



NVIDIA-Certified Associate: Generative AI LLMs 认证考试 学习指南



NVIDIA-Certified Associate: Generative AI LLMs 认证考试 学习指南

目录

核心机器学习和 AI 知识： 考试权重 30%	2
数据分析： 考试权重 14%	3
实验： 考试权重 22%	4
软件开发： 考试权重 24%	5
值得信赖的人工智能： 考试权重 10%	6

针对 NVIDIA-Certified Associate: Generative AI and LLMs (NCA-GENL) 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](#)。

考试适用人员

生成式 AI 大语言模型 (LLM) 助理开发人员参与最先进的生成式 AI LLM 系统的开发、编程和质量保证的相关工作。通过与技术娴熟的 AI 专业人员合作，负责开发数据集、选择要训练的模型、训练模型、测试模型和调试流程等事项。助理开发人员需要了解如何开发和部署应用模型，设计和构建高质量的软件，使用多种语言和平台进行编程，以及维护系统更新。

相关工作职责示例

1. 与 AI 开发团队合作，设计、编码、测试、调试和记录编程应用程序
2. 执行系统分析，以确保软件和系统符合规格要求
3. 帮助将新的 AI 语言模型集成到现有系统中，或根据需要创建新的语言模型
4. 协助评估和解决应用程序和系统性能问题
5. 保持了解新的 AI 模型以及与语言模型的发展
6. 参与制作技术文档和手册
7. 在高级团队成员的指导下进行软件编程和文档开发
8. 执行提示工程
9. 协助模型选择过程
10. 定义、策划、标记和注释 LLM 数据集
11. 执行实验，例如 A/B 测试、评估提示、评估模型和生成 POC

建议具备的知识和经验

1. 会用 Python、C 和 AI 框架 (如 PyTorch、TensorFlow 等)
2. 深刻理解神经网络和深度学习模型
3. 理工科背景，如计算机科学、软件工程、AI 等

认证主题和参考资料

核心机器学习和 AI 知识：考试权重 30%

了解算法、惯例和技术，使计算机能够从数据中学习并根据数据做出预测或决策。

- 1.1 在高级团队成员的监督下协助部署和评估模型的可扩展性、性能和可靠性
- 1.2 了解使用数据挖掘、数据可视化和类似技术从大型数据集中提取见解的过程
- 1.3 构建 LLM 用例，例如检索增强生成 (RAG)、聊天机器人和摘要器
- 1.4 为 RAG 策划和嵌入内容数据集
- 1.5 熟悉机器学习的基础知识 (例如特征工程、模型比较、交叉验证)
- 1.6 熟悉 Python 自然语言包 (spaCy、NumPy、矢量数据库等) 的功能
- 1.7 阅读研究论文 (文章、会议论文等) 以确定新兴的 LLM 趋势和技术
- 1.8 选择并使用模型创建文本嵌入
- 1.9 使用提示工程原理创建提示以实现所需结果
- 1.10 使用 Python 包 (spaCy、NumPy、Keras 等) 实现特定的传统机器学习分析

培训推荐 (可选)

课程：在线自主培训《深度学习新手入门》([查看课程](#))，讲师指导的培训班《深度学习基础——理论与实践入门》([查看课程](#))

- > 学习训练深度学习模型所需的基础技能和工具
- > 了解常见的深度学习数据类型和模型架构
- > 通过模型间的迁移学习，用较少的数据和计算量获得高效的结果

课程：在线自主培训《基于 Transformer 的自然语言处理入门》([查看课程](#))

- > 了解 transformers 如何构建现代大语言模型

课程：讲师指导的培训班《构建基于 Transformer 的自然语言处理应用》([查看课程](#))

- > 了解 transformer 如何构建自然语言处理 (NLP) 应用的现代大语言模型
- > 使用自监督机制优化 BERT、Megatron 和其他变体中的 Transformer 架构，以取得更好的 NLP 结果

课程：在线自主培训《加速端到端的数据科学工作流》([查看课程](#))

- > 利用各种机器学习算法，包括 XGBoost，解决不同的数据科学问题
- > 使用 NetworkX 和 cuGraph 执行强大的图形算法进行复杂网络的分析

课程：讲师指导的培训班《利用提示工程构建大语言模型 (LLM) 应用》([查看课程](#))

- > 理解如何应用迭代式提示工程最佳实践，以创建基于 LLM 的应用程序，用于各种与语言相关的任务

课程：讲师指导的培训班《构建基于大语言模型 (LLM) 的应用》([查看课程](#))

