

PRACTICE QUESTIONS

Databricks Certified Generative AI Engineer Associate

SHARED BY [u/Jazzy-Dolphin-1152](#) on aiexamcerts

EDITION	Free Set 2
---------	------------

QUESTIONS	10
-----------	----

SHARED	October 15, 2026
--------	------------------

EXAM	Databricks Certified Generative AI Engineer Associate
------	---

Shared by [u/Jazzy-Dolphin-1152](#) on aiexamcerts

10 practice questions for the Databricks Certified Generative AI Engineer Associate exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/databricks/databricks-certified-generative-ai-engineer-associate. Spotted a mistake?

Email support@aiexamcerts.com with the exam name and the question number.

1. A notebook job must call an external LLM API that needs a secret key. Which Databricks feature keeps that key out of the notebook code?

- A Delta Sharing recipients
- B Auto Loader checkpoints
- C Databricks secret scopes
- D MLflow run tags

Answer: C

Secret scopes store credentials outside the code and let the job read them at runtime, so the key never appears in the notebook or in version control. Delta Sharing shares data, not API credentials. Auto Loader checkpoints record ingestion progress, and run tags are metadata labels that should never hold secrets.

2. An offline evaluation on a fixed dataset can fully replace monitoring of a deployed LLM application that serves real user traffic.

- A True
- B False

Answer: B (False)

The statement is false. Offline evaluation checks a fixed set before release, but real users ask new questions and inputs drift, so production monitoring of traces and quality signals is still needed. Treating a passing offline score as the end of quality control is the tempting error.

3. Which technique adds a few worked input and output examples to the prompt without changing the model's weights?

- A Few-shot prompting
- B Pretraining from scratch
- C Quantizing the weights to 4-bit precision
- D Fine-tuning the model on the examples

Answer: A

Few-shot prompting places examples in the prompt, so the model itself is unchanged. Pretraining and fine-tuning both update weights, and quantization changes the precision of the weights. Fine-tuning is the tempting distractor because it also uses examples, but it alters the model.

4. Model Serving endpoints can be configured to _____ to zero when idle, which cuts cost at the price of a slower first request after a quiet period.

Answer: scale (also accepted: scale down, scale-down, scaling down)

Scale to zero lets an idle endpoint release its compute, and the next request must wait for it to start again. The trade-off is lower cost against cold-start latency. The blank is the phrase scale down.

5. Which two statements about LLM-as-a-judge evaluation are correct? (Choose two.)
- A A judge's scores should be checked against a sample of human labels before anyone relies on them
 - B A judge that is the same model as the generator never shows any bias
 - C Judge scores replace the need for any ground-truth labels
 - D A rubric with clear criteria makes judge scores more consistent across runs
 - E Judges are useful only for toxicity checks

Answer: A, D

Calibrating a judge against human labels shows whether its scores mean what you think, and a clear rubric reduces run-to-run variation. Using the generator as its own judge invites self-preference bias. Judges can complement ground truth but do not remove the need for it.

6. A product team needs an assistant that answers from a catalog updated hourly and cites its sources. Which architecture is most appropriate?
- A A standalone LLM with no retrieval layer
 - B Retrieval over the catalog index that returns source metadata
 - C A static FAQ page scraped once and embedded permanently in the prompt
 - D A keyword search that returns only the first matching page

Answer: B

Retrieval over a refreshed index keeps answers current, and returning source metadata lets the response cite where each fact came from. A standalone model cannot cite a catalog it has never seen. A scraped, permanently embedded FAQ goes stale within an hour.

7. Which Delta Lake feature lets you query a table as it looked before a bad batch was written?
- A Time travel with VERSION AS OF
 - B Z-ordering on the filtered column
 - C Liquid clustering on the key columns
 - D Bloom filter indexes on the key column

Answer: A

Time travel reads earlier table versions from the transaction log, so the previous state can be queried or restored. Z-ordering and liquid clustering reorganize files to speed up reads and do not recover data. Bloom filters help skip files during a read.

8. The AI Playground in Databricks is a place to chat with and compare foundation models interactively before building an application.

A True
B False

Answer: A (True)

The statement is true. AI Playground lets a builder try prompts against one or more models in the workspace UI, which helps choose a model before any code is written. It is a prototyping surface, not a production endpoint.

9. Which statement about Delta Lake schema enforcement is correct?

A Writes with columns that do not match the table schema are silently dropped
B Schema enforcement is off by default and requires installing a separate library to enable it
C A write that does not match the table schema is rejected unless evolution is enabled
D The schema is checked only when the table is read, never when it is written

Answer: C

Delta Lake validates writes against the table schema and rejects mismatches by default, and schema evolution must be turned on deliberately. Silent dropping of columns is the tempting but dangerous behavior that Delta avoids. Checking only on read would let bad data land in the table.

10. Which approach best measures how often a chatbot's answers are grounded in retrieved text, across a labeled test set?

A Running an offline evaluation with groundedness and correctness judges on a labeled dataset
B Reading the cluster autoscaling logs
C Changing the schema hints used by Auto Loader
D Applying column masks to the evaluation tables and reviewing only the visible rows

Answer: A

An offline run with groundedness and correctness judges scores each answer against its retrieved context across the whole labeled set, which gives a repeatable measure. Autoscaling logs describe compute behavior. Schema hints and column masks have nothing to do with whether answers are supported.