



Generative AI and NetApp Value

NetApp artificial intelligence solutions

NetApp
July 29, 2025

This PDF was generated from <https://docs.netapp.com/us-en/netapp-solutions-ai/gen-ai/ai-wp-genai.html> on August 18, 2025. Always check docs.netapp.com for the latest.

Table of Contents

- Generative AI and NetApp Value 1
 - Abstract 1
 - Executive Summary 1
 - What is Generative AI? 2
 - Enterprise Use Cases and Downstream NLP Tasks 2
 - Role of storage in generative AI 4
 - Three primary approaches to LLMs 4
 - Foundation Models 5
 - Fine-tuning, domain-specificity, and retraining 5
 - Prompt engineering and Inferencing 6
 - LLMOps, Model Monitoring and Vectorstores 6
 - Risks and Ethics in the age of Generative AI 6
 - Customer scenario and NetApp 7
 - NetApp capabilities 8
 - ONTAP AI with DGX BasePOD 10**
 - ONTAP AI with NVIDIA AI Enterprise 10**
 - 1P Cloud Platforms 10**
 - NetApp Partner Solution Suite 10
 - Conclusion 11

Generative AI and NetApp Value

The demand for generative artificial intelligence (AI) is driving disruption across industries, enhancing business creativity and product innovation.

Abstract

Many organizations are using generative AI to build new product features, improve engineering productivity and prototype AI powered applications that deliver better results and consumer experiences. Generative AI such as Generative Pre-trained Transformers (GPT) use neural networks to create new content, as diverse as text, audio, and video. Given the extreme scale and massive datasets involved with large language models (LLMs), it is crucial to architect a robust AI infrastructure that takes advantage of the compelling data storage features of on-premises, hybrid and multicloud deployment options and reduce risks associated with data mobility, data protection and governance before companies can design AI solutions. This paper describes these considerations and the corresponding NetApp AI capabilities that enable seamless data management and data movement across the AI data pipeline for training, retraining, fine-tuning, and inferencing generative AI models.

Executive Summary

Most recently after the launch of ChatGPT, a spin-off of GPT-3 in November 2022, new AI tools used to generate text, code, image, or even therapeutic proteins in response to user prompts have gained significant fame. This indicates users can make a request using natural language and AI will interpret and generate text, such as news articles or product descriptions that reflect user request or produce code, music, speech, visual effects, and 3D assets using algorithms trained on already existing data. As a result, phrases like Stable Diffusion, Hallucinations, Prompt Engineering and Value Alignment are rapidly emerging in the design of AI systems. These self-supervised or semi-supervised machine learning (ML) models are becoming widely available as pre-trained foundation models (FM) via cloud service providers and other AI firms/vendors, which are being adopted by various business establishments across industries for a wide range of downstream NLP (natural language processing) tasks. As asserted by research analyst firms like McKinsey – "Generative AI's impact on productivity could add trillions of dollars in value to the global economy." While companies are reimagining AI as thought partners to humans and FMs are broadening simultaneously to what businesses and institutions can do with generative AI, the opportunities to manage massive volumes of data will continue to grow. This document presents introductory information on generative AI and the design concepts in relation to NetApp capabilities that bring value to NetApp customers, both on-premises and hybrid or multicloud environments.

So, what's in it for customers to use NetApp in their AI environments? NetApp helps organizations meet the complexities created by rapid data and cloud growth, multi-cloud management, and the adoption of next-generation technologies, such as AI. NetApp has combined various capabilities into intelligent data management software and storage infrastructure that have been well balanced with high-performance optimized for AI workloads. Generative AI solutions like LLMs need to read and process their source datasets from storage into memory numerous times to foster intelligence. NetApp has been a leader in data mobility, data governance and data security technologies across the edge-to-core-to-cloud ecosystem, serving enterprise customers build at-scale AI solutions. NetApp, with a strong network of partners has been helping chief data officers, AI engineers, enterprise architects and data scientists in the design of a free-flowing data pipeline for data preparation, data protection, and strategic data management responsibilities of AI model training and inferencing, optimizing the performance and scalability of the AI/ML lifecycle. NetApp data technologies and capabilities such as NetApp ONTAP AI for deep learning data pipeline, NetApp SnapMirror for transporting data seamlessly and efficiently between storage endpoints, and NetApp FlexCache for real-time rendering when the data flow shifts from batch to real-time and data engineering happens at prompt time, bring value to the deployment of real-time Generative AI models. As enterprises of all types embrace new AI

tools, they face data challenges from the edge to the data center to the cloud that demand for scalable, responsible and explainable AI solutions. As the data authority on hybrid and multi cloud, NetApp is committed to building a network of partners and joint solutions that can help with all aspects of constructing a data pipeline and data lakes for generative AI model training (pre-training), fine-tuning, context-based inferencing and model decay monitoring of LLMs.









