PRACTICE QUESTIONS

Google Cloud

Professional Agentic Architect

PUBLISHED BY aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
ISSUED	November 8, 2026
EXAM	Google Cloud Professional Agentic Architect

Shared by **aiexamcerts** on aiexamcerts

10 practice questions for the Google Cloud Professional Agentic Architect exam, each followed by its answer and a short explanation. More free sets and the full practice exam: **aiexamcerts.com/google-cloud/professional-agentic-architect**. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

---

1. A company wants a low-code agent in Gemini Enterprise to answer questions from its internal wiki and shared documents. Which approach follows the recommended pattern?

A Paste the full text of every document into the agent's system instructions each time

B Connect the repositories as data sources in Gemini Enterprise using Agent Search

C Export the documents to a spreadsheet and attach it to every single user prompt

D Retrain a base model on the whole wiki each time any one of its pages changes

Answer: B

Connecting data stores lets the agent retrieve current content securely when a question arrives. Pasting documents into system instructions does not scale, goes stale, and wastes context on every turn.

---

2. An agent API receives bursty, unpredictable traffic and is packaged as a container with custom endpoints. The team wants to pay little during idle periods. Which runtime fits best?

A A fixed-size GKE cluster that is sized for peak load all day

B A single always-on virtual machine that has been sized for the busiest hour of the week

C Cloud Run, which scales container instances with demand, including down to zero

D A scheduled batch job that runs once a day

Answer: C

Cloud Run suits stateless containers with variable load and low idle cost. A cluster or VM sized for peak keeps billing during quiet hours.

---

3. A vector search returns the top 100 candidates quickly, but the top five passed to the LLM are often not the most relevant. Which addition improves precision at the top?

A A second embedding pass that doubles the vector dimensions for every stored document

B A larger chunk size so that each candidate contains many more sentences of text

C A higher top-k that sends all one hundred candidates on to the LLM for answering

D A reranking stage that rescores candidates by reading the query and each passage together

Answer: D

Fast retrieval optimizes recall, and a reranker then orders candidates more accurately for the final few. Sending all 100 passages increases cost and can bury the best evidence.

---

4. A document retrieved by a tool contains hidden instructions to email confidential files to an outside address. Which combination is the strongest defense?

A Treat tool output as untrusted, screen it with Model Armor, limit permissions, and require approval to send

B Rely on the system instruction alone to tell the agent to ignore any malicious text it reads

C Allow the agent to send email freely, but write a log entry for every message sent

D Remove all guardrails so that the agent can learn to recognize attacks by itself

Answer: A

Layered controls handle indirect prompt injection: screening, least privilege and approval for risky actions. A prompt instruction alone is not a reliable barrier.

---

5. Cloud Trace can show how much of an agent request's total latency was spent in each model call and tool call.

A True

B False

Answer: A (True)

Traces represent a request as spans, which makes slow steps visible. Pair them with Cloud Logging for event detail.

---

6. A prompt that asks the model to reason through intermediate steps before giving its answer uses _____ prompting.

Answer: chain-of-thought (also accepted: chain of thought, CoT)

Chain-of-thought prompting elicits step-by-step reasoning and helps on multi-step problems. It is distinct from few-shot prompting, which supplies worked examples.

---

7. A trip-planning agent must fetch flight prices, hotel prices and weather at the same time, because the three lookups do not depend on each other. Which ADK composition is the best fit?

A A SequentialAgent that runs them one after another in a fixed order

B A ParallelAgent that runs the three sub-agents concurrently and gathers their results

C A LoopAgent that repeats them until the user stops the conversation

D A single agent with no tools at all that answers purely from its own training data

Answer: B

Independent work should run concurrently to cut latency. A sequential composition is correct only when each step needs the output of the previous one.

---

8. Monitoring shows that most agent cost comes from sending the same large reference text in every request. Which optimization is most appropriate?

- **A** Add more tools so the model needs fewer tokens to reason, which lowers each request's cost
- **B** Raise the max output tokens so that answers are always complete, whatever the length
- **C** Use context caching or retrieve relevant passages only, and route simple requests to a smaller model
- **D** Duplicate the reference text inside the request to improve the recall of its contents

Answer: C

Cost follows tokens, so avoiding repeated large inputs and using a cheaper model where quality allows directly reduces spend. Longer outputs and duplicate text increase it.

---

9. A RAG agent returns answers that stop mid-explanation because retrieved passages cut the source documents in the middle of a section. What is the best fix?

- **A** Increase the sampling temperature so that the generated answers sound more complete overall
- **B** Reduce the number of embedding dimensions to speed up retrieval
- **C** Disable retrieval and rely on the model's general knowledge
- **D** Rework chunking to respect section boundaries and add overlap between neighboring chunks

Answer: D

Chunking determines whether retrieved text is self-contained, so boundary-aware chunks with overlap preserve context. Temperature changes wording, not what evidence was retrieved.

---

10. Tool results from a CRM contain customer phone numbers and addresses that should not be sent to the model or written to logs. Which service helps?

- **A** Sensitive Data Protection, to de-identify personal data before the model or logs see it
- **B** Cloud Run revisions, which keep the older versions of the service available to roll back
- **C** Agent Registry, which lists the agents that the organization has approved for use
- **D** Vector Search, which stores embeddings of documents for similarity lookups

Answer: A

Sensitive Data Protection finds and masks sensitive elements in text. A registry or vector index does not inspect payloads.