Databricks Databricks-Generative-AI-Engineer-Associate Exam Dumps Free - Real Databricks-Generative-AI-Engineer-Associate Exam

BONUS!!! Download part of Exams-boost Databricks-Generative-AI-Engineer-Associate dumps for free: https://drive.google.com/open?id=1Nh4YzMaCzIgkAeKUzxvzci-jQkPf33w-

For busy candidates who want to study for the Databricks Certified Generative AI Engineer Associate exam on the go via their smartphones, laptops, or tablets, our updated Databricks Databricks-Generative-AI-Engineer-Associate PDF Questions are excellent. Because the PDF file of the latest questions is portable, you can prepare for the Databricks-Generative-AI-Engineer-Associate Exam via a smart device whenever and wherever you like. Additionally, exam PDF questions are printable. You can print these Databricks-Generative-AI-Engineer-Associate exam questions to study when you don't have access to a smart device.

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

Topic	Details
Topic 1	• 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 2	• 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	• Governance: Generative AI Engineers who take the exam get knowledge about masking techniques, guardrail techniques, and legal • licensing requirements in this topic.

Real Databricks-Generative-AI-Engineer-Associate Exam, Latest Databricks-Generative-AI-Engineer-Associate Exam Duration

The Databricks Certified Generative AI Engineer Associate (Databricks-Generative-AI-Engineer-Associate) certification exam is one of the hottest and most industrial-recognized credentials that has been inspiring beginners and experienced professionals since its beginning. With the Databricks-Generative-AI-Engineer-Associate certification exam successful candidates can gain a range of benefits which include career advancement, higher earning potential, industrial recognition of skills and job security, and more career personal and professional growth.

Databricks Certified Generative AI Engineer Associate Sample Questions (Q72-Q77):

NEW QUESTION # 72
A Generative Al Engineer is developing a RAG application and would like to experiment with different embedding models to improve the application performance.
Which strategy for picking an embedding model should they choose?

- A. pick the embedding model ranked highest on the Massive Text Embedding Benchmark (MTEB) leaderboard hosted by HuggingFace
- B. Pick an embedding model trained on related domain knowledge
- C. Pick the most recent and most performant open LLM released at the time
- D. Pick an embedding model with multilingual support to support potential multilingual user questions

Answer: B

Explanation:
The task involves improving a Retrieval-Augmented Generation (RAG) application's performance by experimenting with embedding models. The choice of embedding model impacts retrieval accuracy, which is critical for RAG systems. Let's evaluate the options based on Databricks Generative AI Engineer best practices.
Option A: Pick an embedding model trained on related domain knowledge
Embedding models trained on domain-specific data (e.g., industry-specific corpora) produce vectors that better capture the semantics of the application's context, improving retrieval relevance. For RAG, this is a key strategy to enhance performance.
Databricks Reference: "For optimal retrieval in RAG systems, select embedding models aligned with the domain of your data" ("Building LLM Applications with Databricks," 2023).
Option B: Pick the most recent and most performant open LLM released at the time LLMs are not embedding models; they generate text, not embeddings for retrieval. While recent LLMs may be performant for generation, this doesn't address the embedding step in RAG. This option misunderstands the component being selected.
Databricks Reference: Embedding models and LLMs are distinct in RAG workflows: "Embedding models convert text to vectors, while LLMs generate responses" ("Generative AI Cookbook").
Option C: Pick the embedding model ranked highest on the Massive Text Embedding Benchmark (MTEB) leaderboard hosted by HuggingFace The MTEB leaderboard ranks models across general tasks, but high overall performance doesn't guarantee suitability for a specific domain. A top-ranked model might excel in generic contexts but underperform on the engineer's unique data.
Databricks Reference: General performance is less critical than domain fit: "Benchmark rankings provide a starting point, but domain-specific evaluation is recommended" ("Databricks Generative AI Engineer Guide").
Option D: Pick an embedding model with multilingual support to support potential multilingual user questions Multilingual support is useful only if the application explicitly requires it. Without evidence of multilingual needs, this adds complexity without guaranteed performance gains for the current use case.

Databricks Reference: "Choose features like multilingual support based on application requirements" ("Building LLM-Powered Applications").
Conclusion: Option A is the best strategy because it prioritizes domain relevance, directly improving retrieval accuracy in a RAG system-aligning with Databricks' emphasis on tailoring models to specific use cases.

