



NVIDIA-Certified Professional: Agentic AI 认证考试学习指南



NVIDIA-Certified Professional: Agentic AI 认证考试学习指南

目录

智能体架构与设计：	2
考试权重 15%	
智能体开发：	3
考试权重 15%	
评估和调优：	4
考试权重 13%	
部署和扩展：	5
考试权重 5%	
认知、规划和记忆：	6
考试权重 10%	
知识集成和数据处理：	7
考试权重 10%	
NVIDIA 平台部署：	8
考试权重 7%	
运行、监测和维护：	9
考试权重 7%	
安全、道德与合规性：	10
考试权重 5%	
人机交互和监督：	11
考试权重 5%	

针对 NVIDIA-Certified Professional: Agentic AI 认证考试，本学习指南包含认证所涵盖的各个技术主题的介绍，并推荐了相关培训课程和阅读资料。查看所有 NVIDIA 认证，请单击[此处](#)。

考试适用人群

适用于开发人员的 Professional 级别认证，验证设计、评估和部署自主 AI 系统的能力。能够构建具有弹性、安全且值得信赖的智能代理解决方案，例如客户 / 员工助手、自动化会议助手，以及生产力工具 (如内容生成、数据分析)。您需掌握代理式 AI 架构的基础知识 —— 包括反应式、推理式以及混合系统，并精通目标导向推理、链式思维提示工程，以及高可靠性工具编排与错误处理的能力。

您需要具备专业知识，包括内存管理 (短期和长期上下文)、评估框架设计、容器化工作流部署、监测和日志记录、伦理合规防护 (例如偏差检测、隐私保护和防护措施)、幻觉排查、故障追踪以及复杂多轮任务的状态编排管理。具备结合生成式 AI 与推理系统的专业知识，配合 MLOps 工作流部署和 GPU 加速优化，确保实现过程具备可扩展性与伦理合规性。

相关工作职责示例

1. 端到端智能体开发实践经验
2. 模型和框架的选择与集成
3. 智能体架构和工具创建
4. 智能体工作流编排
5. 评估、评价与迭代改进

建议具备的知识和经验

1. 2 — 3 年从事 AI 或机器学习领域的实践经验
2. 生产级代理式 AI 项目实施经验 (例如聊天机器人、工作流自动化)

认证主题和参考资料

智能体架构和设计：考试权重 15%

代理式 AI 系统的基础结构与设计，重点关注智能体之间的交互、推理及通信方式。	
1.1	能够设计用户界面，实现直观的人机智能体交互。
1.2	实施推理和行动框架 (例如 ReAct)。
1.3	配置用于协作的智能体通信协议。
1.4	管理短期和长期记忆，实现上下文保持。
1.5	编排多智能体工作流与协调。
1.6	应用逻辑树、提示链和有状态编排进行多步骤推理。
1.7	集成知识图谱，实现关系推理。
1.8	确保智能体架构的适应性和可扩展性。

NVIDIA 培训和阅读内容推荐

NVIDIA 培训推荐 (可选)

> 课程：讲师指导的培训班《使用大语言模型构建代理式 AI》([查看课程大纲](#))

阅读内容推荐

- > [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?](#)

智能体开发：考试权重 15%

构建、集成和增强智能体的实践。

- 2.1 设计提示词并构建动态提示链，确保稳定可靠的性能。
- 2.2 集成生成式和多模态模型 (文本、视觉和音频)。
- 2.3 构建并连接自定义工具、API 和函数，实现与外部系统的交互。
- 2.4 实现错误处理 (包括重试逻辑和平滑的故障恢复)。
- 2.5 通过实时流媒体和反馈机制开发动态对话流。
- 2.6 评估并优化代理的决策策略。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

- > 课程：在线自主培训《大语言模型 —— RAG 智能体入门》([查看课程](#)) 或讲师指导的培训班《构建大语言模型 RAG 智能体》([查看课程大纲](#))
- > 课程：讲师指导的培训班《使用大语言模型构建代理式 AI》([查看课程大纲](#))

阅读内容推荐

- > [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](#)

评估和调优：考试权重 13%

测量、比较和优化智能体性能。

- 3.1 实施评估工作流和任务基准测试来衡量性能。
- 3.2 比较智能体在不同任务和数据集上的性能。
- 3.3 收集并整合结构化用户反馈，实现迭代改进。
- 3.4 调整模型参数 (例如模型精度、延迟与效率权衡)。
- 3.5 分析评估结果，指导定向优化。
- 3.6 在 DPU-Arm 上部署 DOCA 服务。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

- > 课程：讲师指导的培训班《使用大语言模型构建代理式 AI》([查看课程大纲](#))
- > 课程：在线自主培训《评估 RAG 和语义检索系统》([查看课程](#))

阅读内容推荐

- > [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](#)

部署和扩展：考试权重 5%

实现智能体系统的落地与规模化。

- 4.1 在生产环境中部署并编排多智能体系统。
- 4.2 应用 MLOps 实践，构建持续集成与持续交付 (CI/CD) 工作流，并实现监控与治理。
- 4.3 在分布式系统负载下分析性能与可靠性。
- 4.4 通过容器化 (Docker、Kubernetes) 和负载均衡技术扩展进行部署。
- 4.5 优化部署成本，同时确保高可用性。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

