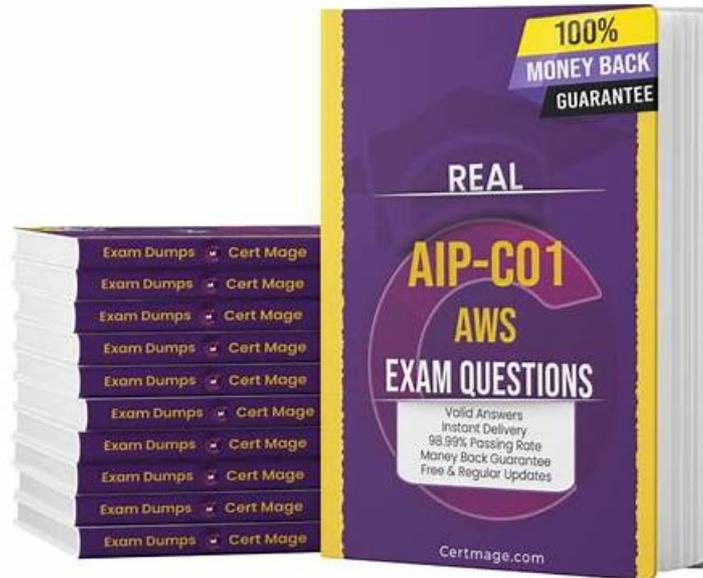


# New Release AIP-C01 Questions - Amazon AIP-C01 Exam Dumps



Our delivery speed is also highly praised by customers. Our AIP-C01 exam dumps won't let you wait for such a long time. As long as you pay at our platform, we will deliver the relevant AIP-C01 test prep to your mailbox within 5-10 minutes. Our company attaches great importance to overall services, if there is any problem about the delivery of AIP-C01 Test Braindumps, please let us know, a message or an email will be available. We are pleased that you can spare some time to have a look for your reference about our AIP-C01 test prep.

Our AWS Certified Generative AI Developer - Professional study question has high quality. So there is all effective and central practice for you to prepare for your test. With our professional ability, we can accord to the necessary testing points to edit AIP-C01 exam questions. It points to the exam heart to solve your difficulty. So high quality materials can help you to pass your exam effectively, make you feel easy, to achieve your goal. With the AIP-C01 Test Guide use feedback, it has 98%-100% pass rate. That's the truth from our customers. And it is easy to use for you only with 20 hours' to 30 hours' practice. After using the AIP-C01 test guide, you will have the almost 100% assurance to take part in an examination. With high quality materials and practices, you will get easier to pass the exam.

>> AIP-C01 Study Material <<

## Reliable AIP-C01 Study Material & Leading Offer in Qualification Exams & Fast Download AIP-C01: AWS Certified Generative AI Developer - Professional

VCE4Plus Amazon AIP-C01 exam questions are compiled according to the latest syllabus and the actual AIP-C01 certification exam. We are also constantly upgrade our training materials so that you could get the best and the latest information for the first time. When you buy our AIP-C01 Exam Training materials, you will get a year of free updates. At any time, you can extend the update subscription time, so that you can have a longer time to prepare for the exam.

### Amazon AIP-C01 Exam Syllabus Topics:

Topic	Details
-------	---------

Topic 1	<ul style="list-style-type: none"> <li>• <b>Testing, Validation, and Troubleshooting:</b> This domain covers evaluating foundation model outputs, implementing quality assurance processes, and troubleshooting GenAI-specific issues including prompts, integrations, and retrieval systems.</li> </ul>
Topic 2	<ul style="list-style-type: none"> <li>• <b>Foundation Model Integration, Data Management, and Compliance:</b> This domain covers designing GenAI architectures, selecting and configuring foundation models, building data pipelines and vector stores, implementing retrieval mechanisms, and establishing prompt engineering governance.</li> </ul>
Topic 3	<ul style="list-style-type: none"> <li>• <b>AI Safety, Security, and Governance:</b> This domain addresses input</li> <li>• output safety controls, data security and privacy protections, compliance mechanisms, and responsible AI principles including transparency and fairness.</li> </ul>
Topic 4	<ul style="list-style-type: none"> <li>• <b>Implementation and Integration:</b> This domain focuses on building agentic AI systems, deploying foundation models, integrating GenAI with enterprise systems, implementing FM APIs, and developing applications using AWS tools.</li> </ul>
Topic 5	<ul style="list-style-type: none"> <li>• <b>Operational Efficiency and Optimization for GenAI Applications:</b> This domain encompasses cost optimization strategies, performance tuning for latency and throughput, and implementing comprehensive monitoring systems for GenAI applications.</li> </ul>

## Amazon AWS Certified Generative AI Developer - Professional Sample Questions (Q35-Q40):

### NEW QUESTION # 35

A book publishing company wants to build a book recommendation system that uses an AI assistant. The AI assistant will use ML to generate a list of recommended books from the company's book catalog. The system must suggest books based on conversations with customers.

