

Amazon-Web-Services

Exam Questions AIP-C01

AWS Certified Generative AI Developer - Professional



NEW QUESTION 1

A financial services company is deploying a generative AI (GenAI) application that uses Amazon Bedrock to assist customer service representatives to provide personalized investment advice to customers. The company must implement a comprehensive governance solution that follows responsible AI practices and meets regulatory requirements.

The solution must detect and prevent hallucinations in recommendations. The solution must have safety controls for customer interactions. The solution must also monitor model behavior drift in real time and maintain audit trails of all prompt-response pairs for regulatory review. The company must deploy the solution within 60 days. The solution must integrate with the company's existing compliance dashboard and respond to customers within 200 ms.

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

- A. Configure Amazon Bedrock guardrails to apply custom content filters and toxicity detection
- B. Use Amazon Bedrock Model Evaluation to detect hallucination
- C. Store prompt-response pairs in Amazon DynamoDB to capture audit trails and set a TTL
- D. Integrate Amazon CloudWatch custom metrics with the existing compliance dashboard.
- E. Deploy Amazon Bedrock and use AWS PrivateLink to access the application securely
- F. Use AWS Lambda functions to implement custom prompt validation
- G. Store prompt-response pairs in an Amazon S3 bucket and configure S3 Lifecycle policies
- H. Create custom Amazon CloudWatch dashboards to monitor model performance metrics.
- I. Use Amazon Bedrock Agents and Amazon Bedrock Knowledge Bases to ground responses
- J. Use Amazon Bedrock Guardrails to enforce content safety
- K. Use Amazon OpenSearch Service to store and index prompt-response pairs
- L. Integrate OpenSearch Service with Amazon QuickSight to create compliance reports and to detect model behavior drift.
- M. Use Amazon SageMaker Model Monitor to detect model behavior drift
- N. Use AWS WAF to filter content
- O. Store customer interactions in an encrypted Amazon RDS database
- P. Use Amazon API Gateway to create custom HTTP APIs to integrate with the compliance dashboard.

Answer: A

NEW QUESTION 2

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 configuration
- 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 search
- 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 documents
- 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 3

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 Bedrock
- B. Use Lambda response streaming to stream response
- C. Enforce token limits within the Lambda function
- 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 Bedrock
- F. Simulate streaming by using client-side polling
- G. Enforce token limits on the frontend
- 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 server
- J. Stream responses by using the WebSocket protocol
- K. Enforce token limits within Amazon EC2
- 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 Bedrock
- N. Use Lambda response streaming to stream response
- O. Enforce token limits within the Lambda function
- P. Implement retry logic by using Lambda and API Gateway timeout configurations.

Answer: A

NEW QUESTION 4

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 endpoint
- 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 endpoint
- 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 5

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 6

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 7

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 infrastru
- 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 8

An ecommerce company is using Amazon Bedrock to build a generative AI (GenAI) application. The application uses AWS Step Functions to orchestrate a multi-agent workflow to produce detailed product descriptions. The workflow consists of three sequential states: a description generator, a technical specifications validator, and a brand voice consistency checker. Each state produces intermediate reasoning traces and outputs that are passed to the next state. The application uses an Amazon S3 bucket for process storage and to store outputs.

During testing, the company discovers that outputs between Step Functions states frequently exceed the 256 KB quota and cause workflow failures. A GenAI Developer needs to revise the application architecture to efficiently handle the Step Functions 256 KB quota and maintain workflow observability. The revised architecture must preserve the existing multi-agent reasoning and acting (ReAct) pattern.

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

- A. Store intermediate outputs in Amazon DynamoD
- B. Pass only references between state
- C. Create a Map state that retrieves the complete data from DynamoDB when required for each agent's processing step.
- D. Configure an Amazon Bedrock integration to use the S3 bucket URI in the input parameters for large output
- E. Use the ResultPath and ResultSelector fields to route S3 references between the agent steps while maintaining the sequential validation workflow.
- F. Use AWS Lambda functions to compress outputs to less than 256 KB before each agent stat
- G. Configure each agent task to decompress outputs before processing and to compress results before passing them to the next state.
- H. Configure a separate Step Functions state machine to handle each agent's processin
- I. Use Amazon EventBridge to coordinate the execution flow between state machine
- J. Use S3 references for the outputs as event data.

