

# Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

<https://www.2passeasy.com/dumps/Databricks-Generative-AI-Engineer-Associate/>



**NEW QUESTION 1**

A Generative AI Engineer at an automotive company would like to build a question- answering chatbot for customers to inquire about their vehicles. They have a database containing various documents of different vehicle makes, their hardware parts, and common maintenance information. Which of the following components will NOT be useful in building such a chatbot?

- A. Response-generating LLM
- B. Invite users to submit long, rather than concise, questions
- C. Vector database
- D. Embedding model

**Answer: B**

**NEW QUESTION 2**

A Generative AI Engineer is developing a RAG system for their company to perform internal document Q&A for structured HR policies, but the answers returned are frequently incomplete and unstructured. It seems that the retriever is not returning all relevant context. The Generative AI Engineer has experimented with different embedding and response generating LLMs but that did not improve results. Which TWO options could be used to improve the response quality? Choose 2 answers.

- A. Add the section header as a prefix to chunks
- B. Increase the document chunk size
- C. Split the document by sentence
- D. Use a larger embedding model
- E. Fine tune the response generation model

**Answer: AB**

**NEW QUESTION 3**

A Generative AI Engineer is building a production-ready LLM system which replies directly to customers. The solution makes use of the Foundation Model API via provisioned throughput. They are concerned that the LLM could potentially respond in a toxic or otherwise unsafe way. They also wish to perform this with the least amount of effort. Which approach will do this?

- A. Host Llama Guard on Foundation Model API and use it to detect unsafe responses
- B. Add some LLM calls to their chain to detect unsafe content before returning text
- C. Add a regex expression on inputs and outputs to detect unsafe responses.
- D. Ask users to report unsafe responses

**Answer: A**

**NEW QUESTION 4**

A Generative AI Engineer has been asked to design an LLM-based application that accomplishes the following business objective: answer employee HR questions using HR PDF documentation. Which set of high level tasks should the Generative AI Engineer's system perform?

- A. Calculate averaged embeddings for each HR document, compare embeddings to user query to find the best document
- B. Pass the best document with the user query into an LLM with a large context window to generate a response to the employee.
- C. Use an LLM to summarize HR documentation
- D. Provide summaries of documentation and user query into an LLM with a large context window to generate a response to the user.
- E. Create an interaction matrix of historical employee questions and HR documentation
- F. Use ALS to factorize the matrix and create embedding
- G. Calculate the embeddings of new queries and use them to find the best HR documentation
- H. Use an LLM to generate a response to the employee question based upon the documentation retrieved.
- I. Split HR documentation into chunks and embed into a vector store
- J. Use the employee question to retrieve best matched chunks of documentation, and use the LLM to generate a response to the employee based upon the documentation retrieved.

**Answer: D**

**NEW QUESTION 5**

A Generative AI Engineer is using the code below to test setting up a vector store:

```
from databricks.vector_search.client import VectorSearchClient

vsc = VectorSearchClient()

vsc.create_endpoint(
    name="vector_search_test",
    endpoint_type="STANDARD"
)
```

Assuming they intend to use Databricks managed embeddings with the default embedding model, what should be the next logical function call?

- A. vsc.get\_index()
- B. vsc.create\_delta\_sync\_index()
- C. vsc.create\_direct\_access\_index()
- D. vsc.similarity\_search()

**Answer:** B

#### NEW QUESTION 6

A Generative AI Engineer has been asked to build an LLM-based question-answering application. The application should take into account new documents that are frequently published. The engineer wants to build this application with the least cost and least development effort and have it operate at the lowest cost possible.

Which combination of chaining components and configuration meets these requirements?

- A. For the application a prompt, a retriever, and an LLM are required.
- B. The retriever output is inserted into the prompt which is given to the LLM to generate answers.
- C. The LLM needs to be frequently updated with the new documents in order to provide most up-to-date answers.
- D. For the question-answering application, prompt engineering and an LLM are required to generate answers.
- E. For the application a prompt, an agent and a fine-tuned LLM are required.
- F. The agent is used by the LLM to retrieve relevant content that is inserted into the prompt which is given to the LLM to generate answers.

**Answer:** A

#### NEW QUESTION 7

A Generative AI Engineer developed an LLM application using the provisioned throughput Foundation Model API. Now that the application is ready to be deployed, they realize their volume of requests are not sufficiently high enough to create their own provisioned throughput endpoint. They want to choose a strategy that ensures the best cost-effectiveness for their application.

What strategy should the Generative AI Engineer use?

