ByteShape

AI MODEL OPTIMIZATION
For model developers, platforms, and deployers

Make AI models faster, smaller, and less expensive to run.

ByteShape enables organizations to deploy more capable AI on existing hardware while reducing latency, memory requirements, energy consumption, and infrastructure cost.

LOWER COST
Improve AI economics
Reduce infrastructure, memory, and energy requirements.

MORE CAPABILITY
Use existing hardware
Deploy larger or higher-quality models on available systems.

MORE CONTROL
Own the deployment
Run in cloud, on-premises, at the edge, or in sovereign infrastructure.

AUTOMATED MODEL + HARDWARE OPTIMIZATION

ByteShape automatically optimizes numerical precision, value representation, and execution code for each target model, workload, and hardware platform.

Our system explores fine-grained datatype and quantization choices across tensors and value groups, balancing model quality against customer-defined deployment objectives.

QUALITY Fidelity	**LATENCY** Response time	**THROUGHPUT** e.g., Tokens / sec
MEMORY Model footprint	**ENERGY** Power use	**COST** Deployment economics

HOW IT WORKS

01 Specify
Define the model, target hardware, quality threshold, and deployment objective.

02 Optimize
ByteShape searches the design space and measures quality and performance.

03 Deploy
Receive an optimized model and execution path that fit your workflow.

WHY BYTESHAPE

AUTOMATED SEARCH
Beyond fixed recipes and manual tuning
Systematically learns a broader optimization space using measured quality and performance.

OBJECTIVE-DRIVEN
Built around your deployment target
Optimizes against the quality, latency, throughput, memory, energy, and cost targets that matter.

BROAD COMPATIBILITY
Models, formats, workloads, and devices
Supports integer, floating-point, and microscaling representations across diverse architectures (not only LLMs) and targets.

FUTURE-READY
Designed to adapt as AI changes
Operates at the level of tensors, datatypes, data movement, and compute kernels - not one fixed model family.

BUILT FOR REPEATABLE DEPLOYMENT ACROSS THE AI STACK

AI MODEL COMPANIES
Ship optimized releases across more devices and price points.

PLATFORMS & INFRASTRUCTURE
Improve utilization and offer stronger performance per dollar.

ENTERPRISES & PUBLIC SECTOR
Deploy capable AI while retaining control of systems and data.

PROVEN TECHNICAL FOUNDATION
Peer-reviewed research at MLSys, demonstrated from edge devices to high-end GPUs.
Commercial model: annual enterprise subscriptions and optimization engagements.

START A CONVERSATION
Bring us a model, target hardware, and deployment objective.
contact@byteshape.com