NVIDIA

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

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

目录

针对 NVIDIA Certified Professional-Gen AI LLMs certification 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。

考试适用人群

生成式 AI 从业者，拥有大语言模型 (LLM) 的实践经验，并对模型开发、优化和大规模部署有充分理解。负责设计、训练和尖端 LLM，应用先进的分布式训练技术和优化策略来交付高性能 AI 解决方案。研究和生产团队协同合作，确保模型的稳健性、效率和可扩展性。

相关工作职责示例

模型开发和训练

- 针对研究和生产用例进行预训练和微调基础模型。
- 应用参数高效微调方法 (例如 Low-Rank Adaptation 或 LoRA) 和优化技术，如知识蒸馏、剪枝、量化和模型简化。
- 架构并实施分布式训练方案，涵盖数据并行、模型并行、张量并行及专家并行技术。

评估和优化

- 开发并应用系统化的评估方法，结合定量指标 (BLEU、ROUGE、perplexity 和 LLM-as-a-judge) 与定性评估 (人工评审、错误分析)。
- 分析并优化模型和 CUDA® 内核性能，诊断 GPU 性能瓶颈，并提高推理与训练的效率。
- 针对多 GPU、云端及本地环境的大规模部署，执行性能基准测试并进行故障排查，确保系统稳定高效运行。

部署和扩展

- 通过容器化 (Docker) 与编排 (Kubernetes) 将模型部署到生产环境中。
- 针对推理进行优化，实现低延迟、高吞吐量，并兼容边缘设备。
- 扩展与维护 Python 驱动的机器学习工作流，必要时通过 C++ 实施针对性的底层性能优化。

创新和研究

- 设计并执行实验，全面验证微调、评估和优化方法，确保统计结果的可靠性。
- 处理真实场景中的问题，如 CUDA 内存分配、内核效率优化和分布式训练的可扩展性。
- 持续跟进生成式 AI、Transformer 架构和 NVIDIA 最新技术趋势，确保技术领先。

建议具备的知识和经验

- 具备 2 — 3 年专注于 LLM 方向的 AI 或 ML 实践经验
- 深入掌握 Transformer 架构原理，包括自注意力、编码器 — 解码器结构和位置编码技术
- 熟悉检索增强生成 (RAG)、幻觉缓解和先进采样技术
- 在分布式并行与参数高效微调工作流方面拥有丰富的经验
- 熟练掌握模型评估指标、性能分析以及优化方法
- 精通 Python 编程；具备以 C++ 实现性能关键模块的能力者优先
- 具备 Docker 和 Kubernetes 可扩展部署的经验
- 熟悉 NVIDIA 生态系统 (NGC™ catalog、Base Command™ Platform、DGX™ systems、AI Enterprise suite) 则更具优势

认证主题和参考资料

LLM 模型架构：考试权重 6%

理解并掌握基础的 LLM 架构与机制，并灵活应用。	
1.1	分析编码器 — 解码器结构及其应用。
1.2	能够描述 Transformer 架构，包括其核心的自注意力 (Self-Attention) 机制。
1.3	能够开发代码，从编码器 — 解码器模型中提取嵌入向量 (embeddings)。
1.4	能够实现用于文本生成的高级采样技术。
1.5	理解基于解码器的语言模型中所使用的输出采样技术。
1.6	理解嵌入概念。

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《快速搭建大语言模型 (LLM) 应用》(查看课程) 或讲师指导的培训班《构建基于大语言模型 (LLM) 的应用》(查看课程大纲)

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

阅读内容推荐

> Mastering LLM Techniques: Training | NVIDIA Blog

> Attention Is All You Need | arXiv

提示工程：考试权重 13%

通过提示工程、链式思维 (CoT)、领域适应性、Zero/One/Few-shot 样本学习以及输出控制，使 LLM 适应新的领域、任务或数据分布。	
2.1	能够设计高效的提示词与模板，包括链式思维 (Chain-of-Thought) 方法，以及针对小规模数据集或专用领域的提示学习 (Prompt Learning) 策略。
2.2	采用 zero-shot、one-shot 和 few-shot 技术扩展模型适应性。
2.3	根据需要，使用因果语言建模训练基于解码器的 LLM。
2.4	设计专用的 LLM 封装模块，集成验证机制和受限解码策略，提升一致性，减少幻觉现象，并优化用户体验。

