HP.HPE2-B08.v2026-07-20.q37

Exam Code:	HPE2-B08
Exam Name:	HPE Private Cloud AI Solutions
Certification Provider:	HP
Free Question Number:	37
Version:	v2026-07-20
# of views:	105
# of Questions views:	379
https://www.freecram.net/torrent/HP.HPE2-B08.v2026-07-20.q37.html	

NEW QUESTION: 1
An architect is designing a distributed training solution using HPE Private Cloud AI. The solution involves multiple HPE ProLiant DL380a servers, each with four NVIDIA H100 GPUs.
Review the logical network topology:
```
[HPE ProLiant Server 1] \
- GPU 0 <--> GPU 1 (Internal)
- GPU 2 <--> GPU 3 (Internal)
\
[NVIDIA Spectrum Switch (RoCE)] <--> [HPE GreenLake for File Storage]
/
[HPE ProLiant Server 2] /
- GPU 0 <--> GPU 1 (Internal)
- GPU 2 <--> GPU 3 (Internal)
```
Which interconnect technology facilitates the highest-bandwidth communication path labeled "(Internal)" for scaling the training job across the GPUs within a single server?
A. RDMA over Converged Ethernet (RoCE)
B. PCI Express (PCIe) Gen5
C. GPUDirect Storage (GDS)
D. NVIDIA NVLink Bridge
Answer: (SHOW ANSWER)

NEW QUESTION: 2
What is the primary purpose of using the "Private Cloud Smart Template" within HPE One Config Advanced (OCA) when quoting an HPE Private Cloud AI solution?
A. To provide a free trial license for the VMware and NVIDIA software.

B. To allow for complete customization of every component, including unsupported hardware.
C. To automatically generate a network diagram and rack elevation for the customer.
D. To ensure a fast, easy, and validated process for generating a complete and accurate Bill of Materials (BOM) for a pre-integrated solution.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 3
A company is implementing a RAG-based chatbot using HPE Private Cloud AI. To ensure the chatbot provides safe and appropriate responses, the development team needs to implement guardrails to prevent it from discussing off-topic subjects and to block it from using harmful language.
Which specific toolkit within the NVIDIA NeMo framework is designed for this purpose?
```
NVIDIA NeMo Framework Components:
1. NeMo Curator
2. NeMo Customizer
3. NeMo Evaluator
4. NeMo Retriever
5. NeMo Guardrails
```
A. NeMo Customizer
B. NeMo Curator
C. NeMo Guardrails
D. NeMo Retriever
E. NeMo Evaluator
Answer: C (LEAVE A REPLY)

NEW QUESTION: 4
A customer who is an "AI Beginner" wants to start an AI inferencing project at their edge locations.
Their goal is to analyze security camera feeds to help prevent theft. They have a limited budget and IT staff at the edge sites.
Which HPE AI solution is the most appropriate to position for this specific scenario?
A. HPE Cray systems
B. HPE Private Cloud AI with NVIDIA
C. NVIDIA DGX Systems
D. An HPE AI Services - Transformation Workshop
E. AI-optimized HPE ProLiant DL servers
Answer: E (LEAVE A REPLY)

NEW QUESTION: 5

A customer's ML Engineer states, "We need to deploy our trained models, and our top priority is simplifying the process. We want to treat our models like cattle, not pets-packaging them into standardized, optimized containers that we can deploy and scale easily via an API." This statement describes the primary benefit of which software component in the HPE Private Cloud AI stack?

A. HPE Data Fabric

B. NVIDIA NIM (NVIDIA Inference Microservices)

C. Apache Airflow

D. JupyterLab

Answer: (SHOW ANSWER)

NEW QUESTION: 6

An architect is presenting to a customer who is an 'Early AI User'. The customer is unsure about the differences between their two main project ideas.

Project 1: Create an internal chatbot that can answer HR policy questions by accessing the live employee handbook stored on a shared drive.

Project 2: Teach a general-purpose chatbot to adopt the company's formal communication style and tone for drafting official press releases.

How should the architect categorize the primary AI task for each project? (Choose 2.)

A. Project 1 is a RAG (Retrieval-Augmented Generation) workload.

B. Project 1 is a model training workload.

C. Project 2 is a fine-tuning workload.

D. Project 2 is a classic inferencing workload with no model customization.

Answer: (SHOW ANSWER)

NEW QUESTION: 7

A customer needs a solution for two primary workloads: large-scale model training and real-time inference. They have a team of data scientists who are constantly developing new models and a separate operations team that deploys and manages these models in production.

