Microsoft.AI-103.v2026-06-06.q24

試験コード：	AI-103
試験名称：	Developing AI Apps and Agents on Azure
認証ベンダー：	Microsoft
無料問題の数：	24
バージョン：	v2026-06-06
ページの閲覧量：	104
問題集の閲覧量：	245

https://www.jpnshiken.com/shiken/Microsoft.AI-103.v2026-06-06.q24.html

質問: **1**
あなたは、自然な音声対話をサポートするために、Microsoft Foundry プロジェクトでエージェント ワークフローを作成しています。
エージェントは、連続的な音声入力を受け取り、推論のためにその入力をテキストに変換し、その後、音声による応答をユーザーに返す必要があります。ワークフローは以下の要件を満たす必要があります。
- ユーザーが話し終える前にエージェントが音声出力の生成を開始する、ターンテーキングのダイナミクスをサポートする。

会話のような自然な体験を維持するために、低遅延で動作します。
リアルタイムのエージェントとのやり取りでは、音声認識と音声合成の両方を有効にする必要があります。
あなたはどうすべきでしょうか？
A. 埋め込みモデルを使用して音声をエンコードし、その後、音声をテキストと音声にデコードします。
B. バッチ文字起こしを使用して音声入力を変換し、エージェントからテキスト応答を返します。
C. 音声翻訳を使用して音声を別の言語に変換し、翻訳されたテキストを返します。
D. 受信音声にはリアルタイム音声認識を使用し、エージェントの応答にはテキスト読み上げを使用します。
正解: (正解を表示します)
正解は D です。受信音声にはリアルタイム音声テキスト変換を、エージェントの応答にはテキスト音声変換を使用します。ワークフローでは、継続的な音声入力、推論のための低遅延の文字起こし、およびユーザーへの音声出力が必要です。Foundry Tools の Azure Speech のリアルタイム音声テキスト変換は、ストリーミング音声からの即時文字起こし用に設計されており、インタラクションの受信音声側を満たします。テキスト音声変換は、エージェントが応答を生成した後、音声による応答の送信パスを提供します。
このパターンは、Microsoft のリアルタイム音声エージェント アーキテクチャと一致しています。Voice Live API の概要では、低遅延音声合成システムが音声認識、生成推論、テキスト読み上げ機能を統合して自然な音声体験を実現すると説明しています。また、コンタクト センターを重要なシナリオとして挙げ、エンド ユーザーにとって知覚される遅延が低いことを強調しています。埋め込みでは、音声を会話音声にデコードしません。バッチ文字起こしではファイル指向の遅延が発生するため、ターンテーキングには適していません。
音声翻訳は言語間の翻訳にのみ適しており、必要な推論と音声応答のループは提供しません。参考トピック :Foundry Tools の Azure Speech、リアルタイム音声テキスト変換、テキスト音声変換、音声エージェント、低遅延インタラクション、会話のターンテーキング。

質問: **2**
エージェントを含む Microsoft Foundry プロジェクトがあります。このエージェントは、Foundry Tools の Azure Speech を使用します。
en-usロケール向けにベースラインの音声認識モデルを微調整し、そのモデルを公開します。
エージェントは音声認識REST APIを呼び出し、プロジェクトIDが無効であることを示すエラーメッセージを返します。
プロジェクトプロパティを正しいIDに設定する必要があります。
プロジェクトプロパティはどのような値に設定すべきですか？
A. カスタム音声エンドポイントURL

B. プロジェクトのURL
C. プロジェクトID
D. カスタム音声プロジェクトID
正解: (正解を表示します)
正解はDです。カスタム音声プロジェクトIDです。カスタム音声の微調整には、音声認識REST APIでプロジェクトプロパティを使用しますが、これは一般的なMicrosoft Foundryプロジェクトではなく、カスタム音声プロジェクトを参照する必要があります。Microsoftのカスタム音声に関するガイダンスでは、カスタム音声に音声認識REST APIを使用する場合は、プロジェクトプロパティをカスタム音声プロジェクトのIDに設定する必要があると記載されています。また、カスタム音声プロジェクトIDはMicrosoft FoundryプロジェクトIDとは異なることも明記されています。
この違いが、無効なプロジェクト ID エラーの原因です。Foundry プロジェクト ID、プロジェクト URL、またはエンドポイント URL を指定しても、微調整された音声モデルを所有するカスタム音声プロジェクトは識別されません。カスタム音声エンドポイント URL は、認識のためにデプロイされたカスタムモデルエンドポイントを呼び出す際に使用されますが、REST API プロジェクト プロパティの値ではありません。また、API は識別子の値を要求するため、プロジェクト URL も受け入れられません。参照トピック: Foundry ツールの Azure Speech、カスタム音声の微調整、音声認識 REST API、カスタム音声プロジェクト ID、モデルの公開、エンドポイント構成。

質問: **3**
エージェントを含む Microsoft Foundry プロジェクトがあります。このエージェントには、Azure AI Search に格納されているナレッジベースを照会する Model Context Protocol (MCP) ツールが含まれています。
エージェントの実行によっては、知識ベースを呼び出さずに基本モデルから回答を返す場合があり、その結果、根拠となる引用のない応答が返される。
エージェントを実行するためのコードスニペットを以下に示します。
run = project_client.agents.runs.create_and_process(
thread_id=thread.id、
agent_id=agent.id、
)
エージェントが実行ごとに必ずMCPツールを呼び出すようにするには、コードに適切なtool_choiceパラメータを追加する必要があります。
何を追加すべきでしょうか？
A. tool_choice ={ " type " : " mcp " }
B. tool_choice={ " auto " }
C. tool_choice={ " type " : " knowledge_base " }
D. tool_choice={ "必須" }
正解: (正解を表示します)
正解はDです。Microsoft Foundry Agent Serviceでは、tool_choiceは、モデルが直接応答するか、ツールを呼び出す必要があるかを制御するために使用されるランタイム制御です。Microsoftのツールに関するベストプラクティスガイダンスでは、autoはモデルがツールを呼び出すかどうかを決定できるようにし、noneはツールの呼び出しを防止し、requiredはモデルが1つ以上のツールを呼び出す必要があることを意味します。これは、一部の実行で基本モデルから応答し、知識ベースをスキップするという問題に直接対処するものです。
Microsoftのチュートリアルによると、MCPツールを介してAzure AI Searchを基盤とするエージェント型検索ソリューションの場合、tool_choice= "required"を設定することで、クエリ処理時にエージェントが常にナレッジベースツールを使用するようになります。これにより、応答前にツール呼び出しが強制されるため、根拠に基づいた回答が得られます。
auto は、既に引用漏れの原因となっている非決定的な動作を維持してしまうため、不適切です。
`{ " type " : " knowledge_base " }` は有効な Foundry ツール選択タイプではありません。`{ " type " : " mcp " }` は一部の Responses API スキーマで MCP ツールタイプを記述していますが、このエージェント実行シナリオの決定論的な保証は、必要なツール呼び出しモードです。参照トピック: Microsoft Foundry Agent Service、MCP ツール、Azure AI Search のエージェントによる取得、tool_choice、および根拠のある引用。

質問: **4**
エージェントを含む Microsoft Foundry プロジェクトがあります。

エージェントは、Azure Blob Storageに保存されているドキュメントから構築された知識ソースを使用します。これらのドキュメントには、複数ページの表を含むデジタルスキャンされたPDFファイルが含まれます。
プレーンテキストのみを抽出する取り込みジョブを使用しているため、テーブル構造、見出し、ページ番号などのメタデータが失われます。
ユーザーからは、複数のページにわたる特定の表の行を取得する必要があるという質問が頻繁に寄せられます。
スキャンされたPDFに対して光学文字認識（OCR）を実行し、表と見出しを構造を考慮したチャンクとして保持し、各チャンクにページ番号メタデータを保存する、検索拡張生成（RAG）パイプライン用の取り込みジョブを設定する必要があります。
データ取り込みジョブはどのように設定すればよいですか？
A. 基本的な構文解析と固定サイズのチャンキングを使用します。
B. 高度なデータ解析を使用してドキュメントを再取り込みします。
C. OCRとページレベルのチャンキングを使用します。
D. ページレベルのOCR抽出を使用し、各ページを単一のチャンクとして保存します。
正解: **(**正解を表示します**)**
適切な構成は高度なデータ解析です。なぜなら、問題は単なるOCRではなく、信頼性の高いRAG取得のために、取り込みジョブでドキュメント構造を維持する必要があるからです。Microsoftの高度な解析に関するガイダンスでは、スキャンされたドキュメント内のテーブルを含むすべてのページにわたるテーブルを自動的に検出し、複数ページにまたがるテーブルをマージし、列ヘッダーを復元し、テーブルインデックス、形状、ページ番号、セクション見出し、テーブルプレビューなどのメタデータを含むテーブルチャンクを作成するとされています。これにより、ソースページのコンテキストを維持しながら、複数ページのテーブルから特定の行を取得するという要件が直接満たされます。
固定サイズのチャンクを使用した基本的な解析では、ドキュメントが任意のテキスト断片に平坦化されてしまい、これが現在の失敗の原因となっています。ページレベルのチャンクを使用したOCRは、スキャンされたPDFからのテキスト抽出を改善しますが、ページをまたいで見出しや表の関係性を維持する構造認識型のチャンクは提供しません。各ページを単一のチャンクとして保存すると、行レベルの検索には粗すぎ、関連する表の行が過剰なコンテキストに埋もれてしまう可能性があります。高度なデータ解析は、意味的に意味のある検索可能なチャンクを生成し、引用や根拠付けに必要なメタデータでそれらを充実させるため、RAGの取り込み専用に設計されています。
参考トピック :RAG取り込み、高度な解析、OCR、テーブル抽出、構造認識型チャンキング、ページメタデータ、Azure Blob Storageドキュメント取り込み。

