

PRACTICE QUESTIONS

Dell Technologies

D-GAI-F-01

Dell Gen AI Foundations

SHARED BY [u/hopeful_f01](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM Dell Technologies D-GAI-F-01, Dell Gen AI Foundations

Shared by [u/hopeful_f01](#) on aiexamcerts

10 practice questions for the Dell Technologies D-GAI-F-01 exam, Dell Gen AI Foundations, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/dell-technologies/d-gai-f-01. Spotted a mistake? Email support@aiexamcerts.com with the exam code D-GAI-F-01 and the question number.

1. A model returns different wording each time the same prompt is sent. Which inference setting most directly increases that variation?

- A Lower maximum output length
- B Higher sampling temperature
- C Smaller context window
- D Fewer fine-tuning epochs

Answer: B

Temperature scales the probability distribution used when choosing each next token, so a higher value gives less likely tokens a better chance and makes output more varied. Lowering the maximum output length only truncates the response and does not change how words are chosen. Fine-tuning epochs affect the weights during training, not the randomness of each individual inference call.

2. The process of splitting raw text into tokens before a model processes it is called _____.

Answer: tokenization (also accepted: text tokenization)

Tokenization maps text into the discrete units, such as subwords, that the model's embedding table can look up. Embedding is tempting as a related step, but it converts tokens into vectors after tokenization has already happened. Quantization is unrelated because it reduces the numeric precision of the weights.

3. In a decoder-only transformer such as a GPT-style model, how is output generated?

- A All output tokens are produced together in one parallel pass.
- B The model copies the closest sentence from its training set.
- C A separate classifier picks each word from a fixed list.
- D One token is generated at a time, each conditioned on the tokens before it.

Answer: D

Decoder-only models are autoregressive, generating one token and feeding it back in before producing the next. The parallel option is tempting because training processes all positions at once, but generation at inference is sequential. Copying training sentences and picking from a fixed word list are not how these models produce text.

4. Encoder-only models such as BERT are generally used to produce contextual representations for tasks like classification and search rather than free-form text generation.

- A True
- B False

Answer: A (True)

BERT-style encoders read the whole input bidirectionally and output representations, which suits classification, search and extraction. Generation tasks usually call for decoder models. The statement is true, and the tempting error is assuming every transformer writes text.

5. Which two factors most affect the quality of retrieved context in a RAG system? (Choose two.)

- A The font used in the user interface
- B The number of GPUs in the inference cluster
- C The chunking strategy applied to source documents
- D The color scheme of the answer panel
- E The choice of embedding model

Answer: C, E

Chunk boundaries determine whether a passage contains a complete answer, and the embedding model determines how well similarity reflects meaning. GPU count is tempting because it shapes response time, but it does not change which passages are retrieved. Interface font and colors have no effect on retrieval.

6. Which prompting approach supplies a few worked input and output pairs before the real request?

- A Zero-shot prompting
- B Few-shot prompting
- C Adversarial prompting
- D Prompt chaining

Answer: B

Few-shot prompting places a handful of example pairs in the prompt so the model can infer the pattern for the new input. Zero-shot prompting is tempting because it is the same idea without examples. Adversarial prompting is a testing or attack technique, and prompt chaining splits a task into sequential prompts.

7. What does chain-of-thought prompting ask the model to do?
- A Reveal its hidden weights so a developer can inspect them
 - B Chain several API keys together to raise the rate limit
 - C Reason through intermediate steps before giving the final answer
 - D Repeat the user prompt verbatim before answering it

Answer: C

Chain-of-thought prompting encourages step-by-step reasoning, which often improves results on multi-step problems. Revealing weights is impossible through a prompt and is tempting only as a misreading of what transparency means. Chaining API keys and repeating the prompt do nothing for reasoning.

8. A vector database returns the stored embeddings nearest to a query vector, which is called similarity _____.

Answer: search

Similarity search ranks stored vectors by distance to the query, which is the core retrieval operation in a vector database. Keyword search is tempting because it is also a retrieval method, but it matches terms rather than vector proximity. Indexing is the preparation step that makes search fast, not the query itself.

9. Which factor most often determines whether a generative AI pilot moves into production?
- A Whether the model is the newest release from its vendor
 - B The color of the user interface
 - C Clear ownership, measured outcomes, and business-approved data access
 - D The number of demo sessions given

Answer: C

Pilots stall when no one owns the outcome, success is not measured, or the needed data is not approved for use. Adopting the newest model is tempting because it feels like progress, but it rarely resolves those gaps. Interface design and demo counts do not determine production readiness.

10. A company wants an assistant that answers HR policy questions. Which knowledge source is most appropriate?
- A Informal employee chat messages from five years ago
 - B The public internet without filtering, review, or HR ownership
 - C Unlabeled photos from the office
 - D Approved, current HR policy documents with named owners

Answer: D

Authoritative, current, owned policies give the assistant reliable answers and let the team update content when rules change. Old chat messages are tempting because they contain real employee questions, but they may be outdated and are not authoritative. Unfiltered web content and photos would produce unreliable or irrelevant answers.