Exam : **CCAR-P**

Title : Claude Certified Architect - Professional

Version : DEMO

1.You are reviewing instrumentation in a multi-agent system.
Which two findings constitute valid observability gaps in the instrumentation? Each correct answer presents a complete solution. (Select two.)
A. Trace spans for each agent step are exported to the shared distributed-tracing backend.
B. Latency and token usage on every span are emitted to the central metrics pipeline.
C. Tool-call payloads and outcomes are recorded with redaction applied to known sensitive fields.
D. Model identity and version on each turn are not recorded with the turn artifacts.
E. Request-scoped correlation identifiers do not propagate across agent and tool calls.
Answer: D, E
Explanation:
Model identity and version are essential diagnostic dimensions. Without them, operators cannot correlate a behavioral change with a model migration, reproduce the conditions of a failed turn, compare performance among model versions, or distinguish model drift from prompt, retrieval, and application defects.
Therefore, the omission described in Option D is a genuine observability gap.
Option E is equally significant. A multi-agent execution normally crosses orchestration, model, agent, MCP, and tool boundaries. If the request-scoped correlation identifier or distributed trace context is lost at any boundary, the resulting spans cannot reliably be reconstructed into one end-to-end transaction.
Anthropic’s monitoring guidance explains that distributed tracing links a user prompt to the API requests and tool executions it initiates; it also documents propagation of W3C trace context to supported subprocesses and outbound MCP requests. Claude Code Monitoring
Options A and B describe healthy telemetry coverage rather than gaps.
Option C is also appropriate because recording tool interactions supports investigation, while redaction reduces the risk of exposing sensitive data in logs.
Study Guide references/topics: Integration—observability challenges at scale; distributed tracing; correlation identifiers; model-version attribution; privacy-aware tool telemetry.

2.You are designing a content moderation classifier that processes high volumes of user-generated comments under a tight per-message latency budget using well-defined classification labels.
Which model selection best aligns with the workload?
A. Opus, because every moderation decision requires maximum reasoning depth regardless of classification complexity.
B. Haiku, because its latency and cost profile align with high-volume classification workloads that require limited reasoning depth.
C. Sonnet, because larger general-purpose models are preferred even when workload latency requirements are strict.
D. Sonnet with extended thinking enabled, because deeper reasoning should be applied to every moderation request to improve edge-case handling.
Answer: B
Explanation:
Haiku is the appropriate starting point because the workload is high-volume, latency-sensitive, and based on a stable closed set of moderation labels. These characteristics favor a fast, cost-efficient model capable of consistent classification without incurring the additional inference time and expense associated with deeper reasoning.

Anthropic’s model-selection guidance requires architects to balance capability, speed, and cost rather than automatically selecting the most capable model. Its content-moderation guidance specifically identifies Haiku as a cost-effective option for processing moderation workloads at substantial scale.
Choosing the Right Model, Content Moderation
Opus is disproportionate to a routine closed-set classification problem. Sonnet may become justified if evaluation demonstrates that Haiku fails materially on complex policy distinctions, multilingual ambiguity, or adversarial edge cases, but it should not be selected merely because it is larger. Enabling extended thinking on every request would further increase latency and token consumption without evidence that the additional reasoning improves the defined success metrics. The correct architectural practice is to establish a representative moderation evaluation set, validate Haiku against accuracy and safety thresholds, and escalate only the cases that genuinely need deeper reasoning.
Study Guide references/topics: Model selection; capability–latency–cost trade-offs; classification workloads; evaluation-driven routing; moderation architecture.