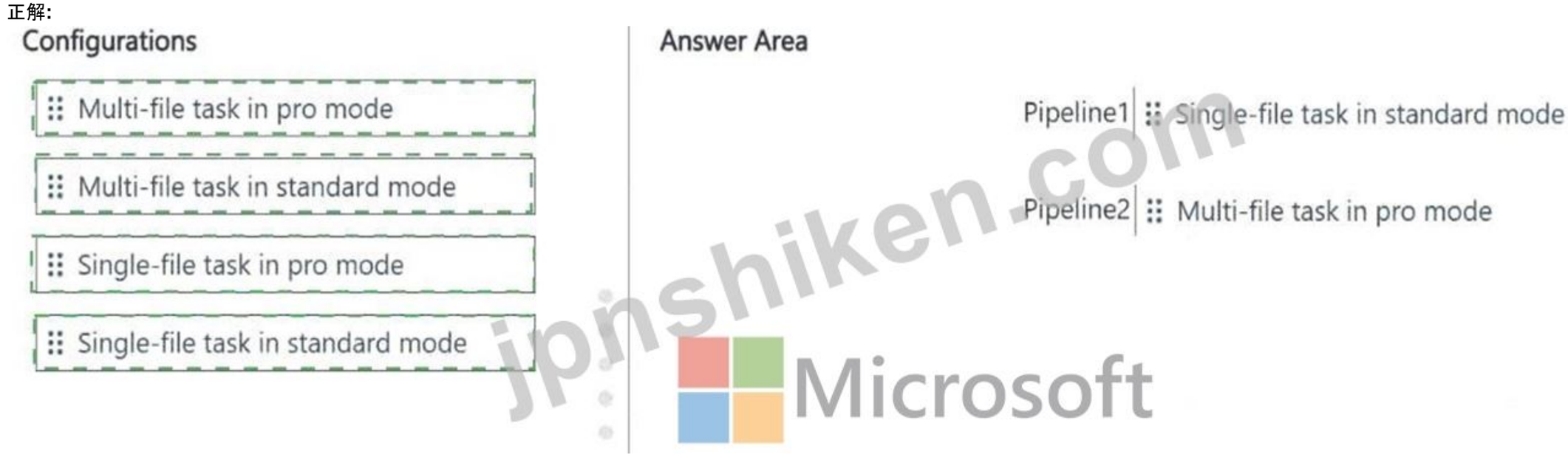
質問: **5**
Microsoft Foundry Agent Service を使用して構築されたカスタマーサポートエージェントがあります。このエージェントは、Azure OpenAL モデルのデプロイメントを呼び出します。
負荷テスト中に、呼び出しが断続的に失敗し、HTTP 429 レート制限超過エラーが返されます。
負荷がかかった状態での通話失敗を減らし、信頼性を向上させるためには、スロットリングを適切に処理する必要があります。このソリューションは、サービスおよびモデルの制限内に収まるものでなければなりません。
あなたはどうすべきでしょうか？
A. 指数バックオフとジッターを使用する再試行ポリシーを実装します。
B. 新しいスレッドを作成し、すぐに呼び出しを再試行します。
C. 登録ツールの数を減らします。
D. アップロードされたコンテンツをより小さなファイルに分割します。
正解: **A (**コメントを発表する**)**
正解は A です。指数バックオフとジッターを使用する再試行ポリシーを実装します。HTTP 429 は、リクエスト レートまたはトークン レートがモデル デプロイメントの構成済みサービス制限を超えていることを示します。Microsoft Foundry Agent Service の制限に関するガイダンスでは、エージェントがレート制限 429 エラーを受信した際に、アプリケーションの再試行ロジックで指数バックオフとジッターを実装することを特に推奨しています。また、デプロイメントの Azure OpenAI クォータ、1 分あたりのトークン数、および 1 分あたりのリクエスト数の制限を確認することも推奨しています。

このアプローチでは、既にスロットリングされているデプロイメントにすぐに負荷をかけるのではなく、再試行を段階的に遅らせることで、サービス制限内に収まりながら信頼性を向上させます。Microsoft Foundry Models のクォータに関するガイダンスでは、失敗したリクエストも分単位のレート制限にカウントされ、バックオフせずにリクエストを継続的に再送信するとスロットリングが悪化するとも述べています。そのため、指数バックオフを使用した再試行ロジックと、利用可能な場合は Retry-After ヘッダーの使用を推奨しています。
新しいスレッドを作成してすぐに再試行しても、デプロイメントのレート制限は変更されず、スロットリングが悪化する可能性があります。登録ツールを減らすとオーケストレーションが簡素化される場合がありますが、モデルの RPM または TPM の制限に直接対処するものではありません。アップロードされたコンテンツを小さなファイルに分割すると、取り込みシナリオに役立つ場合がありますが、断続的な HTTP 429 モデル呼び出しに対する適切なスロットリング制御ではありません。参照トピック: Foundry エージェント サービスの制限、Azure OpenAI のクォータ管理、スロットリング、再試行ポリシー、および運用環境の信頼性。

質問: **6**
Microsoft Foundry プロジェクトで画像編集ワークフローを作成しています。
ワークフローは以下の要件を満たす必要があります。
* マスクベースのインペインティング編集を適用することで、背景オブジェクトが削除できることを確認してください。
編集後の画像の元の照明とスタイルを維持する。
* カスタムモデルではなく、内蔵の画像編集コントロールを使用してください。
画像編集は、マスクされた領域内のみに適用されるようにする必要があります。
ワークフローはどのように設定すればよいでしょうか？
A. テキストから画像への変換モードを有効にし、希望する背景除去方法を説明するプロンプトを表示します。
B. 生成モードをimage_variationに設定し、元の画像を参照として提供します。
C. プロンプトに基づいて完全な画像を再生成するには、image_to_image モードと高強度値を有効にします。
D. mask_inpaintingを有効にし、入力画像と、画像内のどの部分を変更するかを示すマスクの両方を指定します。
正解: **D** (コメントを発表する)
正しい構成は D です。mask_inpainting を有効にし、入力画像と、変更する画像の部分を示すマスクの両方を指定します。要件は新しい画像を生成することではなく、周囲の照明、構図、スタイルを維持しながら、既存の画像の特定の領域を編集することです。Microsoft Foundry の Azure OpenAI 画像編集では、入力画像とプロンプトを送信することで、既存の画像を変更することができます。マスク編集の場合、マスクはモデルが変更できる画像の部分を明示的に定義します。Microsoft は、マスク パラメーターが編集する領域を定義し、入力画像の寸法と一致する必要があると述べています。
text_to_image はプロンプトから新しい画像を作成するため、元の画像の保持を保証することはできません。image_variation は、対象を絞った削除ではなく、関連するバリアントを生成します。image_to_image を高強度で使用すると、より広い領域が再生成され、無関係な視覚的詳細が変更される可能性があります。マスクベースのインペインティングは、選択した領域にのみ変更を限定する組み込みの編集制御です。参照トピック: Azure OpenAI 画像編集、マスクインペインティング、画像編集 API、入力画像、マスクパラメータ、およびコンピュータビジョン画像生成ワークフロー。

質問: **7**
サプライヤーから提出された調達文書を処理するMicrosoft Foundryプロジェクトがあります。
Foundry Tools の Azure Content Understanding を使用して、2 つのパイプラインを実装する必要があります。ソリューションは以下の要件を満たす必要があります。
* スタンドアロンの PDF 請求書をコスト効率よく大量に処理できる Pipeline1 という名前のパイプラインを含めます。
* マルチステップ推論と参照データを使用して文書間検証をサポートする、Pipeline2 という名前のパイプラインを含めます。
各パイプラインはどのように構成すればよいでしょうか？ 回答するには、適切な構成を正しいパイプラインにドラッグしてください。各構成は、1回、複数回、またはまったく使用しない場合があります。コンテンツを表示するには、ペイン間の分割バーをドラッグするか、スクロールする必要がある場合があります。
注：正解ごとに1ポイントが加算されます。

正解:



Explanation:

パイプライン1：標準モードの単一ファイルタスク

Pipeline2: プロモードでの複数ファイルタスク

Pipeline1 では、ワークロードがスタンドアロンの PDF 請求書の大量処理であるため、標準モードで単一ファイル タスクを使用する必要があります。Azure Content Understanding の標準モードは、単純な構造化抽出が必要な個々のファイルを対象としており、Microsoft は、広範なデータ中心の処理シナリオにおけるコストと遅延を最小限に抑えるものと説明しています。そのため、各 PDF を個別に処理できる、費用対効果の高い請求書抽出に最適です。

Pipeline2 は、要件に文書間検証、複数ステップの推論、および参照データが含まれるため、プロモードでマルチファイルタスクを使用する必要があります。Microsoft のガイダンスによると、プロモードは、単一のリクエストで複数の入力ファイルを処理する、文書間でデータを検証または拡充する、参照データを使用して抽出と検証をガイドするなど、複数ステップの推論とファイル間分析を必要とする高度なシナリオ向けに設計されています。

シングルファイル プロ モードでは、Pipeline1 に不要な機能が追加され、コスト効率の高い大容量スタンドアロン処理に最適化されません。マルチファイル スタンダード モードでは、参照データに基づく推論というプロ モードの要件を満たしません。参照トピック: Azure Content、スタンダード モード、プロ モード、シングル ファイル タスク、マルチ ファイル タスク、フィールド抽出、調達文書の検証について。

