

ExecuTorch in the Browser

ExecuTorch is PyTorch's unified solution for deploying AI models on-device, from smartphones to microcontrollers, built for privacy, performance, and portability. This work adds a WebGPU delegate, bringing GPU-accelerated browser-based ML inference to PyTorch.

Browser and Platform Support

- Chrome, Safari, Firefox in Mobile/Laptop
- Metal (macOS), Vulkan (Linux/Windows)

Model Support

Text LLMs: Llama, Qwen, Gemma E2B/E4B/MTP

Image Segmentation: Segment Anything (SAM), Object Detection (YOLO)

Vision: Super-Resolution (Real-ESRGAN), Depth-Anything, Background Removal (MODNet)

VLMS: SmolVLM-500M, FastVLM, Florence-2

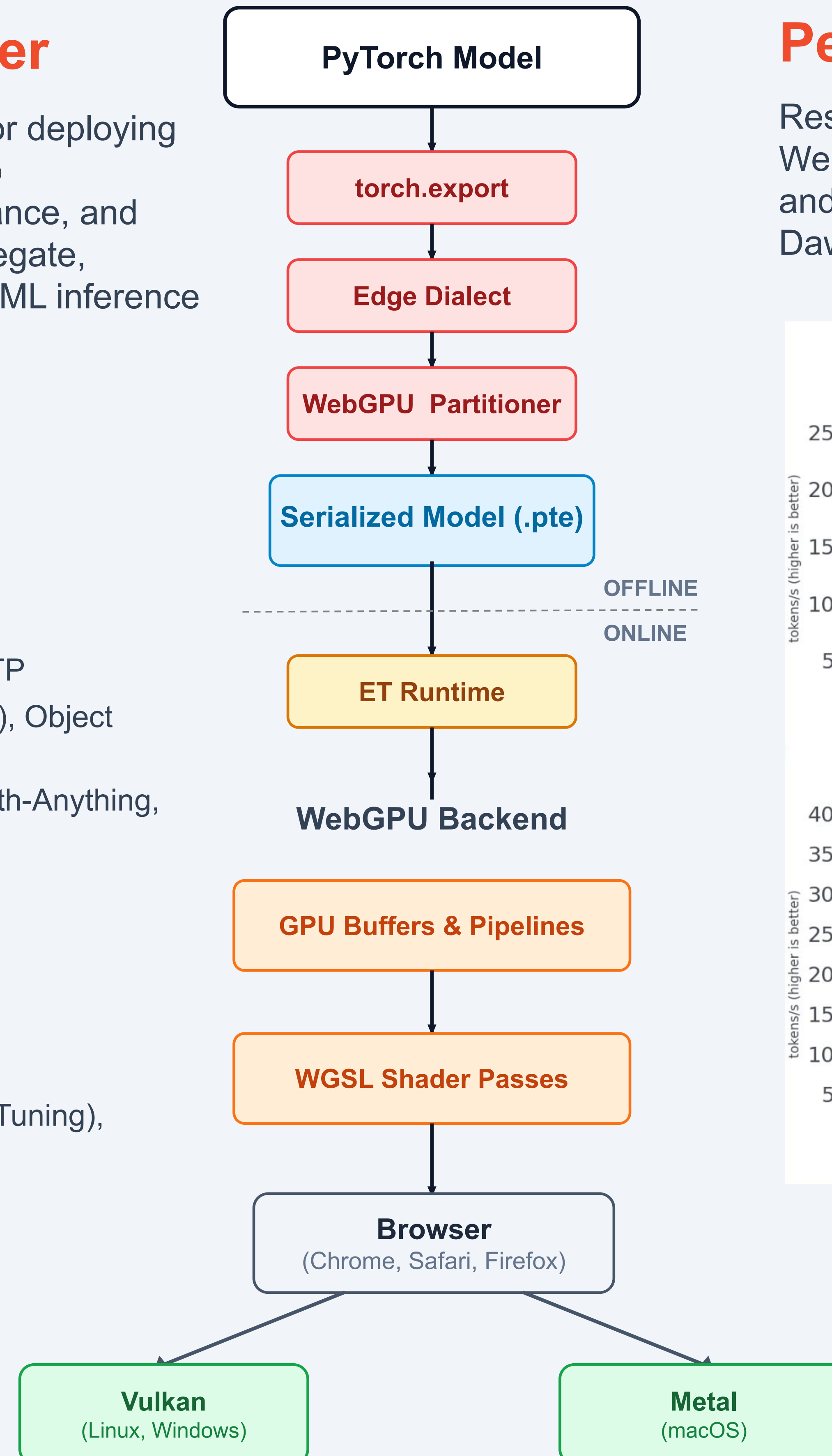
Speech & Audio: Whisper (STT), Voxtral (LLM)

On-Device Fine-Tuning

Models: Llama-3.2-1B, Qwen-3-0.6B

Methods: LoRA, DiReFT (Representation Fine-Tuning), Full-parameter fine-tuning

Use Cases: Personalization, voice adaptation



Performance Analysis

Results are from MacBook M4 Pro (Metal-3). ExecuTorch leverages the WebGPU backend, compared against llama.cpp web, ONNX Runtime Web, and MLC's WebLLM. Both engines run in browser via Chrome Canary through Dawn (Tint).

