Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

About Exambible

Your Partner of IT Exam

Found in 1998

Exambible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, Exambible has its unique advantages that other companies could not achieve.

Our Advances

* 99.9% Uptime

All examinations will be up to date.

* 24/7 Quality Support

We will provide service round the clock.

* 100% Pass Rate

Our guarantee that you will pass the exam.

* Unique Gurantee

If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

NEW QUESTION 1
A Generative AI Engineer is creating an agent-based LLM system for their favorite monster truck team. The system can answer text based questions about the monster truck team, lookup event dates via an API call, or query tables on the team??s latest standings.
How could the Generative AI Engineer best design these capabilities into their system?

A. Ingest PDF documents about the monster truck team into a vector store and query it in a RAG architecture.
B. Write a system prompt for the agent listing available tools and bundle it into an agent system that runs a number of calls to solve a query.
C. Instruct the LLM to respond with ??RAG??, ??API??, or ??TABLE?? depending on the query, then use text parsing and conditional statements to resolve the query.
D. Build a system prompt with all possible event dates and table information in the system promp
E. Use a RAG architecture to lookup generic text questions and otherwise leverage the information in the system prompt.

Answer: B

NEW QUESTION 2
A Generative Al Engineer wants their (inetuned LLMs in their prod Databncks 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 Al Engineer to accomplish their gAi?

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

Answer: D

NEW QUESTION 3
A Generative Al Engineer at an automotive company would like to build a question- answering chatbot for customers to inquire about their vehicles. They have a database containing various documents of different vehicle makes, their hardware parts, and common maintenance information.
Which of the following components will NOT be useful in building such a chatbot?

A. Response-generating LLM
B. Invite users to submit long, rather than concise, questions
C. Vector database
D. Embedding model

Answer: B

NEW QUESTION 4
A Generative Al Engineer is building a system that will answer questions on currently unfolding news topics. As such, it pulls information from a variety of sources including articles and social media posts. They are concerned about toxic posts on social media causing toxic outputs from their system.
Which guardrail will limit toxic outputs?

A. Use only approved social media and news accounts to prevent unexpected toxic data from getting to the LLM.
B. Implement rate limiting
C. Reduce the amount of context Items the system will Include in consideration for its response.
D. Log all LLM system responses and perform a batch toxicity analysis monthly.

Answer: A

NEW QUESTION 5
A Generative Al Engineer is developing an LLM application that users can use to generate personalized birthday poems based on their names.
Which technique would be most effective in safeguarding the application, given the potential for malicious user inputs?

A. Implement a safety filter that detects any harmful inputs and ask the LLM to respond that it is unable to assist
B. Reduce the time that the users can interact with the LLM
C. Ask the LLM to remind the user that the input is malicious but continue the conversation with the user
D. Increase the amount of compute that powers the LLM to process input faster

Answer: A

NEW QUESTION 6
A Generative Al 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. DistilBERT
B. MPT-30B
C. Llama2-70B
D. DBRX

Answer: C

NEW QUESTION 7
A Generative Al Engineer has successfully ingested unstructured documents and chunked them by document sections. They would like to store the chunks in a

Vector Search index. The current format of the dataframe has two columns: (i) original document file name (ii) an array of text chunks for each document.
What is the most performant way to store this dataframe?

A. Split the data into train and test set, create a unique identifier for each document, then save to a Delta table
B. Flatten the dataframe to one chunk per row, create a unique identifier for each row, and save to a Delta table
C. First create a unique identifier for each document, then save to a Delta table
D. Store each chunk as an independent JSON file in Unity Catalog Volum
E. For each JSON file, the key is the document section name and the value is the array of text chunks for that section

Answer: B

NEW QUESTION 8
Generative AI Engineer at an electronics company just deployed a RAG application for customers to ask questions about products that the company carries. However, they received feedback that the RAG response often returns information about an irrelevant product.
What can the engineer do to improve the relevance of the RAG??s response?

A. Assess the quality of the retrieved context
B. Implement caching for frequently asked questions
C. Use a different LLM to improve the generated response
D. Use a different semantic similarity search algorithm

Answer: A

NEW QUESTION 9
A Generative AI Engineer is building an LLM-based application that has an important transcription (speech-to-text) task. Speed is essential for the success of the application
Which open Generative AI models should be used?

