

PRACTICE QUESTIONS

Alibaba Cloud

ACA Generative AI Engineer

SHARED BY [u/Upbeat-Kernel-4233](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM Alibaba Cloud ACA Generative AI Engineer

Shared by [u/Upbeat-Kernel-4233](#) 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. What does pre-training a large language model typically teach it?

- A** To predict the next token over a very large text corpus, which builds general language patterns and world knowledge
- B** To follow a fixed list of company-specific rules that the developer writes into the system prompt by hand
- C** To store every training document verbatim so that it can be retrieved exactly by a search index and quoted back to any user who asks for it later
- D** To classify each incoming request into one of a small set of predefined business categories defined in advance

Answer: A

Pre-training uses next-token prediction on broad text, and that objective is what gives the model its general language and knowledge. Company rules belong in prompts, and a model does not store documents verbatim as a search index would. Classification is a downstream task that usually needs its own fine-tuning or prompting.

2. The architecture that relies on self-attention and is used by most modern LLMs is called the _____.

Answer: transformer

Most modern large language models are transformers, which stack self-attention and feed-forward layers. The name identifies the architecture rather than a product.

3. Which statement about the Qwen model family is accurate?

- A** All Qwen models are closed and can only be used through a single hosted endpoint, with no self-hosting option of any kind for teams
- B** Qwen is a vector database product that is sold separately from any language model or inference service
- C** Qwen is a speech recognition model that only transcribes audio and is not able to generate text responses
- D** Several Qwen models are released with open weights that teams can self-host, and the model service also offers hosted Qwen models

Answer: D

Qwen spans open-weight models that teams can run themselves and hosted models that Alibaba Cloud serves through its model service. License terms differ between releases, so teams should check each one. The distractors wrongly describe Qwen as closed-only, a database, or a transcription-only model.

4. Fine-tuning replaces the need for a retrieval step when the knowledge base changes weekly.
- A True
 - B False

Answer: B (False)

Fine-tuning bakes knowledge into weights, which is slow and costly to repeat every week, and the result cannot be cited to a source. Retrieval updates the evidence by re-indexing without retraining. Fine-tuning and retrieval solve different problems.

5. Which two techniques can improve retrieval recall in a RAG pipeline? (Choose two.)
- A Lowering the number of retrieved chunks to the minimum of a single document
 - B Removing metadata from the index so that it is smaller and faster to scan
 - C Hybrid search that combines keyword results with vector similarity results
 - D Using a larger temperature value when the final answer is generated from the results
 - E Query rewriting or expansion before the retrieval step runs

Answer: C, E

Hybrid search finds both exact-term and paraphrased matches, and query rewriting bridges wording gaps, so more relevant evidence is retrieved. Returning fewer chunks, dropping metadata and raising temperature all reduce recall or do not affect it. Recall is a retrieval property.

6. Which prompting approach shows the model a handful of worked examples before the real input?
- A Sending only a few requests per minute so that the provider does not throttle the account during peak hours
 - B Limiting every answer to a few words so that responses are short enough to display on a mobile screen
 - C Including a small number of worked input and output examples in the prompt so that the model imitates their pattern
 - D Fine-tuning the model on a few thousand examples before it serves any live traffic in production systems

Answer: C

Few-shot prompting places examples in the context and relies on in-context learning, with no change to the model's weights. Request rate limits, answer length and fine-tuning are separate concepts. Fine-tuning changes weights, which few-shot prompting does not.

7. Which technique asks the model to lay out intermediate reasoning before giving its final answer?
- A Chain-of-thought prompting, which elicits step-by-step reasoning before the final answer is produced
 - B Retrieval-augmented generation, which adds documents from a knowledge base to the prompt context of each query
 - C Quantization, which reduces the numeric precision of the model's weights to lower the memory footprint
 - D Speculative decoding, which uses a small draft model to propose tokens that a larger model then verifies

Answer: A

Chain-of-thought prompting asks for worked steps, which often improves accuracy on multi-step problems, at the cost of more output tokens. Retrieval, quantization and speculative decoding are real techniques with different goals. The question names the reasoning style, not the infrastructure.

8. A fine-tuning method that trains small adapter matrices instead of all model weights is called LoRA, short for low-_____ adaptation.

Answer: rank

LoRA stands for low-rank adaptation, which describes the rank-limited updates it learns. The adapters are small enough to store and swap per task.

9. Which dataset property matters most for supervised fine-tuning quality?
- A Accurate, consistent and representative input and output pairs that cover the real variations of the task
 - B The total number of files in the dataset folder, since more files always means a better fine-tuned model
 - C The average length of the file names in the dataset folder used by the training job each time
 - D The age of the data in years, with older records always preferred because they are written more formally

Answer: A

Supervised fine-tuning learns what the examples show, so wrong, inconsistent or unrepresentative pairs teach the wrong behavior no matter how large the dataset is. File counts and name lengths have no effect on learning. Age does not correlate with quality.

10. Which technique reduces model memory use by storing weights at lower numeric precision?
- A Distillation of the training corpus into a smaller set of documents that cover the same topics as the original
 - B Sharding the prompt across several requests so that each individual request carries fewer tokens to process
 - C Quantization, such as 8-bit or 4-bit weights, which trades some accuracy for a smaller memory footprint
 - D Caching the embeddings of each user query in a separate table that grows over time on the database server

Answer: C

Quantization stores weights with fewer bits, which lowers memory and often speeds inference, with a measured accuracy cost that must be evaluated. Corpus distillation, prompt sharding and embedding caches do not change weight precision. Knowledge distillation trains a smaller model and is a different technique.