

# NCP-AAI試験解説 & NCP-AAIコンポーネント



ご存じのように、私たちのNCP-AAI学習教材を利用するユーザーが多いです。NCP-AAI学習教材の質問が表示されない場合は、私たちにご連絡頂きます。私たちのスタッフは毎日多くのことを対処しなければなりません。どのユーザーも無視することはありません。私たちのNCP-AAI学習教材の市場はますます大きくなりました。そして、顧客のサポートがあると、私たちのNCP-AAI学習教材がより良くなると信じています。

## NVIDIA NCP-AAI 認定試験の出題範囲:

トピック	出題範囲
トピック 1	<ul style="list-style-type: none"><li>運用、監視、保守: 展開後のエージェントシステムの継続的な運用、健全性監視、および定期保守について説明します。</li></ul>
トピック 2	<ul style="list-style-type: none"><li>知識統合とデータ処理: エージェントが外部の知識源を統合し、多様なデータタイプを管理して、情報に基づいた意思決定を支援する方法について解説します。</li></ul>
トピック 3	<ul style="list-style-type: none"><li>エージェント開発: ツール、フレームワーク、APIを使用してエージェントを実際に構築、統合、強化することに重点を置きます。</li></ul>
トピック 4	<ul style="list-style-type: none"><li>認知、計画、記憶: インテリジェントエージェントの行動を左右する推論戦略、意思決定プロセス、および記憶管理技術を探求する。</li></ul>
トピック 5	<ul style="list-style-type: none"><li>エージェントアーキテクチャと設計: エージェントAIシステムの構造、および単一エージェント環境と複数エージェント環境におけるエージェントの推論、通信、相互作用について解説します。</li></ul>

>> NCP-AAI試験解説 <<

## NCP-AAIコンポーネント & NCP-AAI資格関連題

NCP-AAI認定試験は現在で本当に人気がある試験ですね。まだこの試験の認定資格を取っていないあなたも試験を受ける予定があるのでしょうか。確かに、これは困難な試験です。しかし、難しいといっても、高い点数を取って楽に試験に合格できないというわけではないです。では、まだ試験に合格するショートカットがわからないあなたは、受験のテクニックを知りたいですか。今教えてあげますよ。それはPassTestのNCP-AAI問題集を利用することです。

## NVIDIA Agentic AI 認定 NCP-AAI 試験問題 (Q39-Q44):

### 質問 # 39

When evaluating GPU utilization inefficiencies in deploying Llama Nemotron models across A100 and H100 clusters, which approaches help identify optimal resource allocation strategies? (Choose two.)

- A. Assess concurrent execution capabilities by employing multi-instance GPU partitioning for varying workload types.
- B. Profile resource utilization for each Nemotron variant and match models to appropriate GPU tiers.
- C. Allow Nemotron variants to profile actual workload characteristics and allocate resources based on observed demands.
- D. Allocate all agents to H100 GPUs, allowing resource profiles to automatically adjust for model size and computational requirements.

正解: A、B

解説:

The decisive point is failure isolation: the combination of Options B and D keeps the agent's decision path observable instead of burying behavior inside one prompt or one service. Together, B states "Profile resource utilization for each Nemotron variant and match models to appropriate GPU tiers."; D states "Assess concurrent execution capabilities by employing multi-instance GPU partitioning for varying workload types.", so the answer covers both sides of the requirement instead of solving only the model or only the infrastructure layer. Profiling each Nemotron variant and using MIG/concurrent execution where appropriate gives resource fit. Sending every workload to H100s wastes premium capacity. The runtime should therefore be built around matching model precision, batch windows, model instances, and GPU memory behavior to the latency service-level objective. The stack-level anchor is clear: TensorRT-LLM and NIM reduce inference overhead, but they still need serving-level tuning to avoid queue buildup under concurrency. The losing choices mostly optimize for short-term convenience; hardware upgrades alone do not fix poor batching, serial ensembles, guardrail overhead, or KV-cache pressure. The answer is therefore about engineered control planes, not simply model capability.

#### 質問 # 40

Your team has built an agent using LangChain and needs to implement guardrails for deployment in a production environment. Which approach represents the MOST effective integration of NVIDIA NeMo Guardrails?

- A. Rebuild the agent using only NeMo Guardrails, thereby reconstructing the LangChain implementation with enhanced safety controls and production-ready guardrail integration.
- B. Wrap the LangChain agent with NeMo Guardrails configuration while maintaining the existing workflow architecture and preserving current development investments.
- C. Configure input filtering to address safety requirements, integrating guardrail mechanisms focused on data validation and moderation within the current framework.
- D. Run the LangChain agent in parallel with NeMo Guardrails, allowing comparison of outputs between systems for comprehensive safety validation and performance optimization.

