

Amazing Databricks-Generative-AI-Engineer-Associate Exam Questions Provide You the Most Accurate Learning Braindumps - ITexamReview

Databricks Generative AI Engineer Associate Exam

Databricks Certified Generative AI Engineer Associate

<https://www.passquestion.com/databricks-generative-ai-engineer-associate.html>



35% OFF on All, Including Databricks Generative AI Engineer Associate Questions and Answers

Pass Databricks Generative AI Engineer Associate Exam with
PassQuestion Databricks Generative AI Engineer Associate
questions and answers in the first attempt.

<https://www.passquestion.com/>

1 / 8

P.S. Free & New Databricks-Generative-AI-Engineer-Associate dumps are available on Google Drive shared by ITexamReview:
https://drive.google.com/open?id=10MMXGgDf_JKvVr3bbWNbg6eRKDSD-vUj

The ITexamReview is one of the top-rated and trusted platforms that are committed to making the Databricks Certified Generative AI Engineer Associate (Databricks-Generative-AI-Engineer-Associate) certification exam journey successful. To achieve this objective ITexamReview has hired a team of experienced and qualified Databricks Databricks-Generative-AI-Engineer-Associate Exam trainers. They work together and put all their expertise to maintain the top standard of Databricks-Generative-AI-Engineer-Associate practice test all the time.

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

Topic	Details
Topic 1	<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 2	<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.
Topic 3	<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 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.

>> Databricks-Generative-AI-Engineer-Associate Reliable Exam Test <<

Guaranteed Success with Real and Updated Databricks Databricks-Generative-AI-Engineer-Associate Exam Questions

In order to meet the needs of all customers, Our Databricks-Generative-AI-Engineer-Associate study torrent has a long-distance aid function. If you feel confused about our Databricks-Generative-AI-Engineer-Associate test torrent when you use our products, do not hesitate and send a remote assistance invitation to us for help, we are willing to provide remote assistance for you in the shortest time. We have professional staff, so your all problems about Databricks-Generative-AI-Engineer-Associate Guide Torrent will be solved by our professional staff. We can make sure that you will enjoy our considerate service if you buy our Databricks-Generative-AI-Engineer-Associate study torrent.

Databricks Certified Generative AI Engineer Associate Sample Questions (Q58-Q63):

NEW QUESTION # 58

Which TWO chain components are required for building a basic LLM-enabled chat application that includes conversational capabilities, knowledge retrieval, and contextual memory?

- A. (Q)
- B. External tools
- C. Conversation Buffer Memory
- D. React Components
- E. Vector Stores
- F. Chat loaders

Answer: C,E

Explanation:

Building a basic LLM-enabled chat application with conversational capabilities, knowledge retrieval, and contextual memory requires specific components that work together to process queries, maintain context, and retrieve relevant information. Databricks' Generative AI Engineer documentation outlines key components for such systems, particularly in the context of frameworks like LangChain or Databricks' MosaicML integrations. Let's evaluate the required components:

* Understanding the Requirements:

* Conversational capabilities: The app must generate natural, coherent responses.

* Knowledge retrieval: It must access external or domain-specific knowledge.

* Contextual memory: It must remember prior interactions in the conversation.

* Databricks Reference:"A typical LLM chat application includes a memory component to track conversation history and a retrieval mechanism to incorporate external knowledge"("Databricks Generative AI Cookbook," 2023).

* Evaluating the Options:

* A. (Q): This appears incomplete or unclear (possibly a typo). Without further context, it's not a valid component.

* B. Vector Stores: These store embeddings of documents or knowledge bases, enabling semantic search and retrieval of relevant

information for the LLM. This is critical for knowledge retrieval in a chat application.

- * Databricks Reference: "Vector stores, such as those integrated with Databricks' Lakehouse, enable efficient retrieval of contextual data for LLMs" ("Building LLM Applications with Databricks").

- * C. Conversation Buffer Memory: This component stores the conversation history, allowing the LLM to maintain context across multiple turns. It's essential for contextual memory.

- * Databricks Reference: "Conversation Buffer Memory tracks prior user inputs and LLM outputs, ensuring context-aware responses" ("Generative AI Engineer Guide").

- * D. External tools: These (e.g., APIs or calculators) enhance functionality but aren't required for a basic chat app with the specified capabilities.

- * E. Chat loaders: These might refer to data loaders for chat logs, but they're not a core chain component for conversational functionality or memory.

