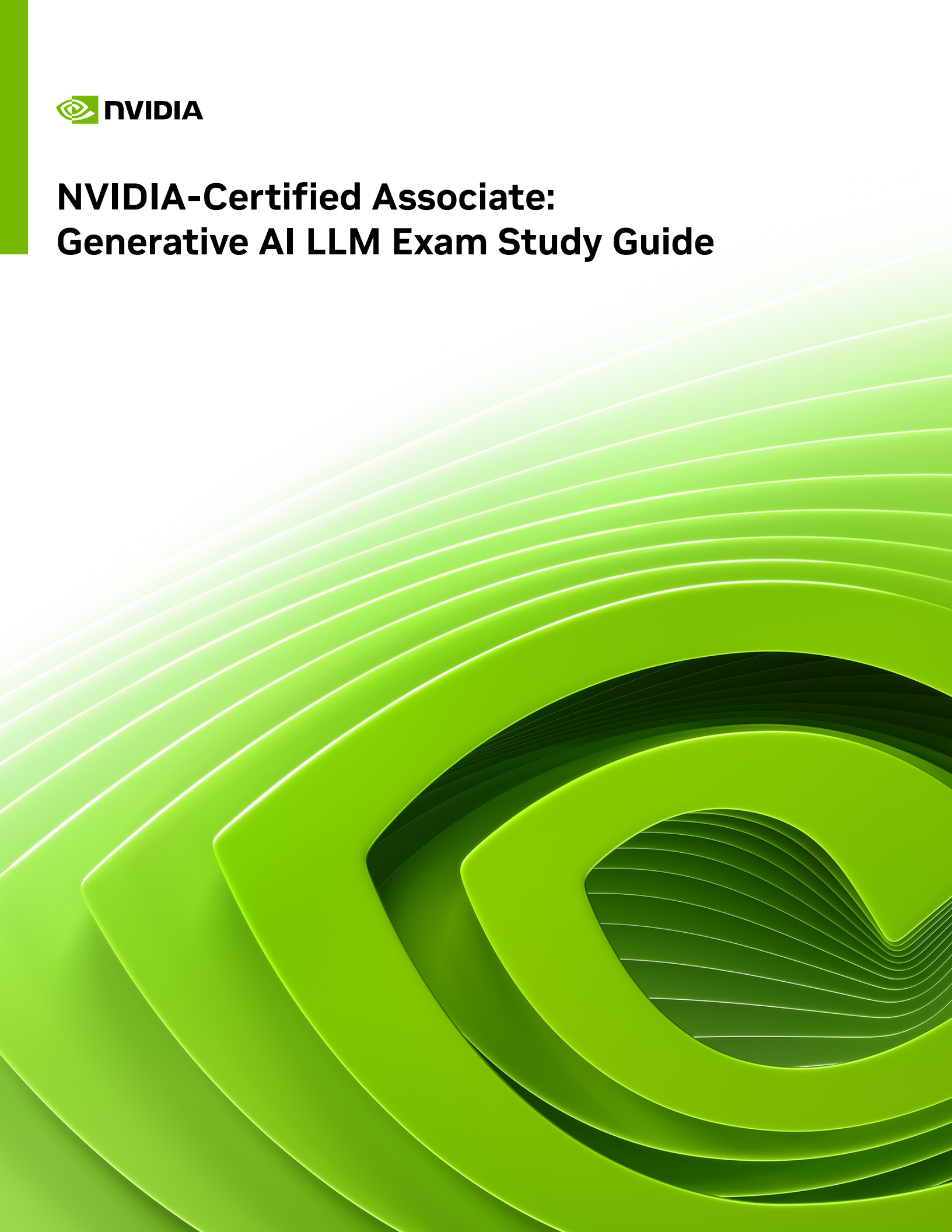
NVIDIA

NVIDIA-Certified Associate: Generative AI LLM Exam Study Guide



NVIDIA-Certified Associate: Generative AI LLM Exam Study Guide

Contents

This study guide provides an overview of each topic covered on the NVIDIA Generative AI LLM certification exam, recommended training, and suggested reading to prepare for the exam.

Information about NVIDIA certifications can be found **here**.

Job Description

The generative AI-large language model (LLM) associate developer is responsible for contributing to the development, programming, and quality assurance of state-of-the-art generative AI LLM systems. They work with a team of skilled AI professionals to develop datasets, select models to train, train models, and implement model testing and debugging processes. The associate should have an understanding of the deployment of models for applications. They'll also be responsible for developing high-quality software design and construction, programming in a variety of languages and platforms, and maintaining system updates.

