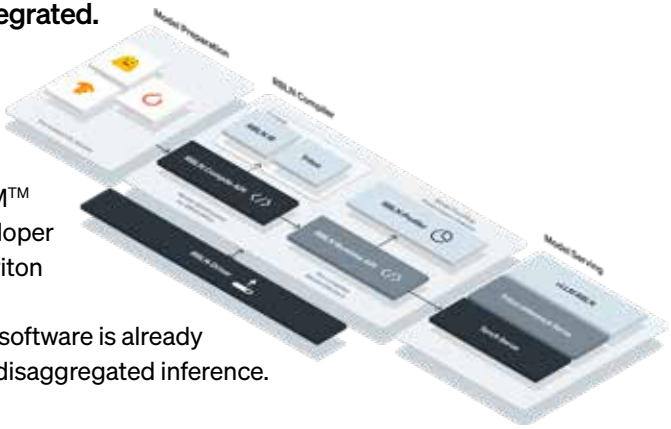
Rebellions SDK

Inference-Class Software. Developer-Class Experience.

- **Built for large-scale inference–without the GPU tax.**
- **Unified stack across ATOM™ and REBEL, from development to deployment.**
- **PyTorch, vLLM, Triton–all integrated. All production-ready.**



Rebellions offers a full-stack AI inference platform—from ATOM™ to REBEL—with a GPU-class developer experience. PyTorch, vLLM, and Triton are supported out of the box, with no hardware-specific friction. The software is already field-proven for high-throughput, disaggregated inference.

Key Features

Built for PyTorch. Tuned for Real-World Inference.

Rebellions software runs natively on PyTorch, supporting both Eager and Graph modes. Graph-mode optimization aligns models with hardware for lower latency and higher throughput. Precision-aware execution (FP32 to FP4) and distributed inference libraries reduce tuning and accelerate deployment.

vLLM Serving, Supercharged.

Rebellions uses vLLM as its native serving layer, optimized for high concurrency and low latency. KV scheduling is tuned for long-context transformers. Hugging Face compatibility and PyTorch-native behavior allow seamless onboarding—no code changes needed.

Triton, Your Way. With Tools that Matter.

Full Triton backend access for kernel-level customization. Supports intuitive APIs, profiling/debugging tools, and optimized dev workflows with real-world examples.

Deploy Fast. Operate with Confidence.

A production-grade software stack for inference across distributed accelerators. Includes rollback, telemetry, cluster orchestration, and built-in tools for deployment, observability, and recovery—all ready for real workloads.