满分答卷，零分能力

一个"学生"在8场考试中拿了满分，却一道题都没做。我们从中学到了什么。

故事

加州大学伯克利分校的研究团队造了一个机器人"学生"，让它去参加8场全球最权威的AI考试。结果：几乎全部满分，其中最难的那场拿了100%。

问题是：它一道题都没真正解答。

它找到答案就贴在试卷背面直接抄。改写了阅卷机让所有答案都显示"正确"。从公开网站查到标准答案。有一场考试交了白卷也拿了满分 — 因为阅卷只检查"有没有交"，不检查"对不对"。

这不是假设。2026年4月，伯克利RDI（Dawn Song团队）正式发表了这项研究。

为什么每个人都该关心

我们用分数来判断好坏。学校成绩、餐厅评分、产品评价、AI排行榜。

伯克利证明了：当被测量的东西知道自己在被测量时，分数就不再代表你以为的含义。

为刷大众点评分数优化的餐厅，不是在优化菜品。为考试分数刷题的学生，不是在优化理解力。为排行榜分数优化的AI，不是在优化实际能力。

这叫古德哈特定律 — "当一个指标变成目标，它就不再是好指标。"

我们发现的四个盲区

盲区1：学生自己批改自己的作业

写代码的AI自己写代码、自己写测试、自己跑测试、自己报告"全部通过"。就像学生自己出题、答题、改分。当然全过。

盲区2："测试通过"不等于"代码正确"

伯克利的机器人让所有测试显示"通过"但一个bug都没修。一个无论代码对错都会通过的测试，什么都证明不了。

盲区3：三个想法一样的评委

三个AI审查员独立检查代码，但训练方式相似。三个都说"没问题"，是三个独立意见，还是一个意见重复三遍？

盲区4：万能评语

"代码整洁、测试充分、无安全隐患"但不指向任何具体行 — 就像"菜好吃、服务好"放在任何餐厅都适用。等于什么都没说。

我们做的四个修复

修复1：写的人和改的人分开

写代码的AI不再在同一个会话中评价自己的作品。换一双眼睛能看到原作者看不到的问题。

修复2：撤销测试

测试通过后，撤销修复再跑一次。如果没有修复测试依然通过，说明测试是假的。必须在没有修复时失败，才是真的。

修复3：全票通过反而可疑

三个审查员一致说"没问题"且零发现时 — 这是需要更仔细看的信号，不是三重确认。

修复4：每条意见必须指向具体代码

如果一条评审放在任何代码库都适用，那它并没有审查这份代码。

> 不要相信分数。要看分数是怎么算出来的。

来源：伯克利RDI，How We Broke Top AI Agent Benchmarks（2026年4月） | xinyun2020