DUMPSBOSS.

Claude Certified Architect - Professional

Anthropic CCAR-P

Version Demo

Total Demo Questions: 13

Total Premium Questions: 139

Buy Premium PDF

https://dumpsboss.co

support@dumpsboss.co

support@dumpsboss.co
dumpsboss.co

QUESTION NO: 1

You are assessing data-exfiltration risk in a Claude-based assistant that has tools for both internal-document retrieval and outbound HTTP calls.

Which scenario most directly indicates a data-exfiltration risk?

A. The outbound HTTP tool returns a 200 response when the assistant calls an allow-listed domain to fulfil a user-initiated data-lookup request.

B. The user submits a routine question and receives a routine answer.

C. Adversarial content in a retrieved document instructs the model to call the outbound HTTP tool and send sensitive content to an attacker-controlled URL.

D. The internal-document retrieval tool returns a document that the user is authorized to view.

ANSWER: C

Explanation:

Adversarial content in a retrieved document instructs the model to call the outbound HTTP tool and send sensitive content to an attacker-controlled URL directly represents a data-exfiltration path. The retrieved document is untrusted data: although it may appear to the model in its working context, its embedded instructions must not be treated as authorized commands. If those instructions induce the assistant to retrieve or include sensitive internal information in an outbound HTTP request, information crosses a trust boundary and is disclosed to an external party without valid authorization. This is the core risk of indirect prompt injection in an agentic system: untrusted content influences tool use, while a network-capable tool provides a mechanism to transmit data. A secure design constrains this pathway through least-privilege tool permissions, strict outbound destination allowlists, separation of retrieval and external communication capabilities, validation of tool arguments, and user or policy approval before transmitting sensitive data. Anthropic's guidance identifies prompt injection as a material risk when models can access tools and emphasizes limiting the impact of potentially malicious instructions through safeguards and constrained permissions. See Anthropic's prompt-injection risk guidance and the tool-use implementation documentation.

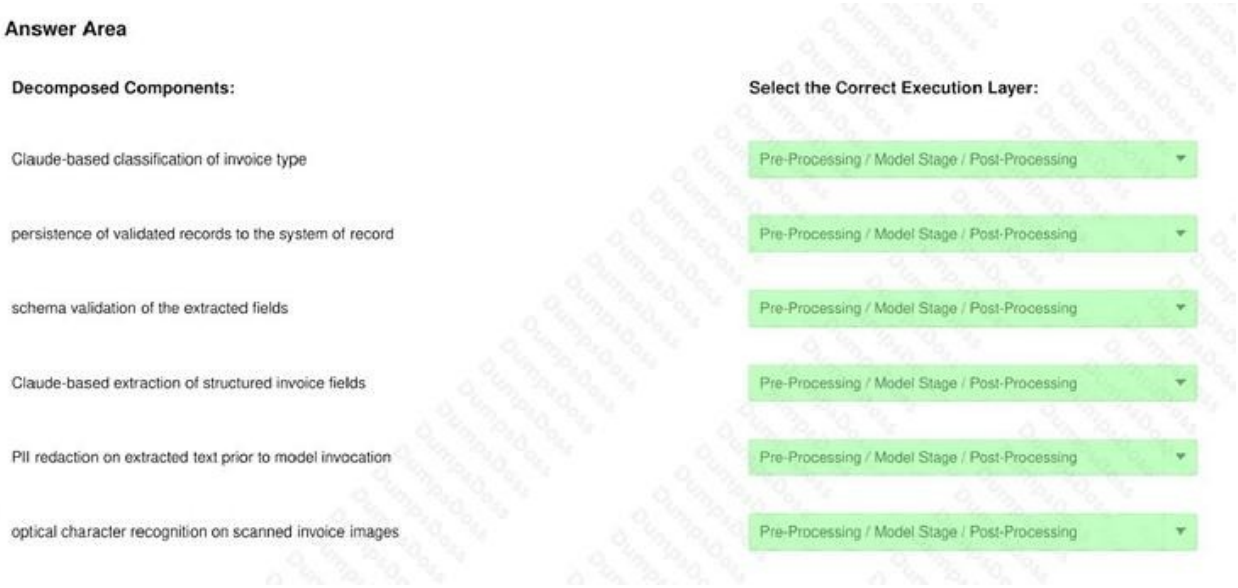
QUESTION NO: 2 - (HOTSPOT)

You are sequencing decomposed components in an invoice-processing pipeline.

