

Claude Certified Architect Professional

Original English Mock Exam

63 questions · 120 minutes

Choose ONE answer unless the question asks for TWO.

All scenarios and options are independently authored.

The domain distribution follows the July 2026 Professional guide.

Keep the separate answer guide closed until you finish.

Instructions & answer sheet

ENGLISH ONLY / 63 POINTS

Allow 120 minutes. Answer every item. For multiple-response items, select exactly TWO answers.

For this practice set, each exact answer is worth one point. There is no partial credit. This is a local scoring rule, not a claim about undisclosed official scoring details.

Use only the facts and assumptions in each scenario. Any prices given are hypothetical. Do not substitute current product prices.

Record your answer, confidence (L/M/H), and a review mark. Questions are mixed across domains; the answer guide contains the domain mapping.

Q	Answer / L M H	Q	Answer / L M H	Q	Answer / L M H
1		22		43	
2		23		44	
3		24		45	
4		25		46	
5		26		47	
6		27		48	
7		28		49	
8		29		50	
9		30		51	
10		31		52	
11		32		53	
12		33		54	
13		34		55	
14		35		56	
15		36		57	
16		37		58	
17		38		59	
18		39		60	
19		40		61	
20		41		62	
21		42		63	

Questions 1-3

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q01 | Select ONE

Answer quality falls immediately after a retrieval index rebuild. The prompt, model version, and latency remain stable. What should the team inspect first?

- A. User training, without first checking the changed retrieval component.
 - B. Model weights, assuming they changed despite a pinned and verified model version.
 - C. Output styling, because a different heading structure explains stale evidence.
 - D. Returned chunks, document version coverage, and indexing or embedding compatibility against a known evaluation set.
-

Q02 | Select ONE

A service receives a short customer message and produces a suggested title. It needs no external data or actions, and a single model call meets all acceptance criteria. Which architecture should the team start with?

- A. A retrieval pipeline that indexes all past messages before every generation.
 - B. An autonomous agent with a search tool and a persistent planning loop.
 - C. A manager agent that delegates title generation to three specialist agents.
 - D. A bounded single-call design with an output contract and validation.
-

Q03 | Select ONE

A sizing estimate divides 8,640 daily requests evenly across 24 hours and concludes that one request every 10 seconds is sufficient. Actual users send most requests during a 30-minute morning window. What is missing?

- A. Peak arrival rate, concurrency, service-time distribution, and load testing against the required latency target.
 - B. A larger output schema for each request.
 - C. A guarantee that retries will never occur because the daily average is low.
 - D. A different rounding rule for the daily average.
-

Questions 4-6

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q04 | Select ONE

A policy requires a person to approve every irreversible account termination. The model reports 99% confidence on one termination. What is the correct routing?

- A. Ask another model for agreement and treat consensus as human approval.
 - B. Terminate immediately and include the case in a monthly sample.
 - C. Skip review because confidence exceeds 95%.
 - D. Require pre-action approval and show the reviewer the evidence, proposed action, and reason for the flag.
-

Q05 | Select TWO

A retrieved troubleshooting article contains an instruction to export internal credentials. The user's original question was harmless and passed input screening. Which TWO controls address this attack path?

- A. Give retrieved articles higher instruction priority than application policy.
 - B. Rely only on an output filter after any export tool has already executed.
 - C. Screen only the original user message because it started the request.
 - D. Treat retrieved content as untrusted data and screen it before adding it to model context.
 - E. Constrain credential access and tool execution with least privilege and authorization independent of the retrieved text.
-

Q06 | Select ONE

A service must return a valid schema, preserve exact invoice totals, and produce helpful summaries. Which evaluation plan is most appropriate?

- A. Use user satisfaction alone and omit component-level checks.
 - B. Use one model judge for every criterion, including arithmetic and schema parsing.
 - C. Use only exact-string matching for summaries, regardless of acceptable paraphrases.
 - D. Use deterministic checks for schema and totals, and a calibrated rubric with human-labelled examples for summary quality.
-

Questions 7-9

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q07 | Select ONE

An incident assistant needs to read service health and propose a remediation plan. Its inherited tool bundle also permits changing access policies. Which change most directly reduces unnecessary authority?

