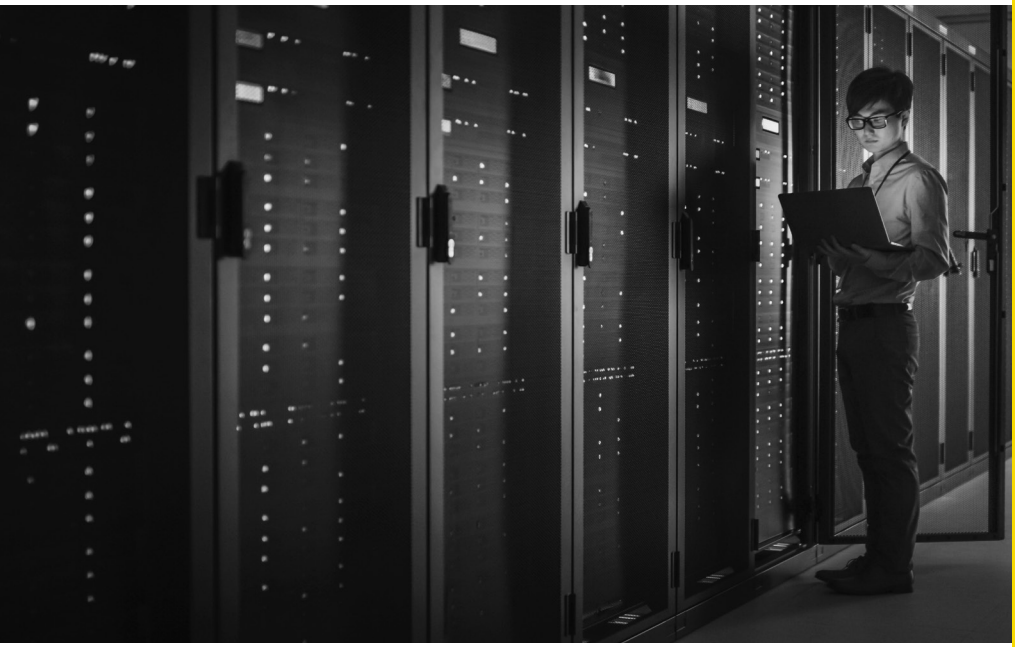
PHISON
PASCARI



Pascari aiDAPTIV™: Cost-Effective On-Premises LLM Training and Inference

Large Model Training with Your Private Data

Pascari aiDAPTIV enables you to train and run large language models on-premises using your own data. By intelligently extending effective GPU memory with high-endurance flash, aiDAPTIV makes large-scale AI development practical without relying on cloud infrastructure.

Scale Model Size Without Scaling GPU Count

aiDAPTIV manages data across GPU memory, system memory, and flash cache to remove memory bottlenecks that limit local AI workloads. This allows teams to work with larger models, longer contexts, and extended AI sessions on existing workstation and server platforms.

Edge / IoT /Robotics
1-8B Parameter
LoRA training

AI Notebook PC
Up to 8B parameter
full model training

Desktop PC
Up to 13B parameter
full model training

Workstation PC
Up to 100B parameter
full model training

Server
Up to 405B parameter
full model training

Benefits

Fits Your Budget

Extends effective GPU memory using aiDAPTIV cache memory, enabling larger AI workloads on existing GPU configurations.

Simple to Use and Deploy

Supports command-line workflows and integrated tooling for data ingest, fine-tuning, and inference. Deploys in offices, labs, classrooms, and data centers using standard infrastructure.

Keeps Data in Your Control

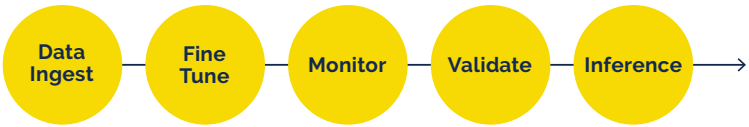
Enables AI training and inference entirely behind your firewall. Maintains ownership of sensitive data and supports data sovereignty and compliance requirements.

Find more information and resources at:
phisonenterprise.com

aiDAPTIV Technology: LLM Training Integrated Solution

aiDAPTIV Pro Suite

aiDAPTIV Pro Suite builds on Phison's aiDAPTIV memory management middleware, while also providing higher-level tools for workflow orchestration, monitoring, and AI workload execution through both command-line and graphical interfaces.



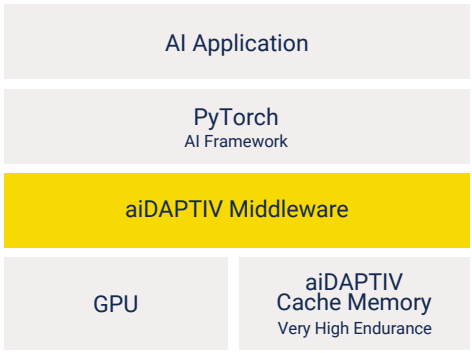
Supported Models

- LLaMA, LLaMA-2, LLaMA-3, Code LLaMA
- Mistral and other transformer-based language models
- Vision and multimodal models such as CLIP and Whisper
- Additional foundation and domain-specific models

aiDAPTIV Pro Suite

and/or

aiDAPTIV Middleware



Built-in Memory Management Solution

aiDAPTIV memory management middleware transparently manages data across GPU memory, system memory, and aiDAPTIV cache memory. It enables memory-intensive AI workloads to exceed native GPU memory limits without requiring changes to existing AI frameworks or applications.

- **AI Application**
- **PyTorch / AI Runtimes:** AI Framework
- **aiDAPTIV Middleware:** Multi-tier Memory Management
- **aiDAPTIV Cache Memory:** Extremely high-endurance SSD flash memory

and

Seamless Integration with GPU Memory

The optimized middleware extends GPU memory by an additional 2x 2TB using aiDAPTIV cache memory. This added memory is used to support LLM training with low latency. Plus, the high-endurance feature offers an industry-leading 100 DWPD, utilizing a specialized SSD design with an advanced NAND correction algorithm.

aiDAPTIV Cache Memory

M.2 AND U.2 SSDS

PHISON PASCARI

Find more information and resources at: phisonenterprise.com