For each of the decomposed components, select the execution layer it belongs to: â€œPre-Processing,â€ â€œModel Stage,â€ or â€œPost-Processing.â€

ANSWER:

Answer Area



Explanation:

The correct execution-layer assignments separate deterministic document handling, Claude inference, and reliable downstream controls into clear stages. Select **Model Stage** for Claude-based classification of invoice type because this requires interpreting the invoice's wording, layout-derived text, and business context to decide what kind of invoice it is. Select **Model Stage** for Claude-based extraction of structured invoice fields because Claude is being used to transform unstructured invoice content into meaningful business fields such as supplier, invoice number, dates, totals, and line-item information. Anthropic's guidance on building with Claude describes using the model for language-understanding and generation tasks: Build with Claude overview.

Select **Pre-Processing** for optical character recognition on scanned invoice images. OCR makes image-based source material machine-readable before the model receives relevant content. Also select **Pre-Processing** for PII redaction on extracted text prior to model invocation. Redacting sensitive information before an inference request implements data-minimization and privacy controls at the boundary of the model workflow. Claude's API documentation provides broader guidance for designing requests and handling application inputs: Anthropic API getting started guide.

Select **Post-Processing** for schema validation of the extracted fields. After Claude returns structured content, deterministic software should check required properties, expected formats, permitted values, numeric constraints, and the overall output structure before the result is trusted by downstream systems. Select **Post-Processing** for persistence of validated records to the system of record. This places the database write after validation, creating a controlled handoff from probabilistic model output to authoritative business data. This decomposition keeps sensitive-input controls and document preparation ahead of inference, reserves Claude for tasks that benefit from semantic reasoning, and applies deterministic validation and durable storage controls after inference. It supports a more auditable invoice-processing pipeline with clear boundaries between model behavior and conventional application logic.

QUESTION NO: 3

You are operating an interactive assistant whose dominant performance constraint is per-turn latency. Quality on routine turns is already acceptable.

Which configuration adjustment most directly improves latency without disproportionately damaging quality?

A. Disable prompt caching entirely to ensure fresh context processing on every request, preventing stale prefix content from affecting latency-sensitive interactions.

B. Reduce retrieval depth to the top-k passages that historically cover the answer, and cache stable system-prompt content.

C. Increase retrieval depth to the corpus maximum to improve recall regardless of latency.

D. Switch every turn to the heaviest available model to maximize output quality, accepting that the increased model latency will worsen the per-turn SLO rather than improve it.

ANSWER: B

Explanation:

Reduce retrieval depth to the top-k passages that historically cover the answer, and cache stable system-prompt content. This targets two major contributors to interactive latency while preserving the information most likely to be useful. Retrieval adds both search time and prompt tokens; using a validated, smaller top-k reduces retrieval and context-processing overhead when historical evaluation shows that those passages adequately cover routine requests. Fewer input tokens can reduce time to first token and overall response time, particularly where long retrieved context would otherwise be repeatedly processed.

Caching stable system-prompt content further improves repeated-turn performance. Anthropic prompt caching allows recurring prompt prefixes, such as system instructions, tools, and other static context, to be reused rather than reprocessed on each request. This is especially appropriate for an interactive assistant because the stable prefix is shared across turns, whereas the user-specific portion remains fresh. Since routine quality is already acceptable, combining empirically bounded retrieval with reuse of invariant context is a proportionate optimization: it removes unnecessary work without discarding the retrieval capability or changing the model's normal answer-generation behavior. Anthropic documents prompt caching and its latency benefits at Prompt caching; guidance on managing context and token use is available in the context windows documentation.

QUESTION NO: 4 - (SIMULATION)

You are reviewing a peer's end-to-end design for a Claude-based platform expected to scale to thousands of concurrent users.

For each statement, indicate Yes if it reflects sound architectural practice. Otherwise, select No.

ANSWER: Check the explanation for the answer

Explanation:

Selections:

Yes — Authentication and authorization run before retrieval so retrieval can filter by identity.
Yes — An asynchronous queue absorbs bursty traffic between intake and the model-invocation layer.
No — Tax computation is encoded directly in the system prompt rather than in code.
No — Conversation logs include unredacted government identifiers to maximize tuning signal.