The company stores the text of the books, customers' and editors' reviews of the books, and extracted book metadata in Amazon S3. The system must support low-latency responses and scale efficiently to handle more than 10,000 concurrent users.

Which solution will meet these requirements?

- A. Use Amazon Bedrock Knowledge Bases to generate embeddings. Store the embeddings as a vector store in Amazon DynamoDB. Create an AWS Lambda function that queries the knowledge base. Configure Amazon API Gateway to invoke the Lambda function when handling user requests.
- **B. Use Amazon Bedrock Knowledge Bases to generate embeddings. Store the embeddings as a vector store in Amazon OpenSearch Service. Create an AWS Lambda function that queries the knowledge base. Configure Amazon API Gateway to invoke the Lambda function when handling user requests.**
- C. Use Amazon SageMaker AI to deploy a pre-trained model to build a personalized recommendation engine for books. Deploy the model as a SageMaker AI endpoint. Invoke the model endpoint by using Amazon API Gateway.
- D. Create an Amazon Kendra GenAI Enterprise Edition index that uses the S3 connector to index the book catalog data stored in Amazon S3. Configure built-in FAQ in the Kendra index. Develop an AWS Lambda function that queries the Kendra index based on user conversations. Deploy Amazon API Gateway to expose this functionality and invoke the Lambda function.

**Answer: B**

Explanation:

Option A best meets the requirements because it directly implements a Retrieval Augmented Generation pattern for conversational recommendations using managed Amazon Bedrock capabilities and a scalable vector store. The company's source data already resides in Amazon S3, which aligns naturally with Amazon Bedrock Knowledge Bases ingestion workflows. A knowledge base can ingest book text, reviews, and metadata, generate embeddings using a supported embedding model, and persist those vectors in a purpose-built vector backend such as Amazon OpenSearch Service. This enables semantic retrieval that is well suited to conversation-driven intent, where user prompts are often descriptive and do not map cleanly to keyword filters.

The requirement to suggest books based on conversations implies the system must interpret natural language context and retrieve relevant passages, reviews, and metadata to ground the recommendation. Knowledge Bases provide managed orchestration for embedding creation and retrieval, which reduces development effort compared to building custom embedding pipelines. OpenSearch Service provides scalable vector search and k-nearest neighbors style similarity retrieval, which supports low-latency responses when properly indexed and sized.

For scaling to more than 10,000 concurrent users, the API layer design in option A is a common AWS pattern: Amazon API Gateway provides a managed front door with throttling and request handling, while AWS Lambda scales horizontally with demand and can invoke the knowledge base retrieval operations. This separates compute scaling from the vector store scaling and helps keep latency predictable under load.

Option B is not the best choice because DynamoDB is not the standard native vector store target for Amazon Bedrock Knowledge Bases in this context and would introduce additional implementation complexity around vector indexing and similarity search behavior. Option C requires substantial ML lifecycle work, model hosting, tuning, and continuous iteration to achieve quality recommendations at scale. Option D provides strong enterprise search, but it focuses on retrieval and FAQs rather than a managed RAG recommendation workflow grounded in embeddings and conversational context for generative responses.

### NEW QUESTION # 36

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 document. Create an Amazon Bedrock knowledge base from the document. Create an Amazon Bedrock evaluation job that uses the retrieval and response generation evaluation type. Specify an Amazon S3 bucket as the output. Run an evaluation job for each FM.
- B. Combine the sample prompts into a single JSONL document. Store the document in an Amazon S3 bucket. Create an Amazon Bedrock evaluation job that uses a judge model. Specify the S3 location as input and Amazon QuickSight as output. Run an evaluation job for each FM and select the FM as the evaluator.
- C. Combine the sample prompts into a single JSONL document. Store the document in an Amazon S3 bucket. Create an Amazon Bedrock evaluation job that uses a judge model. Specify the S3 location as input and a different S3 location as output. Run an evaluation job for each FM and select the FM as the generator.
- D. Combine the sample prompts into a single JSON document. Create an Amazon Bedrock knowledge base with the document. Write a prompt that asks the FM to generate a response to each sample prompt. Use the RetrieveAndGenerate API to generate a report for each model.

