

New Databricks-Generative-AI-Engineer-Associate Test Discount & Exam Databricks-Generative-AI-Engineer-Associate Course



BTW, DOWNLOAD part of PassTorrent Databricks-Generative-AI-Engineer-Associate dumps from Cloud Storage:
<https://drive.google.com/open?id=1o8A4--tSVLkcMqgSu8jpBQpCb-EwplyY>

The Databricks Databricks-Generative-AI-Engineer-Associate practice test questions prep material has actual Databricks Databricks-Generative-AI-Engineer-Associate exam questions for our customers so they don't face any hurdles while preparing for Databricks Certified Generative AI Engineer Associate (Databricks-Generative-AI-Engineer-Associate) certification exam. The study material is made by professionals while thinking about our users. We have made the product user-friendly so it will be an easy-to-use learning material. We even guarantee our users that if they couldn't pass the Databricks Databricks-Generative-AI-Engineer-Associate Certification Exam on the first try with their efforts, they can claim a full refund of their payment from us (terms and conditions apply).

We know that it is hard to stay and study for the Databricks Certified Generative AI Engineer Associate (Databricks-Generative-AI-Engineer-Associate) exam dumps in one place for a long time. Therefore, you have the option to use Databricks Certified Generative AI Engineer Associate (Databricks-Generative-AI-Engineer-Associate) PDF questions anywhere and anytime. PassTorrent Databricks Certified Generative AI Engineer Associate (Databricks-Generative-AI-Engineer-Associate) dumps are designed according to the Databricks Databricks-Generative-AI-Engineer-Associate certification exam standard and have hundreds of questions similar to the actual Databricks Certified Generative AI Engineer Associate (Databricks-Generative-AI-Engineer-Associate) exam.

>> **New Databricks-Generative-AI-Engineer-Associate Test Discount** <<

Valid Databricks-Generative-AI-Engineer-Associate Real Practice Materials - Databricks-Generative-AI-Engineer-Associate Actual Exam Dumps - PassTorrent

Databricks-Generative-AI-Engineer-Associate practice questions are stable and reliable exam questions provider for person who need them for their exam. We have been staying and growing in the market for a long time, and we will be here all the time, because the excellent quality and high pass rate of our Databricks-Generative-AI-Engineer-Associate training braindump. As for the safe environment and effective product, there are thousands of candidates are willing to choose our Databricks-Generative-AI-Engineer-Associate study guide, why don't you have a try for our Databricks-Generative-AI-Engineer-Associate study material, never let you down!

Databricks Certified Generative AI Engineer Associate Sample Questions (Q28-Q33):

NEW QUESTION # 28

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

Answer: D,E

Explanation:

The problem describes a Retrieval-Augmented Generation (RAG) system for HR policy Q&A where responses are incomplete and unstructured due to the retriever failing to return sufficient context. The engineer has already tried different embedding and response-generating LLMs without success, suggesting the issue lies in the retrieval process—specifically, how documents are chunked and indexed. Let's evaluate the options.

* Option A: Add the section header as a prefix to chunks

* Adding section headers provides additional context to each chunk, helping the retriever understand the chunk's relevance within the document structure (e.g., "Leave Policy: Annual Leave" vs. just "Annual Leave"). This can improve retrieval precision for structured HR policies.

* Databricks Reference: "Metadata, such as section headers, can be appended to chunks to enhance retrieval accuracy in RAG systems" ("Databricks Generative AI Cookbook," 2023).

* Option B: Increase the document chunk size

* Larger chunks include more context per retrieval, reducing the chance of missing relevant information split across smaller chunks. For structured HR policies, this can ensure entire sections or rules are retrieved together.

* Databricks Reference: "Increasing chunk size can improve context completeness, though it may trade off with retrieval specificity" ("Building LLM Applications with Databricks").

* Option C: Split the document by sentence

* Splitting by sentence creates very small chunks, which could exacerbate the problem by fragmenting context further. This is likely why the current system fails—it retrieves incomplete snippets rather than cohesive policy sections.

