

Databricks-Generative-AI-Engineer-Associate Desktop Practice Exam Software



P.S. Free & New Databricks-Generative-AI-Engineer-Associate dumps are available on Google Drive shared by TestkingPass:
<https://drive.google.com/open?id=192JV3E7Z2RGOB2FkxUKBSC-dANmDysmv>

With our Databricks-Generative-AI-Engineer-Associate study materials, only should you take about 20 - 30 hours to preparation can you attend the exam. The rest of the time you can do anything you want to do to, which can fully reduce your review pressure. Saving time and improving efficiency is the consistent purpose of our Databricks-Generative-AI-Engineer-Associate Learning Materials. With the help of our Databricks-Generative-AI-Engineer-Associate exam questions, your review process will no longer be full of pressure and anxiety.

The TestkingPass is one of the most in-demand platforms for Databricks Databricks-Generative-AI-Engineer-Associate exam preparation and success. The TestkingPass is offering valid, and real Databricks Databricks-Generative-AI-Engineer-Associate exam dumps. They all used the Databricks Databricks-Generative-AI-Engineer-Associate exam dumps and passed their dream Databricks Databricks-Generative-AI-Engineer-Associate Exam easily. The Databricks Databricks-Generative-AI-Engineer-Associate exam dumps will provide you with everything that you need to prepare, learn and pass the difficult Databricks Databricks-Generative-AI-Engineer-Associate exam.

>> Databricks-Generative-AI-Engineer-Associate Valid Test Simulator <<

Databricks-Generative-AI-Engineer-Associate Test Tutorials & Valid Databricks-Generative-AI-Engineer-Associate Exam Pdf

This format is for candidates who do not have the time or energy to use a computer or laptop for preparation. The Databricks-Generative-AI-Engineer-Associate PDF file includes real Databricks-Generative-AI-Engineer-Associate questions, and they can be easily printed and studied at any time. TestkingPass regularly updates its PDF file to ensure that its readers have access to the updated questions.

Databricks Databricks-Generative-AI-Engineer-Associate Exam Syllabus Topics:

Topic	Details
Topic 1	<ul style="list-style-type: none">• Application Development: In this topic, Generative AI Engineers learn about tools needed to extract data, Langchain• similar tools, and assessing responses to identify common issues. Moreover, the topic includes questions about adjusting an LLM's response, LLM guardrails, and the best LLM based on the attributes of the application.

Topic 2	<ul style="list-style-type: none"> • Design Applications: The topic focuses on designing a prompt that elicits a specifically formatted response. It also focuses on selecting model tasks to accomplish a given business requirement. Lastly, the topic covers chain components for a desired model input and output.
Topic 3	<ul style="list-style-type: none"> • Evaluation and Monitoring: This topic is all about selecting an LLM choice and key metrics. Moreover, Generative AI Engineers learn about evaluating model performance. Lastly, the topic includes sub-topics about inference logging and usage of Databricks features.
Topic 4	<ul style="list-style-type: none"> • Data Preparation: Generative AI Engineers covers a chunking strategy for a given document structure and model constraints. The topic also focuses on filter extraneous content in source documents. Lastly, Generative AI Engineers also learn about extracting document content from provided source data and format.
Topic 5	<ul style="list-style-type: none"> • Governance: Generative AI Engineers who take the exam get knowledge about masking techniques, guardrail techniques, and legal • licensing requirements in this topic.

Databricks Certified Generative AI Engineer Associate Sample Questions (Q46-Q51):

NEW QUESTION # 46

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

Answer: C

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 # 47

A Generative AI Engineer is building a system which will answer questions on latest stock news articles.

Which will NOT help with ensuring the outputs are relevant to financial news?

- A. Implement a comprehensive guardrail framework that includes policies for content filters tailored to the finance sector.
- B. Incorporate manual reviews to correct any problematic outputs prior to sending to the users
- C. Increase the compute to improve processing speed of questions to allow greater relevancy analysis
- D. Implement a profanity filter to screen out offensive language

Answer: C

Explanation:

In the context of ensuring that outputs are relevant to financial news, increasing compute power (option B) does not directly improve the relevance of the LLM-generated outputs. Here's why:

* **Compute Power and Relevancy:** Increasing compute power can help the model process inputs faster, but it does not inherently improve the relevance of the answers. Relevancy depends on the data sources, the retrieval method, and the filtering mechanisms in place, not on how quickly the model processes the query.

* **What Actually Helps with Relevancy:** Other methods, like content filtering, guardrails, or manual review, can directly impact the relevance of the model's responses by ensuring the model focuses on pertinent financial content. These methods help tailor the LLM's responses to the financial domain and avoid irrelevant or harmful outputs.

* **Why Other Options Are More Relevant:**

* **A (Comprehensive Guardrail Framework):** This will ensure that the model avoids generating content that is irrelevant or inappropriate in the finance sector.

* **C (Profanity Filter):** While not directly related to financial relevancy, ensuring the output is clean and professional is still important in maintaining the quality of responses.

* **D (Manual Review):** Incorporating human oversight to catch and correct issues with the LLM's output ensures the final answers are aligned with financial content expectations.

Thus, increasing compute power does not help with ensuring the outputs are more relevant to financial news, making option B the correct answer.

NEW QUESTION # 48

A Generative AI Engineer is tasked with developing an application that is based on an open source large language model (LLM). They need a foundation LLM with a large context window.

Which model fits this need?

