

PRACTICE QUESTIONS

Alibaba Cloud

ACA Generative AI Engineer

SHARED BY [u/zesty_beacon_74](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM Alibaba Cloud ACA Generative AI Engineer

Shared by [u/zesty_beacon_74](#) on aiexamcerts

10 practice questions for the Alibaba Cloud ACA Generative AI Engineer exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/alibaba-cloud/aca-generative-ai-engineer. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. In an LLM request, what does the term token refer to?

- A** A signed credential that authorizes each API call to a model endpoint for a limited period of time
- B** A fixed-length vector stored in a vector database that represents the full meaning of one document
- C** A billing unit that is always exactly one thousand characters of input, regardless of the language used
- D** A unit of text, such as a word piece, that the model reads and generates as one step over its vocabulary

Answer: D

Tokens are the units produced by the tokenizer, and they can be whole words, fragments or single characters depending on the language. Usage and billing are measured in tokens, but the number varies by text. Credentials and embeddings are different concepts that are often confused with tokens.

2. Increasing a model's temperature makes its outputs more deterministic.

- A** True
- B** False

Answer: B (False)

Higher temperature flattens the sampling distribution, which makes output more varied and less deterministic. Lower temperature is the setting that makes output more repeatable. Reversing the direction is the common mistake this statement checks.

3. Why is the system prompt usually kept separate from end-user messages in a chat application?
- A It is stored in the vector database so that it can be retrieved by similarity search for each user query in the chat
 - B It replaces the need for an API key because the provider authenticates the request from its text
 - C It is charged at a higher rate than user messages because it is always the longest part of the request
 - D It defines the assistant's role, constraints and style, and it is sent as its own instruction layer in the request

Answer: D

The system role carries the application's rules and persona, and separating it from user content keeps the structure of the conversation clear. It is not a storage location for retrieval or an authentication mechanism. Billing depends on token counts rather than on the message role.

4. Numeric vectors that represent the meaning of text so that similar items can be found by similarity search are called _____.

Answer: embeddings (also accepted: vector embeddings, embedding vectors)

Embeddings are dense vectors produced by an embedding model, and their distances reflect semantic similarity. They are stored in vector indexes for retrieval.

5. Which two signals suggest that a fine-tuning run should be stopped or revisited? (Choose two.)
- A Validation loss rising while training loss keeps falling, which points to overfitting
 - B Training loss falling smoothly over the first few hundred steps of the run
 - C Evaluation scores on the held-out set falling below the base model's baseline on the same test
 - D The job using more GPU memory than the previous run did with the same batch size
 - E A higher number of checkpoints saved to the storage bucket than was planned in the schedule

Answer: A, C

Diverging validation and training loss signal memorization, and held-out scores below the base model show the tuning made things worse. Smooth early loss and memory or checkpoint counts are expected or operational details. Decisions should follow generalization metrics.

6. A prompt works well on one model but poorly after the model is upgraded. What is the best first step?
- A Revert to the old model permanently and never test the new version against the application again
 - B Rerun a prompt regression set against the new version, then adjust the prompt and record each change
 - C Rewrite the entire application in a different programming language before testing any prompt at all
 - D Ask end users to rephrase every question until the answers look acceptable on the new model version

Answer: B

A fixed regression set shows exactly which cases changed, and the prompt can then be tuned and versioned against that evidence. Permanent reversion and full rewrites are costly and skip diagnosis. Asking users to rephrase shifts the burden of quality onto them.

7. In retrieval-augmented generation, what is the main job of the retriever?
- A To train the language model on the company's documents so that the knowledge is stored in its weights
 - B To generate the final answer from scratch using only the model's memory of the world from pre-training
 - C To find the document chunks most relevant to the user's question so that they can be added to the prompt
 - D To encrypt the user's question before it is sent to the model endpoint for processing by the service

Answer: C

The retriever selects evidence from the knowledge base, and the generator then answers using that evidence. Training on documents is fine-tuning, which RAG avoids for freshness reasons. Encryption is a transport concern, not a retrieval step.

8. LoRA is a method that keeps the base model weights frozen while training small adapter matrices.
- A True
 - B False

Answer: A (True)

LoRA learns low-rank updates in adapter matrices while the original weights stay unchanged, which saves memory and makes adapters easy to swap. The statement describes the defining property of the method. The distractor is that LoRA retrains every parameter.

9. What is the main benefit of caching in LLM serving?

- A** Repeated prefixes or identical requests avoid recomputation, which lowers latency and cost
- B** Cached responses are always more accurate than fresh answers because they were generated at an earlier point in time
- C** Caching removes the need for monitoring, since stored answers cannot change over time during their lifetime
- D** Caching makes the model learn from each request, which gradually improves its weights over the long term

Answer: A

Prefix and response caches reuse computation that would otherwise be repeated, which helps latency and cost when traffic repeats. Cached answers can be stale, so freshness and monitoring still matter. Caching does not change model weights.

10. A deployed model's latency rises sharply under peak traffic. Which action is the best first step?

- A** Turn off logging entirely so that the server has more spare capacity to answer the incoming requests
- B** Increase the prompt length with extra examples so that the model has more context to process each time
- C** Retrain the model on more data, which always reduces the latency of each individual request it serves
- D** Check throughput, queueing and token counts per request, then scale replicas or batch requests where appropriate

Answer: D

Measuring where time goes, such as queueing versus generation and long token counts, points to the right fix before spending money. Scaling replicas or batching then addresses the measured bottleneck. Turning off logs, lengthening prompts and retraining do not target the cause of queueing.