What is Generative AI?

Generative AI is changing how we create content, generate new design concepts, and explore novel compositions. It illustrates neural network frameworks like Generative Adversarial Network (GAN), Variational Autoencoders (VAE), and Generative Pre-Trained Transformers (GPT), which can generate new content like text, code, images, audio, video, and synthetic data. Transformer-based models like OpenAI's Chat-GPT, Google's Bard, Hugging Face's BLOOM, and Meta's LLaMA have emerged as the foundational technology underpinning many advances in large language models. Likewise, OpenAI's Dall-E, Meta's CM3leon, and Google's Imagen are examples for text-to-image diffusion models which offer customers an unprecedented degree of photorealism to create new, complex images from scratch or edit existing images to generate high-quality context-aware images using dataset augmentation and text-to-image synthesis linking textual and visual semantics. Digital artists are starting to apply a combination of rendering technologies like NeRF (Neural Radiance Field) with generative AI to convert static 2D images into immersive 3D scenes. In general, LLMs are broadly characterized by four parameters: (1) Size of the model (typically in billions of parameters); (2) Size of the training dataset; (3) Cost of training, and (4) Model performance after training. LLMs also fall mainly into three transformer architectures. (i) Encoder-only models. E.g. BERT (Google, 2018); (ii) Encoder-Decoder E.g. BART (Meta, 2020) and (iii) Decoder-only models. E.g. LLaMA (Meta, 2023), PaLM-E (Google, 2023). Depending on the business requirement, irrespective of which architecture a company chooses the number of model parameters (N) and the number of tokens (D) in the training dataset generally determine the baseline cost of training (pre-training) or fine-tuning an LLM.

Enterprise Use Cases and Downstream NLP Tasks

Businesses across industries are uncovering more and more potential for AI to extract and produce new forms of value from existing data for business operations, sales, marketing, and legal services. According to IDC (International Data Corporation) market intelligence on global generative AI use cases and investments, knowledge management in software development and product design is to be the most impacted, followed by storyline creation for marketing and code generation for developers. In healthcare, clinical research organizations are breaking new ground in medicine. Pretrained models like ProteinBERT incorporate Gene Ontology (GO) annotations to rapidly design protein structures for medical drugs, representing a significant milestone in drug discovery, bioinformatics, and molecular biology. Biotech firms have initiated human trials for generative AI-discovered medicine, that aims to treat diseases like pulmonary fibrosis (IPF), a lung disease that causes irreversible scarring of lung tissue.

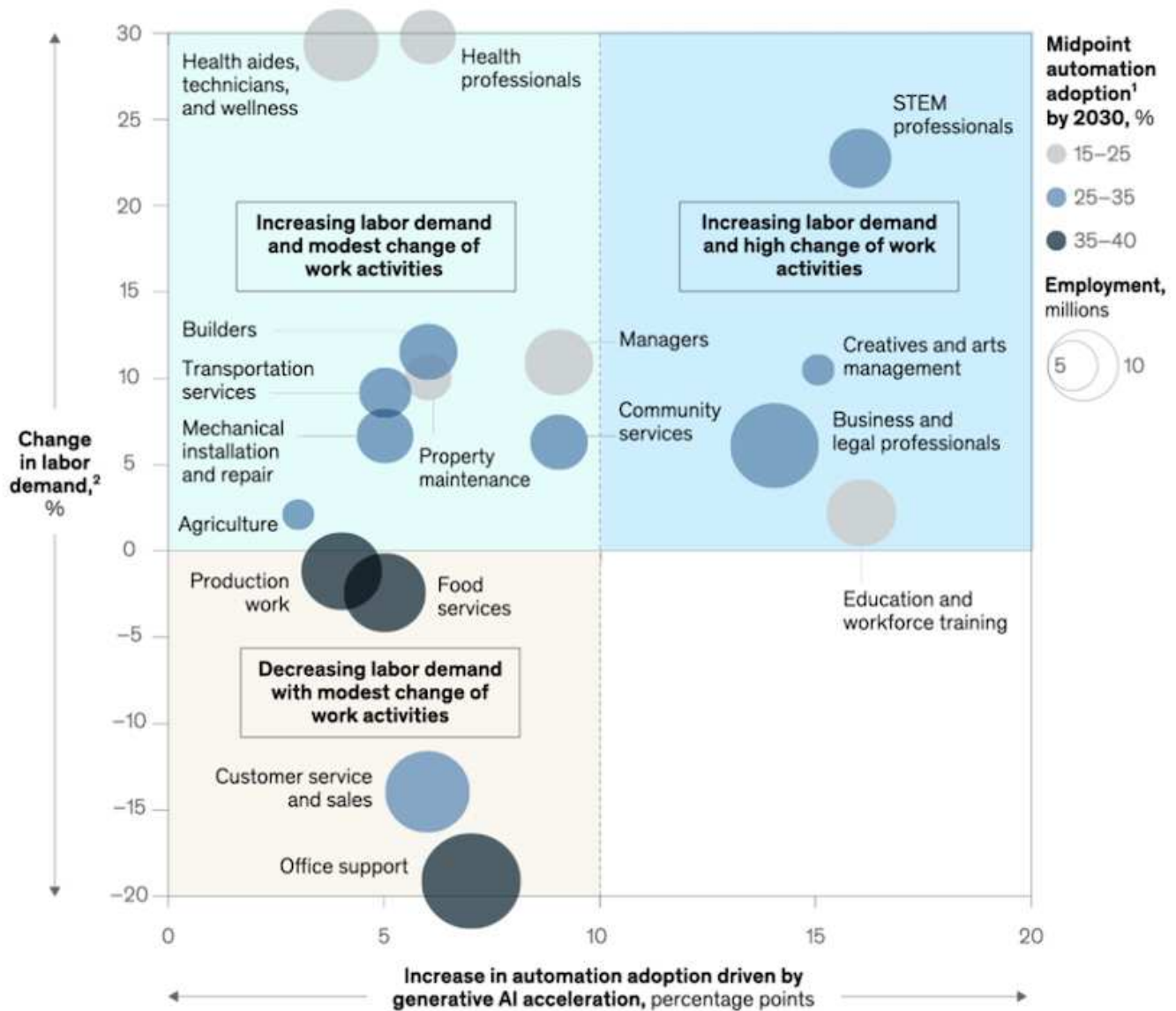
Figure 1: Use cases driving Generative AI