- A. DBRX
- B. Llama2-70B
- C. MPT-30B
- D. DistilBERT

Answer: B

Explanation:

* **Problem Context:** The engineer needs an open-source LLM with a large context window to develop an application.

* **Explanation of Options:**

* **Option A: DistilBERT:** While an efficient and smaller version of BERT, DistilBERT does not provide a particularly large context window.

* **Option B: MPT-30B:** This model, while large, is not specified as being particularly notable for its context window capabilities.

* **Option C: Llama2-70B:** Known for its large model size and extensive capabilities, including a large context window. It is also available as an open-source model, making it ideal for applications requiring extensive contextual understanding.

* **Option D: DBRX:** This is not a recognized standard model in the context of large language models with extensive context windows.

Thus, Option C (Llama2-70B) is the best fit as it meets the criteria of having a large context window and being available for open-source use, suitable for developing robust language understanding applications.

NEW QUESTION # 49

A Generative AI Engineer is setting up a Databricks Vector Search that will lookup news articles by topic within 10 days of the date specified. An example query might be "Tell me about monster truck news around January 5th 1992". They want to do this with the least amount of effort.

How can they set up their Vector Search index to support this use case?

- A. Split articles by 10 day blocks and return the block closest to the query.
- B. Include metadata columns for article date and topic to support metadata filtering.

- C. Create separate indexes by topic and add a classifier model to appropriately pick the best index.
- D. pass the query directly to the vector search index and return the best articles.

Answer: B

Explanation:

The task is to set up a Databricks Vector Search index for news articles, supporting queries like "monster truck news around January 5th, 1992," with minimal effort. The index must filter by topic and a 10-day date range. Let's evaluate the options.

- * Option A: Split articles by 10-day blocks and return the block closest to the query
- * Pre-splitting articles into 10-day blocks requires significant preprocessing and index management (e.g., one index per block). It's effort-intensive and inflexible for dynamic date ranges.
- * Databricks Reference: "Static partitioning increases setup complexity; metadata filtering is preferred" ("Databricks Vector Search Documentation").
- * Option B: Include metadata columns for article date and topic to support metadata filtering
- * Adding date and topic as metadata in the Vector Search index allows dynamic filtering (e.g., date ± 5 days, topic = "monster truck") at query time. This leverages Databricks' built-in metadata filtering, minimizing setup effort.
- * Databricks Reference: "Vector Search supports metadata filtering on columns like date or category for precise retrieval with minimal preprocessing" ("Vector Search Guide," 2023).
- * Option C: Pass the query directly to the vector search index and return the best articles
- * Passing the full query (e.g., "Tell me about monster truck news around January 5th, 1992") to Vector Search relies solely on embeddings, ignoring structured filtering for date and topic. This risks inaccurate results without explicit range logic.
- * Databricks Reference: "Pure vector similarity may not handle temporal or categorical constraints effectively" ("Building LLM Applications with Databricks").
- * Option D: Create separate indexes by topic and add a classifier model to appropriately pick the best index
- * Separate indexes per topic plus a classifier model adds significant complexity (index creation, model training, maintenance), far exceeding "least effort." It's overkill for this use case.
- * Databricks Reference: "Multiple indexes increase overhead; single-index with metadata is simpler" ("Databricks Vector Search Documentation").

Conclusion: Option B is the simplest and most effective solution, using metadata filtering in a single Vector Search index to handle date ranges and topics, aligning with Databricks' emphasis on efficient, low-effort setups.

NEW QUESTION # 50

A Generative AI Engineer is creating an LLM system that will retrieve news articles from the year 1918 and related to a user's query and summarize them. The engineer has noticed that the summaries are generated well but often also include an explanation of how the summary was generated, which is undesirable.

Which change could the Generative AI Engineer perform to mitigate this issue?

- A. Revisit their document ingestion logic, ensuring that the news articles are being ingested properly.
- B. Split the LLM output by newline characters to truncate away the summarization explanation.
- C. Tune the chunk size of news articles or experiment with different embedding models.
- **D. Provide few shot examples of desired output format to the system and/or user prompt.**

Answer: D

Explanation:

To mitigate the issue of the LLM including explanations of how summaries are generated in its output, the best approach is to adjust the training or prompt structure. Here's why Option D is effective:

- * Few-shot Learning: By providing specific examples of how the desired output should look (i.e., just the summary without explanation), the model learns the preferred format. This few-shot learning approach helps the model understand not only what content to generate but also how to format its responses.
 - * Prompt Engineering: Adjusting the user prompt to specify the desired output format clearly can guide the LLM to produce summaries without additional explanatory text. Effective prompt design is crucial in controlling the behavior of generative models.
- Why Other Options Are Less Suitable:

* A: While technically feasible, splitting the output by newline and truncating could lead to loss of important content or create awkward breaks in the summary.

* B: Tuning chunk sizes or changing embedding models does not directly address the issue of the model's tendency to generate explanations along with summaries.

* C: Revisiting document ingestion logic ensures accurate source data but does not influence how the model formats its output.

By using few-shot examples and refining the prompt, the engineer directly influences the output format, making this approach the most targeted and effective solution.

• • • • •

Databricks-Generative-AI-Engineer-Associate Test Tutorials: <https://www.testkingpass.com/Databricks-Generative-AI-Engineer-Associate-testking-dumps.html>

- [illegible]

P.S. Free 2026 Databricks Databricks-Generative-AI-Engineer-Associate dumps are available on Google Drive shared by TestkingPass: <https://drive.google.com/open?id=192JV3E7Z2RGOB2FkxUKBSC-dANmDysmv>