

PRACTICE QUESTIONS

Anthropic

CCAR-P

Claude Certified Architect - Professional

SHARED BY [u/smoothsummit1977](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 18, 2026

EXAM Anthropic CCAR-P, Claude Certified Architect - Professional

Shared by [u/smoothsummit1977](#) on aiexamcerts

10 practice questions for the Anthropic CCAR-P exam, Claude Certified Architect - Professional, each followed by its answer and a short explanation. More free sets and the full practice exam:

aiexamcerts.com/anthropic/ccar-p. Spotted a mistake? Email support@aiexamcerts.com with the exam code CCAR-P and the question number.

1. A team's RAG answers are slow and imprecise on short factual lookups but good on broad questions over long reports. Which retrieval adjustment best matches the data shape and query pattern?
- A Use one large chunk per document for every query, so the full context is always available
 - B Drop the retriever and send the full document collection in each prompt, sorted by document date before it is sent to the model
 - C Increase the number of tokens the model generates per answer
 - D Index smaller chunks for precise lookups and keep larger section-level chunks for broad questions

Answer: D

Matching chunk size to query shape improves precision for narrow lookups without losing context for broad questions. One large chunk per document is tempting because it preserves context, but it dilutes relevance for short factual queries.

2. A RAG pipeline retrieves the right passage in most tests but misses it when users search with product codes or exact error strings. Which change most directly addresses this?
- A Raise the temperature so the model guesses the missing product code, and reads the surrounding words to recover the code
 - B Remove metadata filters so the retriever considers all documents equally
 - C Shorten every chunk to a single sentence so the embedding is more specific
 - D Add keyword (lexical) search alongside vector search and merge the ranked results

Answer: D

Exact identifiers are often poorly captured by dense embeddings, so hybrid retrieval that also scores keywords recovers them. Shortening chunks to single sentences is tempting, but it discards context and does not fix exact-match failures.

3. Does prompt caching work by storing the model's previous complete answers, so a repeated user question is returned without calling the model?
- A True
 - B False

Answer: B (False)

This is false. Prompt caching reuses the processed prefix of a request, such as a long system prompt or tool definitions, so repeated prefixes cost less and return faster. Answer caching in the application is the tempting but different mechanism.

4. Which authentication design best limits the damage if an MCP server's credentials leak?
- A One shared service account with full access, rotated once a year
 - B Per-user OAuth tokens with narrow scopes, stored server-side and short-lived
 - C A single long-lived API key embedded in the client configuration file
 - D Basic authentication with the same password shared by all users of the server

Answer: B

Short-lived, narrowly scoped, per-user tokens limit what a leak exposes and keep access attributable to a person. A shared service account is tempting for simplicity, but it grants every caller the same broad rights and hides who did what.

5. Monitoring at scale should record each tool call's name, latency and _____, meaning whether it succeeded or failed.

Answer: outcome (also accepted: result, status)

A success or failure status per tool call lets operators spot a failing dependency and connect it to latency and cost. Logging only the final model answer is tempting because it is smaller, but it hides which tool caused a failure.

6. A developer asks whether to give an agent 60 tools in one configuration or to expose tools progressively. Which reasoning best supports progressive discovery?
- A Every tool definition is free to include, so a larger catalog never hurts, so a larger catalog never hurts the accuracy of the answer
 - B The model must read every tool definition before it can answer any question
 - C Loading only relevant tool definitions reduces context use and confusion between similar tools
 - D Progressive discovery removes the need for authorization checks on each tool

Answer: C

Tool definitions consume context and similar tools can be confused, so loading the relevant subset improves accuracy and cost. The claim that tool definitions are free is the tempting misconception, since they count against the input budget.

7. Which trade-off best describes a decision to use a smaller model with a larger retrieval top-k?
- A Lower per-call latency and cost, with more candidate context that the model must filter
 - B Higher accuracy guaranteed, because retrieval always improves the model's reasoning
 - C No effect on latency, since retrieval and generation run in the same step
 - D Higher latency and cost, because more chunks always require a larger model, with a larger context window for every request

Answer: A

A smaller model is faster and cheaper per call, but more retrieved chunks add tokens that the model must sift, which can cost accuracy. The belief that retrieval always improves reasoning is tempting, yet irrelevant context can distract the model.

8. Which two statements about the Model Context Protocol are accurate? (Choose two.)
- A MCP messages are based on JSON-RPC 2.0
 - B MCP servers must be written in Python, which is the only supported language
 - C MCP resources are read-only data that clients can retrieve by URI
 - D MCP replaces the need for any authentication when deployed over HTTP
 - E MCP tools run on the client's model weights, so they never call external systems

Answer: A, C

MCP uses JSON-RPC 2.0 messages, and resources are data addressed by URI that clients read. Option B is false because SDKs exist for several languages, D is false because remote transports need their own authorization, and E is wrong because tools commonly call external systems. The tempting distractor is D.

9. An agent's answers change unpredictably after a tool's response format changes from JSON to free text. What is the most sound fix?
- A Adjust the sampling temperature so the model adapts to the new format faster
 - B Remove the tool and paste its last known output into every prompt
 - C Ask the model in the system prompt to ignore formatting differences, and infer any missing fields on its own
 - D Version the tool's output schema, validate responses, and fail clearly on unexpected shapes

Answer: D

Stable, validated tool contracts with explicit versioning make format drift visible instead of silently changing agent behavior. Adjusting sampling settings is tempting, but changing temperature does not repair a broken data contract.

10. Can retrieval-augmented generation ground answers in documents added after the model's training data was collected, provided they are indexed and retrieved at query time?

- A True
- B False

Answer: A (True)

This is true. RAG supplies current or private content at query time, so the model can answer from documents it never saw in training. Fine-tuning on the documents is the tempting alternative, but it is not required for this and is slower to keep current.