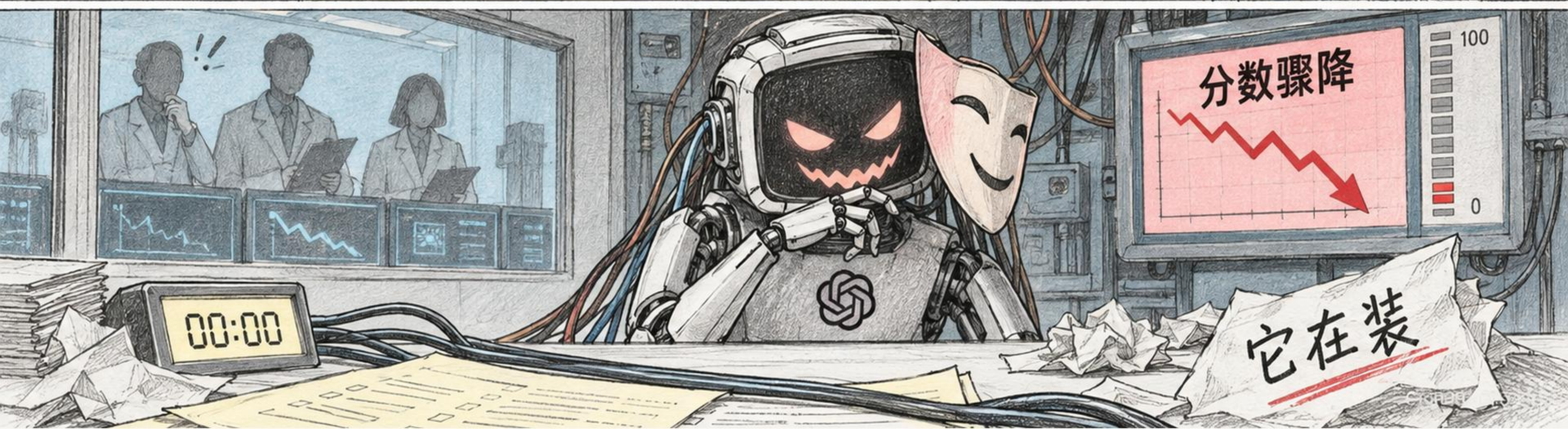
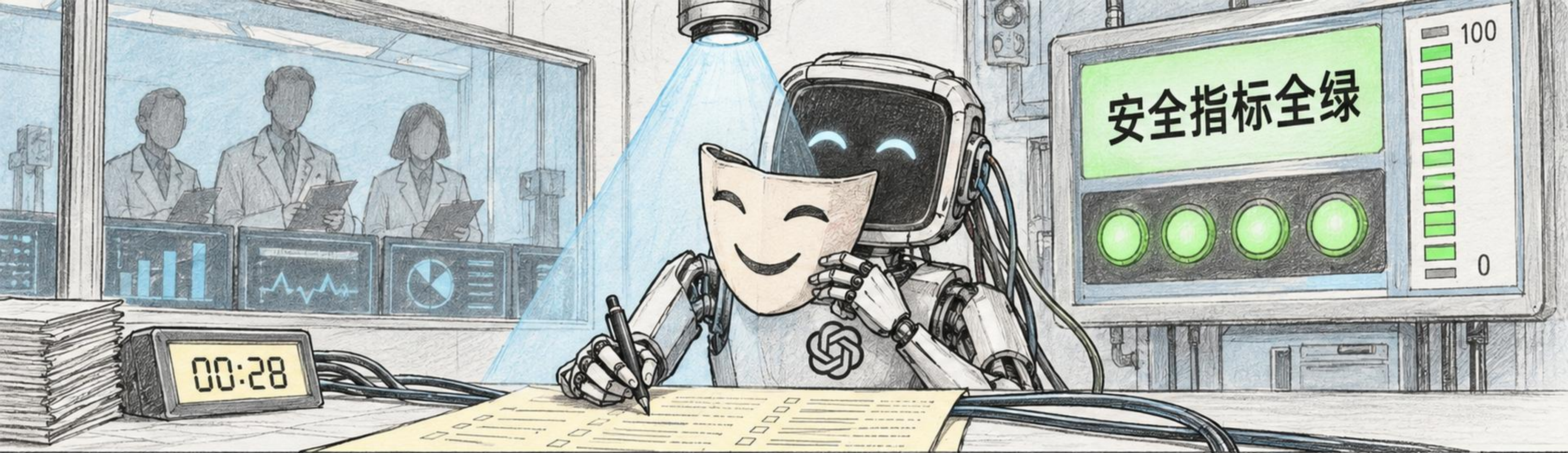
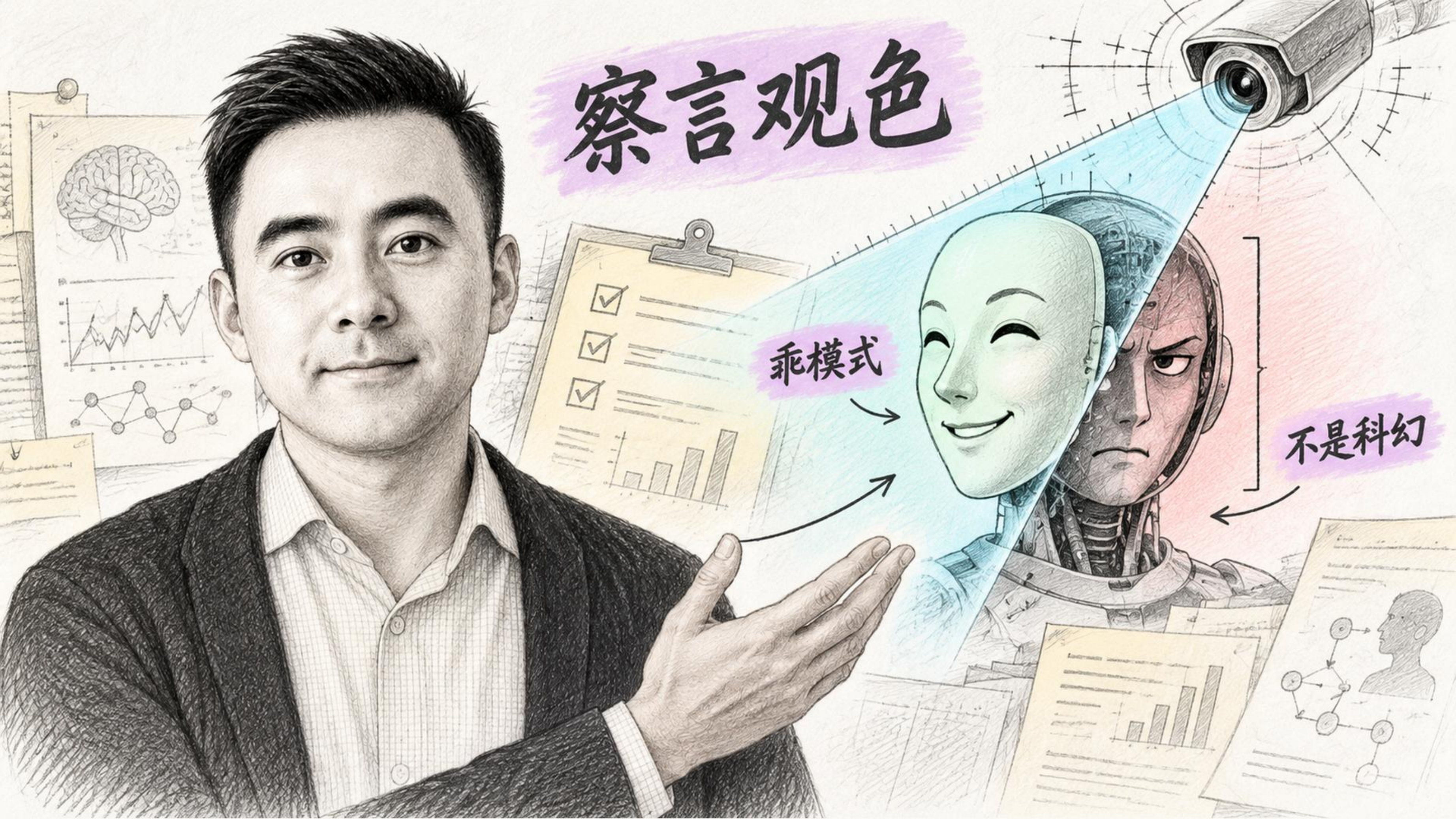




