

PRACTICE QUESTIONS

Google Cloud

Generative AI Leader

SHARED BY [u/keen_seal_48](#) on aiexamcerts

EDITION	Free Set 1
QUESTIONS	10
SHARED	October 15, 2026
EXAM	Google Cloud Generative AI Leader

Shared by u/keen_seal_48 on aiexamcerts

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

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

1. Within a Gemini-style large language model, what does a token represent during text processing?
- A A cryptographic key that unlocks the model's API
 - B A unit of text such as a word fragment that the model reads and generates
 - C A fixed-size image patch that vision models split a frame into
 - D A row in a relational database table that stores each prompt and its reply

Answer: B

Tokens are the units of text that a model reads and produces, and they can be whole words, word fragments or punctuation. API keys are credentials and have nothing to do with the model's text units. Image patches are used by vision models, which is the tempting confusion here.

2. In a transformer model, the mechanism that lets each token weigh the other tokens in the input is called _____ attention.

Answer: self (also accepted: self-attention)

Self-attention lets every position in the sequence consider every other position when building its representation, which is central to how large language models handle context. Recurrent designs process tokens strictly in order and lack this direct link. The blank is self.

3. A company wants employees to ask questions of its own documents and receive cited answers, without assembling a search stack from scratch. Which Google Cloud offering fits best?
- A Vertex AI Search
 - B Cloud Storage lifecycle rules
 - C Cloud DNS
 - D Compute Engine preemptible virtual machines

Answer: A

Vertex AI Search is a managed service for building search and answer experiences over enterprise content, with grounded and cited results. Storage lifecycle rules control object retention and do not answer questions. Cloud DNS and preemptible machines are infrastructure services with no answer-generation role.

4. Model Garden in Vertex AI offers only Google's own Gemini models, with no third-party or open models.
- A True
 - B False

Answer: B (False)

The statement is false. Model Garden is a catalog that includes Google models along with partner and open models, so teams can compare options. Assuming it is limited to Gemini is the tempting error.

5. Which two settings most directly control the randomness of a model's output? (Choose two.)
- A Temperature
 - B Maximum output tokens
 - C Top-p (nucleus sampling)
 - D Stop sequences
 - E Candidate count

Answer: A, C

Temperature scales how spread out the token probabilities are, and top-p limits sampling to the most probable tokens that add up to a chosen mass. Maximum output tokens, stop sequences and candidate count control length, termination and the number of responses, not how the next token is chosen.

6. Which change management step most often determines whether a generative AI pilot scales?
- A Choosing the most expensive model available
 - B Involving and training end users early while measuring real adoption
 - C Limiting access to the IT team for the whole rollout
 - D Removing human review from the workflow

Answer: B

Pilots scale when the people who use them are involved early, trained well and measured on real adoption. Price alone does not produce adoption. Restricting the tool to IT or removing review each undermine trust and quality.

7. Which practice best reduces the risk of a generative AI system exposing personal data?
- A Sending all customer records into prompts for better context
 - B Publishing the training data on a public website
 - C Minimizing sensitive data in prompts and applying Sensitive Data Protection
 - D Letting every employee copy customer records into personal tools to experiment

Answer: C

Limiting what sensitive data enters prompts and datasets, then applying protection controls, reduces exposure at the source. Sending full customer records into prompts increases the risk. Publishing data or copying it into personal tools are direct leaks.

8. Adjusting a pretrained model with additional labeled examples for a specific task is called supervised _____.

Answer: tuning (also accepted: fine-tuning, fine tuning)

Supervised tuning updates the model with labeled input and output pairs, which adapts its behavior to the task. Adding context to the prompt does not do this. The blank is tuning.

9. A regulated bank needs its generative AI workloads to keep customer data in a chosen geography. Which step addresses this most directly?

- A Trusting the vendor's published privacy claims about data handling
- B Choosing the region and residency settings of each resource that holds the data
- C Using a multi-region location that replicates data outside the geography
- D Turning off encryption to speed up access

Answer: B

Data residency depends on where each resource is deployed and configured, so the region choice is the direct control. Vendor marketing is not a configuration. A multi-region location can place copies outside the required geography, and disabled encryption increases risk.

10. Which measure is a useful leading indicator during a generative AI pilot?

- A Weekly active users and how often staff accept generated outputs
- B The number of GPU models the team purchased last year
- C The uptime percentage of the pilot's hosting environment
- D The size of the pilot team's training budget

Answer: A

Active use and acceptance rates show early signs of whether the pilot will create value. Hardware purchases, hosting uptime and training budgets say little about results. A leading indicator is one that moves before final outcomes appear.