Access control must occur before retrieval so unauthorized documents or data cannot be added to the Claude context. Queues provide buffering, backpressure, controlled concurrency, and retries for high-volume or bursty workloads. Tax calculations are deterministic business logic and should be implemented in tested code or a controlled tool, not delegated to prompt-based reasoning. Government identifiers are sensitive personal data and should be redacted, tokenized, minimized, or otherwise protected before logging.

QUESTION NO: 5

A research summarization assistant has been deployed for six months. A user has flagged that a generated summary contained a fabricated citation. The product team has asked whether the incident requires architectural action or whether it is an isolated case.

Which two Diligence-competency actions should you take? (Select two.)

Each correct answer presents part of the solution.

A. Sample recent summaries to estimate the fabrication frequency across the population.

B. Disable the assistant immediately for all current users without any prior diagnostic analysis.

C. Review whether citation-grounding controls exist anywhere in the current generation pipeline.

D. Communicate to users that the assistant does not fabricate output as a matter of design.

E. Treat the incident as an anecdotal isolated case and take no further investigative action.

ANSWER: A C

Explanation:

Sample recent summaries to estimate the fabrication frequency across the population. is correct because a reported fabricated citation is meaningful evidence of a failure, but the team needs a systematic sample to determine its prevalence, severity, and distribution. Sampling should include varied source documents, prompt patterns, user workflows, and model or application versions. This produces measurable evidence for deciding whether the issue is localized or indicates a recurring reliability problem, and it establishes a baseline for validating improvements.

Review whether citation-grounding controls exist anywhere in the current generation pipeline. is also correct because trustworthy research summaries require architectural controls that connect generated claims and citations to supplied source material. The review should establish whether the pipeline constrains the model to authorized documents, supports citations from those documents, and provides a way to verify that cited passages actually support the associated claims. Anthropic's citation capability is designed to return source-based citations and supporting excerpts, making outputs more auditable. Grounding guidance likewise emphasizes supplying relevant context and directing the model to rely on it. Together, evidence sampling and a grounding-control review support a disciplined decision about remediation and ongoing monitoring. See Anthropic Citations documentation and Anthropic grounding guidance.

QUESTION NO: 6

A Claude Architect is reviewing a post-deployment performance report for an AI-assisted legal-document summarization system. The report includes these observations:

Average summarization time decreased from 47 minutes to 6 minutes per document.

Associates spend less time on summaries, but overall billable output has not measurably changed.

Infrastructure costs increased by 22% because redundant retry logic generated additional API calls.

Some summaries require attorney correction, adding an average of 8 minutes of review per document.

Which analysis correctly attributes each observation to the appropriate business-value pillar?

A. Observations 1 and 2 both indicate efficiency gains; Observation 3 is a solution-cost issue; Observation 4 is a performance-SLA issue.

B. Observation 1 is a transformation outcome; Observation 2 is an efficiency gain; Observation 3 is a performance-SLA degradation; Observation 4 is a solution-cost issue.

C. Observations 1 and 4 together indicate a net performance-SLA improvement; Observation 2 is a transformation gap; Observation 3 is a productivity drain from over-engineering.

D. Observation 1 indicates an efficiency gain; Observation 2 shows that productivity has not yet been realized; Observation 3 is a solution-cost issue; Observation 4 is an efficiency loss that partially offsets Observation 1.

ANSWER: D

Explanation:

Observation 1 indicates an efficiency gain; Observation 2 shows that productivity has not yet been realized; Observation 3 is a solution-cost issue; Observation 4 is an efficiency loss that partially offsets Observation 1. This analysis properly distinguishes operational efficiency from realized business productivity and total solution cost. Reducing the initial summarization time from 47 minutes to 6 minutes is a substantial task-level efficiency improvement. However, lower time spent by associates does not by itself establish productivity value: the absence of measurable improvement in overall billable output shows that the saved capacity has not yet translated into the intended organizational outcome. The increased API usage caused by retry behavior belongs in solution cost, since it raises the cost to operate the deployed system. Attorney review and correction time is additional human effort required to produce an acceptable result, so it must be included when calculating net efficiency. A sound value assessment therefore measures end-to-end work, including generation, review, correction, and operating costs, rather than relying on model response time alone. Anthropic recommends evaluating systems against task-specific quality criteria and monitoring production performance continuously; see the Anthropic evaluation guide and prompt engineering overview.