- * F. React Components: These relate to front-end UI development, not the LLM chain's backend functionality.

- * Selecting the Two Required Components:

- * For knowledge retrieval, Vector Stores (B) are necessary to fetch relevant external data, a cornerstone of Databricks' RAG-based chat systems.

- * For contextual memory, Conversation Buffer Memory (C) is required to maintain conversation history, ensuring coherent and context-aware responses.

- * While an LLM itself is implied as the core generator, the question asks for chain components beyond the model, making B and C the minimal yet sufficient pair for a basic application.

Conclusion: The two required chain components are B. Vector Stores and C. Conversation Buffer Memory, as they directly address knowledge retrieval and contextual memory, respectively, aligning with Databricks' documented best practices for LLM-enabled chat applications.

NEW QUESTION # 59

A Generative AI Engineer is building an LLM to generate article summaries in the form of a type of poem, such as a haiku, given the article content. However, the initial output from the LLM does not match the desired tone or style.

Which approach will NOT improve the LLM's response to achieve the desired response?

- A. Include few-shot examples in the prompt to the LLM
- **B. Use a neutralizer to normalize the tone and style of the underlying documents**
- C. Provide the LLM with a prompt that explicitly instructs it to generate text in the desired tone and style
- D. Fine-tune the LLM on a dataset of desired tone and style

Answer: B

Explanation:

The task at hand is to improve the LLM's ability to generate poem-like article summaries with the desired tone and style. Using a neutralizer to normalize the tone and style of the underlying documents (option B) will not help improve the LLM's ability to generate the desired poetic style. Here's why:

- * Neutralizing Underlying Documents: A neutralizer aims to reduce or standardize the tone of input data. However, this contradicts the goal, which is to generate text with a specific tone and style (like haikus). Neutralizing the source documents will strip away the richness of the content, making it harder for the LLM to generate creative, stylistic outputs like poems.

- * Why Other Options Improve Results:

- * A (Explicit Instructions in the Prompt): Directly instructing the LLM to generate text in a specific tone and style helps align the output with the desired format (e.g., haikus). This is a common and effective technique in prompt engineering.

- * C (Few-shot Examples): Providing examples of the desired output format helps the LLM understand the expected tone and structure, making it easier to generate similar outputs.

- * D (Fine-tuning the LLM): Fine-tuning the model on a dataset that contains examples of the desired tone and style is a powerful way to improve the model's ability to generate outputs that match the target format.

Therefore, using a neutralizer (option B) is not an effective method for achieving the goal of generating stylized poetic summaries.

NEW QUESTION # 60

A Generative AI 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 AI Engineer's needs in this situation?

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

Answer: C

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 are confidentiality and sensitivity of 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-large is assumed to be configurable for secure internal use, making it the optimal choice for this scenario.

NEW QUESTION # 61

A Generative AI Engineer is developing a patient-facing healthcare-focused chatbot. If the patient's question is not a medical emergency, the chatbot should solicit more information from the patient to pass to the doctor's office and suggest a few relevant pre-approved medical articles for reading. If the patient's question is urgent, direct the patient to calling their local emergency services.

Given the following user input:

"I have been experiencing severe headaches and dizziness for the past two days." Which response is most appropriate for the chatbot to generate?

- A. Headaches can be tough. Hope you feel better soon!
- B. Please provide your age, recent activities, and any other symptoms you have noticed along with your headaches and dizziness.
- **C. Please call your local emergency services.**
- D. Here are a few relevant articles for your browsing. Let me know if you have questions after reading them.

Answer: C

Explanation:

* Problem Context: The task is to design responses for a healthcare-focused chatbot that appropriately addresses the urgency of a patient's symptoms.

* Explanation of Options:

- * Option A: Suggesting articles might be suitable for less urgent inquiries but is inappropriate for symptoms that could indicate a serious condition.
- * Option B: Given the description of severe symptoms like headaches and dizziness, directing the patient to emergency services is prudent. This aligns with medical guidelines that recommend immediate professional attention for such severe symptoms.
- * Option C: Offering well-wishes does not address the potential seriousness of the symptoms and lacks appropriate action.
- * Option D: While gathering more information is part of a detailed assessment, the immediate need here suggests a more urgent response.

Given the potential severity of the described symptoms, Option B is the most appropriate, ensuring the chatbot directs patients to seek urgent care when needed, potentially saving lives.