- > 课程：在线自主培训《规模化部署 RAG 工作流》([查看课程](#)) 或讲师指导的培训班《使用 NVIDIA NIM 大规模部署 RAG 工作流》([查看课程大纲](#))
- > 课程：讲师指导的培训班《使用大语言模型构建代理式 AI》([查看课程大纲](#))
- > 课程：在线自主培训《大语言模型 —— RAG 智能体入门》([查看课程](#)) 或讲师指导的培训班《构建大语言模型 RAG 智能体》([查看课程大纲](#))

阅读内容推荐

- > [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](#)

认知、规划和记忆：考试权重 10%

了解智能体的核心认知过程，包括推理策略、决策机制和记忆管理。

- 5.1 实现短期和长期上下文保留的记忆机制。
- 5.2 应用推理框架 (思维链和任务分解)。
- 5.3 设计规划策略，以支持顺序性和多步骤的决策制定。
- 5.4 管理有状态的编排，协调复杂任务并实现知识保留。
- 5.5 根据历史经验和反馈调整推理策略。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

- > 课程：讲师指导的培训班《使用大语言模型构建代理式 AI》([查看课程大纲](#))
- > 课程：在线自主培训《大语言模型 —— RAG 智能体入门》([查看课程](#)) 或讲师指导的培训班《构建大语言模型 RAG 智能体》([查看课程大纲](#))

阅读内容推荐

- > [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](#)

知识整合和数据处理：考试权重 10%

外部知识的集成与多元化数据类型的管理。

- 6.1 实现检索管道 (RAG、嵌入式搜索、混合方法)。
- 6.2 配置和优化向量数据库，实现快速检索。
- 6.3 构建提取、转换和加载 (ETL) 工作流来集成企业或客户端数据源。
- 6.4 执行数据质量检查、增强和预处理。
- 6.5 实现对结构化与非结构化知识的实时访问与推理。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

- > 课程：在线自主培训《大语言模型 —— RAG 智能体入门》([查看课程](#)) 或讲师指导的培训班《构建大语言模型 RAG 智能体》([查看课程大纲](#))
- > 课程：在线自主培训《大语言模型压缩：剪枝、蒸馏与量化》([查看课程](#)) 或讲师指导的培训班《为大语言模型添加新知识》([查看课程大纲](#))

阅读内容推荐

- > [How to Make Your LLM More Accurate with RAG and Fine-Tuning | Towards Data Science](#)

NVIDIA 平台部署：考试权重 7%

利用 NVIDIA AI 硬件和软件平台构建代理式 AI 系统。

- 7.1 集成 NVIDIA NeMo Guardrails，实现合规性与安全防护。
- 7.2 部署 NVIDIA NIM 推理微服务，实现高性能推理。
- 7.3 使用 NVIDIA NeMo Agent Toolkit 优化工作流。
- 7.4 利用 NVIDIA TensorRT-LLM 和 Triton 推理服务器降低延迟。
- 7.5 在 NVIDIA 硬件上管理和优化多模态输入工作流。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

- > 课程：在线自主培训《大语言模型 —— RAG 智能体入门》([查看课程](#)) 或讲师指导的培训班《构建大语言模型 RAG 智能体》([查看课程大纲](#))

阅读内容推荐

- > [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](#)

运行、监测和维护：考试权重 7%

部署代理系统后，对其进行持续运营、监控和维护。

- 8.1 定义监控仪表板和可靠性指标。
- 8.2 跟踪日志、错误和异常，进行根因分析。
- 8.3 持续对部署的代理与历史版本进行基准测试。
- 8.4 在生产环境中实现自动调优、再训练和版本控制。
- 8.5 确保实时部署中的持续可用性、透明性与可信性。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

- > 课程：在线自主培训《规模化部署 RAG 工作流》([查看课程](#)) 或讲师指导的培训班《使用 NVIDIA NIM 大规模部署 RAG 工作流》([查看课程大纲](#))

阅读内容推荐

- > [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](#)

安全、道德与合规：考试权重 5%

确保代理式 AI 系统可靠运行、坚持道德标准并遵守法律和监管框架的原则与实践。

- 9.1 设计并执行系统安全和审计跟踪。
- 9.2 集成合规防护机制 (隐私保护、企业政策)。
- 9.3 降低输出中的偏见与有害内容。
- 9.4 部署分层安全框架 (过滤机制、升级处理协议)。
- 9.5 确保遵守许可和监管标准。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

- > 课程：在线自主培训《大语言模型 —— RAG 智能体入门》([查看课程](#)) 或讲师指导的培训班《构建大语言模型 RAG 智能体》([查看课程大纲](#))

阅读内容推荐

- > [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](#)

人机交互和监督：考试权重 5%

设计并实现能够促进人类对智能体进行有效监督与交互的系统。

10.1 构建直观的用户界面，实现用户参与的交互流程。

10.2 设计结构化反馈循环来指导智能体迭代改进。

10.3 实现透明机制 (可解释推理、决策可追溯)。

10.4 启用人工监督与干预，确保责任与信任。

NVIDIA 培训推荐课程及更多资源

NVIDIA 推荐培训 (可选)

> 课程：讲师指导的培训班《使用大语言模型构建代理式 AI》([查看课程大纲](#))

阅读内容推荐

- > [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](#)

更多咨询

请邮件至 dlichina@nvidia.com