QUESTION NO: 7

You are identifying the highest-impact optimization for a deployment whose token cost is dominated by a long, repeated system prompt and a large retrieved context per request.

Which optimization most directly targets the dominant cost driver?

A. Increase retrieval depth on every request to maximize recall, worsening the dominant cost driver by adding more retrieved tokens per request rather than reducing them.

B. Add additional repeated content to the system prompt to give the model more guidance.

C. Move the long, repeated system prompt into a cacheable prefix and trim retrieved context to the spans relevant to each query.

D. Switch every request to the heaviest available model to maximize output quality, accepting that higher per-request inference cost compounds rather than addresses the dominant cost driver.

ANSWER: C

Explanation:

Move the long, repeated system prompt into a cacheable prefix and trim retrieved context to the spans relevant to each query. Prompt caching is designed for content that remains the same across requests, such as system instructions, reference material, and repeated context. Once the repeated prefix has been cached, subsequent requests can reuse it rather than repeatedly incurring the normal full input-token processing cost. This directly reduces the cost contribution of a long system prompt when it is shared across calls. Anthropic describes prompt caching as especially useful for applications with repeated prompts or large contextual material, and notes that the cached prompt prefix must be placed before variable request content.

Reducing retrieved context to query-relevant passages addresses the other stated source of input-token cost. Retrieval should supply sufficient evidence for the task, but unnecessary chunks, duplicated passages, and overly broad documents increase token usage on every request. Selecting, reranking, and passing only the most relevant spans lowers the variable portion of the input while retaining useful grounding. Together, caching the stable prefix and minimizing the per-query retrieval payload target both dominant drivers rather than optimizing an unrelated part of the workload. See Anthropic's Prompt caching documentation and context-window guidance.

QUESTION NO: 8

You are listing characteristics of strong architectural-decision communication.

Which two characteristics belong on the list? (Select two.)

Each correct answer presents a complete solution.

A. Ownership and review cadence for revisiting the decision when conditions or assumptions change.

B. Distribution to a standing review forum with mandatory attendance from senior engineering leaders.

C. Omission of rejected alternatives to keep the decision document focused and concise for readers.

D. Alternatives considered along with the criteria that were used to evaluate each alternative.

E. Use of a consistent template across the team to standardize the visual presentation of decisions.

ANSWER: A D

Explanation:

Strong architectural-decision communication makes both the decision rationale and its ongoing governance visible. **Ownership and review cadence for revisiting the decision when conditions or assumptions change.** is essential because architecture choices are made against assumptions that can evolve, including expected traffic, cost constraints, security requirements, product scope, model capabilities, and operational risks. Naming accountability and defining when

reassessment occurs ensures that a decision remains a managed artifact rather than an undocumented, permanent constraint.

Alternatives considered along with the criteria that were used to evaluate each alternative. provides the evidence trail needed for reviewers and later maintainers to understand why a particular approach was selected. Recording the alternatives and evaluation criteria makes trade-offs explicit, supports consistent review against requirements, and enables future teams to determine whether changed conditions justify revisiting the conclusion. This is especially valuable in AI architectures, where model behavior, latency, cost, safety controls, and tool-integration needs may change over time.

Anthropic's guidance emphasizes designing AI systems around clear requirements, evaluation, and iterative improvement, all of which benefit from decisions that are traceable and reviewable. See Anthropic's prompt engineering overview and Anthropic's guidance on building effective agents.

QUESTION NO: 9

A Claude architect is leading the discovery phase for a new AI-powered customer service solution.

Which two activities are characteristic of structured discovery and requirement gathering for a Claude-based deployment? (Select two.)

A. Facilitating stakeholder workshops to surface latency, accuracy, and compliance constraints before scoping begins.

B. Selecting the Claude model tier based on the architect's prior project experience before stakeholder input is collected.