Answer: B

NEW QUESTION 9

A financial services company uses multiple foundation models (FMs) through Amazon Bedrock for its generative AI (GenAI) applications. To comply with a new regulation for GenAI use with sensitive financial data, the company needs a token management solution.

The token management solution must proactively alert when applications approach model- specific token limits. The solution must also process more than 5,000 requests each minute and maintain token usage metrics to allocate costs across business units.

Which solution will meet these requirements?

- A. Develop model-specific tokenizers in an AWS Lambda functio
- B. Configure the Lambda function to estimate token usage before sending requests to Amazon Bedroc
- C. Configure the Lambda function to publish metrics to Amazon CloudWatch and trigger alarms when requests approach threshold
- D. Store detailed token usage in Amazon DynamoDB to report costs.
- E. Implement Amazon Bedrock Guardrails with token quota policie
- F. Capture metrics on rejected request
- G. Configure Amazon EventBridge rules to trigger notifications based on Amazon Bedrock Guardrails metric
- H. Use Amazon CloudWatch dashboards to visualize token usage trends across models.
- I. Deploy an Amazon SQS dead-letter queue for failed request
- J. Configure an AWS Lambda function to analyze token-related failure
- K. Use Amazon CloudWatch Logs Insights to generate reports on token usage patterns based on error logs from Amazon Bedrock API responses.
- L. Use Amazon API Gateway to create a proxy for all Amazon Bedrock API call
- M. Configure request throttling based on custom usage plans with predefined token quota
- N. Configure API Gateway to reject requests that will exceed token limits.

Answer: A

NEW QUESTION 10

A wildlife conservation agency operates zoos globally. The agency uses various sensors, trackers, and audiovisual recorders to monitor animal behavior. The agency wants to launch a generative AI (GenAI) assistant that can ingest multimodal data to study animal behavior.

The GenAI assistant must support natural language queries, avoid speculative behavioral interpretations, and maintain audit logs for ethical research audits.

Which solution will meet these requirements?

- A. Ingest raw videos into Amazon Rekognition to detect animal postures and expression
- B. Use Amazon Data Firehose to stream sensor and GPS data into Amazon S3. Prompt an Amazon Bedrock FM using basic templates stored in AWS Systems Manager Parameter Stor
- C. Use IAM for access contro
- D. Use AWS CloudTrail for audit logging.
- E. Use Amazon SageMaker Processing and Amazon Transcribe to pre-process multimodal dat
- F. Ingest curated summaries into an Amazon Bedrock Knowledge Base
- G. Apply Amazon Bedrock guardrails to restrict speculative output
- H. Use AWS AppConfig to manage prompt template
- I. Use AWS CloudTrail to log research activity for audits.
- J. Use Amazon OpenSearch Serverless to index behavioral logs and telemetr
- K. Use Amazon Comprehend to extract entitie
- L. Use Amazon Bedrock to answer questions over indexed dat
- M. Use IAM for access control and CloudTrail for audit logging.
- N. Configure Amazon O Business to federate data across Amazon S3, Amazon Kinesis, and Amazon SageMaker Feature Stor
- O. Use EventBridge for ingestion orchestratio
- P. Use custom AWS Lambda functions to filter LLM outputs for ethical compliance.

Answer: B

NEW QUESTION 10

A company is developing a generative AI (GenAI)-powered customer support application that uses Amazon Bedrock foundation models (FMs). The application must maintain conversational context across multiple interactions with the same user. The application must run clarification workflows to handle ambiguous user queries. The company must store encrypted records of each user conversation to use for personalization. The application must be able to handle thousands of

concurrent users while responding to each user quickly.
Which solution will meet these requirements?

- A. Use an AWS Step Functions Express workflow to orchestrate conversation flow
- B. Invoke AWS Lambda functions to run clarification logic
- C. Store conversation history in Amazon RDS and use session IDs as the primary key.
- D. Use an AWS Step Functions Standard workflow to orchestrate clarification workflow
- E. Include Wait for a Callback patterns to manage the workflow
- F. Store conversation history in Amazon DynamoDB
- G. Purchase on-demand capacity and configure server-side encryption.
- H. Deploy the application by using an Amazon API Gateway REST API to route user requests to an AWS Lambda function to update and retrieve conversation context
- I. Store conversation history in Amazon S3 and configure server-side encryption
- J. Save each interaction as a separate JSON file.
- K. Use AWS Lambda functions to call Amazon Bedrock inference API
- L. Use Amazon SQS queues to orchestrate clarification step
- M. Store conversation history in an Amazon ElastiCache (Redis OSS) cluster
- N. Configure encryption at rest.

