

# NVIDIA

# NCA-GENL

NCA - Generative AI LLMs

For More Information – Visit link below:

<https://www.examsempire.com/>

**Product Version**

1. Up to Date products, reliable and verified.
2. Questions and Answers in PDF Format.



<https://examsempire.com/>

Visit us at: <https://www.examsempire.com/nca-genl>

# Latest Version: 6.0

## Question: 1

Which deployment strategy is best suited for minimizing latency in inference for real-time applications using Generative AI LLMs, considering the scalability and cost factors?

- A. Deploying the model on edge devices with model pruning and quantization.
- B. Using a serverless architecture with autoscaling capabilities in the cloud.
- C. Deploying the model on a dedicated GPU cluster with direct low-latency network connections.
- D. Utilizing hybrid on-premise and cloud deployment to leverage local processing and cloud scaling.

**Answer: C**

Explanation:

Deploying the model on a dedicated GPU cluster with direct low-latency network connections ensures minimal latency and optimal performance for real-time applications. While initially expensive, it provides significant advantages in speed and responsiveness, critical for real-time systems.

## Question: 2

When evaluating the performance of a Large Language Model (LLM) in a production environment with dynamic input distributions, which of the following strategies is the most effective in ensuring robust performance across varying tasks?

- A. Utilize a static benchmark dataset that reflects the common use cases during model training.
- B. Implement continuous learning from user interactions and adjust model parameters in real-time as inputs change.
- C. Rely solely on pre-trained models without considering domain-specific fine-tuning.
- D. Conduct periodic re-training using a static sample of production data without monitoring real-time user feedback.

**Answer: B**

Explanation:

In highly dynamic environments, models must adapt quickly to new data distributions. Employing continuous learning from real-time user interactions ensures that the model performance remains high by quickly accommodating shifts in input patterns.

## Question: 3

You are tasked with implementing a generative AI use case in a highly regulated industry that requires strict auditing and compliance. Which of the following NVIDIA tools and strategies would be most crucial to ensure both efficiency and regulatory compliance when deploying your generative AI model?

- A. Utilize NVIDIA TensorRT to optimize inference while implementing robust logging and monitoring systems to ensure compliance.
- B. Employ NVIDIA Triton Inference Server for scalable deployment and leverage NVIDIA FLARE for federated learning to meet privacy requirements.
- C. Use NVIDIA NeMo for model tuning and leverage pre-trained secure models while implementing an end-to-end encryption and audit trail system.
- D. Deploy NVIDIA Morpheus for cybersecurity monitoring and integrate NVIDIA Gauge for ensuring precise model explainability and detailed logging.

**Answer: D**

Explanation:

Option 4 combines the benefits of multiple NVIDIA tools that focus specifically on critical aspects of cybersecurity monitoring, explainability, and logging which align with auditing and compliance needs in a regulated environment. This makes it the most comprehensive choice for ensuring compliance while deploying generative AI solutions.

### Question: 4

During the training of a Generative AI language model, you encounter an exploding gradient problem. Which of the following approaches is most effective in mitigating this issue?

- A. Increase the learning rate significantly to stabilize gradients.
- B. Use gradient clipping to restrict the magnitude of gradients.
- C. Implement batch normalization to normalize the inputs of layers.
- D. Switch to a simpler, non-recurrent neural architecture.

**Answer: B**

Explanation:

Gradient clipping is a well-established method to deal with exploding gradients, particularly in the context of training deep neural networks like RNNs and LSTMs, by keeping gradients within a specific range.

### Question: 5

In the context of implementing a generative AI solution using NVIDIA's framework, what is the most critical consideration for optimizing end-to-end system performance when deploying a large language model (LLM) in a multi-node GPU setup?

- A. Minimizing data transfer latency by optimizing the inter-GPU communication.
- B. Ensuring uniform memory allocation across all GPU nodes to prevent bottlenecks.

- C. Implementing advanced model parallelism techniques that match the LLM's architecture for efficient resource utilization.
- D. Utilizing dynamic monitoring tools to adapt the LLM's configuration at runtime based on workload variations.