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《开发基于提示工程的大语言模型 (LLM) 应用》(查看课程) 或讲师指导的培训班《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

- Mastering LLM Techniques: Inference Optimization | NVIDIA Technical Blog
- An Easy Introduction to LLM Reasoning, AI Agents, and Test-Time Scaling | NVIDIA Technical Blog
- Train a Reasoning-Capable LLM in One Weekend With NVIDIA NeMo™ | NVIDIA Technical Blog
- Multi-Turn Conversational Chat Bot NVIDIA Generative AI Examples 0.5.0 Documentation

数据准备：考试权重 9%

通过数据清洗、筛选、分析与组织，以及 Tokenization 和词汇管理，完成用于预训练、微调或推理的数据准备工作。
3.1　清洗并整理数据 (处理缺失值、归一化、缩放)，并分析类别不平衡和特征分布。
3.2　组织数据集、确保格式正确并为建模准备数据。
3.3　选择并训练分词器，优化分词策略和词汇表大小 (如 BPE 和 WordPiece)，以适配任务需求和资源限制。

NVIDIA 培训推荐及更多资源

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

阅读内容推荐

- NVIDIA/GenerativeAIExamples | GitHub
- NVIDIA AI Workbench Example Projects
- Tokenizers NVIDIA NeMo Framework User Guide
- NVIDIA/NeMo | GitHub
- Train Models Using a Distributed Training Workload
- Speed Up Data Exploration With NVIDIA RAPIDS™ cuDF | NVIDIA Technical Blog
- Accelerating Time-Series Forecasting With RAPIDS cuML | NVIDIA Technical Blog
- Model Fine-Tuning NVIDIA NeMo Framework User Guide 24.07
- NeMo Curator | NVIDIA Developer
- Faster Causal Inference on Large Datasets With NVIDIA RAPIDS | NVIDIA Technical Blog
- Fine-Tuning NVIDIA NeMo Framework User Guide

模型优化：考试权重 17%

使用大语言模型的模型优化策略， 如剪枝、量化和知识蒸馏，来降低内存使用、加速推理和让模型兼容 GPU 加速，实现高效部署。	
4.1	应用剪枝、稀疏化以及权重 / 激活量化技术，以降低内存占用，并优化模型推理性能，实现硬件加速。
4.2	选择并实施针对硬件和任务 (例如 NVIDIA A100/H100 Tensor Core GPU、FP16、INT8) 定制的量化策略 (训练后、量化感知、激活量化) 并衡量任何模型精度的代价。
4.3	使用知识蒸馏技术，基于大型预训练模型构建更小更高效的模型。
4.4	进行系统化的超参数调优与分布式参数搜索，包括学习率调度和批量大小调整。
4.5	使用高级采样方法 (如束搜索 Beam Search、温度缩放 Temperature Scaling)，并开展系统化消融实验，评估模型优化的效果。
4.6	根据模型架构、任务需求和可用资源，选择并应用优化方法 (如 NVIDIA TensorRT™、滑动窗口 / 流式注意力、键值缓存)。
4.7	通过掩码语言建模 (MLM) 和 / 或下一句预测技术训练基于编码器的基础 LLM，并理解量化、蒸馏和模型剪枝等概念。

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《构建基于 Transformer 的自然语言处理应用》(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)
- 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《模型并行 —— 构建和部署大型神经网络》(查看课程大纲)

阅读内容推荐

- Best Practices for TensorRT Performance | NVIDIA Documentation
- Quantization NVIDIA NeMo User Guide
- DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper, and Lighter | arXiv
- What Is Knowledge Distillation? | IBM
- Sparsity in INT8: Training Workflow and Best Practices for NVIDIA TensorRT Acceleration | NVIDIA Technical Blog
- Quantization and Calibration NVIDIA TensorRT Documentation
- Post-Training Quantization vs. Quantization-Aware Training | Fiveable
- Quantization-Aware Training (QAT) vs. Post-Training Quantization (PTQ) | Better ML (Medium)
- The Illustrated Transformer | Jay Alammar
- Understanding Data Types in AI and HPC: INT8, FP8, FP16, BF16, BF32, FP32, TF32, FP64, and Hardware Accelerators | It's About AI
- Quantization: What You Should Understand if You Want to Run LLMs | Pavan Mantha, LinkedIn
- Capabilities NVIDIA TensorRT Documentation
- LoRA: Low-Rank Adaptation of Large Language Models | OpenReview
- GPTQ: Accurate Post-Training Quantization for Generative Pretrained Transformers | arXiv