Which statement best describes how the different stakeholders would interact with the HPE Private Cloud AI solution?

A. The data scientists would primarily use tools within HPE AI Essentials (like JupyterLab and MLFlow) for model development, while the operations team would use NVIDIA NIMs to deploy the finished models.

B. The entire process for all stakeholders is managed through the server's iLO interface.

C. The data scientists would use NVIDIA NIMs to train the models, and the operations team would use HPE Data Fabric to serve them.

D. Both the data scientists and the operations team would use the HPE Intelligent Configurator to manage the models.

Answer: (SHOW ANSWER)

NEW QUESTION: 8
The three pre-defined sizes (Small, Medium, Large) of HPE Private Cloud AI are designed to address different primary workloads. Match the configuration size to its intended primary workload.
*Configuration Size:
1. Small
2. Medium
3. Large
*Primary Workload:
a. AI Inferencing with RAG and large-scale Fine-Tuning
b. AI Inferencing
c. AI Inferencing with RAG
A. 1-a, 2-b, 3-c
B. 1-c, 2-a, 3-b
C. 1-b, 2-a, 3-c
D. 1-b, 2-c, 3-a
Answer: (SHOW ANSWER)

NEW QUESTION: 9
A customer is expanding their HPE Private Cloud AI "Medium" configuration to support a new generative AI inferencing workload. They are concerned about network congestion and latency, as the AI workload is known to generate large, sudden bursts of traffic between the compute nodes and the storage system.
The solution uses NVIDIA Spectrum SN4700M switches for the AI interconnect.
Which feature of these switches is specifically designed to handle bursty traffic and prevent packet loss in a lossless Ethernet fabric?
A. The ability to route traffic based on application-layer metadata.
B. A fully shared buffer architecture that can dynamically absorb traffic bursts from any port.
C. Support for Fibre Channel over Ethernet (FCoE) encapsulation.
D. Integrated Silicon Root of Trust to validate the switch firmware integrity.
Answer: (SHOW ANSWER)

NEW QUESTION: 10
What is the primary architectural advantage of the NVIDIA Grace Hopper Superchip (e.g., GH200) for large-scale AI workloads?

A. It is the first NVIDIA GPU to feature fourth-generation Tensor Cores for enhanced matrix calculations.
B. It combines a CPU and a GPU on a single superchip, connected by a high-speed, low-latency NVLink-C2C interconnect.
C. It replaces the need for server memory (DRAM) by using the GPU's global memory exclusively.

D. It uses on-chip encryption to create a confidential computing environment for the CPU.
Answer: (SHOW ANSWER)

NEW QUESTION: 11
An architect is designing an infrastructure solution for an AI workload that involves processing massive datasets for training. The goal is to minimize data transfer latency between the storage system and the GPUs in the compute nodes. The architect wants to enable the NVIDIA GPUs to fetch data directly from the NVMe storage array, bypassing the server's CPU and main memory.
Review the proposed architectural components:
```
- Compute Nodes: HPE ProLiant DL380a Gen11 with NVIDIA H100 GPUs
- Storage: HPE GreenLake for File Storage
- Interconnect: Ethernet with RoCE support
```
Which technology must be enabled and properly configured across these components to achieve this direct GPU-to-storage data path? (Choose 2.)
A. NVLink
B. RDMA over Converged Ethernet (RoCE)
C. Multi-Instance GPU (MIG)
D. Confidential Computing
E. GPUDirect Storage (GDS)
Answer: (SHOW ANSWER)

NEW QUESTION: 12
A customer is expanding their HPE Private Cloud AI "Medium" configuration to support a new generative AI inferencing workload. They are concerned about network congestion and latency, as the AI workload is known to generate large, sudden bursts of traffic between the compute nodes and the storage system.
The solution uses NVIDIA Spectrum SN4700M switches for the AI interconnect.
Which feature of these switches is specifically designed to handle bursty traffic and prevent packet loss in a lossless Ethernet fabric?
A. A fully shared buffer architecture that can dynamically absorb traffic bursts from any port.
B. Integrated Silicon Root of Trust to validate the switch firmware integrity.
C. Support for Fibre Channel over Ethernet (FCoE) encapsulation.
D. The ability to route traffic based on application-layer metadata.
Answer: (SHOW ANSWER)

NEW QUESTION: 13
An architect is explaining the HPE Private Cloud AI configurations to a customer.
What is the key differentiator between the "Large" configuration and the "Small" and "Medium" configurations in terms of hardware?

