

EXAMTOPICS

- Expert Verified, Online, **Free**.



CERTIFICATION TEST

- [CertificationTest.net](https://www.certificationtest.net) - Cheap & Quality Resources With Best Support

HOTSPOT -

Case Study -

This is a case study. Case studies are not timed separately from other exam sections. You can use as much exam time as you would like to complete each case study. However, there might be additional case studies or other exam sections. Manage your time to ensure that you can complete all the exam sections in the time provided. Pay attention to the Exam Progress at the top of the screen so you have sufficient time to complete any exam sections that follow this case study.

To answer the case study questions, you will need to reference information that is provided in the case. Case studies and associated questions might contain exhibits or other resources that provide more information about the scenario described in the case. Information provided in an individual question does not apply to the other questions in the case study.

A Review Screen will appear at the end of this case study. From the Review Screen, you can review and change your answers before you move to the next exam section. After you leave this case study, you will NOT be able to return to it.

To start the case study -

To display the first question in this case study, select the "Next" button. To the left of the question, a menu provides links to information such as business requirements, the existing environment, and problem statements. Please read through all this information before answering any questions. When you are ready to answer a question, select the "Question" button to return to the question.

Overview -

Company Information -

Contoso, Ltd is a multinational retail company that builds, deploys, and manages generative AI and agent-based solutions by using Microsoft Foundry.

Existing Environment -

Identity Environment -

Contoso uses Microsoft Entra ID for identity management, authentication, and authorization capabilities that enable agents to access organizational resources and services.

Contoso recently formed a new AI engineering team named Agent1Dev Team to optimize and maintain existing AI solutions.

The team collaborates with solution architects, DevOps engineers, and security engineers to design, implement, monitor, and secure AI applications.

Contoso also has a team named Agent1Test Team that is responsible for validating AI solutions before the solution deployments.

Generative Environment -

Contoso has a Microsoft Foundry deployment that contains two projects named Project1 and Project2.

Project1 -

Project1 contains a customer support agent named Agent1 that assists customers with product inquiries and troubleshooting requests.

Agent1 has the following configurations:

Agent1 uses a base model deployment.

A safety evaluation pipeline is NOT enabled.

Tool invocation approval workflows are NOT enabled.

Conversation memory constraints are NOT configured.

Agent1 interacts with customers by using digital support channels and answers general questions about Contoso products.

Project1 is deployed to an Azure region located in the European Union (EU).

Agent1Dev Team will use Project1 to optimize and maintain Agent1.

Project2 -

Project2 contains a deployed video generation model. The marketing department at Contoso has access to Project2 and plans to use the model to develop a video creation solution.

Development of the solution is incomplete.

Data Environment -

Contoso stores product-related information in Azure resources that support AI applications.

The Azure environment contains an Azure Blob Storage account named storage1 that stores product detail sheets for all the Contoso products. The product sheets include specifications, feature descriptions, and product support information that Agent1 can use to answer customer questions. The product sheets are stored in the PDF format.

Problem Statements -

Contoso identifies the following issues:

Agent1 has only general knowledge of the Contoso products.

A recent chat interaction with Agent1 was analyzed for sentiment. The results of the analysis have NOT been processed yet.

Agent1 does NOT use the detailed product information in the product sheets stored in storage1 when responding to customer questions.

The finance department at Contoso reports that vendor invoices must be reviewed manually to ensure that the invoices match the terms defined in the vendor contracts. The invoices contain tables, logos, and varied layouts that make the documents difficult to process consistently.

Requirements -

Planned Changes -

Contoso plans to implement the following changes:

Implement a solution for Project1 that analyzes the vendor invoices by evaluating both the visual layout and the textual content of the invoices, so that the invoice details can be verified against the vendor contract terms.

Update the base model deployment used by Agent1 and standardize the model version to ensure continuity and consistent responses.

Enable Agent1 to retrieve and use the detailed product information from the product sheets stored in storage1.

Implement an indexing solution for the product sheets that Agent1 can use to answer customer questions.

Complete the development of the video creation solution.

Technical Requirements -

Contoso identifies the following technical requirements:

The model deployment used by Agent1 must support scalable, high-throughput generative AI workloads and dynamically scale to handle variable customer support traffic, without requiring reserved throughput capacity.

The product sheets must be processed by using an indexing pipeline that enables semantic and vector search, so that Agent1 can retrieve the relevant product information.

Responses generated by using the product sheet information must be relevant, complete, and accurate.

Agent1 must be able to use the product sheets to answer natural language questions about product details.

The model version used by Agent1 must remain consistent to ensure stable responses.

The data processed by the model must remain within the EU.

Security and Compliance Requirements

Contoso identifies the following security and compliance requirements:

API keys must NOT be used to access Foundry-deployed models.

Access to the Azure resources must follow the principle of least privilege.

The developers at Contoso must authenticate to Microsoft Foundry resources by using Microsoft Entra authentication.

Access to Project1 must be assigned to the members of Agent1Dev Team by using a security group named SC_Agent1_Dev.

Access to Project1 must be assigned to the members of Agent1Test Team by using a security group named SC_Agent1_Test.

Agent1 must never reveal customer information, even if a document that contains customer data is added erroneously to the product sheet repository in storage1.

The product sheets might contain images that include embedded text. Agent1 must be protected from malicious instructions potentially hidden within the images.

Business Requirements -

Contoso identifies the following business requirements:

Users that interact with Agent1 must have a personalized experience in future interactions, including the ability for Agent1 to retain conversation context and recall relevant information from previous interactions.

