# NVIDIA-Certified Professional: Agentic AI Exam Study Guide

# NVIDIA-Certified Professional: Agentic AI Exam Study Guide

Contents

This study guide provides an overview of each topic covered on the NVIDIA-Certified Professional: Agentic AI certification exam, recommended training, and suggested reading to prepare for the exam.

## Job Description

The agentic AI professional is an intermediate practitioner adept at designing, evaluating, and deploying autonomous AI systems. Responsibilities include constructing resilient, secure, and trustworthy agentic solutions such as customer/employee assistants, automated meeting companions, and productivity tools (e.g., content generation, analytics). The candidate must master agentic AI architecture fundamentals—including reactive, deliberative, and hybrid systems—and excel in goal-oriented reasoning, chain-of-thought prompt engineering, and tool orchestration with robust error handling.

Expertise includes memory management (short- and long-term context), evaluation frameworks, deployment of containerized workflows, monitoring/logging, ethical safeguards (e.g., bias detection, privacy preservation, guardrails), troubleshooting hallucinations, tracing failures, and managing stateful orchestration for complex, multi-turn tasks. They combine generative AI and reasoning systems knowledge with MLOps workflow deployment and GPU-optimized operations to ensure scalable, ethical implementation.

## Key Responsibilities

1. End-to-end agent development exposure
2. Model and framework selection and integration
3. Agent structure and tool creation
4. Orchestration of agent workflows
5. Assessment, evaluation, and iterative improvement

## Recommended Qualifications and Experience

1. 2-3 years in AI and machine learning roles
2. Experience with production-level agentic AI projects (e.g., chatbots, workflow automation)

# Certification Topics and References

## Agent Architecture and Design: Exam Weight 15%

Foundational structuring and design of agentic AI systems, focusing on how agents interact, reason, and communicate within their environments.

1.1 Design user interfaces for intuitive human-agent interaction.

1.2 Implement reasoning and action frameworks (e.g., ReAct).

1.3 Configure agent-to-agent communication protocols for collaboration.

1.4 Manage short-term and long-term memory for context retention.

1.5 Orchestrate multi-agent workflows and coordination.

1.6 Apply logic trees, prompt chains, and stateful orchestration for multi-step reasoning.

1.7 Integrate knowledge graphs to enable relational reasoning.

1.8 Ensure adaptability and scalability of the agent's architecture.

### Recommended NVIDIA Course and Suggested Readings

### NVIDIA Course

- Building Agentic AI Applications with LLMs

### Suggested Readings

- Agentic AI in the Factory
- Building Autonomous AI with NVIDIA Agentic NeMo
- Three Building Blocks for Creating AI Virtual Assistants for Customer Service with an NVIDIA NIM Agent Blueprint
- Agentic AI: Towards Autonomous Artificial Intelligence Agents
- Catch Me If You Can: A Multi-Agent Framework for Financial Fraud Detection
- What Are Multi-Agent Systems?

## **Agent Development:** Exam Weight 15%

Practical building, integration, and enhancement of agents.

2.1 Engineer prompts and dynamic prompt chains for reliable performance.

2.2 Integrate generative and multimodal models (text, vision, audio).

2.3 Build and connect custom tools, APIs, and functions for external system interaction.

2.4 Implement error handling (retry logic, graceful failure recovery).

2.5 Develop dynamic conversation flows with real-time streaming and feedback mechanisms.

2.6 Evaluate and refine agent decision-making strategies.

### Recommended NVIDIA Course and Suggested Readings

### NVIDIA Courses

- Building RAG Agents With LLMs
- Building Agentic AI Applications with LLMs

### Suggested Readings

- Optimization—NVIDIA Triton™ Inference Server
- NVIDIA Agent Intelligence Toolkit Overview—NVIDIA Agent Intelligence Toolkit (1.1.0)
- An Introduction to Large Language Models: Prompt Engineering and P-Tuning | NVIDIA Technical Blog
- Building Multimodal AI RAG With LlamaIndex, NVIDIA NIM™, and Milvus
- Design Considerations of Advanced Agentic AI for Real-World Applications
- Transient Fault Handling—Azure Architecture Center | Microsoft Learn
- Circuit Breaker Pattern—Azure Architecture Center | Microsoft Learn
- Retry Pattern—Azure Architecture Center | Microsoft Learn

## **Evaluation and Tuning:** Exam Weight 13%

Measuring, comparing, and optimizing agent performance.

