the model after it is re-trained on the updated customer datase
F. Use Amazon EventBridge to create an event bu
G. Set the Lambda function event as the source and SageMaker Pipelines as the target.
H. Create an AWS Step Functions Express workflow with AWS SDK integrations to retrieve the customer data from the S3 bucket when an employee uploads a new file to the S3 bucke
I. Use a SageMaker Data Wrangler flow to export the data from the S3 bucket to SageMaker Autopilo
J. Use the SageMaker Autopilot to re-deploy the model after it has been re-trained on the updated customer dataset.
K. Create an AWS Step Functions Standard workflo
L. Configure the first state to call an AWS Lambda function to respond when an employee uploads a new file to the S3 bucke
M. Use a pipeline in SageMaker Pipelines to re-deploy the model after it has been re-trained on the updated customer datase
N. Use the next state in the workflow to run the pipeline when the first state receives a response.

**Answer:** D

**NEW QUESTION 117**
A legal research company has a Retrieval Augmented Generation (RAG) application that uses Amazon Bedrock and Amazon OpenSearch Service. The application stores 768- dimensional vector embeddings for 15 million legal documents, including statutes, court rulings, and case summaries.
The company's current chunking strategy segments text into fixed-length blocks of 500 tokens. The current chunking strategy often splits contextually linked information such as legal arguments, court opinions, or statute references across separate chunks. Researchers report that generated outputs frequently omit key context or cite outdated legal information.
Recent application logs show a 40% increase in response times. The p95 latency metric exceeds 2 seconds. The company expects storage needs for the application to grow from 90 GB to 360 GB within a year.
The company needs a solution to improve retrieval relevance and system performance at scale.
Which solution will meet these requirements?

A. Increase the embedding vector dimensionality from 768 to 4,096 without changing the existing chunking or pre-processing strategy.
B. Replace dynamic retrieval with static, pre-written summaries that are stored in Amazon S3. Use Amazon CloudFront to serve the summaries to reduce compute demand and improve predictability.
C. Update the chunking strategy to use semantic boundaries such as complete legal arguments, clauses, or sections rather than fixed token limit
D. Regenerate vector embeddings to align with the new chunk structure.
E. Migrate from OpenSearch Service to Amazon DynamoD
F. Implement keyword-based indexes to enable faster lookups for legal concepts.

**Answer:** C

**NEW QUESTION 120**
A company is using Amazon Bedrock and Anthropic Claude 3 Haiku to develop an AI assistant. The AI assistant normally processes 10,000 requests each hour but experiences surges of up to 30,000 requests each hour during peak usage periods. The AI assistant must respond within 2 seconds while operating across multiple AWS Regions.
The company observes that during peak usage periods, the AI assistant experiences throughput bottlenecks that cause increased latency and occasional request timeouts. The company must resolve the performance issues.
Which solution will meet this requirement?

A. Purchase provisioned throughput and sufficient model units (MUs) in a single Regio
B. Configure the application to retry failed requests with exponential backoff.
C. Implement token batching to reduce API overhea
D. Use cross-Region inference profiles to automatically distribute traffic across available Regions.
E. Set up auto scaling AWS Lambda functions in each Regio
F. Implement client-side round-robin request distributio
G. Purchase one model unit (MU) of provisioned throughput as a backup.
H. Implement batch inference for all requests by using Amazon S3 buckets across multiple Region
I. Use Amazon SQS to set up an asynchronous retrieval process.

**Answer:** B

**NEW QUESTION 124**
A company is building an AI advisory application by using Amazon Bedrock. The application will provide recommendations to customers. The company needs the application to explain its reasoning process and cite specific sources for data. The application must retrieve information from company data sources and show step-by-step reasoning for recommendations. The application must also link data claims to source documents and maintain response latency under 3 seconds.
Which solution will meet these requirements with the LEAST operational overhead?

A. Use Amazon Bedrock Knowledge Bases with source attribution enable
B. Use the Anthropic Claude Messages API with RAG to set high-relevance thresholds for sourcedocument
C. Store reasoning and citations in Amazon S3 for auditing purposes.
D. Use Amazon Bedrock with Anthropic Claude models and extended thinkin
E. Configure a 4,000-token thinking budge
F. Store reasoning traces and citations in Amazon DynamoDB for auditing purposes.
G. Configure Amazon SageMaker AI with a custom Anthropic Claude mode
H. Use the model??s reasoning parameter and AWS Lambda to process response
I. Add source citations from a separate Amazon RDS database.
J. Use Amazon Bedrock with Anthropic Claude models and chain-of-thought reasonin
K. Configure custom retrieval tracking with the Amazon Bedrock Knowledge Bases AP
L. Use Amazon CloudWatch to monitor response latency metrics.