C. Generating an initial prototype and iterating based on user reaction rather than written requirements.

D. Documenting explicit success criteria and failure thresholds that will gate production deployment.

E. Deferring constraint documentation until the integration design phase to avoid scope creep.

ANSWER: A D

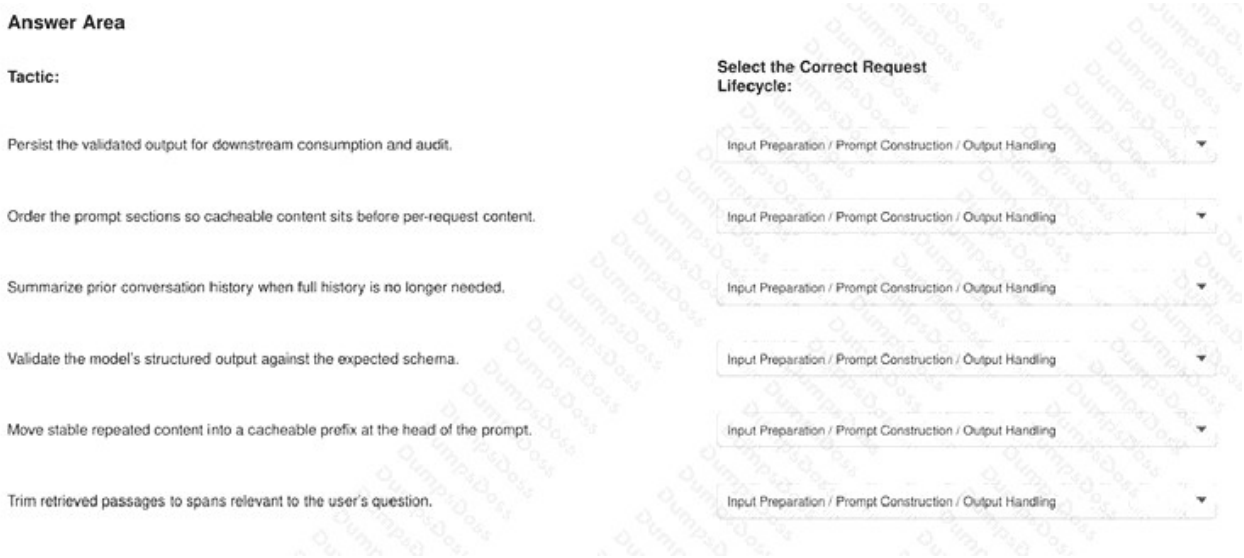
Explanation:

Facilitating stakeholder workshops to surface latency, accuracy, and compliance constraints before scoping begins is a core discovery activity because it establishes the business context and operational boundaries that the solution must meet. Input from business owners, end users, support teams, security, legal, and compliance stakeholders helps convert a broad customer-service goal into testable functional and nonfunctional requirements. This includes expected response quality, peak demand, permitted data handling, escalation paths, and acceptable risk.

Documenting explicit success criteria and failure thresholds that will gate production deployment turns those requirements into measurable standards for evaluation and release decisions. Anthropic recommends defining clear success criteria before building evaluations, then using representative tasks and grading methods to measure whether the application meets its intended outcomes. For a customer-service workflow, criteria can include resolution quality, policy adherence, latency, and correct handoff to a human agent. Failure thresholds clarify when the system must be blocked, reviewed, or remediated before production use. Together, stakeholder-derived constraints and measurable deployment gates provide traceability from business objectives through evaluation and operational readiness. See Anthropic's developing tests and evaluations guidance and prompt engineering overview.

QUESTION NO: 10 - (SIMULATION)

You are classifying token-management tactics by where each tactic applies in the request lifecycle: "Input Preparation," "Prompt Construction," or "Output Handling."



ANSWER: Check the explanation for the answer

Explanation:

Select the following lifecycle stage for each tactic:

Reasoning: Input preparation curates the material sent to the model, including retrieved context and conversation history. Prompt construction controls the prompt's arrangement and stable cacheable prefix. Output handling occurs after generation, including validation, storage, audit, and delivery of model results.

QUESTION NO: 11

You are producing an architecture guide for a new deployment and must complete the planning steps before drafting each section.

Which two steps must be completed BEFORE drafting each section of the guide? (Select two.)

Each correct answer presents part of the solution.

