

# Exam NCA-GENL Bible - NCA-GENL Certification Practice



P.S. Free 2026 NVIDIA NCA-GENL dumps are available on Google Drive shared by Pass4sures: <https://drive.google.com/open?id=1vk9EougONQXjvtETEQHTOn9fEQLW8uZK>

Our NCA-GENL learning guide is for the world and users are very extensive. In order to give users a better experience, we have been constantly improving. The high quality and efficiency of NCA-GENL exam prep has been recognized by users. The high passing rate of our NCA-GENL test materials are its biggest feature. As long as you use NCA-GENL Exam Prep, you can certainly harvest what you want thing. Not only you can pass the NCA-GENL exam in the shortest time, but also you can obtain the dreaming NCA-GENL certification to have a brighter future.

In use process, if you have some problems on our NCA-GENL study materials provide 24 hours online services, you can email or contact us on the online platform. In addition, our backstage will also help you check whether the NCA-GENL exam prep is updated in real-time. If there is an update, our system will send to the customer automatically. Our NCA-GENL Learning Materials also provide professional staff for remote assistance, to help users immediate effective solve the existing problems if necessary. So choosing our NCA-GENL study materials make you worry-free.

>> Exam NCA-GENL Bible <<

## NCA-GENL Certification Practice | NCA-GENL Reliable Dump

Customer first, service first is our principle of service. If you buy our NCA-GENL study guide, you will find our after sale service is so considerate for you. We are glad to meet your all demands and answer your all question about our NCA-GENL Training Materials. So do not hesitate and buy our NCA-GENL study guide, we believe you will find surprise from our products. you should have the right to enjoy the perfect after sale service and the high quality products!

### NVIDIA NCA-GENL Exam Syllabus Topics:

Topic	Details
Topic 1	<ul style="list-style-type: none"><li>• Prompt engineering: Focuses on techniques for designing and refining input prompts to effectively guide LLM outputs toward desired results.</li></ul>
Topic 2	<ul style="list-style-type: none"><li>• LLM integration and deployment: Addresses connecting LLMs into real-world applications and deploying them reliably across production environments.</li></ul>
Topic 3	<ul style="list-style-type: none"><li>• Alignment: Addresses methods for ensuring LLM behavior is safe, accurate, and consistent with human intentions and values.</li></ul>
Topic 4	<ul style="list-style-type: none"><li>• Experiment design: Focuses on structuring controlled tests and workflows to systematically evaluate LLM performance and outcomes.</li></ul>
Topic 5	<ul style="list-style-type: none"><li>• 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.</li></ul>

### NVIDIA Generative AI LLMs Sample Questions (Q49-Q54):

### NEW QUESTION # 49

In the context of language models, what does an autoregressive model predict?

- A. The probability of the next token in a text given the previous tokens.
- B. The probability of the next token using a Monte Carlo sampling of past tokens.
- C. The probability of the next token by looking at the previous and future input tokens.
- D. The next token solely using recurrent network or LSTM cells.

**Answer: A**

Explanation:

Autoregressive models are a cornerstone of modern language modeling, particularly in large language models (LLMs) like those discussed in NVIDIA's Generative AI and LLMs course. These models predict the probability of the next token in a sequence based solely on the preceding tokens, making them inherently sequential and unidirectional. This process is often referred to as "next-token prediction," where the model learns to generate text by estimating the conditional probability distribution of the next token given the context of all previous tokens. For example, given the sequence "The cat is," the model predicts the likelihood of the next word being "on," "in," or another token. This approach is fundamental to models like GPT, which rely on autoregressive decoding to generate coherent text. Unlike bidirectional models (e.g., BERT), which consider both previous and future tokens, autoregressive models focus only on past tokens, making option D incorrect. Options B and C are also inaccurate, as Monte Carlo sampling is not a standard method for next-token prediction in autoregressive models, and the prediction is not limited to recurrent networks or LSTM cells, as modern LLMs often use Transformer architectures. The course emphasizes this concept in the context of Transformer-based NLP: "Learn the basic concepts behind autoregressive generative models, including next-token prediction and its implementation within Transformer-based models." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

### NEW QUESTION # 50

Why might stemming or lemmatizing text be considered a beneficial preprocessing step in the context of computing TF-IDF vectors for a corpus?

- A. It enhances the aesthetic appeal of the text, making it easier for readers to understand the document's content.
- B. It guarantees an increase in the accuracy of TF-IDF vectors by ensuring more precise word usage distinction.
- C. It increases the complexity of the dataset by introducing more unique tokens, enhancing the distinctiveness of each document.
- D. It reduces the number of unique tokens by collapsing variant forms of a word into their root form, potentially decreasing noise in the data.

