NCA GENERATIVE AI LLM (NCA-GENL) PRACTICE EXAM (SAMPLE)

STUDY GUIDE



SAMPLE STUDY GUIDE WITH LIMITED QUESTIONS

VISIT US ONLINE FOR THE FULL VERSION STUDY GUIDE

https://ncagenl.examzify.com

Copyright © 2026 by Examzify - A Kaluba Technologies Inc. product.

ALL RIGHTS RESERVED.

No part of this book may be reproduced or transferred in any form or by any means, graphic, electronic, or mechanical, including photocopying, recording, web distribution, taping, or by any information storage retrieval system, without the written permission of the author.

Notice: Examzify makes every reasonable effort to obtain accurate, complete, and timely information about this product from reliable sources.

SAMPLE

Table of Contents

Copyright 1

Table of Contents 2

Introduction 3

How to Use This Guide 4

Questions 5

Answers 8

Explanations 10

Next Steps 16

SAMPLE

Introduction

Preparing for a certification exam can feel overwhelming, but with the right tools, it becomes an opportunity to build confidence, sharpen your skills, and move one step closer to your goals. At Examzify, we believe that effective exam preparation isn't just about memorization, it's about understanding the material, identifying knowledge gaps, and building the test-taking strategies that lead to success.

This guide was designed to help you do exactly that.

Whether you're preparing for a licensing exam, professional certification, or entry-level qualification, this book offers structured practice to reinforce key concepts. You'll find a wide range of multiple-choice questions, each followed by clear explanations to help you understand not just the right answer, but why it's correct.

The content in this guide is based on real-world exam objectives and aligned with the types of questions and topics commonly found on official tests. It's ideal for learners who want to:

- Practice answering questions under realistic conditions,
- Improve accuracy and speed,
- Review explanations to strengthen weak areas, and
- Approach the exam with greater confidence.

We recommend using this book not as a stand-alone study tool, but alongside other resources like flashcards, textbooks, or hands-on training. For best results, we recommend working through each question, reflecting on the explanation provided, and revisiting the topics that challenge you most.

Remember: successful test preparation isn't about getting every question right the first time, it's about learning from your mistakes and improving over time. Stay focused, trust the process, and know that every page you turn brings you closer to success.

Let's begin.

SAMPLE

How to Use This Guide

This guide is designed to help you study more effectively and approach your exam with confidence. Whether you're reviewing for the first time or doing a final refresh, here's how to get the most out of your Examzify study guide:

1. Start with a Diagnostic Review

Skim through the questions to get a sense of what you know and what you need to focus on. Your goal is to identify knowledge gaps early.

2. Study in Short, Focused Sessions

Break your study time into manageable blocks (e.g. 30 – 45 minutes). Review a handful of questions, reflect on the explanations.

3. Learn from the Explanations

After answering a question, always read the explanation, even if you got it right. It reinforces key points, corrects misunderstandings, and teaches subtle distinctions between similar answers.

4. Track Your Progress

Use bookmarks or notes (if reading digitally) to mark difficult questions. Revisit these regularly and track improvements over time.

5. Simulate the Real Exam

Once you're comfortable, try taking a full set of questions without pausing. Set a timer and simulate test-day conditions to build confidence and time management skills.

6. Repeat and Review

Don't just study once, repetition builds retention. Re-attempt questions after a few days and revisit explanations to reinforce learning. Pair this guide with other Examzify tools like flashcards, and digital practice tests to strengthen your preparation across formats.

There's no single right way to study, but consistent, thoughtful effort always wins. Use this guide flexibly, adapt the tips above to fit your pace and learning style. You've got this!

