Pacific Compute

연구소 팀을 위한 안내
샌프란시스코 · 싱가포르

출시 전 안전 평가.

오픈 가중치 공개 전 위험 역량 테스트는 싱가포르 허브의 핵심 임무입니다. 모델 출시 전에 연구소 및 평가 파트너와 협력하면 오픈 프런티어가 더 안전해집니다.

> 연구소의 출시 평가는 무료입니다.



엔드포인트 기반 평가에서는 연구소가 모델을 호스팅하고 접근을 제어합니다. 미공개 가중치는 기본적으로 연구소 측에 두며, 신뢰할 수 있는 협력 관계에서만 더 깊은 접근을 허용합니다. 우리 컴퓨팅은 주변 평가 작업을 실행합니다.

연구소 제어 테넌시

연구소 제어 테넌시는 연구소 엔지니어가 파티션, 모델 키, 승인 사용자를 제어하도록 설계됩니다. 작업 시작 전에 연구소와 직접 구성, 가중치 보관 방식, 보안 요구 사항을 합의합니다.

대응에 필요한 평가 결과.

3–4
주, 합의된 시작일부터

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

> 평가 결과는 연구소가 지정한 팀에 비공개로 전달합니다. 참여를 위해 결과를 공개하거나 협력 사실을 공표할 의무는 없습니다.

결과 보고서의 구성

질문
모델 버전, 접근 방식, 조사한 안전성 질문.

테스트 범위
완료한 테스트, 시행 횟수, 조건과 미실시 항목.

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

한계
불확실성, 채점 검토, 근거만으로 판단할 수 없는 사항.

재현과 재평가
재현 기록, 합의된 재평가 결과 또는 재평가하지 못한 이유.

자원
컴퓨팅·엔드포인트 사용량, 인력 시간, 저장·전송량과 정산 비용.

문제가 발견되면.

대응은 연구소가 결정합니다. 연구소의 요청에 따라 후속 후보 모델을 재평가할 수 있습니다.

기본 원칙은 기울기 업데이트를 하지 않는 것입니다. 평가에서 긴급한 완화가 필요한 위험한 문제가 발견되는 드문 경우, 클러스터 내 연구소가 제어하는 파티션에서 제한적인 지도 미세조정을 지원하고 이후 다시 평가할 수 있습니다.

평가 자료는 연구소와 합의한 목적과 기간에 한해서만 보관합니다. 최소한의 운영 기록은 별도로 취급합니다.

후원자에게 연구소의 평가 결과나 운영 기록 접근권을 제공하지 않습니다.

안전 연구 범위와 기밀 유지 ↗

다음 출시를 함께 논의하세요.

info@pacificcompute.org

신청 양식에서 이메일 또는 WeChat 연락을 선택할 수 있습니다.