

# 便利Databricks Databricks-Generative-AI-Engineer-Associate | 正確的なDatabricks-Generative-AI-Engineer-Associate試験対応試験 | 試験の準備方法Databricks Certified Generative AI Engineer Associate最新試験情報



P.S. JpexamがGoogle Driveで共有している無料かつ新しいDatabricks-Generative-AI-Engineer-Associateダンプ: <https://drive.google.com/open?id=18q62aRQV-n6Z5p7TUa5m3-Yek5J2Vnbs>

最高のサービスを提供することを義務と考えています。そのため、患者の同僚が24時間年中無休でサポートを提供し、Databricks-Generative-AI-Engineer-Associate実践教材に関する問題をすべて解決します。あなたが私たちを必要とする限り、私たちは思いやりのあるサービスを提供しています。それに、一生懸命努力しながら失敗することは不名誉ではありません。残念ながらDatabricks-Generative-AI-Engineer-Associateスタディガイドで試験に不合格になった場合、他のバージョンに切り替えるか、今回は不合格であると仮定して全額返金し、不合格書類で証明します。あなたの能力を過小評価しないでください。Databricks-Generative-AI-Engineer-Associateの実際のテストを試みている間、私たちはあなたの最強のバックアップになります。

## Databricks Databricks-Generative-AI-Engineer-Associate 認定試験の出題範囲:

トピック	出題範囲
トピック 1	<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>
トピック 2	<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>
トピック 3	<ul style="list-style-type: none"> <li>• 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.</li> </ul>

>> Databricks-Generative-AI-Engineer-Associate試験対応 <<

## Databricks-Generative-AI-Engineer-Associate最新試験情報、Databricks-Generative-AI-Engineer-Associate参考資料

なぜみんなが順調にDatabricksのDatabricks-Generative-AI-Engineer-Associate試験に合格できることに対する好奇心がありますか。DatabricksのDatabricks-Generative-AI-Engineer-Associate試験に合格したいんですか。実は、彼らが試験に合格したコツは我々Jpexamの提供するDatabricksのDatabricks-Generative-AI-Engineer-Associate試験ソフトを利用したんです。豊富の問題集、専門的な研究と購入の後の一年間の無料更新、ソフトで復習して、自分の能力の高めを感じられます。DatabricksのDatabricks-Generative-AI-Engineer-Associate試験に合格することができます。

### Databricks Certified Generative AI Engineer Associate 認定 Databricks-Generative-AI-Engineer-Associate 試験問題 (Q55-Q60):

#### 質問 # 55

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

正解: B

解説:

In the context of designing a chatbot to engage users on a gaming platform, diversity of responses (option B) is a key metric to increase user engagement and retention. Here's why:

\* **Diverse and Engaging Interactions:** A chatbot that provides varied and interesting responses will keep users engaged, especially in an interactive environment like a gaming platform. Gamers typically enjoy dynamic and evolving conversations, and diversity of responses helps prevent monotony, encouraging users to interact more frequently with the bot.

\* **Increasing Retention:** By offering different types of responses to similar queries, the chatbot can create a sense of novelty and excitement, which enhances the user's experience and makes them more likely to return to the platform.

\* **Why Other Options Are Less Effective:**

\* **A (Randomness):** Random responses can be confusing or irrelevant, leading to frustration and reducing engagement.

\* **C (Lack of Relevance):** If responses are not relevant to the user's queries, this will degrade the user experience and lead to disengagement.

\* **D (Repetition of Responses):** Repetitive responses can quickly bore users, making the chatbot feel uninteresting and reducing the likelihood of continued interaction.

Thus, diversity of responses (option B) is the most effective way to keep users engaged and retain them on the platform.

#### 質問 # 56

A Generative AI Engineer interfaces with an LLM with prompt/response behavior that has been trained on customer calls inquiring about product availability. The LLM is designed to output "In Stock" if the product is available or only the term "Out of Stock" if not. Which prompt will work to allow the engineer to respond to call classification labels correctly?

- **A. You will be given a customer call transcript where the customer asks about product availability. The outputs are either "In Stock" or "Out of Stock". Format the output in JSON, for example: {"call\_id": "123", "label": "In Stock"}.**
- B. You will be given a customer call transcript where the customer inquires about product availability. Respond with "In Stock" if the product is available or "Out of Stock" if not.
- C. Respond with "In Stock" if the customer asks for a product.
- D. Respond with "Out of Stock" if the customer asks for a product.

正解: A

解説:

\* **Problem Context:** The Generative AI Engineer needs a prompt that will enable an LLM trained on customer call transcripts to classify and respond correctly regarding product availability. The desired response should clearly indicate whether a product is "In Stock" or "Out of Stock," and it should be formatted in a way that is structured and easy to parse programmatically, such as JSON.

\* Explanation of Options:

\* Option A: Respond with "In Stock" if the customer asks for a product. This prompt is too generic and does not specify how to handle the case when a product is not available, nor does it provide a structured output format.

