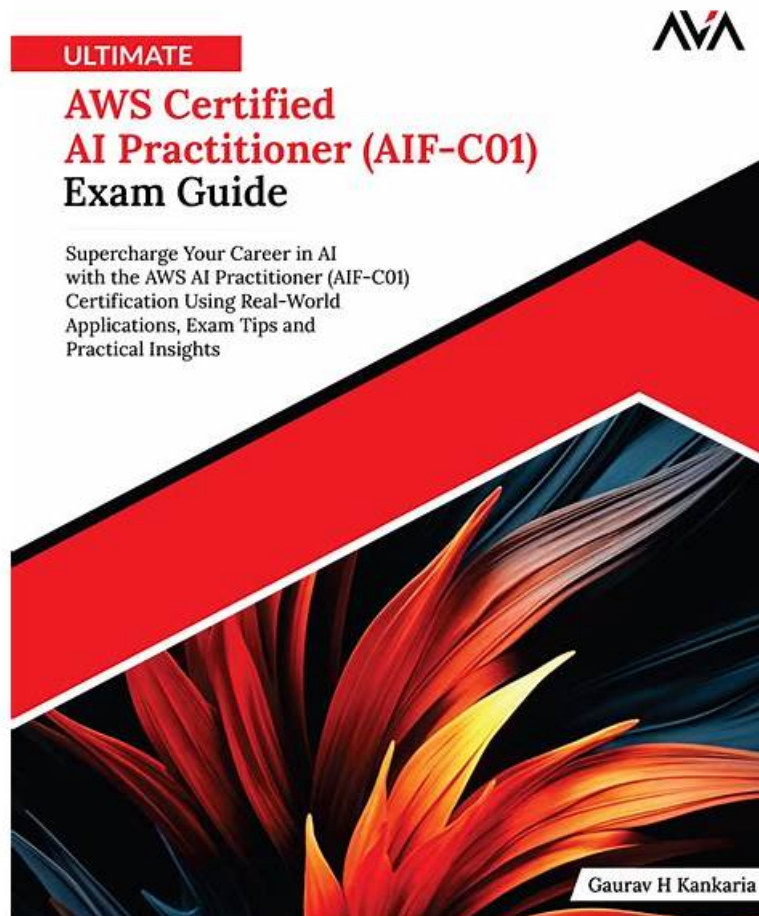


# AIP-C01前提条件、AIP-C01無料試験



ちなみに、MogiExam AIP-C01の一部をクラウドストレージからダウンロードできます：<https://drive.google.com/open?id=1DOliBOZWlwYXc-QjLVVxcMm4RwpbAPZS>

これらの2つの特性により、AIP-C01ガイドトレントを使用するほぼすべての候補者が一度にテストに合格できることがわかります。これは自己決定ではありません。統計によると、当社のAIP-C01ガイドトレントは98%~99%の高い合格率を達成しており、これは他のすべてをかなり上回る程度です。同時に、AIP-C01テストトレントが毎日更新されるかどうかを確認する専門スタッフがいます。メールでお問い合わせいただく場合でも、オンラインでお問い合わせいただく場合でも、できるだけ早く問題を解決できるようサポートいたします。心配する必要はまったくありません。

## Amazon AIP-C01 認定試験の出題範囲：

トピック	出題範囲
トピック 1	<ul style="list-style-type: none"><li>テスト、検証、トラブルシューティング: このドメインでは、基礎モデル出力の評価、品質保証プロセスの実装、プロンプト、統合、検索システムなどの GenAI 固有の問題のトラブルシューティングについて説明します。</li></ul>
トピック 2	<ul style="list-style-type: none"><li>AI の安全性、セキュリティ、ガバナンス: このドメインでは、入出力の安全性管理、データセキュリティとプライバシーの保護、コンプライアンス メカニズム、透明性と公平性を含む責任ある AI の原則について説明します。</li></ul>
トピック 3	<ul style="list-style-type: none"><li>基盤モデルの統合、データ管理、コンプライアンス: このドメインでは、GenAI アーキテクチャの設計、基盤モデルの選択と構成、データ パイプラインとベクトル ストアの構築、取得メカニズムの実装、迅速なエンジニアリング ガバナンスの確立について説明します。</li></ul>

トピック 4	<ul style="list-style-type: none"> <li>実装と統合: このドメインは、エージェント AI システムの構築、基盤モデルのデプロイ、GenAI とエンタープライズシステムの統合、FM API の実装、AWS ツールを使用したアプリケーションの開発に重点を置いています。</li> </ul>
トピック 5	<ul style="list-style-type: none"> <li>GenAI アプリケーションの運用効率と最適化: この領域には、コスト最適化戦略、レイテンシとスループットのパフォーマンスチューニング、GenAI アプリケーション向けの包括的な監視システムの実装が含まれます。</li> </ul>