Job Responsibilities

1. Collaborate with the AI development team to design, code, test, debug, and document programming applications.
2. Perform system analysis to ensure software and systems meet required specifications.
3. Aid in integrating new AI language models into existing systems or creating new ones as needed.
4. Assist in the assessment and resolution of application and system performance issues.
5. Stay updated on new AI models and other developments related to language learning models.
6. Contribute to the production of technical documents and manuals.
7. Conduct software programming and documentation development under the direction of senior staff.
8. Perform prompt engineering.
9. Select models.
10. Define, curate, label, and annotate LLM datasets.
11. Perform experimentation like A/B testing, evaluating prompts, evaluating models, and producing POCs.

Recommended Qualifications and Experience

1. Bachelor's degree in computer science, software engineering, AI, or a related field
2. Knowledge of Python, C, and AI frameworks (PyTorch, TensorFlow, etc.)
3. Solid understanding of neural networks and deep learning models

Certification Topics and References

Core Machine Learning and AI Knowledge: Exam Weight 30%

Knowledge of algorithms, conventions, and techniques that allow computers to learn from and make predictions or decisions based on data.

1.1	Assist in deployment and evaluation of model scalability, performance, and reliability under the supervision of senior team members.
1.2	Awareness of the process of extracting insights from large datasets using data mining, data visualization, and similar techniques.
1.3	Build LLM use cases such as retrieval-augmented generation (RAG), chatbots, and summarizers.
1.4	Curate and embed content datasets for RAGs.
1.5	Familiarity with the fundamentals of machine learning (e.g., feature engineering, model comparison, cross validation).
1.6	Familiarity with the capabilities of Python natural language packages (spaCy, NumPy, vector databases, etc.).
1.7	Read research papers (articles, conference papers, etc.) to identify emerging LLM trends and technologies.
1.8	Select and use models to create text embeddings.
1.9	Use prompt engineering principles to create prompts to achieve desired results.
1.10	Use Python packages (spaCy, NumPy, Keras, etc.) to implement specific traditional machine learning analyses.

NVIDIA Course Objectives

Course reference: **Fundamentals of Deep Learning**, **Getting Started With Deep Learning**

- Learn the fundamental techniques and tools required to train a deep learning model.
- Gain experience with common deep learning data types and model architectures.
- Leverage transfer learning between models to achieve efficient results with less data and computation.

Course reference: **Introduction to Transformer-Based Natural Language Processing**

- Learn how transformers are used as the building blocks of modern LLMs.

Course reference: **Building Transformer-Based Natural Language Processing Applications**

- Understand how transformers are used as the basic building blocks of modern LLMs for NLP applications.
- See how self-supervision improves upon the transformer architecture in BERT, Megatron, and other LLM variants for superior NLP results.

Course reference: **Fundamentals of Accelerated Data Science**

- Utilize a wide variety of machine learning algorithms, including XGBoost, for different data science problems.
- Learn and apply powerful graph algorithms to analyze complex networks with NetworkX and cuGraph.

Course reference: **Building LLM Applications With Prompt Engineering**

- Understand how to apply iterative prompt engineering best practices to create LLM-based applications for various language-related tasks.

Course reference: **Rapid Application Development With Large Language Models (LLMs)**

- Use encoder models for tasks like semantic analysis, embedding, question-answering, and zero-shot classification.
- Work with conditioned decoder-style models to take in and generate interesting data formats, styles, and modalities.

Suggested Readings

- **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**

Data Analysis: Exam Weight 14%

Inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.	
2.1	Awareness of the process of extracting insights from large datasets using data mining, data visualization, and similar techniques.
2.2	Compare models using statistical performance metrics, such as loss functions or proportion of explained variance.
2.3	Conduct data analysis under the supervision of a senior team member.
2.4	Create graphs, charts, or other visualizations to convey the results of data analysis using specialized software.
2.5	Identify relationships and trends or any factors that could affect the results of research.

NVIDIA Course Objectives

Course reference: **Fundamentals of Deep Learning, Getting Started With Deep Learning**

- Enhance datasets through data augmentation to improve model accuracy.

Course reference: **Introduction to Transformer-Based Natural Language Processing**

- Use transformer-based models for text classification.
- Apply LLMs for named-entity recognition (NER).
- Utilize transformer models for author attribution.

Course reference: **Building Transformer-Based Natural Language Processing Applications**

- How to leverage pretrained, modern LLM models to solve multiple NLP tasks, such as text classification, named-entity recognition, and question-answering.

Course reference: **Building LLM Applications With Prompt Engineering**

- Be proficient in using LangChain to organize and compose LLM workflows.

Course reference: **Rapid Application Development With Large Language Models (LLMs)**

- Explore using LangChain and LangGraph for orchestrating data pipelines and environment-enabled agents.

Course reference: **Fundamentals of Accelerated Data Science**

- Use cuDF to accelerate pandas, Polars, and Dask for analyzing datasets of all sizes efficiently.
- Perform multiple analysis tasks on massive datasets to stave off a simulated epidemic outbreak affecting the UK.

