CLOUDYETI / FREE PDF / CCAR-P

Claude Certified Architect - Professional (CCAR-P) Questions and Answers

25 practice questions with answers and explanations, covering every exam domain. Taken from our bank of 150 reviewed questions.

Looking for exam dumps? Use these original practice questions instead. These are original practice questions written by CloudYeti. They are NOT real exam questions and are not affiliated with or endorsed by Anthropic.

Take the full timed practice test free:
https://learn.cloudyeti.io/free/claude-certified-architect-professional-practice-test

DOMAIN 1 / SOLUTION DESIGN AND ARCHITECTURE

1. An executive asks for an autonomous agent to 'eliminate support backlog.' What should the architect establish before selecting a model or framework?

A) A multi-agent diagram

B) A list of every Claude feature

C) The maximum available context window

D) A measurable target, task boundaries, current process baseline, failure costs, decision rights, and escalation requirements

Answer: D. Architecture should follow a measurable operating problem and its constraints. Selecting technology before defining success and risk produces an untestable solution.

2. A firm needs a low-volume internal policy assistant in six weeks. Its content changes weekly and must be cited. Which direction is the strongest starting point?

A) Build a fully autonomous multi-agent organization

B) Train a foundation model from scratch

C) Use a managed Claude endpoint with retrieval over governed documents, citations, and a human escalation path

D) Fine-tune policy facts into model weights every week

Answer: C. Managed inference plus governed retrieval fits a short timeline, changing knowledge, and citation needs. Training or repeated fine-tuning adds cost and makes freshness harder.

3. A customer-service assistant depends on a CRM, order system, and knowledge index. How should the solution behave when the CRM is unavailable?

A) Invent customer context to keep latency low

B) Enter a defined degraded mode, disclose the limitation, prevent CRM-dependent actions, and offer retry or human handoff

C) Retry forever

D) Disable all services permanently

Answer: B. Graceful degradation preserves safe capabilities while blocking actions that require unavailable truth. Clear user communication and handoff protect trust.

4. An environmental monitoring agency is deploying a Claude-powered anomaly-detection assistant that processes real-time sensor telemetry for 200 regional stations. The agency requires a gradual canary release so that a small percentage of live traffic is routed to the new model version while the legacy version remains the authoritative path. During the canary window, architects must be able to roll back within minutes if error rates spike. Which architecture pattern best satisfies these constraints?

A) Run the new model version in a shadow mode that receives a copy of all requests but whose responses are discarded; promote to production only after an offline evaluation batch confirms accuracy thresholds.

B) Use a feature-flag service to enable the new model version for a pre-selected list of station IDs; all other stations continue to use the legacy version until the flag is globally toggled.

C) Deploy both model versions behind a weighted-routing layer that splits traffic by configurable percentage; monitor per-version error rates and adjust weights or redirect all traffic to the stable version without redeploying application code.

D) Deploy the new model version to a separate environment and replay the last 24 hours of recorded sensor events against it; if replay error rates are acceptable, cut over all live traffic at once.

Answer: C. A weighted-routing canary splits live traffic by configurable percentage, enabling real-time error-rate comparison and instant rollback by adjusting weights - satisfying both gradual exposure and sub-minute recovery. Shadow mode discards responses so it cannot measure live user impact. A feature-flag approach tied to station IDs is a fixed cohort, not a percentage-based canary, and toggling is not sub-minute rollback. Replay-then-cutover is a batch validation strategy that still performs a full cutover, eliminating the gradual exposure requirement entirely.

5. A global platform wants to route requests among Claude models by task difficulty. What must accompany the router before production release?

A) Routing based only on prompt length

B) A promise that cheaper models are good enough

C) Per-route eval thresholds, fallback behavior, observability, and change control for model or policy updates

D) A random fallback model

Answer: C. Routing introduces a policy layer that can change quality and risk. It needs validated thresholds, monitored decisions, controlled fallbacks, and versioned changes.

6. A regulated assistant serves multiple business units whose policy documents may conflict. What is the best context strategy?

A) Put all units' documents into every request

B) Filter retrieval by authenticated tenant and policy scope, rank current authoritative sources, and preserve citations and access checks

C) Let users name any document path

D) Use the longest model context and skip retrieval

Answer: B. Context assembly must enforce tenancy and authority before generation. Retrieval quality and access control are separate controls and both are required.

