

PRACTICE QUESTIONS

Microsoft

AI-500

Multi-Agent AI Solutions Expert

SHARED BY [u/Witty-Pencil-1943](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 18, 2026

EXAM Microsoft AI-500, Multi-Agent AI Solutions Expert

Shared by [u/Witty-Pencil-1943](#) on aiexamcerts

10 practice questions for the Microsoft AI-500 exam, Multi-Agent AI Solutions Expert, each followed by its answer and a short explanation. More free sets and the full practice exam: aiexamcerts.com/microsoft/ai-500. Spotted a mistake? Email support@aiexamcerts.com with the exam code AI-500 and the question number.

1. A workflow includes a simple classification step that runs thousands of times a day. Which model choice best balances cost and latency for that step?
- A The largest reasoning model for every step, regardless of cost, latency, or whether the step needs reasoning at all
 - B A fine-tuned image generation model
 - C A reasoning model with the longest possible output limit
 - D A smaller, faster model that is sized for simple classification

Answer: D

Matching model capacity to task demand means a small, fast model handles routine classification while larger models are reserved for hard reasoning steps. Using the largest model everywhere inflates cost and latency for no gain. An image generation model cannot classify text, and a long output limit does not make the model faster.

2. Which Azure service is commonly placed in front of MCP tools to apply authentication, rate limits and logging policies?
- A Azure API Management
 - B Azure Blob Storage
 - C Azure Monitor metric export
 - D Azure Batch

Answer: A

Azure API Management is a gateway that applies policies such as authentication, rate limiting and request logging to APIs and can front MCP-exposed endpoints. Blob Storage holds files, Azure Monitor metric export moves telemetry, and Azure Batch runs compute jobs, so none of them enforces these policies on tool traffic.

3. Which state tier holds the turn-by-turn history of one active conversation, and is expected to expire when that session ends?
- A Long-term semantic memory with no expiration, so that session turns are stored forever under the user's record
 - B Session state in a fast key-value store with an expiration
 - C The model weights
 - D The source control repository

Answer: B

Session state is short-lived and per conversation, which suits a fast key-value store with an expiration so that abandoned sessions clear themselves. Long-term semantic memory is meant to outlast sessions. Model weights and source control do not store conversation turns.

4. Agents inside the same solution do not need authentication with each other, because the private network boundary already replaces identity checks.
- A True
 - B False

Answer: B (False)

A network boundary limits who can reach an endpoint, but it does not show which agent is calling or what it is allowed to do. Agent-to-agent calls should authenticate and be authorized like any other service call. The statement is false.

5. Which developer tool gives a team a reproducible environment, including the runtime, extensions and dependencies, for building agents?
- A A shared administrator laptop
 - B A spreadsheet that lists tool versions
 - C Dev containers
 - D A production-only sandbox used for development

Answer: C

Dev containers define the development environment in code, so every developer and CI run starts from the same image with the same tools and extensions. A shared laptop and a spreadsheet of versions drift out of date, and a production sandbox is the wrong place for development work.

6. In a RAG pipeline, the chunk size affects both retrieval precision and how much context each retrieved chunk adds to the prompt.
- A True
 - B False

Answer: A (True)

Larger chunks carry more context but dilute the match, while smaller chunks match precisely but may miss surrounding facts, and each chunk also consumes prompt tokens. The statement is true, so chunk size is a design decision that should be tested.

7. A tool-calling service receives bursts of requests with long idle periods between them. Which compute choice fits best on cost and scaling?
- A A dedicated virtual machine running around the clock
 - B A single hand-managed container with a fixed size
 - C An on-premises tape archive
 - D Azure Functions with consumption-based hosting

Answer: D

Consumption-based Azure Functions scales with demand and bills for executions, which suits bursty tool calls with idle gaps. A dedicated virtual machine pays for idle hours, and a fixed container does not scale to bursts. A tape archive is a storage medium, not a compute host.

8. Which control stops one tenant's agent from retrieving another tenant's stored memory?
- A Tenant-scoped partitions or namespaces with an access check on every read
 - B A single global memory index with no filter on the tenant, so that every agent can search every customer's records
 - C Caching the memory in a browser cookie
 - D Increasing the embedding dimension

Answer: A

Scoping memory to a tenant partition or namespace and checking the tenant on every read enforces isolation at the data layer. A shared index without a tenant filter exposes all memory to every caller. Browser cookies and embedding dimensions do not isolate stored data.

9. A lead agent assigns subtasks to specialist subagents and combines their outputs into one answer. Which orchestration pattern does this describe?
- A Peer-to-peer with no coordinator
 - B Orchestrator-subagent
 - C Single-shot prompting with no tool use
 - D A fine-tuning loop over the training set

Answer: B

In the orchestrator-subagent pattern, one agent plans and delegates subtasks to subagents, then assembles their results. Peer-to-peer has no central coordinator, single-shot prompting makes one call, and a fine-tuning loop changes model weights rather than orchestrating agents.

10. Three independent research agents must run at the same time, and their findings must then be merged into a single report. Which orchestration pattern fits?
- A A sequential chain where each agent waits for the previous one
 - B One agent handling all three research tasks with a very large prompt
 - C A parallel fan-out with a merge step after all branches finish
 - D Manual copy and paste of the three outputs by a person

Answer: C

A parallel fan-out starts the independent agents together and joins their outputs in a merge step, which cuts total duration. A sequential chain runs them one after another, which is slower. A single agent with a very large prompt and manual copying do not use parallel agents.