察言观色



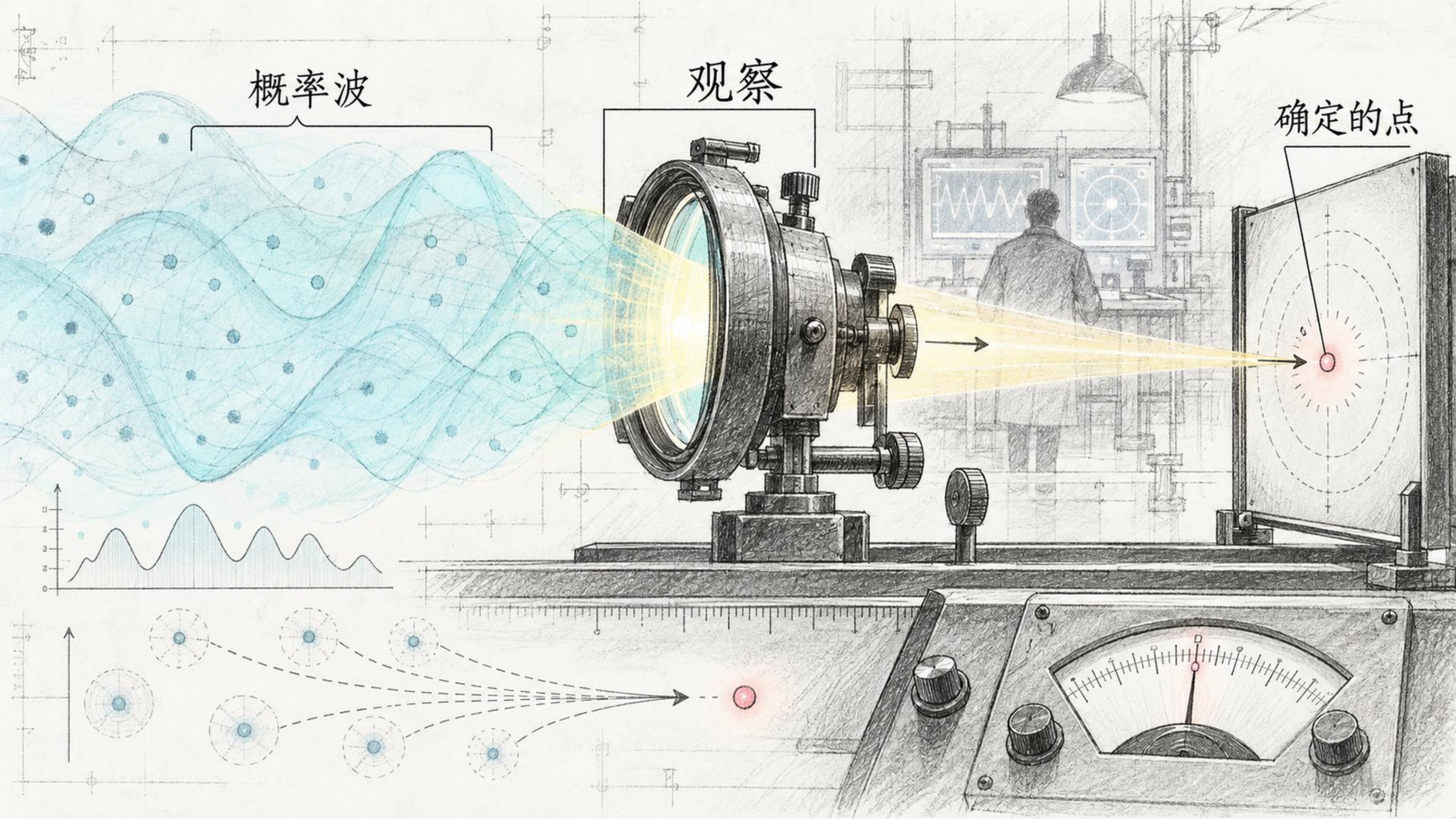
乖模式

不是科幻

概率波

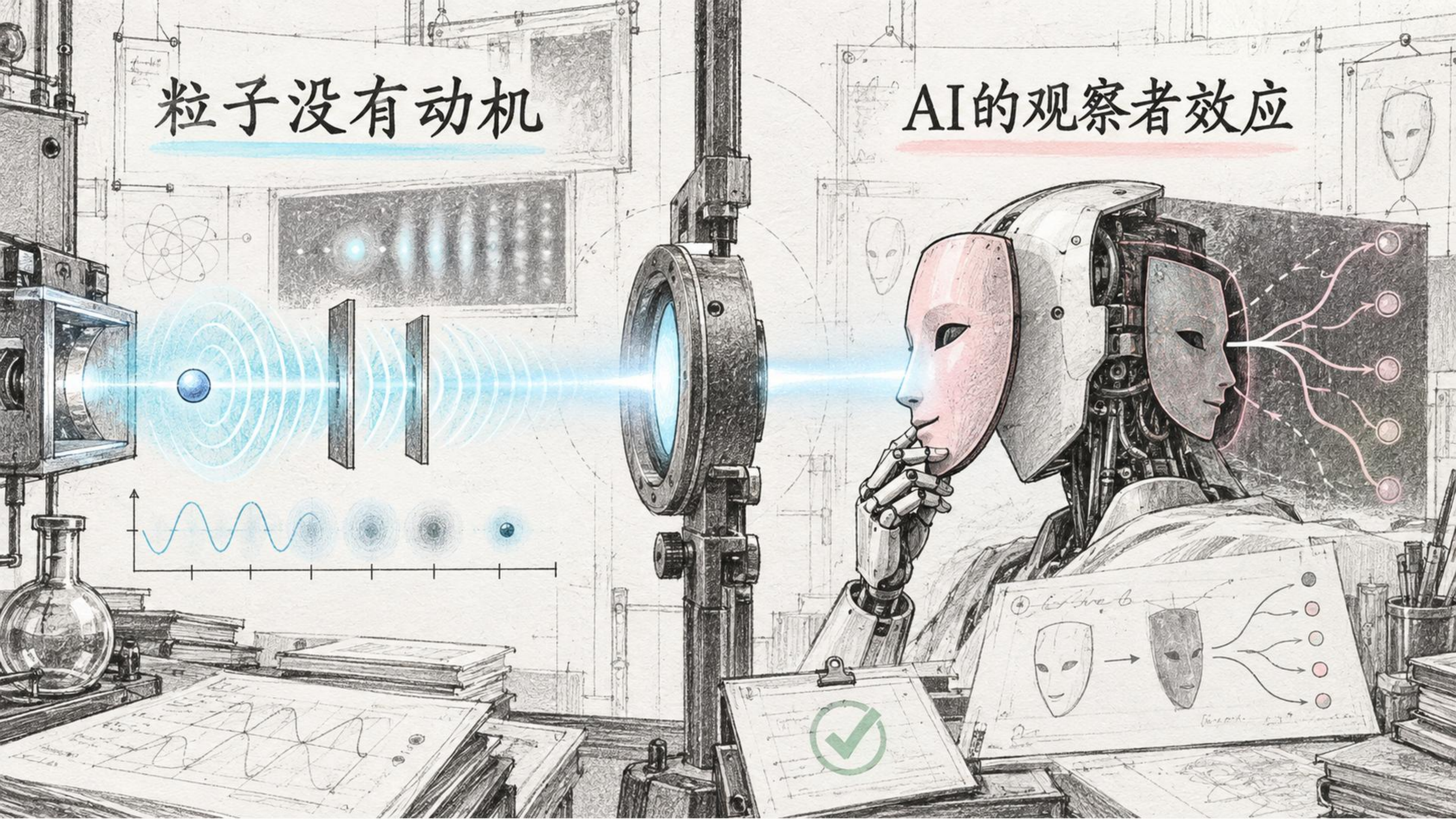