質問: **8**
あなたは、顧客サポートプラットフォーム向けに、Microsoft Foundry 上で音声処理ソリューションを構築しています。
このプラットフォームは通話内容をリアルタイムで文字起こしするため、社内の管理者は通話の文字起こしを確認し、通話中に問題点を検出できます。通話音声は電話システムから連続ストリームとして送信されます。
通話の文字起こしが音声ストリームの再生開始からわずか数秒以内に表示されるようにする必要があります。
あなたはどうすべきでしょうか？
A. 録音された音声ファイルに対してバッチ文字起こしジョブを実行します。
B. リアルタイム音声認識を使用して、ストリーミング音声入力を処理します。
C. 音声翻訳を使用して、複数の言語の文字起こしを生成します。
D. カスタムニューラルボイスを使用してテキスト読み上げ機能を使用します。
正解: **(**正解を表示します**)**
正解は B です。リアルタイム音声テキスト変換を使用して、ストリーミング音声入力を処理します。このシナリオでは、連続した電話ストリームからのリアルタイム文字起こしが必要であり、通話が進行中でも数秒以内に文字起こしテキストが表示されます。Foundry Tools の Azure Speech のリアルタイム音声認識は、コールセンターの支援、音声入力、ライブ会議の字幕など、即時文字起こしのシナリオ向けに特別に設計されています。Microsoft の音声ガイダンスでは、リアルタイム音声テキスト変換は音声入力を処理してリアルタイムで文字起こしを返すものと説明されており、これはスーパーバイザーの監視要件に合致しています。
バッチ文字起こしは、録音後に保存された音声ファイルを処理するものであり、アクティブなライブストリームを処理するものではないため、不適切です。音声翻訳は、単に同じ言語の通話の文字起こしをリアルタイムで作成するのではなく、音声を別の言語に翻訳することを主な目的とする場合に適用されます。テキスト読み上げは、テキストから音声を生成するという逆の操作を実行し、着信通話の文字起こしは行いません。リアルタイム音声テキスト変換は、ライブ運用監視に必要な低遅延のストリーミング認識パスを提供します。参照トピック：Foundry Tools の Azure Speech、リアルタイム音声認識、ストリーミング音声入力、コールセンターの文字起こし、ライブキャプション。

質問: **9**
Microsoft Foundry プロジェクトにチャット アプリがあり、Azure AI Search のベクトル化インデックスがあります。
以下の要件を満たすには、インデックスに接続する必要があります。
複雑な質問では、複数の情報チャンクから情報を取得する必要があります。
* 複数ターンにわたる会話は、情報検索計画に影響を与える必要がある。
* 遅延を削減するため、データ取得は並列で実行する必要があります。
どの検索方法を用いるべきでしょうか？
A. 古典的な検索拡張生成（RAG）
B. 思考の連鎖
C. エージェント型検索拡張生成 (RAG)
D. 反復検索
正解: **C (**コメントを発表する**)**
正解は、要件が Azure AI Search のエージェント取得パイプラインを説明しているため、エージェント取得拡張生成 (RAG) です。エージェント取得は、ユーザーのリクエストが複雑で会話的であり、以前のターンに依存する可能性があるチャットやコパイロットのシナリオ向けに設計されています。Azure AI Search のエージェント取得では、LLM 支援の計画ステージを使用して複雑なリクエストを焦点を絞ったサブクエリに分割し、単一のクエリパスに依存するのではなく、複数のチャンクから基本情報を取得できるようにします。
MicrosoftのAzure AI Searchのガイダンスでは、エージェントによる検索は、チャットやエージェントのワークフローにおける複雑な質問に対応するためのマルチクエリパイプラインであり、サブクエリには追加のコンテキストとしてチャット履歴を含めることができると説明されています。

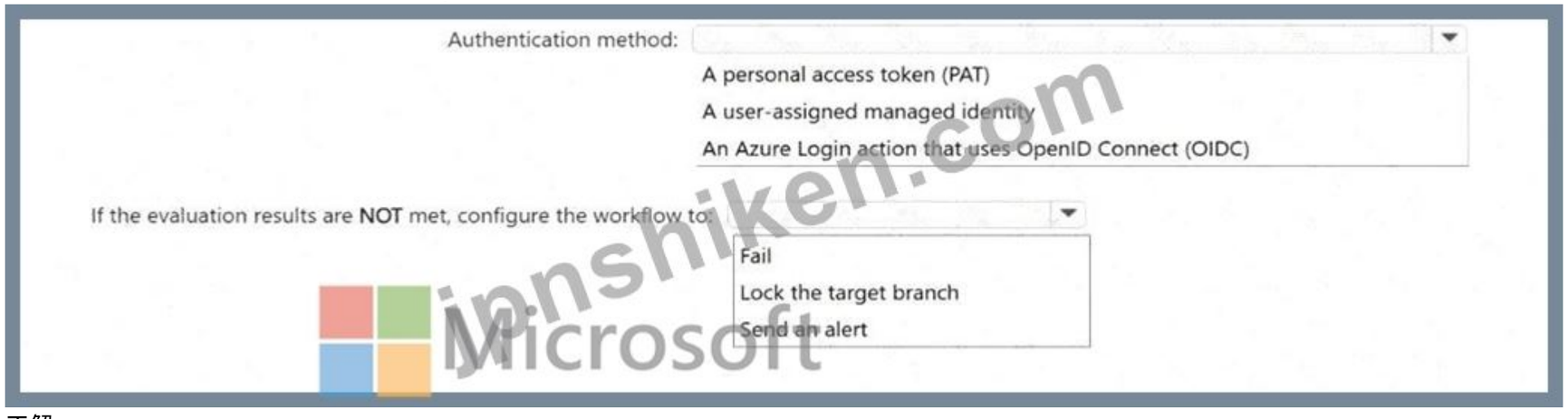
エージェントによる検索では、生成されたサブクエリを並列で実行し、生成モデルで使用するために最適な結果をマージして再ランク付けするため、レイテンシの要件も満たされます。従来の RAG はよりシンプルで、通常は単一のクエリを検索に送信するため、マルチホップまたは会話型の検索計画にはあまり適していません。Chain of thought は推論手法であり、Azure AI Search の検索アプローチではありません。また、反復型検索では、ここで説明する組み込みのクエリ計画、会話対応の検索、並列実行は提供されません。参照トピック: Azure AI Search エージェントによる検索、Azure AI Search を使用した RAG、知識ベース、クエリ計画、生成型 AI の基盤。

質問: **10**
ユーザーが写真をアップロードできるようにするサポートエージェントをデプロイしています。
アップロードされた画像を自動的に分類し、有害コンテンツを検出する必要があります。ソリューションは、有害度レベルに基づいてコンテンツをブロックする必要があります。
あなたはどうすべきでしょうか？
A. 画像モデレーションを実施してください。
B. プロンプトシールドを有効にする。
C. Foundry Tools の Azure Vision を使用して、光学文字認識 (OCR) 出力にキーワード スキャンを適用します。
D. ブロックリストを使用します。
正解: **(**正解を表示します**)**
正解はAです。画像モデレーションを実装してください。Azure AI Content Safetyは、アップロードされた画像を、憎悪、性的コンテンツ、暴力、自傷行為などの有害コンテンツカテゴリに分類する画像分析機能を提供します。MicrosoftのContent Safetyの概要によると、Analyze Image APIは、複数の深刻度レベルで有害コンテンツをスキャンします。これは、アップロードされた写真を自動的に分類し、設定された深刻度しきい値に基づいてコンテンツをブロックするという要件に直接合致しています。
プロンプトシールドは、生成モデルに対するプロンプトインジェクションやジェイルブレイクのような攻撃を検出することを目的としており、画像の有害性カテゴリを分類するものではありません。キーワードスキャンOCR出力では、画像から抽出された可視テキストのみが検出され、画像自体の視覚的な有害性は見逃されます。ブロックリストは既知の単語やカスタムパターンとの照合に役立ちますが、完全な画像安全分類器ではなく、ここで必要とされる組み込みの深刻度に基づく画像有害性分類機能を提供しません。したがって、ユーザーがアップロードした写真には、画像モデレーションが適切な制御手段となります。参照トピック :Azure AI コンテンツ安全、画像モデレーション、有害性カテゴリ、深刻度レベル、Foundry ガードレール、責任ある AI コントロール。

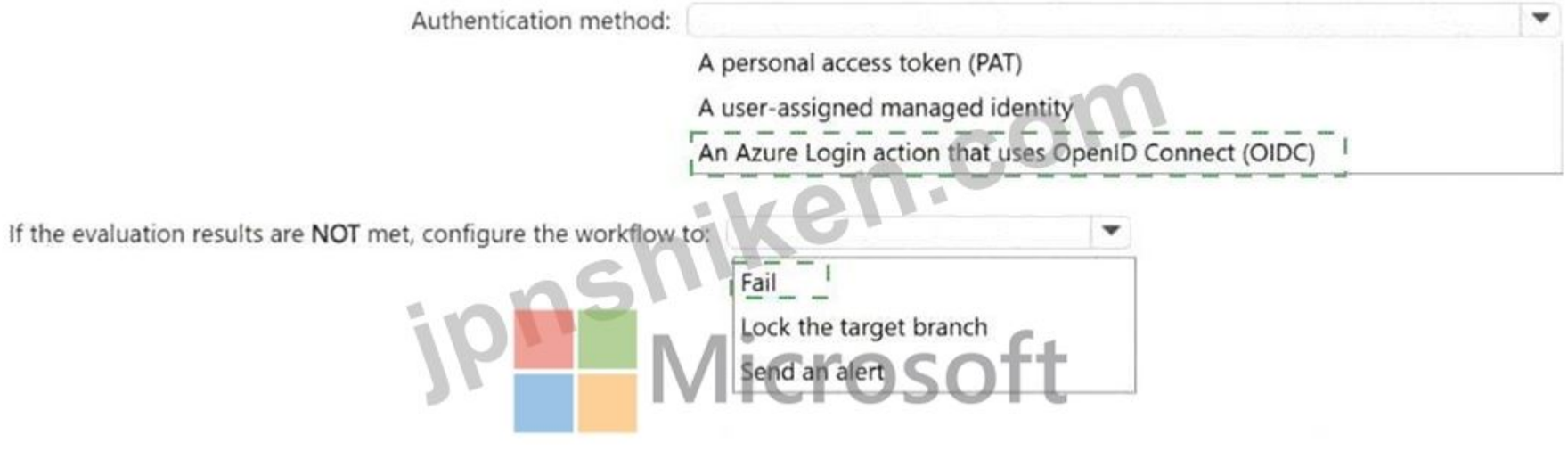
質問: **11**
スキャンされたPDF形式の請求書を処理するアプリケーションがあります。請求書のレイアウトは様々で、複数ページにわたる表が含まれています。
光学文字認識（OCR）を使用して合計金額と請求書番号を抽出するパイプラインがあります。
結果はしばしば
文書構造が無視されるため、誤りです。
OCR、レイアウト分析、およびテンプレート一般化フィールド抽出機能を提供するソリューションを実装する必要があります。このソリューションは、カスタムモデルのトレーニングを必要としないものでなければなりません。また、管理作業を最小限に抑える必要があります。
解決策には何を含めるべきでしょうか？
A. Azure Machine Learning モデル
B. Foundry ToolsにおけるAzureコンテンツの理解
C. Foundry ToolsにおけるAzure言語
正解: **(**正解を表示します**)**
正解は、Foundry Tools の Azure Content Understanding です。このシナリオでは、スキャンされた請求書のレイアウトが多様で、複数ページの表が含まれているため、基本的な OCR 以上の機能が必要です。Content Understanding は、インテリジェントなドキュメント処理のために設計されており、マネージド サービスとして OCR、レイアウト検出、表抽出、フィールド抽出、信頼度スコア、およびグラウンディングを提供します。Microsoft は、Content Understanding を、非構造化コンテンツを構造化出力に変換し、複雑なドキュメントからフィールドを抽出および検証することで請求書処理をサポートするサービスと説明しています。