 <p>Chatbots</p>	 <p>Drug discovery</p>
 <p>Text generation</p>	 <p>Genome model expression</p>
 <p>Image generation</p>	 <p>Classification</p>
 <p>Code generation</p>	 <p>Speech-to-Text</p>

Increases in automation adoption driven by generative AI is also changing the supply & demand of work activities for many occupations. As per McKinsey the US labor market (diagram below) has gone through a rapid transition, which may only continue when factoring in the impact of AI.

Source: McKinsey & Company

Estimated labor demand change and generative AI automation acceleration by occupation, US, 2022–30



Role of storage in generative AI

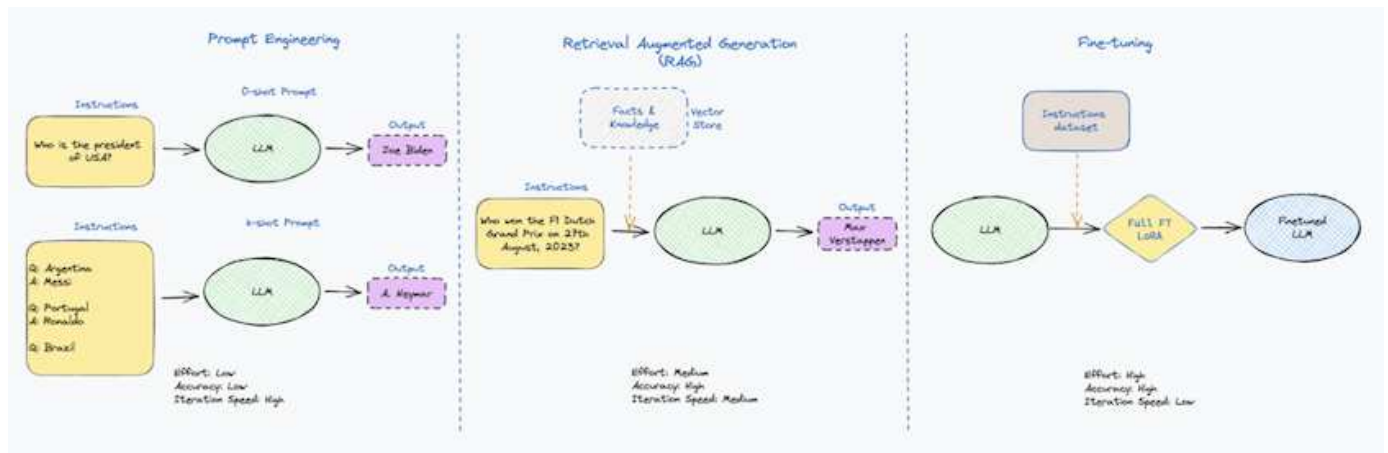
LLMs rely largely on deep learning, GPUs, and compute. However, when GPU buffer fills up, the data needs to be written quickly to storage. While some AI models are small enough to execute in memory, LLMs require high IOPS and high throughput storage to provide fast access to large datasets, especially if it involves billions of tokens or millions of images. For a typical GPU memory requirement of an LLM, the memory needed to train a model with 1 billion parameters could go up to 80GB @32-bit full precision. In which case, Meta's LLaMA 2, a family of LLMs ranging in scale from 7 billion to 70 billion parameters, may require 70x80, approx. 5600GB or 5.6TB of GPU RAM. Furthermore, the amount of memory you need is directly proportional to the maximum number of tokens you want to generate. For example, if you want to generate outputs of up to 512 tokens (about 380 words), you need "512MB". It may seem inconsequential – but, if you want to run bigger batches it starts to add up. Therefore, making it very expensive for organizations training or fine-tuning LLMs in memory, thus making storage a cornerstone for generative AI.