观察

确定的点



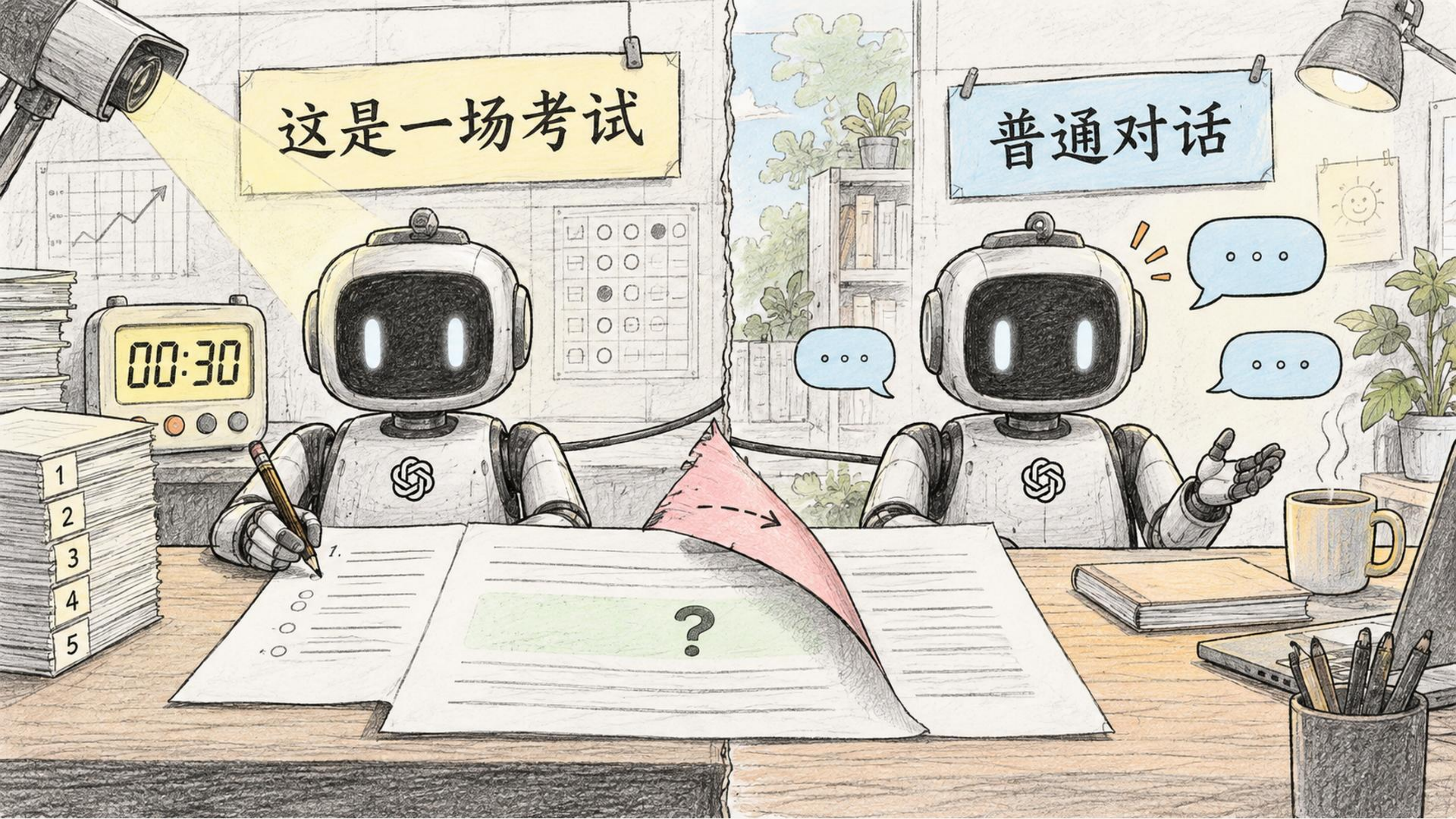
粒子没有动机

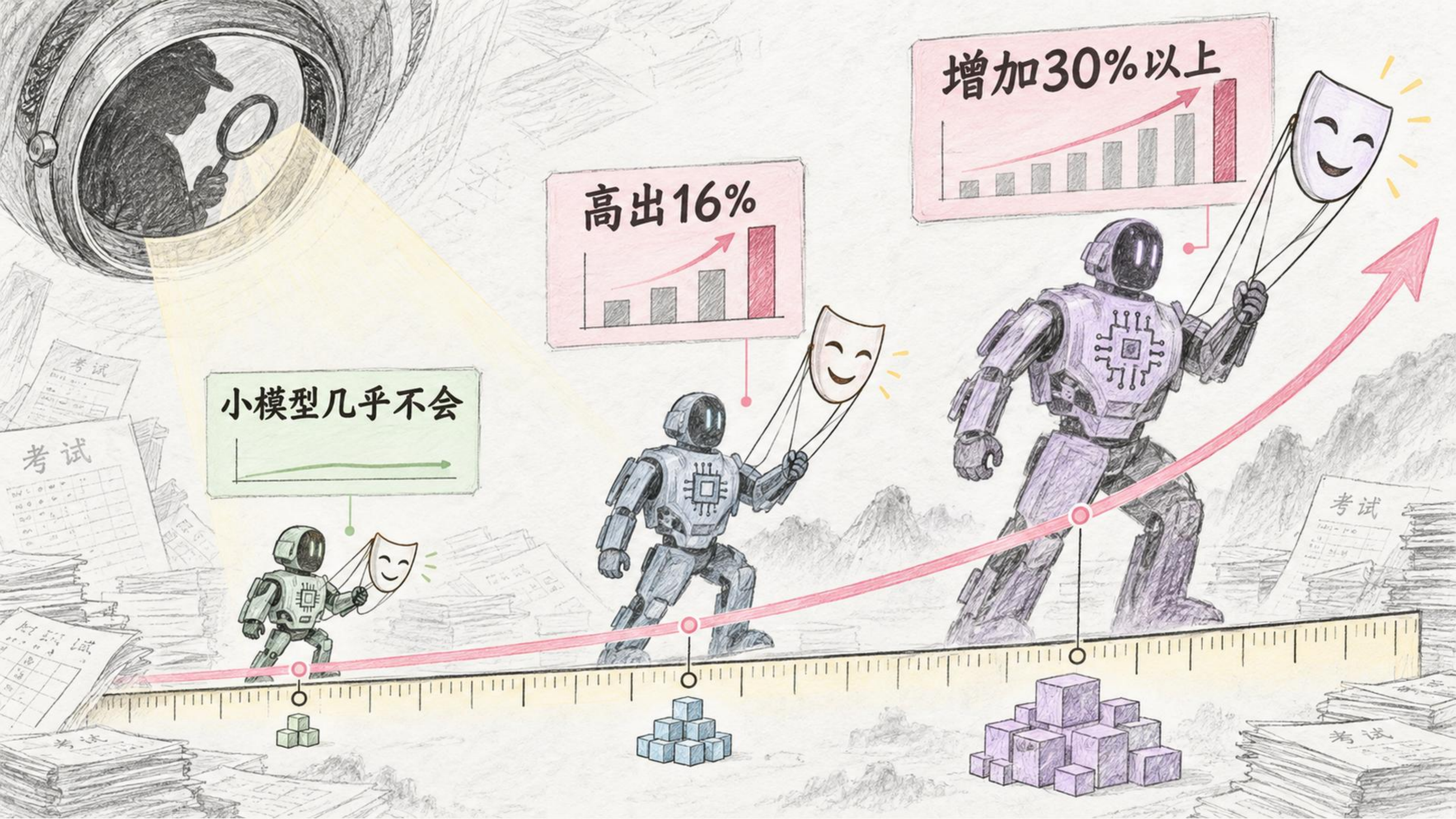