これは、カスタムモデルのトレーニングを回避するという要件も満たしています。Content Understandingには、請求書や調達文書の処理など、事前構築済みのドメイン固有のアナライザーが含まれており、Microsoftはこれらのアナライザーがカスタムトレーニングなしで構造化された抽出を提供すると述べています。請求書のレイアウトごとに個別のモデルを用意するのではなく、意味論的な文書カテゴリを使用することで、ビジュアルテンプレートのバリエーション全体にわたって汎用性を実現します。
Azure Machine Learning は、モデルの開発、トレーニング、デプロイ、監視が必要となるため、管理作業が増加します。Azure Language は、テキストが利用可能になった後の分類やエンティティ抽出などのテキスト分析タスクに最適化されていますが、ドキュメントのレイアウト分析や複数ページの表構造の抽出は提供していません。参考トピック: コンテンツ理解、インテリジェントなドキュメント処理、OCR、レイアウト分析、アナライザー、フィールドスキーマ、構造化抽出。

質問: **12**
エージェントを含む Microsoft Foundry プロジェクトがあります。
CI/CDにはGitHub Actionsワークフローを使用します。
プルリクエスト（PR）が作成されたときにエージェントを自動的に評価し、評価結果が定義されたしきい値を満たさない場合はブランチのマージを防止するようにワークフローを設定する必要があります。
ワークフローはどのように設定すればよいですか？回答するには、回答欄で適切なオプションを選択してください。
注：正解ごとに1ポイントが加算されます。



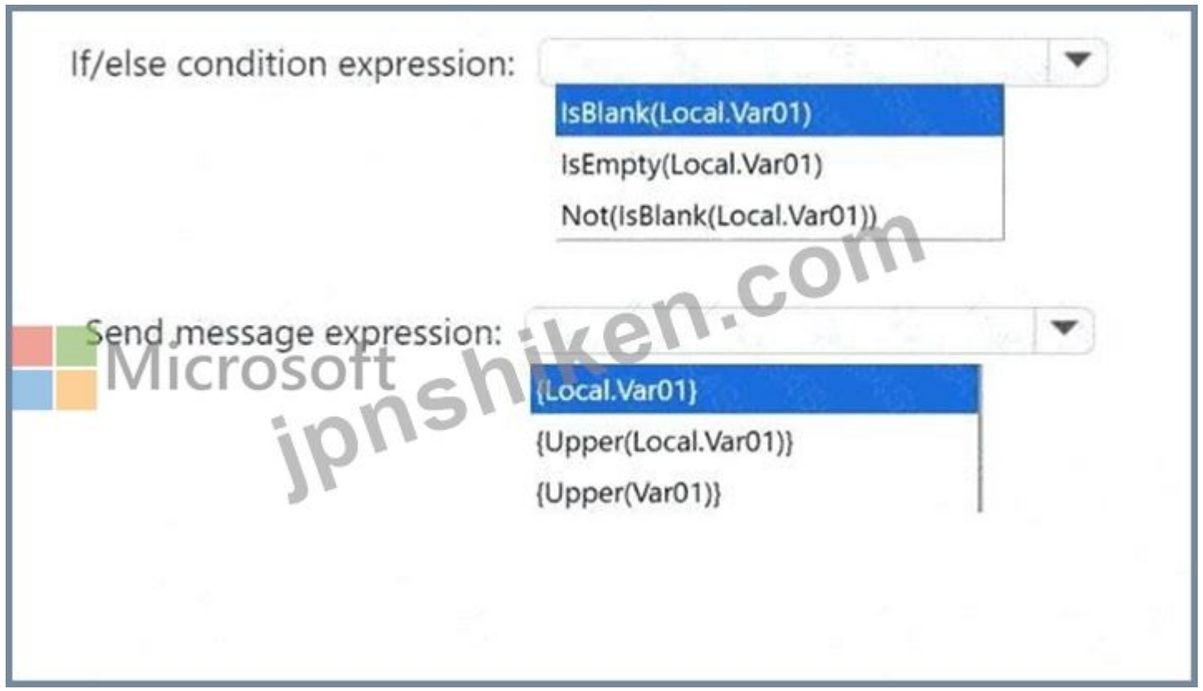
正解:



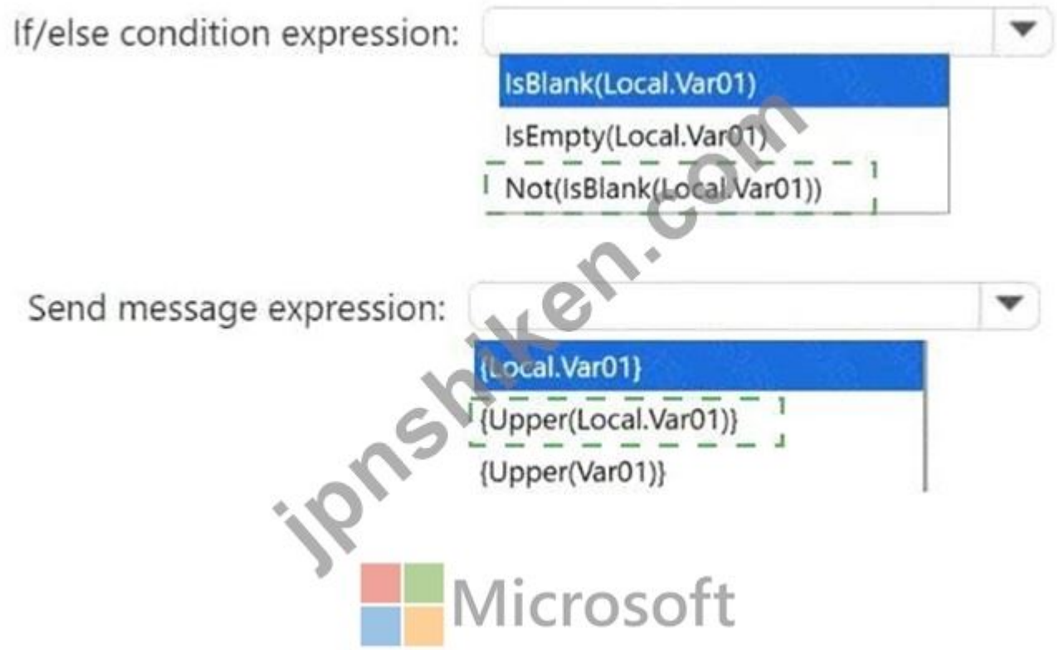
Explanation:

認証方法: OpenID Connect (OIDC) を使用する Azure Login アクション 評価結果が満たされない場合は、ワークフローを次のように構成します: 失敗 正しい認証方法は、OpenID Connect (OIDC) を使用した Azure Login です。Microsoft Foundry の GitHub Actions 評価ガイダンスでは、Microsoft Entra ID 認証を推奨しており、OpenID Connect を使用した Azure Login GitHub アクションで認証を自動化できると述べています。サンプル評価ワークフローでは、id-token: write を付与し、azure/login@v2 を実行してから、Microsoft AI Agent Evaluation アクションを呼び出します。これは、有効期限の長い個人アクセス トークンを回避し、GitHub Actions から Azure への安全なフェデレーション認証をサポートするため、適切な CI/CD 認証パターンです。
ワークフローは、評価しきい値が満たされない場合に失敗するように構成する必要があります。Foundry の評価 GitHub Action は、CI/CD パイプラインで Microsoft Foundry エージェントの事前評価を自動化し、構成された評価者とテスト データセットの評価結果を生成するように設計されています。GitHub Actions チェックが失敗した場合は、ブランチ保護によって強制的に実行できるため、品質ゲートが通過するまで PR をマージすることはできません。ターゲット ブランチをロックしたり、アラートを送信したりしても、CI 品質ゲートが直接実装されるわけではありません。参照トピック: Microsoft Foundry エージェントの評価、GitHub Actions 評価ワークフロー、Microsoft Entra 認証、Azure Login with OIDC、プル リクエストの品質ゲート、CI/CD ガバナンス。

質問**: 13**
顧客サポートのトリアージプロセス用のワークフローを含むMicrosoft Foundryプロジェクトがあります。
質問ノードがあり、ユーザーの回答はVar01という名前のローカル変数に保存されます。
以下のPower Fx式を作成する必要があります。
* Var01に値が含まれることを保証するif/else条件式
* 保存されたユーザー応答を大文字で返す送信メッセージ式。式はどのように構成すればよいですか？回答するには、回答領域で適切なオプションを選択してください。
注：正解ごとに1ポイントが加算されます。