正解: B

解説:

Option B is the right call because it gives the platform team levers to tune behavior without rewriting the entire agent loop. The selected option specifically B states "Wrap the LangChain agent with NeMo Guardrails configuration while maintaining the existing workflow architecture and preserving current development investments.", which matches the operational requirement rather than a superficial wording match. Wrapping LangChain with NeMo Guardrails preserves the existing agent while adding policy enforcement. Rebuilding the workflow is unnecessary risk. The implementation detail that matters is multi-layer controls that combine semantic checks, topic control, content safety, jailbreak detection, and logged decisions. Within the NVIDIA stack, the guardrail layer should emit enough telemetry to show which policy triggered, which content was blocked or modified, and where the decision occurred. The losing choices mostly optimize for short-term convenience; unlogged guardrail decisions leave compliance teams unable to reconstruct what happened during an incident. That is the difference between an agent that works in a notebook and an agent that remains reliable in production.

#### 質問 # 41

Which two validation approaches are MOST critical for ensuring agent reliability in production deployments? (Choose two.)

- A. Random sampling of agent interactions for manual review
- B. Automated consistency checking across multiple agent runs
- C. Performance testing during development phases
- D. Structured output validation with Pydantic schemas
- E. User satisfaction surveys as the primary quality metric

正解: B、D

解説:

Together, C states "Structured output validation with Pydantic schemas"; E states "Automated consistency checking across multiple agent runs", so the answer covers both sides of the requirement instead of solving only the model or only the infrastructure layer. Pydantic-style structured validation catches malformed outputs; consistency checks detect nondeterministic behavior across runs. Surveys are secondary quality signals. the combination of Options C and E wins because it optimizes the system boundary around the risky component rather than hoping the base model behaves consistently. The NVIDIA implementation angle is not cosmetic here: NVIDIA evaluation tooling emphasizes whole-agent behavior, including tool selection order, final outcome quality, throughput, latency, and traceability. That matters because closed-loop evaluation where benchmark results, user feedback, and parameter changes are versioned together. That is why the other options are traps: looking only at speed can reward broken behavior, while looking only at accuracy can ignore cost and reliability failures. The result is a system that can be benchmarked, traced, and revised without destabilizing the whole agent fabric.

#### 質問 # 42

What is RAG Fusion primarily designed to achieve?

- A. Automatically translating and integrating all retrieved chunks into a single language.
- **B. Blending information from multiple retrieved chunks into a single response generated by the LLM.**
- C. Minimizing the need for retrieval, allowing the LLM to generate responses directly from its internal knowledge.
- D. Creating a separate, dedicated database for storing all the retrieved chunks.

正解: B

解説:

RAG Fusion improves generation by blending evidence from multiple retrieved chunks. It is about combining retrieved context, not eliminating retrieval. In a GPU-backed agent deployment, Option C maps closest to how the NVIDIA stack expects orchestration, inference, and control policies to be separated. The selected option specifically C states "Blending information from multiple retrieved chunks into a single response generated by the LLM.", which matches the operational requirement rather than a superficial wording match.

The correct implementation surface is retriever isolation, vector index quality, reranking, freshness-aware ingestion, query expansion, and retrieval guardrails. This lines up with NVIDIA guidance because NeMo Guardrails can add retrieval rails around RAG context, while the serving layer remains independent from the vector database. The distractors fail because keyword-only retrieval misses semantic matches, while unfiltered concatenation can pollute the answer with weak evidence. This choice gives engineering teams the knobs they need for continuous tuning after deployment. The retrieval layer should be independently measured for recall, relevance, freshness, and latency before blaming the generator.

#### 質問 # 43

When evaluating coordination failures in a multi-agent system managing distributed manufacturing workflows, which analysis approach best identifies state management and planning synchronization issues?

- **A. Deploy distributed state tracing across agents, analyze transition timing, study communication overhead, and verify synchronization accuracy.**
- B. Monitor agent outputs individually to confirm local correctness and examine results of specific workflow steps.
- C. Track workflow throughput and task completions to measure performance trends and highlight workflow outcomes.
- D. Assess synchronization methods during design reviews and use simulations to evaluate coordination across representative workflow scenarios.

正解: A

解説:

The rejected options are weaker because single-loop agents and isolated workers collapse planning, memory, and validation into one failure domain, which is brittle under real-time enterprise load. Coordination failures are temporal failures. You need transition timing, state visibility, and message-path analysis, not just local agent output review. Option B wins because it optimizes the system boundary around the risky component rather than hoping the base model behaves consistently. The selected option specifically B states "Deploy distributed state tracing across agents, analyze transition timing, study communication overhead, and verify synchronization accuracy.", which matches the operational requirement rather than a superficial wording match. The NVIDIA implementation angle is not cosmetic here: specialized agents can be served, evaluated, and replaced independently when their role or model changes. That matters because clear boundaries between planning, execution, validation, and escalation rather than one LLM attempting every responsibility. The result is a system that can be benchmarked, traced, and revised without destabilizing the whole agent fabric.