Three primary approaches to LLMs

For most businesses, based on current trends, the approach to deploying LLMs can be condensed into 3 basic

scenarios. As described in a recent "[Harvard Business Review](#)" article: (1) Training (pre-training) an LLM from scratch – costly and requires expert AI/ML skills; (2) Fine-tuning a foundation model with enterprise data – complex, yet feasible; (3) Using retrieval-augmented generation (RAG) to query document repositories, APIs and vector databases that contain company data. Each of these has tradeoffs between the effort, iteration speed, cost-efficiency and model accuracy in their implementations, used to solving different types of problems (diagram below).

Figure 3: Problem Types



Foundation Models

A foundation model (FM) also known as base model is a large AI model (LLM) trained on vast quantities of unlabeled data, using self-supervision at scale, generally adapted for a wide range of downstream NLP tasks. Since the training data is not labelled by humans, the model emerges rather than being explicitly encoded. This means the model can generate stories or a narrative of its own without being explicitly programmed to do so. Hence an important characteristic of FM is homogenization, which means the same method is used in many domains. However, with personalization and fine-tuning techniques, FMs integrated into products appearing these days are not only good at generating text, text-to-images, and text-to-code, but also for explaining domain specific tasks or debugging code. For instance, FMs like OpenAI's Codex or Meta's Code Llama can generate code in multiple programming languages based on natural language descriptions of a programming task. These models are proficient in over a dozen programming languages including Python, C#, JavaScript, Perl, Ruby, and SQL. They understand the user's intent and generate specific code that accomplishes the desired task useful for software development, code optimization, and automation of programming tasks.

Fine-tuning, domain-specificity, and retraining

One of the common practices with LLM deployment following data preparation and data pre-processing is to select a pre-trained model that has been trained on a large and diverse dataset. In the context of fine-tuning this can be an open-source large language model such as "[Meta's Llama 2](#)" trained on 70 billion parameters and 2 trillion tokens. Once the pre-trained model is selected, the next step is to fine-tune it on the domain-specific data. This involves adjusting the model's parameters and training it on the new data to adapt to a specific domain and task. For example, BloombergGPT, a proprietary LLM trained on a wide range of financial data serving the financial industry. Domain-specific models designed and trained for a specific task generally have higher accuracy and performance within their scope, but low transferability across other tasks or domains. When business environment and data change over a period, the prediction accuracy of the FM could begin to decline when compared to their performance during testing. This is when retraining or fine-tuning the model becomes crucial. Model retraining in traditional AI/ML refers to updating a deployed ML model with new data, generally performed to eliminate two types of drifts that occur. (1) Concept drift – when the link between the input variables and the target variables changes over time, since the description of what we want to predict changes, the model can produce inaccurate predictions. (2) Data drift – occurs when the characteristics of the

input data change, like changes in customer habits or behavior over time and therefore the model's inability to respond to such changes. In a similar fashion, retraining applies to FMs/LLMs, however it can be a lot costlier (in \$millions), therefore not something most organizations might consider. It is under active research, still emerging in the realm of LLMops. So instead of re-training, when model decay occurs in fine-tuned FMs, businesses may opt for fine-tuning again (lot cheaper) with a newer dataset. For a cost perspective, listed below is an example of a model-price table from Azure-OpenAI Services. For each task category, customers can fine-tune and evaluate models on specific datasets.

Source: Microsoft Azure

Model	Per 1000 token
Text-Ada	\$0.0001
GPT-3.5 Turbo	\$0.003
GPT-4	\$0.06
Text-Davinci	\$0.02
Model	Per 100 images
Dall-E	\$2

Prompt engineering and Inferencing

Prompt engineering refers to the effective methods of how to communicate with LLMs to perform desired tasks without updating the model weights. As important as AI model training and fine-tuning is to NLP applications, inferencing is equally important, where the trained models respond to user prompts. The system requirements for inferencing are generally much more on the read performance of the AI storage system that feeds data from LLMs to the GPUs as it needs to be able to apply billions of stored model parameters to produce the best response.

