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
What is the most suitable library for building a multi-step LLM-based workflow?

A. Pandas
B. TensorFlow
C. PySpark
D. LangChain

Answer: D

NEW QUESTION 2
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. Inform the user of the expected RAG behavior
C. Restrict access to the data sources to a limited number of users
D. Curate upstream data properly that includes manual review before it is fed into the RAG system

Answer: D

NEW QUESTION 3
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. Vector Stores
C. Conversation Buffer Memory
D. External tools
E. Chat loaders
F. React Components

Answer: BC

NEW QUESTION 4
A Generative AI 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 5
A Generative AI Engineer is creating an LLM-powered application that will need access to up-to-date news articles and stock prices.
The design requires the use of stock prices which are stored in Delta tables and finding the latest relevant news articles by searching the internet.
How should the Generative AI Engineer architect their LLM system?

A. Use an LLM to summarize the latest news articles and lookup stock tickers from the summaries to find stock prices.
B. Query the Delta table for volatile stock prices and use an LLM to generate a search query to investigate potential causes of the stock volatility.
C. Download and store news articles and stock price information in a vector stor
D. Use a RAG architecture to retrieve and generate at runtime.
E. Create an agent with tools for SQL querying of Delta tables and web searching, provide retrieved values to an LLM for generation of response.

Answer: D

NEW QUESTION 6
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 7
A Generative Al Engineer needs to design an LLM pipeline to conduct multi-stage reasoning that leverages external tools. To be effective at this, the LLM will need to plan and adapt actions while performing complex reasoning tasks.
Which approach will do this?

A. Tram the LLM to generate a single, comprehensive response without interacting with any external tools, relying solely on its pre-trained knowledge.
B. Implement a framework like ReAct which allows the LLM to generate reasoning traces and perform task-specific actions that leverage external tools if necessary.
C. Encourage the LLM to make multiple API calls in sequence without planning or structuring the calls, allowing the LLM to decide when and how to use external tools spontaneously.
D. Use a Chain-of-Thought (CoT) prompting technique to guide the LLM through a series of reasoning steps, then manually input the results from external tools for the final answer.

Answer: B

NEW QUESTION 8
A Generative Al Engineer is creating an LLM-based application. The documents for its retriever have been chunked to a maximum of 512 tokens each. The Generative Al Engineer knows that cost and latency are more important than quality for this application. They have several context length levels to choose from.
Which will fulfill their need?

A. context length 514; smallest model is 0.44GB and embedding dimension 768
B. context length 2048: smallest model is 11GB and embedding dimension 2560
C. context length 32768: smallest model is 14GB and embedding dimension 4096
D. context length 512: smallest model is 0.13GB and embedding dimension 384

Answer: D

NEW QUESTION 9
A Generative Al 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. Increase the compute to improve processing speed of questions to allow greater relevancy analysisC Implement a profanity filter to screen out offensive language
C. Incorporate manual reviews to correct any problematic outputs prior to sending to the users

Answer: B

NEW QUESTION 10
A Generative AI Engineer has created a RAG application which can help employees retrieve answers from an internal knowledge base, such as Confluence pages or Google Drive. The prototype application is now working with some positive feedback from internal company testers. Now the Generative AI Engineer wants to formally evaluate the system??s performance and understand where to focus their efforts to further improve the system.
How should the Generative AI Engineer evaluate the system?

A. Use cosine similarity score to comprehensively evaluate the quality of the final generated answers.
B. Curate a dataset that can test the retrieval and generation components of the system separatel
C. Use MLflow??s built in evaluation metrics to perform the evaluation on the retrieval and generation components.
D. Benchmark multiple LLMs with the same data and pick the best LLM for the job.
E. Use an LLM-as-a-judge to evaluate the quality of the final answers generated.

Answer: B

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 11
A Generative AI Engineer developed an LLM application using the provisioned throughput Foundation Model API. Now that the application is ready to be deployed, they realize their volume of requests are not sufficiently high enough to create their own provisioned throughput endpoint. They want to choose a strategy that ensures the best cost- effectiveness for their application.
What strategy should the Generative AI Engineer use?