**Answer: D**

Explanation:

Stemming and lemmatizing are preprocessing techniques in NLP that reduce words to their root or base form, as discussed in NVIDIA's Generative AI and LLMs course. In the context of computing TF-IDF (Term Frequency-Inverse Document Frequency) vectors, these techniques are beneficial because they collapse variant forms of a word (e.g., "running," "ran" to "run") into a single token, reducing the number of unique tokens in the corpus. This decreases noise and dimensionality, improving the efficiency and effectiveness of TF-IDF representations for tasks like document classification or clustering. Option B is incorrect, as stemming and lemmatizing are not about aesthetics but about data preprocessing. Option C is wrong, as these techniques reduce, not increase, the number of unique tokens. Option D is inaccurate, as they do not guarantee accuracy improvements but rather reduce noise. The course states: "Stemming and lemmatizing reduce the number of unique tokens in a corpus by normalizing word forms, improving the quality of TF-IDF vectors by minimizing noise and dimensionality." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

### NEW QUESTION # 51

In Natural Language Processing, there are a group of steps in problem formulation collectively known as word representations (also word embeddings). Which of the following are Deep Learning models that can be used to produce these representations for NLP tasks? (Choose two.)

- A. Kubernetes
- B. TensorRT
- C. BERT
- D. Word2vec

- E. WordNet

**Answer: C,D**

Explanation:

Word representations, or word embeddings, are critical in NLP for capturing semantic relationships between words, as emphasized in NVIDIA's Generative AI and LLMs course. Word2vec and BERT are deep learning models designed to produce these embeddings. Word2vec uses shallow neural networks (CBOW or Skip-Gram) to generate dense vector representations based on word co-occurrence in a corpus, capturing semantic similarities. BERT, a Transformer-based model, produces contextual embeddings by considering bidirectional context, making it highly effective for complex NLP tasks. Option B, WordNet, is incorrect, as it is a lexical database, not a deep learning model. Option C, Kubernetes, is a container orchestration platform, unrelated to NLP or embeddings. Option D, TensorRT, is an inference optimization library, not a model for embeddings. The course notes: "Deep learning models like Word2vec and BERT are used to generate word embeddings, enabling semantic understanding in NLP tasks, with BERT leveraging Transformer architectures for contextual representations." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

### NEW QUESTION # 52

What metrics would you use to evaluate the performance of a RAG workflow in terms of the accuracy of responses generated in relation to the input query? (Choose two.)

- A. Response relevancy
- B. Context precision
- C. Retriever latency
- D. Generator latency
- E. Tokens generated per second

**Answer: A,B**

Explanation:

In a Retrieval-Augmented Generation (RAG) workflow, evaluating the accuracy of responses relative to the input query focuses on the quality of the retrieved context and the generated output. As covered in NVIDIA's Generative AI and LLMs course, two key metrics are response relevancy and context precision. Response relevancy measures how well the generated response aligns with the input query, often assessed through human evaluation or automated metrics like ROUGE or BLEU, ensuring the output is pertinent and accurate.

Context precision evaluates the retriever's ability to fetch relevant documents or passages from the knowledge base, typically measured by metrics like precision@k, which assesses the proportion of retrieved items that are relevant to the query. Options A (generator latency), B (retriever latency), and C (tokens generated per second) are incorrect, as they measure performance efficiency (speed) rather than accuracy. The course notes:

"In RAG workflows, response relevancy ensures the generated output matches the query intent, while context precision evaluates the accuracy of retrieved documents, critical for high-quality responses." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

### NEW QUESTION # 53

In evaluating the transformer model for translation tasks, what is a common approach to assess its performance?

- A. Evaluating the consistency of translation tone and style across different genres of text.
- B. Measuring the syntactic complexity of the model's translations against a corpus of professional translations.
- C. Comparing the model's output with human-generated translations on a standard dataset.
- D. Analyzing the lexical diversity of the model's translations compared to source texts.

**Answer: C**

Explanation:

A common approach to evaluate Transformer models for translation tasks, as highlighted in NVIDIA's Generative AI and LLMs course, is to compare the model's output with human-generated translations on a standard dataset, such as WMT (Workshop on Machine Translation) or BLEU-evaluated corpora. Metrics like BLEU (Bilingual Evaluation Understudy) score are used to quantify the similarity between machine and human translations, assessing accuracy and fluency. This method ensures objective, standardized evaluation.