**Answer:** A

**NEW QUESTION 125**
A company is developing a customer support application that uses Amazon Bedrock foundation models (FMs) to provide real-time AI assistance to the company??s employees. The application must display AI-generated responses character by character as the responses are generated. The application needs to support thousands of concurrent users with minimal latency. The responses typically take 15 to 45 seconds to finish.
Which solution will meet these requirements?

A. Configure an Amazon API Gateway WebSocket API with an AWS Lambda integratio
B. Configure the WebSocket API to invoke the Amazon Bedrock InvokeModelWithResponseStream API and stream partial responses through WebSocket connections.
C. Configure an Amazon API Gateway REST API with an AWS Lambda integratio
D. Configure the REST API to invoke the Amazon Bedrock standard InvokeModel API and implement frontend client-side polling every 100 ms for complete response chunks.
E. Implement direct frontend client connections to Amazon Bedrock by using IAM user credentials and the InvokeModelWithResponseStream API without any intermediate gateway or proxy layer.
F. Configure an Amazon API Gateway HTTP API with an AWS Lambda integratio
G. Configure the HTTP API to cache complete responses in an Amazon DynamoDB table and serve the responses through multiple paginated GET requests to frontend clients.

**Answer:** A

**NEW QUESTION 129**
A company is developing a generative AI (GenAI) application by using Amazon Bedrock. The application will analyze patterns and relationships in the company??s data. The application will process millions of new data points daily across AWS Regions in Europe, North America, and Asia before storing the data in Amazon S3. The application must comply with local data protection and storage regulations. Data residency and processing must occur within the same continent. The application must also maintain audit trails of the application??s decision-making processes and provide data classification capabilities.
Which solution will meet these requirements?

A. Deploy the application in each Region with local IAM policie
B. Use Amazon Bedrock cross-Region inference to distribute the workloa
C. Use Amazon CloudWatch to log AI decision-making processe
D. Manually track compliance certifications across Regions.
E. Use SCPs with AWS Organizations to manage location-specific permission
F. Use AWS CloudTrail immutable logs to audit decision-making processe
G. Import a custom model into Amazon Bedrock and deploy the model to each Region.
H. Use Amazon S3 Object Lock with Region-specific S3 bucket policie
I. Pre-process the data points within the Region based on geographic origin before sending the data points to Amazon Bedroc
J. Use Amazon Macie to classify the dat
K. Use AWS CloudTrail immutable logs to audit the decision-making processes.
L. Create separate AWS accounts for each Region with individual compliance framework
M. Use Amazon SageMaker AI with custom monitorin
N. Create manual compliance reports for each regulatory jurisdiction.

**Answer:** C

**NEW QUESTION 133**
A company is developing a customer communication platform that uses an AI assistant powered by an Amazon Bedrock foundation model (FM). The AI assistant summarizes customer messages and generates initial response drafts.
The company wants to use Amazon Comprehend to implement layered content filtering. The layered content filtering must prevent sharing of offensive content, protect customer privacy, and detect potential inappropriate advice solicitation. Inappropriate advice solicitation includes requests for unethical practices, harmful activities, or manipulative behaviors.
The solution must maintain acceptable overall response times, so all pre-processing filters must finish before the content reaches the FM.
Which solution will meet these requirements?

A. Use parallel processing with asynchronous API call
B. Use toxicity detection for offensive conten
C. Use prompt safety classification for inappropriate advice solicitatio
D. Use personally identifiable information (PII) detection without redaction.
E. Use custom classification to build an FM that detects offensive content and inappropriate advice solicitatio
F. Apply personally identifiable information (PII) detection as a secondary filter only when messages pass the custom classifier.
G. Deploy a multi-stage proces
H. Configure the process to use prompt safety classification first, then toxicity detection on safe prompts only, and finally personally identifiable information (PII) detection in streaming mod
I. Route flagged messages through Amazon EventBridge for human review.
J. Use toxicity detection with thresholds configured to 0.5 for all categorie
K. Use parallel processing for both prompt safety classification and personally identifiable information (PII) detection with entity redactio
L. Apply Amazon CloudWatch alarms to filter metrics.

**Answer:** D

**NEW QUESTION 138**
.....

# THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual AIP-C01 Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the AIP-C01 Product From:

**https://www.2passeasy.com/dumps/AIP-C01/**

## Money Back Guarantee

## AIP-C01 Practice Exam Features:

* AIP-C01 Questions and Answers Updated Frequently

* AIP-C01 Practice Questions Verified by Expert Senior Certified Staff

* AIP-C01 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* AIP-C01 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year