

# Databricks Databricks-Generative-AI-Engineer-Associate Related Exams - Databricks-Generative-AI-Engineer-Associate Reliable Test Answers



P.S. Free & New Databricks-Generative-AI-Engineer-Associate dumps are available on Google Drive shared by Pass4cram [https://drive.google.com/open?id=1bjXoPRZAwVxwVswclW3\\_wyg\\_MpouqV4m](https://drive.google.com/open?id=1bjXoPRZAwVxwVswclW3_wyg_MpouqV4m)

Preparing for the Databricks-Generative-AI-Engineer-Associate exam can be a daunting task, but with real Databricks-Generative-AI-Engineer-Associate exam questions, it can be a lot easier. The importance of actual Databricks Certified Generative AI Engineer Associate (Databricks-Generative-AI-Engineer-Associate) questions cannot be overemphasized. Databricks-Generative-AI-Engineer-Associate Real Questions are crucial for passing the Databricks-Generative-AI-Engineer-Associate exam. When candidates have access to the updated Databricks Databricks-Generative-AI-Engineer-Associate practice test questions, they are better prepared to succeed.

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

Topic	Details
Topic 1	<ul style="list-style-type: none"> <li>Application Development: In this topic, Generative AI Engineers learn about tools needed to extract data, Langchain</li> <li>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.</li> </ul>
Topic 2	<ul style="list-style-type: none"> <li>Governance: Generative AI Engineers who take the exam get knowledge about masking techniques, guardrail techniques, and legal</li> <li>licensing requirements in this topic.</li> </ul>
Topic 3	<ul style="list-style-type: none"> <li>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.</li> </ul>
Topic 4	<ul style="list-style-type: none"> <li>Assembling and Deploying Applications: In this topic, Generative AI Engineers get knowledge about coding a chain using a pyfunc mode, coding a simple chain using langchain, and coding a simple chain according to requirements. Additionally, the topic focuses on basic elements needed to create a RAG application. Lastly, the topic addresses sub-topics about registering the model to Unity Catalog using MLflow.</li> </ul>

# 100% Pass 2026 Efficient Databricks Databricks-Generative-AI-Engineer-Associate Related Exams

Our specialists check whether the contents of Databricks-Generative-AI-Engineer-Associate real exam are updated every day. If there are newer versions, they will be sent to users in time to ensure that users can enjoy the latest resources in the first time. In such a way, our Databricks-Generative-AI-Engineer-Associate guide materials can have such a fast update rate that is taking into account the needs of users. Users using our Databricks-Generative-AI-Engineer-Associate Study Materials must be the first group of people who come into contact with new resources. When you receive an update reminder from Databricks-Generative-AI-Engineer-Associate practice questions, you can update the version in time and you will never miss a key message.

## Databricks Certified Generative AI Engineer Associate Sample Questions (Q38-Q43):

### NEW QUESTION # 38

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 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.
- B. 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.
- C. 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.
- D. 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.

**Answer: C**

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 # 39

A Generative AI Engineer has already trained an LLM on Databricks and it is now ready to be deployed.

Which of the following steps correctly outlines the easiest process for deploying a model on Databricks?

- A. Save the model along with its dependencies in a local directory, build the Docker image, and run the Docker container
- **B. Log the model using MLflow during training, directly register the model to Unity Catalog using the MLflow API, and start a serving endpoint**
- C. Wrap the LLM's prediction function into a Flask application and serve using Gunicorn
- D. Log the model as a pickle object, upload the object to Unity Catalog Volume, register it to Unity Catalog using MLflow, and start a serving endpoint

**Answer: B**

### NEW QUESTION # 40

A Generative AI Engineer is tasked with improving the RAG quality by addressing its inflammatory outputs.

Which action would be most effective in mitigating the problem of offensive text outputs?

