ClaudeExamPrep.com

FREE SAMPLE — 10 CCA Practice Questions

Claude Certified Architect — Foundations Exam

10 representative questions from the full 152-question set — covering all six CCA exam domains.
Get a feel for the question style, depth, and explanations before you buy.

This is a free sample. The full set contains 152 questions.

claudeexamprep.com

Sample Questions

Question 1 — Claude API Architecture

A developer needs to build a customer support chatbot that handles complex multi-turn conversations with documents up to 500 pages. Which Claude model and configuration is most appropriate?

A) Claude Haiku 4.5 with 200k context window for fastest response times

B) Claude Opus 4.6 with 1M token context window and extended thinking enabled

C) Claude Sonnet 4.6 with 1M token context window for the best balance of speed and intelligence

D) Claude Opus 4.6 with 200k context window and prompt caching

Correct Answer: C

Claude Sonnet 4.6 with the 1M token context window provides the optimal balance of speed and intelligence for a customer support chatbot handling large documents. Opus would be overkill for most support scenarios and slower, while Haiku's 200k context window may be insufficient for 500-page documents.

Question 2 — Claude API Architecture

A team is building a RAG application that sends the same 15,000-token system prompt with every request. Which optimization would reduce costs by the largest margin?

A) Switching from Opus to Haiku for all requests

B) Enabling prompt caching for the static system prompt

C) Reducing the system prompt to under 5,000 tokens

D) Using the Batch API for all requests

Correct Answer: B

Prompt caching can reduce costs by 70-80% for repeated context. Since the 15,000-token system prompt is sent with every request, caching it eliminates redundant processing and provides the most dramatic cost reduction for this pattern.

Question 3 — Prompt Engineering

When designing a system prompt for a production application, what is the recommended structure?

A) Keep it as short as possible to minimize token usage

B) Use XML tags to organize sections like role, rules, examples, and output format

C) Write it as a continuous narrative paragraph

D) Use JSON format for all system prompt content

Correct Answer: B

Anthropic recommends using XML tags (like , ,) to structure system prompts. This provides clear separation of concerns, makes the prompt easier to maintain, and helps Claude distinguish between different types of instructions.

Question 4 — Prompt Engineering

A developer wants Claude to AVOID generating certain types of content in a customer-facing app. What is the most effective prompting technique?

A) Tell Claude what NOT to do in the system prompt

B) Combine positive instructions (what TO do) with specific constraints, plus examples of correct behavior

C) Only use positive instructions and trust Claude to infer what to avoid

D) Use post-processing to filter unwanted content from responses

Correct Answer: B

The most effective approach combines positive instructions (telling Claude what to do) with specific constraints (what to avoid), backed by examples showing correct behavior. Post-processing alone is fragile and doesn't prevent the generation of unwanted content.

Question 5 — Safety & Constitutional AI

What is Constitutional AI (CAI), Anthropic's approach to AI alignment?

A) A legal framework governing AI development in the United States

B) A training method where the AI is guided by a set of principles to be helpful, harmless, and honest, using self-critique and revision

C) A hardware architecture designed for safe AI computation

D) A regulatory body that certifies AI systems as safe

Correct Answer: B

Constitutional AI is Anthropic's training approach where Claude is guided by a set of written principles (a 'constitution') to be helpful, harmless, and honest. During training, the model critiques its own outputs against these principles and revises them.

Question 6 — Safety & Constitutional AI

When deploying Claude in a production environment, which practice BEST addresses content safety?

A) Relying entirely on Claude's built-in safety training

B) Implementing defense in depth: system prompt guardrails, input validation, output filtering, and monitoring

C) Disabling all safety features for maximum helpfulness

D) Only allowing pre-approved prompts through a whitelist

Correct Answer: B

Defense in depth is the recommended approach: multiple layers of safety including system prompt guardrails, input validation/sanitization, output filtering, human review for high-risk decisions, and monitoring/alerting. No single safety measure is sufficient on its own.

Question 7 — Multi-Agent Systems

What is the Model Context Protocol (MCP) and what problem does it solve?

A) A database protocol for storing conversation history

B) An open protocol that standardizes how AI models connect to external tools, data sources, and services through a unified interface

C) A network protocol for distributing model inference across multiple GPUs

D) A compression protocol for reducing context window size

Correct Answer: B

MCP (Model Context Protocol) is an open protocol developed by Anthropic that standardizes how AI models connect to external tools, data sources, and services. It provides a unified interface for tool integration, replacing the need for custom connectors for each service.

Question 8 — Production Deployment

A production Claude application is experiencing high latency during peak hours. Which optimization should be attempted FIRST?

A) Switching to a larger Claude model for faster processing

B) Implementing streaming responses, prompt caching, and evaluating whether a faster model can handle the workload

C) Adding more API keys to increase throughput

D) Reducing the quality of responses to speed up generation

Correct Answer: B

The first optimizations should be: enabling streaming (so users see responses incrementally), implementing prompt caching (reducing processing time for repeated context), and evaluating whether a faster model can handle the workload. Streaming alone dramatically improves perceived latency.

Question 9 — Production Deployment

What is 'context rot' and how should production systems address it?

A) Physical degradation of server hardware over time

B) The degradation of model accuracy and recall as the context window fills with more tokens, addressed through context curation and summarization

C) API key expiration over time

D) Gradual decrease in model capabilities with each API update

Correct Answer: B

Context rot refers to the degradation of model accuracy as more tokens fill the context window. Production systems should address this by curating what goes into context, using summarization for older history, and placing critical information at the beginning and end of the context.

Question 10 — Model Evaluation

When selecting a Claude model for a production application, what is the recommended decision framework?

A) Always choose the newest and most capable model

B) Evaluate based on task complexity, latency requirements, cost constraints, and quality thresholds — then select the minimum capable model that meets all requirements

C) Choose based solely on benchmark scores

D) Start with Opus and never consider switching

Correct Answer: B

The recommended approach is to define requirements across four dimensions: task complexity, latency, cost, and quality. Then select the least expensive model that meets all requirements. This prevents over-provisioning while ensuring adequate performance.

Want all 152 questions?

The full set covers all 6 CCA exam domains with detailed explanations for every answer.

Get the full set at claudeexamprep.com

- 152 questions across all CCA domains
- Weighted toward high-frequency exam topics
- Detailed explanation for every answer
- Organized by domain for targeted study
- PDF format — study on any device