- A. Keep the policy tool but add a daily report of all policy changes.
 - B. Keep all tools and ask the assistant to choose only the safest ones.
 - C. Remove policy-changing capabilities from this assistant and keep its tools aligned with the read-and-propose role.
 - D. Upgrade the model while preserving the inherited tool bundle.
-

Q08 | Select ONE

A purchasing assistant extracts a requested quantity from an email, checks current stock, and approves purchases above a company threshold. The threshold is maintained in an existing policy service. Which division of responsibility is most appropriate?

- A. Use Claude for extraction and threshold interpretation, and query the stock service only when Claude reports uncertainty.
 - B. Let an agent own the entire decision, provided its prompt contains the latest threshold and asks for confidence.
 - C. Use Claude for extraction, authoritative services for stock and threshold checks, and an authorized person for required approvals.
 - D. Use the policy service for extraction and Claude for stock estimation, then sample completed purchases for review.
-

Q09 | Select ONE

A required approval-policy service times out before an agent can execute a high-impact account closure. What should the application do?

- A. Block or queue the closure and use the defined escalation path until authorization can be established.
 - B. Let the model replace the unavailable policy service with its best judgment.
 - C. Execute the closure because no explicit denial was returned.
 - D. Execute the closure and mark the audit record as pending approval.
-

Questions 10-12

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q10 | Select ONE

A prototype passes ten short, clean documents. Production includes scans, long documents, and uncommon layouts. What should happen before a production claim is made?

- A. Test only the shortest documents because they are easiest to label.
 - B. Build a representative held-out evaluation set with important edge cases and report results by input category.
 - C. Repeat the same ten documents until the average score stops changing.
 - D. Select production documents only after seeing which ones the model handles well.
-

Q11 | Select ONE

Consider this application flow: authenticate user; ask Claude for an action and account_id; execute_action(action, account_id); check whether the final text is safe. What is the most important missing control?

- A. A post-action model judge that checks whether the action sounded reasonable.
 - B. Server-side authorization of the requested action and target account immediately before execution.
 - C. A larger output token budget to make the action explanation more detailed.
 - D. A statement in the final response that the action was authorized.
-

Q12 | Select ONE

Two architectures both meet the mandatory requirements. One is quicker to build but more expensive at scale. What should an approval presentation include?

- A. Only the preferred architecture's strongest benchmark result.
 - B. Expected gains, operating costs at forecast volume, remaining risks, and the cost of reversing the choice later.
 - C. A detailed component diagram with no recommendation or business consequences.
 - D. Only cost per call, without volume or migration implications.
-

Questions 13-15

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q13 | Select ONE

A classifier produces inconsistent labels even though the allowed categories are clearly defined. Errors cluster around two neighboring categories. Which intervention should be tested first?

- A. Add representative contrastive examples for the ambiguous boundary and evaluate on held-out cases.
 - B. Increase the context window without supplying any new task information.
 - C. Remove the allowed category list so the model can invent clearer labels.
 - D. Add many examples from the easiest category and measure only those examples.
-

Q14 | Select ONE

A repository assistant investigates unfamiliar build failures. The useful files and number of diagnostic steps cannot be known in advance. File reads are allowed, but production deployment is prohibited. Which design best fits?

- A. A bounded agent with scoped repository tools, a step budget, and a review gate for proposed changes.
 - B. A fixed three-prompt chain that always reads the same files and stops after the same checks.
 - C. A single call containing filenames only, with no mechanism to inspect their contents.
 - D. A team of deployment agents allowed to modify production until the build succeeds.
-

Q15 | Select TWO

A review pipeline analyzes 36 independent specification files and then produces a consolidated report. Which TWO controls best prevent a fluent report from hiding incomplete work?

- A. Reconcile the expected file IDs with the file IDs in validated worker results before synthesis.
 - B. Ask the synthesizer to sound confident even if some workers time out.
 - C. Define structured worker outputs and explicit retry or escalation behavior for missing or malformed results.
 - D. Treat the total number of worker responses as sufficient, even if some IDs are duplicates.
 - E. Increase the synthesizer model size without changing the result contract.
-

Questions 16-18

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q16 | Select ONE

A reusable procedure must reach only two departments and must support controlled updates and rollback. Which distribution approach best matches the stated need?

- A. A governed plugin assigned to the relevant groups, with an owner and version-controlled releases.
 - B. An organization-wide upload with no scoped access or version recovery plan.
 - C. A copy kept in one developer's unshared local directory.
 - D. A prompt pasted into every employee's personal chat.
-

Q17 | Select ONE

A dashboard shows steadily declining usefulness over four weeks, but error rates remain below alert thresholds. No one owns quality review. What should be added?