3.1 Implement evaluation pipelines and task benchmarks to measure performance.

3.2 Compare agent performance across tasks and datasets.

3.3 Collect and integrate structured user feedback for iterative improvements.

3.4 Tune model parameters (e.g., accuracy, latency-efficiency trade-offs).

3.5 Analyze evaluation results to guide targeted optimization.

### Recommended NVIDIA Course and Suggested Readings

### NVIDIA Courses

- Building Agentic AI Applications With LLMs
- Evaluating RAG and Semantic Search Systems

### Suggested Readings

- Powering the Next Generation of AI Agents
- NVIDIA Agent Intelligence Toolkit Overview
- NVIDIA Agent Intelligence Toolkit Tutorials
- NVIDIA Agent Intelligence Toolkit FAQ
- Launching the NVIDIA Agent Intelligence Toolkit API Server and User Interface—NVIDIA Agent Intelligence Toolkit (1.1)
- NVIDIA NeMo™ Agent Toolkit | GitHub
- Agentic AI: The Next Big Thing in Artificial Intelligence
- Agentic AI: The Top 5 Challenges and How to Overcome Them
- AI Agents for Beginners—Production Patterns | Microsoft
- Navigating the Challenges: 5 Common Pitfalls in Agentic AI Adoption

## **Deployment and Scaling:** Exam Weight 5%

Operationalizing and scaling agentic systems.

4.1 Deploy and orchestrate multi-agent systems at production scale.

4.2 Apply MLOps practices for continuous integration and continuous delivery (CI/CD) workflows, monitoring, and governance.

4.3 Profile performance and reliability under distributed system loads.

4.4 Scale deployments using containerization (Docker, Kubernetes) with load balancing.

4.5 Optimize deployment costs while ensuring high availability.

### Recommended NVIDIA Courses and Suggested Readings

### NVIDIA Courses

> Deploying RAG Pipelines for Production at Scale
> Building Agentic AI Applications With LLMs
> Building RAG Agents With LLMs

### Suggested Readings

> Agentic AI in the Factory | NVIDIA Whitepaper
> NVIDIA TensorRT™-LLM | GitHub
> Measure and Improve AI Workload Performance With NVIDIA DGX™ Cloud Benchmarking
> Kubernetes Glossary | NVIDIA
> NVIDIA Nsight™ Systems
> Kube Prometheus for GPU Telemetry | NVIDIA Docs
> Scaling LLMs With NVIDIA Triton and TensorRT-LLM Using Kubernetes
> TensorRT-LLM Performance Analysis Documentation

## **Cognition, Planning, and Memory:** Exam Weight 10%

Core cognitive processes underlying intelligent agent behavior, including reasoning strategies, decision-making, and memory management.

5.1 Implement memory mechanisms for short- and long-term context retention.

5.2 Apply reasoning frameworks (chain-of-thought, task decomposition).

5.3 Engineer planning strategies for sequential and multi-step decision-making.

5.4 Manage stateful orchestration to coordinate complex tasks and knowledge retention.

5.5 Adapt reasoning strategies based on prior experiences and feedback.

### Recommended NVIDIA Course and Suggested Readings

### NVIDIA Courses

- Building Agentic AI Applications With LLMs
- Building RAG Agents with LLMs

### Suggested Readings

- NVIDIA NeMo
- Large Language Models Are in Context Learners | arXiv:2310.10501
- NeMo RL Documentation
- Jamba 1.5 LLMs Leverage Hybrid Architecture
- Understanding the Planning of LLM Agents: A Survey | HTML version
- AI Agent Memory | IBM
- MCP Agent Memory Types, Management, Implementation
- Understanding the Planning of LLM Agents: A Survey | arXiv:2402.02716

## **Knowledge Integration, and Data Handling:** Exam Weight 10%

Integration of external knowledge and the management of diverse data types.

6.1 Implement retrieval pipelines (RAG, embedded search, hybrid approaches).

6.2 Configure and optimize vector databases for fast retrieval.

6.3 Build extract, transform, and load (ETL) pipelines to integrate enterprise or client data sources.

6.4 Conduct data quality checks, augmentation, and preprocessing.

6.5 Enable real-time access and reasoning over structured and unstructured knowledge.

### Recommended NVIDIA Course and Suggested Readings

### NVIDIA Courses

- Building RAG Agents With LLMs
- Adding New Knowledge to LLMs

### Suggested Readings

- How to Make Your LLM More Accurate with RAG and Fine-Tuning | Towards Data Science

## **NVIDIA Platform Implementation:** Exam Weight 7%

| Leveraging NVIDIA's AI hardware and software platforms for agentic AI systems. |
|---|
| 7.1 Integrate NVIDIA NeMo Guardrails for compliance and safety enforcement. |
| 7.2 Deploy NVIDIA NIM microservices for high-performance inference. |
| 7.3 Optimize workflows with the NVIDIA NeMo Agent Toolkit. |
| 7.4 Leverage NVIDIA TensorRT-LLM and Triton Inference Server for latency reduction. |
| 7.5 Manage and optimize multimodal input pipelines on NVIDIA hardware. |