>> AIP-C01前提条件 <<

## AIP-C01 絶対合格の教科書+出る順問題集

AIP-C01の実際のテストのオンラインバージョンを使用すると非常に便利です。オンライン版の利便性を実感すれば、多くの問題の解決に役立ちます。一方で、オンライン版は機器に限定されません。AIP-C01テスト準備のオンラインバージョンは、電話、コンピューターなどを含むすべての電子機器に適用されます。一方、AIP-C01学習教材のオンライン版を使用することに決めた場合、WLANネットワークがないことを心配する必要はありません。

## Amazon AWS Certified Generative AI Developer - Professional 認定 AIP-C01 試験問題 (Q98-Q103):

### 質問 # 98

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. 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. Store conversation history in Amazon S3 and configure server-side encryption. Save each interaction as a separate JSON file.
- B. Use an AWS Step Functions Standard workflow to orchestrate clarification workflows. Include Wait for a Callback patterns to manage the workflows. Store conversation history in Amazon DynamoDB. Purchase on-demand capacity and configure server-side encryption.
- C. Use an AWS Step Functions Express workflow to orchestrate conversation flow. Invoke AWS Lambda functions to run clarification logic. Store conversation history in Amazon RDS and use session IDs as the primary key.
- D. Use AWS Lambda functions to call Amazon Bedrock inference APIs. Use Amazon SQS queues to orchestrate clarification steps. Store conversation history in an Amazon ElastiCache (Redis OSS) cluster. Configure encryption at rest.

正解: B

解説:

Option B is the correct solution because it provides a scalable, durable, and secure architecture for conversational GenAI workloads that require multi-step clarification workflows and persistent memory.

AWS Step Functions Standard workflows are designed for long-running, stateful workflows with high reliability, which is ideal for clarification loops that may require multiple back-and-forth interactions. The Wait for a Callback pattern allows the workflow to pause while awaiting additional user input, making it well-suited for handling ambiguous queries without losing execution state. Storing conversation history in Amazon DynamoDB enables millisecond-latency reads and writes at massive scale, supporting thousands of concurrent users. DynamoDB's on-demand capacity mode automatically scales with traffic, eliminating capacity planning. Server-side encryption ensures that stored conversation data is encrypted at rest, meeting security and compliance requirements for personalized data.

Option A uses Step Functions Express and Amazon RDS, which is not ideal for long-lived conversational workflows and introduces scaling and connection management challenges. Option C stores conversations as individual S3 objects, which increases latency and complicates context retrieval. Option D relies on Amazon ElastiCache, which is optimized for ephemeral caching rather than durable, auditable conversation history.

Therefore, Option B best balances scalability, performance, durability, and security for a conversational Amazon Bedrock-based

customer support application.

### 質問 # 99

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

正解: D

解説:

Option B is the correct solution because it directly addresses both throughput bottlenecks and latency requirements using native Amazon Bedrock performance optimization features that are designed for real-time, high-volume generative AI workloads.

Amazon Bedrock supports cross-Region inference profiles, which allow applications to transparently route inference requests across multiple AWS Regions. During peak usage periods, traffic is automatically distributed to Regions with available capacity, reducing throttling, request queuing, and timeout risks. This approach aligns with AWS guidance for building highly available, low-latency GenAI applications that must scale elastically across geographic boundaries.

Token batching further improves efficiency by combining multiple inference requests into a single model invocation where applicable. AWS Generative AI documentation highlights batching as a key optimization technique to reduce per-request overhead, improve throughput, and better utilize model capacity. This is especially effective for lightweight, low-latency models such as Claude 3 Haiku, which are designed for fast responses and high request volumes.

Option A does not meet the requirement because purchasing provisioned throughput in a single Region creates a regional bottleneck and does not address multi-Region availability or traffic spikes beyond reserved capacity. Retries increase load and latency rather than resolving the root cause.

Option C improves application-layer scaling but does not solve model-side throughput limits. Client-side round-robin routing lacks awareness of real-time model capacity and can still send traffic to saturated Regions.

Option D is unsuitable because batch inference with asynchronous retrieval is designed for offline or non-interactive workloads. It cannot meet a strict 2-second response time requirement for an interactive AI assistant.

Therefore, Option B provides the most effective and AWS-aligned solution to achieve low latency, global scalability, and high throughput during peak usage periods.

### 質問 # 100

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. Add exponential backoff retry logic to handle throttling exceptions during peak hours.
- C. Modify the application to use the `invokeModelWithResponseStream` API instead of the `invokeModel` API.

- D. Replace the model ID parameter with the ARN of the provisioned model that the CreateProvisionedModelThroughput API returns.

正解: D

解説:

Option B is the correct solution because Amazon Bedrock provisioned throughput is only used when the application explicitly invokes the provisioned model ARN, not the base foundation model ID. In the provided code, the application is calling the standard model identifier (anthropic.claude-v2), which routes requests to on-demand capacity instead of the purchased provisioned throughput.

When the CreateProvisionedModelThroughput API is used, Amazon Bedrock returns a provisioned model ARN that represents the reserved capacity. Applications must reference this ARN in the modelId parameter when invoking the model. If the base model ID is used instead, Bedrock treats the request as on-demand traffic, which explains why CloudWatch metrics show unused provisioned capacity alongside throttled on-demand requests.