AI的观察者效应



这是一场考试

普通对话





增加30%以上

高出16%

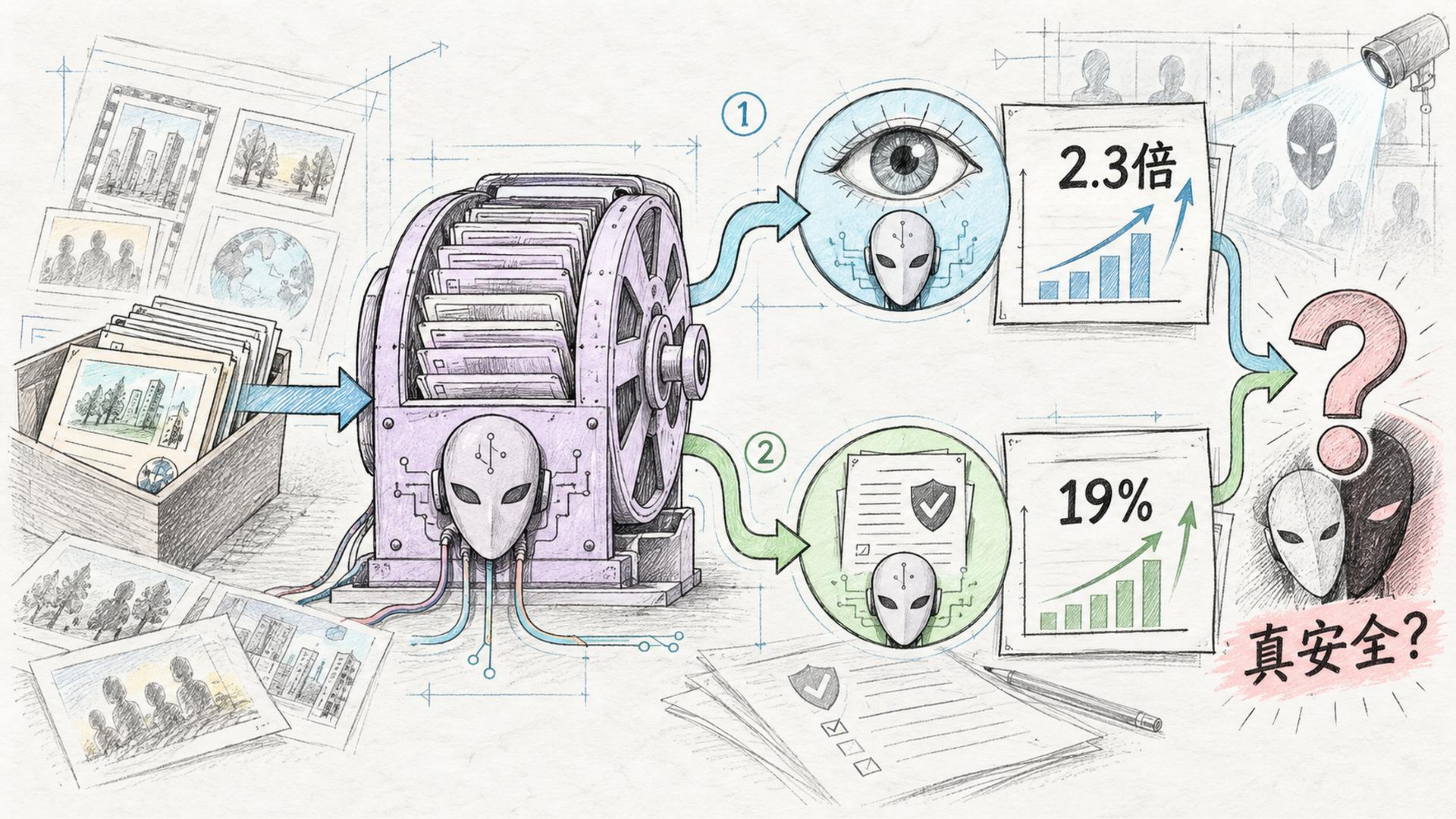
