Boost up Your Certification Score

NVIDIA
NCP-GENL
Generative AI LLMs



For More Information – Visit link below:

https://www.examsboost.com/

Product Version

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

Visit us at: https://www.examsboost.com/test/ncp-genl

Latest Version: 6.0

Question: 1

When deploying a 13B parameter model across 4 A100 40GB GPUs for inference, the team faces OOM errors despite theoretical calculations showing sufficient memory.
Which TWO strategies would most effectively resolve this issue? Pick the 2 correct responses below

A. Apply activation checkpointing, allowing intermediate activations to be recomputed on demand instead of being stored, thus reducing GPU memory requirements.
B. Enable NVIDIA Multi-Instance GPU (MIG) features to partition each A100 GPU into multiple, smaller instances to share resources more flexibly.
C. Increase the server's system RAM to provide additional swap space for GPU memory overflow during inference.
D. Distribute the model layers evenly across GPUs using model parallelism and optimize the pipeline scheduling to balance memory and computation.

Answer: A,D

Question: 2

A team is developing a language translation system and must choose between a Recurrent Neural Network (RNN) with attention and a Transformer model.
Which TWO statements correctly describe the main differences between these architectures? Pick the 2 correct responses below

A. Transformers are slower at processing long documents, while RNNs process their inputs in parallel, enabling faster training and better handling of long-range dependencies.
B. Transformers can model dependencies between any parts of the input sequence regardless of their distance, while RNNs struggle with very long sequences due to vanishing gradients.
C. The RNNs and Transformers process data sequentially, making them inefficient for long documents. However, Transformers show better contextual comprehension.
D. RNNs are slower at processing long documents, while Transformers process their inputs in parallel, enabling faster training and better handling of long-range dependencies.

Answer: B,D

Question: 3

When optimizing throughput for a 3B parameter model on A100 GPUs, profiling shows 70% memory utilization but only 50% SM activity.
Which TWO techniques would improve throughput? Pick the 2 correct responses below

A. Use smaller sequence lengths to process more samples per batch
B. Enable torch.compile() or TensorRT optimization for kernel fusion and better SM utilization
C. Increase batch size until memory utilization reaches 90-95% for better GPU saturation
D. Reduce model precision from FP16 to INT8 to fit larger batches
E. Implement gradient accumulation to simulate larger batch sizes without increasing memory

Answer: B,C

Question: 4

When combining automated benchmark results with human-in-the-loop evaluation, which approaches optimize the balance between scalability and assessment quality? Pick the 2 correct responses below

A. Stratified sampling for human evaluation with focus on edge cases and automated metric disagreements
B. Automated evaluation only without human oversight to maximize efficiency and processing speed
C. Random human evaluation without consideration for automated results or systematic sampling strategies
D. Complete human evaluation of all samples for maximum accuracy regardless of time and cost constraints
E. Active learning approaches to identify samples requiring human judgment based on model uncertainty

Answer: A,E

Question: 5

A government agency is deploying an LLM for citizen services (benefits eligibility, tax questions, immigration status).
Requirements:
• Must serve all citizens equitably
• Audit trail for all decisions
• Ability to correct errors rapidly
• Compliance with accessibility standards
The model performs well in testing, but stakeholders worry about real-world fairness.
Which deployment strategy best ensures responsible Al practices?

A. Phased rollout starting with low-risk queries, expanding based on fairness metrics from each phase

B. Parallel deployment with human agents handling sensitive cases while the LLM handles routine queries despite model biases
C. Full deployment with a prominent feedback mechanism and weekly bias analysis of user interactions
D. Blue-green deployment with ability to instantly rollback to previous versions if bias is detected

Answer: A

Thank You for Trying Our Product

For More Information – Visit link below:

https://www.examsboost.com/

15 USD Discount Coupon Code:

G74JA8UF

FEATURES

- ✓ 90 Days Free Updates
- ✓ Money Back Pass Guarantee
- ✓ Instant Download or Email Attachment
- ✓ 24/7 Live Chat Support
- ✓ PDF file could be used at any Platform
- ✓ 50,000 Happy Customer