- > 使用编码器模型执行语义分析、嵌入、问答和零样本分类等任务
- > 使用有条件的解码器模型接收并生成有趣的数据格式、风格和模态

阅读内容推荐

- > [Attention Is All You Need](#)
- > [End-to-End AI for NVIDIA-Based PCs: Transitioning AI Models With ONNX](#), NVIDIA Technical Blog
- > [Generative AI—What Is It and How Does it Work?](#)

- > [Activation Function](#)
- > [Implementing Deep Learning Methods and Feature Engineering for Text Data](#)
- > [Autoregressive Model](#)
- > [What Are Foundation Models?](#), NVIDIA Blog
- > [LoRA: Low-Rank Adaptation of Large Language Models](#)
- > [Generative AI Research Spotlight: Demystifying Diffusion-Based Models](#), NVIDIA Technical Blog
- > [Training Hidden Units With Back Propagation](#)

数据分析：考试权重 14%

检查、清理、转换和建模数据，目的是发现有用的信息、得出结论和支持决策。

- 2.1 了解使用数据挖掘、数据可视化和类似技术从大型数据集中提取见解的过程
- 2.2 使用统计性能指标 (例如损失函数或解释方差比例) 比较模型
- 2.3 在高级团队成员的指导下进行数据分析
- 2.4 使用专门的软件创建图形、图表或其他可视化效果来传达数据分析的结果
- 2.5 识别关系和趋势或任何可能影响研究结果的因素

培训推荐 (可选)

课程：在线自主培训《深度学习新手入门》([查看课程](#))，讲师指导的培训班《深度学习基础——理论与实践入门》([查看课程](#))

- > 通过数据增强优化数据集，提高模型精准度

课程：在线自主培训《基于 Transformer 的自然语言处理入门》([查看课程](#))

- > 使用基于 transformer 的模型进行文本分类
- > 应用 LLM 构建命名实体识别 (NER) 项目
- > 利用 transformer 模型进行作者归属识别

课程：讲师指导的培训班《构建基于 Transformer 的自然语言处理应用》([查看课程](#))

- > 利用经过预训练的现代 NLP 模型来完成多个任务，如文本分类、NER 和问答系统

课程：讲师指导的培训班《利用提示工程构建大语言模型 (LLM) 应用》([查看课程](#))

- > 熟练使用 LangChain 来组织和编排 LLM 工作流

课程：讲师指导的培训班《构建基于大语言模型 (LLM) 的应用》([查看课程](#))

- > 探索使用 LangChain 和 LangGraph 来编排数据工作流和感知环境的智能体

课程：在线自主培训《加速端到端的数据科学工作流》([查看课程](#))

- > 使用 cuDF 加速 pandas、Polars 和 Dask 实现高效分析各种规模的数据集
- > 在大型数据集上执行多个分析任务，来避免英国人口的 (模拟的) 流行病爆发

阅读内容推荐

- > [RAPIDS](#)
- > [cuML 24.04.00 documentation](#)
- > [GPU Accelerated Data Science With RAPIDS](#)
- > [Data Exploration](#)
- > [Stemming and Lemmatizing With sklearn Vectorizers](#)

实验：考试权重 22%

研究如何执行、评估和解释实验，包括 AI 模型评估以及使用人类受试者进行标记或从人类反馈中进行强化学习 (RLHF)。

3.1 了解使用数据挖掘、数据可视化和类似技术从大型数据集中提取见解的过程

3.2 使用统计性能指标 (例如损失函数或解释方差比例) 比较模型

3.3 在高级团队成员的指导下进行数据分析

3.4 使用专门的软件创建图形、图表或其他可视化效果来传达数据分析的结果

3.5 确识别关系和趋势或任何可能影响研究结果的因素

培训推荐 (可选)

课程：在线自主培训《深度学习新手入门》([查看课程](#))，讲师指导的培训班《深度学习基础——理论与实践入门》([查看课程](#))

- > 利用先进的深度学习框架自信地运作自己的项目
- > 通过模型间的迁移学习，用较少的数据和计算量获得高效的结果

课程：在线自主培训《基于 Transformer 的自然语言处理入门》([查看课程](#))

- > 使用基于 Transformer 的模型完成各种自然语言处理任务
- > 测试并比较模型在问答任务上的性能

课程：讲师指导的培训班《构建基于 Transformer 的自然语言处理应用》([查看课程](#))

- > 利用经过预训练的现代 NLP 模型来完成多个任务，如文本分类、NER 和问答系统

课程：在线自主培训《加速端到端的数据科学工作流》([查看课程](#))

- > 使用 NetworkX 和 cuGraph 执行强大的图形算法进行复杂网络的分析

课程：讲师指导的培训班《构建基于大语言模型 (LLM) 的应用》([查看课程](#))