- A. Increase the frequency of upstream data updates
- **B. Curate upstream data properly that includes manual review before it is fed into the RAG system**
- C. Inform the user of the expected RAG behavior
- D. Restrict access to the data sources to a limited number of users

**Answer: B**

Explanation:

Addressing offensive or inflammatory outputs in a Retrieval-Augmented Generation (RAG) system is critical for improving user experience and ensuring ethical AI deployment. Here's why Dis is the most effective approach:

\* Manual data curation: The root cause of offensive outputs often comes from the underlying data used to train the model or populate the retrieval system. By manually curating the upstream data and conducting thorough reviews before the data is fed into the RAG system, the engineer can filter out harmful, offensive, or inappropriate content.

\* Improving data quality: Curating data ensures the system retrieves and generates responses from a high-quality, well-vetted dataset. This directly impacts the relevance and appropriateness of the outputs from the RAG system, preventing inflammatory content from being included in responses.

\* Effectiveness: This strategy directly tackles the problem at its source (the data) rather than just mitigating the consequences (such as informing users or restricting access). It ensures that the system consistently provides non-offensive, relevant information.

Other options, such as increasing the frequency of data updates or informing users about behavior expectations, may not directly mitigate the generation of inflammatory outputs.

### NEW QUESTION # 41

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

**Answer: A**

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.

## NEW QUESTION # 42

A Generative AI Engineer wants their (in)etuned LLMs in their prod Databricks workspace available for testing in their dev workspace as well. All of their workspaces are Unity Catalog enabled and they are currently logging their models into the Model Registry in MLflow.

What is the most cost-effective and secure option for the Generative AI Engineer to accomplish their goal?

- A. Setup a duplicate training pipeline in dev, so that an identical model is available in dev.
- B. Use MLflow to log the model directly into Unity Catalog, and enable READ access in the dev workspace to the model.
- C. Use an external model registry which can be accessed from all workspaces
- D. Setup a script to export the model from prod and import it to dev.

**Answer: B**

Explanation:

The goal is to make fine-tuned LLMs from a production (prod) Databricks workspace available for testing in a development (dev) workspace, leveraging Unity Catalog and MLflow, while ensuring cost-effectiveness and security. Let's analyze the options.

Option A: Use an external model registry which can be accessed from all workspaces. An external registry adds cost (e.g., hosting fees) and complexity (e.g., integration, security configurations) outside Databricks' native ecosystem, reducing security compared to Unity Catalog's governance.

Databricks Reference: "Unity Catalog provides a centralized, secure model registry within Databricks" ("Unity Catalog Documentation," 2023).

Option B: Setup a script to export the model from prod and import it to dev. Export/import scripts require manual effort, storage for model artifacts, and repeated execution, increasing operational cost and risk (e.g., version mismatches, unsecured transfers). It's less efficient than a native solution.

Databricks Reference: Manual processes are discouraged when Unity Catalog offers built-in sharing: "Avoid redundant workflows with Unity Catalog's cross-workspace access" ("MLflow with Unity Catalog").

Option C: Setup a duplicate training pipeline in dev, so that an identical model is available in dev. Duplicating the training pipeline doubles compute and storage costs, as it retrains the model from scratch. It's neither cost-effective nor necessary when the prod model can be reused securely.

Databricks Reference: "Re-running training is resource-intensive; leverage existing models where possible" ("Generative AI Engineer Guide").

Option D: Use MLflow to log the model directly into Unity Catalog, and enable READ access in the dev workspace to the model. Unity Catalog, integrated with MLflow, allows models logged in prod to be centrally managed and accessed across workspaces with fine-grained permissions (e.g., READ for dev). This is cost-effective (no extra infrastructure or retraining) and secure (governed by Databricks' access controls).

Databricks Reference: "Log models to Unity Catalog via MLflow, then grant access to other workspaces securely" ("MLflow Model



What's more, part of that Pass4cram Databricks-Generative-AI-Engineer-Associate dumps now are free:  
[https://drive.google.com/open?id=1bjXoPRZAwVxwVswclW3\\_wyg\\_MpouqV4m](https://drive.google.com/open?id=1bjXoPRZAwVxwVswclW3_wyg_MpouqV4m)