



Fits Your Budget

Offloads expensive HBM & GDDR memory to cost-effective Flash memory. Significantly reduces the need for large numbers of high-cost and power-hungry GPU cards.



Simple to Use and Deploy

Offers all-in-one AI toolset enabling ingest to fine-tuning to inference using an intuitive graphical user interface. Deploys in your home, office, classroom or data center using commonplace power.



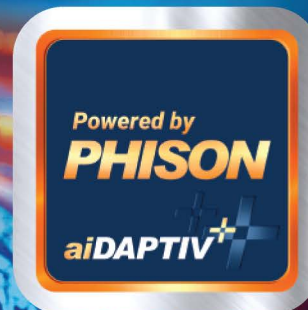
Keeps Your Data Private

Enables LLM training behind your firewall. Gives you full control over your private data and peace of mind over data sovereignty compliance.

PHISON aiDAPTIV+

- ✓ Budget-Friendly
- ✓ Private
- ✓ Smarter AI

Significantly Improves LLM Training Capacity and Inference Capabilities



Cost-Effective On-Site LLM Training and Inference

Any Model Size On-Premises

aiDAPTIV+ allows businesses to scale-up or scale-out nodes to increase training data size and reduce training time.

Phison's aiDAPTIV+ enables organizations to tackle the largest AI processing challenges on-premises. Running on validated platforms from Edge/IoT devices to a single AI PC or workstation to data center servers, aiDAPTIV+ provides a cost-effective approach to model training and inferencing on LLMs such as Llama-3 70B and Falcon 180B parameter models.



Large Model Training with Your Private Data

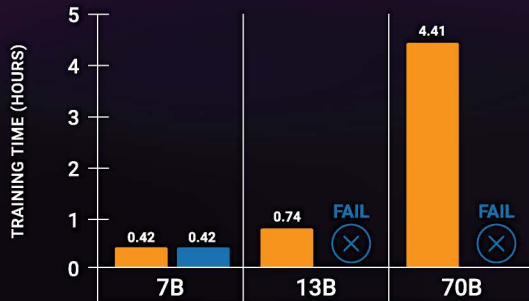
aiDAPTIV+ provides a turnkey solution for organizations to train large data models on-site at a price they can afford. It enhances foundation LLMs by incorporating an organization's own data enabling better decision making and innovation.

Boosts Inference Performance and Accuracy

aiDAPTIV+ delivers a better inferencing experience on-premises. It does this by extending the token length which enables you to create more complex and detailed prompts leading to lengthier and more accurate responses. Furthermore, aiDAPTIV+ provides faster time to first token (TTFT) recall for subsequent prompts helping you get to better results more quickly.

Unlock Large Model Training

Phison's aiDAPTIV+ solution enables significantly larger training models, giving you the opportunity to run AI processing that was previously too expensive to run on-premises or only reserved for the public cloud.



Training Set Size	140 GB	260 GB	1400 GB
HBM Pool (Usage%)	192 GB (73%)	192 GB (120%)	192 GB (729%)
Minimum GPU Count	4 / 4	4 / 6	4 / 30

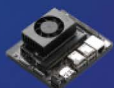
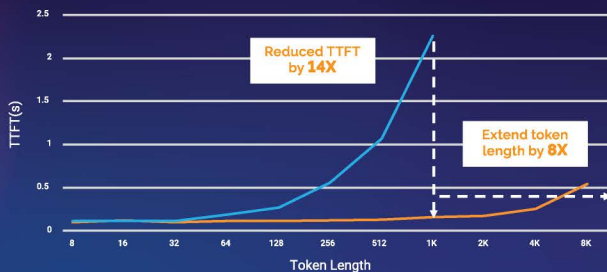
Note: Scaling is linear based on GPU count and model size

aiDAPTIV+	System Configuration
GPU Only	• RAM: 512 GB
	• GPU: 4x RTX 6000 ada
	• HBM: 192 GB

From Training to Chat

The software interface allows you to proceed from data ingest to fine-tune and RAG training to chat. In addition to enhanced LLM capacity, aiDAPTIV+ also improves inference context and recall performance for a better user experience.

Faster Prompt Recall Performance and Improved Context for Better Answers



Note:
1. Estimated figures are based on Llama 3.1-8B
2. System: NVIDIA Jetson Orin Nano Super 8GB

Phison aiDAPTIV+ LLM Training Integrated Solution

Use a Command Line or leverage the intuitive All-in-One aiDAPTIVPro Suite to perform LLM Training



Supported Models

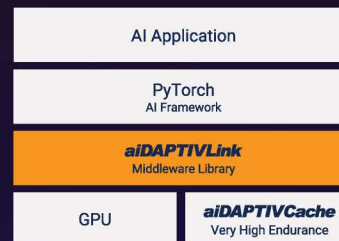
- Llama, Llama-2, Llama-3, CodeLlama
- Vicuna, Falcon, Whisper, Clip Large
- Metaformer, Resnet, Deit base, Mistral, TAIDE
- And many more being continually added

aiDAPTIVPro Suite



and/or

aiDAPTIVLink



Built-in Memory Management Solution

Experience seamless PyTorch compliance that eliminates the need to modify your AI application. You can effortlessly add nodes as needed. System vendors have access to AI100E SSD, middleware library licenses, and full Phison support to facilitate smooth system integration.

and

Seamless Integration with GPU Memory

The optimized middleware extends GPU memory by an additional 80-320GB for IoT devices, 320GB-2TB for PCs, and 1-8TB for workstations and servers using aiDAPTIVCache. This added memory is used to support LLM training with low latency. Furthermore, the high endurance feature offers an industry-leading 100 DWPD, utilizing a specialized SSD design with an advanced NAND correction algorithm.

aiDAPTIVCache Family



AI100E M.2 SSD