SAMPLE

Questions

SAMPLE

1. Which activation function is frequently avoided in certain scenarios due to the vanishing gradient problem?

A. Sigmoid

B. Tanh

C. ReLU

D. GeLu

2. What is Pipeline Parallelism primarily used for?

A. Processing data in parallel through shards

B. Reducing unnecessary weight updates

C. Managing data storage effectively

D. Integrating with cloud services

3. What technique involves distributed computing for training large language models?

A. Model Parallelism

B. Data Parallelism

C. Vertical Scaling

D. Task Scheduling

4. Which technique is known for ensuring data privacy through federated learning?

A. Hotfix Deployment

B. Nvidia FLARE

C. Holistic Model Compression

D. MoE-FT

5. Which implementation in Kubernetes limits deployment to a single cluster?

A. Pod Affinity

B. Node Affinity

C. Cluster Autoscaler

D. Service Mesh

SAMPLE

6. What key paper from 2017 introduced the principles of modern large language model architecture?

A. Attention is All You Need

B. The Limitations of Deep Learning

C. Understanding LSTM Networks

D. The Evolution of AI Models

7. What does INT4 Quantization primarily focus on?

A. Speeding up graphical applications

B. Reducing model size

C. Maintaining accuracy while reducing complexity

D. Enhancing user experience in systems

8. What is the full name of the ELU activation function?

A. Exponential Linear Unit

B. Enhanced Linear Unit

C. Elastic Linear Unit

D. Exponential Limit Unit

9. Which activation function is defined as a hyperbolic tangent and is used commonly in neural networks?

A. ReLU

B. Tanh

C. Sigmoid

D. GeLu

10. In the context of AI models, what does model pruning specifically optimize?

A. Model Complexity

B. Training Time

C. Inference Speed

D. Data Variability

SAMPLE

Answers

SAMPLE

1. A
2. A
3. A
4. B
5. B
6. A
7. C
8. A
9. B
10. A

SAMPLE

Explanations

SAMPLE

1. Which activation function is frequently avoided in certain scenarios due to the vanishing gradient problem?

A. Sigmoid

B. Tanh

C. ReLU

D. GeLu

The sigmoid activation function is often avoided in certain scenarios because it can lead to the vanishing gradient problem. The vanishing gradient problem occurs when gradients become very small during backpropagation, making it difficult for the network to learn. This is particularly evident in deep networks where layers can be stacked and the gradients diminish as they propagate backward through subsequent layers. The sigmoid function maps input values to a range between 0 and 1, and its output saturates at these extremes. When activations are very close to either 0 or 1, the gradients of the sigmoid function become nearly zero. As a result, updates to the neural network weights for neurons that are saturated do not happen effectively during training, slowing down the overall learning process and making it challenging for the model to converge. In contrast, other activation functions such as Tanh, ReLU, and GeLU have properties that help mitigate the vanishing gradient issue, especially in deeper architectures. Therefore, in many deep learning applications, the sigmoid function is less preferred in hidden layers, though it is still commonly used in the output layer for binary classification tasks. Understanding the nuances of activation functions like sigmoid is essential for effective neural network design and optimization.

2. What is Pipeline Parallelism primarily used for?

A. Processing data in parallel through shards

B. Reducing unnecessary weight updates

C. Managing data storage effectively

D. Integrating with cloud services

Pipeline Parallelism is primarily used for processing data in parallel through multiple stages of computation, which allows different parts of a model to be trained simultaneously. This technique enhances computational efficiency and speeds up the training process by dividing the model into segments, each responsible for a specific part of the computation. In the context of large models, Pipeline Parallelism enables the training of deep learning models that may be too large to fit into the memory of a single GPU or machine. By splitting the model into smaller chunks and processing these chunks in tandem across available compute resources, it effectively utilizes the hardware, thus optimizing performance. While the other choices may pertain to aspects of model training and data management, they do not specifically describe the function or application of Pipeline Parallelism. For instance, reducing unnecessary weight updates pertains more to techniques that streamline learning or optimizations in training algorithms, and managing data storage effectively or integrating with cloud services are broader aspects that do not specifically leverage the core principles of Pipeline Parallelism.

SAMPLE

3. What technique involves distributed computing for training large language models?

A. Model Parallelism

B. Data Parallelism

C. Vertical Scaling

D. Task Scheduling

The technique that involves distributed computing for training large language models is model parallelism. This approach is particularly beneficial when a single model is too large to fit into the memory of a single machine. With model parallelism, the model is split across multiple devices (such as GPUs or machines), allowing different parts of the model to be processed simultaneously. This is essential for efficiently handling very large models, as it enables the sharing of the computational load, leveraging the strengths of distributed resources. While data parallelism, another commonly used technique in training large models, focuses on distributing training data across multiple processors to perform the same operations, model parallelism specifically addresses the challenges posed by large models themselves. Vertical scaling refers to increasing the power of a single machine (e.g., adding more memory or CPU cores), which is not as effective for extremely large models. Task scheduling, while important for managing when and how tasks are executed across multiple systems, does not directly relate to the concept of distributing model components for training.

