Microsoft

AI-500

Designing and Implementing Multi-Agent AI Solutions

Latest Version: 6.0

Question: 1

You need to define a strategy to meet the business requirements for emergency room visits.
Which workflow node should you add to the Lead Orchestrator workflow?

A. Deliver a message
B. Ask a question
C. Agent
D. Go to

Answer: C

Question: 2

HOTSPOT:
You are validating the outcome of the consultant proposal for the Patient Intake agent.
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 proposed solution resolves the Patient Intake agent issues.	○	○
The proposed solution meets the safety requirements for the Patient Intake agent.	○	○
The proposed solution meets the business requirements for the Patient Intake agent.	○	○

Answer:

Answer Area

Statements	Yes	No
The proposed solution resolves the Patient Intake agent issues.	○	○
The proposed solution meets the safety requirements for the Patient Intake agent.	○	○
The proposed solution meets the business requirements for the Patient Intake agent.	○	○

Question: 3

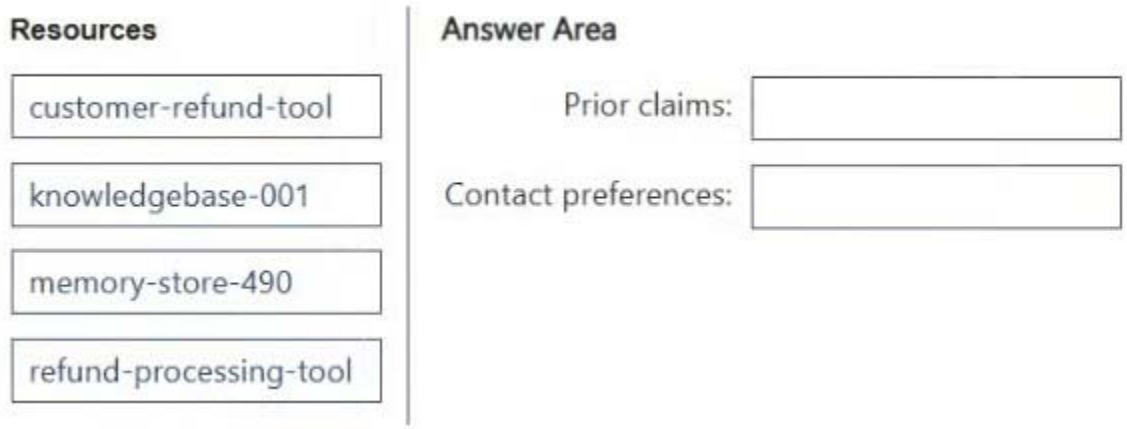
You need to implement the business rule for Claim Approval.
Which workflow node should you use?

A. Ask a question
B. Go to
C. Deliver a message
D. Invoke agent

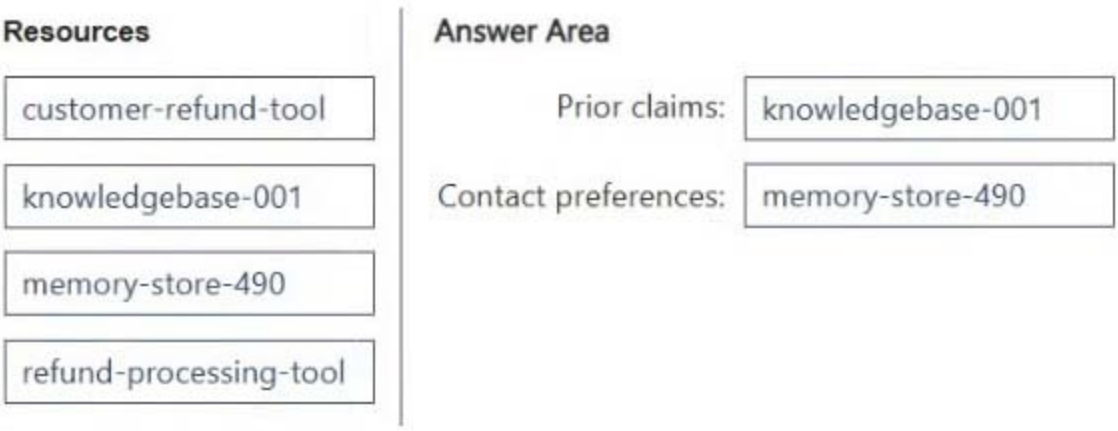
Answer: D

Question: 4

DRAGDROP:
You need to recommend a memory retrieval strategy to support the planned changes for Claim Approval.
What should you recommend? To answer, drag the appropriate resources to the correct requirements. Each resource 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.