7. Several teams edit a production system prompt directly in the console, making incidents hard to reproduce. What operating model should replace this?

A) Version prompts and schemas in source control, test them against evals, approve changes, and attach deployed versions to traces

B) Ask teams to remember their edits

C) Freeze the prompt forever

D) Allow more people to edit it

Answer: A. Prompts are production artifacts. Versioning, evaluation, approval, and trace linkage make behavior reproducible and changes auditable.

DOMAIN 3 / INTEGRATION

8. A RAG assistant gives fluent wrong answers. Retrieval logs show that the correct passage was never returned. Which component should the team improve first?

A) The response tone

B) The ingestion and retrieval layer, using relevance judgments and retrieval metrics before retuning generation

C) The user interface animation

D) The maximum output token count

Answer: B. Generation cannot ground itself in evidence it never receives. Isolating retrieval quality from answer quality prevents prompt changes from masking the root cause.

9. An enterprise agent uses MCP tools to read and modify records. Where should authorization decisions be enforced?

A) At the identity-aware application and tool boundary for every request, with least privilege and audit logs

B) In the user's natural-language request

C) By hiding tool names

D) Only in the model prompt

Answer: A. The model is not an authorization authority. Each tool call needs enforcement based on authenticated identity, resource, action, and policy.

10. A platform team must connect Claude to an internal service used only by one backend and to a reusable enterprise tool catalog used across clients. What is the best split?

A) Expose both databases directly

B) Put integration logic in prompts

C) Use MCP for everything by policy

D) Use a direct API or SDK where the integration is private and tightly coupled, and MCP where a standardized reusable client boundary adds value

Answer: D. Protocol choice should follow reuse and boundary needs. MCP is valuable for standardized discovery across clients but is not mandatory for every private service call.

11. A team wants full traces for debugging, but prompts can contain customer secrets. What observability design is best?

A) Send traces to personal accounts

B) Log everything indefinitely

C) Capture structured operational metadata and sampled redacted content under access, retention, and incident policies

D) Disable all telemetry

Answer: C. Useful observability can coexist with privacy through minimization, redaction, access controls, sampling, and retention. Either extreme loses necessary protection or diagnostic evidence.

12. A consumer electronics retailer is preparing to release a Claude-powered product recommendation agent. The compliance team requires that every agent decision affecting a customer's cart be accompanied by a human-readable rationale that can be reviewed post-hoc. During staging, the team notices that the agent sometimes returns recommendations without surfacing its reasoning steps. Which release strategy best satisfies the explainable decision evidence constraint before the agent goes to production?

A) Deploy to production immediately and rely on customer feedback tickets to identify cases where reasoning is absent, then patch the prompt retroactively.

B) Enable sampled tracing at 10% of traffic so that most decisions are logged with rationale, accepting that the remaining 90% will be unaudited.

C) Gate production release on a structured evaluation harness that verifies reasoning traces are present and coherent for a representative set of test cases, and block rollout if the pass rate falls below a defined threshold.

D) Add a post-processing filter that appends a generic disclaimer to every recommendation, satisfying the audit requirement without changing the agent's core behavior.

Answer: C. Gating on a structured evaluation harness that checks for reasoning traces directly addresses the explainable decision evidence constraint before any customer exposure. Anthropic's agent evaluation guidance emphasizes building eval sets that test specific behavioral properties - such as the presence of coherent reasoning - and using pass/fail thresholds to gate releases. Relying on customer feedback is reactive and exposes customers to unexplained decisions. Sampled tracing at 10% leaves 90% of decisions unaudited, violating the compliance requirement. Appending a generic disclaimer fabricates evidence rather than capturing genuine rationale.

13. An assistant's overall success rate rises from 84% to 89%, but policy-violation failures double from 0.2% to 0.4%. Should the release proceed?

A) Not until the risk-specific release threshold is met and the violation regression is understood

B) Yes, if the new model is cheaper

C) Decide from one stakeholder demo

D) Yes, because average success improved

Answer: A. Averages must not hide critical regressions. Risk-weighted gates should block a release that worsens an explicitly protected safety outcome.

14. A research agent is evaluated only by an LLM judge that scores writing quality. What is missing?

A) A longer answer limit

B) A second identical writing judge

C) More decorative formatting

D) A mixed evaluation that also checks source validity, claim support, coverage, tool behavior, and expert review for disputed cases

Answer: D. One style-oriented judge cannot measure factual grounding or process correctness. Combining deterministic checks, task-specific judges, and expert review gives broader evidence.