4. Which technique is known for ensuring data privacy through federated learning?

A. Hotfix Deployment

B. Nvidia FLARE

C. Holistic Model Compression

D. MoE-FT

Federated learning is a technique that allows machine learning models to be trained across multiple decentralized devices or servers holding local data samples, without exchanging them. This approach is particularly significant for ensuring data privacy, as it enables models to learn from data without exposing the raw data itself. Nvidia FLARE (Federated Learning Application Runtime Environment) is specifically designed to facilitate federated learning by providing a framework that supports secure and efficient training of models in a distributed environment. It emphasizes privacy by keeping the training data on local devices and only sharing model updates or gradients, thereby minimizing the risk of personal data exposure. This means that sensitive information stays on the user's device, aligning with privacy regulations and protecting individual data. Other techniques mentioned do not directly pertain to the principle of federated learning and the specific emphasis on data privacy in the same way. Hotfix deployment refers to quick patches in software systems, holistic model compression focuses on optimizing model size without particular regard for privacy, and MoE-FT (Mixture of Experts - Fine Tuning) deals with model efficiency rather than data privacy aspects. Thus, Nvidia FLARE stands out as the key technique relevant to ensuring data privacy through federated learning.

SAMPLE

5. Which implementation in Kubernetes limits deployment to a single cluster?

A. Pod Affinity

B. Node Affinity

C. Cluster Autoscaler

D. Service Mesh

The implementation that limits deployment to a single cluster is Node Affinity. Node Affinity is a set of rules used by Kubernetes to control which nodes your pods can be scheduled on, based on labels and conditions set on the nodes. When a pod is configured with Node Affinity, it ensures that the pod will only be deployed on nodes that meet certain criteria within the same Kubernetes cluster, thereby restricting deployment to that particular cluster. In contrast, Pod Affinity pertains to how pods can be scheduled in relation to other pods, determining their placement based on labels; it doesn't restrict the deployment to a single cluster. The Cluster Autoscaler focuses on automatically adjusting the size of your cluster based on the current resource demand but doesn't impose restrictions on deployment across multiple clusters. Service Mesh refers to a dedicated infrastructure layer for managing service-to-service communication, which also does not limit deployment to a single cluster. Therefore, Node Affinity is correctly identified as the mechanism that enforces pod scheduling within a specific Kubernetes cluster.

6. What key paper from 2017 introduced the principles of modern large language model architecture?

A. Attention is All You Need

B. The Limitations of Deep Learning

C. Understanding LSTM Networks

D. The Evolution of AI Models

The key paper from 2017 that introduced the principles of modern large language model architecture is "Attention is All You Need." This groundbreaking work introduced the Transformer architecture, which fundamentally changed the way natural language processing tasks are approached. The fundamental innovation in the paper is the self-attention mechanism, which allows models to weigh the importance of different words relative to each other, irrespective of their distance in the text. This allows for more nuanced understanding and generation of language, enabling significantly better performance on various language tasks compared to previous models that relied primarily on recurrent architectures. The implications of this architecture extend to enabling parallelization in training processes, managing longer contexts more effectively, and leading to the development of subsequent models like BERT and GPT, which are built upon the principles set out in this paper. Other options do not refer to the foundational architectural changes that propelled language models to their current state of sophistication. For example, "The Limitations of Deep Learning" focuses on identifying challenges within the field, while "Understanding LSTM Networks" discusses a specific type of recurrent neural network, and "The Evolution of AI Models" may generally cover the progression of models without focusing on the specific architectural innovation of the Transformer. Thus, "Attention is All You Need

