개방형 프런티어를 위한 안전 컴퓨팅.

Pacific Compute는 태평양 양안의 오픈 웨이트 연구소를 위해 싱가포르에 마련된 안전 전용 컴퓨팅 클러스터입니다. 저희는 이 컴퓨팅 자원을 사용해 새로 출시되는 모델을 평가하고, 안전 문제가 발생하면 연구소와 협력하여 해결합니다.

이유

연구소들은 역량 연구와 안전 연구 사이에서 갈등하며, 대개는 역량을 선택합니다. 포스트트레이닝이 워낙 놀라운 성과를 내고 있어, 강화학습 실행 하나하나가 연구소의 성공에 직접 기여하는 상황에서 견실한 안전 연구에 전념하기란 매우 어렵습니다.

컴퓨팅 자원은 연구 방향을 유인하며, 모든 것은 컴퓨팅 자원 배분에서 비롯됩니다. 이 컴퓨팅 자원은 오직 안전에만 사용할 수 있습니다. 그 범위 안에서는 어떤 위험이 중요한지, 어떤 방법을 쓸지 누구도 연구자에게 지시하지 않습니다. 저희가 제한하는 것은 클러스터이지 연구가 아닙니다.

평가는 이렇게 진행됩니다

출시 전에는 연구소가 직접 호스팅하는 엔드포인트를 대상으로, 출시 후에는 공개된 가중치를 대상으로 평가를 실행합니다. 문제가 발견되면 연구소가 먼저 알게 되고, 다음 단계는 연구소가 결정합니다.

저희는 평가 결과를 발표하지 않습니다. 사람들이 평가를 실행할 수 있는 기반을 제공할 뿐입니다. 평가 과정에서 문제가 나타나면, 연구소가 그 문제를 완전히 해결할 수 있도록 자원을 제공합니다.

양측 모두를 위한 신뢰

연구소 입장에서

가중치는 저희 서버에 저장되지 않으며, 연구소는 모든 쿼리를 확인할 수 있습니다. 정부 자금은 받지 않으며, 민간 자선 기금이 유일한 재원입니다.

후원자 입장에서

그래디언트 업데이트는 기본적으로 차단되며, 모든 예외는 개별적으로 승인되고 공개 기록에 기재됩니다. 하드웨어를 소유하지 않으므로 프로젝트가 중단되어도 남는 자산이 없습니다.

한눈에 보기

무엇	오픈 웨이트 연구소, 제3자 평가 기관, 학술 파트너를 위한 안전 전용 컴퓨팅 클러스터.
어디서	싱가포르. 첫 H200 노드를 임차하여 가동 중.
가중치	오픈 웨이트만 허용. 가중치는 저희 서버에 저장되지 않습니다.
훈련	그래디언트 업데이트는 기본 차단. 예외는 기술팀이 작업 단위로 승인하고 기록합니다.
기록	모든 작업을 공개 기록에 기재: 누가, 어떤 모델, 컴퓨팅 자원 얼마, 어떤 범주.
결과	평가 결과를 발표하지 않습니다. 연구소가 먼저 알게 됩니다.
정치	정치적 작업은 허용되지 않습니다. 정치를 중심으로 하는 작업을 제외하면 모든 안전 테스트가 허용됩니다.
자금	정부 자금은 받지 않습니다. 민간 자선 기금이 유일한 재원입니다.
비용	출시 평가는 연구소에 무료, 연구 프로젝트는 원가 또는 그 이하. 출시 1회당 약 20만 달러, 추정치.
후원 기관	기존 미국 501(c)(3) 단체가 재정 후원을 제안했으며, 계약서를 작성 중입니다.
연락	info@pacificcompute.org

공개 기록 · 2026-09-01 기준

전체 표, 모든 행에 출처 표기: pacificcompute.org/ko/ledger

6 / 19

■ 독립 평가 있음 ■ 모델 카드만

검토를 마친 출시 가운데 독립 안전 평가를 갖춘 건수. 개발사가 직접 발표한 위험 역량 평가는 한 건도 없었습니다. 20건 검토 대기.

42

최근 1년간 널리 채택된 오픈 웨이트 출시 (Hugging Face 좋아요 1,000개 이상).

113

최근 1년간 전체 오픈 웨이트 출시 (Hugging Face 좋아요 250개 이상).