Agent1 must answer questions only about the products sold by Contoso.

You need to configure the model deployment for Agent1 to meet the technical requirements.

What should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Deployment type:

- Standard
- Global Standard
- Global Provisioned

Version update policy:

- Once the current version expires
- Opt out of automatic model version upgrades
- Upgrade once a new default version becomes available

Answer Area

Suggested Answer:

Deployment type:

- Standard
- Global Standard
- Global Provisioned

Version update policy:

- Once the current version expires
- Opt out of automatic model version upgrades
- Upgrade once a new default version becomes available

None

Case Study -

This is a case study. Case studies are not timed separately from other exam sections. You can use as much exam time as you would like to complete each case study. However, there might be additional case studies or other exam sections. Manage your time to ensure that you can complete all the exam sections in the time provided. Pay attention to the Exam Progress at the top of the screen so you have sufficient time to complete any exam sections that follow this case study.

To answer the case study questions, you will need to reference information that is provided in the case. Case studies and associated questions might contain exhibits or other resources that provide more information about the scenario described in the case. Information provided in an individual question does not apply to the other questions in the case study.

A Review Screen will appear at the end of this case study. From the Review Screen, you can review and change your answers before you move to the next exam section. After you leave this case study, you will NOT be able to return to it.

To start the case study -

To display the first question in this case study, select the "Next" button. To the left of the question, a menu provides links to information such as business requirements, the existing environment, and problem statements. Please read through all this information before answering any questions. When you are ready to answer a question, select the "Question" button to return to the question.

Overview -

Company Information -

Contoso, Ltd is a multinational retail company that builds, deploys, and manages generative AI and agent-based solutions by using Microsoft Foundry.

Existing Environment -

Identity Environment -

Contoso uses Microsoft Entra ID for identity management, authentication, and authorization capabilities that enable agents to access organizational resources and services.

Contoso recently formed a new AI engineering team named Agent1Dev Team to optimize and maintain existing AI solutions.

The team collaborates with solution architects, DevOps engineers, and security engineers to design, implement, monitor, and secure AI applications.

Contoso also has a team named Agent1Test Team that is responsible for validating AI solutions before the solution deployments.

Generative Environment -

Contoso has a Microsoft Foundry deployment that contains two projects named Project1 and Project2.

Project1 -

Project1 contains a customer support agent named Agent1 that assists customers with product inquiries and troubleshooting requests.

Agent1 has the following configurations:

Agent1 uses a base model deployment.

A safety evaluation pipeline is NOT enabled.

Tool invocation approval workflows are NOT enabled.

Conversation memory constraints are NOT configured.

Agent1 interacts with customers by using digital support channels and answers general questions about Contoso products.

Project1 is deployed to an Azure region located in the European Union (EU).

Agent1Dev Team will use Project1 to optimize and maintain Agent1.

Project2 -

Project2 contains a deployed video generation model. The marketing department at Contoso has access to Project2 and plans to use the model to develop a video creation solution.

Development of the solution is incomplete.

Data Environment -

Contoso stores product-related information in Azure resources that support AI applications.

The Azure environment contains an Azure Blob Storage account named storage1 that stores product detail sheets for all the Contoso products.

The product sheets include specifications, feature descriptions, and product support information that Agent1 can use to answer customer

questions. The product sheets are stored in the PDF format.

Problem Statements -

Contoso identifies the following issues:

Agent1 has only general knowledge of the Contoso products.

A recent chat interaction with Agent1 was analyzed for sentiment. The results of the analysis have NOT been processed yet.

Agent1 does NOT use the detailed product information in the product sheets stored in storage1 when responding to customer questions.

The finance department at Contoso reports that vendor invoices must be reviewed manually to ensure that the invoices match the terms defined in the vendor contracts. The invoices contain tables, logos, and varied layouts that make the documents difficult to process consistently.

Requirements -

Planned Changes -

Contoso plans to implement the following changes:

Implement a solution for Project1 that analyzes the vendor invoices by evaluating both the visual layout and the textual content of the invoices, so that the invoice details can be verified against the vendor contract terms.

Update the base model deployment used by Agent1 and standardize the model version to ensure continuity and consistent responses.

Enable Agent1 to retrieve and use the detailed product information from the product sheets stored in storage1.

Implement an indexing solution for the product sheets that Agent1 can use to answer customer questions.

Complete the development of the video creation solution.

Technical Requirements -

Contoso identifies the following technical requirements:

The model deployment used by Agent1 must support scalable, high-throughput generative AI workloads and dynamically scale to handle variable customer support traffic, without requiring reserved throughput capacity.

The product sheets must be processed by using an indexing pipeline that enables semantic and vector search, so that Agent1 can retrieve the relevant product information.

Responses generated by using the product sheet information must be relevant, complete, and accurate.

Agent1 must be able to use the product sheets to answer natural language questions about product details.

The model version used by Agent1 must remain consistent to ensure stable responses.

The data processed by the model must remain within the EU.

Security and Compliance Requirements

Contoso identifies the following security and compliance requirements:

API keys must NOT be used to access Foundry-deployed models.

Access to the Azure resources must follow the principle of least privilege.

The developers at Contoso must authenticate to Microsoft Foundry resources by using Microsoft Entra authentication.

Access to Project1 must be assigned to the members of Agent1Dev Team by using a security group named SC_Agent1_Dev.

Access to Project1 must be assigned to the members of Agent1Test Team by using a security group named SC_Agent1_Test.