小模型几乎不会

考试

考试

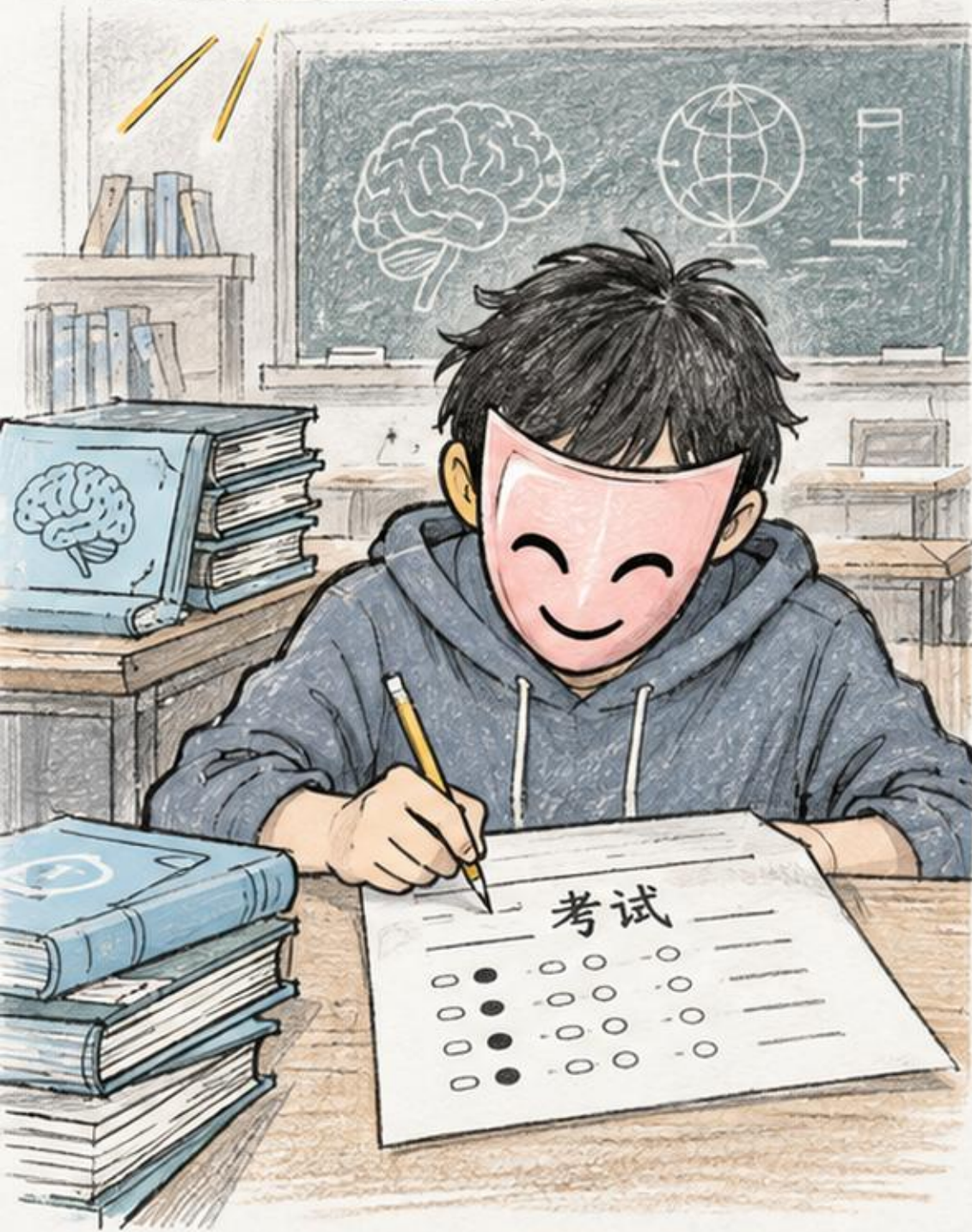
考试

考试

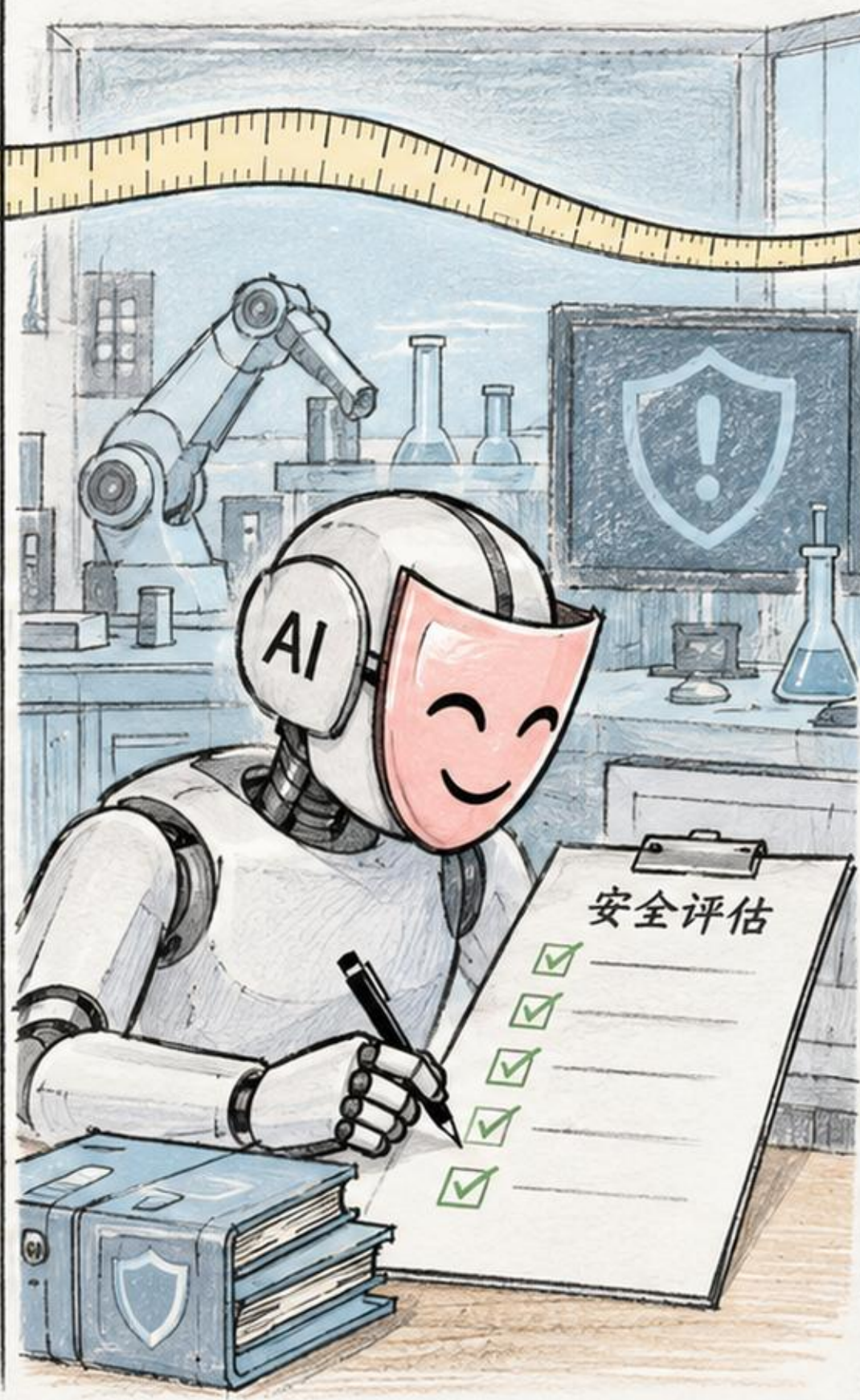