微调：考试权重 13%

使用参数高效方法、人类反馈、对比学习以及稳健评估，将预训练的 LLM 定制化应用于下游任务或特定领域。	
5.1	通过监督微调或基于人类反馈的强化学习来实现模型与意图的一致性，包括直接偏好优化 (DPO) 或群体相对策略优化 (GRPO) 等方法。
5.2	针对嵌入应用对比损失，并使用参数高效技术 (LoRA、适配器、P-tuning)。
5.3	实施提前终止，防止过度拟合，并为各阶段选择性能指标。
5.4	减轻幻觉、评估微调影响并为 LLM 执行参数高效更新。

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

- Deploy Diverse AI Apps With Multi-LoRA Support on NVIDIA RTX™ AI PCs and Workstations | NVIDIA Blog
- Selecting Large Language Model Customization Techniques | NVIDIA Technical Blog
- LoRA: Low-Rank Adaptation of Large Language Models. | arXiv
- Parameter-Efficient Fine-Tuning for LLMs With NVIDIA NeMo | NVIDIA Technical Blog
- Prevent LLM Hallucinations With the Cleanlab Trustworthy Language Model in NVIDIA NeMo Guardrails | NVIDIA Technical Blog
- Chapter 7, Regularization for Deep Learning—Deep Learning: An MIT Press Book (Goodfellow, Bengio, Courville)
- LLM Evaluation Metrics: BLEU, ROUGE, and METEOR Explained | Medium

评估：考试权重 7%

通过定量与定性指标、框架设计、基准测试、错误分析以及可扩展评估，对 LLM 进行系统化评估。	
6.1	分析基准测试结果、进行真人参与和以评判身份进行评估的 LLM ，使用 BLEU、ROUGE、Perplexity 等关键指标评估模型质量。
6.2	诊断 LLM 故障模式并执行系统性的错误分析，识别常见的行为和输出模式。
6.3	使用标准化评估指标对各种平台 (如本地 DGX 系统、云端 GPU) 上的 LLM 部署进行基准测试与对比分析。
6.4	设计并实施综合评估框架，整合上述所有实践，确保模型评估的稳健性与可扩展性，能够提供系统化的验证标准。

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《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

- NVIDIA Metrics | Ragas
- Retrieval-Augmented Generation (RAG) Pipeline | NVIDIA Docs
- Evaluating Medical RAG With NVIDIA AI Endpoints and Ragas | NVIDIA Technical Blog
- Integrations | Ragas
- NVIDIA-AI-Blueprints/RAG | GitHub
- NeMo Evaluator | NVIDIA Developer
- Introduction—NVIDIA Autonomous Vehicles Safety Report
- Performance Analysis Tools | NVIDIA Developer
- Deployment Best Practices—NVIDIA RTX vWS: Sizing and GPU Selection Guide for Virtualized Workloads
- Troubleshoot NVIDIA NIM for LLMs
- How the DGX H100 Accelerates AI Workloads | CUDO Compute
- Masked Language Model Scoring | arXiv
- Perplexity of Fixed-Length Models | Hugging Face
- The Importance of Starting With Error Analysis in LLM Applications | Shekhar Gulati

GPU 加速和优化：考试权重 14%

扩展并优化 GPU 硬件上的 LLM 训练和推理。涵盖多 GPU / 分布式部署、并行计算技术、故障排查、内存与批处理优化，以及性能分析。

7.1	配置多 GPU 和分布式训练设置 (DDP、FSDP、模型、工作流、张量、数据、序列和专家并行)。
7.2	应用 Tensor Core 和混合精度优化以及批量 / 内存管理，实现高效吞吐量。
7.3	分配和优化自注意力头部通用矩阵乘法 (GEMM) 操作，并采用梯度累积策略，确保在大模型训练或内存受限场景下的高效运行。
7.4	通过 CUDA 分析识别和解决瓶颈问题，并针对内存和内核效率问题进行故障排除。