NEW QUESTION # 73
A Generative Al Engineer is tasked with developing a RAG application that will help a small internal group of experts at their company answer specific questions, augmented by an internal knowledge base. They want the best possible quality in the answers, and neither latency nor throughput is a huge concern given that the user group is small and they're willing to wait for the best answer. The topics are sensitive in nature and the data is highly confidential and so, due to regulatory requirements, none of the information is allowed to be transmitted to third parties.
Which model meets all the Generative Al Engineer's needs in this situation?

- A. BGE-large
- B. OpenAI GPT-4
- C. Dolly 1.5B
- D. Llama2-70B

Answer: A

Explanation:
Problem Context: The Generative AI Engineer needs a model for a Retrieval-Augmented Generation (RAG) application that provides high-quality answers, where latency and throughput are not major concerns. The key factors areconfidentialityandsensitivityof the data, as well as the requirement for all processing to be confined to internal resources without external data transmission.
Explanation of Options:
* Option A: Dolly 1.5B: This model does not typically support RAG applications as it's more focused on image generation tasks.
* Option B: OpenAI GPT-4: While GPT-4 is powerful for generating responses, its standard deployment involves cloud-based processing, which could violate the confidentiality requirements due to external data transmission.
* Option C: BGE-large: The BGE (Big Green Engine) large model is a suitable choice if it is configured to operate on-premises or within a secure internal environment that meets regulatory requirements.
Assuming this setup, BGE-large can provide high-quality answers while ensuring that data is not transmitted to third parties, thus aligning with the project's sensitivity and confidentiality needs.
* Option D: Llama2-70B: Similar to GPT-4, unless specifically set up for on-premises use, it generally relies on cloud-based services, which might risk confidential data exposure.
Given the sensitivity and confidentiality concerns,BGE-largeis assumed to be configurable for secure internal use, making it the optimal choice for this scenario.

NEW QUESTION # 74
A team uses Mosaic AI Vector Search to retrieve documents for their Retrieval-Augmented Generation (RAG) pipeline. The search query returns five relevant documents, and the first three are added to the prompt as context. Performance evaluation with Agent Evaluation shows that some lower-ranked retrieved documents have higher context relevancy scores than higher-ranked documents.
Which option should the team consider to optimize this workflow?

- A. Increase the number of documents added to the prompt to improve context relevance.
- B. Use a different embedding model for computing document embeddings.
- C. Use a reranker to order the documents based on the relevance scores.
- D. Modify the prompt to instruct the LLM to order the documents based on the relevance scores.

Answer: C

Explanation:
The scenario describes a common "retrieval gap" where the initial bi-encoder (embedding model) used for vector search identifies relevant documents but does not rank them perfectly. This happens because embedding models represent entire documents as a single vector, which can lose nuance. The standard engineering solution is to implement a Reranker (Cross-Encoder). Unlike embedding models, a reranker processes the query and a candidate document simultaneously, allowing it to capture deep semantic interactions between the two. In a Mosaic AI workflow, after the vector search retrieves the top k documents, the reranker evaluates those specific k documents to produce a more accurate relevance score. This ensures that the most contextually relevant documents are placed at the top of the list (and thus the top of the LLM prompt), which is crucial because LLMs are sensitive to

document order and often prioritize information found at the beginning of the context.

NEW QUESTION # 75
A Generative AI Engineer is designing a chatbot for a gaming company that aims to engage users on its platform while its users play online video games.
Which metric would help them increase user engagement and retention for their platform?

- A. Randomness
- B. Repetition of responses
- C. Lack of relevance
- D. Diversity of responses

Answer: D

Explanation:
In the context of designing a chatbot to engage users on a gaming platform,diversity of responses(option B) is a key metric to increase user engagement and retention. Here's why:
* Diverse and Engaging Interactions:A chatbot that provides varied and interesting responses will keep users engaged, especially in an interactive environment like a gaming platform. Gamers typically enjoy dynamic and evolving conversations, anddiversity of responseshelps prevent monotony, encouraging users to interact more frequently with the bot.
* Increasing Retention:By offering different types of responses to similar queries, the chatbot can create a sense of novelty and excitement, which enhances the user's experience and makes them more likely to return to the platform.
* Why Other Options Are Less Effective:
* A (Randomness): Random responses can be confusing or irrelevant, leading to frustration and reducing engagement.
* C (Lack of Relevance): If responses are not relevant to the user's queries, this will degrade the user experience and lead to disengagement.
* D (Repetition of Responses): Repetitive responses can quickly bore users, making the chatbot feel uninteresting and reducing the likelihood of continued interaction.
Thus,diversity of responses(option B) is the most effective way to keep users engaged and retain them on the platform.