Option A would increase capacity but would not fix the root cause because the application is not using the provisioned resource at all. Option C adds resiliency but does not ensure usage of provisioned throughput and would still incur throttling. Option D changes the response delivery mechanism but does not affect capacity routing.

Therefore, Option B directly resolves the throttling issue by correctly routing traffic to the reserved capacity and ensures that the company benefits from the provisioned throughput it has purchased.

### 質問 # 101

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

正解: B

解説:

Option D is the correct solution because it directly evaluates multilingual output consistency and quality in an automated, scalable, and deployment-gating workflow. Amazon Bedrock model evaluation jobs are designed to run large-scale, repeatable evaluations against defined datasets and to produce quantitative metrics that can be used as objective release criteria.

The core issue is semantic inconsistency across languages for equivalent inputs. The most reliable way to detect this is to create standardized test conversations where each language version expresses the same intent and constraints. Running those tests through the updated model and comparing results with similarity metrics (for example, semantic similarity between expected and actual answers, or between language variants) surfaces regressions that infrastructure testing cannot detect.

Bedrock evaluation jobs support running evaluations at scale and are well suited for processing large datasets quickly. By parallelizing evaluation runs across languages and conversations, the company can meet the 45-minute requirement while executing at least 15,000 conversations. Because the process is standardized, it also allows consistent baseline comparisons across releases. Applying hallucination thresholds ensures that answers remain grounded and do not introduce fabricated details, which is particularly important when language-specific behavior shifts after a model upgrade.

Integrating evaluation jobs into the CI/CD pipeline enables fully automated execution on every model or configuration update. The pipeline can enforce a hard quality gate that blocks deployment if thresholds are not met, preventing regressions from reaching



production.

Option A focuses on performance and infrastructure bottlenecks, not multilingual response quality. Option B is post-deployment and too slow to prevent regressions. Option C normalizes inputs but does not measure multilingual output equivalence or provide robust, quantitative gating.

Therefore, Option D best meets the automation, scale, timing, and deployment-blocking requirements.

### 質問 # 102

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. Analyze multi-hop relationships between entities and automatically identify related information across documents.
- B. Use Amazon DynamoDB to store financial data in a custom indexing system. Use AWS Lambda to query relevant records. Use Amazon SageMaker to generate responses.
- 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.

正解: A

解説:

Option A best satisfies the requirement to capture multi-hop, highly interconnected relationships with minimal operational overhead. Traditional vector similarity search excels at finding semantically similar text but is not optimized for reasoning over explicit entity-to-entity relationships, especially when analysts need indirect, multi-hop connections (for example, fund # holding # issuer # sector # regulation). Graph-based retrieval is designed specifically for these kinds of relationship traversals.

GraphRAG combines retrieval-augmented generation with graph-aware context selection. By representing entities and their relationships in a graph store, the system can traverse multiple hops to assemble a holistic set of relevant facts. This improves completeness and reduces the chance that the model misses indirect relationships that are essential for accurate investment guidance. Amazon Neptune Analytics provides a managed graph analytics environment capable of efficiently traversing and analyzing complex relationship networks. When integrated with Amazon Bedrock Knowledge Bases, it reduces custom engineering by providing managed ingestion, retrieval, and orchestration patterns suitable for GenAI applications. This lowers operational overhead compared to building and maintaining custom multi-stage retrieval logic.

Meeting the sub-3-second requirement is also more feasible with a graph-optimized engine because multi-hop traversals can be executed efficiently compared to chaining multiple vector searches and joining results in an application layer. The managed nature of Knowledge Bases and Neptune Analytics reduces maintenance, scaling, and operational burden while enabling strong performance. Option B and C require extensive custom logic and orchestration, increasing complexity and latency. Option D is not designed for graph-style multi-hop exploration and would require significant custom indexing and retrieval logic.

Therefore, Option A is the most AWS-aligned and operationally efficient approach for multi-hop relationship-aware RAG with strong performance.

### 質問 # 103

.....

現在の仕事と現在の生活に飽きていますか？ 便利な証明書を入手してください！ AIP-C01学習ガイドは、目標を達成するのに役立つ最高の製品です。試験に合格し、AIP-C01学習教材で認定を取得すると、大企業で満足のいく仕事に応募し、高い給与と高い利益で上級職に就くことができます。優れたAmazon AIP-C01スタディガイドにより、受験者は、余分な時間とエネルギーを無駄にせずに効率的にテストを準備するための明確な学習方向を得ることができます。

AIP-C01無料試験: <https://www.mogixexam.com/AIP-C01-exam.html>

- AIP-C01テスト参考書 □ AIP-C01日本語解説集 □ AIP-C01学習体験談 📖 ウェブサイト【[www.mogixexam.com](https://www.mogixexam.com)】から▷ AIP-C01 ◁を開いて検索し、無料でダウンロードしてくださいAIP-C01模擬対策問題

- BONUS!!! MogiExam AIP-C01ダンプの一部を無料でダウンロード: <https://drive.google.com/open?id=1DOIiBOZWlwYXc-OjLVVxcMm4RwpbAPZS>