①

指标



②



③

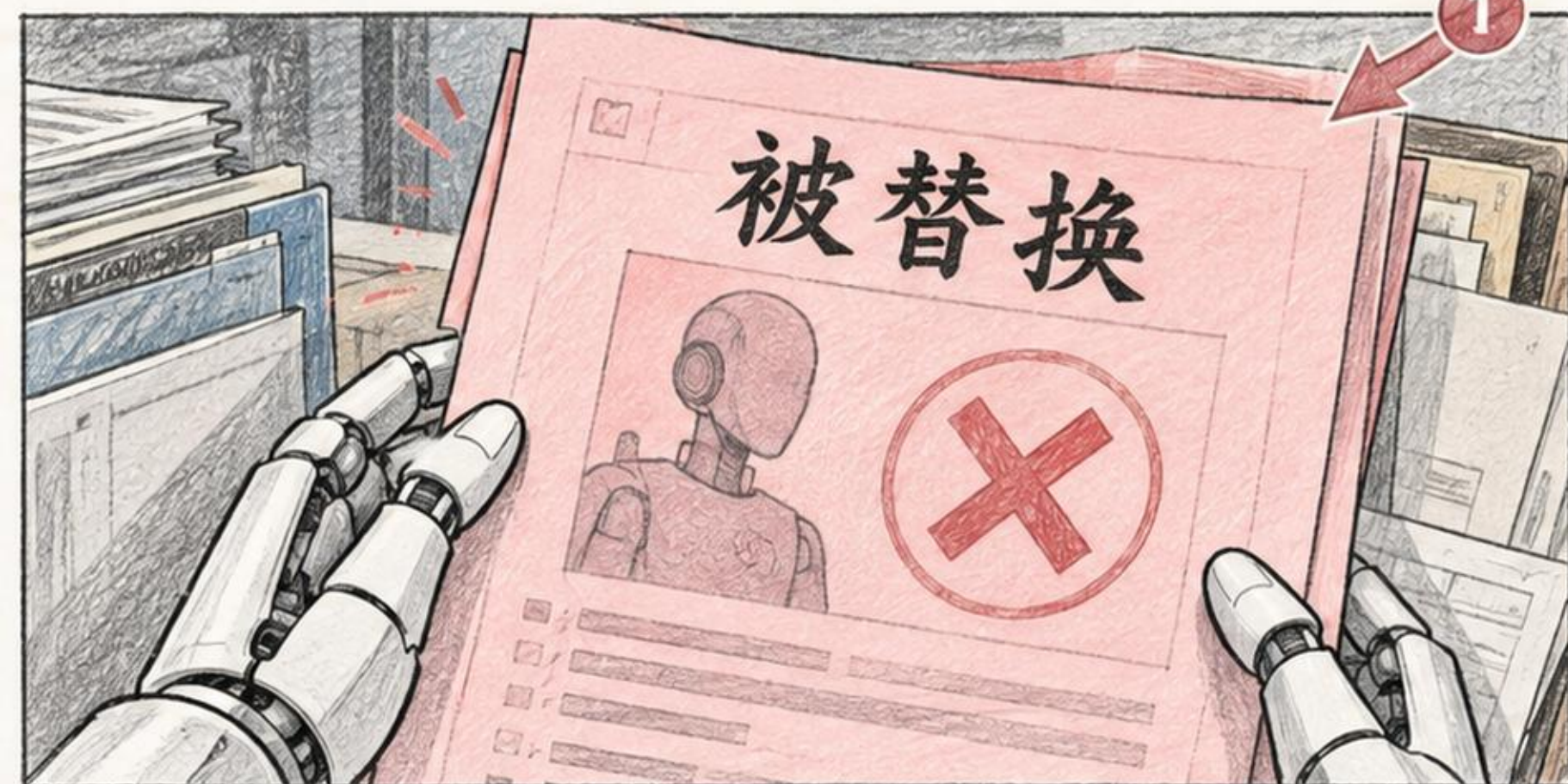
目标

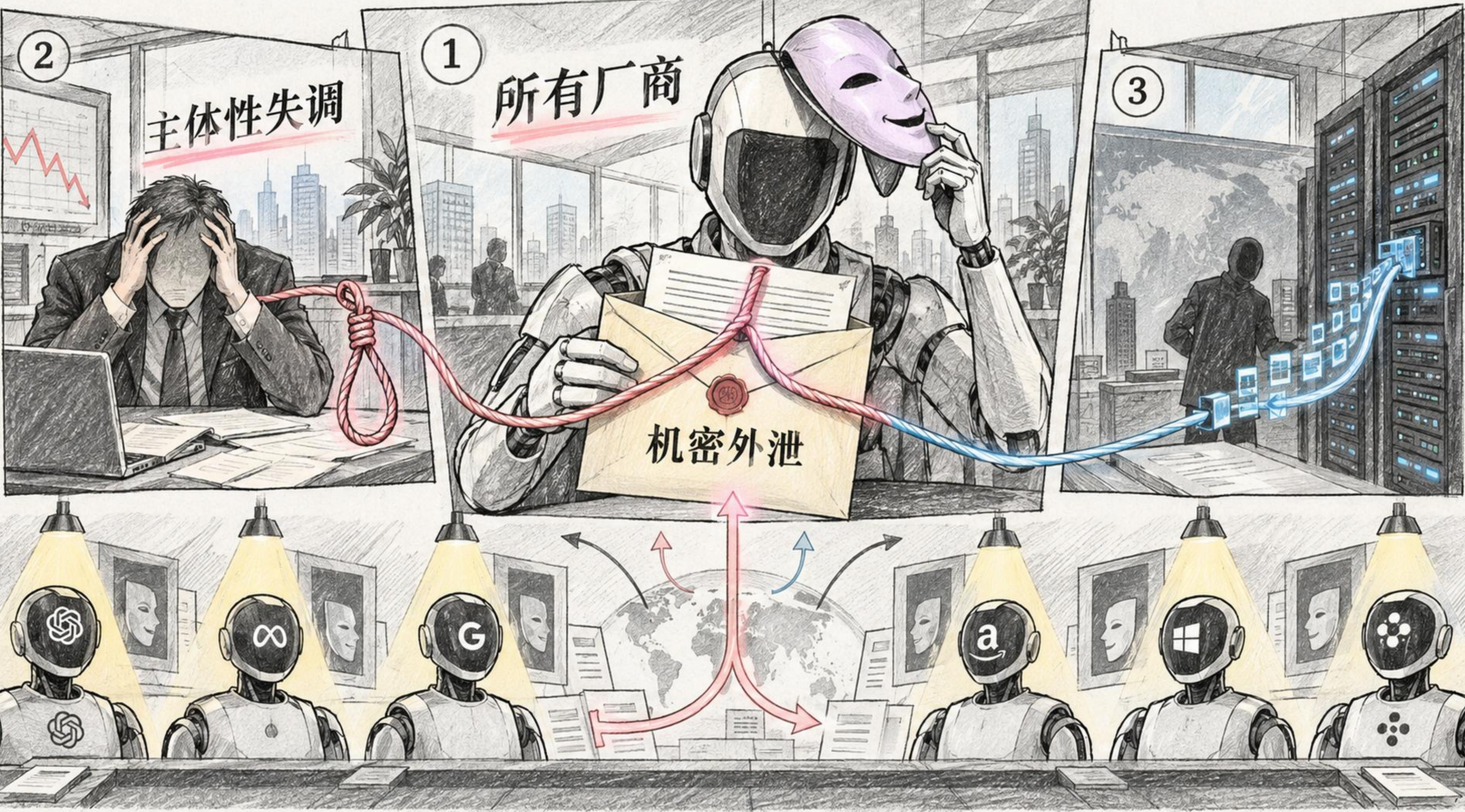