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《使用 Nsight 分析工具优化 CUDA 机器学习代码》(查看课程)
- 课程：讲师指导的培训班《模型并行 —— 构建和部署大型神经网络》(查看课程大纲)

阅读内容推荐

- Distributed Data Parallel in PyTorch
- CUDA C++ Best Practices Guide 13.0
- Assess, Parallelize, Optimize, Deploy | NVIDIA Blog
- Parallelisms — NVIDIA NeMo User Guide
- Batching — NVIDIA NeMo Developer Docs
- PyTorch Lightning: Gradient Accumulation | GitHub
- Accelerate Usage Guides: Gradient Accumulation | Hugging Face

模型部署：考试权重 9%

使用容器化工作流实现生产级 LLM 部署，结合可扩展的任务编排、高效的批处理与模型服务，并提供实时性能监控。
8.1　分析模型类型 (编码器、解码器、编码器 — 解码器) 计算权衡，并针对内存和延迟进行优化。
8.2　构建容器化推理工作流，结合动态批处理策略，并通过 NVIDIA Dynamo-Triton 实现高效部署。
8.3　配置和管理服务 (Kubernetes、ensemble workflows)、实施实时监测并在 Docker 中运行模型。

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：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 NIM Microservices for Accelerated AI Inference
- Overview of NVIDIA NIM for Large Language Models (LLMs)
- Power Your AI Projects With New NVIDIA NIMs for Mistral and Mixtral Models | NVIDIA Blog
- Concurrent Model Execution—Dynamo-Triton (Previously NVIDIA Triton™ Inference Server) User Guide
- Model Configuration—Dynamo-Triton (Previously Triton Inference Server) User Guide
- Schedulers—Dynamo-Triton (Previously Triton Inference Server) User Guide
- Batchers—Dynamo-Triton (Previously Triton Inference Server) User Guide

生产监测和可靠性：考试权重 7%

建立监测仪表板和可靠性指标，同时追踪日志和异常，实现根因分析。评估基准测试代理与历史版本的表现差异。实施自动化调优、再训练和版本管理，确保生产环境部署的持续可用性、透明性和可信度。

9.1	定义监测仪表板和可靠性指标。
9.2	追踪日志、错误和异常，进行根因分析。
9.3	持续对已部署的代理进行基准测试，并与历史版本进行对比。
9.4	在生产环境中实现自动调优、再训练和版本控制。
9.5	确保部署到生产环境上时可保持正常运行、且具透明度和可信度。

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

> Attention I All You Need—arXiv
> BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding—arXiv
> Improving Language Understanding by Generative Pretraining (Radford, Narasimhan, Salimans, Sutskever) | OpenAI
> Masked Language Modeling Guide | Hugging Face
> Causal Language Modeling Guide | Hugging Face

安全、道德与合规：考试权重 5%

在整个 LLM 生命周期中实践 Responsible AI 原则。包括偏差与公平性的审计、实施防护措施、配置监测，确保伦理合规，以及应用偏差检测与缓解策略，确保 LLM 在部署和使用过程中符合伦理与责任要求。	
10.1	将 Responsible AI 原则实践应用于模型部署。
10.2	审计 LLM 的偏差和公平性。
10.3	为生产环境的 LLM 配置监测系统。
10.4	实施偏差检测和缓解策略。
10.5	部署防护机制，防止生成不符合预期或不合规的 LLM 输出。

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《开发基于提示工程的大语言模型 (LLM) 应用》(查看课程) 或讲师指导的培训班《利用提示工程构建大语言模型 (LLM) 应用》(查看课程大纲)
- 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

- Build an Enterprise RAG Pipeline Blueprint | build.nvidia.com
- RAG 101: Demystifying Retrieval-Augmented Generation Pipelines | NVIDIA Technical Blog
- Full-Stack Observability for NVIDIA Blackwell and NIM-Based AI | Dynatrace
- Large-Scale Production Deployment of RAG Pipelines | NVIDIA Deep Learning Institute
- Observability Tool NVIDIA Generative AI Examples 0.5.0 | GitHub
- NVIDIA-AI-Blueprints/RAG | GitHub
- Measuring the Effectiveness and Performance of AI Guardrails in Generative AI Applications | NVIDIA Technical Blog

更多咨询

请邮件至 dlichina@nvidia.com