SAMPLE

7. What does INT4 Quantization primarily focus on?

A. Speeding up graphical applications

B. Reducing model size

C. Maintaining accuracy while reducing complexity

D. Enhancing user experience in systems

INT4 quantization primarily focuses on maintaining accuracy while reducing complexity within machine learning models. This method compresses the size of the model by converting the weights from floating-point representations to a lower-bit representation, specifically 4 bits. While reducing model size is an inherent benefit of this quantization technique, the primary aim is to ensure that the model's performance does not significantly degrade as a result of this compression. By using a lower precision format like INT4, the model can operate with less memory and computational resources, making it more efficient without sacrificing the accuracy needed for its tasks. This balance between reducing complexity (in terms of memory usage and required computational power) and preserving the effectiveness of the model aligns with the challenges in deploying machine learning models, especially in resource-constrained environments. The other options primarily address aspects that are secondary to the main goal of INT4 quantization. While speeding up graphical applications or enhancing user experience may relate to performance improvements, they do not capture the core intent of INT4 quantization as accurately as maintaining model accuracy while reducing complexity does.

8. What is the full name of the ELU activation function?

A. Exponential Linear Unit

B. Enhanced Linear Unit

C. Elastic Linear Unit

D. Exponential Limit Unit

The full name of the ELU activation function is Exponential Linear Unit. This activation function is widely used in neural networks due to its ability to alleviate the vanishing gradient problem and its smooth, differentiable nature. The ELU combines benefits from both linear and non-linear activation functions, allowing for faster learning and better performance in deep learning models compared to traditional activations like ReLU. The "Exponential" part refers to the way the negative inputs are transformed, which follows an exponential curve, providing a non-zero output for negative values and helping to maintain the mean of the activations nearer to zero. This capability of producing negative outputs, unlike some other activation functions, contributes to effective training dynamics in deeper networks. This characteristic is key in enabling the model to learn more complex patterns. Other options do not represent established names for the ELU function within the context of machine learning, which is why they are not correct.

SAMPLE

9. Which activation function is defined as a hyperbolic tangent and is used commonly in neural networks?

A. ReLU

B. Tanh

C. Sigmoid

D. GeLu

The hyperbolic tangent function, commonly referred to as "tanh," is an activation function that is widely used in neural networks due to its properties that can significantly improve the model's learning capability. Tanh returns values between -1 and 1, which helps to center the data. This centering can lead to faster convergence during training compared to other activation functions that are not zero-centered, such as the sigmoid function. Using tanh can also mitigate issues related to saturated outputs. As the outputs lie within the range of -1 to 1, tanh tends to have steeper gradients compared to the sigmoid function, whose outputs range from 0 to 1. This can help in propagating gradients during backpropagation, making it less likely for the model to suffer from the vanishing gradient problem compared to those that use sigmoid. In practice, tanh is particularly beneficial in hidden layers of a neural network because it allows for more expressive power by capturing negative values, thus enabling the network to learn more complex patterns. Its mathematical formulation, defined as (e^x - e^-x) / (e^x + e^-x), ensures that the output can represent a wider variety of potential responses useful in decision-making processes in AI

10. In the context of AI models, what does model pruning specifically optimize?

A. Model Complexity

B. Training Time

C. Inference Speed

D. Data Variability

Model pruning specifically optimizes model complexity by reducing the number of parameters in a neural network while maintaining its performance. The primary objective of pruning is to streamline the model, making it less resource-intensive and easier to deploy in real-world applications. This is achieved by removing less important weights or neurons from the model, which simplifies the architecture without significantly degrading its predictive capabilities. By focusing on reducing the number of parameters, pruning can lead to a more efficient model that requires less memory and computational power, which is essential for deploying AI in environments with limited resources. Although pruning can indirectly influence aspects like training time and inference speed, its main goal is to reduce the overall complexity of the model. Thus, the emphasis on optimizing model complexity defines what pruning aims to achieve.

SAMPLE

Next Steps

Congratulations on reaching the final section of this guide. You've taken a meaningful step toward passing your certification exam and advancing your career.

As you continue preparing, remember that consistent practice, review, and self-reflection are key to success. Make time to revisit difficult topics, simulate exam conditions, and track your progress along the way.

If you need help, have suggestions, or want to share feedback, we'd love to hear from you. Reach out to our team at hello@examzify.com.

Or visit your dedicated course page for more study tools and resources:

https://ncagenl.examzify.com

We wish you the very best on your exam journey. You've got this!

SAMPLE