Answer: B

NEW QUESTION 14

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 testing
- C. Switch to block mode for production deployment.
- D. Configure an Amazon Bedrock guardrail to apply sensitive information filter
- E. Set the guardrail to mask mode during development and testing
- F. Switch to block mode for production deployment
- 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 production
- 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 testing
- M. Switch to mask mode for production deployment.

Answer: B

NEW QUESTION 15

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 chunking
- B. Set a 300-token maximum chunk size and a 10% overlap between chunks
- C. Use an appropriate Amazon Bedrock embedding model.
- D. Configure the knowledge base to use hierarchical chunking
- E. Use parent chunks that contain 1,000 tokens and child chunks that contain 200 tokens
- F. Set a 50-token overlap between chunks.
- G. Configure the knowledge base to use semantic chunking
- 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 chunking
- J. Manually split each document into separate files before ingestion
- K. Apply post-processing reranking during retrieval.

Answer: B

NEW QUESTION 20

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 21

A company developed a multimodal content analysis application by using Amazon Bedrock. The application routes different content types (text, images, and code) to specialized foundation models (FMs).

The application needs to handle multiple types of routing decisions. Simple routing based on file extension must have minimal latency. Complex routing based on content semantics requires analysis before FM selection. The application must provide detailed history and support fallback options when primary FMs fail.

Which solution will meet these requirements?

- A. Configure AWS Lambda functions that call Amazon Bedrock FMs for all routing logi
- B. Use conditional statements to determine the appropriate FM based on content type and semantics.
- C. Create a hybrid solutio
- D. Handle simple routing based on file extensions in application cod
- E. Handle complex content-based routing by using an AWS Step Functions state machine with JSONata for content analysis and the InvokeModel API for specialized FMs.
- F. Deploy separate AWS Step Functions workflows for each content type with routing logic in AWS Lambda function
- G. Use Amazon EventBridge to coordinate between workflows when fallback to alternate FMs is required.
- H. Use Amazon SQS with different SQS queues for each content typ
- I. Configure AWS Lambda consumers that analyze content and invoke appropriate FMs based on message attributes by using Amazon Bedrock with an AWS SDK.

Answer: B

NEW QUESTION 22

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 26

A medical device company wants to feed reports of medical procedures that used the company's devices into an AI assistant. To protect patient privacy, the AI assistant must expose patient personally identifiable information (PII) only to surgeons. The AI assistant must redact PII for engineers. The AI assistant must reference only medical reports that are less than 3 years old.

The company stores reports in an Amazon S3 bucket as soon as each report is published. The company has already set up an Amazon Bedrock Knowledge Bases. The AI assistant uses Amazon Cognito to authenticate users.

Which solution will meet these requirements?

- A. Enable Amazon Macie PII detection on the S3 bucke
- B. Use an S3 trigger to invoke an AWS Lambda function that redacts PII from the report
- C. Configure the Lambda function to delete outdated documents and invoke knowledge base syncing.
- D. Invoke an AWS Lambda function to sync the S3 bucket and the knowledge base when a new report is uploade
- E. Use a second Lambda function with Amazon Comprehend to redact PII for engineer
- F. Use S3 Lifecycle rules to remove reports older than 3 years.
- G. Set up an S3 Lifecycle configuration to remove reports that are older than 3 year
- H. Schedule an AWS Lambda function to run daily syncs between the bucket and the knowledge bas
- I. When users interact with the AI assistant, apply a guardrail configuration selected based on the user's Cognito user group to redact PII from responses when required.
- J. Create a second knowledge bas
- K. Use Lambda and Amazon Comprehend to redact PII before syncing to the second knowledge bas
- L. Route users to the appropriate knowledge base based on Cognito group membership.