\* Option B: This option is correctly formatted and explicit. It instructs the LLM to respond based on the availability mentioned in the customer call transcript and to format the response in JSON.

This structure allows for easy integration into systems that may need to process this information automatically, such as customer service dashboards or databases.

\* Option C: Respond with "Out of Stock" if the customer asks for a product. Like option A, this prompt is also insufficient as it only covers the scenario where a product is unavailable and does not provide a structured output.

\* Option D: While this prompt correctly specifies how to respond based on product availability, it lacks the structured output format, making it less suitable for systems that require formatted data for further processing.

Given the requirements for clear, programmatically usable outputs, Option B is the optimal choice because it provides precise instructions on how to respond and includes a JSON format example for structuring the output, which is ideal for automated systems or further data handling.

### 質問 # 57

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

正解: C、E

解説:

\* Problem Context: After switching to a model with a shorter context length, the error message indicating that the prompt token count has exceeded the limit suggests that the input to the model is too large.

\* Explanation of Options:

\* Option A: Use a smaller embedding model to generate- This wouldn't necessarily address the issue of prompt size exceeding the model's token limit.

\* Option B: Reduce the maximum output tokens of the new model- This option affects the output length, not the size of the input being too large.

\* Option C: Decrease the chunk size of embedded documents- This would help reduce the size of each document chunk fed into the model, ensuring that the input remains within the model's context length limitations.

\* Option D: Reduce the number of records retrieved from the vector database- By retrieving fewer records, the total input size to the model can be managed more effectively, keeping it within the allowable token limits.

\* Option E: Retrain the response generating model using ALiBi- Retraining the model is contrary to the stipulation not to change the response generating model.

Options C and E are the most effective solutions to manage the model's shorter context length without changing the model itself, by adjusting the input size both in terms of individual document size and total documents retrieved.

### 質問 # 58

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

正解: C

解説:

In this case, the Generative AI Engineer is developing an application to generate personalized birthday poems, but there's a need to safeguard against malicious user inputs. The best solution is to implement a safety filter (option A) to detect harmful or inappropriate inputs.

\* Safety Filter Implementation: Safety filters are essential for screening user input and preventing inappropriate content from being processed by the LLM. These filters can scan inputs for harmful language, offensive terms, or malicious content and intervene before the prompt is passed to the LLM.

\* Graceful Handling of Harmful Inputs: Once the safety filter detects harmful content, the system can provide a message to the user, such as "I'm unable to assist with this request," instead of processing or responding to malicious input. This protects the system from generating harmful content and ensures a controlled interaction environment.

\* Why Other Options Are Less Suitable:

\* B (Reduce Interaction Time): Reducing the interaction time won't prevent malicious inputs from being entered.

\* C (Continue the Conversation): While it's possible to acknowledge malicious input, it is not safe to continue the conversation with harmful content. This could lead to legal or reputational risks.

\* D (Increase Compute Power): Adding more compute doesn't address the issue of harmful content and would only speed up processing without resolving safety concerns.

Therefore, implementing a safety filter that blocks harmful inputs is the most effective technique for safeguarding the application.

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

正解: D

解説:

When deploying an LLM application for customer service inquiries, the primary focus is on measuring the operational efficiency and quality of the responses. Here's why A is the correct metric:

\* Number of customer inquiries processed per unit of time: This metric tracks the throughput of the customer service system, reflecting how many customer inquiries the LLM application can handle in a given time period (e.g., per minute or hour). High throughput is crucial in customer service applications where quick response times are essential to user satisfaction and business efficiency.

\* Real-time performance monitoring: Monitoring the number of queries processed is an important part of ensuring that the model is performing well under load, especially during peak traffic times. It also helps ensure the system scales properly to meet demand.

Why other options are not ideal:

\* B. Energy usage per query: While energy efficiency is a consideration, it is not the primary concern for a customer-facing application where user experience (i.e., fast and accurate responses) is critical.

\* C. Final perplexity scores for the training of the model: Perplexity is a metric for model training, but it doesn't reflect the real-time operational performance of an LLM in production.

\* D. HuggingFace Leaderboard values for the base LLM: The HuggingFace Leaderboard is more relevant during model selection and benchmarking. However, it is not a direct measure of the model's performance in a specific customer service application in production.

Focusing on throughput (inquiries processed per unit time) ensures that the LLM application is meeting business needs for fast and efficient customer service responses.

### 質問 # 60

.....

時代に対応するために、科学技術は人々の学習方法を向上させると信じています。特にこのようなペースの速い生活テンポでは、効率の高い学習を非常に重視しています。したがって、当社のDatabricks-Generative-AI-Engineer-Associate学習資料は、過去の試験問題と現在の試験の傾向に基づいており、実際の試験環境に配置するためのこのような効果的なシミュレーション機能を設計します。高度なDatabricks-Generative-AI-Engineer-Associate