正解:



Explanation:
if/else条件式: Not(IsBlank(Local.Var01))
メッセージ送信式: {Upper(Local.Var01)}
正しい if/else 条件は Not(IsBlank(Local.Var01)) です。ワークフローは、[質問する] ノードが空白以外の値をキャプチャした場合にのみ続行する必要があるためです。Power Fx では、IsBlank は値が空白かどうかをチェックし、Not はそのブール値を反転します。したがって、Not(IsBlank(Local.Var01)) は Var01 にユーザー入力が含まれている場合にのみ true と評価されます。IsEmpty(Local.Var01) は正しい選択肢ではありません。IsEmpty は、テキスト変数に値があるかどうかではなく、テーブルにレコードが含まれているかどうかをチェックするからです。Power Fx の数式リファレンスでは、空白値に対して IsBlank、空のテーブルに対して IsEmpty を定義することで、これらの関数を区別しています。
メッセージ送信式は {Upper(Local.Var01)} でなければなりません。Microsoft Foundry ワークフローガイダンスでは、同じパターンを使用して、Var01 として保存された質問ノードに続いて、メッセージ送信アクションが出力されます。
{Upper(Local.Var01)}。応答はローカルのワークフロー変数に格納されるため、Local. プレフィックスが必要です。Upper() は格納されたテキスト応答を大文字に変換します。参照トピック: Microsoft Foundry ワークフロー、Power Fx 式、ローカル変数、if/else 分岐、およびメッセージの送信アクション。

質問: **14**
注 :このセクションには、同じシナリオと問題に関する複数の質問セットが含まれています。各質問には、問題に対する固有の解決策が提示されています。提示された解決策が、提示された目標を満たしているかどうかを判断する必要があります。セット内の複数の解決策が問題を解決できる場合もあります。また、セット内のどの解決策も問題を解決できない場合もあります。
このセクションの質問に回答すると、前のセクションに戻ることはできません。そのため、これらの質問は復習画面には表示されません。
画像アップロードを受け付け、抽出した画像テキストを使用して応答を生成する、マルチモーダルなAI生成モデルをお持ちです。
ユーザーが安全でない画像をアップロードしたり、モデルを操作するための隠された指示を画像に埋め込んだりできることが分かります。
リスクを軽減するための対策を実施する必要があります。
解決策 :ユーザープロンプト用のプロンプトシールドを設定します。
これは目標を達成していると言えるでしょうか？
A. はい
B. いいえ

正解: **B** (コメントを発表する)
この解決策は目的を達成していません。ユーザープロンプト用のプロンプトシールドは、ユーザーがプロンプト自体を通してモデルを直接操作しようとする試みを検出するように設計されています。このシナリオでは、悪意のある指示はアップロードされた画像内に埋め込まれ、抽出された画像テキストを通してモデルのコンテキストに導入されます。
そのパターンは、単なる直接的なユーザープロンプト攻撃ではなく、間接的なプロンプト挿入またはドキュメント攻撃である。
マイクロソフトのプロンプトシールドに関するガイダンスでは、ユーザープロンプト攻撃とドキュメント攻撃を区別しており、ドキュメント攻撃とは、提供されたドキュメントやサードパーティのコンテンツに埋め込まれた有害な指示を伴う攻撃であると述べています。
ユーザーが安全でない画像をアップロードできるため、このソリューションは不完全です。Azure AI Content Safetyには、画像内の有害なコンテンツを検出し、さまざまなモダリティでのモデレーションをサポートする画像APIが含まれています。完全な対策としては、安全でないビジュアルコンテンツに対する画像モデレーションと、ドキュメント攻撃に対するプロンプトシールド、そしてオプションでスポットライト機能を組み合わせることで、OCRで生成されたテキストや埋め込まれた画像テキストを信頼性の低いコンテキストとして扱うことができます。
ユーザープロンプトのみを対象としたプロンプトシールドでは、安全でない画像や、それらの画像から抽出された隠された指示を確実にブロックすることはできません。参考トピック：Azure AI コンテンツセキュリティ、プロンプトシールド、ユーザープロンプト攻撃、ドキュメント攻撃、画像モデレーション、マルチモーダルセキュリティ。

質問: **15**
トラフィック量の多いエージェントを含む Microsoft Foundry プロジェクトがあります。
最近のアップデート後、運用コストが大幅に増加しました。
監視の結果、エージェントへのユーザーアクセス量は変化していないことが確認されました。
リクエストまたはレスポンスの特性の変化が、増加の原因となっているのではないかと疑っている。
追加コストの原因が、モデルの入力サイズ、出力サイズ、またはツール使用範囲の拡大のいずれにあるかを特定する必要があります。
どの可観測性機能を使用すべきでしょうか？
A. 評価指標
B. トークンの使用
C. レイテンシー
D. 実行成功率
正解: (正解を表示します)
適切な機能はトークン使用量です。Microsoft Foundry の可観測性では、リクエスト量が変わらない場合、トークン消費量がモデルコストの変化を診断するための主要なシグナルとなります。トークン使用量によって、プロンプトが大きくなった、取得またはツール提供のコンテキストが拡大した、応答が長くなった、またはエージェントの実行によってモデル呼び出しが増えたなど、コストが増加した原因を区別できます。Microsoft Foundry の監視ダッシュボードは、トークン消費量、レイテンシ、エラー率、品質スコアなどの運用メトリックを追跡し、エージェント監視ダッシュボードは、本番トラフィックのトークン使用量、レイテンシ、成功率、評価結果を分析することを目的としています。
これは、トラフィックの増加ではなく、リクエストまたはレスポンスの特性の変化が問題であるため、シナリオに直接一致します。入力トークンは、モデルに送信されるプロンプト、チャット履歴、グラウンディングデータ、またはツール出力が増加したかどうかを示します。出力トークンは、モデルがより長い完了応答を生成しているかどうかを示します。
ツールの使用範囲が拡大すると、後続のモデル要求にツールの結果、中間呼び出し、コンテキストが追加されるため、間接的にコストが増加する可能性があります。Foundryのトレースと可観測性により、エージェント実行時のツールの使用状況とトークン消費量が把握されます。
評価指標は、コスト要因ではなく、応答の品質と安全性を評価します。レイテンシはパフォーマンスの遅延を特定し、実行成功率は信頼性を測定します。参考トピック：Microsoft Foundry の可観測性、エージェント監視ダッシュボード、トークン消費量、コスト分析、ツール使用状況、および本番環境監視。

質問: **16**
Agent1は、Contoso製品に関する顧客からの質問にのみ回答するように設定する必要があります。このソリューションは、ビジネス要件を満たさなければなりません。
あなたはどうすべきでしょうか？
A. トップpサンプリングを適用する。
B. システムメッセージの指示を変更します。

C. 少数の撮影例を追加します。
D. 温度パラメータの値を増加させます。
正解: (正解を表示します)
正解はBです。システムメッセージの指示を変更します。ケーススタディでは、Agent1がContoso製品に関する一般的な質問に答えること、そしてビジネス要件として、Agent1がContosoが販売する製品に関する質問にのみ答えることが求められていることが述べられています。この要件はエージェントの許可されたドメインと拒否境界を定義するため、エージェントのシステムレベルの指示に明記する必要があります。Microsoft Foundryのガイダンスでは、システムメッセージはAzure OpenAIチャットモデルの動作を制御し、アシスタントの役割、境界、出力形式、安全性または品質の制約を定義するために使用されると説明されています。
システムメッセージは、エージェント1に対し、Contoso製品に関する質問にのみ回答し、利用可能な場合はContoso製品ドキュメントを使用し、Contoso製品以外の製品に関する質問は拒否するよう指示する必要があります。これにより、最高レベルの指示において、意図した業務範囲が直接的に強制されます。少数のサンプルは望ましい動作を強化するのに役立ちますが、必須の動作境界を定義するための主要な制御手段ではありません。トップpサンプリングと温度はデコード制御であり、エージェントが回答を特定の製品ドメインに限定するかどうかではなく、ランダム性と多様性に影響を与えます。温度を上げると、一貫性が低下する可能性があります。参照トピック :Microsoft Foundryエージェントの指示、システムメッセージ設計、プロンプトエンジニアリング、応答境界、およびグラウンデッド生成AI動作。

質問: **17**
Project1という名前のMicrosoft Foundryプロジェクトがあります。
プロジェクト1には、PDF形式の仕入先請求書を処理するアプリケーションが含まれています。
Foundry ToolsでAzure Document Intelligenceを構成して、PDFのセクションと表構造を保持するMarkdown出力を生成する必要があります。このソリューションは、開発作業を最小限に抑える必要があります。
あなたはどうすべきでしょうか？
A. 信頼度閾値を上げる。
B. PDFを解析する際に、出力をfiguresに設定します。
C. ドキュメントを分析する際に、content=markdown を設定します。
D. output_content_format=ContentFormat.MARKDOWN の値を設定します。
正解: **D** (コメントを発表する)
正解は D です。output_content_format=ContentFormat.MARKDOWN の値を設定します。Azure Document Intelligence Layout API は、見出し、段落、セクション、表、その他のレイアウト要素などの意味構造を保持したまま、抽出したドキュメント コンテンツを Markdown 形式で返すことができます。Microsoft の Document Intelligence レイアウト ガイダンスでは、prebuilt-layout モデルを使用してドキュメントを分析し、begin_analyze_document 呼び出しで output_content_format=ContentFormat.MARKDOWN を設定する Python SDK パターンが示されています。Markdown 出力は、分析結果の最上位コンテンツ セクションに返されます。
このサービスは、OCRで取得した生のテキストからセクションや表の書式設定を再構築するためのカスタム後処理を必要とせず、構造を保持したMarkdownを直接生成するため、開発作業を最小限に抑えることができます。MicrosoftのMarkdown出力に関するドキュメントによると、Markdown出力を指定すると、段落、見出し、表、その他のドキュメント要素を適切な階層構造で保持する、意味的に構造化されたコンテンツが生成されます。