- > 查找、引入和使用 HuggingFace 模型库及其依赖的 Transformers API

课程：讲师指导的培训班《利用提示工程构建大语言模型 (LLM) 应用》([查看课程](#))

- > 熟练使用 LangChain 来组织和编排 LLM 工作流

阅读内容推荐

- > [How to Conduct A/B Testing in Machine Learning?](#)
- > [Inference Optimization](#)
- > [Zero-Shot Testing](#)
- > [Speech and Language Processing](#)
- > [Machine Translation methods](#)
- > [Hallucinations in Large Language Models](#)
- > [General Language Understanding Evaluation](#)
- > [Evaluating RAG Applications](#)
- > [Cross-Validation in Machine Learning](#)
- > [Benchmarking Elementary Language Tasks](#)
- > [Building Transformer-Based Natural Language Processing Applications](#)

软件开发：考试权重 24%

创建、维护和测试软件。

- 4.1 在高级团队成员的指导下协助部署和评估模型的可扩展性、性能和可靠性
- 4.2 构建 LLM 用例，例如 RAG、聊天机器人和摘要器
- 4.3 熟悉 Python 自然语言包 (spaCy、NumPy、矢量数据库等) 的功能
- 4.4 确定满足用户需求所需的系统数据、硬件或软件组件
- 4.5 监控数据收集、实验和其他软件流程的运行情况
- 4.6 使用 Python 包 (spaCy、NumPy、Keras 等) 实现特定的传统机器学习分析
- 4.7 在高级团队成员的指导下编写软件组件或脚本

培训推荐 (可选)

课程：在线自主培训《深度学习新手入门》([查看课程](#))，讲师指导的培训班《深度学习基础——理论与实践入门》([查看课程](#))

- > 了解常见的深度学习数据类型和模型架构
- > 利用先进的深度学习框架自信地运作自己的项目

课程：在线自主培训《基于 Transformer 的自然语言处理入门》([查看课程](#))

- > 使用基于 Transformer 的模型完成各种自然语言处理任务
- > 使用大语言模型开发文本分类任务、命名实体识别任务、作者归属和问答系统的解决方案

课程：讲师指导的培训班《构建基于 Transformer 的自然语言处理应用》([查看课程](#))

- > 管理推理挑战，以及为实时应用部署经过优化的模型

课程：在线自主培训《加速端到端的数据科学工作流》([查看课程](#))

- > 在 Triton Inference Server 上部署机器学习模型，实现最佳性能

课程：讲师指导的培训班《构建基于大语言模型 (LLM) 的应用》([查看课程](#))

- > 安全、有效和可扩展地处理自然数据任务的人工智能解决方案

课程：讲师指导的培训班《利用提示工程构建大语言模型 (LLM) 应用》([查看课程](#))

- > 编写应用代码，利用 LLM 执行生成任务，比如文档分析、聊天机器人等

阅读内容推荐

- > [TensorRT—Get Started](#), NVIDIA Blog
- > [Best Practices—NVIDIA NeMo](#)
- > [Mastering LLM Techniques: Customization](#), NVIDIA Technical Blog
- > [Achieving FP32 Accuracy for INT8 Inference Using Quantization-Aware Training With NVIDIA TensorRT](#)
- > [NCCL: Accelerated Multi-GPU Collective Communications](#)
- > [Technologies Behind Distributed Deep Learning: AllReduce](#), Preferred Networks Research & Development
- > [Visual Intuition on Ring—Allreduce for Distributed Deep Learning](#), by Edir Garcia Lazo, Towards Data Science
- > [Big Data? Datasets to the Rescue!](#), Hugging Face NLP Course
- > [Deep Learning Scaling Is Predictable, Empirically](#)
- > [BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding](#)

值得信赖的人工智能：考试权重 10%

创建和评估合乎道德、节能且可靠的人工智能系统，该系统能够解释和整合各种形式的数据，确保它们的设计和应用方式透明、公平且可验证。

5.1 描述可信赖人工智能的道德原则性

5.2 描述数据隐私与数据同意重要性之间的平衡

5.3 描述如何使用 NVIDIA 和其他技术来提高人工智能的可信度

5.4 描述如何最大限度地减少人工智能系统中的偏见

培训推荐 (可选)

课程：讲师指导的培训班《构建基于大语言模型 (LLM) 的应用》([查看课程](#))

- > 安全、有效和可扩展地处理自然数据任务的人工智能解决方案

阅读内容推荐

- > [Trustworthy AI for A Better World](#), NVIDIA
- > [What Is Trustworthy AI?](#), NVIDIA Blog
- > [What Is Retrieval-Augmented Generation aka RAG?](#), NVIDIA Blog

问题咨询，请发邮件至 dlichina@nvidia.com