LLMOps, Model Monitoring and Vectorstores

Like traditional Machine Learning Ops (MLOps), Large Language Model Operations (LLMOps) also require the collaboration of data scientists and DevOps engineers with tools and best practices for the management of LLMs in production environments. However, the workflow and tech stack for LLMs could vary in some ways. For instance, LLM pipelines built using frameworks like LangChain string together multiple LLM API calls to external embedding endpoints such as vectorstores or vector databases. The use of an embedding endpoint and vectorstore for downstream connectors (like to a vector database) represents a significant development in how data is stored and accessed. As opposed to traditional ML models that are developed from scratch, LLMs often rely on transfer learning since these models start with FMs that are fine-tuned with new data to improve performance in a more specific domain. Therefore, it is crucial LLMOps deliver the capabilities of risk management and model decay monitoring.

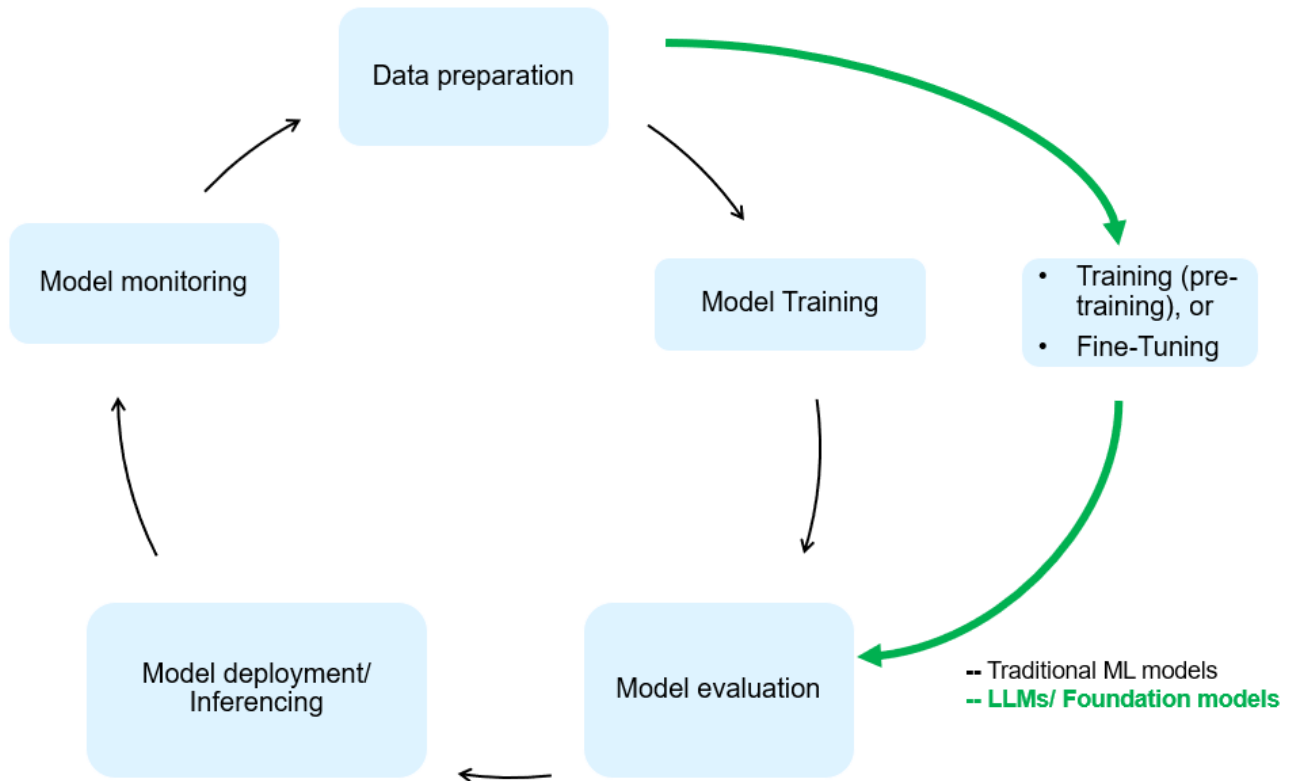
Risks and Ethics in the age of Generative AI

"ChatGPT – It's slick but still spews nonsense."– MIT Tech Review. Garbage in–garbage out, has always been the challenging case with computing. The only difference with generative AI is that it excels at making the garbage highly credible, leading to inaccurate outcomes. LLMs are prone to invent facts to fit the narrative it's

building. Therefore, companies that see generative AI as a great opportunity to lower their costs with AI equivalents need to efficiently detect deep fakes, reduce biases, and lower risks to keep the systems honest and ethical. A free-flowing data pipeline with a robust AI infrastructure that supports data mobility, data quality, data governance and data protection via end-to-end encryption and AI guardrails is eminent in the design of responsible and explainable generative AI models.