Suggested Readings

- **RAPIDS**
- **cuML 24.04.00 documentation**
- **GPU Accelerated Data Science With RAPIDS**
- **Data Exploration**
- **Stemming and Lemmatizing With sklearn Vectorizers**

Experimentation: Exam Weight 22%

The study of how to perform, evaluate, and interpret experiments, including AI model evaluation and the use of human subjects in labeling or reinforcement learning from human feedback (RLHF).
3.1 Awareness of the process of extracting insights from large datasets using data mining, data visualization, and similar techniques.
3.2 Compare models using statistical performance metrics, such as loss functions or proportion of explained variance.
3.3 Conduct data analysis under the supervision of a senior team member.
3.4 Create graphs, charts, or other visualizations to convey the results of data analysis using specialized software.
3.5 Identify relationships and trends or any factors that could affect the results of research.

NVIDIA Course Objectives

Course reference: **Fundamentals of Deep Learning, Getting Started With Deep Learning**
- Build confidence to take on your own project with a modern deep learning framework.
- Leverage transfer learning between models to achieve efficient results with less data and computation.

Course reference: **Introduction to Transformer-Based Natural Language Processing**
- Experiment with transformer-based models for various NLP tasks.
- Test and compare model performance on question-answering tasks.

Course reference: **Building Transformer-Based Natural Language Processing Applications**
- Leverage pretrained, modern NLP models to solve multiple tasks, such as text classification, NER, and question-answering.

Course reference: **Fundamentals of Accelerated Data Science**
- Learn and apply powerful graph algorithms to analyze complex networks with NetworkX and cuGraph.

Course reference: **Rapid Application Development With Large Language Models (LLMs)**
- Find, pull in, and experiment with the HuggingFace model repository and Transformers API.

Course reference: **Building LLM Applications With Prompt Engineering**
- Be proficient in using LangChain to organize and compose LLM workflows.

Suggested Reading List

- **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**

Software Development: Exam Weight 24%

Create, maintain, and test software.
4.1 Assist in the deployment and evaluations of model scalability, performance, and reliability under the supervision of senior team member.
4.2 Build LLM use cases such as RAGs, chatbots, and summarizers.
4.3 Familiarity with the capabilities of Python natural language packages (spaCy, NumPy, vector databases, etc.).
4.4 Identify system data, hardware, or software components required to meet user needs.
4.5 Monitor functioning of data collection, experiments, and other software processes.
4.6 Use Python packages (spaCy, NumPy, Keras, etc.) to implement specific traditional machine learning analyses.
4.7 Write software components or scripts under the supervision of a senior team member.

NVIDIA Course Objectives

Course reference: **Fundamentals of Deep Learning, Getting Started With Deep Learning**
- Gain experience with common deep learning data types and model architectures.
- Build confidence to take on your own project with a modern deep learning framework.

Course reference: **Introduction to Transformer-Based Natural Language Processing**
- Implement transformer-based models for different NLP applications.
- Develop solutions for text classification, NER, author attribution, and question-answering using LLMs.

Course reference: **Building Transformer-Based Natural Language Processing Applications**
- Manage inference challenges and deploy refined models for live applications.

Course reference: **Fundamentals of Accelerated Data Science**
- Deploy machine learning models on a Triton Inference Server to deliver optimal performance.

Course reference: **Rapid Application Development With Large Language Models (LLMs)**
- Kick-start and guide generative AI solutions for safe, effective, and scalable natural data tasks.

Course reference: **Building LLM Applications With Prompt Engineering**
- Write application code to harness LLMs for generative tasks, document analysis, chatbot applications, and more.

Suggested Readings

- **TensorRT—Get Started**, NVIDIA Developer
- **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**

Trustworthy AI: Exam Weight 10%

Creation and assessment of ethical, energy-conscious, and reliable artificial intelligence systems capable of interpreting and integrating various forms of data, ensuring that they're designed and applied in a manner that's transparent, fair, and verifiable.
5.1 Describe the ethical principles of trustworthy AI.
5.2 Describe the balance between data privacy and the importance of data consent.
5.3 Describe how to use NVIDIA and other technologies to improve AI trustworthiness.
5.4 Describe how to minimize bias in AI systems.

NVIDIA Course Objectives

Course reference: **Rapid Application Development With Large Language Models (LLMs)**

> Kick-start and guide generative AI solutions for safe, effective, and scalable natural data tasks..

Suggested Readings

> **Trustworthy AI for A Better World**, NVIDIA
> **What Is Trustworthy AI?**, NVIDIA Blog
> **What Is Retrieval-Augmented Generation aka RAG?**, NVIDIA Blogs

Questions?

Contact us **here**.

NVIDIA