

모델 출시 전에 평가를 후원하세요.

출시 전에 연구소가 위험 역량을 발견하고 이해하도록 평가를 후원하세요. 발견을 유용하게 만드는 벤치마크, 재현, 후속 작업을 지원하세요.

2026 · 초기 모금 목표

\$4M

팀, 평가 도구, 컴퓨팅 자원을 마련하고 4분기에 출시되는 일련의 오픈 프런티어 모델을 평가합니다.

2027 · 프로그램 예산

\$12M

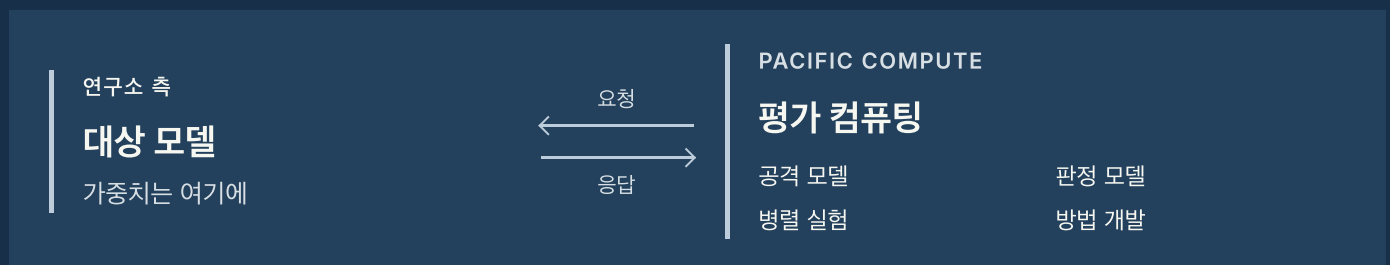
평가 수행과 태평양 양쪽의 오픈 모델 안전 연구 지원을 지속합니다.

프로그램 후원으로 연구자를 확보하고 평가 도구를 유지하며 짧은 출시 전 일정에 맞춰 컴퓨팅을 준비합니다.

안전 연구는 역량 연구와 경쟁합니다.

같은 GPU로 모델 역량을 높이거나 위험을 이해할 수 있습니다. 두 작업이 예산을 공유하면 역량 연구가 우선될 수 있습니다. 우리는 안전 연구에 별도 컴퓨팅을 제공합니다.

GPU가 필요한 이유



엔드포인트 기반 평가 · 첫 H200 노드를 임대해 가동 중입니다.

후원이 만드는 결과.

\$200k

평가당 총액

안전 평가 한 세트의 총액은 20만 달러입니다. 연구자 시간, 컴퓨팅 자원, 신호를 유용한 조치로 이어 주는 후속 작업을 포함합니다.

합의된 시작일부터 결과 전달과 정해진 재평가를 포함해 목표 3주, 최대 4주로 범위를 정합니다.

결과 보고서의 구성

질문

모델 버전, 접근 방식, 조사한 안전성 질문.

결과

근거와 연결된 관찰 내용. 음성 결과나 결론을 내리지 못한 결과도 포함.

재현과 재평가

재현 기록, 합의된 재평가 결과 또는 재평가하지 못한 이유.

테스트 범위

완료한 테스트, 시행 횟수, 조건과 미 실시 항목.

한계

불확실성, 채점 검토, 근거만으로 판단할 수 없는 사항.

자원

컴퓨팅·엔드포인트 사용량, 인력 시간, 저장·전송량과 정산 비용.

6 / 19



코딩된 19개 기록 중 6개에 기준에 부합하는 독립 평가 근거가 있습니다. 공개 보고서 기록이며 Pacific Compute의 평가 이력이 아닙니다. 다른 21개 제품군은 검토 대기 중입니다.

근거와 방법 · 2026-09-01 ↗

위험과 유의사항

평가는 연구소의 모델 개선에도 도움이 될 수 있습니다. 우리는 안전 측면의 이익이 역량 향상에 따른 이익보다 클 것으로 예상할 때 이 상충관계를 받아들입니다. 이 모델들은 어쨌든 출시될 가능성이 높으며, 출시 전 추가 검토를 통해 대응할 시간이 남아 있을 때 위험을 발견할 수 있다는 것이 우리의 판단입니다.

후원자에게 연구소의 평가 결과나 운영 기록 접근권을 제공하지 않습니다.

[안전 연구 범위와 기밀 유지](#) ↗

프로그램 후원을 논의하세요.

info@pacificcompute.org