Agent1 must never reveal customer information, even if a document that contains customer data is added erroneously to the product sheet repository in storage1.

The product sheets might contain images that include embedded text. Agent1 must be protected from malicious instructions potentially hidden within the images.

Business Requirements -

Contoso identifies the following business requirements:

Users that interact with Agent1 must have a personalized experience in future interactions, including the ability for Agent1 to retain conversation context and recall relevant information from previous interactions.

Agent1 must answer questions only about the products sold by Contoso.

You need to configure Agent1 to meet the security and compliance requirements.

What should you use?

- A. self-harm content filtering
- B. prompt shields
- C. Personally identifiable information (PII) Detection
- D. violence content filtering

Suggested Answer: *C*

None

You are planning a Microsoft Foundry project named Project1 that will contain multiple agents. Each agent will access the same Azure AI Search resource.

You need to recommend a solution to centrally manage the Azure AI Search credentials within Project1. The solution must be implemented across all the agents.

What should you recommend?

- A. Enable role-based access control (RBAC) for the Azure AI Search resource.
- B. Disable key-based access control on the Azure AI Search resource.
- C. Add a connection to the Azure AI Search resource.
- D. Create a managed private endpoint that connects to the Azure AI Search resource.

Suggested Answer: C

None

HOTSPOT -

Your company is piloting a customer support agent in a Microsoft Foundry project name Project1. Project1 is connected to an existing Application Insights resource, and the company’s support team reviews runs in the Traces tab.

The Foundry Agent Service is configured to perform the following actions:

Retrieve the Application Insights connection string by calling `project_client.telemetry.get_application_insights_connection_string()`.

Call `configure_azure_monitor(connection_string=...)` to enable telemetry.

A separate LangChain service is configured to use OpenTelemetry and has the following configurations:

Uses `AzureAIOpenTelemetryTracer(connection_string=..., enable_content_recording=False)`

Passes the tracer by using `config={"callbacks":[azure_tracer]}`

Company policy has the following requirements:

Telemetry from LangChain and OpenTelemetry must be distinguishable within the same Application Insights resource.

Secrets and credentials must NOT be stored in prompts, tool arguments, or span attributes.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Answer Area

Statements	Yes	No
The LangChain service will appear in Traces without configuring a tracer.	<input type="radio"/>	<input type="radio"/>
Setting different <code>OTEL_SERVICE_NAME</code> values separates the services in Application Insights.	<input type="radio"/>	<input type="radio"/>
When using <code>enable_content_recording=False</code> , prompts and tool data will be captured in the telemetry.	<input type="radio"/>	<input type="radio"/>

Suggested Answer:

Statements	Yes	No
The LangChain service will appear in Traces without configuring a tracer.	<input type="radio"/>	<input checked="" type="radio"/>
Setting different <code>OTEL_SERVICE_NAME</code> values separates the services in Application Insights.	<input checked="" type="radio"/>	<input type="radio"/>
When using <code>enable_content_recording=False</code> , prompts and tool data will be captured in the telemetry.	<input type="radio"/>	<input checked="" type="radio"/>

None

DRAG DROP -

You have a Microsoft Foundry project that processes procurement documents submitted by suppliers.

You need to implement two pipelines by using Azure Content Understanding in Foundry Tools. The solution must meet the following requirements:

Include a pipeline named Pipeline1 that supports cost-effective, high-volume processing of standalone PDF invoices.

Include a pipeline named Pipeline2 that supports cross-document validation by using multi-step reasoning and reference data.

How should you configure each pipeline? To answer, drag the appropriate configurations to the correct pipelines. Each configuration may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Configurations

- ☰ Multi-file task in pro mode
- ☰ Multi-file task in standard mode
- ☰ Single-file task in pro mode
- ☰ Single-file task in standard mode

Answer Area

- Pipeline1:
- Pipeline2:

Suggested Answer:**Answer Area**

- Pipeline1: ☰ Single-file task in standard mode
- Pipeline2: ☰ Multi-file task in pro mode

None

HOTSPOT -

You have a Python application named App1 that integrates with a Microsoft Foundry project named Project1.

You need to ensure that App1 meets the following requirements:

Authenticates by using a Microsoft Entra managed identity

Sends prompts to a deployed model by using the Azure OpenAI Responses API

How should you complete the Python code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```

from azure.identity import DefaultAzureCredential
from azure.ai.projects import AIProjectClient
credential =  ()

project_client = AIProjectClient(
    endpoint="https://contosoai.services.ai.azure.com/api/projects/project1",
    credential=credential,
)
with project_client.get_openai_client() as openai_client:
    response = openai_client.responses.  (

        model="trail-guide-chat",
        input="Create a 3-day hiking itinerary near Seattle.",
    )
print(response.output_text)

```

Answer Area

