

PRACTICE QUESTIONS

Microsoft

DP-800

Microsoft Certified: SQL AI Developer Associate

SHARED BY [u/throwaway_dp800](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	Microsoft DP-800, Microsoft Certified: SQL AI Developer Associate

Shared by [u/throwaway_dp800](#) on aiexamcerts

10 practice questions for the Microsoft DP-800 exam, Microsoft Certified: SQL AI Developer Associate, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/microsoft/dp-800. Spotted a mistake? Email support@aiexamcerts.com with the exam code DP-800 and the question number.

1. Which prompt instruction most reduces unsupported answers in a retrieval-augmented system?
 - A Answer with as much detail as possible from general knowledge when the passages are thin
 - B Ignore the passages if they seem long
 - C Always answer in exactly 3,000 words
 - D Answer only from the provided passages and say when the answer is not present

Answer: D

Constraining the model to the retrieved passages and allowing it to say the answer is missing reduces invented facts. Encouraging general knowledge invites unsupported claims, and ignoring passages defeats retrieval. A fixed word count does not improve accuracy.

2. Query Store can help identify vector queries whose performance regressed after a plan change.
 - A True
 - B False

Answer: A (True)

Query Store keeps runtime statistics and plan history, so it shows when a query's plan changed and how its duration moved. Administrators can then force a previous plan if it performed better. This makes regressions after statistics updates or deployments visible.

3. Which statement about execution plans is accurate?
 - A An execution plan is stored as a column in the table and is never shown to the user
 - B An estimated plan shows the intended operators, and an actual plan adds runtime row counts
 - C Execution plans cannot be viewed in Azure SQL Database, only in SQL Server
 - D An estimated plan includes actual row counts from production runs

Answer: B

An estimated plan shows the optimizer's intended operators and row estimates without running the query, and an actual plan adds the runtime counts. Plans are not stored as table columns, they can be viewed in Azure SQL Database, and an estimated plan by definition has no actual counts.

4. Azure OpenAI _____ models produce the vectors that are stored in the VECTOR column.

Answer: embedding

Embedding models convert text into fixed-length numeric vectors, and the output size must match the column definition. Chat models generate answers rather than vectors, so they are not used for this column. Teams should record which embedding model produced each vector.

5. Which two statements about combining full-text and vector retrieval in Azure SQL are accurate? (Choose two.)

- A It replaces the need for any embedding column
- B It combines keyword ranking from full-text search with vector similarity
- C It works only when the query and documents share the same words
- D It merges two ranked lists with a fusion method such as reciprocal rank fusion
- E It can run only on documents shorter than 100 characters

Answer: B, D

Hybrid search runs keyword and vector retrieval side by side and merges their rankings, often with reciprocal rank fusion. It still needs embeddings for the vector half. It does not require identical wording and it is not limited to short documents.

6. Which constraint prevents two chunks from the same document from sharing the same chunk number?

- A A check constraint that allows NULL chunk numbers
- B A default of GETDATE() on the chunk number column
- C A unique constraint on DocumentId and ChunkNumber
- D A trigger that silently deletes the duplicate rows

Answer: C

A unique constraint over DocumentId and ChunkNumber enforces one chunk per position in each document. A check constraint that allows nulls does not prevent duplicates, a date default is unrelated to ordering, and a silent delete trigger hides data problems. The constraint makes the rule explicit.

7. What does a vector distance of zero between two embeddings indicate for the chosen metric?

- A The source texts are unrelated in meaning and topic
- B The vectors have zero dimensions
- C The vector index is corrupt
- D The vectors are identical under that metric

Answer: D

A distance of zero means the vectors are identical under the metric, such as cosine, so the texts are as similar as the model can express. It does not say the texts are unrelated, and it does not imply an empty vector or an index fault. Interpretation depends on the model used.

8. A row-level security predicate is an inline table-valued function that returns a row for each record the current user is allowed to see.
- A True
 - B False

Answer: A (True)

A predicate function returns a row when the filter condition is met, and the security policy applies it to queries on the table. It can read session context such as the user identity or tenant. Predicates are central to multi-tenant RAG isolation.

9. A batch embedding job fails halfway through. Which design allows it to resume safely?
- A Track which rows have embeddings and process only missing or stale rows
 - B Restart from the first row and insert duplicate chunks
 - C Delete the whole table and reload every document from the source
 - D Disable the vector column until the job finishes and then re-enable it

Answer: A

Recording the embedding state per row lets the job skip finished work and retry only failed or outdated rows. Restarting from the beginning creates duplicates, and deleting and reloading wastes cost. Disabling the column does not make the job resumable.

10. Which technique reorders the top retrieved candidates with a stronger relevance scorer before they are sent to the model?
- A Sharding
 - B Reranking
 - C Normalization
 - D Tokenization

Answer: B

Reranking applies a more expensive scorer to a small candidate set, which often improves the order of the passages sent to the model. Sharding splits data across databases, normalization restructures tables, and tokenization splits text into model tokens. None of those reorders results.