

# NVIDIA Generative AI LLMs Exam Questions Pdf & NCA-GENL Test Training Demo & NVIDIA Generative AI LLMs Test Online Engine



P.S. Free 2026 NVIDIA NCA-GENL dumps are available on Google Drive shared by ExamsLabs: [https://drive.google.com/open?id=139X59Icuqnb8JdqR\\_QJAbhW2bQsCt8d](https://drive.google.com/open?id=139X59Icuqnb8JdqR_QJAbhW2bQsCt8d)

Constantly updated multiple mock exams with a great number of questions that will help you in better self-assessment. Memorize all your previous NVIDIA Generative AI LLMs (NCA-GENL) exam questions attempts and display all the changes in your results at the end of each NVIDIA NCA-GENL Practice Exam attempt. Users will be able to customize the NCA-GENL practice test software by time or question types. Supported on all Windows-based PCs.

## NVIDIA NCA-GENL Exam Syllabus Topics:

Topic	Details
Topic 1	<ul style="list-style-type: none"> <li>Python Libraries for LLMs: This section of the exam measures skills of LLM Developers and covers using Python tools and frameworks like Hugging Face Transformers, LangChain, and PyTorch to build, fine-tune, and deploy large language models. It focuses on practical implementation and ecosystem familiarity.</li> </ul>

Topic 2	<ul style="list-style-type: none"> <li>• <b>LLM Integration and Deployment:</b> This section of the exam measures skills of AI Platform Engineers and covers connecting LLMs with applications or services through APIs, and deploying them securely and efficiently at scale. It also includes considerations for latency, cost, monitoring, and updates in production environments.</li> </ul>
Topic 3	<ul style="list-style-type: none"> <li>• <b>Data Preprocessing and Feature Engineering:</b> This section of the exam measures the skills of Data Engineers and covers preparing raw data into usable formats for model training or fine-tuning. It includes cleaning, normalizing, tokenizing, and feature extraction methods essential to building robust LLM pipelines.</li> </ul>
Topic 4	<ul style="list-style-type: none"> <li>• <b>Software Development:</b> This section of the exam measures the skills of Machine Learning Developers and covers writing efficient, modular, and scalable code for AI applications. It includes software engineering principles, version control, testing, and documentation practices relevant to LLM-based development.</li> </ul>
Topic 5	<ul style="list-style-type: none"> <li>• <b>Experiment Design</b></li> </ul>
Topic 6	<ul style="list-style-type: none"> <li>• <b>Data Analysis and Visualization:</b> This section of the exam measures the skills of Data Scientists and covers interpreting, cleaning, and presenting data through visual storytelling. It emphasizes how to use visualization to extract insights and evaluate model behavior, performance, or training data patterns.</li> </ul>

>> **NCA-GENL Test Question** <<

## Reliable NCA-GENL Braindumps Free, NCA-GENL Reliable Exam Vce

By taking our NVIDIA NCA-GENL practice exam, which is customizable, you can find and strengthen your weak areas. Additionally, we provide a specialized 24/7 customer support team to assist you with any problems you may run into while using our NVIDIA Generative AI LLMs exam questions. Our NVIDIA NCA-GENL desktop-based practice exam software's ability to be used without an active internet connection is another incredible feature.

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

#### NEW QUESTION # 79

Which tool would you use to select training data with specific keywords?

- **A. Regular expression filter**
- B. JSON parser
- C. Tableau dashboard
- D. ActionScript

**Answer: A**

Explanation:

Regular expression (regex) filters are widely used in data preprocessing to select text data containing specific keywords or patterns. NVIDIA's documentation on data preprocessing for NLP tasks, such as in NeMo, highlights regex as a standard tool for filtering datasets based on textual criteria, enabling efficient data curation. For example, a regex pattern like `.*keyword.*` can select all texts containing "keyword." Option A (ActionScript) is a programming language for multimedia, not data filtering. Option B (Tableau) is for visualization, not text filtering. Option C (JSON parser) is for structured data, not keyword-based text selection.

References:

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

#### NEW QUESTION # 80

Which metric is primarily used to evaluate the quality of the text generated by language models?

- **A. Perplexity**
- B. Recall
- C. Precision

- D. Accuracy

**Answer: A**

Explanation:

Perplexity is the primary metric used to evaluate the quality of text generated by language models, as emphasized in NVIDIA's Generative AI and LLMs course. Perplexity measures how well a language model predicts a sequence of tokens, with lower values indicating better performance, as the model is less

"surprised" by the data. It is calculated as the exponentiated average negative log-likelihood of the tokens in a test set, reflecting the model's ability to assign high probabilities to correct sequences. In generative tasks, perplexity is widely used because it directly assesses the model's fluency and coherence. Option B, Precision, and Option C, Recall, are metrics for classification tasks, not text generation. Option D, Accuracy, is also irrelevant for evaluating generative quality, as it applies to categorical predictions. The course notes:

"Perplexity is a key metric for evaluating language models, measuring how well the model predicts text sequences, with lower perplexity indicating higher-quality generation." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

### NEW QUESTION # 81

Which of the following claims is correct about TensorRT and ONNX?

- A. TensorRT is used for model creation and ONNX is used for model interchange.
- **B. TensorRT is used for model deployment and ONNX is used for model interchange.**
- C. TensorRT is used for model deployment and ONNX is used for model creation.
- D. TensorRT is used for model creation and ONNX is used for model deployment.

**Answer: B**

Explanation:

NVIDIA TensorRT is a deep learning inference library used to optimize and deploy models for high- performance inference, while ONNX (Open Neural Network Exchange) is a format for model interchange, enabling models to be shared across different frameworks, as covered in NVIDIA's Generative AI and LLMs course. TensorRT optimizes models (e.g., via layer fusion and quantization) for deployment on NVIDIA GPUs, while ONNX ensures portability by providing a standardized model representation. Option B is incorrect, as ONNX is not used for model creation but for interchange. Option C is wrong, as TensorRT is not for model creation but optimization and deployment. Option D is inaccurate, as ONNX is not for deployment but for model sharing. The course notes: "TensorRT optimizes and deploys deep learning models for inference, while ONNX enables model interchange across frameworks for portability." References: NVIDIA Building Transformer-Based Natural Language Processing Applications course; NVIDIA Introduction to Transformer-Based Natural Language Processing.

### NEW QUESTION # 82

Which model deployment framework is used to deploy an NLP project, especially for high-performance inference in production environments?

- A. NeMo
- **B. NVIDIA Triton**
- C. HuggingFace
- D. NVIDIA DeepStream

**Answer: B**

Explanation:

NVIDIA Triton Inference Server is a high-performance framework designed for deploying machine learning models, including NLP models, in production environments. It supports optimized inference on GPUs, dynamic batching, and integration with frameworks like PyTorch and TensorFlow. According to NVIDIA's Triton documentation, it is ideal for deploying LLMs for real-time applications with low latency. Option A (DeepStream) is for video analytics, not NLP. Option B (HuggingFace) is a library for model development, not deployment. Option C (NeMo) is for training and fine-tuning, not production deployment.

References:

NVIDIA Triton Inference Server Documentation: <https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/index.html>



