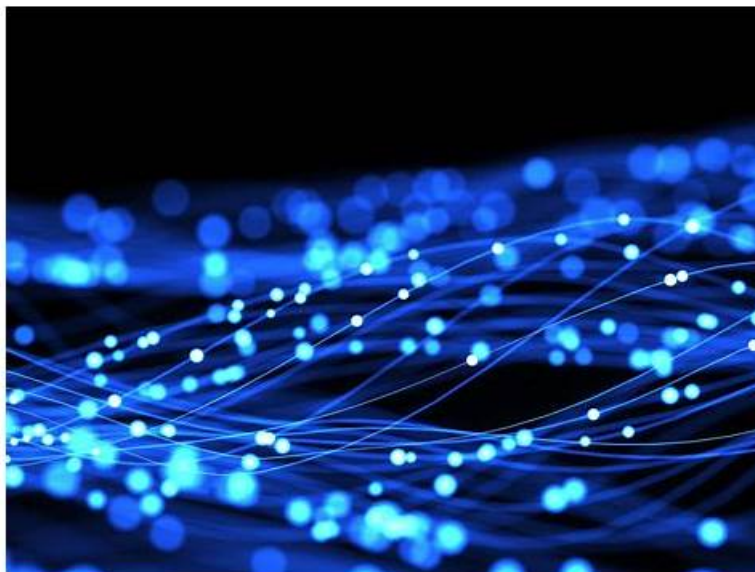


NVIDIA NCA-GENL Valid Exam Voucher | NCA-GENL New Study Guide



What's more, part of that TrainingQuiz NCA-GENL dumps now are free: https://drive.google.com/open?id=1nrBCQ9Oe-qVHjqih0yimzc_Rx_StuID

Based on the research results of the examination questions over the years, the experts give more detailed explanations of the contents of the frequently examined contents and difficult-to-understand contents, and made appropriate simplifications for infrequently examined contents. NCA-GENL test questions make it possible for students to focus on the important content which greatly shortens the students' learning time. With NCA-GENL Exam Torrent, you will no longer learn blindly but in a targeted way. With NCA-GENL exam guide, you only need to spend 20-30 hours to study and you can successfully pass the exam. You will no longer worry about your exam because of bad study materials. If you decide to choose and practice our NCA-GENL test questions, our life will be even more exciting.

NVIDIA NCA-GENL Exam Syllabus Topics:

Topic	Details
Topic 1	<ul style="list-style-type: none">• Prompt engineering: Focuses on techniques for designing and refining input prompts to effectively guide LLM outputs toward desired results.
Topic 2	<ul style="list-style-type: none">• Experiment design: Focuses on structuring controlled tests and workflows to systematically evaluate LLM performance and outcomes.
Topic 3	<ul style="list-style-type: none">• Python libraries for LLMs: Covers key Python frameworks and tools — such as LangChain, Hugging Face, and similar libraries — used to build and interact with LLMs.
Topic 4	<ul style="list-style-type: none">• LLM integration and deployment: Addresses connecting LLMs into real-world applications and deploying them reliably across production environments.
Topic 5	<ul style="list-style-type: none">• Data analysis and visualization: Covers interpreting datasets and presenting insights through visual tools to support informed model development decisions.
Topic 6	<ul style="list-style-type: none">• Alignment: Addresses methods for ensuring LLM behavior is safe, accurate, and consistent with human intentions and values.

NVIDIA NCA-GENL New Study Guide & Practice NCA-GENL Test Engine

In the such a brilliant era of IT industry in the 21st century competition is very fierce. Naturally, NVIDIA Certification NCA-GENL Exam has become a very popular exam in the IT area. More and more people register for the exam and passing the certification exam is also those ambitious IT professionals' dream.

NVIDIA Generative AI LLMs Sample Questions (Q79-Q84):

NEW QUESTION # 79

Which metric is commonly used to evaluate machine-translation models?

- A. BLEU score
- B. ROUGE score
- C. F1 Score
- D. Perplexity

Answer: A

Explanation:

The BLEU (Bilingual Evaluation Understudy) score is the most commonly used metric for evaluating machine-translation models. It measures the precision of n-gram overlaps between the generated translation and reference translations, providing a quantitative measure of translation quality. NVIDIA's NeMo documentation on NLP tasks, particularly machine translation, highlights BLEU as the standard metric for assessing translation performance due to its focus on precision and fluency. Option A (F1 Score) is used for classification tasks, not translation. Option C (ROUGE) is primarily for summarization, focusing on recall. Option D (Perplexity) measures language model quality but is less specific to translation evaluation.