15. A model change saves 25% per request but increases human review time by 40%. What should determine whether it is an optimization?

A) The vendor's marketing claim

B) Inference price alone

C) Total cost per successful outcome, including review, retries, latency, incidents, and quality

D) The model's benchmark ranking

Answer: C. System economics are measured at the successful business outcome, not the API line item. Downstream labor and failure costs can erase nominal savings.

16. A public transit authority runs a Claude-powered passenger information agent on Amazon Bedrock. The system serves three separate transit agencies - each with distinct route data, fare rules, and compliance obligations - from a single deployment. An architect proposes enabling prompt caching to reduce per-request cost by reusing large static context blocks. A security review flags that cached prefixes must never leak across agency tenants. Which implementation approach satisfies the cost-reduction goal while enforcing the required security boundary?

A) Cache only the generic system instructions shared across all agencies and exclude all tenant-specific context from caching to avoid any isolation risk.

B) Scope each cache prefix to a per-tenant static block (agency route data + fare rules) injected before any shared instructions, ensuring each tenant's prefix is populated and retrieved only within that tenant's request path.

C) Enable prompt caching globally on the shared system prompt and rely on Bedrock's request-level IAM authorization to prevent cross-tenant data access.

D) Disable prompt caching for all tenants and instead reduce cost by switching to a smaller Claude model for all agencies, accepting a quality trade-off.

Answer: B. Prompt caching works by matching an exact prefix; scoping the cache prefix to each tenant's unique static block guarantees that a cache hit can only occur on that tenant's own prefix, enforcing isolation by construction. Relying solely on IAM authorization does not prevent a shared prefix from being reused across tenants if the prefix bytes are identical. Disabling caching entirely forfeits the cost benefit and the model-downgrade trade-off is a separate quality decision, not a security boundary solution. Caching only generic shared instructions leaves the large tenant-specific context uncached, eliminating most of the cost savings.

DOMAIN 5 / GOVERNANCE, SAFETY, AND RISK MANAGEMENT

17. A business wants to retain every agent conversation forever 'in case it becomes useful.' What should the architect recommend?

A) Store copies in every region

B) Let each engineer decide

C) Approve because storage is cheap

D) Define purpose-based retention by data class, minimize stored content, support deletion, and document access and legal requirements

Answer: D. Indefinite collection increases privacy, security, and compliance risk. Retention should be justified, classified, controlled, and enforceable.

18. A procurement agent reads vendor websites and can create purchase orders. What is the most important architectural risk?

A) The model may write concise summaries

B) Web pages may use unusual fonts

C) Untrusted content can inject instructions that influence a consequential tool call

D) Purchase orders have identifiers

Answer: C. The system combines untrusted external content with write authority. Separation of data and instructions, constrained tools, validation, approvals, and monitoring are required.

19. A company wants one approval policy for all agent actions. Which alternative is more effective?

A) Require no approvals

B) Classify actions by reversibility, financial or legal impact, data sensitivity, and confidence, then assign proportional controls

C) Require the CEO to approve every read

D) Ask the model to choose whether it needs approval

Answer: B. Risk-tiered controls preserve usability for low-risk work and add stronger guarantees for consequential actions. The model should not be the sole arbiter of its own oversight.

20. An education platform deploys a Claude-powered tutoring agent that ingests student-submitted essays, external reading materials, and third-party quiz APIs. After a reported grading anomaly, the incident team cannot determine whether the root cause was a malicious prompt in a student essay, a corrupted quiz API response, or a model hallucination. Which architectural change most directly provides the explainable decision evidence needed to diagnose future failures?

A) Switch to a multi-agent pipeline where one Claude instance grades and a second Claude instance reviews the grade independently.

B) Emit a structured trace for every agent turn that records the exact tool inputs, raw tool outputs, and the model's reasoning step before each action.

C) Add a content-moderation classifier that blocks student submissions containing injection patterns before they reach the agent.

D) Enable prompt caching on the system prompt so that the baseline context is frozen and cannot be altered by student content.

Answer: B. Structured per-turn traces that capture tool inputs, raw tool outputs, and intermediate reasoning give investigators a complete causal chain to replay and pinpoint the failure source. A content-moderation classifier prevents some attacks but produces no post-hoc evidence trail for anomalies that slip through. A dual-model review pipeline adds redundancy but still lacks the granular trace needed to distinguish injection from hallucination. Prompt caching freezes the system prompt but does not record what happened during execution. Anthropic's agent-building guidance explicitly recommends logging inputs and outputs at every step for auditability.