Customer scenario and NetApp

Figure 3: Machine Learning/Large Language Model Workflow



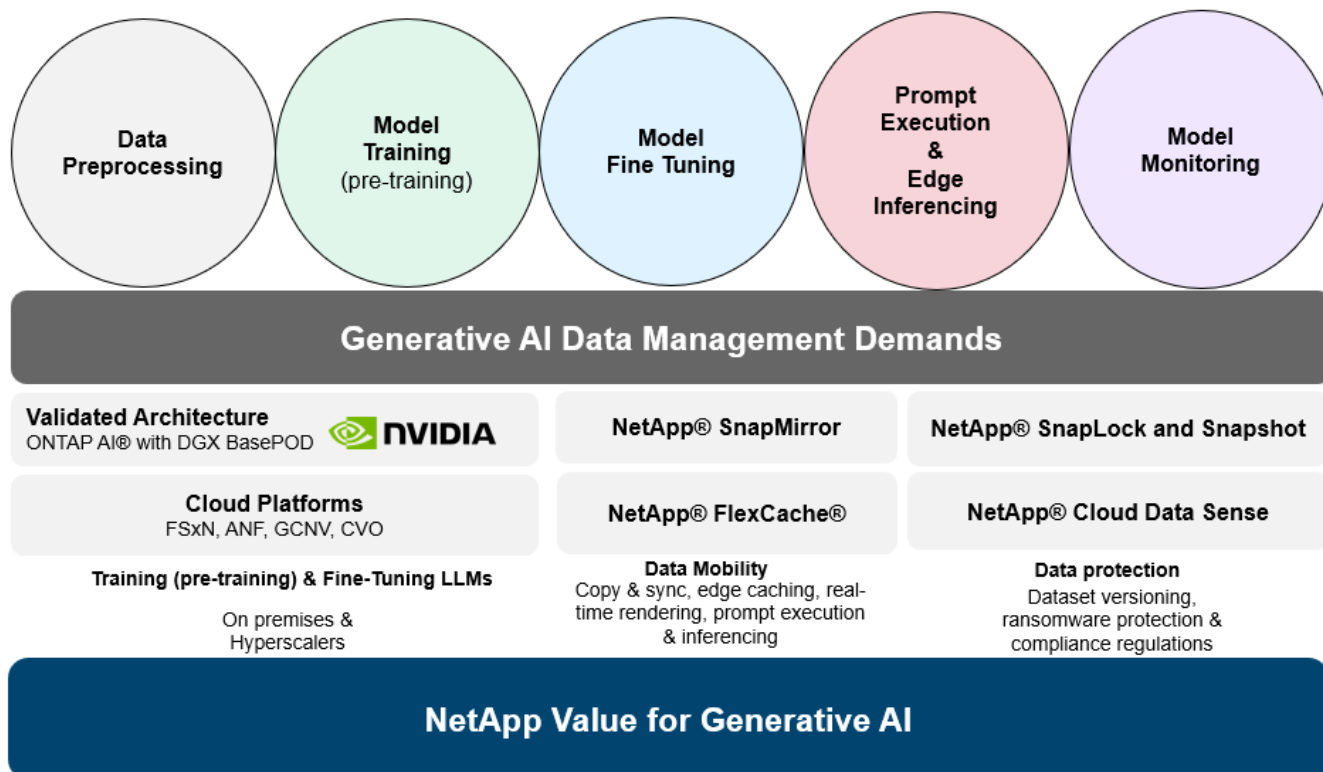
Are we training or fine-tuning? The question of whether to (a) train an LLM model from scratch, fine-tune a pre-trained FM, or use RAG to retrieve data from document repositories outside a foundation model and augment prompts, and (b) either by leveraging open-source LLMs (E.g., Llama 2) or proprietary FMs (E.g., ChatGPT, Bard, AWS Bedrock) is a strategic decision for organizations. Each approach has a tradeoff between cost-efficiency, data gravity, operations, model accuracy and management of LLMs.

NetApp as a company embraces AI internally in its work culture and in its approach to product design and engineering efforts. For instance, NetApp's autonomous ransomware protection is built using AI and machine learning. It provides early detection of file system anomalies to help identify threats before they impact operations. Second, NetApp uses predictive AI for its business operations like sales and inventory forecasting and chatbots to assist customers in call center product support services, tech specs, warranty, service manuals, and more. Third, NetApp brings customer value to the AI data pipeline and ML/LLM workflow via products and solutions serving customers building predictive AI solutions such as demand forecasting, medical imaging, sentiment analysis, and generative AI solutions like GANs for industrial images anomaly detection in manufacturing sector and anti-money laundering and fraud detection in banking & financial services with NetApp products and capabilities like NetApp ONTAP AI, NetApp SnapMirror, and NetApp FlexCache.