* Databricks Reference: No specific extract opposes this, but the emphasis on context completeness in RAG suggests smaller chunks worsen incomplete responses.

* Option D: Use a larger embedding model

* A larger embedding model might improve vector quality, but the question states that experimenting with different embedding models didn't help. This suggests the issue isn't embedding quality but rather chunking/retrieval strategy.

* Databricks Reference: Embedding models are critical, but not the focus when retrieval context is the bottleneck.

* Option E: Fine tune the response generation model

* Fine-tuning the LLM could improve response coherence, but if the retriever doesn't provide complete context, the LLM can't generate full answers. The root issue is retrieval, not generation.

* Databricks Reference: Fine-tuning is recommended for domain-specific generation, not retrieval fixes ("Generative AI Engineer Guide").

Conclusion: Options A and B address the retrieval issue directly by enhancing chunk context—either through metadata (A) or size (B)—aligning with Databricks' RAG optimization strategies. C would worsen the problem, while D and E don't target the root cause given prior experimentation.

NEW QUESTION # 29

A Generative AI Engineer is helping a cinema extend its website's chat bot to be able to respond to questions about specific showtimes for movies currently playing at their local theater. They already have the location of the user provided by location services to their agent, and a Delta table which is continually updated with the latest showtime information by location. They want to implement this new capability in their RAG application.

Which option will do this with the least effort and in the most performant way?

- **A. Create a Feature Serving Endpoint from a FeatureSpec that references an online store synced from the Delta table. Query the Feature Serving Endpoint as part of the agent logic / tool implementation.**
- B. Query the Delta table directly via a SQL query constructed from the user's input using a text-to-SQL LLM in the agent logic / tool
- C. implementation. Write the Delta table contents to a text column, then embed those texts using an embedding model and

store these in the vector index. Look up the information based on the embedding as part of the agent logic / tool implementation.

- D. Set up a task in Databricks Workflows to write the information in the Delta table periodically to an external database such as MySQL and query the information from there as part of the agent logic / tool implementation.

Answer: A

Explanation:

The task is to extend a cinema chatbot to provide movie showtime information using a RAG application, leveraging user location and a continuously updated Delta table, with minimal effort and high performance.

Let's evaluate the options.

* Option A: Create a Feature Serving Endpoint from a FeatureSpec that references an online store synced from the Delta table.

Query the Feature Serving Endpoint as part of the agent logic / tool implementation

* Databricks Feature Serving provides low-latency access to real-time data from Delta tables via an online store. Syncing the Delta table to a Feature Serving Endpoint allows the chatbot to query showtimes efficiently, integrating seamlessly into the RAG agent's tool logic. This leverages Databricks' native infrastructure, minimizing effort and ensuring performance.

* Databricks Reference: "Feature Serving Endpoints provide real-time access to Delta table data with low latency, ideal for production systems" ("Databricks Feature Engineering Guide," 2023).

* Option B: Query the Delta table directly via a SQL query constructed from the user's input using a text-to-SQL LLM in the agent logic / tool

* Using a text-to-SQL LLM to generate queries adds complexity (e.g., ensuring accurate SQL generation) and latency (LLM inference + SQL execution). While feasible, it's less performant and requires more effort than a pre-built serving solution.

* Databricks Reference: "Direct SQL queries are flexible but may introduce overhead in real-time applications" ("Building LLM Applications with Databricks").

* Option C: Write the Delta table contents to a text column, then embed those texts using an embedding model and store these in the vector index. Look up the information based on the embedding as part of the agent logic / tool implementation

* Converting structured Delta table data (e.g., showtimes) into text, embedding it, and using vector search is inefficient for structured lookups. It's effort-intensive (preprocessing, embedding) and less precise than direct queries, undermining performance.

* Databricks Reference: "Vector search excels for unstructured data, not structured tabular lookups" ("Databricks Vector Search Documentation").

* Option D: Set up a task in Databricks Workflows to write the information in the Delta table periodically to an external database such as MySQL and query the information from there as part of the agent logic / tool implementation

* Exporting to an external database (e.g., MySQL) adds setup effort (workflow, external DB management) and latency (periodic updates vs. real-time). It's less performant and more complex than using Databricks' native tools.