- A. Switch to using External Models instead
- B. Deploy the model using pay-per-token throughput as it comes with cost guarantees
- C. Change to a model with a fewer number of parameters in order to reduce hardware constraint issues
- D. Throttle the incoming batch of requests manually to avoid rate limiting issues

**Answer:** B

#### NEW QUESTION 8

A Generative AI Engineer is developing a chatbot designed to assist users with insurance-related queries. The chatbot is built on a large language model (LLM) and is conversational. However, to maintain the chatbot's focus and to comply with company policy, it must not provide responses to questions about politics.

Instead, when presented with political inquiries, the chatbot should respond with a standard message:

??Sorry, I cannot answer that. I am a chatbot that can only answer questions around insurance.??

Which framework type should be implemented to solve this?

- A. Safety Guardrail
- B. Security Guardrail
- C. Contextual Guardrail
- D. Compliance Guardrail

**Answer:** A

#### NEW QUESTION 9

A Generative AI Engineer has developed an LLM application to answer questions about internal company policies. The Generative AI Engineer must ensure that the application doesn't hallucinate or leak confidential data.

Which approach should NOT be used to mitigate hallucination or confidential data leakage?

- A. Add guardrails to filter outputs from the LLM before it is shown to the user
- B. Fine-tune the model on your data, hoping it will learn what is appropriate and not
- C. Limit the data available based on the user's access level
- D. Use a strong system prompt to ensure the model aligns with your needs.

**Answer:** B

#### NEW QUESTION 10

A small and cost-conscious startup in the cancer research field wants to build a RAG application using Foundation Model APIs.

Which strategy would allow the startup to build a good-quality RAG application while being cost-conscious and able to cater to customer needs?

- A. Limit the number of relevant documents available for the RAG application to retrieve from
- B. Pick a smaller LLM that is domain-specific
- C. Limit the number of queries a customer can send per day
- D. Use the largest LLM possible because that gives the best performance for any general queries

**Answer:** B

#### NEW QUESTION 10

What is an effective method to preprocess prompts using custom code before sending them to an LLM?

- A. Directly modify the LLM's internal architecture to include preprocessing steps
- B. It is better not to introduce custom code to preprocess prompts as the LLM has not been trained with examples of the preprocessed prompts

- C. Rather than preprocessing prompts, it's more effective to postprocess the LLM outputs to align the outputs to desired outcomes
- D. Write a MLflow PyFunc model that has a separate function to process the prompts

**Answer:** D

### NEW QUESTION 13

A Generative AI Engineer is responsible for developing a chatbot to enable their company's internal HelpDesk Call Center team to more quickly find related tickets and provide resolution. While creating the GenAI application work breakdown tasks for this project, they realize they need to start planning which data sources (either Unity Catalog volume or Delta table) they could choose for this application. They have collected several candidate data sources for consideration:

call\_rep\_history: a Delta table with primary keys representative\_id, call\_id. This table is maintained to calculate representatives' call resolution from fields call\_duration and call\_start\_time.

transcript Volume: a Unity Catalog Volume of all recordings as \*.wav files, but also a text transcript as \*.txt files.

call\_cust\_history: a Delta table with primary keys customer\_id, call\_id. This table is maintained to calculate how much internal customers use the HelpDesk to make sure that the charge back model is consistent with actual service use.

call\_detail: a Delta table that includes a snapshot of all call details updated hourly. It includes root\_cause and resolution fields, but those fields may be empty for calls that are still active.

maintenance\_schedule – a Delta table that includes a listing of both HelpDesk application outages as well as planned upcoming maintenance downtimes.

They need sources that could add context to best identify ticket root cause and resolution. Which TWO sources do that? (Choose two.)

- A. call\_cust\_history
- B. maintenance\_schedule
- C. call\_rep\_history
- D. call\_detail
- E. transcript Volume

**Answer:** DE

### NEW QUESTION 17

.....

## THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual Databricks-Generative-AI-Engineer-Associate Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the Databricks-Generative-AI-Engineer-Associate Product From:

<https://www.2passeasy.com/dumps/Databricks-Generative-AI-Engineer-Associate/>

## Money Back Guarantee

### **Databricks-Generative-AI-Engineer-Associate Practice Exam Features:**

- \* Databricks-Generative-AI-Engineer-Associate Questions and Answers Updated Frequently
- \* Databricks-Generative-AI-Engineer-Associate Practice Questions Verified by Expert Senior Certified Staff
- \* Databricks-Generative-AI-Engineer-Associate Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- \* Databricks-Generative-AI-Engineer-Associate Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year