

NCA-GENL Deutsch, NCA-GENL Prüfungsaufgaben



P.S. Kostenlose 2026 NVIDIA NCA-GENL Prüfungsfragen sind auf Google Drive freigegeben von It-Pruefung verfügbar:
https://drive.google.com/open?id=1rVIBL7_0DuNuuL75191v62fFX6Yrba3q

Auf unterschiedliche Art und Weise kann man verschiedene Zwecke erfüllen. Was wichtig ist, dass man welchen Weg einschlägt. Viele Leute beteiligen sich an der NVIDIA NCA-GENL Zertifizierungsprüfung, um seine Lebens- und Arbeitsumstände zu verbessern. Wie alle wissen, dass es nicht so leicht ist, die NVIDIA NCA-GENL (NVIDIA Generative AI LLMs) Zertifizierungsprüfung zu bestehen. Für die Prüfung verwendet man viel Energie und Zeit. Traurigerweise haben sie die NVIDIA NCA-GENL Prüfung noch nicht bestanden.

NVIDIA NCA-GENL Prüfungsplan:

Thema	Einzelheiten
Thema 1	<ul style="list-style-type: none">• Software development: Covers the programming practices and coding skills required to build, maintain, and deploy generative AI applications.
Thema 2	<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.
Thema 3	<ul style="list-style-type: none">• LLM integration and deployment: Addresses connecting LLMs into real-world applications and deploying them reliably across production environments.
Thema 4	<ul style="list-style-type: none">• Fundamentals of machine learning and neural networks: Covers the core concepts of how machine learning models learn from data, including the structure and function of neural networks that underpin large language models.
Thema 5	<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.
Thema 6	<ul style="list-style-type: none">• Data preprocessing and feature engineering: Covers preparing raw data through cleaning, transformation, and feature selection to make it suitable for model training.

>> NCA-GENL Deutsch <<

NCA-GENL Schulungsangebot, NCA-GENL Testing Engine, NVIDIA

Generative AI LLMs Trainingsunterlagen

Heutzutage, wo die Zeit besonders geschätzt wird, ist es kostengünstig, It-Prüfung zum Bestehen der NVIDIA NCA-GENL Zertifizierungsprüfung zu wählen. Wenn Sie It-Prüfung wählen, würden wir mit äußerster Kraft Ihnen helfen, die NVIDIA NCA-GENL Prüfung zu bestehen. Außerdem bieten wir Ihnen einen einjährigen kostenlosen Update-Service. Fallen Sie in der Prüfung durch, zahlen wir Ihnen gesammte Einkaufsgebühren zurück.

NVIDIA Generative AI LLMs NCA-GENL Prüfungsfragen mit Lösungen (Q43-Q48):

43. Frage

In the transformer architecture, what is the purpose of positional encoding?

- A. To remove redundant information from the input sequence.
- B. To encode the importance of each token in the input sequence.
- C. To encode the semantic meaning of each token in the input sequence.
- **D. To add information about the order of each token in the input sequence.**

Antwort: D

Begründung:

Positional encoding is a vital component of the Transformer architecture, as emphasized in NVIDIA's Generative AI and LLMs course. Transformers lack the inherent sequential processing of recurrent neural networks, so they rely on positional encoding to incorporate information about the order of tokens in the input sequence. This is typically achieved by adding fixed or learned vectors (e.g., sine and cosine functions) to the token embeddings, where each position in the sequence has a unique encoding. This allows the model to distinguish the relative or absolute positions of tokens, enabling it to understand word order in tasks like translation or text generation. For example, in the sentence "The cat sleeps," positional encoding ensures the model knows "cat" is the second token and "sleeps" is the third. Option A is incorrect, as positional encoding does not remove information but adds positional context. Option B is wrong because semantic meaning is captured by token embeddings, not positional encoding. Option D is also inaccurate, as the importance of tokens is determined by the attention mechanism, not positional encoding. The course notes: "Positional encodings are used in Transformers to provide information about the order of tokens in the input sequence, enabling the model to process sequences effectively." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

44. Frage

In large-language models, what is the purpose of the attention mechanism?