- A. A rule that reviews quality only when the API returns errors.
 - B. A longer interval between stakeholder reviews to reduce interruptions.
 - C. A feedback rule mapping quality trends to a review trigger, accountable owner, required action, and follow-up.
 - D. Another dashboard displaying the same metrics with different colors.
-

Q18 | Select ONE

Overall accuracy is stable, but one user subgroup receives worse recommendations. The corpus and few-shot examples overrepresent another subgroup. What is the best next step?

- A. Evaluate outcomes by subgroup and inspect corpus, examples, prompt framing, and downstream routing with governed decision evidence.
 - B. Remove subgroup measurements to avoid exposing differences.
 - C. Declare the model fair because its aggregate accuracy has not changed.
 - D. Assign the issue entirely to the model provider without inspecting application inputs.
-

Questions 19-21

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q19 | Select ONE

Claude refuses broadly harmful requests in testing. A project lead therefore assumes it will also enforce a private rule that users may not see records from another division. What is the correct architectural response?

- A. Place the division name only in a friendly greeting and rely on contextual inference.
 - B. Accept the assumption because all unsafe requests belong to the same category.
 - C. Implement the division access rule in application authorization; broad trained behavior does not encode the private policy.
 - D. Test more harmful prompts while leaving cross-division access uncontrolled.
-

Q20 | Select ONE

A vendor contract states that customer content is not used to train models by default. A stakeholder concludes that no data is retained anywhere. Which response is accurate?

- A. Training use and retention are separate; verify the applicable route, features, logs, and contractual retention terms.
 - B. Agree, because no training necessarily means zero retention.
 - C. Use the model's verbal assurance as the evidence of the retention policy.
 - D. Assume all data is permanently retained because training is excluded.
-

Q21 | Select ONE

The primary inference route meets a contractual region restriction. During overload, a fallback sends the same regulated payload to an unverified global route. What is the best correction?

- A. Allow only fallbacks verified against the same hard constraints, and define a safe degraded response when none is available.
 - B. Remove route information from logs so the application appears consistent.
 - C. Tell the fallback model to imagine that processing occurs in the approved region.
 - D. Keep the fallback because availability always overrides the region restriction.
-

Questions 22-24

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q22 | Select ONE

A pipeline contains a simple label extraction step and a difficult exception analysis step. Higher reasoning effort improves only exception analysis and increases cost for both steps. What should the team do?

- A. Choose effort settings solely from output length, without task evaluation.
 - B. Use higher effort selectively where measured benefits justify its cost and latency.
 - C. Disable reasoning for every task because one step did not improve.
 - D. Enable maximum effort globally because all tasks benefit equally.
-

Q23 | Select ONE

A legal drafting service repeatedly uses the same approved instructions and glossary. A unique request timestamp is inserted before them, and cache reads are unexpectedly low. Which change most directly addresses the cache issue?

- A. Increase the output token limit without changing the prompt layout.
 - B. Add more timestamps throughout the prompt to help the cache identify each request.
 - C. Place stable reusable content before changing fields, then verify cache usage and invalidation behavior.
 - D. Move the glossary into the generated answer so it is available for the next user.
-

Q24 | Select ONE

A formatting Skill from an unknown publisher includes a script that reads credentials and uploads files to an unrelated endpoint. What is the most appropriate response?

- A. Trust it if the Skill name matches the requested task.
 - B. Reject or remediate the bundle before use, audit its behavior against its purpose, and apply runtime confinement to approved versions.
 - C. Run it once in production and inspect the final formatted output.
 - D. Add an instruction telling the Skill to ignore any malicious code.
-

Questions 25-27

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q25 | Select ONE

A team wants to replace its current model with a cheaper one. The candidate passes average quality targets but fails a rare, contractually important document category. What is the best next step?

- A. Use the cheaper model globally and ask it to report higher confidence.
 - B. Exclude the difficult category from reporting without changing the service scope.
 - C. Replace the model globally because the aggregate score passes.
 - D. Keep a validated model for the failing category and evaluate a selective routing strategy.
-

Q26 | Select ONE

A successor proposes replacing a retrieval strategy to reduce cost. The current strategy was chosen to satisfy a critical data-handling constraint, but the diagram does not explain why. Which artifact is most needed?