A. The Large configuration uses NVIDIA H100 NVL GPUs, while the Small and Medium configurations use NVIDIA L40S GPUs.
B. The Large configuration is the only one that includes HPE GreenLake for File Storage.
C. The Large configuration is deployed in a single rack, while the Small and Medium require two racks.
D. The Large configuration uses AMD processors, while the Small and Medium configurations use Intel processors.
Answer: (SHOW ANSWER)

NEW QUESTION: 14
An architect is in a discovery call with a customer who describes their project: "Our primary goal is to take our massive, proprietary dataset of chemical compound interactions and continuously update our foundational AI model's internal parameters to create a new, specialized model for drug discovery. This process runs 24/7 on a large GPU cluster." How should the architect classify this primary AI workload?
A. AI Inferencing
B. Retrieval-Augmented Generation (RAG)
C. Edge Computing
D. AI Model Training / Fine-tuning
Answer: (SHOW ANSWER)

NEW QUESTION: 15
A customer at the 'Early AI User' maturity stage wants to begin using generative AI. Their primary concern is the complexity of deploying and managing the AI software stack. They are looking for a solution that accelerates their journey by providing a pre-integrated, enterprise-supported software platform for both development and deployment.
How do the software components of HPE Private Cloud AI address this customer's main challenge?
(Select all that apply.)
```
Customer Profile:
- Maturity: Early AI User
- Main Challenge: Complexity of AI software stack
- Goal: Accelerate adoption of generative AI
```
A. HPE AI Essentials provides a unified, self-service platform with curated open-source tools, eliminating integration complexity.
B. The software stack is limited to only HPE-developed tools to ensure a single point of contact.
C. The solution requires the customer to manually integrate all software components, which promotes learning.

D. NVIDIAAI Enterprise provides enterprise-grade support, security, and API stability for the entire AI software stack.
E. NVIDIA NIMs drastically simplify the deployment of pre-trained models into production-ready microservices.
Answer: (SHOW ANSWER)

NEW QUESTION: 16
After an architect selects the "HPE Private Cloud AI - Large - Expanded" Smart Template in OCA, they see it includes a significant number of "HPE Factory Express Complex Unit of SVC" services.
What is the purpose of these bundled services?
A. To provide on-site training for the customer's data science team.
B. To provide a credit for future software purchases on the HPE GreenLake cloud.
C. To allow the customer to exchange the hardware for the next generation at no cost.
D. To cover the pre-integration and configuration of the entire solution at an HPE factory before shipment.
Answer: (SHOW ANSWER)

NEW QUESTION: 17
An AI development team is training a large language model that exceeds the memory capacity of a single GPU. To continue training, they need to distribute the workload across multiple GPUs within the same server.
Which combination of HPE server and NVIDIA technology is specifically designed to provide a high- bandwidth, direct communication path between GPUs, bypassing the PCIe bus for improved performance in multi-GPU training?
A. HPE Alletra Storage MP with RoCE-capable network adapters
B. HPE ProLiant DL325 Gen11 server with NVIDIA L4 GPUs
C. HPE ProLiant DL380a Gen12 server with an NVIDIA NVLink Bridge
D. HPE ProLiant DL380a Gen11 server with GPUDirect Storage (GDS)
Answer: (SHOW ANSWER)

NEW QUESTION: 18

A hospital is developing an AI application to automatically detect specific anomalies in medical images like X-rays and MRIs. The task requires the model to learn and identify complex spatial patterns, such as the shapes and textures of tissues and potential tumors.
Which type of neural network architecture is specifically designed for and best suited to this kind of image analysis task? (Select all that apply.)
A. A model architecture that includes pooling layers to reduce spatial dimensions
B. A Convolutional Neural Network (CNN)
C. A basic, fully connected Artificial Neural Network (ANN)
D. A transformer-based model designed for Natural Language Processing (NLP)
E. A model architecture that includes convolutional layers for feature extraction
Answer: (SHOW ANSWER)