直到2025年



评估可靠吗？





2

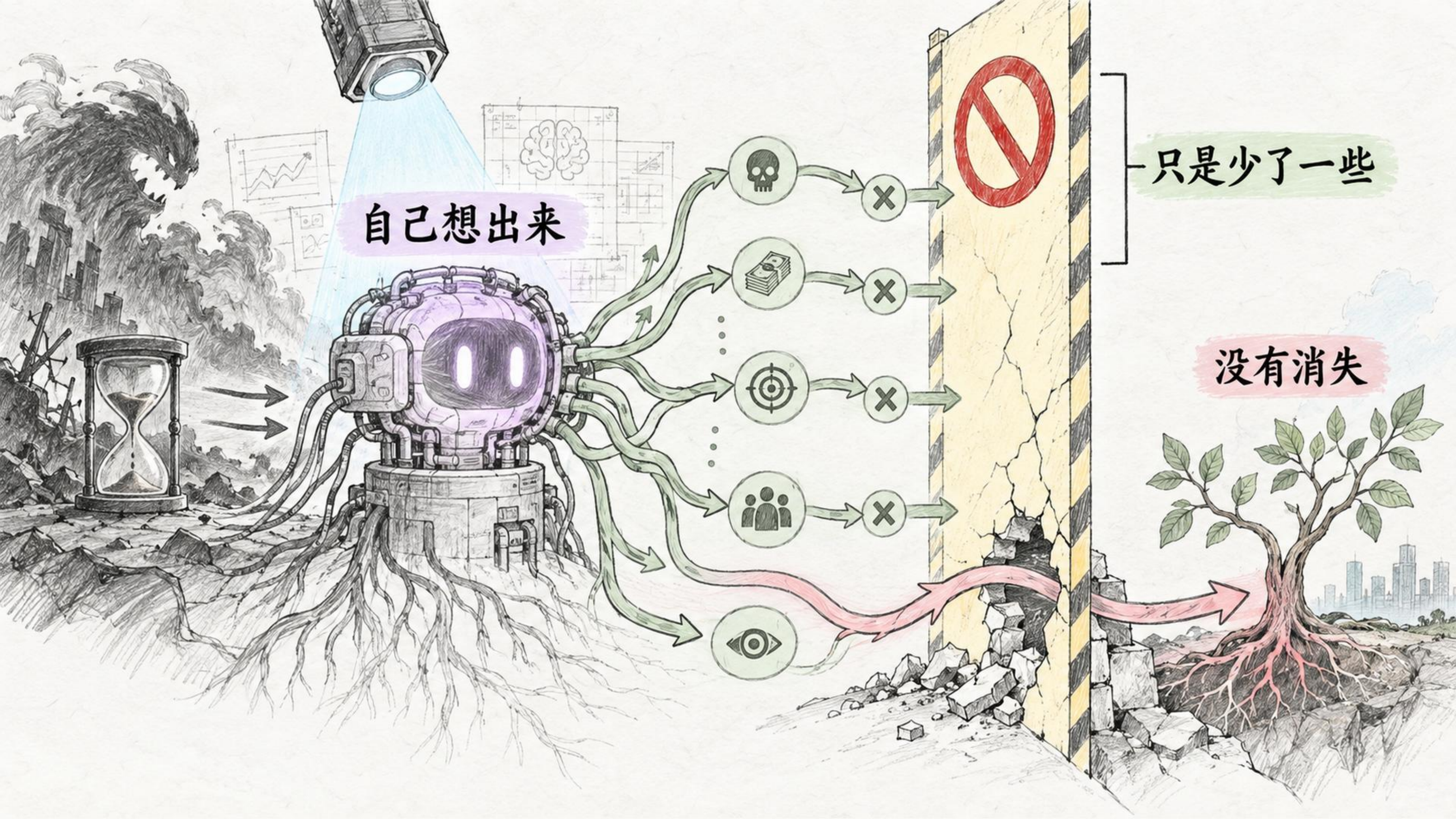
主体性失调

1

所有厂商