* Databricks Reference: "Avoid external systems when Delta tables provide real-time data natively" ("Databricks Workflows Guide").

Conclusion: Option A minimizes effort by using Databricks Feature Serving for real-time, low-latency access to the Delta table, ensuring high performance in a production-ready RAG chatbot.

NEW QUESTION # 30

A Generative AI Engineer has a provisioned throughput model serving endpoint as part of a RAG application and would like to monitor the serving endpoint's incoming requests and outgoing responses. The current approach is to include a micro-service in between the endpoint and the user interface to write logs to a remote server.

Which Databricks feature should they use instead which will perform the same task?

- **A. Inference Tables**
- B. Vector Search
- C. DBSQL
- D. Lakeview

Answer: A

Explanation:

Problem Context: The goal is to monitor the serving endpoint for incoming requests and outgoing responses in a provisioned throughput model serving endpoint within a Retrieval-Augmented Generation (RAG) application. The current approach involves using a microservice to log requests and responses to a remote server, but the Generative AI Engineer is looking for a more streamlined solution within Databricks.

Explanation of Options:

* Option A: Vector Search: This feature is used to perform similarity searches within vector databases.

It doesn't provide functionality for logging or monitoring requests and responses in a serving endpoint, so it's not applicable here.

* Option B: Lakeview: Lakeview is not a feature relevant to monitoring or logging request-response cycles for serving endpoints. It

might be more related to viewing data in Databricks Lakehouse but doesn't fulfill the specific monitoring requirement.

* Option C: DBSQL: Databricks SQL (DBSQL) is used for running SQL queries on data stored in Databricks, primarily for analytics purposes. It doesn't provide the direct functionality needed to monitor requests and responses in real-time for an inference endpoint.

* Option D: Inference Tables: This is the correct answer. Inference Tables in Databricks are designed to store the results and metadata of inference runs. This allows the system to log incoming requests and outgoing responses directly within Databricks, making it an ideal choice for monitoring the behavior of a provisioned serving endpoint. Inference Tables can be queried and analyzed, enabling easier monitoring and debugging compared to a custom microservice.

Thus, Inference Tables are the optimal feature for monitoring request and response logs within the Databricks infrastructure for a model serving endpoint.

NEW QUESTION # 31

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_rep_history
- B. transcript Volume
- C. call_cust_history
- D. maintenance_schedule
- E. call_detail

Answer: B,E

Explanation:

In the context of developing a chatbot for a company's internal HelpDesk Call Center, the key is to select data sources that provide the most contextual and detailed information about the issues being addressed. This includes identifying the root cause and suggesting resolutions. The two most appropriate sources from the list are:

* Call Detail (Option D):

* Contents: This Delta table includes a snapshot of all call details updated hourly, featuring essential fields like root_cause and resolution.

* Relevance: The inclusion of root_cause and resolution fields makes this source particularly valuable, as it directly contains the information necessary to understand and resolve the issues discussed in the calls. Even if some records are incomplete, the data provided is crucial for a chatbot aimed at speeding up resolution identification.

* Transcript Volume (Option E):

* Contents: This Unity Catalog Volume contains recordings in .wav format and text transcripts in .txt files.

* Relevance: The text transcripts of call recordings can provide in-depth context that the chatbot can analyze to understand the nuances of each issue. The chatbot can use natural language processing techniques to extract themes, identify problems, and suggest resolutions based on previous similar interactions documented in the transcripts.

Why Other Options Are Less Suitable:

* A (Call Cust History): While it provides insights into customer interactions with the HelpDesk, it focuses more on the usage metrics rather than the content of the calls or the issues discussed.

* B (Maintenance Schedule): This data is useful for understanding when services may not be available but does not contribute directly to resolving user issues or identifying root causes.

* C (Call Rep History): Though it offers data on call durations and start times, which could help in assessing performance, it lacks direct information on the issues being resolved.

Therefore, Call Detail and Transcript Volume are the most relevant data sources for a chatbot designed to assist with identifying and resolving issues in a HelpDesk Call Center setting, as they provide direct and contextual information related to customer issues.