Answer:



Question: 5

HOTSPOT:
You have a multi-agent solution in a Microsoft Foundry project. The project is connected to an Application Insights resource.
You have the following code that implements tracing.

```
project = AIProjectClient(
    endpoint=os.environ["AZURE_AI_PROJECT_ENDPOINT"],
    credential=DefaultAzureCredential(),
)

openai_client = project.get_openai_client()
os.environ["AZURE_EXPERIMENTAL_ENABLE_GENAI_TRACING"] = "true"
connection_string = project.telemetry.get_connection_string()
configure_azure_monitor(connection_string=connection_string)

AIProjectInstrumentor().instrument(
    enable_content_recording=False,
    enable_trace_context_propagation=True,
)

@trace_function()
def lookup_customer(email: str, loyalty_tier: str) -> str:
    return f"{email} ({loyalty_tier})"

def main() -> None:
    customer = lookup_customer(email="ada@example.com", loyalty_tier="gold")
    response = openai_client.responses.create(
        model="gpt-5.4",
        input=f"Write a short thank-you note for customer {customer}.",
    )
...
```

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
`responses.create()` will record the prompt text and the model's generated response as a span attribute.	○	○
The stored OpenAI client will inject `traceparent` and `tracestate` headers into the `responses.create()` request.	○	○
The `lookup_customer` span will include `code.function.parameter.email` and `code.function.parameter.loyalty_tier` attributes.	○	○

Answer:

Answer Area

Statements	Yes	No
`responses.create()` will record the prompt text and the model's generated response as a span attribute.	○	■
The stored OpenAI client will inject `traceparent` and `tracestate` headers into the `responses.create()` request.	■	○
The `lookup_customer` span will include `code.function.parameter.email` and `code.function.parameter.loyalty_tier` attributes.	■	○

Question: 6

You have a Microsoft Foundry multi-agent solution that generates an equity research brief based on a given stock ticker symbol. The workflow includes the following agents:

Agent	Description	Duration (seconds)
DataExtraction	Prepares normalized financial data	8
FundamentalAnalysis	Requires the normalized financial data	15
TechnicalAnalysis	Uses only the ticker symbol	30
SentimentAnalysis	Uses only the ticker symbol	12
Summary	Runs after all the analysis outputs are available	6

The model deployment quota supports a maximum of three agents simultaneously.
You discover that the current implementation runs all the agents sequentially and exceeds the response-time target.
You need to reduce the total workflow duration to less than 40 seconds without exceeding the quota.
Which orchestration design should you use?

A. Run DataExtraction, and then run FundamentalAnalysis, TechnicalAnalysis, and SentimentAnalysis concurrently; run Summary after all the analysis outputs complete.
B. Run DataExtraction and TechnicalAnalysis concurrently; start FundamentalAnalysis after DataExtraction completes, and then run SentimentAnalysis after FundamentalAnalysis completes; run Summary after all the analysis outputs complete.
C. Run DataExtraction and SentimentAnalysis concurrently; start FundamentalAnalysis and TechnicalAnalysis after DataExtraction completes; run Summary after all the analysis outputs complete.
D. Run DataExtraction, TechnicalAnalysis, and SentimentAnalysis concurrently; start FundamentalAnalysis after DataExtraction completes; run Summary after all the analysis outputs complete.