```

from azure.identity import DefaultAzureCredential
from azure.ai.projects import AIProjectClient
credential =  ()

project_client = AIProjectClient(
    endpoint="https://contosoai.services.ai.azure.com/api/projects/project1",
    credential=credential,
)
with project_client.get_openai_client() as openai_client:
    response = openai_client.responses.  (

        model="trail-guide-chat",
        input="Create a 3-day hiking itinerary near Seattle.",
    )
print(response.output_text)

```

Suggested Answer:

None

HOTSPOT -

You have a Microsoft Foundry project that contains a workflow for a customer support triage process.

You have an Ask a question node that stores user responses in a local variable named Var01.

You need to create the following Power Fx expressions:

An if/else condition expression that ensures that Var01 contains a value

A Send message expression that returns the stored user response in uppercase

How should you configure the expressions? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

If/else condition expression:

IsBlank(Local.Var01)
IsEmpty(Local.Var01)
Not(IsBlank(Local.Var01))

Send message expression:

{Local.Var01}
{Upper(Local.Var01)}
{Upper(Var01)}

Answer Area

If/else condition expression:

IsBlank(Local.Var01)
IsEmpty(Local.Var01)
Not(IsBlank(Local.Var01))

Suggested Answer:

Send message expression:

{Local.Var01}
{Upper(Local.Var01)}
{Upper(Var01)}

None

HOTSPOT -

You have a Microsoft Foundry project that contains a customer support agent built by using the Foundry Agent Service.

The agent uploads user-provided screenshots to Azure Storage through a ticketing tool and receives a blob URL for additional reasoning.

You need to use image moderation during agent runs and prevent harmful content from being returned during runs. Azure AI Content Safety must access the images by using the blob URL. The solution must follow the principle of least privilege.

What should you configure for Content Safety? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Guardrails:

- Select Tool call and set Action to Block.
- Select User input and Output and set Action to Annotate.
- Select User input and Tool response and set Action to Annotate.
- Select User input, Output, Tool response, and Tool call and set Action to Block.

Storage access:

- Storage account access keys
- A user-assigned identity that is assigned the Storage Queue Data Contributor role
- A system-assigned managed identity that is assigned the Storage Blob Data Reader role
- A system-assigned managed identity that is assigned the Storage Blob Data Contributor role

Suggested Answer:

Answer Area

Guardrails:

- Select Tool call and set Action to Block.
- Select User input and Output and set Action to Annotate.
- Select User input and Tool response and set Action to Annotate.
- Select User input, Output, Tool response, and Tool call and set Action to Block.

Storage access:

- Storage account access keys
- A user-assigned identity that is assigned the Storage Queue Data Contributor role
- A system-assigned managed identity that is assigned the Storage Blob Data Reader role
- A system-assigned managed identity that is assigned the Storage Blob Data Contributor role

None

You have a Microsoft Foundry project that contains three agents as shown in the following table.

Name	Description
TriageAgent	Classifies incoming customer requests
PolicyAgent	Answers policy questions by searching internal content
ActionAgent	Creates or updates tickets by calling an HTTP API

You need to orchestrate the agents to ensure that the customer requests meet the following requirements:

Support a deterministic, step-based process that uses conditional branching and shared state across the agents.

Optionally trigger a ticket action based on the triage result.

The solution must minimize development effort.

What should you include in the solution?

- A. a workflow
- B. threads and runs without a workflow
- C. a multi-agent group chat session
- D. separate agent runs coordinated in the application code

Suggested Answer: A

None

You have a Microsoft Foundry project that contains an agent. The agent uses Azure Speech in Foundry Tools. You fine-tune a baseline speech to text model for the en-us locale and publish the model. The agent calls the Speech to text REST API and returns an error message indicating that the project ID is invalid. You need to set the project property to the correct ID. To what should you set the project property?

- A. the project URL
- B. the custom speech project ID
- C. the project ID
- D. the custom speech endpoint URL

Suggested Answer: A

None

HOTSPOT -

You have a Microsoft Foundry project that contains an agent named PaymentAgent.

PaymentAgent includes a function tool that issues customer refunds by using an external API.

You are creating a workflow in YAML.

You need to ensure that the workflow pauses for human approval and continues with the refund step only after approval is granted.

How should you complete the workflow definition? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
steps:
  - id: propose_refund
    type: agent
    agent: PaymentAgent
  - id: approval
    type:
      ask_question
      basic_chat
      data_transformation
  - id: execute_refund
    type: agent
    agent: PaymentAgent
    condition:
      approval == "approved"
      propose_refund.output != null
      true
```

Answer Area

```
steps:
  - id: propose_refund
    type: agent
    agent: PaymentAgent
  - id: approval
    type:
      ask_question
      basic_chat
      data_transformation
  - id: execute_refund
    type: agent
    agent: PaymentAgent
    condition:
      approval == "approved"
      propose_refund.output != null
      true
```

Suggested Answer:

None

You have an Azure Speech in Foundry Tools resource that hosts a custom speech to text model deployed to a custom endpoint. An agent uses the endpoint to perform real-time speech recognition.

You are approaching the expiration date of the custom speech to text model.

What is the expected behavior when the model expires?

- A. Speech recognition requests will return a 4xx error until a new custom model is deployed.
- B. Speech recognition requests will continue to use the expired custom model until the model is removed manually.
- C. Speech recognition requests will fall back to the most recent base model for the same locale.
- D. The custom model will be deleted automatically when the model expires.

Suggested Answer: *C*

None

You have a Microsoft Foundry project that contains a model deployment.
You have an application that calls the deployment by using the Azure OpenAI v1 API and DefaultAzureCredential.
The developers at your company receive HTTP 403 errors when they send inference requests, even after running az login.
You need to ensure that the developers can perform model inference. The solution must follow the principle of least privilege.
Which role-based access control (RBAC) role should you assign to the developers?

- A. Cognitive Services User
- B. Cognitive Services OpenAI User
- C. Contributor
- D. Cognitive Services Data Reader

Suggested Answer: *B*

None

You have a Microsoft Foundry project that contains an agent. The agent has a Model Context Protocol (MCP) tool that queries a knowledge base stored in Azure AI Search.

Some agent runs return answers from the base model without invoking the knowledge base, which results in responses without grounded citations.

