Infrastructure Stack

Modal Cloud
Serverless GPUs

NVIDIA A100
40GB VRAM | CUDA 12.x

PyTorch 2.1+
Transformers 4.36+

BitsAndBytes
4-bit Quantization Library