References:

NVIDIA NeMo Documentation: <https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/intro.html>

Papineni, K., et al. (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation."

NEW QUESTION # 80

"Hallucinations" is a term coined to describe when LLM models produce what?

- A. Correct sounding results that are wrong.
- B. Grammatically incorrect or broken outputs.
- C. Images from a prompt description.
- D. Outputs are only similar to the input data.

Answer: A

Explanation:

In the context of LLMs, "hallucinations" refer to outputs that sound plausible and correct but are factually incorrect or fabricated, as emphasized in NVIDIA's Generative AI and LLMs course. This occurs when models generate responses based on patterns in training data without grounding in factual knowledge, leading to misleading or invented information. Option A is incorrect, as hallucinations are not about similarity to input data but about factual inaccuracies. Option B is wrong, as hallucinations typically refer to text, not image generation. Option D is inaccurate, as hallucinations are grammatically coherent but factually wrong. The course states: "Hallucinations in LLMs occur when models produce correct-sounding but factually incorrect outputs, posing challenges for ensuring trustworthy AI." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

NEW QUESTION # 81

Imagine you are training an LLM consisting of billions of parameters and your training dataset is significantly larger than the available RAM in your system. Which of the following would be an alternative?

- A. Using a memory-mapped file that allows the library to access and operate on elements of the dataset without needing to fully load it into memory.
- B. Eliminating sentences that are syntactically different by semantically equivalent, possibly reducing the risk of the model hallucinating as it is trained to get to the point.

- C. Discarding the excess of data and pruning the dataset to the capacity of the RAM, resulting in reduced latency during inference.
- D. Using the GPU memory to extend the RAM capacity for storing the dataset and move the dataset in and out of the GPU, using the PCI bandwidth possibly.

Answer: A

Explanation:

When training an LLM with a dataset larger than available RAM, using a memory-mapped file is an effective alternative, as discussed in NVIDIA's Generative AI and LLMs course. Memory-mapped files allow the system to access portions of the dataset directly from disk without loading the entire dataset into RAM, enabling efficient handling of large datasets. This approach leverages virtual memory to map file contents to memory, reducing memory bottlenecks. Option A is incorrect, as moving large datasets in and out of GPU memory via PCI bandwidth is inefficient and not a standard practice for dataset storage. Option C is wrong, as discarding data reduces model quality and is not a scalable solution. Option D is inaccurate, as eliminating semantically equivalent sentences is a specific preprocessing step that does not address memory constraints.

The course states: "Memory-mapped files enable efficient training of LLMs on large datasets by accessing data from disk without loading it fully into RAM, overcoming memory limitations." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

NEW QUESTION # 82

Why do we need positional encoding in transformer-based models?

- A. To reduce the dimensionality of the input data.
- **B. To represent the order of elements in a sequence.**
- C. To prevent overfitting of the model.
- D. To increase the throughput of the model.

Answer: B

Explanation:

Positional encoding is a critical component in transformer-based models because, unlike recurrent neural networks (RNNs), transformers process input sequences in parallel and lack an inherent sense of word order.

Positional encoding addresses this by embedding information about the position of each token in the sequence, enabling the model to understand the sequential relationships between tokens. According to the original transformer paper ("Attention is All You Need" by Vaswani et al., 2017), positional encodings are added to the input embeddings to provide the model with information about the relative or absolute position of tokens. NVIDIA's documentation on transformer-based models, such as those supported by the NeMo framework, emphasizes that positional encodings are typically implemented using sinusoidal functions or learned embeddings to preserve sequence order, which is essential for tasks like natural language processing (NLP). Options B, C, and D are incorrect because positional encoding does not address overfitting, dimensionality reduction, or throughput directly; these are handled by other techniques like regularization, dimensionality reduction methods, or hardware optimization.

References:

Vaswani, A., et al. (2017). "Attention is All You Need."

NVIDIA NeMo Documentation:<https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/intro.html>

NEW QUESTION # 83

You are in need of customizing your LLM via prompt engineering, prompt learning, or parameter-efficient fine-tuning. Which framework helps you with all of these?

- A. NVIDIA Triton
- **B. NVIDIA NeMo**
- C. NVIDIA TensorRT
- D. NVIDIA DALI

Answer: B

Explanation:

The NVIDIA NeMo framework is designed to support the development and customization of large language models (LLMs), including techniques like prompt engineering, prompt learning (e.g., prompt tuning), and parameter-efficient fine-tuning (e.g., LoRA),

