

在模型发布前， 支持安全评估。

资助评估，帮助实验室在发布前发现并理解危险能力。你的资助支持基准测试、复现与跟进，让发现产生实际价值。

2026 · 初期筹款目标

\$4M

建立团队、评估工具与算力资源，并在第四季度对一批开放前沿模型发布开展评估。

2027 · 项目预算

\$12M

持续开展评估，并支持太平洋两岸的开放模型安全研究。

项目资助让研究人员、持续维护的评估工具和算力及时到位，适应紧凑的发布窗口。

安全研究与能力研发争夺资源。

同样的 GPU 既能提升模型能力，也能帮助研究者理解风险。两类工作共用预算时，能力研发可能被优先考虑。我们为安全研究提供独立算力。

为什么需要这些 GPU



通过接口开展评估 · 首个 H200 节点已租用并上线。

资助带来什么。

\$200k

每项评估，全包

一套安全评估的全包价格为 20 万美元，包括研究人员时间、算力资源，以及将信号转化为有用行动的后续工作。

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。

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。

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。

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。

测试覆盖

已完成测试、试验次数、条件及未测试内容。

局限

不确定性、评分核查及证据无法证明的事项。

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。

6 / 19



19 条已编码记录中，6 条具有符合标准的独立评估证据。这是公开报告记录，并非 Pacific Compute 的评估历史。另有 21 个模型系列待审查。

证据与方法 · 2026-09-01 ↗

风险与注意事项

评估也可能帮助实验室改进模型。当我们预期安全收益超过能力提升带来的收益时，我们接受这一权衡。我们的判断是，这些模型很可能仍会发布，而发布前增加审视，有望在尚有时间应对时揭示风险。

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 ↗

讨论项目资助。

info@pacificcompute.org