

PRACTICE QUESTIONS

# Databricks Certified Generative AI Engineer Associate

SHARED BY [u/fuzzy\\_seal\\_42](#) on aiexamcerts

---

EDITION Free Set 1

---

QUESTIONS 10

---

SHARED October 15, 2026

---

EXAM Databricks Certified Generative AI Engineer Associate

---

Shared by [u/fuzzy\\_seal\\_42](#) 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](https://www.aiexamcerts.com/databricks/databricks-certified-generative-ai-engineer-associate). Spotted a mistake? Email [support@aiexamcerts.com](mailto:support@aiexamcerts.com) with the exam name and the question number.

1. Which chunking strategy is most sensible for long policy manuals when the goal is to retrieve the exact paragraph that answers a question?
  - A One chunk per entire document so that nothing is ever split
  - B Single-word chunks, so every term is indexed separately
  - C Whole chapters as single chunks regardless of their length
  - D Chunks of a few hundred tokens with a small overlap

**Answer: D**

Moderate chunks stay focused enough to match a specific question, and a small overlap keeps sentences that straddle a boundary retrievable. Whole documents or chapters dilute the embedding and waste context window. Single-word chunks lose the sentence-level meaning that embeddings depend on.

2. In Unity Catalog, a table is addressed with a three-level name, catalog. \_\_\_\_\_.table\_name, where the middle level is the schema.

**Answer: schema**

Unity Catalog uses a three-level namespace of catalog, schema and table, so the middle part of the name is the schema. Confusing the middle level with the catalog would point a query at the wrong container. The blank is therefore the schema name.

3. A chatbot must answer from a product catalog that changes daily. Which pattern keeps answers current without retraining the LLM?
  - A Keep the LLM fixed and refresh the vector index from the source Delta table
  - B Fine-tune the LLM on the catalog once a year
  - C Paste the entire catalog into every prompt
  - D Hard-code current product facts into the system prompt and redeploy on each change

**Answer: A**

Refreshing the retrieval index keeps knowledge current while the model stays fixed, and a Delta Sync index can follow a Delta table with change data feed enabled. Yearly fine-tuning leaves daily changes stale. Pasting a full catalog exceeds the context window, and hard-coded prompts need a redeploy for every change.

4. Row filters and column masks in Unity Catalog are two ways to restrict sensitive data returned to a user without creating separate copies of the table.
- A True
  - B False

**Answer: A** (True)

The statement is true. A row filter limits which rows a principal sees, and a column mask transforms column values for a principal, both applied at query time. Neither requires maintaining a duplicate table.

5. Which two actions keep a vector index in sync with a source table that receives frequent updates? (Choose two.)
- A Enable Change Data Feed on the source Delta table so that a Delta Sync index can read the row changes
  - B Export the table to CSV once a month and load the index from the file by hand
  - C Drop and recreate the source table each night under the same name
  - D Use a Delta Sync index, which syncs from the source table through a continuous or triggered pipeline, instead of rebuilding it manually
  - E Keep the embeddings in a notebook variable that is reloaded on demand

**Answer: A, D**

Change Data Feed records row-level changes that a Delta Sync index can consume, and a Delta Sync index keeps itself updated through a managed pipeline. Monthly exports and manual reloads leave the index stale between runs. Dropping and recreating the source table breaks its history and the index that depends on it.

6. An extraction task must return the same fields in the same format on every run. Which setting and prompt combination is most likely to give repeatable output?
- A A high temperature with top\_p close to 1 so fields are sampled widely
  - B A low temperature and a prompt that specifies the exact JSON schema
  - C A new random seed on every request
  - D A much larger maximum output length

**Answer: B**

A low temperature reduces sampling randomness, and an explicit schema constrains the structure of the reply. A high temperature increases variation between runs. A larger output limit controls how long the answer can be, not whether its structure is consistent.

7. Which tool declares a project's jobs, pipelines and deployment settings in a `databricks.yml` file so the same project can be deployed to dev and prod targets?
- A MLflow Tracing
  - B Unity Catalog volumes
  - C Delta Live Tables expectations
  - D Databricks Asset Bundles

**Answer: D**

Databricks Asset Bundles describe resources in `databricks.yml` and deploy them with the Databricks CLI to named targets. Delta Live Tables expectations check row-level data quality rather than deployment. Unity Catalog volumes govern file storage and are not a deployment tool.

8. Auto Loader stores its progress in a \_\_\_\_\_ location, so a restarted stream does not reprocess files it has already loaded.

**Answer: checkpoint** (also accepted: checkpoint directory, checkpointLocation)

The checkpoint records which files have been processed, which is what makes restarts incremental. Without it, a stream would have to rescan the whole landing zone. The blank is the word checkpoint.

9. A RAG app answers questions about a 2,000-page manual, but users say the answers miss relevant sections. Which change should you try first?
- A Switch to a larger generator LLM
  - B Tune the chunking and top-k settings and re-measure retrieval
  - C Lower the temperature to zero and keep the retrieval settings
  - D Increase max\_tokens so the answer can include more sections

**Answer: B**

Missed sections point to a retrieval problem, so chunking and top-k settings address the cause directly, and re-measuring retrieval shows whether the change helped. A larger generator cannot answer from passages it never receives. Changing the temperature or the output length does not bring the missing sections into the prompt.

10. What does a drift metric in Lakehouse Monitoring typically show?
- A Which user last edited a notebook in the workspace
  - B How a column's value distribution has shifted compared with a baseline
  - C How many clusters are currently running in the workspace region
  - D The size of each Delta file in megabytes on disk

**Answer: B**

Drift metrics compare the distribution of values over time, which is how changes in input data are detected. The other options describe operational facts, not shifts in data distributions. File sizes are a storage detail that drift analysis does not use.