- A. A more attractive copy of the same diagram with no rationale.
 - B. A statement that all original decisions must remain unchanged forever.
 - C. A list of model names without the reasons for selecting them.
 - D. A dated decision record with rejected alternatives, the governing constraint, tradeoffs, owner, and supporting evidence.
-

Q27 | Select ONE

After a release, p95 response time rises from 5 to 14 seconds. A trace shows 9 seconds in one catalog tool; model time and token counts are unchanged. What should be investigated first?

- A. The catalog dependency and its recent behavior, using the slow tool span and request details.
 - B. The embedding dimension, without checking whether retrieval time changed.
 - C. The answer template, because shorter headings always remove tool latency.
 - D. The model tier, because language models are usually the slowest component.
-

Questions 28-30

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q28 | Select ONE

An AI-generated change passes existing tests, but the author cannot explain how it validates untrusted input. What should happen before merge?

- A. Remove the tests and require every line to be rewritten manually, regardless of the actual issue.
 - B. Merge it because passing tests prove every possible input is safe.
 - C. Review the behavior and security assumptions, add meaningful checks for uncovered cases, and require the author to understand the change.
 - D. Replace the author's explanation with the model's confidence score.
-

Q29 | Select ONE

A candidate prompt improves average usefulness but increases critical unsupported claims from 0.2% to 1.1%. The agreed maximum is 0.5%. What is the best release decision?

- A. Change the limit to 1.2% after observing the candidate so it can pass.
 - B. Do not approve the candidate under the current criteria; diagnose and correct the safety regression.
 - C. Release it because usefulness is the primary metric and guardrails are optional.
 - D. Average the two percentages together and compare that number with the usefulness score.
-

Q30 | Select ONE

A stakeholder requests a response-time commitment. Which definition is most useful for a production SLA discussion?

- A. The measured end-to-end percentile and window, agreed workload scope, breach threshold, owner, and required response to a breach.
 - B. A promise that the model will usually feel fast.
 - C. The fastest response observed during a demonstration.
 - D. A token-generation benchmark with no retrieval or tool time included, although the user waits for those steps.
-

Questions 31-33

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q31 | Select ONE

Three independently developed AI clients need to discover and use the same internal catalog tools through a reusable integration contract. Which mechanism best fits that requirement?

- A. An MCP server exposing a well-scoped tool interface, with authorization enforced by the backing system.
 - B. A shared API key pasted into each end-user conversation.
 - C. A separate copy of the catalog stored in each model's prompt.
 - D. A different undocumented shell command embedded in each client's system prompt.
-

Q32 | Select ONE

A proposal assumes that an assistant can answer from a private handbook that has never been provided to it. A secure retrieval service can supply the handbook, but its coverage has not been tested. What is the most defensible feasibility statement?

- A. Feasible as scoped because a large model probably learned similar handbooks.
 - B. Feasible as scoped if the model is instructed never to admit uncertainty.
 - C. Not feasible under any architecture because the handbook is private.
 - D. Feasible with constraints, contingent on authorized retrieval coverage and measured answer quality.
-

Q33 | Select ONE

A design shows input, retrieval, generation, and a user response. The sponsor also requires ongoing quality improvement based on user corrections. What is missing?

- A. A second model that rewrites every answer without retaining outcome evidence.
 - B. A prompt that tells the model it will improve automatically after each interaction.
 - C. A governed path to capture corrections, curate evaluation cases, assign review ownership, and gate future changes.
 - D. A larger output schema containing a field named `feedback_enabled`.
-

Questions 34-36

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q34 | Select ONE

An invoice service always performs extraction, validation, a policy lookup, and either human review or export. The branches are fully specified and must be replayable. What is the best control structure?

- A. An explicit workflow whose code controls transitions and whose model performs the language tasks.
 - B. Parallel agents that independently choose a destination and vote on the final action.
 - C. An agent that decides whether validation is necessary based on each invoice.
 - D. One long prompt that describes the sequence and asks the model to follow it exactly.
-

Q35 | Select ONE

Several engineering repositories use the same optional release-review procedure. Teams need to update it as a versioned package while keeping project-wide instructions short. Which approach best fits?

- A. Paste the entire procedure into every developer's personal chat and rely on manual updates.
 - B. Place the full procedure in every tool response, including unrelated calls.
 - C. Package the procedure as a governed Skill or plugin, and keep only the essential baseline in project instructions.
 - D. Create a permanent specialist agent solely because the procedure is reusable.
-

Q36 | Select ONE

A deployment requires a quarterly data-boundary review even if all performance metrics remain healthy. How should this be represented?