Answer: D

Question: 7

HOTSPOT:
You have the following persistence configuration for a Microsoft Foundry multitenant, multi-agent solution.

```
{
  "tenantModel": {
    "teamIdScope": "uniqueWithinTenantOnly"
  },
  "foundryAgentService": {
    "runOptions": {
      "store": false
    },
    "managedMemoryFeature": "notConfigured"
  },
  "sessionState": {
    "tier": "durableWorkflowState",
    "store": "Azure Cosmos DB for NoSQL",
    "checkpointPolicy": "afterEachRun",
    "checkpointSource": "applicationTranscriptBuffer",
    "restorePattern": "createNewThreadAndReplayMessages"
  },
  "sharedTeamState": {
    "store": "Azure Managed Redis",
    "scope": "sharedAcrossAllTenants",
    "keyPattern": "team:{teamId}:artifact:{artifactId}"
  },
  "longTermSemanticMemory": {
    "tier": "durableSemanticMemory",
    "store": "Azure Cosmos DB for NoSQL",
    "partitionKey": "/tenantId"
  }
}
```

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 `sharedTeamState` tier separates cached artifacts for tenants that use the same `teamId` value.	○	○
The `longTermSemanticMemory` tier stores tenant-isolated durable memory outside the service-managed Foundry Memory feature.	○	○
When service-managed history is disabled, the `sessionState` tier can support workflow resumption after an application process restart.	○	○

Answer:

Answer Area

Statements	Yes	No
The `sharedTeamState` tier separates cached artifacts for tenants that use the same `teamId` value.	○	●
The `longTermSemanticMemory` tier stores tenant-isolated durable memory outside the service-managed Foundry Memory feature.	●	○
When service-managed history is disabled, the `sessionState` tier can support workflow resumption after an application process restart.	●	○

Question: 8

You are designing a multitenant software as a service (SaaS) platform that uses multiple agents. Users will send latency-sensitive inference requests to the platform by using a shared API.
Initially, there will be 20 tenants, and the platform will expand to 200 tenants.
You need to identify the compute component for a production agent runtime. The solution must meet the following requirements:
Isolate workloads for each tenant by using containerization.
Dynamically scale based on demand.
Minimize administrative effort.
What should you use?

A. Azure Container Instances
B. Azure Kubernetes Service (AKS)
C. Microsoft Foundry Agent Service
D. GPU-enabled Azure virtual machines

Answer: B

Question: 9

You have a Microsoft Agent Framework workflow processor that receives customer support requests from a queue. Each request is evaluated by three independent Microsoft Foundry agents.

You discover that the current processor dequeues 30 requests at a time and starts all agent runs immediately. During peak load, as many as 90 agent runs execute simultaneously, and the downstream API receives partial participant messages.
A single agent run completes in five seconds at the 95th percentile (95p), and the target throughput is 120 requests per minute.
You need to change the orchestration to ensure that it meets the throughput target and prevents more than 30 agent runs from executing simultaneously. The solution must produce one consolidated downstream payload for each request.
What should you do?

A. Dequeue 10 requests per batch. Start 10 group chat workflows that include the three agents and a manager.
B. Dequeue 10 requests per batch. Start 10 sequential workflows that include the three agents as ordered steps.
C. Dequeue 10 requests per batch. Start 10 ConcurrentBuilder workflows that include the three agents as participants.
D. Dequeue 30 requests per batch. Start a single ConcurrentBuilder workflow that includes 90 agent executors as participants.

Answer: B

Question: 10

You have a Microsoft Foundry multi-agent solution for loan applications. Each agent scores a full application independently and does NOT require output from other agents.
You need to recommend an orchestration pattern that meets the following requirements:
Produces one aggregated recommendation
Preserves independent scoring -
Minimizes end-to-end latency -
Minimize development effort -
What should you recommend?

A. group chat
B. magnetic
C. sequential
D. concurrent

Answer: D

Thank You for Trying Our Product

For More Information – Visit link below:

https://www.examsboost.com/

15 USD Discount Coupon Code:

G74JA8UF

FEATURES

- ✓ 90 Days Free Updates
- ✓ Money Back Pass Guarantee
- ✓ Instant Download or Email Attachment
- ✓ 24/7 Live Chat Support
- ✓ PDF file could be used at any Platform
- ✓ 50,000 Happy Customer