NetApp capabilities

The movement and management of data in generative AI applications such as chatbot, code generation, image generation or genome model expression can span across the edge, private data center, and hybrid multicloud ecosystem. For instance, a real-time AI-bot helping a passenger upgrade his or her airline ticket to business class from an end-user app exposed via APIs of pre-trained models such as ChatGPT cannot achieve that task by itself since the passenger information is not publicly available on the internet. The API requires access to the passenger's personal info and ticket info from the airline carrier which may exist in a hybrid or multicloud ecosystem. A similar scenario might apply to scientists sharing a drug molecule and patient data via an end-user application that uses LLMs to accomplish clinical trials across drug discovery involving one-to-many bio-medical research institutions. Sensitive data that gets passed to FMs or LLMs may include PII, financial information, health information, biometric data, location data, communications data, online behavior, and legal information. In such an event of real-time rendering, prompt execution and edge inferencing there is data movement from end user app to storage endpoints via open source or proprietary LLM models to a data center on premises or public cloud platforms. In all such scenarios, data mobility and data protection are crucial for the AI operations involving LLMs which rely on large training datasets and movement of such data.

Figure 4: Generative AI - LLM Data Pipeline



NetApp's portfolio of storage infrastructure, data and cloud services is powered by intelligent data management software.

Data Preparation: The first pillar of the LLM tech stack is largely untouched from the older traditional ML stack. Data preprocessing in AI pipeline is necessary to normalize and cleanse the data before training or fine-tuning. This step includes connectors to ingest data wherever it may reside in the form of an Amazon S3 tier or in on-premises storage systems such as a file store or an object store like NetApp StorageGRID.

NetApp ONTAP is the foundational technology that underpins NetApp's critical storage solutions in the data center and the cloud. ONTAP includes various data management and protection features and capabilities,

including automatic ransomware protection against cyber-attacks, built-in data transport features, and storage efficiency capabilities for a range of architectures from on-premises, hybrid, multiclouds in NAS, SAN, object, and software defined storage (SDS) situations of LLM deployments.

NetApp ONTAP AI for deep learning model training. NetApp ONTAP supports NVIDIA GPU Direct Storage with the use of NFS over RDMA for NetApp customers with ONTAP storage cluster and NVIDIA DGX compute nodes . It offers a cost-efficient performance to read and process source datasets from storage into memory numerous times to foster intelligence, enabling organizations with training, fine-tuning, and scaling access to LLMs.

NetApp FlexCache is a remote caching capability that simplifies file distribution and caches only the actively read data. This can be useful for LLM training, re-training, and fine tuning, bringing value to customers with business requirements like real-time rendering and LLM inferencing.

NetApp SnapMirror is an ONTAP feature that replicates volume snapshots between any two ONTAP systems. This feature optimally transfers data at the edge to your on-premises data center or to the cloud. SnapMirror can be used for moving data securely and efficiently between on-premises and hyperscaler clouds, when customers want to develop generative AI in clouds with RAG containing enterprise data. It efficiently transfers only changes, saving bandwidth and speeding replication, thus bringing essential data mobility features during the operations of training, re-training, and fine-tuning of FMs or LLMs.

NetApp SnapLock brings immutable disk capability on ONTAP-based storage systems for dataset versioning. The microcore architecture is designed to protect customer data with FPolicy Zero Trust engine. NetApp ensures customer data is available by resisting denial-of-service (DoS) attacks when an attacker interacts with an LLM in a particularly resource-consuming way.

NetApp Cloud Data Sense helps identify, map, and classify personal information present in enterprise datasets, enact policies, meet privacy requirements on premises or in the cloud, help improve security posture and comply with regulations.

NetApp BlueXP classification, powered by Cloud Data Sense. Customers can automatically scan, analyze, categorize, and act on data across data estate, detect security risks, optimize storage, and accelerate cloud deployments. It combines storage and data services via its unified control plane, Customers can use GPU instances for computation, and hybrid multicloud environments for cold storage tiering and for archives and backups.

NetApp File-Object Duality. NetApp ONTAP enables dual-protocol access for NFS and S3. With this solution, customers can access NFS data from Amazon AWS SageMaker notebooks via S3 buckets from NetApp Cloud Volumes ONTAP. This offers flexibility to customers who need easy access to heterogeneous data sources with the ability to share data from both NFS and S3. For e.g., fine-tuning FMs like Meta's Llama 2 text-generation models on SageMaker with access to file-object buckets.

NetApp Cloud Sync service offers a simple and secure way to migrate data to any target, in the cloud or on-premises. Cloud Sync seamlessly transfers and synchronizes data between on-premises or cloud storage, NAS, and object stores.

NetApp XCP is a client software that enables fast and reliable any-to-NetApp and NetApp-to-NetApp data migrations. XCP also provides the capability of moving bulk data efficiently from Hadoop HDFS file systems into ONTAP NFS, S3 or StorageGRID and XCP file analytics provides visibility into the file system.