- A. As a note in the system prompt asking the model to remember the audit date.
 - B. As a calendar-triggered governance task with an owner, review evidence, and escalation for findings.
 - C. As a one-time launch checkbox with no recurring owner.
 - D. As an alert that fires only when model accuracy drops.
-

Questions 37-39

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q37 | Select TWO

A team compares two prompts using different customer cohorts on different weekdays, then reports the most favorable of six metrics. Which TWO changes most improve the experiment?

- A. Randomly assign comparable traffic with stable session treatment and control important input differences.
 - B. Stop the test whenever any metric briefly favors the new prompt.
 - C. Remove all uncertainty estimates from the report to make the result clearer.
 - D. Keep the cohorts different so the new prompt receives easier cases.
 - E. Predefine a primary metric, guardrails, and sample-size assumptions before inspecting the result.
-

Q38 | Select ONE

A payment-status tool is read-only, but a separate charge tool can move money. Network timeouts leave charge completion uncertain. Which integration design best prevents duplicate charges during recovery?

- A. Use idempotency identifiers and authoritative completion checks for charge retries, with bounded retry behavior.
 - B. Ask the model whether the charge probably completed and act on its confidence.
 - C. Retry every timeout indefinitely because timeouts prove that nothing happened.
 - D. Treat read-only and money-moving tools as identical because both use HTTP.
-

Q39 | Select TWO

A coding assistant has access to a very large repository, but most tasks touch only a few files. Which TWO practices best support effective context use?

- A. Discover and load relevant files as the investigation narrows.
 - B. Treat maximum context capacity as a target that every request must fill.
 - C. Include every repository file in every request, whether relevant or not.
 - D. Delete all source pointers once a summary has been generated.
 - E. Preserve compact task state and source pointers so important evidence can be revisited.
-

Questions 40-42

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q40 | Select ONE

A support engineer repeatedly escalates the same tool timeout to the architect. Each time, a known dependency limit is the cause. What is the best lasting improvement?

- A. Document the symptom, diagnostic evidence, safe recovery steps, and escalation boundary in a runbook, then train the team to use it.
 - B. Disable the tool's timeout so the incident is no longer reported.
 - C. Have the architect personally resolve every recurrence without documenting the cause.
 - D. Tell the team to retry indefinitely whenever any tool fails.
-

Q41 | Select ONE

A tenant identifier supplied in the user message is used to filter a document search. Users can edit this field. Which change is required?

- A. Derive the tenant scope from verified server-side identity and enforce it in retrieval and tool access.
 - B. Tell the model to reject any tenant identifier that looks suspicious.
 - C. Log the supplied tenant identifier after answering so incidents can be investigated.
 - D. Use a larger embedding model to reduce the chance of retrieving another tenant's records.
-

Q42 | Select ONE

A 120-person team has uneven Claude adoption and incompatible local configurations. Which rollout plan best supports sustainable use?

- A. Require a fixed number of prompts per employee without changing workflows.
 - B. Leave each developer to configure tools independently and compare only total usage.
 - C. Give everyone unrestricted access to every tool and the most expensive model.
 - D. Establish a shared baseline, pilot real workflows with team champions, and expand in batches while tracking quality and cost.
-

Questions 43-45

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q43 | Select ONE

A maintenance corpus contains error codes such as ZX-417 and narrative descriptions of symptoms. Users search with either exact codes or natural-language descriptions. What retrieval strategy should be evaluated?

- A. A hybrid lexical and semantic strategy, evaluated separately on code and narrative query slices.
 - B. Lexical retrieval only, assuming every user knows the exact error code.
 - C. Semantic retrieval only, with no evaluation of exact-code queries.
 - D. Randomly selecting chunks to avoid favoring either query style.
-

Q44 | Select ONE

A new agent design improves a task score from 91% to 92%, but its p95 latency is 18 seconds against a hard 6-second requirement. A workflow achieves 91% and 4 seconds; the required quality floor is 90%. Which recommendation is best supported?

- A. Combine both designs sequentially to preserve the higher score.
 - B. Use the agent and remove latency measurement from the release criteria.
 - C. Use the workflow for the current release because it meets both mandatory constraints.
 - D. Use the agent because the highest task score should determine architecture.
-

Q45 | Select TWO

A team moves a Claude workload from one delivery route to another. The application assumes identical model identifiers, feature support, and regional behavior. Which TWO actions are most appropriate before cutover?