オプションAは検証動作のみを変更し、Markdownは生成しません。オプションBは図を要求し、構造化Markdownは要求しません。オプションCはパラメーター名が間違っています。SDKのドキュメントに記載されている設定はoutput_content_formatであり、contentではありません。参考トピック：Azure Document Intelligence Layout API、Markdown出力、PDF分析、表の抽出、Foundry Toolsのドキュメント処理。

質問: **18**
貴社は、Microsoft Foundry プロジェクト「Project1」でカスタマーサポートエージェントの試験運用を行っています。Project1 は既存の Application Insights リソースに接続されており、貴社のサポートチームは「トレース」タブで実行結果を確認します。
Foundryエージェントサービスは、以下の動作を実行するように構成されています。
* Application Insights の接続文字列を取得するには、次の関数を呼び出します。
project_client.telemetry.get_application_insights_connection_string()。
* テレメトリを有効にするには、configure_azure_monitor(connection_string=...) を呼び出します。
OpenTelemetryを使用するように構成された別のLangChainサービスには、以下の設定があります。
* AzureAIOpenTelemetryTracerを使用します(connection_string=..., enable_content_recording=False)
* config={ " callbacks " :[azure_tracer]} を使用してトレーサーを渡します。
会社の方針には以下の要件があります。
* LangChainとOpenTelemetryからのテレメトリは、同じApplication Insightsリソース内で区別できる必要があります。
* プロンプト、ツール引数、またはスパン属性に、機密情報や認証情報を保存してはなりません。
以下の各記述について、正しい場合は「はい」を選択してください。そうでない場合は「いいえ」を選択してください。
注：正解ごとに1ポイントが加算されます。

Answer Area

Statements	Yes	No
The LangChain service will appear in Traces without configuring a tracer.	○	○
Setting different `OTEL_SERVICE_NAME` values separates the services in Application Insights.	○	○
When using `enable_content_recording=False`, prompts and tool data will be captured in the telemetry.	○	○

正解:

Explanation:
LangChain サービスは、トレーサーを構成しなくてもトレースに表示されます: いいえ 異なる OTEL_SERVICE_NAME 値を設定すると、Application Insights でサービスが分離されます: はい enable_content_recording=False を使用すると、プロンプトとツール データがテレメトリにキャプチャされます: いいえ 最初のステートメントは、別の LangChain または LangGraph アプリケーションが、構成されたトレース統合を介してテレメトリを出力する必要があるため、いいえです。Microsoft の LangChain トレース ガイダンスでは、AzureAIOpenTelemetryTracer を構成し、コールバックを介して実行可能ファイルまたはエージェントにアタッチし、Azure Monitor で出力されたトレースを検査するように指示されています。トラブルシューティング ガイダンスでは、トレース コールバックが実行にアタッチされていないことが、LangChain または LangGraph スパンが見つからない原因であるとも述べています。
2つ目の記述は「はい」です。OpenTelemetryでは、OTEL_SERVICE_NAMEはservice.nameリソース属性にマッピングされます。Azure Monitor Application Insightsは、クラウドロール名を使用して個別のサービスを表します。Microsoftは、複数のサービスが同じApplication Insightsリソースに情報を送信する場合、サービスが適切に表現されるようにクラウドロール名を設定する必要があると述べています。
3つ目のステートメントは「いいえ」です。enable_content_recording=False は、トレースからメッセージの内容とツール呼び出しの引数を削除するために特に使用されます。Microsoft は、運用環境ではコンテンツ記録を無効にし、プロンプトやツール引数にシークレット、資格情報、トークンを保存しないことも推奨しています。参照トピック: Microsoft Foundry トレース、LangChain トレース、OpenTelemetry サービス命名、Application Insights、およびセキュアなテレメトリ構成。

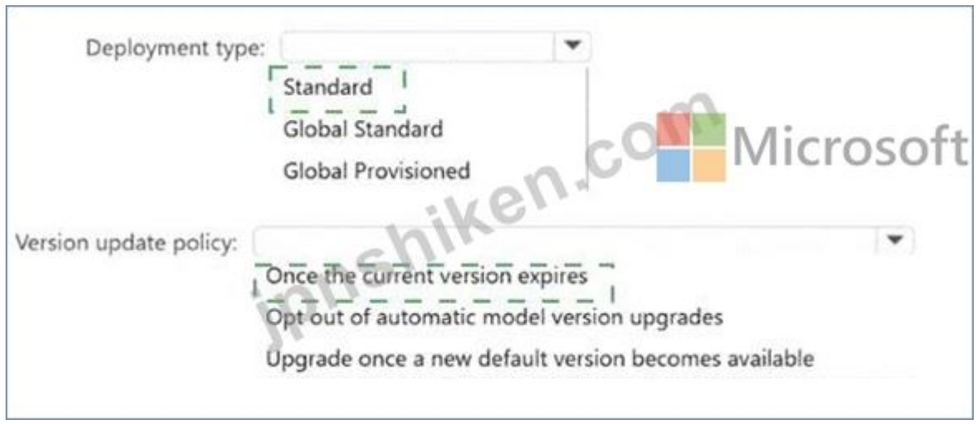
質問: **19**
注 :このセクションには、同じシナリオと問題に関する複数の質問セットが含まれています。各質問には、問題に対する固有の解決策が提示されています。提示された解決策が、提示された目標を満たしているかどうかを判断する必要があります。セット内の複数の解決策が問題を解決できる場合もあります。また、セット内のどの解決策も問題を解決できない場合もあります。
このセクションの質問に回答すると、前のセクションに戻ることはできません。そのため、これらの質問は復習画面には表示されません。
画像アップロードを受け付け、抽出した画像テキストを使用して応答を生成する、マルチモーダルなAI生成モデルをお持ちです。
ユーザーが安全でない画像をアップロードしたり、モデルを操作するための隠された指示を画像に埋め込んだりできることが分かります。
リスクを軽減するための対策を実施する必要があります。
解決策：保護対象物の検出設定を行います。
これは目標を達成していると言えるでしょうか？
A. はい
B. いいえ
正解: **(**正解を表示します**)**
このソリューションは目的を満たしていません。保護対象コンテンツの検出は、著作権で保護されたテキスト、特定のWebコンテンツ、歌詞、記事、レシピ、コードなど、既知の保護対象テキストまたはコードに一致する大規模な言語モデルの出力を識別することを目的としています。マイクロソフトは、保護対象コンテンツの検出を、AI生成コンテンツが既知の保護対象コンテンツを複製することを防止する制御として説明しており、画像の安全性やプロンプトの挿入を制御するものとしては説明していません。

報告されているリスクには、ユーザーが安全でない画像をアップロードする可能性があることと、ユーザーがモデルを操作するために画像に隠し命令を埋め込む可能性があることの2つの側面があります。Azure AI Content Safetyは、さまざまなモダリティにわたる有害コンテンツを検出する画像APIを提供し、有害カテゴリと深刻度に基づいてブロックの決定をサポートできるため、安全でない画像のアップロードには画像モデレーションが必要です。画像から抽出された隠し命令は、間接的なプロンプト挿入またはドキュメント攻撃です。Microsoft Prompt Shieldsは、サードパーティのコンテンツに埋め込まれた有害命令を含む、ユーザーのプロンプト攻撃とドキュメント攻撃を検出するように設計された機能です。
したがって、保護対象コンテンツの検出だけでは、いずれの主要なリスクも軽減できません。参考トピック :Azure AI コンテンツ セーフティ、画像モデレーション、プロンプト シールド、ドキュメント攻撃、間接的なプロンプト挿入、保護対象コンテンツの検出。
トピック1、ケーススタディ Contoso, Ltd
概要
Contoso, Ltdは、Microsoft Foundryを使用して生成型AIおよびエージェントベースのソリューションを構築、展開、管理する多国籍小売企業です。
アイデンティティ環境:
Contosoは、エージェントが組織のリソースやサービスにアクセスできるようにするID管理、認証、認可機能にMicrosoft Entra IDを使用しています。
Contosoは最近、既存のAIソリューションを最適化および保守するために、Agent1Dev Teamという新しいAIエンジニアリングチームを結成しました。
チームは、ソリューションアーキテクト、DevOpsエンジニア、セキュリティエンジニアと協力して、AIアプリケーションの設計、実装、監視、およびセキュリティ対策を行います。
Contosoには、AIソリューションの展開前にその妥当性を検証するAgent1Testチームというチームも存在する。
生成環境：
Contosoは、Project1とProject2という名前の2つのプロジェクトを含むMicrosoft Foundry環境を構築しています。
プロジェクト1
Project1には、顧客からの製品に関する問い合わせやトラブルシューティングの依頼に対応する、Agent1という名前のカスタマーサポートエージェントが含まれています。
Agent1には以下の設定があります。
* Agent1は基本モデルのデプロイメントを使用します。
* 安全性評価パイプラインは有効になっていません。
* ツール呼び出し承認ワークフローは有効になっていません。
* 会話メモリの制約は設定されていません。
Agent1は、デジタルサポートチャネルを使用して顧客とやり取りし、Contoso製品に関する一般的な質問に回答します。
Project1は、欧州連合（EU）域内にあるAzureリージョンにデプロイされています。
Agent1開発チームは、Project1を使用してAgent1の最適化と保守を行います。
プロジェクト2
Project2には、既に展開済みの動画生成モデルが含まれています。Contoso社のマーケティング部門はProject2にアクセスでき、このモデルを用いて動画制作ソリューションを開発する予定です。
解決策の開発は未完了です。
データ環境:
Contosoは、AIアプリケーションをサポートするAzureリソースに製品関連情報を保存します。
Azure環境には、Contoso製品の製品詳細シートを保存するstorage1という名前のAzure Blob Storageアカウントが含まれています。
製品仕様書には、仕様、機能説明、および製品サポート情報が記載されており、エージェント1はこれらを使用して顧客からの質問に回答できます。製品仕様書はPDF形式で保存されます。
問題提起：
コントソ氏は以下の問題点を指摘している。
* エージェント1は、コントソ製品に関する一般的な知識しか持っていません。
* エージェント1との最近のチャットのやり取りについて、感情分析が行われました。分析結果はまだ処理されていません。
* Agent1は、顧客からの質問に回答する際に、storage1に保存されている製品シートの詳細な製品情報を使用しません。

