

PRACTICE QUESTIONS

IABAC

Generative AI Specialist

SHARED BY [u/zestydolphin1994](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM IABAC Generative AI Specialist

Shared by [u/zestydolphin1994](#) on aiexamcerts

10 practice questions for the IABAC Generative AI Specialist exam, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/iabac/generative-ai-specialist. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. Which parameter most directly controls the randomness of a language model's token selection?
- A Batch size, which sets how many prompts are grouped during a training step
 - B Gradient clipping threshold, which caps the size of weight updates during training
 - C Embedding dimension, which sets the length of each token's vector representation
 - D Temperature, where higher values flatten the probability distribution and vary the output

Answer: D

Temperature rescales the token probabilities at generation time, so higher values produce more varied text. Batch size and the gradient clipping threshold are training settings, not inference controls. Embedding dimension fixes the vector size and does not change sampling behavior at generation.

2. The mechanism that lets each token weigh the relevance of every other token in the sequence is called self-_____.

Answer: attention

Self-attention computes weighted relationships between tokens within the same sequence, and it is the core operation of transformer models. The accepted answer is attention. Recurrence processes tokens in order and is not the same mechanism.

3. What is the main purpose of a system prompt in a chat-based LLM application?
- A To set persistent instructions about role, constraints and style that apply across the conversation
 - B To store the user's credentials so the model can log in to external services on their behalf
 - C To replace the model's pretrained weights with a smaller set of new weights
 - D To send the conversation transcript to a human reviewer after every answer

Answer: A

A system prompt gives the model standing instructions about role, scope and format for the whole conversation. It does not store credentials, change weights or route transcripts. Those options describe unrelated mechanisms.

4. Few-shot examples included in a prompt permanently change the model's weights for all future users.
- A True
 - B False

Answer: B (False)

In-context examples affect only the current request and do not update the model's parameters. Permanent changes require training, such as fine-tuning. The statement is false.

5. Which two conditions argue for fine-tuning rather than prompting alone? (Choose two.)
- A A stable, labeled dataset of high-quality examples is available for the task
 - B The business needs the newest product price list every hour without changes to the model
 - C A consistent output style or format is needed across thousands of requests
 - D The model already performs perfectly on the task with a clear prompt
 - E The team wants the model to answer a single one-off question tomorrow

Answer: A, C

Fine-tuning pays off for consistent style and for tasks where labeled examples exist at scale. One-off questions and fast-changing facts favor prompting or retrieval. If the model already performs well, extra tuning adds cost without benefit.

6. In a RAG pipeline, what happens to the documents before they are indexed?
- A They are merged into one long string and sent to the model as a single prompt
 - B They are split into chunks, embedded as vectors and stored with metadata for retrieval
 - C They are sent to the model one at a time so that it can memorize each sentence
 - D They are deleted once the first user query has been answered by the model

Answer: B

Indexing typically chunks documents, creates embeddings and stores them with metadata for filtering. Merging into one string loses retrievability, and memorization is not how RAG works. Deleting documents after a query would make the system useless.

7. Which similarity measure is most commonly used to compare two text embeddings in a vector search?
- A Cosine similarity, which compares the angle between two vectors regardless of their length
 - B Character count difference, which treats the smallest length gap as the best semantic match
 - C Alphabetical order distance, which measures how far apart the first letters are
 - D Font weight similarity, which compares the boldness of the text in each document

Answer: A

Cosine similarity is the standard measure for comparing embedding direction, which captures semantic closeness. Character counts and alphabetical positions ignore meaning. Font weight is a display property and has no bearing on retrieval.

8. Quantizing a base model to low-precision numbers and training low-rank adapters on top of it is known as _____.

Answer: QLoRA (also accepted: Q-LoRA)

QLoRA combines a quantized base model with LoRA adapters to reduce memory use during fine-tuning. The method keeps adaptation quality close to standard LoRA while fitting on much smaller GPUs.

9. Which evaluation approach best checks whether a RAG answer is grounded in the retrieved passages?
- A Check each claim in the answer against the retrieved context, scoring whether the context supports it
 - B Count how many words the answer contains and reward the longest answers
 - C Measure how quickly the answer appeared on screen for the end user
 - D Ask the model whether it is confident and accept any answer it rates above fifty percent confidence

Answer: A

Groundedness evaluation checks each claim against the retrieved context. Word counts, display speed and self-rated confidence do not show that the claims are supported. Self-confidence is a known unreliable signal.

10. Why should an evaluation set be kept separate from the data used to tune prompts or train the model?
- A The evaluation software requires separate files, and tuning examples must be deleted after each run
 - B The evaluation set must be stored on a different cloud region for speed
 - C Evaluation data becomes invalid if it is ever read by a human reviewer
 - D Reusing the same examples for tuning and testing inflates results and hides real failures

Answer: D

Reusing tuning examples for testing makes performance look better than it is on new inputs. The separation is a methodological requirement, not a file format or storage rule. Human review of evaluation data does not invalidate it.