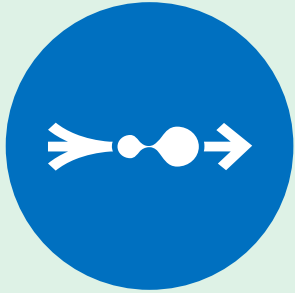


# FlexPod AI

**“A simple, automated, secure platform for runtime AI”**

# FlexPod AI:

A simple, automated, secure platform for all your workloads



## Simple

Reduce complexity with a flexible IT infrastructure



## Automated

Save time and reduce risk



## Secure

Save time and reduce risk

# Use Cases with FlexPod AI



## Operationalize Gen AI LLM Models

- Content Generation
- Chatbots
- Knowledge Management
- Language Translation



## Inferencing

- Product Recommendations
- Disease Diagnosis
- Drug Discovery
- Smart Devices
- Quality Control
- IoT
- Autonomous Vehicles



## Operationalize Computer Vision Models

- Object Recognition
- Facial Recognition
- Image and Video Analysis
- Medical Diagnosis
- Surveillance and Security

# FlexPod AI Validated Designs

Simple



**Model Provider**


 Hugging Face  NVIDIA  PyTorch  Meta



**Dev & Deployment Suite**

 NVIDIA AI Enterprise  Red Hat  Kubernetes

 DOMINO DATA LAB  NetApp DataOps Toolkit 

**Infrastructure**

**UCS Servers**  vmware<sup>®</sup> by Broadcom

**NetApp Storage Controller**  NVIDIA  NVIDIA GPUs

## Reference Architectures Solutions for AI Use Cases:

1. [FlexPod with Red Hat OpenShift, NVAIE, Nvidia GPUs](#)
2. [FlexPod with Suse Rancher, NVAIE, Nvidia GPUs](#)
3. [FlexPod Scaling and Benchmarking for GPU Intensive Applications](#)

# Large Language Models with FlexPod AI

## Proprietary LLMs

- High Cost for access and usage
- Limited Customization
- Data Privacy Concerns
- Vendor Lock in
- Integration Challenges
- Lack of Transparency
- Limited community support

## Open-source LLMs

- **Cost-Effective:** generally free or low-cost
- **Customizability:** high levels of customization
- **Transparency:** access to the source code and training methodologies
- **Community Support:** a community of developers and users contributes to continuous improvement
- **Data Privacy and Security:** Users have more control over their data
- **No Vendor Lock-in:** Users are not tied to a single vendor

# Prompting vs Fine-tuning vs RAG with FlexPod AI

## Prompting:

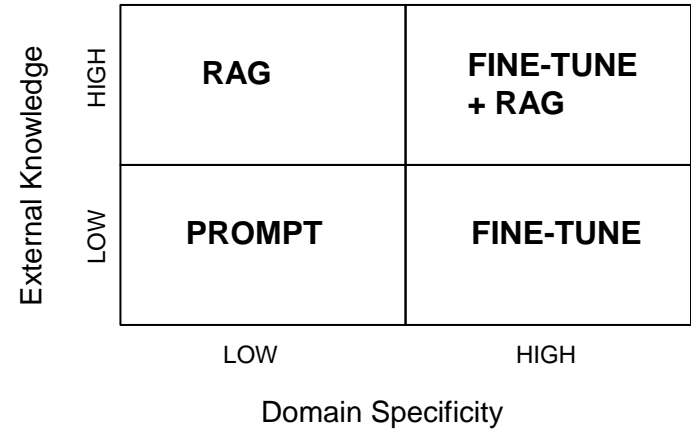
- Involves providing a pre-trained language model with a specific text input (prompt) to guide its response.
- Used to extract or generate information based on the context provided in the prompt.
- Effective for tasks like question answering, text completion, and creative writing.
- No model training required; relies on the model's existing knowledge and capabilities.

## Fine-tuning:

- A process of further training a pre-trained model on a specific dataset to specialize its performance.
- Enhances model accuracy on tasks or datasets different from the original training data.
- Requires a dataset representative of the target task or domain.
- Involves a risk of overfitting if the fine-tuning dataset is not sufficiently diverse.

## Retrieval-Augmented Generation (RAG):

- Combines pre-trained language models with a retrieval mechanism to augment generation with external information.
- Retrieves relevant documents or data from a large corpus, which the model uses to inform its responses.
- Enhances the model's ability to provide up-to-date, factual, and detailed answers.
- Useful for tasks requiring external knowledge or fact-checking.



# Free NetApp AI tools

Simple

Included at no additional cost to empower FlexPod users



## NetApp AI Control Plane: Full Stack AI Data and Experiment Management

- Comprehensive management of AI, ML, and deep learning data and experiments
- Enables AI workload scalability across regions and sites using Kubernetes
- Ensures physical storage space is utilized efficiently
- Utilizes Kubeflow to simplify the deployment of AI workflows



## NetApp DataOps Tool Kit: a full-stack AI data and experiment management reference architecture

- Simplifies the management of development/training workspaces and inference servers
- Python-based tool to rapidly provisions new JupyterLab workspaces and NVIDIA Triton Inference Servers
- Facilitates collaboration between data scientists and AI teams through shared access to datasets, experiments, and models



## NetApp BlueXP: Modern data estate operations made simple

- Classification provides data analysis and categorization to ensure data models meet privacy compliance requirements, detect security vulnerabilities, optimize costs, and accelerate data ingest.
- Create and configure storage on the cloud of your choice, Azure, Google or FSx
- Uncover and address risk factors or find opportunities to improve system availability, security, and performance; Handle storage data growth challenges efficiently

# FlexPod: A Secure, modern foundation for all your applications

