

✓ Private

✓ Easy-to-Use

✓ Budget-Friendly

Significantly Boosts LLM Training Size and Inference Performance

aiDAPTIV Enables Private, Bigger, Faster LLM Training and Inference

Any Model Size On-Premises

With aiDAPTIV, you can scale-up or scale-out nodes to increase training data size and reduce training time.

Phison's aiDAPTIV technology enables you 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
Inference & 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 671B Parameter
Full Model Training

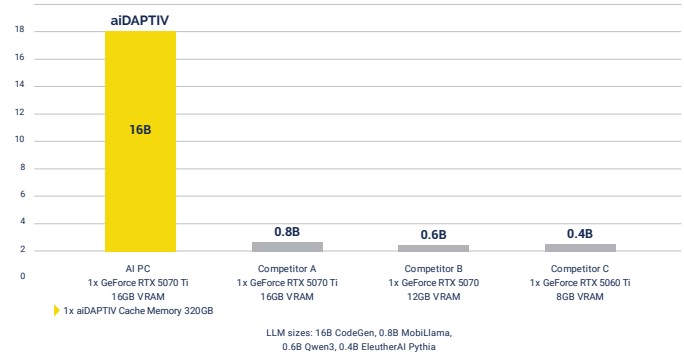


Storage Systems
Over 1T Parameter
Full Model Training

Unlock Large Model Training

Our 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

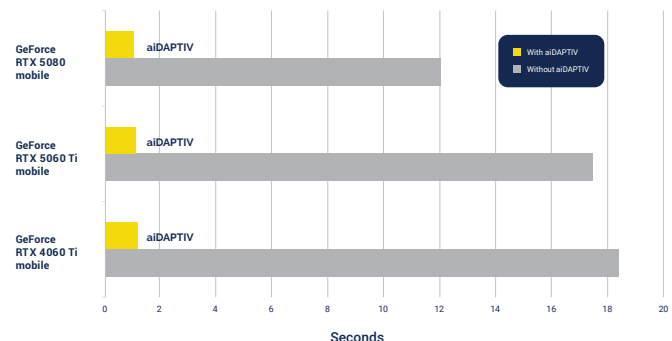


From Training to Chat

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.

Large Model Training with Your Private Data

aiDAPTIV provides a turnkey solution that enables you to train large data models on-site more affordably. It enhances foundation LLMs by incorporating your own data to enable 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 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.

Pascari aiDAPTIV AI Processing Integrated Solution

aiDAPTIV Pro Suite

Use a command line or leverage the intuitive all-in-one Pro Suite to perform LLM training.



Supported Models

- Llama, Llama-2, Llama-3, CodeLlama
- Mistral, Mixtral, Qwen, DeepSeek
- Phi, Gemma, Vicuna, Falcon
- Whisper, LLaVA, Qwen-VL, InternVL
- TAIDE (Breeze series)
- And many more added continually



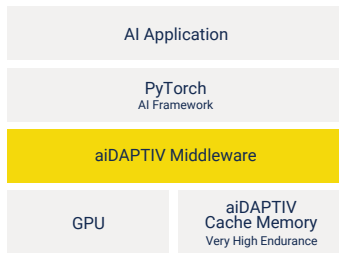
and/or

and

Memory Management Middleware

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 aiDAPTIV cache memory, middleware licenses, and full Phison support to facilitate smooth system integration.

aiDAPTIV Middleware

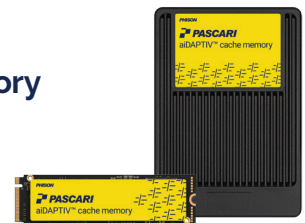


Seamless Integration with GPU Memory

The optimized middleware technology offers a range of capacities to extend GPU memory 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



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.



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.



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.