A. L!ama-2-70b-chat-hf
B. MPT-30B-Instruct
C. DBRX
D. whisper-large-v3 (1.6B)

Answer: D

NEW QUESTION 10
A Generative AI Engineer has been asked to design an LLM-based application that accomplishes the following business objective: answer employee HR questions using HR PDF documentation.
Which set of high level tasks should the Generative AI Engineer's system perform?

A. Calculate averaged embeddings for each HR document, compare embeddings to user query to find the best documen
B. Pass the best document with the user query into an LLM with a large context window to generate a response to the employee.
C. Use an LLM to summarize HR documentatio
D. Provide summaries of documentation and user query into an LLM with a large context window to generate a response to the user.
E. Create an interaction matrix of historical employee questions and HR documentatio
F. Use ALS to factorize the matrix and create embedding
G. Calculate the embeddings of new queries and use them to find the best HR documentatio
H. Use an LLM to generate a response to the employee question based upon the documentation retrieved.
I. Split HR documentation into chunks and embed into a vector stor
J. Use the employee question to retrieve best matched chunks of documentation, and use the LLM to generate a response to the employee based upon the documentation retrieved.

Answer: D

NEW QUESTION 10
A Generative AI Engineer has been asked to build an LLM-based question-answering application. The application should take into account new documents that are frequently published. The engineer wants to build this application with the least cost and least development effort and have it operate at the lowest cost possible.
Which combination of chaining components and configuration meets these requirements?

A. For the application a prompt, a retriever, and an LLM are require
B. The retriever output is inserted into the prompt which is given to the LLM to generate answers.
C. The LLM needs to be frequently with the new documents in order to provide most up-to- date answers.
D. For the question-answering application, prompt engineering and an LLM are required to generate answers.
E. For the application a prompt, an agent and a fine-tuned LLM are require
F. The agent is used by the LLM to retrieve relevant content that is inserted into the prompt which is given to the LLM to generate answers.

Answer: A

NEW QUESTION 12
A team wants to serve a code generation model as an assistant for their software developers. It should support multiple programming languages. Quality is the primary objective.
Which of the Databricks Foundation Model APIs, or models available in the Marketplace, would be the best fit?

A. Llama2-70b
B. BGE-large
C. MPT-7b
D. CodeLlama-34B

Answer: D

NEW QUESTION 15
A Generative AI Engineer is developing a chatbot designed to assist users with insurance- related queries. The chatbot is built on a large language model (LLM) and is conversational. However, to maintain the chatbot??s focus and to comply with company policy, it must not provide responses to questions about politics. Instead, when presented with political inquiries, the chatbot should respond with a standard message:
??Sorry, I cannot answer that. I am a chatbot that can only answer questions around insurance.??
Which framework type should be implemented to solve this?

A. Safety Guardrail
B. Security Guardrail
C. Contextual Guardrail
D. Compliance Guardrail

Answer: A

NEW QUESTION 17
A Generative AI Engineer has developed an LLM application to answer questions about internal company policies. The Generative AI Engineer must ensure that the application doesn??t hallucinate or leak confidential data.
Which approach should NOT be used to mitigate hallucination or confidential data leakage?

A. Add guardrails to filter outputs from the LLM before it is shown to the user
B. Fine-tune the model on your data, hoping it will learn what is appropriate and not
C. Limit the data available based on the user??s access level
D. Use a strong system prompt to ensure the model aligns with your needs.

Answer: B

NEW QUESTION 19
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. Here are a few relevant articles for your browsin
B. Let me know if you have questions after reading them.
C. Please call your local emergency services.
D. Headaches can be toug
E. Hope you feel better soon!
F. Please provide your age, recent activities, and any other symptoms you have noticed along with your headaches and dizziness.

Answer: B

NEW QUESTION 21
A Generative AI Engineer has built an LLM-based system that will automatically translate user text between two languages. They now want to benchmark multiple LLM's on this task and pick the best one. They have an evaluation set with known high quality translation examples. They want to evaluate each LLM using the evaluation set with a performant metric.
Which metric should they choose for this evaluation?

A. ROUGE metric
B. BLEU metric
C. NDCG metric
D. RECALL metric

Answer: B

NEW QUESTION 23
When developing an LLM application, it??s crucial to ensure that the data used for training the model complies with licensing requirements to avoid legal risks.
Which action is NOT appropriate to avoid legal risks?

A. Reach out to the data curators directly before you have started using the trained model to let them know.
B. Use any available data you personally created which is completely original and you can decide what license to use.
C. Only use data explicitly labeled with an open license and ensure the license terms are followed.
D. Reach out to the data curators directly after you have started using the trained model to let them know.