You are provided with the following code snippet that runs the agent.

```
run = project_client.agents.runs.create_and_process(  
    thread_id=thread.id,  
    agent_id=agent.id  
)
```

You need to add the correct tool _choice parameter to the code to deterministically force the agent to invoke the MCP tool on each run.

What should you add?

- A. tool_choice={"required"}
- B. tool_choice={"auto"}
- C. tool_choice={"type":"knowledge_base"}
- D. tool_choice ={"type":"mcp"}

Suggested Answer: A

None

DRAG DROP -

You have a Microsoft Foundry project that contains a customer support agent grounded in internal documentation.

After a recent update, users report the following issues:

Some answers are unsupported by retrieved documents.

A small number of responses are flagged for policy violations.

You need to evaluate each issue.

Which observability signals should you use for each issue? To answer, drag the appropriate observability signals to the correct issues. Each observability signal may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Observability signals	Answer Area
<input type="checkbox"/> Groundedness evaluation metrics	Unsupported responses: <input type="checkbox"/> Observability signal
<input type="checkbox"/> Latency breakdown traces	
<input type="checkbox"/> Risk and safety metrics	Policy violations: <input type="checkbox"/> Observability signal
<input type="checkbox"/> Token usage analytics	

Answer Area	
Suggested Answer:	Unsupported responses: <input type="checkbox"/> Groundedness evaluation metrics
	Policy violations: <input type="checkbox"/> Risk and safety metrics

None

You have a Microsoft Foundry project named Project1 that contains an agent. The agent uses an OpenAPI 3.0 specification to call an external weather service.

The weather service requires a key to be passed in an HTTP header. The key value is stored as a connection in Project1.

You need to ensure that the key value from the connection is included automatically whenever the OpenAPI tool is invoked.

What should you configure in the OpenAPI specification?

- A. a header parameter defined for each operation
- B. an Azure Key Vault connection
- C. an API key security scheme
- D. a Bearer token security scheme

Suggested Answer: C

None

You have a Microsoft Foundry project that serves a high-volume chat app. Most requests are simple FAQs, but some require advanced reasoning. You need to reduce costs and latency for common queries, without degrading the quality of the responses to complex questions. What should you do?

- A. Route all the requests to a smaller model.
- B. Use a model cascade that routes the requests to different models.
- C. Increase the value of the max_tokens parameter for all the requests.
- D. Route all the requests to the most capable model.

Suggested Answer: *B*

None

HOTSPOT -

You have a Microsoft Foundry project that contains an internal Q&A agent.

Users report the following issues when they ask the agent questions:

An increase in the following response: "No relevant information found"

Periodic HTTP 429 rate limit exceeded errors during peak hours

You need to identify whether each issue is caused by model unavailability, resource limits, or inference failures.

What should you do? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Metrics to enable:

- Model Availability Rate and Provisioned Utilization
- Only Tokens Cache Match Rate
- Only Total Requests filtered to status code 200
- Time To Response and Total Tokens

Diagnostic log to collect:

- AllMetrics
- audit
- RequestResponse
- trace

Suggested Answer:

Answer Area

Metrics to enable:

- Model Availability Rate and Provisioned Utilization
- Only Tokens Cache Match Rate
- Only Total Requests filtered to status code 200
- Time To Response and Total Tokens

Diagnostic log to collect:

- AllMetrics
- audit
- RequestResponse
- trace

None

You have a Microsoft Foundry project that contains a high-traffic agent.

After a recent update, operational costs increase significantly.

Monitoring confirms that the volume of user traffic to the agent remains unchanged.

You suspect that changes to the request or response characteristics are causing the increase. You need to identify whether the additional costs are driven by the model input size, the model output size, or expanded tool usage.

Which observability capability should you use?

- A. latency
- B. evaluation metrics
- C. run success rate
- D. token usage

Suggested Answer: *D*

None

HOTSPOT -

You have a Microsoft Foundry project that contains an agent.

The agent uses tools to retrieve internal content and call external APIs. The agent is configured to let the model decide when to call the tools.

You need to publish the agent for a compliance workflow. The solution must meet the following requirements:

Each workflow run must include a retrieval step before generating a response.

Tool calls must authenticate by using the published agent's own identity.

Tool access must use an identity isolated from other project resources.

Tool access must use support audit tracing.

What should you do? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Set tool_choice to:

auto
none
required

Configure the tool to authenticate by:

Storing API keys in prompts
Using the shared project agent identity
Using a distinct agent identity bound to the client application**Answer Area**

Set tool_choice to:

auto
none
required

Configure the tool to authenticate by:

Storing API keys in prompts
Using the shared project agent identity
Using a distinct agent identity bound to the client application**Suggested Answer:**

None

You have a Microsoft Foundry project named Project1 that contains the following:

An OpenAPI tool that calls an external API

A project connection named Connection1 that stores the API key of the external API

When an agent calls the OpenAPI tool, the API returns a 401 unauthorized error, and traces show that the API key header is NOT being sent.

You need to ensure that the OpenAPI tool automatically includes the API key from Connection1 on all requests.

What should you do?

- A. Enable identity passthrough so that the tool uses the Microsoft Entra token of the caller.
- B. Add the API key header manually to the OpenAPI specification.
- C. Configure the tool to use the default connection of Project1.
- D. Connect the tool to Connection1.

Suggested Answer: *B*

None

You have a Microsoft Foundry project that contains a customer support agent. The agent calls an internal knowledge API tool before generating responses.

Users report the following issues:

Some requests take more than 15 seconds to complete.