NEW QUESTION # 62

A Generative AI Engineer is building a Generative AI system that suggests the best matched employee team member to newly scoped projects. The team member is selected from a very large team. The match should be based upon project date availability and

how well their employee profile matches the project scope. Both the employee profile and project scope are unstructured text. How should the Generative AI Engineer architect their system?

- A. Create a tool for finding available team members given project dates. Embed all project scopes into a vector store, perform a retrieval using team member profiles to find the best team member.
- **B. Create a tool for finding available team members given project dates. Embed team profiles into a vector store and use the project scope and filtering to perform retrieval to find the available best matched team members.**
- C. Create a tool for finding team member availability given project dates, and another tool that uses an LLM to extract keywords from project scopes. Iterate through available team members' profiles and perform keyword matching to find the best available team member.
- D. Create a tool to find available team members given project dates. Create a second tool that can calculate a similarity score for a combination of team member profile and the project scope. Iterate through the team members and rank by best score to select a team member.

Answer: B

Explanation:

* Problem Context: The problem involves matching team members to new projects based on two main factors:

* Availability: Ensure the team members are available during the project dates.

* Profile-Project Match: Use the employee profiles (unstructured text) to find the best match for a project's scope (also unstructured text).

The two main inputs are the employee profiles and project scopes, both of which are unstructured. This means traditional rule-based systems (e.g., simple keyword matching) would be inefficient, especially when working with large datasets.

* Explanation of Options: Let's break down the provided options to understand why D is the most optimal answer.

* Option A suggests embedding project scopes into a vector store and then performing retrieval using team member profiles. While embedding project scopes into a vector store is a valid technique, it skips an important detail: the focus should primarily be on embedding employee profiles because we're matching the profiles to a new project, not the other way around.

* Option B involves using a large language model (LLM) to extract keywords from the project scope and perform keyword matching on employee profiles. While LLMs can help with keyword extraction, this approach is too simplistic and doesn't leverage advanced retrieval techniques like vector embeddings, which can handle the nuanced and rich semantics of unstructured data. This approach may miss out on subtle but important similarities.

* Option C suggests calculating a similarity score between each team member's profile and project scope. While this is a good idea, it doesn't specify how to handle the unstructured nature of data efficiently. Iterating through each member's profile individually could be computationally expensive in large teams. It also lacks the mention of using a vector store or an efficient retrieval mechanism.

* Option D is the correct approach. Here's why:

* Embedding team profiles into a vector store: Using a vector store allows for efficient similarity searches on unstructured data.

Embedding the team member profiles into vectors captures their semantics in a way that is far more flexible than keyword-based matching.

* Using project scope for retrieval: Instead of matching keywords, this approach suggests using vector embeddings and similarity search algorithms (e.g., cosine similarity) to find the team members whose profiles most closely align with the project scope.

* Filtering based on availability: Once the best-matched candidates are retrieved based on profile similarity, filtering them by availability ensures that the system provides a practically useful result.

This method efficiently handles large-scale datasets by leveraging vector embeddings and similarity search techniques, both of which are fundamental tools in Generative AI engineering for handling unstructured text.

* Technical References:

* Vector embeddings: In this approach, the unstructured text (employee profiles and project scopes) is converted into high-dimensional vectors using pretrained models (e.g., BERT, Sentence-BERT, or custom embeddings). These embeddings capture the semantic meaning of the text, making it easier to perform similarity-based retrieval.

* Vector stores: Solutions like FAISS or Milvus allow storing and retrieving large numbers of vector embeddings quickly. This is critical when working with large teams where querying through individual profiles sequentially would be inefficient.

* LLM Integration: Large language models can assist in generating embeddings for both employee profiles and project scopes. They can also assist in fine-tuning similarity measures, ensuring that the retrieval system captures the nuances of the text data.

* Filtering: After retrieving the most similar profiles based on the project scope, filtering based on availability ensures that only team members who are free for the project are considered.

This system is scalable, efficient, and makes use of the latest techniques in Generative AI, such as vector embeddings and semantic search.

NEW QUESTION # 63

.....

- BONUS!!! Download part of ITexamReview Databricks-Generative-AI-Engineer-Associate dumps for free: https://drive.google.com/open?id=10MMXGgDf_JKvVr3bbWNbg6eRKDSD-vUj