太平洋
算力

面向实验室团队
旧金山 · 新加坡

发布前
安全评估。

在开放权重发布前测试危险能力，是我们新加坡中心的核心使命。在模型发布前与实验室和评估合作伙伴协作，意味着更安全的开放模型前沿。

> **实验室免费参与发布评估。**



通过接口开展评估时，模型由实验室托管，访问权限也由实验室控制。未发布权重默认留在你方，仅在可信合作安排下授予更深入的访问权限。我们的算力运行配套评估工作。

实验室控制的租户环境

实验室控制的租户环境旨在让你的工程师掌控分区、模型密钥及获准用户。工作开始前，我们直接与你的实验室商定配置、权重保管和安全要求。

让发现转化为行动。

3–4
周，从商定启动起

评估按目标三周、最长四周规划，从商定启动到交付发现并完成约定复测。

> 评估发现仅向实验室指定团队私下交付。参与不以公开结果或公开确认合作为条件。

发现报告包含什么

问题
模型版本、访问方式及所调查的安全问题。

测试覆盖
已完成测试、试验次数、条件及未测试内容。

发现
附支持证据的观察结果，包括阴性结果或无法得出结论的结果。

局限
不确定性、评分核查及证据无法证明的事项。

复现与复测
复现记录、约定范围内的复测结果，或无法复测的原因。

资源
算力与接口用量、人员时间、存储与传输，以及核对后的成本。

发现问题之后。

实验室决定如何应对。 可应实验室要求对后续候选版本复测。

默认规则是不进行梯度更新。在极少数情况下，如果评估发现需要紧急缓解的危险问题，我们可以在集群中由实验室控制的分区内支持有限的监督微调，随后重新评估。

评估材料仅按与实验室约定的用途和期限保留。最少必要的运营记录另行处理。

资助方不获得实验室评估发现或运营记录的访问权。

安全范围与保密 ↗

讨论你的下一次发布。

info@pacificcompute.org

申请表支持邮箱或微信联系方式。