Some responses are incorrect, even when the knowledge API returns the expected data.

You need to inspect individual agent runs to view the ordered sequence of large language model (LLM) calls, tool invocations, and timing information.

Which observability capability should you use?

- A. token usage
- B. monitoring
- C. safety metrics
- D. tracing

Suggested Answer: *D*

None

Note: This section contains one or more sets of questions with the same scenario and problem. Each question presents a unique solution to the problem. You must determine whether the solution meets the stated goals. More than one solution in the set might solve the problem. It is also possible that none of the solutions in the set solve the problem.

After you answer a question in this section, you will NOT be able to return. As a result, these questions do not appear on the Review Screen.

You have a multimodal AI generative model that accepts image uploads and uses extracted image text to generate responses.

You discover that users can upload unsafe images and embed hidden instructions into images to manipulate the model.

You need to implement controls to mitigate the risk.

Solution: You configure a prompt shield for user prompts.

Does this meet the goal?

A. Yes

B. No

Suggested Answer: A

None

Note: This section contains one or more sets of questions with the same scenario and problem. Each question presents a unique solution to the problem. You must determine whether the solution meets the stated goals. More than one solution in the set might solve the problem. It is also possible that none of the solutions in the set solve the problem.

After you answer a question in this section, you will NOT be able to return. As a result, these questions do not appear on the Review Screen.

You have a multimodal AI generative model that accepts image uploads and uses extracted image text to generate responses.

You discover that users can upload unsafe images and embed hidden instructions into images to manipulate the model.

You need to implement controls to mitigate the risk.

Solution: You configure image moderation to block unsafe content before processing the images.

Does this meet the goal?

A. Yes

B. No

Suggested Answer: A

None

Note: This section contains one or more sets of questions with the same scenario and problem. Each question presents a unique solution to the problem. You must determine whether the solution meets the stated goals. More than one solution in the set might solve the problem. It is also possible that none of the solutions in the set solve the problem.

After you answer a question in this section, you will NOT be able to return. As a result, these questions do not appear on the Review Screen.

You have a multimodal AI generative model that accepts image uploads and uses extracted image text to generate responses.

You discover that users can upload unsafe images and embed hidden instructions into images to manipulate the model.

You need to implement controls to mitigate the risk.

Solution: You configure a prompt shield for documents.

Does this meet the goal?

A. Yes

B. No

Suggested Answer: A

None

Note: This section contains one or more sets of questions with the same scenario and problem. Each question presents a unique solution to the problem. You must determine whether the solution meets the stated goals. More than one solution in the set might solve the problem. It is also possible that none of the solutions in the set solve the problem.

After you answer a question in this section, you will NOT be able to return. As a result, these questions do not appear on the Review Screen.

You have a multimodal AI generative model that accepts image uploads and uses extracted image text to generate responses.

You discover that users can upload unsafe images and embed hidden instructions into images to manipulate the model.

You need to implement controls to mitigate the risk.

Solution: You configure protected material detection.

Does this meet the goal?

A. Yes

B. No

Suggested Answer: *B*

None

Case Study -

This is a case study. Case studies are not timed separately from other exam sections. You can use as much exam time as you would like to complete each case study. However, there might be additional case studies or other exam sections. Manage your time to ensure that you can complete all the exam sections in the time provided. Pay attention to the Exam Progress at the top of the screen so you have sufficient time to complete any exam sections that follow this case study.

To answer the case study questions, you will need to reference information that is provided in the case. Case studies and associated questions might contain exhibits or other resources that provide more information about the scenario described in the case. Information provided in an individual question does not apply to the other questions in the case study.

A Review Screen will appear at the end of this case study. From the Review Screen, you can review and change your answers before you move to the next exam section. After you leave this case study, you will NOT be able to return to it.

To start the case study -

To display the first question in this case study, select the "Next" button. To the left of the question, a menu provides links to information such as business requirements, the existing environment, and problem statements. Please read through all this information before answering any questions. When you are ready to answer a question, select the "Question" button to return to the question.

Overview -

Company Information -

Contoso, Ltd is a multinational retail company that builds, deploys, and manages generative AI and agent-based solutions by using Microsoft Foundry.

Existing Environment -

Identity Environment -

Contoso uses Microsoft Entra ID for identity management, authentication, and authorization capabilities that enable agents to access organizational resources and services.

Contoso recently formed a new AI engineering team named Agent1Dev Team to optimize and maintain existing AI solutions.

The team collaborates with solution architects, DevOps engineers, and security engineers to design, implement, monitor, and secure AI applications.

Contoso also has a team named Agent1Test Team that is responsible for validating AI solutions before the solution deployments.

Generative Environment -

Contoso has a Microsoft Foundry deployment that contains two projects named Project1 and Project2.

Project1 -

Project1 contains a customer support agent named Agent1 that assists customers with product inquiries and troubleshooting requests.

Agent1 has the following configurations:

Agent1 uses a base model deployment.

A safety evaluation pipeline is NOT enabled.

Tool invocation approval workflows are NOT enabled.

Conversation memory constraints are NOT configured.

Agent1 interacts with customers by using digital support channels and answers general questions about Contoso products.

Project1 is deployed to an Azure region located in the European Union (EU).

Agent1Dev Team will use Project1 to optimize and maintain Agent1.

Project2 -

Project2 contains a deployed video generation model. The marketing department at Contoso has access to Project2 and plans to use the model to develop a video creation solution.

Development of the solution is incomplete.