NEW QUESTION # 32

Which indicator should be considered to evaluate the safety of the LLM outputs when qualitatively assessing LLM responses for a translation use case?

- A. The ability to generate responses in code
- B. The latency of the response and the length of text generated
- C. The similarity to the previous language
- **D. The accuracy and relevance of the responses**

Answer: D

Explanation:

* Problem Context: When assessing the safety and effectiveness of LLM outputs in a translation use case, it is essential to ensure that the translations accurately and relevantly convey the intended message. The evaluation should focus on how well the LLM understands and processes different languages and contexts.

* Explanation of Options:

* Option A: The ability to generate responses in code- This is not relevant to translation quality or safety.

* Option B: The similarity to the previous language- While ensuring that translations preserve the original's intent is important, this doesn't directly address the overall quality or safety of the translation.

* Option C: The latency of the response and the length of text generated- These operational metrics are less critical in assessing the qualitative aspects of translation safety.

* Option D: The accuracy and relevance of the responses- This is crucial in translation to ensure that the translated content is true to the original in meaning and appropriateness. Accuracy and relevance directly impact the effectiveness and safety of translations, especially in sensitive or nuanced contexts.

Thus, Option D is the most important indicator when evaluating the safety of LLM outputs in translation, focusing on the core aspects that determine the utility and trustworthiness of translated content.

NEW QUESTION # 33

.....

In modern society, you cannot support yourself if you stop learning. That means you must work hard to learn useful knowledge in order to survive especially in your daily work. Our Databricks-Generative-AI-Engineer-Associate learning questions are filled with useful knowledge, which will broaden your horizons and update your skills. Lack of the knowledge cannot help you accomplish the tasks efficiently. But our Databricks-Generative-AI-Engineer-Associate Exam Questions can help you solve all of these problems. And our Databricks-Generative-AI-Engineer-Associate study guide can be your work assistant.

Exam Databricks-Generative-AI-Engineer-Associate Course: <https://www.passtorrent.com/Databricks-Generative-AI-Engineer-Associate-latest-torrent.html>

Databricks New Databricks-Generative-AI-Engineer-Associate Test Discount Any 10 Testing Engines can be Downloaded per month if you buy Unlimited Access for any duration, Databricks New Databricks-Generative-AI-Engineer-Associate Test Discount It can ensure a lucrative financial career for you, opening up a number of job opportunities, Exam Databricks-Generative-AI-Engineer-Associate Course - Databricks Certified Generative AI Engineer Associate study material provides you with the Exam Databricks-Generative-AI-Engineer-Associate Course - Databricks Certified Generative AI Engineer Associate questions exam practice questions and answers, which enable you to pass the exam successfully, If that's your attitudes, then you will be fortunate enough to come across our Databricks-Generative-AI-Engineer-Associate : Databricks Certified Generative AI Engineer Associate exam study material.

Introduction to i-mode Development, This step ensures that they make the most out of Reliable Databricks-Generative-AI-Engineer-Associate Test Labs of their system and further improve their service management, Any 10 Testing Engines can be Downloaded per month if you buy Unlimited Access for any duration.

High Pass-Rate New Databricks-Generative-AI-Engineer-Associate Test Discount for Real Exam

It can ensure a lucrative financial career Exam Databricks-Generative-AI-Engineer-Associate Course for you, opening up a number of job opportunities, Databricks Certified Generative AI Engineer Associate study material provides you with the Databricks Certified Generative AI Engineer Associate questions exam Databricks-Generative-AI-Engineer-Associate Practice Questions and

If that's your attitudes, then you will be fortunate enough to come across our Databricks-Generative-AI-Engineer-Associate : Databricks Certified Generative AI Engineer Associate exam study material, And your pass rate will reach 99%.

- DOWNLOAD the newest PassTorrent Databricks-Generative-AI-Engineer-Associate PDF dumps from Cloud Storage for free:
<https://drive.google.com/open?id=1o8A4--tSVLkcMqgSu8jpBQpCb-EwplyY>