3.You are compiling continuity practices that span the deployment lifecycle.
Which two practices belong on the list? Each correct answer presents a complete solution. (Select two.)
A. Maintain a stakeholder register and notify the listed parties at every phase transition event.
B. Carry the evaluation framework and reference set forward across iterations rather than rebuilding each time.
C. Archive every phase deliverable in long-term storage to preserve a record of what was produced.
D. Capture lessons learned at the end of each phase and surface them as inputs to the next phase.
E. Lock decisions made in early phases to prevent revisiting them as later phases begin.
Answer: B, D
Explanation:
Lifecycle continuity depends on preserving validated knowledge and feeding operational learning back into subsequent phases. Carrying the evaluation framework and reference set forward, as stated in Option B, creates a stable baseline across prompt changes, model migrations, retrieval adjustments, and architectural revisions. Rebuilding the evaluation system each time would undermine longitudinal comparison because changes in the test framework could be mistaken for changes in solution performance.
Option D establishes the second essential continuity mechanism: lessons from discovery, design, implementation, deployment, and production monitoring become explicit inputs to the next phase. This closes the feedback loop and prevents recurring defects, invalid assumptions, and operational findings from being lost at organizational handoffs.
Option A is overly mechanical. Stakeholders should receive communications relevant to their responsibilities and decision rights, not indiscriminate notifications at every transition.
Option C confuses comprehensive archival with lifecycle continuity; retention must follow business, regulatory, security, and records-management requirements rather than an unconditional “archive everything” policy.
Option E is directly contrary to iterative architecture. Decisions should be documented and governed, but material evidence or changed requirements must be allowed to reopen them.
Study Guide references/topics: Lifecycle phases; evaluation continuity; reference datasets; feedback loops; lessons learned; decision records; iterative architecture governance.

4.A revenue projection assistant has missed its monthly cost target by 38 percent. Profiling shows three contributors: a 6,000-token policy preamble repeated on every call (45 percent of cost), retrieval of historical sales chunks averaging 3,000 tokens per call (30 percent), and inference on a flagship-tier model (25 percent). Stakeholders require that projection accuracy remain unchanged.
Which two optimizations should you sequence first to reduce cost without affecting accuracy? Each correct answer presents part of the solution. (Select two.)
A. Reduce the number of historical sales chunks retrieved across each query run.
B. Truncate the policy preamble to remove non-essential clauses from the prompt.
C. Enable prompt caching on the static policy preamble across the recurring calls.
D. Switch the workload to a smaller, faster Claude model tier across all queries.
E. Cache common retrieved sales chunks accessed across many of the daily queries.
Answer: C, E
Explanation:
The required sequence must reduce repeated computation without changing the information or model capability used to generate projections. Prompt caching the static 6,000-token policy preamble directly addresses the largest cost contributor while preserving the complete instruction set. Anthropic states that cache reads cost substantially less than uncached input tokens, making repeated, stable prompt prefixes an ideal caching target. Prompt Caching
Caching frequently reused historical-sales chunks applies the same principle to the retrieval layer. When identical, version-controlled chunks are repeatedly fetched and supplied to the model, caching their retrieval or reusable representation eliminates redundant work while maintaining the same evidence available to the projection process. The cache must use appropriate invalidation or source-version keys so updated sales data cannot be replaced by stale content.
Options A and B modify the information supplied to Claude. Fewer sales chunks could remove relevant historical evidence, while truncating policy instructions could alter constraints or projection behavior.
Option D introduces a model-capability change and therefore cannot guarantee unchanged accuracy without a comparative evaluation. Those interventions may be considered later, but only after representative regression testing establishes equivalence.
Study Guide references/topics: Cost profiling; prompt caching; retrieval caching; cache invalidation; accuracy-preserving optimization; model and context trade-offs.

5.HOTSPOT
You are evaluating prompting claims in a peer’s design document.
For each claim, select yes if the claim reflects sound practice. Otherwise, select no.

Answer Area

Claim:	Select Yes or No:
Zero-shot is appropriate as a starting point for closed-set classification when categories and descriptions are provided in the pron	Yes / No
Adding a few-shot example set with exact field names anchors output formatting on extraction tasks.	Yes / No
Chain-of-thought should be the default on every task regardless of complexity.	Yes / No
Leading-question framing is best mitigated by asking for an even-handed comparison with stated criteria.	Yes / No
Behavioral differences across model versions can be ignored once a prompt has been validated on one version.	Yes / No

Answer:

Answer Area

Claim:	Select Yes or No:
Zero-shot is appropriate as a starting point for closed-set classification when categories and descriptions are provided in the pron	Yes / No
Adding a few-shot example set with exact field names anchors output formatting on extraction tasks.	Yes / No
Chain-of-thought should be the default on every task regardless of complexity.	Yes / No
Leading-question framing is best mitigated by asking for an even-handed comparison with stated criteria.	Yes / No
Behavioral differences across model versions can be ignored once a prompt has been validated on one version.	Yes / No