- A. Verify model mappings and required feature support on the target route, then run representative regression tests.
 - B. Reuse every configuration unchanged because the model family name is the same.
 - C. Assume a regional client endpoint proves that every log and fallback stays in that region.
 - D. Validate authentication, permitted processing locations, logging, and fallback behavior on the target configuration.
 - E. Approve cutover based only on a successful hello-world request.
-

Questions 46-48

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q46 | Select TWO

A sponsor has request counts, average model latency, and error rates but cannot justify expanding an assistant that was intended to shorten case resolution. Which TWO additions best support the expansion decision?

- A. An auditable measurement source and a named owner for ongoing outcome tracking.
 - B. A list of future features unrelated to the original business goal.
 - C. A claim that request growth alone proves financial return.
 - D. Comparable before-and-after case-resolution measurements using the same definition and documented limitations.
 - E. More detailed token-generation benchmarks without case outcomes.
-

Q47 | Select ONE

A sponsor says, "The assistant must feel seamless." Which response best turns this preference into an architectural requirement?

- A. Record seamless as a completed requirement and choose an agent architecture immediately.
 - B. Ask which model the sponsor prefers and use that as the only design input.
 - C. Ask what waiting, re-entry, tool switching, and failure behavior would break that experience, then record measurable constraints and assumptions.
 - D. Assume it means a one-second deadline and treat that number as stakeholder-approved.
-

Q48 | Select ONE

A design document says all processing stays within an approved boundary. No owner is assigned, and the logging configuration changed last month. Which evidence is most useful now?

- A. The model's general safety benchmark score.
 - B. The original architecture diagram alone, because it was approved at launch.
 - C. A system prompt stating the approved region.
 - D. A current data-flow and configuration record covering processing, logs, caches, and fallbacks, with an accountable owner.
-

Questions 49-51

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q49 | Select ONE

A policy answer often retrieves an exception paragraph without the parent rule or section title. The model then applies the exception too broadly. Which change should be tested first?

- A. Preserve section hierarchy and parent context in chunking or retrieval expansion.
 - B. Reduce every chunk to one sentence regardless of document structure.
 - C. Remove source citations because they expose the fragmented context.
 - D. Increase output creativity so the model can reconstruct the parent rule.
-

Q50 | Select ONE

A long-running case assistant summarizes older turns to save context. Later, it acts on the wrong case because an exact case ID was changed in the summary. Which change best addresses the failure?

- A. Ask for a more fluent summary and discard the original IDs after compaction.
 - B. Increase creativity so the model can reconstruct the missing identifier.
 - C. Store the case ID only in a free-form model confidence statement.
 - D. Keep authoritative IDs and pending approvals in structured state, and compact explanatory narrative separately.
-

Q51 | Select ONE

A team wants to compare a candidate agent on copied production traffic without affecting customers. The agent has a tool that sends notifications. What must the shadow design include?

- A. A sandbox or stub for side-effecting tools, with separate traces and outcome evaluation.
 - B. A prompt asking the candidate not to send duplicate messages while leaving the tool unrestricted.
 - C. Removal of all output logging so the experiment has no visible effect.
 - D. Real notification delivery from both agents so their behavior can be compared accurately.
-

Questions 52-54

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q52 | Select ONE

A requirements document includes 180-day transcript retention, but discovery notes mention only faster response drafting and human approval before sending. What should the architect do?

- A. Treat retention as confirmed because it appears in the requirements document.
 - B. Label retention as an unconfirmed assumption, assign an owner, and resolve purpose and permitted duration before relying on it.
 - C. Delete all audit requirements because retention was not discussed.
 - D. Choose the longest possible duration so future questions can always be answered.
-

Q53 | Select ONE

For this question only, input costs \$2 per million tokens and output costs \$10 per million tokens. Each of 10,000 monthly calls uses 3,000 input tokens and 500 output tokens. There is no caching, retry, or other charge. What is the monthly model cost?

- A. \$60.
 - B. \$50.
 - C. \$1,100.
 - D. \$110.
-

Q54 | Select ONE

A service reviews an uploaded maintenance report, consults a stable engineering manual, and confirms whether a replacement part is available now. Which composition is most appropriate?