Answer: D

NEW QUESTION 25
A Generative AI Engineer is deciding between using LSH (Locality Sensitive Hashing) and HNSW (Hierarchical Navigable Small World) for indexing their vector database Their top priority is semantic accuracy
Which approach should the Generative AI Engineer use to evaluate these two techniques?

A. Compare the cosine similarities of the embeddings of returned results against those of a representative sample of test inputs
B. Compare the Bilingual Evaluation Understudy (BLEU) scores of returned results for a representative sample of test inputs
C. Compare the Recall-Onented-Understudy for Gistmg Evaluation (ROUGE) scores of returned results for a representative sample of test inputs
D. Compare the Levenshtein distances of returned results against a representative sample of test inputs

Answer: A

NEW QUESTION 28
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. Diversity of responses
C. Lack of relevance
D. Repetition of responses

Answer: B

NEW QUESTION 31
A Generative AI Engineer would like an LLM to generate formatted JSON from emails. This will require parsing and extracting the following information: order ID, date, and sender email. Here??s a sample email:

```
Date: April 23, 2024
Time: 4:22 PM
From: anjali.thayer@computex.org
To: cust_service@realtek.com
Subject: Shipment details

Hey there,

I have a shipment (order ID is CD34RFT) can you please send me an update?

Thank you,
Anjali
```

They will need to write a prompt that will extract the relevant information in JSON format with the highest level of output accuracy.
Which prompt will do that?

A. You will receive customer emails and need to extract date, sender email, and order I
B. You should return the date, sender email, and order ID information in JSON format.
C. You will receive customer emails and need to extract date, sender email, and order I
D. Return the extracted information in JSON format.Here??s an example: {??date??: ??April 16, 2024??, ??sender_email??: ??sarah.lee925@gmail.com??,??order_id??: ??RE987D??}
E. You will receive customer emails and need to extract date, sender email, and order I
F. Return the extracted information in a human-readable format.
G. You will receive customer emails and need to extract date, sender email, and order I
H. Return the extracted information in JSON format.

Answer: B

NEW QUESTION 32
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. Dolly 1.5B
B. OpenAI GPT-4
C. BGE-large
D. Llama2-70B

Answer: C

NEW QUESTION 33
A Generative AI Engineer is building a RAG application that will rely on context retrieved from source documents that are currently in PDF format. These PDFs can contain both text and images. They want to develop a solution using the least amount of lines of code.
Which Python package should be used to extract the text from the source documents?

A. flask
B. beautifulsoup
C. unstructured
D. numpy

Answer: B

NEW QUESTION 38

A Generative AI Engineer has created a RAG application to look up answers to questions about a series of fantasy novels that are being asked on the author??s web forum. The fantasy novel texts are chunked and embedded into a vector store with metadata (page
number, chapter number, book title), retrieved with the user??s query, and provided to an LLM for response generation. The Generative AI Engineer used their intuition to pick the chunking strategy and associated configurations but now wants to more methodically choose the best values.
Which TWO strategies should the Generative AI Engineer take to optimize their chunking strategy and parameters? (Choose two.)

A. Change embedding models and compare performance.
B. Add a classifier for user queries that predicts which book will best contain the answe
C. Use this to filter retrieval.
D. Choose an appropriate evaluation metric (such as recall or NDCG) and experiment with changes in the chunking strategy, such as splitting chunks by paragraphs or chapter
E. Choose the strategy that gives the best performance metric.
F. Pass known questions and best answers to an LLM and instruct the LLM to provide the best token coun
G. Use a summary statistic (mean, median, etc.) of the best token counts to choose chunk size.
H. Create an LLM-as-a-judge metric to evaluate how well previous questions are answered by the most appropriate chun
I. Optimize the chunking parameters based upon the values of the metric.

Answer: CE

NEW QUESTION 40
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 planningwhich 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 a *.wav files, but also a text transcript as *.txt files.
call_cust_history: a Delta table with primary keys customer_id, cal1_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_cust_history
B. maintenance_schedule
C. call_rep_history
D. call_detail
E. transcript Volume

Answer: DE

NEW QUESTION 42
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. Vector Search
B. Lakeview
C. DBSQL
D. Inference Tables

Answer: D

NEW QUESTION 44
.....

Relate Links

100% Pass Your Databricks-Generative-AI-Engineer-Associate Exam with Exambible Prep Materials

https://www.exambible.com/Databricks-Generative-AI-Engineer-Associate-exam/

Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - https://www.exambible.com/