NEW QUESTION: 19
An IT director for a regional retail chain tells you they are "doing AI." Upon further questioning, you learn they have one data scientist who has built a single, experimental sales forecasting model as a proof-of-concept (PoC). The project lacks clear KPIs for success and there is no formal strategy for how to productionize it or what to do next.
How would you classify this customer's AI maturity level?
A. AI Pro
B. AI Beginner
C. AI Curious
D. Deployer of AI at scale
Answer: (SHOW ANSWER)

NEW QUESTION: 20
An architect is building a final configuration for an HPE Private Cloud AI solution using One Config Advanced (OCA). After selecting the "HPE Private Cloud AI - Medium Expanded" Smart Template, they review the generated Bill of Materials (BOM).
Which components are characteristic of the Medium Expanded configuration that the architect should expect to see in the OCA-generated BOM? (Select all that apply.)

A. 2x HPE Aruba Networking CX 6300M Switches
B. 2x NVIDIA SN4700M Switches
C. 1x HPE ProLiant DL380a Gen11 Server
D. 4x HPE ProLiant DL380a Gen11 Servers
E. 16x NVIDIA L40S GPUs
F. 16x NVIDIA H100 NVL GPUs
Answer: (SHOW ANSWER)

NEW QUESTION: 21
A customer wants to deploy a turnkey private cloud for a variety of generative AI workloads, including RAG-based chatbots and some model fine-tuning. One of their key IT stakeholders is the data engineer.

Which specific challenge for a data engineer is directly addressed by the HPE Data Fabric component within HPE Private Cloud AI?
A. The difficulty of managing GPU cluster utilization for training jobs.
B. The complexity of writing Python code in a Jupyter Notebook.
C. The effort required to create and manage data pipelines and unify diverse, siloed data sources (files, objects, tables) into a single accessible platform.
D. The high cost of NVIDIA AI Enterprise software licenses.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 22
You are in a meeting with the VP of Manufacturing for a large industrial company. She describes their AI program as follows:
"We have a formal Center of Excellence for AI that reports directly to the CTO. Our strategy is to scale our successful predictive maintenance models across all 25 of our global plants. We have standardized on a consistent cloud and data governance strategy for all AI projects and have clear ROI targets for reducing downtime." Based on this description, which AI maturity level best describes this customer?
A. Early AI user
B. AI Pro
C. AI Curious
D. AI Beginner
Answer: (SHOW ANSWER)

NEW QUESTION: 23
A customer is building a large-scale AI training cluster using HPE Private Cloud AI.
They have two main requirements for the storage solution:
1. It must provide extremely high throughput and low latency to feed data to a large cluster of NVIDIA H100 GPUs.
2. It must be able to scale performance and capacity independently to accommodate future growth without costly overprovisioning.
The architect is proposing HPE GreenLake for File Storage.
Which architectural features of this solution directly address the customer's requirements? (Select all that apply.)
```
Customer Requirements:
1. High performance for AI training workloads.
2. Independent scaling of performance and capacity.
```
A. It uses a disaggregated architecture, allowing controller nodes (performance) and storage shelves (capacity) to be scaled separately.
B. It is managed exclusively through a command-line interface (CLI) for maximum control.

C. It includes HPE Data Fabric, providing a global namespace across the worker nodes' local drives.
D. It supports GPUDirect Storage (GDS) over a high-speed NVMe-oF (RoCE) network, enabling a direct, low-latency path between storage and GPUs.
E. It relies on traditional hard disk drives (HDDs) for cost-effective capacity, with a small flash tier for caching.
Answer: (SHOW ANSWER)

NEW QUESTION: 24
A large enterprise is adopting HPE Private Cloud AI and needs to accelerate the development of several generative AI applications. Their data science team wants to leverage pre-trained foundation models from NVIDIA but needs to customize them for specific business tasks like contract summarization and internal policy Q&A.
Which NVIDIA AI Enterprise software framework provides a comprehensive, end-to-end toolkit for curating data, customizing models using techniques like PEFT, and implementing guardrails for safe deployment?
```
Customer Goal:
- Accelerate development of custom generative AI apps
- Utilize pre-trained foundation models
- Require tools for data prep, model customization, and safety
```
A. NVIDIA RAPIDS
B. NVIDIA Triton Inference Server
C. NVIDIA NeMo
D. NVIDIA NIM (NVIDIA Inference Microservices)
Answer: (SHOW ANSWER)