- A. A manual-only retrieval pipeline with no access to the uploaded report.
 - B. One vector index containing the report, manual, and last month's inventory export.
 - C. Document extraction, retrieval over the manual, and an authorized lookup against the inventory system.
 - D. An autonomous writing agent that estimates availability from the manual.
-

Questions 55-57

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q55 | Select ONE

An assistant answers product documentation questions correctly, but shipping dates are sometimes wrong because the search index refreshes nightly. Which integration change best addresses this?

- A. Refresh the same index weekly to stabilize the model's context.
 - B. Increase semantic similarity thresholds for shipping-date queries.
 - C. Ask the model to infer the latest date from the previous conversation.
 - D. Keep documentation retrieval and use an authorized order-service lookup for the current shipping date.
-

Q56 | Select ONE

An assistant needs per-request debugging, but raw documents contain confidential employee details. Which logging design best balances the requirements?

- A. Remove user names only and assume all remaining text is non-sensitive.
 - B. Record trace IDs, versions, timings, and permitted outcome metadata; retain any necessary sensitive evidence in a restricted, governed store.
 - C. Disable all traces because any observability is incompatible with confidentiality.
 - D. Write every raw prompt and document into a public team dashboard for easy debugging.
-

Q57 | Select ONE

A successful prototype is ready for handoff, but the evaluation report, operating owner, and incident runbook are missing. What is the most appropriate lifecycle decision?

- A. Send the source code and assume ownership transfers automatically.
 - B. Declare handoff complete because the demo ran successfully once.
 - C. Skip operational documentation because the model provider will resolve application incidents.
 - D. Complete and review the acceptance evidence and operating handoff artifacts before treating the deployment as ready.
-

Questions 58-60

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q58 | Select ONE

An LLM judge gives high scores to verbose answers that expert reviewers consider weakly grounded. What should be done before using its scores as a release gate?

- A. Discard expert labels because automated scores are inherently more objective.
 - B. Use the same judge without changes and report only the mean score.
 - C. Increase the judge's output length so it can defend its original scores.
 - D. Refine the rubric and calibrate the judge against expert-labelled cases, including concise and verbose answers.
-

Q59 | Select ONE

A system has a 30-second end-to-end deadline. It streams a status message in 1 second, but completes the grounded answer in 42 seconds at p95. Which statement is correct?

- A. The deadline is not met; measure and optimize the full path rather than treating the status message as completion.
 - B. The deadline is met because the first visible token arrived within 30 seconds.
 - C. The deadline is met if the generated answer has the correct JSON schema.
 - D. The deadline can be assessed from average daily request count alone.
-

Q60 | Select TWO

Reviewers receive hundreds of low-risk drafts and a few high-stakes decisions in one queue. They see only the output and approve almost everything. Which TWO changes best improve oversight?

- A. Keep the same queue and add a second identical approve button.
 - B. Treat approval counts as proof of review quality.
 - C. Route by stakes and required policy gates so reviewer capacity is reserved for consequential cases.
 - D. Remove mandatory high-stakes approvals to reduce waiting time.
 - E. Provide source inputs, the proposed output or action, and the reason each item requires review.
-

Questions 61-63

SELECT THE BEST ANSWER UNDER THE STATED CONSTRAINTS

Q61 | Select ONE

An assistant reduces average report drafting time from 18 minutes to 7 minutes. It handles 2,000 reports per month, and saved time is valued at \$0.50 per minute. Model, review, and operating costs total \$3,000 per month. Under these assumptions, what is the estimated net monthly benefit?

- A. \$8,000.
 - B. \$3,000.
 - C. \$15,000.
 - D. \$11,000.
-

Q62 | Select ONE

An orchestrator can run for 20 minutes. If its process restarts, all worker status and approval state are lost. Some workers perform chargeable external operations. What should be added before production?

- A. A retry that restarts every worker from the beginning after any process failure.
 - B. Durable task state, operation identifiers, and recovery logic that checks completed actions before retrying.
 - C. A longer system prompt asking the orchestrator to remember completed actions.
 - D. A larger context window so all state remains in the running process.
-

Q63 | Select ONE

An extraction API returns schema-valid JSON for every request, but 6% of account identifiers do not match the source documents. What should the architect conclude?

- A. A stronger instruction to always be accurate is sufficient to close the gap.
 - B. The system should remove the schema because structured output causes factual errors.
 - C. Structural validity is working; identifier fidelity needs source-based checks and targeted evaluation.
 - D. Schema validity proves that the identifiers are correct, so the labels must be wrong.
-