NEW QUESTION # 76
A Generative AI Engineer is designing a RAG application for answering user questions on technical regulations as they learn a new sport.
What are the steps needed to build this RAG application and deploy it?

- A. Ingest documents from a source -> Index the documents and saves to Vector Search -> User submits queries against an LLM -> LLM retrieves relevant documents -> Evaluate model -> LLM generates a response -> Deploy it using Model Serving
- B. Ingest documents from a source -> Index the documents and save to Vector Search -> User submits queries against an LLM -> LLM retrieves relevant documents -> LLM generates a response -> Evaluate model -> Deploy it using Model Serving
- C. Ingest documents from a source -> Index the documents and save to Vector Search -> Evaluate model -> Deploy it using Model Serving
- D. User submits queries against an LLM -> Ingest documents from a source -> Index the documents and save to Vector Search -> LLM retrieves relevant documents -> LLM generates a response -> Evaluate model -> Deploy it using Model Serving

Answer: B

Explanation:
The Generative AI Engineer needs to follow a methodical pipeline to build and deploy a Retrieval-Augmented Generation (RAG) application. The steps outlined in option B accurately reflect this process:
Ingest documents from a source: This is the first step, where the engineer collects documents (e.g., technical regulations) that will be used for retrieval when the application answers user questions.
Index the documents and save to Vector Search: Once the documents are ingested, they need to be embedded using a technique like embeddings (e.g., with a pre-trained model like BERT) and stored in a vector database (such as Pinecone or FAISS). This enables fast retrieval based on user queries.
User submits queries against an LLM: Users interact with the application by submitting their queries. These queries will be passed to the LLM.

LLM retrieves relevant documents: The LLM works with the vector store to retrieve the most relevant documents based on their vector representations.
LLM generates a response: Using the retrieved documents, the LLM generates a response that is tailored to the user's question.
Evaluate model: After generating responses, the system must be evaluated to ensure the retrieved documents are relevant and the generated response is accurate. Metrics such as accuracy, relevance, and user satisfaction can be used for evaluation.
Deploy it using Model Serving: Once the RAG pipeline is ready and evaluated, it is deployed using a model-serving platform such as Databricks Model Serving. This enables real-time inference and response generation for users.
By following these steps, the Generative AI Engineer ensures that the RAG application is both efficient and effective for the task of answering technical regulation questions.

NEW QUESTION # 77
.....

www.troytecdumps.com◂ and search for □ Databricks-Generative-AI-Engineer-Associate □ to download for free □ □Databricks-Generative-AI-Engineer-Associate Updated Dumps

- Features of Pdfvce Databricks Databricks-Generative-AI-Engineer-Associate Web-Based Practice Questions □ Copy URL ▹ www.pdfvce.com ◃ open and search for ▹ Databricks-Generative-AI-Engineer-Associate ◃ to download for free □ □Exam Databricks-Generative-AI-Engineer-Associate Quick Prep
- Reliable Databricks-Generative-AI-Engineer-Associate Exam Dumps Free bring you Verified Real Databricks-Generative-AI-Engineer-Associate Exam for Databricks Databricks Certified Generative AI Engineer Associate □ The page for free download of ▹ Databricks-Generative-AI-Engineer-Associate ◃ on □ www.prep4away.com □ will open immediately □ □Databricks-Generative-AI-Engineer-Associate Certification Materials
- my-social-box.com, bookmarks4seo.com, zaynabfyqe688422.bloggip.com, sashalhyx377035.thelateblog.com, kingbookmark.com, ilovebookmark.com, ariabookmarks.com, bookmarklogin.com, nikolasymoh822557.smblogsites.com, socialbuzzfeed.com, Disposable vapes

P.S. Free 2026 Databricks Databricks-Generative-AI-Engineer-Associate dumps are available on Google Drive shared by Exams-boost: https://drive.google.com/open?id=1Nh4YzMaCzIgkAeKUzxvzci-jQkPf33w-