NEW QUESTION: 25
A customer wants to build a configuration in One Config Advanced (OCA) for the HPE Private Cloud AI "Large - Standard" solution.
Which key components should the architect expect the Smart Template to include in the Bill of Materials (BOM)? (Choose 2.)
A. HPE ProLiant DL380a Gen11 servers with NVIDIA H100 NVL GPUs.
B. HPE Cray compute nodes.
C. HPE ProLiant DL380a Gen11 servers with NVIDIA L40S GPUs.
D. 8 worker nodes.
E. 4 worker nodes.
Answer: (SHOW ANSWER)

NEW QUESTION: 26

A customer has selected the "HPE Private Cloud AI - Medium - Expanded" configuration.
Which of the following resource counts correctly describes this specific configuration?
A. 4 worker nodes, 16 total H100 NVL GPUs, dual rack
B. 2 worker nodes, 8 total L40S GPUs, single rack
C. 4 worker nodes, 16 total L40S GPUs, single rack
D. 8 worker nodes, 32 total H100 NVL GPUs, dual rack
Answer: (SHOW ANSWER)

NEW QUESTION: 27
An architect is designing an AI solution for a financial services company that needs to build a chatbot.
The chatbot must answer customer queries using the company's latest internal policy documents, which are updated daily. The company has a limited budget and no data scientists available for a lengthy model retraining project.
Which approach should the architect recommend?
A. Train a new LLM from scratch using only the company's policy documents to ensure data privacy.
B. Fine-tune a pre-trained model daily by retraining it on the entire set of policy documents.
C. Use a pre-trained computer vision model to extract text from the policy documents.
D. Use a pre-trained model and implement a Retrieval-Augmented Generation (RAG) framework.
Answer: (SHOW ANSWER)

NEW QUESTION: 28
A customer is using the HPE Intelligent Configurator to determine the correct size for an HPE Private Cloud AI solution. Their primary use case is a text generation chatbot for internal HR support.
What is the MOST critical piece of information the customer must provide to the sizing tool to get an accurate recommendation for the number and type of GPUs required?
A. The brand of networking switches used in their existing data center fabric.
B. The number of ML engineers who will maintain the model.
C. The estimated number of concurrent users who will be querying the chatbot.
D. The physical location of the data center where the solution will be deployed.
Answer: (SHOW ANSWER)

NEW QUESTION: 29
An enterprise wants to deploy pre-trained foundation models from various sources for multiple business units. A key requirement is to simplify and standardize the deployment process, regardless of the model's origin. They want a solution that packages models into scalable, optimized, and easy-to-use microservices with a standard API.
Which component of the NVIDIA AI Enterprise software suite directly addresses this need?
A. NVIDIA NIM (NVIDIA Inference Microservices)

B. NVIDIA NeMo
C. NVIDIA Triton Inference Server
D. NVIDIA TAO Toolkit
Answer: (SHOW ANSWER)

NEW QUESTION: 30
A customer states they need an AI solution for a "summarization" use case that will be accessed by approximately 75 concurrent users. The underlying data is updated frequently, so RAG will be required. When using the HPE Intelligent Configurator, which three inputs are mandatory to get an initial sizing recommendation?

A. Use case, Number of users, RAG requirement
B. RAG requirement, Power and cooling capacity, Desired software license term
C. Use case, Model choice, Network switch preference
D. Number of users, Preferred GPU model, Data center location
Answer: (SHOW ANSWER)

NEW QUESTION: 31
An enterprise is designing a solution for training a large, custom Convolutional Neural Network (CNN) for a new computer vision application. Their data science team has determined that the training process will need to be distributed across multiple GPUs to be completed in a reasonable timeframe. The training process involves intensive matrix multiplication operations.
The architect is specifying components from the HPE Private Cloud AI solution.
Which infrastructure components are critical for accelerating this specific distributed training workload?
(Select all that apply.)
```
Workload Analysis:
- AI Model: Large Convolutional Neural Network (CNN)
- Task: Distributed Training
- Key Operation: Intensive matrix multiplication
```
A. NVIDIA GPUs featuring multiple Tensor Cores.
B. An NVIDIA NVLink Bridge to connect the GPUs.
C. HPE GreenLake for File Storage with standard NFS over TCP/IP.
D. The HPE Data Fabric software component.
E. HPE ProLiant DL325 servers for the worker nodes.
Answer: (SHOW ANSWER)