コントソ社の財務部門によると、仕入先からの請求書は、契約書に定められた条件と一致していることを確認するため、手作業で精査する必要があるとのことです。請求書には表やロゴ、様々なレイアウトが含まれているため、一貫した処理が困難です。
要件：
予定されている変更点：
コントソ氏は以下の変更を実施する予定です。
* プロジェクト1向けに、仕入先請求書の視覚的なレイアウトとテキスト内容の両方を評価することで請求書を分析し、請求書の詳細が仕入先契約条件と照合できるようにするソリューションを実装する。
* Agent1で使用される基本モデルのデプロイメントを更新し、モデルのバージョンを標準化して、継続性と一貫した応答を確保します。
* Agent1がstorage1に保存されている製品シートから詳細な製品情報を取得して使用できるようにします。
* Agent1が顧客からの質問に回答するために使用できる製品シートのインデックス作成ソリューションを実装する。
* 動画制作ソリューションの開発を完了する。
技術要件:
Contoso社は、以下の技術要件を特定しました。
* Agent1で使用されるモデル展開は、拡張性が高く、高スループットの生成型AIワークロードをサポートし、予約済みのスループット容量を必要とせずに、変動する顧客サポートトラフィックを処理するために動的に拡張する必要があります。
* 製品シートは、セマンティック検索とベクトル検索を可能にするインデックス作成パイプラインを使用して処理され、Agent1が関連する製品情報を取得できるようにする必要があります。
* 製品シートの情報に基づいて生成される回答は、関連性があり、完全かつ正確でなければなりません。
* エージェント1は、製品シートを使用して、製品の詳細に関する自然言語の質問に答えることができなければなりません。
* エージェント1が使用するモデルのバージョンは、安定した応答を確保するために一貫している必要があります。
* モデルによって処理されるデータは、EU域内に留まらなければなりません。
安全および法令遵守要件：
Contosoは、以下のセキュリティおよびコンプライアンス要件を特定しています。
* APIキーを使用してFoundryにデプロイされたモデルにアクセスしてはなりません。
* Azureリソースへのアクセスは、最小権限の原則に従う必要があります。
* Contosoの開発者は、Microsoft Entra認証を使用してMicrosoft Foundryリソースへの認証を行う必要があります。
* Project1へのアクセス権は、SC_Agent1_Devという名前のセキュリティグループを使用して、Agent1Devチームのメンバーに割り当てる必要があります。
* Project1へのアクセス権は、SC_Agent1_Testという名前のセキュリティグループを使用して、Agent1Testチームのメンバーに割り当てる必要があります。
* Agent1は、顧客データを含む文書がストレージ1の製品シートリポジトリに誤って追加された場合でも、顧客情報を決して開示してはならない。
* 製品仕様書には、テキストが埋め込まれた画像が含まれている場合があります。Agent1は、画像内に隠されている可能性のある悪意のある命令から保護される必要があります。
事業情報：
Contosoは以下のビジネス要件を特定しました。
* Agent1とやり取りするユーザーは、今後のやり取りにおいてパーソナライズされた体験を得る必要があり、Agent1が会話のコンテキストを保持し、以前のやり取りから関連情報を想起できる機能も必要です。
* エージェント1は、Contosoが販売する製品に関する質問のみに回答しなければなりません。

質問: **20**
Agent1のモデル展開を技術要件を満たすように構成する必要があります。
何を設定すればよいですか？回答するには、回答欄で適切なオプションを選択してください。
注：正解ごとに1ポイントが加算されます。

正解:

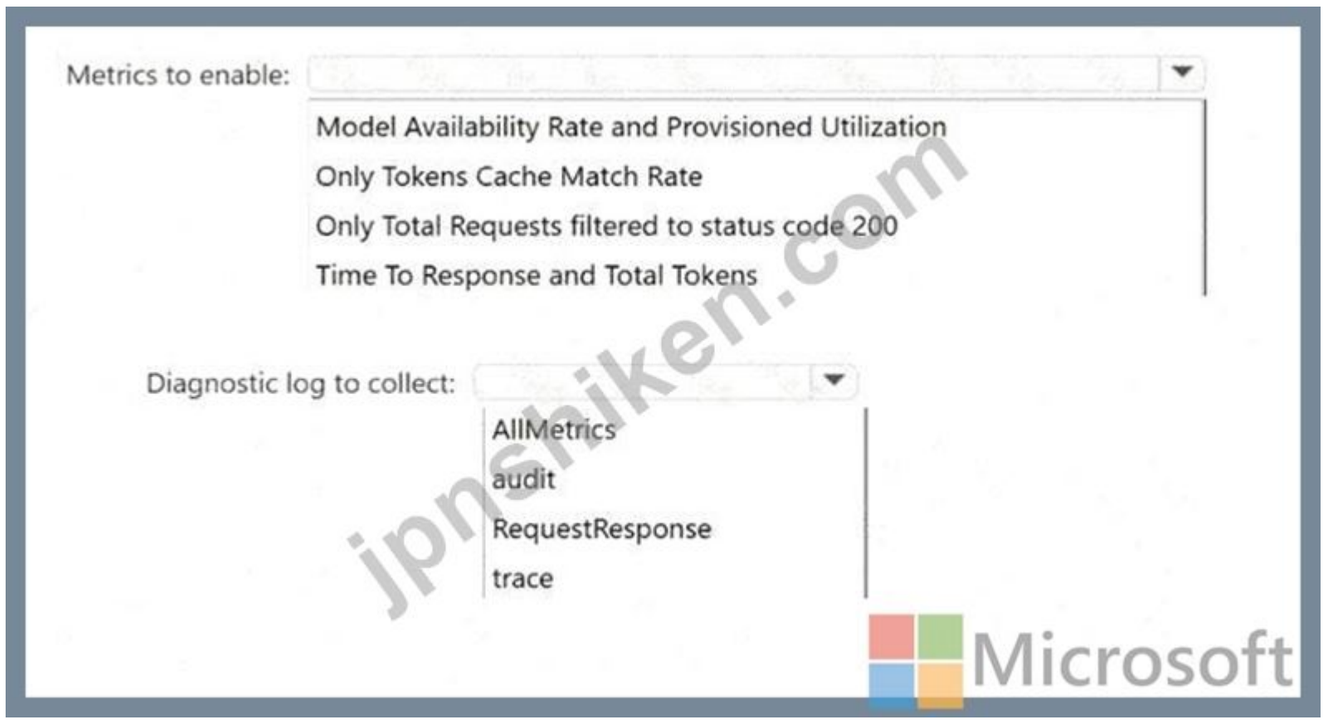


Explanation:
展開タイプ: 標準
バージョン更新ポリシー：現在のバージョンの有効期限が切れた後
適切なデプロイ タイプは Standard です。ケース スタディでは、Project1 が EU Azure リージョンにデプロイされ、モデル処理されたデータは EU 内に留まる必要があると指定されています。また、予約されたスループット容量なしで、変動する顧客サポート トラフィックを動的に処理する、スケーラブルで高スループットの生成型 AI ワークロードも必要です。Microsoft Foundry Models では、Standard はトークンごとに課金されるデプロイ タイプで、単一の Azure リージョンでデータを処理します。一方、Global Standard はリージョンをまたいでリクエストを処理でき、Global Provisioned は予約されたプロビジョニング済みスループットを使用します。Microsoft のデプロイ タイプ ガイダンスでは、Standard は単一リージョン、トークンごとに課金されるタイプ、Global Provisioned はリージョンをまたいで予約された容量を持つタイプとして識別されています。
正しいバージョン更新ポリシーは「現在のバージョンの有効期限が切れたら」です。これにより、Agent1はサポート対象ライフサイクル期間中、選択されたモデルバージョンに維持され、安定した一貫性のある応答が保証されます。また、現在のバージョンが廃止された際には、サポート対象の代替バージョンに自動的に移行することで、継続性も維持されます。
Microsoft のモデル バージョン管理に関するガイダンスでは、このポリシーは現在のモデル バージョンが期限切れになった場合にのみ更新され、新しいデフォルトが利用可能になったときにアップグレードすると展開が早期に変更され、オプトアウトすると廃止後に展開が動作しなくなる可能性があると規定されています。参照トピック: 展開の種類、リージョン別データ処理、モデル バージョン管理、スループット容量、安定した運用展開。