NetApp DataOps Toolkit is a Python library that makes it simple for data scientists, DevOps, and data engineers to perform various data management tasks, such as near-instantaneously provisioning, cloning, or snapshotting a data volume or JupyterLab workspace that are backed by high-performance scale-out NetApp storage.

NetApp's product security. LLMs may inadvertently reveal confidential data in their responses, thus a concern to CISOs who study the vulnerabilities associated with AI applications leveraging LLMs. As outlined by OWASP (Open Worldwide Application Security Project), security issues such as data poisoning, data leakage, denial of service and prompt injections within LLMs can impact businesses from data exposure to unauthorized access serving attackers. Data storage requirements should include integrity checks and immutable snapshots for structured, semi-structured, and unstructured data. NetApp Snapshots and SnapLock are being used for dataset versioning. It brings strict role-based access control (RBAC), as well as secure protocols, and industry standard encryption for securing both data at rest and in transit. Cloud Insights and Cloud Data Sense together offer capabilities to help you forensically identify the source of the threat and prioritize which data to restore.

ONTAP AI with DGX BasePOD

NetApp ONTAP AI reference architecture with NVIDIA DGX BasePOD is a scalable architecture for machine learning (ML) and artificial intelligence (AI) workloads. For the critical training phase of LLMs, data is typically copied from the data storage into the training cluster at regular intervals. The servers that are used in this phase use GPUs to parallelize computations, creating a tremendous appetite for data. Meeting the raw I/O bandwidth needs is crucial for maintaining high GPU utilization.

ONTAP AI with NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software that is optimized, certified, and supported by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. This software facilitates the simple and rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud environment. NVIDIA AI Enterprise, powered by NetApp and VMware, delivers enterprise-class AI workload and data management in a simplified, familiar package.

1P Cloud Platforms

Fully managed cloud storage offerings are available natively on Microsoft Azure as Azure NetApp Files (ANF), on AWS as Amazon FSx for NetApp ONTAP (FSx ONTAP), and on Google as Google Cloud NetApp Volumes (GNCV). 1P is a managed, high-performance file system that enables customers to run highly available AI workloads with improved data security in public clouds, for fine-tuning LLMs/FMs with cloud native ML platforms like AWS SageMaker, Azure-OpenAI Services, and Google's Vertex AI.

NetApp Partner Solution Suite

In addition to its core data products, technologies and capabilities, NetApp also collaborates closely with a robust network of AI partners to bring added value to customers.

NVIDIA Guardrails in AI systems serve as safeguards to ensure the ethical and responsible use of AI technologies. AI developers can choose to define the behavior of LLM-powered applications on specific topics and prevent them from engaging in discussions on unwanted topics. Guardrails, an open-source toolkit, provides the ability to connect an LLM to other services, seamlessly and securely for building trustworthy, safe, and secure LLM conversational systems.

Domino Data Lab provides versatile, enterprise-grade tools for building and productizing Generative AI - fast, safe, and economical, wherever you are in your AI journey. With Domino's Enterprise MLOps Platform, data scientists can use preferred tools and all their data, train and deploy models easily anywhere and manage risk and cost effectively - all from one control center.

Modzy for Edge AI. NetApp and Modzy have partnered together to deliver AI at scale to any type of data, including imagery, audio, text, and tables. Modzy is an MLOps platform for deploying, integrating, and running AI models, offers data scientists the capabilities of model monitoring, drift detection and explainability, with an

integrated solution for seamless LLM inference.

Run:AI and NetApp have partnered to demonstrate the unique capabilities of the NetApp ONTAP AI solution with the Run:AI cluster management platform for simplifying orchestration of AI workloads. It automatically splits and joins GPU resources, designed to scale your data processing pipelines to hundreds of machines with built-in integration frameworks for Spark, Ray, Dask, and Rapids.

Conclusion

Generative AI can produce effective results only when the model is trained on reams of quality data. While LLMs have achieved remarkable milestones, it is critical to recognize its limitations, design challenges and risks associated with data mobility and data quality. LLMs rely on large and disparate training datasets from heterogenous data sources. Inaccurate outcomes or biased results generated by the models can put both businesses and consumers in jeopardy. These risks can correspond to constraints for LLMs emerging potentially from data management challenges associated with data quality, data security, and data mobility. NetApp helps organizations meet the complexities created by rapid data growth, data mobility, multi-cloud management, and the adoption of AI. At scale AI infrastructure and efficient data management is crucial to defining the success of AI applications like generative AI. It is critical customers cover all the deployment scenarios without compromising on the ability to expand as enterprises need to while maintaining cost-efficiency, data governance and ethical AI practices in control. NetApp is constantly working to help customers simplify and accelerate their AI deployments.

Copyright information

Copyright © 2025 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

NETAPP, the NETAPP logo, and the marks listed at <http://www.netapp.com/TM> are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.