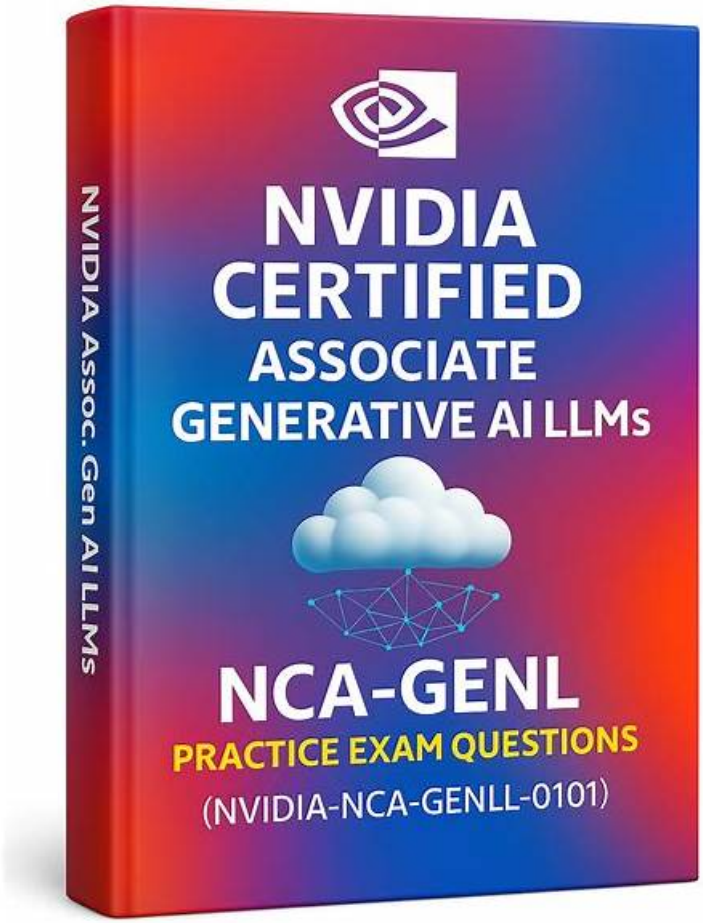
Latest NCA-GENL Mock Test - Exam NCA-GENL Registration



P.S. Free & New NCA-GENL dumps are available on Google Drive shared by TestkingPass: https://drive.google.com/open?id=1YC9-PnP2AZ87BW2CfBPa-GCm2LDUAazz

If you are looking to be NVIDIA NCA-GENL certified. TestkingPass is here to provide you with the best NVIDIA Generative AI LLMs (NCA-GENL) exam dumps through which you can clear your NVIDIA Generative AI LLMs (NCA-GENL) certification exam. We are providing practice exams in three formats including PDF which is the downloadable file from which you can study for your NVIDIA Generative AI LLMs (NCA-GENL) exam questions and our Web-based application provides you the facility to assess yourself without installing any software on your device to prepare you for NVIDIA Generative AI LLMs (NCA-GENL)exam dumps.

As you know, the low-quality latest NCA-GENL exam torrent may do harmful influence on you which may causes results past redemption. Whether you have experienced that problem or not was history by now. The free demos do honor to the perfection of our latest NCA-GENL exam torrent, and also a performance of our considerate after sales services. Those demos serve as epitomes of real NCA-GENL Quiz guides for your reference. In our demos, some examples or question points were enumerated as some representatives of our NCA-GENL test prep. How convenient and awesome of it!

>> Latest NCA-GENL Mock Test <<

2026 100% Free NCA-GENL –Updated 100% Free Latest Mock Test | Exam NVIDIA Generative AI LLMs Registration

The unmatched and the most workable study guides of TestkingPass are your real destination to achieve your goal. The pathway to pass NCA-GENL was not so easy and perfectly reliable as it has become now with the help of our products. Just you need to spend a few hours daily for two week and you can surely get the best insight of the syllabus and command over it. The NCA-GENL Questions and answers in the guide are meant to deliver you simplified and the most up to date information in as fewer words as possible.

NVIDIA Generative AI LLMs Sample Questions (Q64-Q69):

NEW QUESTION # 64
What is the fundamental role of LangChain in an LLM workflow?

- A. To orchestrate LLM components into complex workflows.
- B. To act as a replacement for traditional programming languages.
- C. To reduce the size of AI foundation models.
- D. To directly manage the hardware resources used by LLMs.

Answer: A

Explanation:
LangChain is a framework designed to simplify the development of applications powered by large language models (LLMs) by orchestrating various components, such as LLMs, external data sources, memory, and tools, into cohesive workflows. According to NVIDIA's documentation on generative AI workflows, particularly in the context of integrating LLMs with external systems, LangChain enables developers to build complex applications by chaining together prompts, retrieval systems (e.g., for RAG), and memory modules to maintain context across interactions. For example, LangChain can integrate an LLM with a vector database for retrieval-augmented generation or manage conversational history for chatbots. Option A is incorrect, as LangChain complements, not replaces, programming languages. Option B is wrong, as LangChain does not modify model size. Option D is inaccurate, as hardware management is handled by platforms like NVIDIA Triton, not LangChain.
References:
NVIDIA NeMo Documentation: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp
/intro.html
LangChain Official Documentation: https://python.langchain.com/docs/get_started/introduction

NEW QUESTION # 65
Which of the following is a parameter-efficient fine-tuning approach that one can use to fine-tune LLMs in a memory-efficient fashion?

- A. TensorRT
- B. Chinchilla
- C. NeMo
- D. LoRA

Answer: D

Explanation:
LoRA (Low-Rank Adaptation) is a parameter-efficient fine-tuning approach specifically designed for large language models (LLMs), as covered in NVIDIA's Generative AI and LLMs course. It fine-tunes LLMs by updating a small subset of parameters through low-rank matrix factorization, significantly reducing memory and computational requirements compared to full fine-tuning. This makes LoRA ideal for adapting large models to specific tasks while maintaining efficiency. Option A, TensorRT, is incorrect, as it is an inference optimization library, not a fine-tuning method. Option B, NeMo, is a framework for building AI models, not a specific fine-tuning technique. Option C, Chinchilla, is a model, not a fine-tuning approach. The course emphasizes: "Parameter-efficient fine-tuning methods like LoRA enable memory-efficient adaptation of LLMs by updating low-rank approximations of weight matrices, reducing resource demands while maintaining performance." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

NEW QUESTION # 66

When comparing and contrasting the ReLU and sigmoid activation functions, which statement is true?

- A. ReLU is less computationally efficient than sigmoid, but it is more accurate than sigmoid.
- B. ReLU and sigmoid both have a range of 0 to 1.
- C. ReLU is more computationally efficient, but sigmoid is better for predicting probabilities.
- D. ReLU is a linear function while sigmoid is non-linear.

Answer: C

Explanation:
ReLU (Rectified Linear Unit) and sigmoid are activation functions used in neural networks. According to NVIDIA's deep learning documentation (e.g., cuDNN and TensorRT), ReLU, defined as f(x) = max(0, x), is computationally efficient because it involves simple thresholding, avoiding expensive exponential calculations required by sigmoid, f(x) = 1/(1 + e