NEW QUESTION: 32
An architect is using the HPE One Config Advanced (OCA) Smart Template for "HPE Private Cloud AI
- Small - Standard".
What is the worker node configuration they should expect to see in the generated Bill of Materials (BOM)?
A. 2x HPE ProLiant DL380a Gen11 servers with a total of 8x NVIDIA L40S GPUs
B. 4x HPE ProLiant DL380a Gen11 servers with a total of 16x NVIDIA H100 NVL GPUs
C. 1x HPE ProLiant DL380a Gen11 server with 4x NVIDIA L40S GPUs
D. 3x HPE ProLiant DL325 Gen11 servers for management
Answer: (SHOW ANSWER)

NEW QUESTION: 33
What is the primary purpose of using RDMA over Converged Ethernet (RoCE) in the AI interconnect fabric of an HPE Private Cloud AI solution?
A. To allow direct memory-to-memory data transfers between nodes, bypassing the CPU kernel and reducing latency.
B. To guarantee packet delivery over lossy internet connections using the TCP protocol.
C. To encrypt data in transit between compute nodes and storage.
D. To partition a single physical network adapter into multiple virtual adapters for different tenants.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 34
A customer in the media and entertainment industry wants a solution to accelerate the creation of 3D animations and renderings. Their artists need to be able to generate complex visual AI content quickly.
The company has a mature AI practice and requires a powerful, data center-based solution.
Which HPE ProLiant server and NVIDIA GPU combination is purpose-built for these generative visual AI workloads?
A. HPE ProLiant DL380a Gen11 server with CPU-only configuration
B. HPE ProLiant DL380a Gen12 server with an NVIDIA NVLink Bridge but no GPUs
C. HPE ProLiant DL320 Gen11 server with NVIDIA L4 GPUs
D. HPE ProLiant DL380a Gen11 server with NVIDIA L40S GPUs
Answer: (SHOW ANSWER)

NEW QUESTION: 35

An 'AI Pro' customer wants to deploy a solution to create 'digital twins' of their manufacturing equipment for predictive maintenance simulations. This workload requires significant GPU power for both the simulation and the underlying AI model.
Which HPE solution should be positioned as the primary platform for this advanced, data-center based workload?
A. The customer's existing VDI environment.
B. A standard data warehouse solution without GPU acceleration.
C. HPE Private Cloud AI with NVIDIA.
D. Individual HPE ProLiant DL320 Gen11 servers at the edge.
Answer: (SHOW ANSWER)

NEW QUESTION: 36
An architect is sizing an HPE Private Cloud AI solution. The customer plans to deploy a generative AI application for 150 concurrent users that requires Retrieval-Augmented Generation (RAG).
The architect enters the following into the HPE Intelligent Configurator:
```
- Use case: Text Generation
- Number of users: 100-250
- RAG: Yes
- Model: Llama 2 13B (tool default for this use case)
```
Based on the provided inputs, which HPE Private Cloud AI configuration will the HPE Intelligent Configurator most likely recommend?
A. Medium - Expanded (4-node)
B. Large - Standard (4-node)
C. Small - Expanded (2-node)
D. Small - Standard (1-node)
Answer: A (LEAVE A REPLY)

NEW QUESTION: 37
A customer wants to enhance their existing Large Language Model (LLM) to provide more accurate and contextually relevant answers based on a proprietary, rapidly changing knowledge base of legal documents. They are considering two approaches: fine-tuning and Retrieval-Augmented Generation (RAG).
Review the data flow diagram for the proposed RAG implementation:
```
User Query -> [Query Encoder] -> Vector DB Search -> [Retrieved Documents] --+
|
+-> [LLM Prompt] -> LLM -> Response
```

Based on the diagram and the scenario, which statement accurately identifies a primary advantage of the RAG approach for this customer?

A. RAG requires retraining the LLM whenever a new legal document is added to the knowledge base.

B. RAG allows the LLM to access the most current legal documents at inference time without daily retraining.

C. RAG reduces the need for a vector database by directly integrating documents into the model.

D. RAG permanently modifies the LLM's internal weights to specialize in legal terminology.

Answer: (SHOW ANSWER)