Data Environment -

Contoso stores product-related information in Azure resources that support AI applications.

The Azure environment contains an Azure Blob Storage account named storage1 that stores product detail sheets for all the Contoso products.

The product sheets include specifications, feature descriptions, and product support information that Agent1 can use to answer customer

questions. The product sheets are stored in the PDF format.

Problem Statements -

Contoso identifies the following issues:

Agent1 has only general knowledge of the Contoso products.

A recent chat interaction with Agent1 was analyzed for sentiment. The results of the analysis have NOT been processed yet.

Agent1 does NOT use the detailed product information in the product sheets stored in storage1 when responding to customer questions.

The finance department at Contoso reports that vendor invoices must be reviewed manually to ensure that the invoices match the terms defined in the vendor contracts. The invoices contain tables, logos, and varied layouts that make the documents difficult to process consistently.

Requirements -

Planned Changes -

Contoso plans to implement the following changes:

Implement a solution for Project1 that analyzes the vendor invoices by evaluating both the visual layout and the textual content of the invoices, so that the invoice details can be verified against the vendor contract terms.

Update the base model deployment used by Agent1 and standardize the model version to ensure continuity and consistent responses.

Enable Agent1 to retrieve and use the detailed product information from the product sheets stored in storage1.

Implement an indexing solution for the product sheets that Agent1 can use to answer customer questions.

Complete the development of the video creation solution.

Technical Requirements -

Contoso identifies the following technical requirements:

The model deployment used by Agent1 must support scalable, high-throughput generative AI workloads and dynamically scale to handle variable customer support traffic, without requiring reserved throughput capacity.

The product sheets must be processed by using an indexing pipeline that enables semantic and vector search, so that Agent1 can retrieve the relevant product information.

Responses generated by using the product sheet information must be relevant, complete, and accurate.

Agent1 must be able to use the product sheets to answer natural language questions about product details.

The model version used by Agent1 must remain consistent to ensure stable responses.

The data processed by the model must remain within the EU.

Security and Compliance Requirements

Contoso identifies the following security and compliance requirements:

API keys must NOT be used to access Foundry-deployed models.

Access to the Azure resources must follow the principle of least privilege.

The developers at Contoso must authenticate to Microsoft Foundry resources by using Microsoft Entra authentication.

Access to Project1 must be assigned to the members of Agent1Dev Team by using a security group named SC_Agent1_Dev.

Access to Project1 must be assigned to the members of Agent1Test Team by using a security group named SC_Agent1_Test.

Agent1 must never reveal customer information, even if a document that contains customer data is added erroneously to the product sheet repository in storage1.

The product sheets might contain images that include embedded text. Agent1 must be protected from malicious instructions potentially hidden within the images.

Business Requirements -

Contoso identifies the following business requirements:

Users that interact with Agent1 must have a personalized experience in future interactions, including the ability for Agent1 to retain conversation context and recall relevant information from previous interactions.

Agent1 must answer questions only about the products sold by Contoso.

You need to recommend a solution to assess the responses generated by Agent1 when the agent uses the product information stored in storage1.

The solution must meet the technical requirements.

What should you include in the recommendation?

- A. a Retrieval Augmented Generation (RAG) evaluator
- B. a custom guardrail
- C. model fine-tuning
- D. a groundedness evaluator

Suggested Answer: A

None

Case Study -

This is a case study. Case studies are not timed separately from other exam sections. You can use as much exam time as you would like to complete each case study. However, there might be additional case studies or other exam sections. Manage your time to ensure that you can complete all the exam sections in the time provided. Pay attention to the Exam Progress at the top of the screen so you have sufficient time to complete any exam sections that follow this case study.

To answer the case study questions, you will need to reference information that is provided in the case. Case studies and associated questions might contain exhibits or other resources that provide more information about the scenario described in the case. Information provided in an individual question does not apply to the other questions in the case study.

A Review Screen will appear at the end of this case study. From the Review Screen, you can review and change your answers before you move to the next exam section. After you leave this case study, you will NOT be able to return to it.

To start the case study -

To display the first question in this case study, select the "Next" button. To the left of the question, a menu provides links to information such as business requirements, the existing environment, and problem statements. Please read through all this information before answering any questions. When you are ready to answer a question, select the "Question" button to return to the question.

Overview -

Company Information -

Contoso, Ltd is a multinational retail company that builds, deploys, and manages generative AI and agent-based solutions by using Microsoft Foundry.

Existing Environment -

Identity Environment -

Contoso uses Microsoft Entra ID for identity management, authentication, and authorization capabilities that enable agents to access organizational resources and services.

Contoso recently formed a new AI engineering team named Agent1Dev Team to optimize and maintain existing AI solutions.

The team collaborates with solution architects, DevOps engineers, and security engineers to design, implement, monitor, and secure AI applications.

Contoso also has a team named Agent1Test Team that is responsible for validating AI solutions before the solution deployments.

Generative Environment -

Contoso has a Microsoft Foundry deployment that contains two projects named Project1 and Project2.

Project1 -

Project1 contains a customer support agent named Agent1 that assists customers with product inquiries and troubleshooting requests.

Agent1 has the following configurations:

Agent1 uses a base model deployment.

A safety evaluation pipeline is NOT enabled.

Tool invocation approval workflows are NOT enabled.

Conversation memory constraints are NOT configured.

Agent1 interacts with customers by using digital support channels and answers general questions about Contoso products.

Project1 is deployed to an Azure region located in the European Union (EU).

Agent1Dev Team will use Project1 to optimize and maintain Agent1.