21. A sponsor demands 99.99% answer accuracy but will not fund source curation or human review. What should the architect do?

A) Replace accuracy with response length

B) Promise the target to secure approval

C) Show measured baseline performance, identify the controls and cost required, and negotiate a scoped service level tied to risk

D) Hide known errors

Answer: C. Responsible architecture makes constraints and evidence visible. Service levels require investment and scope; they should not be promised without a credible operating design.

22. A new case-management agent affects 2,000 employees and changes approval responsibilities. What rollout is strongest?

A) Enable it for everyone with no warning

B) Pilot with representative users, train affected roles, publish decision rights and fallbacks, monitor outcomes, and expand by gated cohorts

C) Launch only a marketing page

D) Measure adoption but not errors

Answer: B. A staged rollout tests both technical behavior and operating-model change. Training, explicit ownership, monitoring, and fallbacks reduce organizational risk.

23. An agent makes an unauthorized external change. What should the incident plan prioritize?

A) Contain tool access, preserve evidence, assess and reverse impact where possible, notify accountable owners, and convert findings into controls and evals

B) Change the UI color

C) Ask the agent not to repeat it

D) Delete all logs immediately

Answer: A. Agent incidents require conventional containment and evidence handling plus system-specific remediation. New controls and regression cases make the lesson durable.

24. Multiple teams are building Claude applications with inconsistent security and evaluation practices. What should a platform team provide first?

A) An unrestricted API key shared in chat

B) Approved templates with identity, logging, eval, tool-policy, and deployment defaults plus a paved path and exception process

C) A mandate to use identical prompts

D) No shared guidance so teams can move faster

Answer: B. A paved path turns governance into reusable engineering defaults while preserving an explicit route for justified exceptions. Shared secrets and prompt mandates do not create safe operations.

25. An agricultural cooperative runs a multi-tenant Claude API platform where each member organization (grain, dairy, livestock) must never see another tenant's proprietary yield data or pricing models. The platform team is designing a shared prompt-template library. Developers on each tenant team will instantiate templates from this library when building their own Claude-powered features. Which governance control most directly enforces tenant isolation at the template layer, ensuring that every instantiated prompt carries the correct data-scope declaration before reaching the model?

A) Require each tenant to maintain a completely separate codebase and prompt library with no shared templates, accepting duplicated maintenance overhead as the cost of isolation.

B) Allow tenants to share a single generic system prompt and rely on Claude's built-in refusal behavior to block requests that reference another tenant's data.

C) Design each template in the shared library to require a mandatory, validated tenant-scope parameter that is injected into the system prompt at instantiation time, with platform-level validation rejecting any template render that omits or mismatches the tenant identifier.

D) Store all tenant context in a shared vector database and use retrieval similarity thresholds to filter results before they reach the model, treating threshold tuning as the isolation guarantee.

Answer: C. Designing shared templates to require a mandatory, validated tenant-scope parameter that is injected at instantiation time is a template-layer governance control: isolation is enforced structurally by the template itself, not by downstream infrastructure or probabilistic model behavior. Every rendered prompt carries the correct data-scope declaration before the model processes any user input, and the platform rejects renders that omit or mismatch the identifier - making misconfiguration a build-time or request-time error rather than a silent runtime risk. This directly supports the goal of standardizing safe developer workflows because tenant teams consume a shared, pre-hardened library rather than each building their own isolation logic. Relying on Claude's refusal behavior is not a platform control; model behavior is probabilistic and cannot substitute for deterministic structural enforcement in a multi-tenant system. Using a shared vector database with similarity-threshold filtering is a retrieval heuristic, not an isolation guarantee - threshold tuning errors can expose cross-tenant data, and the control lives outside the template layer. Requiring fully separate codebases per tenant eliminates the shared-library benefit, duplicates maintenance effort, and does not constitute a standardized platform control.

Want the rest?

The full CCAR-P practice test on learn.cloudyeti.io has 150 questions, a timed exam mode, and a score by domain so you know what to study next.

Start the free CCAR-P practice test

These are original practice questions written by CloudYeti. They are NOT real exam questions and are not affiliated with or endorsed by Anthropic.