**Answer: C**

Explanation:

Option B is correct because it uses the managed evaluation capability in Amazon Bedrock that is intended specifically for comparing foundation models using a consistent prompt set and producing structured results with minimal custom tooling. In a Bedrock evaluation workflow, you provide an input dataset of prompts, typically in JSON Lines format so each line represents one evaluation record. Storing the JSONL file in Amazon S3 allows Bedrock to read the dataset at scale and write standardized evaluation outputs back to S3 for downstream analysis, sharing, and retention.

The key requirement is to assess both quality and safety across multiple models. A Bedrock evaluation job can use a judge model to score the generated outputs against defined criteria. This approach supports repeatable, apples-to-apples comparisons because the same judge model and scoring rubric can be applied to every candidate foundation model. The candidate models are configured as generators, meaning each evaluation job run uses one selected FM to produce answers for the same prompt set, and the judge model evaluates those answers. That matches the requirement to generate evaluation reports that help a data scientist select the best FM.

Option A does not use Bedrock evaluation jobs, and a knowledge base plus RetrieveAndGenerate is a RAG pattern, not an evaluation framework. It would produce responses but not standardized scoring and reporting suitable for model selection. Option C is incorrect because Bedrock evaluation outputs are delivered to S3, not directly to a BI destination, and selecting the candidate FM as the evaluator conflicts with the intended pattern of using a stable judge model. Option D misuses knowledge bases and retrieval evaluation types when the requirement is prompt-based model assessment rather than evaluating retrieval quality.

### NEW QUESTION # 37

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 with partial results enabled to deliver fragments of transcribed text before customers finish speaking. Forward text fragments to Amazon Bedrock by using the InvokeModelWithResponseStream API. Stream responses to call center agents through an Amazon API Gateway WebSocket API.
- B. Use the Amazon Transcribe streaming API with an AWS Lambda function to transcribe each audio segment. Call the Amazon Titan Embeddings model on Amazon Bedrock by using the InvokeModel API. Publish results to Amazon SNS.
- C. Use Amazon Transcribe batch processing to convert calls to text. Pass complete transcripts to Anthropic Claude on Amazon Bedrock by using the ConverseStream API. Return responses through an Amazon Lex chatbot interface.
- D. Use Amazon Transcribe streaming to transcribe calls. Pass the text to Amazon Comprehend for sentiment analysis. Feed the results to Anthropic Claude on Amazon Bedrock by using the InvokeModel API. Store results in Amazon DynamoDB. Use a WebSocket API to display the results.

**Answer: A**

Explanation:

Option B is the only solution that satisfies all strict real-time, streaming, and latency requirements. Amazon Transcribe streaming with partial results allows transcription fragments to be delivered before the speaker finishes a sentence. This significantly reduces perceived latency and enables downstream processing to begin immediately, which is essential for maintaining sub-1-second end-to-end response times.

Using Amazon Bedrock's InvokeModelWithResponseStream API enables token-level or chunk-level streaming responses from the foundation model. This allows the GenAI assistant to begin delivering suggestions to call center agents incrementally instead of waiting for a full model response. This streaming inference capability is critical for interactive, real-time agent assistance use cases. Amazon API Gateway WebSocket APIs provide fully managed, bidirectional communication between backend services and agent dashboards. This ensures that updates flow continuously to agents as new transcription fragments and model outputs become available, preserving real-time responsiveness without requiring custom socket infrastructure.

Option A introduces additional synchronous processing layers and storage writes that increase latency. Option C uses batch transcription and post-call processing, which cannot meet real-time requirements. Option D uses embeddings and asynchronous messaging, which are not suitable for live incremental suggestions and bidirectional streaming.

Therefore, Option B best aligns with AWS real-time GenAI architecture patterns by combining streaming transcription, streaming model inference, and managed bidirectional communication while maintaining low latency and operational simplicity.

### NEW QUESTION # 38

A specialty coffee company has a mobile app that generates personalized coffee roast profiles by using Amazon Bedrock with a three-stage prompt chain. The prompt chain converts user inputs into structured metadata, retrieves relevant logs for coffee roasts, and generates a personalized roast recommendation for each customer.

Users in multiple AWS Regions report inconsistent roast recommendations for identical inputs, slow inference during the retrieval step, and unsafe recommendations such as brewing at excessively high temperatures. The company must improve the stability of outputs for repeated inputs. The company must also improve app performance and the safety of the app's outputs. The updated solution must ensure 99.5% output consistency for identical inputs and achieve inference latency of less than 1 second. The solution must also block unsafe or hallucinated recommendations by using validated safety controls.

Which solution will meet these requirements?