Project2 -

Project2 contains a deployed video generation model. The marketing department at Contoso has access to Project2 and plans to use the model to develop a video creation solution.

Development of the solution is incomplete.

Data Environment -

Contoso stores product-related information in Azure resources that support AI applications.

The Azure environment contains an Azure Blob Storage account named storage1 that stores product detail sheets for all the Contoso products.

The product sheets include specifications, feature descriptions, and product support information that Agent1 can use to answer customer

questions. The product sheets are stored in the PDF format.

Problem Statements -

Contoso identifies the following issues:

Agent1 has only general knowledge of the Contoso products.

A recent chat interaction with Agent1 was analyzed for sentiment. The results of the analysis have NOT been processed yet.

Agent1 does NOT use the detailed product information in the product sheets stored in storage1 when responding to customer questions.

The finance department at Contoso reports that vendor invoices must be reviewed manually to ensure that the invoices match the terms defined in the vendor contracts. The invoices contain tables, logos, and varied layouts that make the documents difficult to process consistently.

Requirements -

Planned Changes -

Contoso plans to implement the following changes:

Implement a solution for Project1 that analyzes the vendor invoices by evaluating both the visual layout and the textual content of the invoices, so that the invoice details can be verified against the vendor contract terms.

Update the base model deployment used by Agent1 and standardize the model version to ensure continuity and consistent responses.

Enable Agent1 to retrieve and use the detailed product information from the product sheets stored in storage1.

Implement an indexing solution for the product sheets that Agent1 can use to answer customer questions.

Complete the development of the video creation solution.

Technical Requirements -

Contoso identifies the following technical requirements:

The model deployment used by Agent1 must support scalable, high-throughput generative AI workloads and dynamically scale to handle variable customer support traffic, without requiring reserved throughput capacity.

The product sheets must be processed by using an indexing pipeline that enables semantic and vector search, so that Agent1 can retrieve the relevant product information.

Responses generated by using the product sheet information must be relevant, complete, and accurate.

Agent1 must be able to use the product sheets to answer natural language questions about product details.

The model version used by Agent1 must remain consistent to ensure stable responses.

The data processed by the model must remain within the EU.

Security and Compliance Requirements

Contoso identifies the following security and compliance requirements:

API keys must NOT be used to access Foundry-deployed models.

Access to the Azure resources must follow the principle of least privilege.

The developers at Contoso must authenticate to Microsoft Foundry resources by using Microsoft Entra authentication.

Access to Project1 must be assigned to the members of Agent1Dev Team by using a security group named SC_Agent1_Dev.

Access to Project1 must be assigned to the members of Agent1Test Team by using a security group named SC_Agent1_Test.

Agent1 must never reveal customer information, even if a document that contains customer data is added erroneously to the product sheet repository in storage1.

The product sheets might contain images that include embedded text. Agent1 must be protected from malicious instructions potentially hidden within the images.

Business Requirements -

Contoso identifies the following business requirements:

Users that interact with Agent1 must have a personalized experience in future interactions, including the ability for Agent1 to retain conversation context and recall relevant information from previous interactions.

Agent1 must answer questions only about the products sold by Contoso.

You need to configure Agent1 to answer customer questions about only the Contoso products. The solution must meet the business requirements.

What should you do?

- A. Modify the system message instructions.
- B. Add few-shot examples.
- C. Apply top-p sampling.
- D. Increase the value of the temperature parameter.

Suggested Answer: A

None

You have a Microsoft Foundry project.

You plan to build a customer support solution that contains an agent. The solution must meet the following requirements:

Provide accurate, context-aware responses grounded in internal product documentation stored in Azure AI Search.

Require deep, multi-step reasoning across long contexts.

Generate detailed natural language responses.

Which type of model should you use to power the agent?

- A. a multimodal model
- B. a small language model (SLM)
- C. a key phrase extraction model
- D. a large language model (LLM)

Suggested Answer: *D*

None

DRAG DROP -

You have a Microsoft Foundry project that contains a deployed ticket-triage agent.

You discover that sometimes the agent responds without calling any tools, even when a tool is required.

You need to ensure that the agent calls a tool during execution.

How should you complete the Python code? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Values

⋮ "auto"

⋮ "required"

⋮ "response_format"

⋮ "tool_choice"

⋮ "tools"

⋮ "type"

Answer Area

```
run_payload = {  
    "assistant_id": agent_id,  
    [Value] : [Value] ,  
    "metadata": {  
        "scenario": "ticket-triage"  
    }  
}
```

Answer Area

Suggested Answer:

```
run_payload = {  
    "assistant_id": agent_id,  
    [⋮ "tool_choice"] : [⋮ "required"]  
    "metadata": {  
        "scenario": "ticket-triage"  
    }  
}
```

None

You are building a web app named App1 that generates responses by using a model deployed to a Microsoft Foundry project named Project1. Before sending the prompts to the model, App1 must retrieve documents by using Azure AI Search. You need to integrate Project1 and App1. The solution must meet the following requirements:

- Multiple client applications must use the same search configuration.
- A security policy must prevent key-based authentication.
- Administrative effort must be minimized.

What should you do?

- A. Create a custom HTTP connection in Foundry and manually configure Azure AI Search endpoints per application.
- B. Configure an Azure AI Search connection in Project1 and reference the connection in each application.
- C. Call Azure AI Search directly from each application by using Microsoft Entra authentication.
- D. Enable a managed identity for each application and call Azure AI Search directly.

Suggested Answer: *B*

None