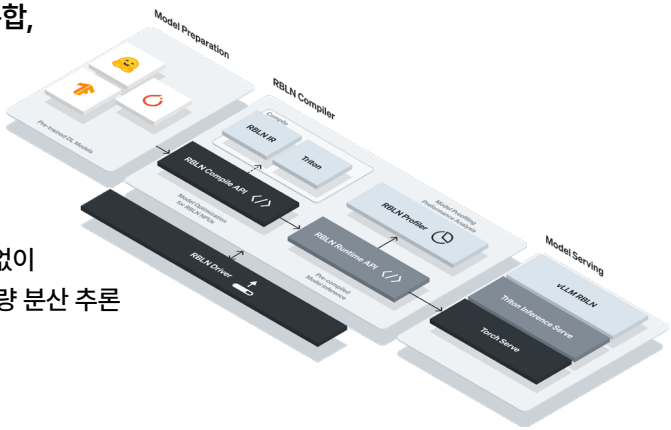
rebellions_

Rebellions SDK

Inference-Class Software. Developer-Class Experience.

- GPU 비용 부담 없이 대규모 추론을 위해 설계
- ATOM™부터 REBEL까지 개발부터 배포까지 아우르는 통합 스택
- PyTorch, vLLM, Triton 완전 통합, 즉시 상용 서비스 투입 가능

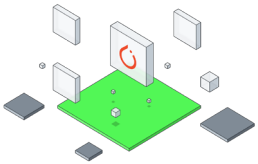


ATOM™부터 REBEL까지 아우르는 풀스택 AI 추론 플랫폼을 제공합니다. GPU급 개발 경험을 그대로 지원하며, PyTorch·vLLM·Triton을 별도 수정 없이 바로 사용할 수 있습니다. 이미 고처리량 분산 추론 환경에서 검증된 소프트웨어입니다.

Key Features

PyTorch 최적화, 실서비스 추론까지 최적화

리벨리온의 소프트웨어는 PyTorch 네이티브 환경에서 동작하며 Eager 모드와 Graph 모드를 모두 지원합니다. Graph 모드 최적화를 통해 지연시간을 줄이고 처리량을 높입니다. FP32부터 FP4까지 정밀도 기반 실행과 분산 추론 라이브러리가 튜닝 시간을 단축하고 배포 속도를 높입니다.



vLLM 서빙, 한 단계 더 강화

vLLM을 네이티브 서빙 레이어로 채택해 고동시성과 저지연 성능을 제공합니다. KV 스케줄링은 장문 컨텍스트 트랜스포머에 맞춰 최적화되었습니다. Hugging Face 호환성과 PyTorch 네이티브 동작으로 코드 수정 없이 손쉽게 도입할 수 있습니다.

Triton, 원하는 방식 그대로

Triton 백엔드를 완전히 개방해 커널 수준 커스터마이징이 가능합니다. 직관적인 API, 프로파일링/디버깅 툴, 최적화된 개발 워크플로와 실제 사례 지원으로 생산성을 높입니다.

빠른 배포, 안정적인 운영

분산 가속기 전반에서 동작하는 프로덕션급 소프트웨어 스택을 제공합니다. 롤백, 텔레메트리, 클러스터 오케스트레이션, 배포·관측·복구용 내장 툴까지 갖추어 실제 워크로드에 즉시 대응할 수 있습니다.

rebellions_