- A. Deploy Amazon Bedrock with provisioned throughput to stabilize inference latency. Apply Amazon Bedrock guardrails with semantic denial rules to block unsafe outputs. Use Amazon Bedrock Prompt Management to manage prompts by using approval workflows.
- B. Use Amazon Kendra to improve roast log retrieval accuracy. Store normalized prompt metadata within Amazon DynamoDB. Use AWS Step Functions to orchestrate multi-step prompts.
- C. Use Amazon Bedrock Agents to manage chaining. Log model inputs and outputs to Amazon CloudWatch Logs. Use logs from CloudWatch to perform A/B testing for prompt versions.
- D. Cache prompt results in Amazon ElastiCache. Use AWS Lambda functions to pre-process metadata and to trace end-to-end latency. Use AWS X-Ray to identify and remediate performance bottlenecks.

**Answer: A**

Explanation:

Option A is the only choice that simultaneously addresses all three requirements: (1) higher output consistency for identical inputs, (2) sub-1-second performance, and (3) validated safety controls that block unsafe or hallucinated recommendations.

Provisioned throughput in Amazon Bedrock reserves capacity for the chosen model, which helps stabilize latency and reduces the chance of throttling or variable response times across Regions. This is important for a mobile app with strict latency goals and users distributed across multiple Regions. While provisioned throughput primarily improves performance predictability, it also reduces variability caused by contention during peak demand.

Amazon Bedrock guardrails provide validated safety controls to filter or block unsafe content. Semantic denial rules are appropriate for preventing dangerous brewing guidance (for example, excessively high temperatures) and for reducing hallucinated instructions that violate safety policies. Guardrails can be enforced consistently regardless of prompt-chain complexity, providing a uniform safety layer around the model outputs.

Amazon Bedrock Prompt Management supports controlled prompt versioning and approval workflows. By standardizing prompts, controlling changes, and ensuring the same prompt version is used for identical inputs, the company improves output stability and reduces drift caused by unmanaged prompt edits. Combined with strict configuration control (including fixed inference parameters such as temperature where appropriate), this improves repeatability and increases the likelihood of achieving the 99.5% consistency target.

Option B improves observability and experimentation but does not provide strong safety enforcement or latency stabilization. Option C improves performance through caching and tracing but does not provide validated safety controls and does not directly address cross-Region output consistency. Option D may improve retrieval but does not enforce safety controls or ensure repeatable outputs. Therefore, Option A best meets the stability, performance, and safety requirements using AWS-native controls.

### NEW QUESTION # 39

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. Set up an S3 Lifecycle configuration to remove reports that are older than 3 years. Schedule an AWS Lambda function to run daily syncs between the bucket and the knowledge base. 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.
- B. Create a second knowledge base. Use Lambda and Amazon Comprehend to redact PII before syncing to the second knowledge base. Route users to the appropriate knowledge base based on Cognito group membership.
- C. Invoke an AWS Lambda function to sync the S3 bucket and the knowledge base when a new report is uploaded. Use a second Lambda function with Amazon Comprehend to redact PII for engineers. Use S3 Lifecycle rules to remove reports older than 3 years.
- D. Enable Amazon Macie PII detection on the S3 bucket. Use an S3 trigger to invoke an AWS Lambda function that redacts PII from the reports. Configure the Lambda function to delete outdated documents and invoke knowledge base syncing.

**Answer: A**

Explanation:

Option C is the correct solution because it enforces privacy controls at inference time, not at ingestion time, which is required when different user roles require different visibility into the same underlying data.

Using an S3 Lifecycle configuration ensures that documents older than 3 years are automatically removed, guaranteeing that the knowledge base references only compliant, recent medical reports. Scheduling Lambda-based syncs keeps the knowledge base aligned with the bucket contents without introducing complex per-upload orchestration.

The most important requirement is role-based PII exposure. Amazon Bedrock guardrails support dynamic application at inference time, allowing the system to select a guardrail configuration based on the authenticated user's Amazon Cognito group. Surgeons can receive full responses, while engineers receive responses with PII masked-without duplicating data or maintaining multiple knowledge bases.

This approach preserves a single source of truth for medical reports while enforcing privacy through response-level controls. It also maintains full auditability of access and redaction behavior.

Option A permanently removes PII and violates surgeon access requirements. Option B redacts data inconsistently and couples privacy logic to ingestion. Option D doubles storage, increases cost, and introduces data drift risk.

Therefore, Option C best meets privacy, compliance, scalability, and operational efficiency requirements.

### NEW QUESTION # 40

.....

So rest assured that with the VCE4Plus AIP-C01 exam questions you will get everything that is necessary for AIP-C01 exam preparation and success. Take a decision right now and just get registered in the Amazon AIP-C01 certification exam and start preparation with VCE4Plus AIP-C01 Exam Questions. You do not need to get worried about it choose the right VCE4Plus AWS Certified Generative AI Developer - Professional exam questions formats and start this journey without wasting further time.

