

PRACTICE QUESTIONS

NVIDIA

NCP-AAI

NVIDIA-Certified Professional: Agentic AI

SHARED BY [u/upbeatcompass2004](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM NVIDIA NCP-AAI, NVIDIA-Certified Professional: Agentic AI

Shared by u/upbeatcompass2004 on aiexamcerts

10 practice questions for the NVIDIA NCP-AAI exam, NVIDIA-Certified Professional: Agentic AI, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/nvidia/ncp-ai. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCP-AAI and the question number.

-
1. Which control is most effective for limiting what a compromised agent can do?
- A Giving every agent the organization's admin token for convenience during development and testing
 - B Sharing one service account across all agents
 - C Storing credentials in the prompt so the model can use them
 - D Least-privilege credentials scoped to the specific tools and data each agent role needs

Answer: D

Scoped credentials limit the damage if an agent misbehaves or is manipulated. Shared admin or prompt-stored credentials remove that boundary, which is why those choices are tempting but dangerous.

-
2. Two agents writing to the same shared-state key without locking or versioning can silently overwrite each other's updates.
- A True
 - B False

Answer: A (True)

Concurrent writes without ownership rules can lose one agent's update with no error raised. Locks, version checks or single-writer rules prevent this.

-
3. Why should the agent's worker processes be kept stateless when running several replicas?
- A Stateless workers use less disk space in every case
 - B Stateless workers cannot call tools unless every replica holds the full session in memory
 - C Any replica can then serve any turn, because session state lives in a shared store
 - D Statelessness makes the model more accurate

Answer: C

Keeping session state in a shared store lets the load balancer send turns anywhere and survive replica failures. Tool access and accuracy are unaffected, which is why those statements are wrong.

-
4. The serving technique that admits new sequences into the running batch as earlier ones finish is called _____ batching.

Answer: continuous (also accepted: in-flight, rolling)

Continuous, or in-flight, batching keeps the GPU busy by swapping finished sequences for new ones at each step. Static batching waits for a whole batch to finish, which is the less efficient alternative.

5. Which pair of checks best reduces harmful outputs from an agent that writes customer emails? (Choose two.)

- A A human approval step for messages that mention pricing or legal terms
- B Letting the model choose recipients without any allowlist
- C Turning off output logging to reduce review noise
- D An output classifier or policy check before the email is sent
- E Using a higher temperature so the email sounds more natural

Answer: A, D

Automated output checks and human review for sensitive topics catch problems before they reach customers. Free choice of recipients and disabled logs add risk, which is why those options are tempting but wrong.

6. An agent loop has no iteration limit and occasionally repeats the same failing tool call for hours. Which control addresses this most directly?

- A A larger embedding dimension
- B Lowering the top-k of the retriever to 1 so each step reads only one document at a time
- C A maximum step budget with a stop condition and escalation when the budget is exhausted
- D Switching the tokenizer to a byte-level vocabulary

Answer: C

Agent loops need a bounded number of steps, a repeated-action check and an exit path to a human or a fallback. Retriever settings and tokenization do not stop a loop, which is what makes them distractors for this failure.

7. A team uses an LLM as a judge to score agent answers. Which precaution is most important?

- A Use the same prompt for the judge as for the agent, with no scoring rubric and no reference answers
- B Trust the judge score without sampling any outputs
- C Calibrate the judge against human labels and check it for position and self-preference biases
- D Score only answers the judge already liked

Answer: C

LLM judges can favor certain positions or their own style, so agreement with human labels should be measured and biases tested. Trusting the judge without checks is the common shortcut, and it produces misleading metrics.

8. Summarizing old conversation turns can lose details, so critical constraints are often pinned verbatim in the prompt.

- A True
- B False

Answer: A (True)

Summaries compress information, and a dropped constraint can change the agent's behavior without warning. Pinning the key rules keeps them intact across long sessions.

9. A research team wants several agents to debate and vote on an answer before it is returned. What is the main trade-off?

- A Higher accuracy on some tasks in exchange for more tokens, latency and cost per request
- B Lower cost and latency with no change in quality
- C Guaranteed correctness whenever at least two independent agents agree on the same final answer
- D No change in cost because the agents share one call

Answer: A

Debate and voting can reduce individual errors but multiply the calls, so cost and latency rise and agreement is not proof of correctness. The claim that agreement guarantees correctness is the trap in this question.

10. Which retrieval setup is most appropriate for an agent that must answer questions over private, frequently updated documents?

- A Fine-tuning the model once a year on a static snapshot
- B A retrieval pipeline with embeddings, a vector index, metadata filters and a re-ranking step over fresh chunks
- C Storing all documents in the system prompt
- D A random sample of ten documents per query

Answer: B

Retrieval keeps answers current because the index is refreshed, and metadata filters and re-ranking improve precision. A yearly fine-tune is stale between updates, which is why it is a tempting but poor fit.