Answer: C

NEW QUESTION 31

A medical company is building a generative AI (GenAI) application that uses Retrieval Augmented Generation (RAG) to provide evidence-based medical information. The application uses Amazon OpenSearch Service to retrieve vector embeddings. Users report that searches frequently miss results that contain exact medical terms and acronyms and return too many semantically similar but irrelevant documents. The company needs to improve retrieval quality and maintain low end-user latency, even as the document collection grows to millions of documents.

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

- A. Configure hybrid search by combining vector similarity with keyword matching to improve semantic understanding and exact term and acronym matching.
- B. Increase the dimensions of the vector embeddings from 384 to 1536. Use a post- processing AWS Lambda function to filter out irrelevant results after retrieval.
- C. Replace OpenSearch Service with Amazon Kendr
- D. Use query expansion to handle medical acronyms and terminology variants during pre-processing.
- E. Implement a two-stage retrieval architecture in which initial vector search results are re- ranked by an ML model hosted on Amazon SageMaker.

Answer: A

NEW QUESTION 33

A company is developing a generative AI (GenAI) application that analyzes customer service calls in real time and generates suggested responses for human customer service agents. The application must process 500,000 concurrent calls during peak hours with less than 200 ms end-to-end latency for each suggestion. The company uses existing architecture to transcribe customer call audio streams. The application must not exceed a predefined monthly compute budget and must maintain auto scaling capabilities.

Which solution will meet these requirements?

- A. Deploy a large, complex reasoning model on Amazon Bedrock
- B. Purchase provisioned throughput and optimize for batch processing.
- C. Deploy a low-latency, real-time optimized model on Amazon Bedrock
- D. Purchase provisioned throughput and set up automatic scaling policies.
- E. Deploy a large language model (LLM) on an Amazon SageMaker real-time endpoint that uses dedicated GPU instances.
- F. Deploy a mid-sized language model on an Amazon SageMaker serverless endpoint that is optimized for batch processing.

Answer: B

NEW QUESTION 34

A pharmaceutical company is developing a Retrieval Augmented Generation (RAG) 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 35

A company uses AWS Lake Formation to set up a data lake that contains databases and tables for multiple business units across multiple AWS Regions. The company wants to use a foundation model (FM) through Amazon Bedrock to perform fraud detection. The FM must ingest sensitive financial data from the data lake. The data includes some customer personally identifiable information (PII).

The company must design an access control solution that prevents PII from appearing in a production environment. The FM must access only authorized data subsets that have PII redacted from specific data columns. The company must capture audit trails for all data access.

Which solution will meet these requirements?

- A. Create a separate dataset in a separate Amazon S3 bucket for each business unit and Region combinatio
- B. Configure S3 bucket policies to control access based on IAM roles that are assigned to FM training instance
- C. Use S3 access logs to track data access.
- D. Configure the FM to authenticate by using AWS Identity and Access Management roles and Lake Formation permissions based on LF-Tag expression
- E. Define business units and Regions as LF-Tags that are assigned to databases and table
- F. Use AWS CloudTrail to collect comprehensive audit trails of data access.
- G. Use direct IAM principal grants on specific databases and tables in Lake Formatio
- H. Create a custom application layer that logs access requests and further filters sensitive columns before sending data to the FM.
- I. Configure the FM to request temporary credentials from AWS Security Token Servic
- J. Access the data by using presigned S3 URLs that are generated by an API that applies business unit and Regional filter
- K. Use AWS CloudTrail to collect comprehensive audit trails of data access.

Answer: B

NEW QUESTION 38

A company has a customer service application that uses Amazon Bedrock to generate personalized responses to customer inquiries. The company needs to establish a quality assurance process to evaluate prompt effectiveness and model configurations across updates. The process must automatically compare outputs from multiple prompt templates, detect response quality issues, provide quantitative metrics, and allow human reviewers to give feedback on responses. The process must prevent configurations that do not meet a predefined quality threshold from being deployed.

Which solution will meet these requirements?

