

PRACTICE QUESTIONS

Microsoft

AI-103

Azure AI Apps and Agents Developer Associate

SHARED BY [u/throwaway_ai103](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	Microsoft AI-103, Azure AI Apps and Agents Developer Associate

Shared by [u/throwaway_ai103](#) on aiexamcerts

10 practice questions for the Microsoft AI-103 exam, Azure AI Apps and Agents Developer Associate, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/microsoft/ai-103. Spotted a mistake? Email support@aiexamcerts.com with the exam code AI-103 and the question number.

1. A team plans an Azure OpenAI solution and needs a model version that is not offered in its preferred region. What is the most appropriate step?

- A Deploy the model into another resource group in the same region
- B Create the AI resource in a region where that model is offered
- C Stay in the same region and lower the temperature
- D Switch the app to a managed identity

Answer: B

Model availability varies by region, so the AI resource has to be created where the model is offered, and dependent services should be placed nearby to limit latency. Lowering the temperature is the tempting fix, but it changes output randomness and cannot make an unavailable model appear. Managed identities and resource groups have no effect on regional availability.

2. A pure vector query in Azure AI Search matches documents by counting shared keywords, rather than by comparing embeddings.

- A True
- B False

Answer: B (False)

A pure vector query compares the query embedding with stored document embeddings and returns the nearest neighbors. Counting shared keywords is what a full-text query does, which is the tempting confusion when comparing query types.

3. Which chat completion parameter controls how random the model output is?

- A max_tokens
- B stop sequences
- C presence_penalty
- D temperature

Answer: D

Temperature scales sampling randomness, and lower values produce more deterministic output. The presence penalty is the tempting distractor because it also shapes wording, but it discourages repeated topics rather than controlling randomness. The max_tokens parameter only caps response length.

4. Prompt Shields in Azure AI Content Safety helps detect user _____ attempts that try to override the system instructions of an app.

Answer: prompt injection (also accepted: prompt-injection, jailbreak)

Prompt Shields inspects user prompts and retrieved documents for jailbreak and indirect injection attempts. Category filtering for hate or violence is the tempting alternative, but it does not look for instruction override. Treat any retrieved text as untrusted input.

5. Which two practices reduce hallucinations in a RAG answer? (Choose two.)

- A Raise the temperature to 1.5 so the model explores more options
- B Remove the grounding instructions from the system prompt
- C Instruct the model to answer only from the retrieved passages and say when they are insufficient
- D Increase max_tokens to 16000 so the model has more room
- E Retrieve the top passages and cite their source IDs in the prompt context

Answer: C, E

Restricting answers to retrieved passages and requiring citations tie each claim to evidence. Raising temperature is the tempting way to get fuller answers, but it increases variability and makes invented details more likely. A larger max_tokens value only allows longer output and does nothing for grounding.

6. A team must turn a product manual PDF into searchable text and extract the embedded images during indexing. Which Azure AI Search feature performs this work?

- A A skillset attached to an indexer
- B A synonym map
- C A scoring profile
- D A data source with a SAS token

Answer: A

A skillset runs AI enrichment steps such as OCR and text extraction as the indexer processes documents. A synonym map is the tempting choice for search quality, but it changes how query terms match and does not extract anything. A data source only tells the indexer where the files are.

7. Which Foundry agent tool lets an agent run short Python code to analyze an uploaded CSV file and produce a chart?
- A Bing grounding
 - B Function calling
 - C Code interpreter
 - D Azure AI Search tool

Answer: C

Code interpreter runs Python in a sandbox against uploaded files and returns outputs such as charts. Function calling is the tempting choice because it also extends an agent, but it runs your own application code rather than a sandbox. Bing grounding and the Azure AI Search tool retrieve content and do not execute analysis code.

8. Microsoft Foundry projects can restrict which users can use a model deployment through Microsoft Entra role assignments.
- A True
 - B False

Answer: A (True)

Role assignments in Entra ID control which users and groups can access a project and its deployments. Shared API keys are the tempting alternative for access control, but they bypass per-user assignments and cannot be audited by person.

9. Which evaluator checks whether the response actually addresses the user's question, rather than drifting into unrelated content?
- A Fluency
 - B Relevance
 - C Groundedness
 - D Token count

Answer: B

Relevance judges how well the response addresses the query. Groundedness is the tempting choice in a RAG app, but it checks support from the context and not whether the answer addresses the question. Fluency and token count do not measure topical fit.

10. A RAG solution using Azure AI Search and Azure OpenAI must keep customer data within a required geography. Which step best meets this requirement?
- A Enable a CDN in front of the web app
 - B Increase the search service replica count
 - C Turn on verbose application logging and send it to a shared mailbox
 - D Deploy the search and OpenAI resources in the required region

Answer: D

Data residency is determined by where each resource is created, so placing the search service and the OpenAI resource in the required region is the control that matters. A CDN is the tempting performance fix, but it caches content and does not decide where data is stored. Replicas and logging affect availability and diagnostics, not residency.