**Answer: C**

Explanation:

In a multi-node GPU setup, optimizing end-to-end performance for large language models primarily involves implementing advanced model parallelism techniques. These techniques help ensure efficient distribution of model components across GPUs, effectively utilizing hardware capabilities and maximizing throughput.

### Question: 6

In the context of generative AI large language models (LLMs), which of the following evaluation techniques is most effective for ensuring robust model performance across diverse languages and contexts?

- A. Cross-lingual transfer evaluation using bilingual translation tasks.
- B. Fine-tuning models on specific language datasets followed by monolingual perplexity measurement.
- C. Using a combination of benchmark datasets like GLUE and multilingual evaluation datasets like XNLI.
- D. Relying solely on syntactic and semantic correctness metrics across a very large monolingual corpus.

**Answer: C**

Explanation:

Evaluating a model's performance across diverse languages and contexts requires both general language understanding and multilingual transfer capability. Utilizing a combination of benchmark datasets like GLUE and multilingual evaluation datasets such as XNLI provides a comprehensive measure of a model's robustness across these dimensions.

### Question: 7

In the context of understanding large language models, which of the following statements best describes the role of attention mechanisms in transformer architectures?

- A. Attention mechanisms encode positional information to differentiate between various token positions.
- B. Attention mechanisms allow the model to focus on specific parts of the input sequence, facilitating the computation of context-dependent word representations.
- C. Attention mechanisms are solely responsible for reducing model size and improving computational efficiency.
- D. Attention mechanisms transform the transformer model's output into a fixed-size vector for classification tasks.

**Answer: B**

Explanation:

In transformer models, attention mechanisms are crucial for allowing the model to weigh the relevance of different words in the input sequence, enabling it to produce more nuanced and contextually relevant word embeddings.

### Question: 8

In the context of training large language models (LLMs), which of the following techniques is most effective for reducing the problem of catastrophic forgetting during continuous training?

- A. Data augmentation with diverse datasets.
- B. Elastic Weight Consolidation (EWC).
- C. Hyperparameter tuning.
- D. Dropout regularization.

**Answer: B**

Explanation:

Catastrophic forgetting occurs when a model trained on new data overwrites knowledge acquired from previous data. Elastic Weight Consolidation (EWC) addresses this by selectively preserving certain learned weights essential for past tasks, thereby minimizing the loss of pre-existing knowledge.

### Question: 9

While deploying a large language model (LLM) using NVIDIA hardware, you observe that the inference throughput is significantly lower than expected. Which of the following is the most likely cause of this issue?

- A. Insufficient GPU memory capacity.
- B. Inefficient batching during inference.
- C. Incorrect CUDA driver version installed.
- D. Insufficient CPU cores allocated for the workload.

**Answer: B**

Explanation:

The most common cause of reduced throughput in deploying LLMs on NVIDIA hardware is inefficient batching, as it fails to fully leverage the parallel processing power of the GPUs. Correcting the batching strategy often leads to improved performance.

### Question: 10

When training a large language model (LLM) on a distributed GPU cluster, which of the following methods is most effective in minimizing the impact of network communication overhead?

- A. Using synchronous data parallelism with a large batch size.
- B. Implementing model parallelism to split the model across multiple GPUs.
- C. Utilizing mixed precision training to reduce the amount of data transferred.
- D. Employing asynchronous gradient updates with gradient compression techniques.

**Answer: D**

Explanation:

Distributed training of LLMs often involves significant communication overhead among GPUs, which can bottleneck scaling. Asynchronous updates mitigate this by allowing non-blocking updates, and gradient compression reduces the data size being communicated. This combination is generally effective at minimizing the impact of network communication overhead.

**Thank You for Trying Our Product**

**Special 16 USD Discount Coupon: NSZUBG3X**

**Email:** [support@examsempire.com](mailto:support@examsempire.com)

**Check our Customer Testimonials and ratings  
available on every product page.**

**Visit our website.**

**<https://examsempire.com/>**