ByteShape

COMPANY OVERVIEW
For model developers, platforms, and deployers
Toronto, Canada · September 2026

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

ByteShape is a University of Toronto spinout. Our technology, ShapeLearn, learns how to represent AI models with fewer bits, so organizations can deploy more capable AI on the hardware they already have, with lower latency, memory, energy use, 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 the systems you have.

MORE CONTROL
Own the deployment.
Run in the cloud, on-premises, or at the edge, and keep control of your systems and data.

PUBLISHED LANGUAGE-MODEL RESULT

A 35B model on one desktop graphics card.

24 GB
RTX 4090
graphics memory

285.53
tokens per second
with multi-token prediction

99.27%
of its own full-precision
benchmark score

Qwen3.6-35B-A3B · MTP-GPU-5 · llama.cpp. Mixture of experts: 35B total parameters, 3B active. Model file: 18.61 GB, separate from total runtime memory. Score is normalized to this model's BF16 baseline on the release's own suite. Published results [1].

The technology

ShapeLearn learns numerical formats per group of weights. It searches a wider space than fixed quantization recipes or manual tuning, guided by measured quality and performance on the target hardware. It optimizes against the targets that matter to a deployment: quality, latency, throughput, memory, energy, and cost.

How it works

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

02 Learn. ShapeLearn searches the design space and allocates precision across the model.

03 Evaluate and deliver. We measure quality and speed on the target hardware, then deliver the optimized model files, benchmark results, and runtime settings that fit your workflow.

The product

ShapeLearn is an automated model optimization library and SDK. We are building toward independent use within teams' workflows, with collaborative integration and project support around their deployment requirements. Public language, vision-language, and image builds make the approach inspectable today. Longer term, the same optimization extends across AI software and hardware. Technology direction [4].

Beyond language

Qwen-Image-2512 diffusion weights shrink from **40.86 GB to 7.77 GB (5.26× smaller)** in our 3.04-bit GGUF build. This is weight storage, excluding text encoder, VAE, and runtime memory. Published image pairs show the quality tradeoffs. [2] ShapeLearn is not tied to one model family: it works across integer and floating-point formats and across devices, from a Raspberry Pi 5 to workstation GPUs.

The team and research

Founders **Andreas Moshovos, Miloš Nikolić, Enrique Torres Sanchez, and Ali Hadi Zadeh** bring backgrounds in AI acceleration and computer engineering at the University of Toronto. ShapeLearn builds on their peer-reviewed research in numerical representations at MLSys, ISCA, MICRO, and EMNLP.

Backed by **Two Small Fish Ventures** and **UTEST**. Team and research [3].

Who it is for

AI model companies: ship optimized releases across more devices and price points.

Platforms and infrastructure: raise utilization and serve more from the same hardware.

Enterprises and public sector: deploy capable AI while keeping control of systems and data.

Bring us a model, target hardware, and a deployment objective.

byteshape.com/company
ShapeLearn & integration · Investors & partners

Evidence and further reading
[1] Qwen3.6-35B-A3B release · [2] Qwen-Image-2512 release
[3] Company and research · [4] Technology
Public model catalog · Evaluation methodology