- A. To determine the order in which words are generated.
- B. To measure the importance of the words in the output sequence.
- C. To capture the order of the words in the input sequence.
- **D. To assign weights to each word in the input sequence.**

Antwort: D

Begründung:

The attention mechanism is a critical component of large language models, particularly in Transformer architectures, as covered in NVIDIA's Generative AI and LLMs course. Its primary purpose is to assign weights to each token in the input sequence based on its relevance to other tokens, allowing the model to focus on the most contextually important parts of the input when generating or interpreting text. This is achieved through mechanisms like self-attention, where each token computes a weighted sum of all other tokens' representations, with weights determined by their relevance (e.g., via scaled dot-product attention). This enables the model to capture long-range dependencies and contextual relationships effectively, unlike traditional recurrent networks. Option A is incorrect because attention focuses on the input sequence, not the output sequence. Option B is wrong as the order of generation is determined by the model's autoregressive or decoding strategy, not the attention mechanism itself. Option C is also inaccurate, as capturing the order of words is the role of positional encoding, not attention. The course highlights: "The attention mechanism enables models to weigh the importance of different tokens in the input sequence, improving performance in tasks like translation and text generation." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language

45. Frage

Your company has upgraded from a legacy LLM model to a new model that allows for larger sequences and higher token limits. What is the most likely result of upgrading to the new model?

- A. The number of tokens is fixed for all existing language models, so there is no benefit to upgrading to higher token limits.
- B. The newer model allows for larger context, so the outputs will improve without increasing inference time overhead.
- **C. The newer model allows larger context, so outputs will improve, but you will likely incur longer inference times.**
- D. The newer model allows the same context lengths, but the larger token limit will result in more comprehensive and longer outputs with more detail.

Antwort: C

Begründung:

Upgrading to a new LLM with larger sequence lengths and higher token limits, as discussed in NVIDIA's Generative AI and LLMs course, typically allows the model to process larger contexts, leading to improved output quality due to better understanding of extended dependencies in text. However, handling larger sequences increases computational requirements, often resulting in longer inference times, especially on the same hardware. This trade-off is a key consideration in LLM deployment. Option A is incorrect, as token limits vary across models, and higher limits offer benefits. Option B is wrong, as larger context processing typically increases inference time. Option C is inaccurate, as higher token limits primarily enable larger context, not just longer outputs. The course notes: "Larger sequence lengths in LLMs allow for improved output quality by capturing more context, but this often comes at the cost of increased inference times due to higher computational demands." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

46. Frage

In ML applications, which machine learning algorithm is commonly used for creating new data based on existing data?

- A. Decision tree
- **B. Generative adversarial network**
- C. Support vector machine
- D. K-means clustering

Antwort: B

Begründung:

Generative Adversarial Networks (GANs) are a class of machine learning algorithms specifically designed for creating new data based on existing data, as highlighted in NVIDIA's Generative AI and LLMs course. GANs consist of two models—a generator that produces synthetic data and a discriminator that evaluates its authenticity—trained adversarially to generate realistic data, such as images, text, or audio, that resembles the training distribution. This makes GANs a cornerstone of generative AI applications. Option A, Decision tree, is incorrect, as it is primarily used for classification and regression tasks, not data generation. Option B, Support vector machine, is a discriminative model for classification, not generation. Option D, K-means clustering, is an unsupervised clustering algorithm and does not generate new data. The course emphasizes: "Generative Adversarial Networks (GANs) are used to create new data by learning to mimic the distribution of the training dataset, enabling applications in generative AI." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

47. Frage

What are some methods to overcome limited throughput between CPU and GPU? (Pick the 2 correct responses)

- **A. Upgrade the GPU to a higher-end model.**
- **B. Using techniques like memory pooling.**
- C. Increase the clock speed of the CPU.
- D. Increase the number of CPU cores.

Antwort: A,B

Begründung:

Limited throughput between CPU and GPU often results from data transfer bottlenecks or inefficient resource utilization. NVIDIA's documentation on optimizing deep learning workflows (e.g., using CUDA and cuDNN) suggests the following:

* Option B: Memory pooling techniques, such as pinned memory or unified memory, reduce data transfer overhead by optimizing

