



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 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 AI 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 2

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

Answer: B

NEW QUESTION 5

A Generative AI 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 6

A Generative AI 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

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 8

A Generative AI 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 9

A Generative AI Engineer is creating an LLM-based application. The documents for its retriever have been chunked to a maximum of 512 tokens each. The Generative AI 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 10

A Generative AI 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 analysis
- C 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 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 15

A Generative AI Engineer is using an LLM to classify species of edible mushrooms based on text descriptions of certain features. The model is returning accurate responses in testing and the Generative AI Engineer is confident they have the correct list of possible labels, but the output frequently contains additional reasoning in the answer when the Generative AI Engineer only wants to return the label with no additional text. Which action should they take to elicit the desired behavior from this LLM?

- A. Use few snot prompting to instruct the model on expected output format
- B. Use zero shot prompting to instruct the model on expected output format
- C. Use zero shot chain-of-thought prompting to prevent a verbose output format
- D. Use a system prompt to instruct the model to be succinct in its answer

Answer: D

NEW QUESTION 18

A Generative AI Engineer is ready to deploy an LLM application written using Foundation Model APIs. They want to follow security best practices for production scenarios

Which authentication method should they choose?

- A. Use an access token belonging to service principals
- B. Use a frequently rotated access token belonging to either a workspace user or a service principal
- C. Use OAuth machine-to-machine authentication
- D. Use an access token belonging to any workspace user

Answer: A

NEW QUESTION 23

A Generative AI Engineer I using the code below to test setting up a vector store:

```
from databricks.vector_search.client import VectorSearchClient

vsc = VectorSearchClient()

vsc.create_endpoint(
    name= "vector_search_test",
    endpoint_type= "STANDARD"
)
```

Assuming they intend to use Databricks managed embeddings with the default embedding model, what should be the next logical function call?

- A. vsc.get_index()
- B. vsc.create_delta_sync_index()
- C. vsc.create_direct_access_index()
- D. vsc.similarity_search()

Answer: B

NEW QUESTION 24

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 25

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 29

A Generative AI Engineer is setting up a Databricks Vector Search that will lookup news articles by topic within 10 days of the date specified An example query might be "Tell me about monster truck news around January 5th 1992". They want to do this with the least amount of effort.

How can they set up their Vector Search index to support this use case?

- A. Split articles by 10 day blocks and return the block closest to the query.
- B. Include metadata columns for article date and topic to support metadata filtering.
- C. pass the query directly to the vector search index and return the best articles.
- D. Create separate indexes by topic and add a classifier model to appropriately pick the best index.

Answer: B

NEW QUESTION 32

A Generative AI Engineer is designing an LLM-powered live sports commentary platform. The platform provides real-time updates and LLM-generated analyses for any users who would like to have live summaries, rather than reading a series of potentially outdated news articles.

Which tool below will give the platform access to real-time data for generating game analyses based on the latest game scores?