- A. Create an AWS Lambda function that sends sample customer inquiries to multiple Amazon Bedrock model configurations and stores responses in Amazon S3. Use Amazon QuickSight to visualize response pattern
- B. Manually review outputs dail
- C. Use AWS CodePipeline to deploy configurations that meet the quality threshold.
- D. Use Amazon Bedrock evaluation jobs to compare model outputs by using custom prompt dataset
- E. Configure AWS CodePipeline to run the evaluation jobs when prompt templates chang
- F. Configure CodePipeline to deploy only configurations that exceed the predefined quality threshold.
- G. Set up Amazon CloudWatch alarms to monitor response latency and error rates from Amazon Bedroc
- H. Use Amazon EventBridge rules to notify teams when thresholds are excee
- I. Configure a manual approval workflow in AWS Systems Manager.
- J. Use AWS Lambda functions to create an automated testing framework that samples production traffic and routes duplicate requests to the updated model versio
- K. Use Amazon Comprehend sentiment analysis to compare result
- L. Block deployment if sentiment scores decrease.

Answer: B

NEW QUESTION 42

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 data
- 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 system
- F. Use AWS Lambda to query relevant records
- G. Use Amazon SageMaker to generate responses.

Answer: A

NEW QUESTION 47

A media company is launching a platform that allows thousands of users every hour to upload images and text content. The platform uses Amazon Bedrock to process the uploaded content to generate creative compositions.

The company needs a solution to ensure that the platform does not process or produce inappropriate content. The platform must not expose personally identifiable information (PII) in the compositions. The solution must integrate with the company's existing Amazon S3 storage workflow.

Which solution will meet these requirements with the LEAST infrastructure management overhead?

- A. Enable the Enhanced Monitoring tool
- B. Use an Amazon CloudWatch alarm to filter traffic to the platform
- C. Use Amazon Comprehend PII detection to pre-process the data
- D. Create a CloudWatch alarm to monitor for Amazon Comprehend PII detection events
- E. Create an AWS Step Functions workflow that includes an Amazon Rekognition image moderation step.
- F. Use an Amazon API Gateway HTTP API with request validation templates to screen content before storing the uploaded content in Amazon S3. Use Amazon SageMaker AI to build custom content moderation models that process content before sending the processed content to Amazon Bedrock.
- G. Create an Amazon Cognito user pool that uses pre-authentication AWS Lambda functions to run content moderation checks
- H. Use Amazon Textract to filter text content and Amazon Rekognition to filter image content before allowing users to upload content to the platform.
- I. Create an AWS Step Functions workflow that uses built-in Amazon Bedrock guardrails to filter content
- J. Use Amazon Comprehend PII detection to pre-process the content
- K. Use Amazon Rekognition image moderation.

Answer: D

NEW QUESTION 49

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% overlap
- 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 tokens
- D. Use Amazon Bedrock Knowledge Bases built-in retrieval to automatically select relevant parent chunks based on query context
- 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 sentences
- 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 tokenizer
- I. Configure the Lambda function to split documents into equal segments that fit within 80% of the context window
- J. Configure the Lambda function to process each segment independently before aggregating the results.

Answer: C

NEW QUESTION 54

A company is using AWS Lambda and REST APIs to build a reasoning agent to automate support workflows. The system must preserve memory across interactions, share relevant agent state, and support event-driven invocation and synchronous invocation. The system must also enforce access control and session-based permissions.

Which combination of steps provides the MOST scalable solution? (Select TWO.)

- A. Use Amazon Bedrock AgentCore to manage memory and session-aware reasoning
- B. Deploy the agent with built-in identity support, event handling, and observability.
- C. Register the Lambda functions and REST APIs as actions by using Amazon API Gateway and Amazon EventBridge
- D. Enable Amazon Bedrock AgentCore to invoke the Lambda functions and REST APIs without custom orchestration code.
- E. Use Amazon Bedrock Agents for reasoning and conversation management
- F. Use AWS Step Functions and Amazon SQS for orchestration
- G. Store agent state in Amazon DynamoDB.
- H. Deploy the reasoning logic as a container on Amazon ECS behind API Gateway
- I. Use Amazon Aurora to store memory and identity data.
- J. Build a custom RAG pipeline by using Amazon Kendra and Amazon Bedrock
- K. Use AWS Lambda to orchestrate tool invocation

L. Store agent state in Amazon S3.

Answer: AB

NEW QUESTION 58

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 61

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 a PostgreSQL database.

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, feature, 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 the 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: D

NEW QUESTION 64

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 68

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

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

100% Actual & Verified — Instant Download, Please Click
[Order The AIP-C01 Practice Test Here](#)