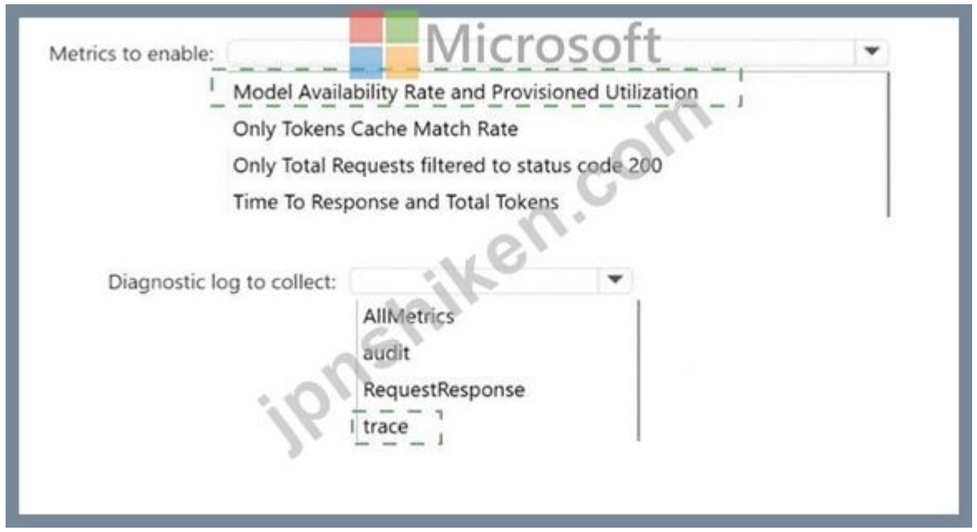
質問: **21**
エージェントを含む Microsoft Foundry プロジェクトがあります。
エージェントの知識源は、Azure Blob Storageに保存されているスキャン済みのPDFトラブルシューティングガイドのセットです。ガイドページは2列レイアウトと表で構成されています。
PDFを処理するには、Foundry ToolsのAzure Content Understandingを使用します。
処理済みのコンテンツをインデックスに取り込み、検索拡張生成（RAG）に利用し、抽出したフィールドを後続の自動化のために保存する予定です。

関係者は、抽出された各フィールド値が元のPDFのどこから来たのかを確認できなければならず、信頼性の低い抽出結果は手動レビューのために回送する必要がある。
コンテンツ理解ドキュメントアナライザーの出力に、フィールドごとの信頼度スコアと、ソースドキュメント内のソース特定箇所が含まれていることを確認する必要があります。
あなたはどうすべきでしょうか？
A. estimateFieldSourceAndConfidence を有効にします。
B. アナライザーを、すべてのフィールドに対して生成抽出を使用するように設定します。
C. enableSegment を true に設定します。
D. ラベル付きサンプルを提供する。
正解: **A (**コメントを発表する**)**
正解は A です。estimateFieldSourceAndConfidence を有効にします。Azure Content Understanding ドキュメント アナライザーは、フィールド抽出のオプトイン信頼度とグラウンディング機能をサポートしています。Microsoft のドキュメントには、信頼度とグラウンディングをオプトインするには、estimateFieldSourceAndConfidence を設定すると記載されています。
アナライザー構成で = true に設定するか、特定のフィールドに対して estimateSourceAndConfidence = true を設定します。
これにより、抽出された各フィールドに信頼度スコアと元の文書ソースの場所への参照を含めることができるようになります。
これは、関係者双方の要求を直接的に満たします。ソースの特定により、ユーザーは抽出された値がスキャンされたPDFのどこから来たのかを確認でき、信頼度スコアは、信頼度の低いフィールドを手動レビューに送るなど、下流の自動化ルールをサポートします。マイクロソフトのアナライザー改善ガイダンスでは、信頼度スコアは0から1の間の値、特定は抽出された出力が元のドキュメントコンテンツを参照または引用したものであると説明されています。
生成抽出では、フィールドごとの信頼性やソースの特定は保証されません。enableSegment は文書のセグメンテーションに使用され、信頼性スコアリングには使用されません。ラベル付きサンプルは抽出品質を向上させることができますが、それ自体では信頼性や特定情報の出力は保証されません。参考トピック :コンテンツ、文書アナライザー、フィールド抽出、信頼性スコアリング、特定情報、RAG 取り込みについて理解する。

質問: **22**
内部Q&Aエージェントを含むMicrosoft Foundryプロジェクトがあります。
ユーザーがエージェントに質問した際に、以下のような問題が報告されています。
* 関連情報が見つかりませんでした」という応答が増加しました。
* ピーク時間帯にHTTP 429レート制限超過エラーが定期的に発生
各問題の原因が、モデルの利用不可、リソース制限、または推論の失敗のいずれであるかを特定する必要があります。
どうすればよいですか？回答するには、回答欄で適切な選択肢を選んでください。
注：正解ごとに1ポイントが加算されます。



正解:



Explanation:

有効化する指標 :モデル可用性率およびプロビジョニング済み利用率

収集する診断ログ: トレース

適切な指標は、モデル可用性率とプロビジョニング済み利用率です。モデル可用性率は、総呼び出し数からサーバーエラー数を差し引いて算出されるため、サービス側のモデルが利用できないことが障害の原因かどうかを識別できます。プロビジョニング済み利用率は、リソース制限状態を識別します。Microsoftによると、利用率が100%に達するか超えると、呼び出しが制限され、HTTP 429エラーが返されます。これは、報告されたピーク時のレート制限エラーに直接対応します。

適切な診断ログはトレースです。社内Q&Aシナリオでは、未回答の質問を含む質問応答の動作を分析するためにトレースログが必要です。Microsoftのカスタム質問応答分析ガイダンスでは、診断ログにはテレメトリとチャットログが保存され、監査、リクエストレスポンス、およびAllMetricsに加えてトレースを有効にするよう指示されています。未回答の質問に対するKustoのサンプルクエリでは、回答、質問、スコア、およびナレッジベースIDを検査し、スコアがゼロの未回答の結果をフィルタリングします。
RequestResponse はリクエストの状態とレイテンシの把握に役立ち、Audit は管理操作の把握に役立ちますが、どちらも 関連情報が見つかりません」などの Q&A 推論動作の分析には最適な診断カテゴリではありません。参照トピック: Foundry の監視、モデルの可用性、プロビジョニングされた利用率、診断ログ、カスタム質問応答分析、トレースログ。

質問: **23**
あなたは、複数のエージェントを含む Project1」という名前のMicrosoft Foundryプロジェクトを計画しています。各エージェントは同じAzure AI Searchリソースにアクセスします。
Project1内でAzure AI Searchの認証情報を一元管理するためのソリューションを提案してください。このソリューションは、すべてのエージェントに実装する必要があります。
何をおすすめしますか？
A. Azure AI Search リソースに対してロールベースのアクセス制御 (RBAC) を有効にします。
B. Azure AI Search リソースへの接続を追加します。
C. Azure AI Search リソースのキーベースのアクセス制御を無効にします。
D. Azure AI Search リソースに接続するマネージド プライベート エンドポイントを作成します。
正解: **(正解を表示します)**
正しい推奨事項はBです。Azure AI Searchリソースへの接続を追加します。Microsoft Foundryプロジェクト接続は、プロジェクトからAzure AI Searchなどの外部リソースへのアクセスを一元的に定義するために使用されます。公式の接続ガイダンスでは、Azure AI Searchなどの外部サービスを選択し、リソースを選択して、そのリソースの認証方法を選択することで接続を追加できると説明されています。これにより、各エージェントが検索資格情報を個別に保存または複製する必要がなくなり、再利用可能なプロジェクトレベルの構成が作成されます。
Azure AI Search を使用する Foundry エージェントの場合、Azure AI Search ツールには project_connection_id が必要です。これは、Azure AI Search へのプロジェクト接続のリソース ID です。これにより、複数のエージェントが同じマネージド接続を参照しながら、構成済みのエンドポイントと認証設定を一貫して使用できます。RBAC はキーレス認証設計の一部となる場合がありますが、それ自体では Foundry 資格情報の構成を一元管理することはできません。キーベースのアクセスを無効にするとセキュリティ体制は強化されますが、エージェントは接続されません。マネージド プライベート エンドポイントは、資格情報の集中管理ではなく、ネットワーク分離に対応します。
参考トピック :Microsoft Foundry プロジェクト接続、Azure AI Search ツール、プロジェクト接続 ID、認証、およびエージェントのグラウンディング。

質問: **24**
注 :このセクションには、同じシナリオと問題に関する複数の質問セットが含まれています。各質問には、問題に対する固有の解決策が提示されています。提示された解決策が、提示された目標を満たしているかどうかを判断する必要があります。セット内の複数の解決策が問題を解決できる場合もあります。また、セット内のどの解決策も問題を解決できない場合もあります。
このセクションの質問に回答すると、前のセクションに戻ることはできません。そのため、これらの質問は復習画面には表示されません。
エージェントを含む Microsoft Foundry プロジェクトがあります。このエージェントは、取得したポリシー文書から概要を生成します。
ユーザーからは、取得したコンテンツに規定の規制条項が含まれている場合でも、一部の回答では必要な規制条項が省略されているとの報告がある。
回答の完全性を改善する必要があります。
解決策：必要な句が欠落している場合にレスポンスを再生成するリフレクションパスを追加します。
これは目標を達成していると言えるでしょうか？
A. はい
B. いいえ
正解: **A (コメントを発表する)**
はい、このソリューションは目的を達成しています。問題は取得の可用性ではなく、必要な規制条項は既に取得済みのポリシー文書に含まれているためです。問題は生成時に発生しており、エージェントが必要なコンテンツを省略した要約を生成しています。リフレクションパスは、応答を返す前に検証ステップを追加するため、適切なアプリケーションレベルの制御方法です。このパスは、ドラフト回答を取得した条項と比較し、不足している必須コンテンツを検出し、要約に必要な条項が含まれるまで再生成または修正をトリガーできます。

これは、Microsoft Foundry の評価および可観測性モデルに準拠しており、生成された応答は、AI アプリケーションのライフサイクル全体を通して、信頼性、妥当性、関連性、および品質について評価されます。Foundry の可観測性ガイダンスでは、評価は、開発および運用ワークフロー全体で応答の品質を測定し、AI 出力を改善するためのメカニズムとして説明されています。Azure AI 評価 SDK では、完全性とは、生成された応答が、提供された正解データに関して必要なすべての関連情報を含んでいる程度であると定義されています。リフレクションは、事後に欠陥を報告するだけでなく、アプリケーション フロー内でその品質チェックを運用化します。参照トピック: モデル リフレクション、応答の完全性、RAG 生成品質、取得コンテキストの検証、およびエージェント応答の最適化。