PHISON aiDAPTIV+

RED DATA

aiDAPTIV+ Delivers Accessible, Custom-Trained AI

☑ Private

☑ Easy-to-Use

☑ Budget-Friendly

Powered by PHISON aiDAPTIV+

Significantly Boosts LLM Training Size and Inference Performance

Challenges Facing the Public Sector

- Cloud compute and data transfer fees make large-scale AI operations unsustainable
- Sensitive or classified datasets cannot enter multi-tenant cloud environments
- Agencies and universities must maintain strict data sovereignty and chain-of-custody
- Air-gapped or SCIF environments require fully on-prem AI solutions
- Training staff in cloud-based systems is expensive and slow
- Mission workloads need local, reliable AI compute at the edge or in secure facilities
- Budgets and procurement cycles restrict the ability to buy massive GPU clusters

Private, Bigger, Faster LLM Training and Inference

Any Model Size On-Premises

aiDAPTIV+ allows organizations 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 systems, aiDAPTIV+ provides a less expensive approach to model training and inferencing on LLMs such as Llama-3 70B and Falcon 180B parameter models.

Edge/IoT/Robotics
Up to 1B Parameter
Full Model 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 671B Parameter
Full Model Training

Storage Systems
Over 1T Parameter
Full Model Training

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.

> **"Train and do inference in secure environments"**

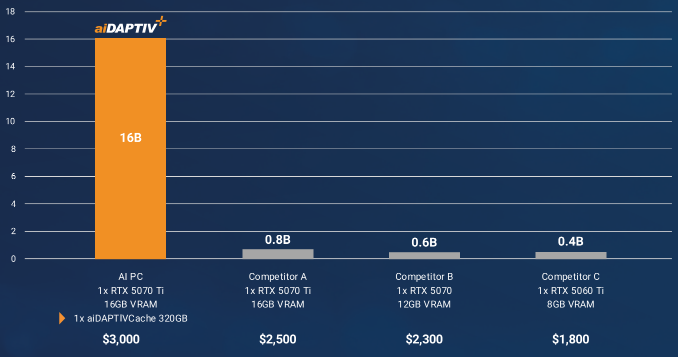
Boosts Inference Performance and Accuracy

aiDAPTIV+ delivers a better inferencing experience on-premises. It does this by extending the token length which enables you to create more detailed prompts leading to more accurate responses. Furthermore, aiDAPTIV+ provides faster time to first token (TTFT) when the KV cache is full, helping you get 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.

Capacity Boost for a 1-GPU Desktop PC with aiDAPTIV+



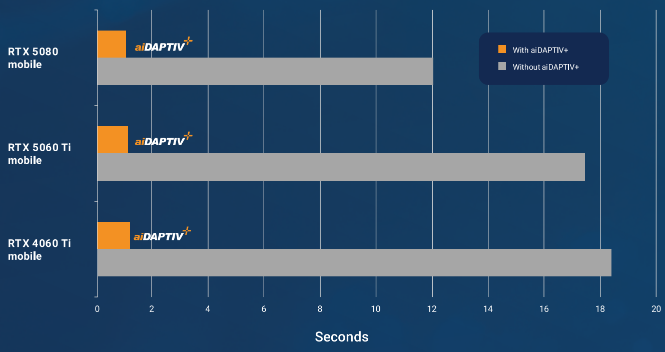
LLM sizes: 16B CodeGen, 0.8B MobiLlama, 0.6B Qwen3, 0.4B EleutherAI Pythia

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 performance for a better user experience.

10X Faster Inference performance with aiDAPTIV+

Time to First Token after KV Cache Pre-Fill (Smaller is Better)

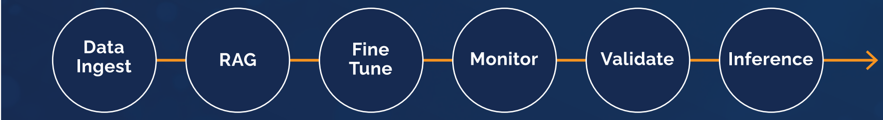


Llama 3.1 8B Q4, the KV cache size of 4GB for 32K token length.
TTFT with aiDAPTIV+ is minimally affected by computing power or number of GPUs.

Phison aiDAPTIV+ AI Processing Integrated Solution

aiDAPTIVPro Suite

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



Supported Models

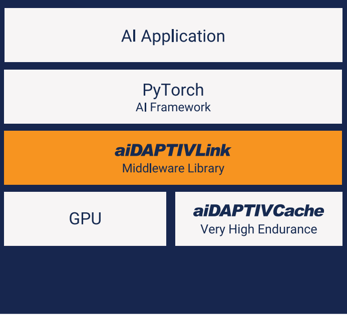
- Llama, Llama-2, Llama-3, CodeLlama
- Vicuna, Falcon, Whisper, Clip Large
- Metaformer, Resnet, Deit base, Mistral, TAIDE
- And many more being continually added

and/or

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 aiDAPTIVCache SSDs, middleware library licenses, and full Phison support to facilitate smooth system integration.

aiDAPTIVLink



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 AND U.2 SSDS