A. Switch to using External Models instead
B. Deploy the model using pay-per-token throughput as it comes with cost guarantees
C. Change to a model with a fewer number of parameters in order to reduce hardware constraint issues
D. Throttle the incoming batch of requests manually to avoid rate limiting issues

Answer: B

NEW QUESTION 16
After changing the response generating LLM in a RAG pipeline from GPT-4 to a model with a shorter context length that the company self-hosts, the Generative AI

Engineer is getting the following error:

```
{"error_code": "BAD_REQUEST", "message": "Bad request: rpc error:
code = InvalidArgument desc = prompt token count (4595) cannot
exceed 4096..."}
```

What TWO solutions should the Generative AI Engineer implement without changing the response generating model? (Choose two.)

A. Use a smaller embedding model to generate
B. Reduce the maximum output tokens of the new model
C. Decrease the chunk size of embedded documents
D. Reduce the number of records retrieved from the vector database
E. Retrain the response generating model using ALiBi

Answer: CD

NEW QUESTION 20
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 21
A Generative Al 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 26
A small and cost-conscious startup in the cancer research field wants to build a RAG application using Foundation Model APIs.
Which strategy would allow the startup to build a good-quality RAG application while being cost-conscious and able to cater to customer needs?

A. Limit the number of relevant documents available for the RAG application to retrieve from
B. Pick a smaller LLM that is domain-specific
C. Limit the number of queries a customer can send per day
D. Use the largest LLM possible because that gives the best performance for any general queries

Answer: B

NEW QUESTION 28
A Generative Al Engineer is working with a retail company that wants to enhance its customer experience by automatically handling common customer inquiries. They are working on an LLM-powered Al solution that should improve response times while maintaining a personalized interaction. They want to define the appropriate input and LLM task to do this.
Which input/output pair will do this?

A. Input: Customer reviews; Output Group the reviews by users and aggregate per-user average rating, then respond
B. Input: Customer service chat logs; Output Group the chat logs by users, followed by summarizing each user's interactions, then respond
C. Input: Customer service chat logs; Output: Find the answers to similar questions and respond with a summary
D. Input: Customer reviews: Output Classify review sentiment

Answer: C

NEW QUESTION 30
A Generative Al Engineer is tasked with deploying an application that takes advantage of a custom MLflow Pyfunc model to return some interim results.
How should they configure the endpoint to pass the secrets and credentials?

A. Use spark.conf.set ()
B. Pass variables using the Databricks Feature Store API
C. Add credentials using environment variables
D. Pass the secrets in plain text

Answer: C

NEW QUESTION 35
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 36
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 39
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. Split the LLM output by newline characters to truncate away the summarization explanation.
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. Provide few shot examples of desired output format to the system and/or user prompt.

Answer: D

NEW QUESTION 40
A Generative AI Engineer just deployed an LLM application at a digital marketing company that assists with answering customer service inquiries.
Which metric should they monitor for their customer service LLM application in production?

A. Number of customer inquiries processed per unit of time
B. Energy usage per query
C. Final perplexity scores for the training of the model
D. HuggingFace Leaderboard values for the base LLM

Answer: A

NEW QUESTION 41
A Generative AI Engineer is testing a simple prompt template in LangChain using the code below, but is getting an error.

```
from langchain.chains import LLMChain
from langchain_community.llms import OpenAI
from langchain_core.prompts import PromptTemplate

prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
    input_variables=["adjective"],
    template=prompt_template
)

llm = LLMChain(prompt=prompt)
llm.generate([{"adjective": "funny"}])
```

Assuming the API key was properly defined, what change does the Generative AI Engineer need to make to fix their chain?

A)

```
prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
    input_variables=["adjective"],
    template=prompt_template
)

llm = LLMChain(prompt=prompt)
llm.generate("funny")
```

B)

```
prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
    input_variables=["adjective"],
    template=prompt_template
)

llm = LLMChain(prompt=prompt.format("funny"))
llm.generate()
```

C)

```
prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
    input_variables=["adjective"],
    template=prompt_template
    llm=OpenAI()
)

llm = LLMChain(prompt=prompt)
llm.generate([{"adjective": "funny"}])
```

D)

```
prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
    input_variables=["adjective"],
    template=prompt_template
)

llm = LLMChain(llm=OpenAI(), prompt=prompt)
llm.generate([{"adjective": "funny"}])
```

A. Option A
B. Option B
C. Option C
D. Option D

Answer: C

NEW QUESTION 46

.....

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/