- A. DatabricksIQ
- B. Foundation Model APIs
- C. Feature Serving
- D. AutoML

Answer: C

NEW QUESTION 33

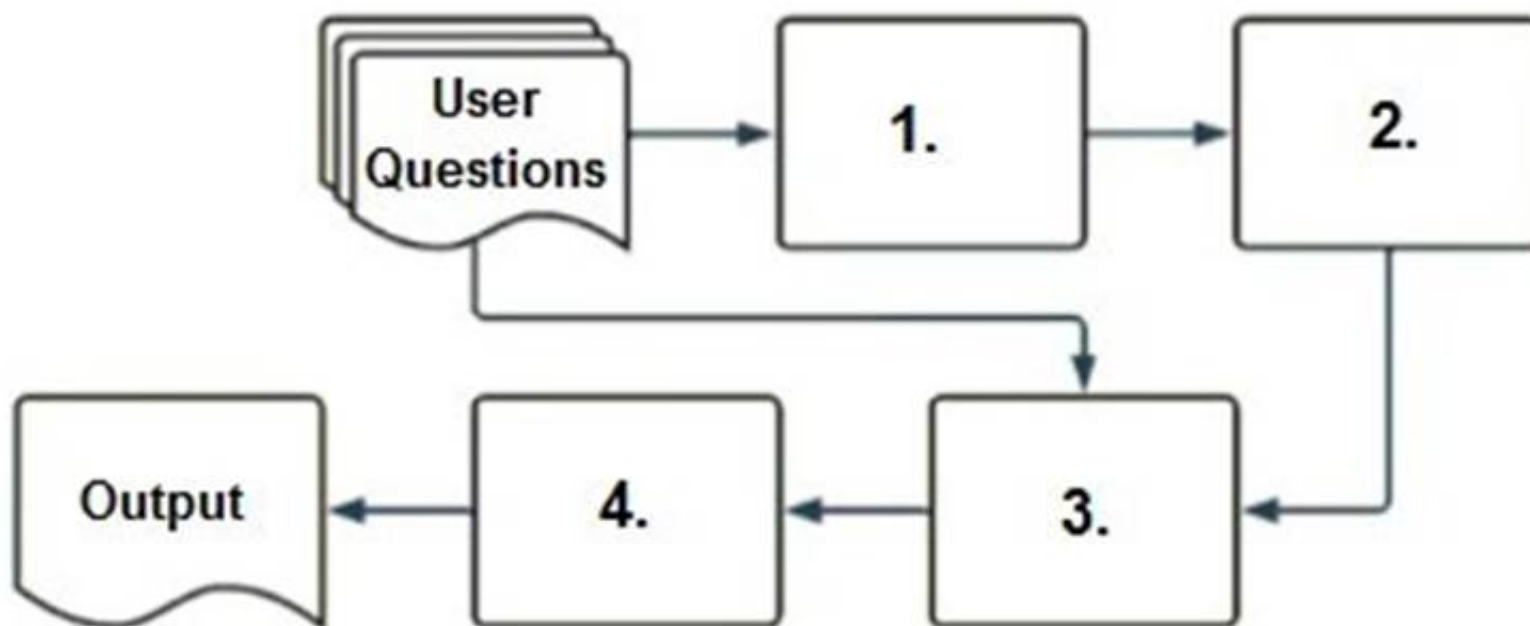
A Generative AI Engineer is building a Generative AI system that suggests the best matched employee team member to newly scoped projects. The team member is selected from a very large team. The match should be based upon project date availability and how well their employee profile matches the project 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 date
- B. 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 team member availability given project dates, and another tool that uses an LLM to extract keywords from project scope
- D. Iterate through available team members?? profiles and perform keyword matching to find the best available team member.
- E. Create a tool to find available team members given project date
- F. Create a second tool that can calculate a similarity score for a combination of team member profile and the project scop
- G. Iterate through the team members and rank by best score to select a team member.
- H. Create a tool for finding available team members given project date
- I. 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.

Answer: D

NEW QUESTION 35

A company has a typical RAG-enabled, customer-facing chatbot on its website.



Select the correct sequence of components a user's questions will go through before the final output is returned. Use the diagram above for reference.

- A. 1.embedding model, 2.vector search, 3.context-augmented prompt, 4.response- generating LLM
- B. 1.context-augmented prompt, 2.vector search, 3.embedding model, 4.response- generating LLM
- C. 1.response-generating LLM, 2.vector search, 3.context-augmented prompt, 4.embedding model
- D. 1.response-generating LLM, 2.context-augmented prompt, 3.vector search, 4.embedding model

Answer: A

NEW QUESTION 38

A Generative AI 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 42

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 47

What is an effective method to preprocess prompts using custom code before sending them to an LLM?

- A. Directly modify the LLM's internal architecture to include preprocessing steps
- B. It is better not to introduce custom code to preprocess prompts as the LLM has not been trained with examples of the preprocessed prompts
- C. Rather than preprocessing prompts, it's more effective to postprocess the LLM outputs to align the outputs to desired outcomes
- D. Write a MLflow PyFunc model that has a separate function to process the prompts

Answer: D

NEW QUESTION 50

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 52

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 answer.
- 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 count
- 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 chunk
- I. Optimize the chunking parameters based upon the values of the metric.

Answer: CE

NEW QUESTION 57

A Generative AI Engineer is helping a cinema extend its website's chat bot to be able to respond to questions about specific showtimes for movies currently playing at their local theater. They already have the location of the user provided by location services to their agent, and a Delta table which is continually updated with the latest showtime information by location. They want to implement this new capability in their RAG application.

Which option will do this with the least effort and in the most performant way?

- A. Create a Feature Serving Endpoint from a FeatureSpec that references an online store synced from the Delta table
- B. Query the Feature Serving Endpoint as part of the agent logic/ tool implementation.
- C. Query the Delta table directly via a SQL query constructed from the user's input using a text-to-SQL LLM in the agent logic / tool implementation
- D. Write the Delta table contents to a text column, then embed those texts using an embedding model and store these in the vector index. Lookup the information based on the embedding as part of the agent logic / tool implementation.
- E. Set up a task in Databricks Workflows to write the information in the Delta table periodically to an external database such as MySQL and query the information from there as part of the agent logic / tool implementation.

Answer: A

NEW QUESTION 59

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 63

.....

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/>