### Recommended NVIDIA Course and Suggested Readings

### NVIDIA Course

- Building RAG Agents With LLMs

### Suggested Readings

- Best Practices—NVIDIA TensorRT Documentation
- Batchers—NVIDIA Triton Inference Server
- Triton Inference Server Backend | NVIDIA Documentation
- NeMo Guardrails | NVIDIA Developer
- NVIDIA NeMo Guardrails | GitHub
- Performance Tuning Guide—NVIDIA NeMo Framework User Guide
- Best Practices—NVIDIA NeMo Framework User Guide
- Optimization—NVIDIA Triton Inference Server
- NVIDIA NeMo Agent Toolkit
- NVIDIA Agent Intelligence Toolkit
- NVIDIA AIQ Toolkit
- Mastering LLM Techniques: Inference Optimization
- Deploy Inference Workloads With NVIDIA NIM
- How to Use the NVIDIA Llama Nemotron API for Advanced AI Agents
- How to Deploy Llama-3.1-Nemotron-70B-Instruct on a Virtual Machine in the Cloud
- AI Agents Blueprint: Designing Foundation Models and Agents for the Next Wave of AI
- Improve AI Code Generation Using NVIDIA Agent Intelligence Toolkit
- NVIDIA NeMo: A Scalable Generative AI Framework | GitHub

## **Run, Monitor, and Maintain:** Exam Weight 7%

Ongoing operation, monitoring, and maintenance of agentic systems post-deployment.

8.1 Define monitoring dashboards and reliability metrics.

8.2 Track logs, errors, and anomalies for root cause diagnosis.

8.3 Continuously benchmark deployed agents against prior versions.

8.4 Implement automated tuning, retraining, and versioning in production.

8.5 Ensure continuous uptime, transparency, and trust in live deployments.

### Recommended NVIDIA Course and Suggested Readings

### NVIDIA Course

- Deploying RAG Pipelines in Production at Scale

### Suggested Readings

- What Is AI Agent Evaluation?
- Log, Trace, and Monitor
- Time-Weighted Retriever
- Troubleshooting
- LangChain Tracing Concepts
- LangChain Structured Outputs Concepts
- Smith LangChain Model Evaluation: Rate Limiting
- A Guide to Monitoring Machine Learning Models in Production
- Monitoring Machine Learning Models in Production: How to Track Data Quality and Integrity

## Safety, Ethics, and Compliance: Exam Weight 5%

| Principles and practices that ensure agentic AI systems operate responsibly, uphold ethical standards, and comply with legal and regulatory frameworks. |
|---|
| 9.1 Design and enforce system security and audit trails. |
| 9.2 Integrate compliance guardrails (privacy, enterprise policy). |
| 9.3 Mitigate bias and toxicity in outputs. |
| 9.4 Deploy layered safety frameworks (filters, escalation protocols). |
| 9.5 Ensure compliance with licensing and regulatory standards. |

### Recommended NVIDIA Course and Suggested Readings

### NVIDIA Course

> Building RAG Agents with LLMs

### Suggested Readings

> Building Safer LLM Apps With LangChain Templates and NVIDIA NeMo Guardrails
> NVIDIA NeMo Guardrails
> Artificial Intelligence and Machine Learning in Software as a Medical Device
> Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence | Artificial Intelligence Act
> Ethically Aligned Design: A Vision for Prioritizing Human Well-Being With Autonomous and Intelligent Systems
> NVIDIANeMo Guardrails | GitHub
> Securing Generative AI Deployments With NVIDIA NIM and NVIDIA NeMo Guardrails
> Metrics for Agentic AI
> AI for Regulatory Compliance
> Responsible AI Revisited

## **Human-AI Interaction and Oversight:** Exam Weight 5%

The design and implementation of systems that facilitate effective human oversight and interaction with agents.

10.1 Build intuitive UIs with user-in-the-loop interaction.

10.2 Design structured feedback loops that guide iterative agent improvements.

10.3 Implement transparency mechanisms (explainable reasoning, decision traceability).

10.4 Enable human oversight and intervention for accountability and trust.

### Recommended NVIDIA Course and Suggested Readings

#### NVIDIA Course

> Building Agentic AI Applications with LLMs

#### Suggested Readings

> NVIDIA Agent Intelligence Toolkit

> NVIDIA Data Flywheel Glossary

> AI Agents With Human-in-the-Loop | Medium

> Human-in-the-Loop AI | HolisticAI

> Human-in-the-Loop Agentic AI Systems | OneReach.ai

> Aporia: AI Guardrails

> Improve AI Code Generation Using NVIDIA Agent Intelligence Toolkit

> Chain-of-Thought (CoT) Prompting | Codecademy

## Questions?

Contact us here.

NVIDIA