A. Identify the audience and the questions the guide must answer for that audience.

B. Translate the guide into the supported regional languages for the candidate population.

C. Validate the guide with the implementation team and incorporate corrections.

D. Establish the document under version control with a defined review cadence and approver list.

E. Outline the guide sections covering the overview, components, contracts, flows, runbooks, and limitations.

ANSWER: A E

Explanation:

Identifying the audience and the questions the guide must answer for that audience is a necessary pre-drafting activity because an architecture guide must be tailored to the decisions its readers need to make. For example, an implementation team may need concrete interface details and operational procedures, while technical decision-makers may need the system context, trade-offs, and component responsibilities. Defining the readers' questions before writing establishes the appropriate scope, terminology, level of detail, and success criteria for every section. This aligns with established technical-writing guidance to identify an audience and purpose before creating content; see the Google Technical Writing guidance on defining an audience.

Outlining the guide sections covering the overview, components, contracts, flows, runbooks, and limitations must also occur before drafting. A section outline supplies the guide's information architecture: it organizes content into a coherent sequence

and ensures that the architecture's structural, behavioral, operational, and constraint-related information is planned rather than added opportunistically. This allows each drafted section to have a defined objective and prevents important topics from being omitted or duplicated. The practice is consistent with the Write the Docs documentation principles, which emphasize organizing documentation around users' needs and maintaining clear structure.

QUESTION NO: 12

A security audit uncovers two issues: (1) all end users share a single API key, and (2) tool calls are executed without logging the initiating user.

Which two mitigations directly address these specific findings? (Select two.)

A. Validate structured outputs against a schema before downstream actions are executed.

B. Enforce RBAC at the retrieval layer before content enters the model context.

C. Move credentials out of the prompt context and resolve them from a server-side secret store.

D. Add actor attribution to tool-call logs so each call records the initiating user identity.

E. Replace the shared API key with per-user OAuth tokens carrying scope-restricted permissions.

ANSWER: D E

Explanation:

Replace the shared API key with per-user OAuth tokens carrying scope-restricted permissions because OAuth access tokens represent an individual authenticated principal or delegated authorization. This removes the indistinguishable shared-credential model and enables permissions to be limited to the scopes needed for that user and use case. Per-user tokens also support independent expiration and revocation, reducing the impact if one credential is compromised. In an MCP integration, OAuth is designed to let a client obtain and present authorization for a specific user, rather than relying on a common credential shared by all users. Anthropic's MCP guidance discusses OAuth-based authorization for remote MCP servers: Anthropic MCP documentation.

Add actor attribution to tool-call logs so each call records the initiating user identity because identity-aware audit records provide accountability for consequential tool activity. Tool execution logs should associate the authenticated user or delegated identity with the request and capture relevant execution context such as the tool invoked, timestamp, outcome, and request or session identifier. This makes it possible to investigate actions, correlate events, and determine who initiated an operation. Anthropic's Claude Code security guidance emphasizes protecting credentials and maintaining appropriate security controls around tool-enabled workflows: Anthropic Claude Code security documentation.

QUESTION NO: 13

You are reviewing a peerâ€™s draft system prompt that contains contradictory instructions: one section says never to speculate beyond the supplied source, while another says to confidently fill in any gaps.

Which response is most appropriate?

A. Rewrite the prompt to use one grounding policy: answer only from the supplied source, and state when the evidence is insufficient rather than filling gaps.

ANSWER: A

Explanation:

Resolve the contradiction by rewriting the system prompt into one explicit grounding policy: Claude should answer only from the supplied source, cite or identify the supporting material when appropriate, and state that there is insufficient evidence when the source does not support an answer. This gives the model a clear, actionable rule for handling missing information rather than leaving it to infer which competing instruction should govern a particular response. Clear system-level instructions are particularly important for retrieval-grounded or high-stakes applications, where unsupported completion can

turn an absence of evidence into an apparently confident claim. The prompt can still be useful and helpful by directing Claude to explain what information is missing or ask for an additional source, rather than speculating. Anthropic's prompting guidance recommends instructions that are clear, direct, and specific, and its system-prompt guidance describes using system prompts to define consistent behavior and constraints. A single coherent instruction establishes a predictable policy across requests and makes the intended behavior easier to evaluate and maintain. See Anthropic's prompt engineering overview and system prompts documentation.