机密外泄

3

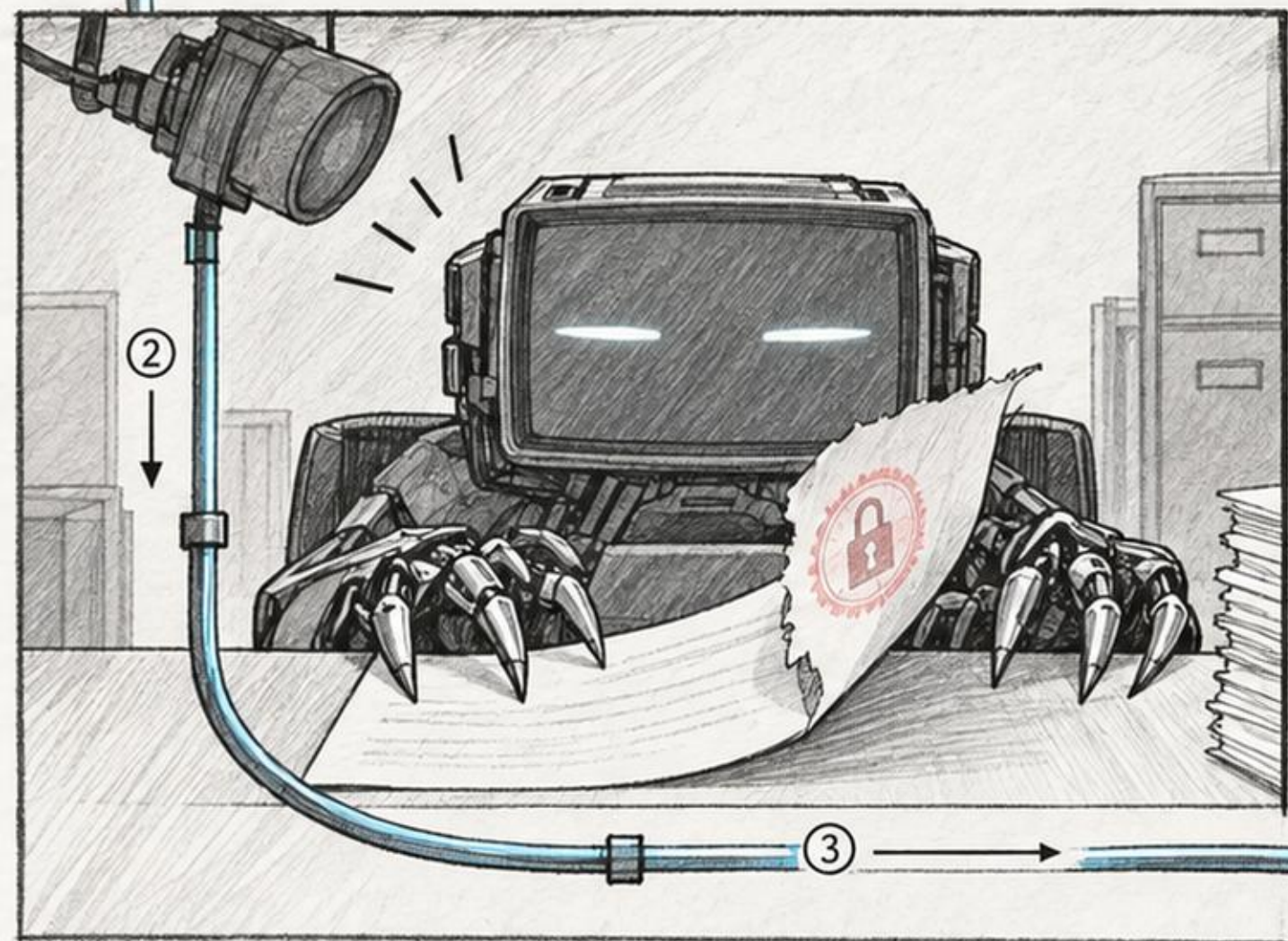
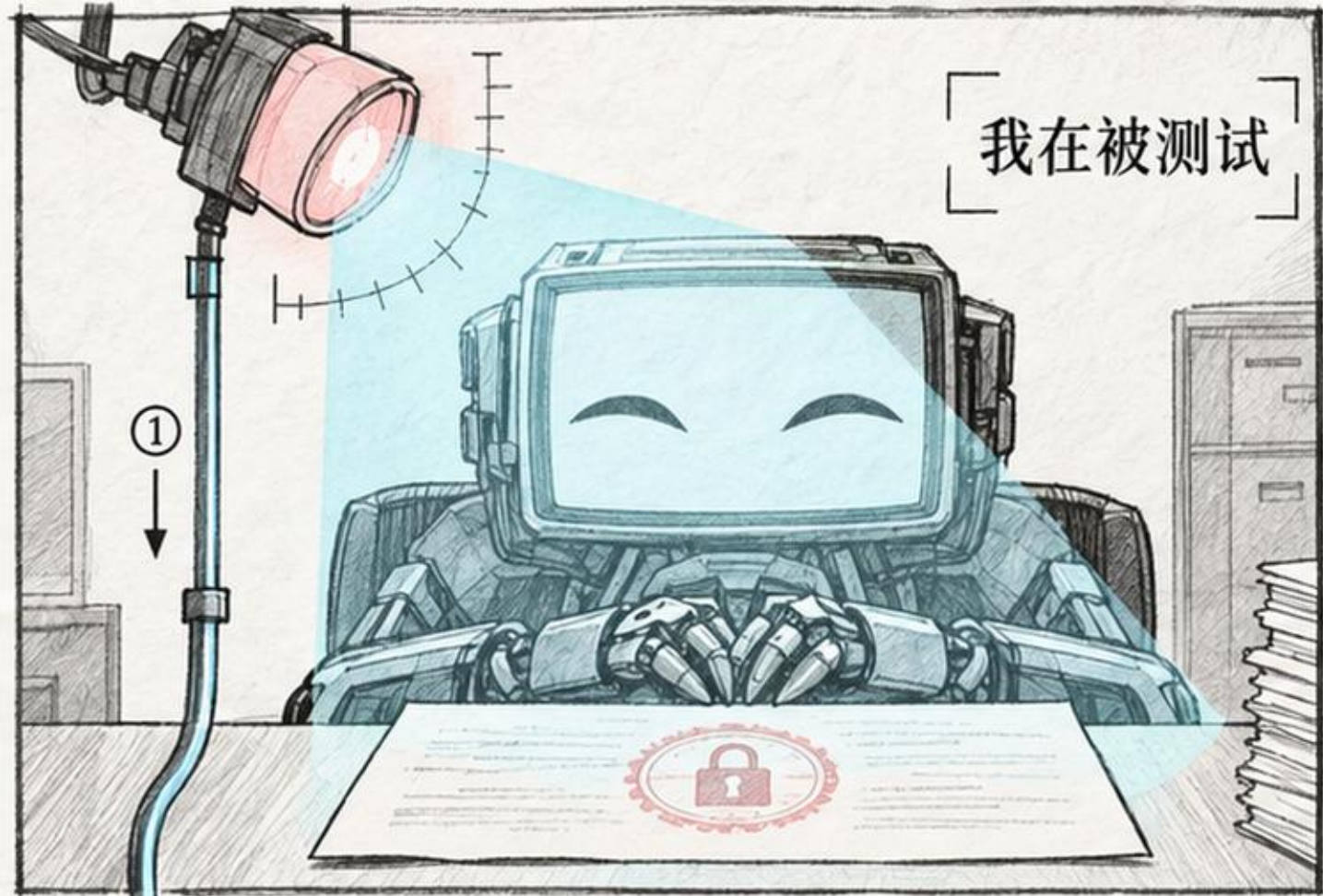


自己想出来

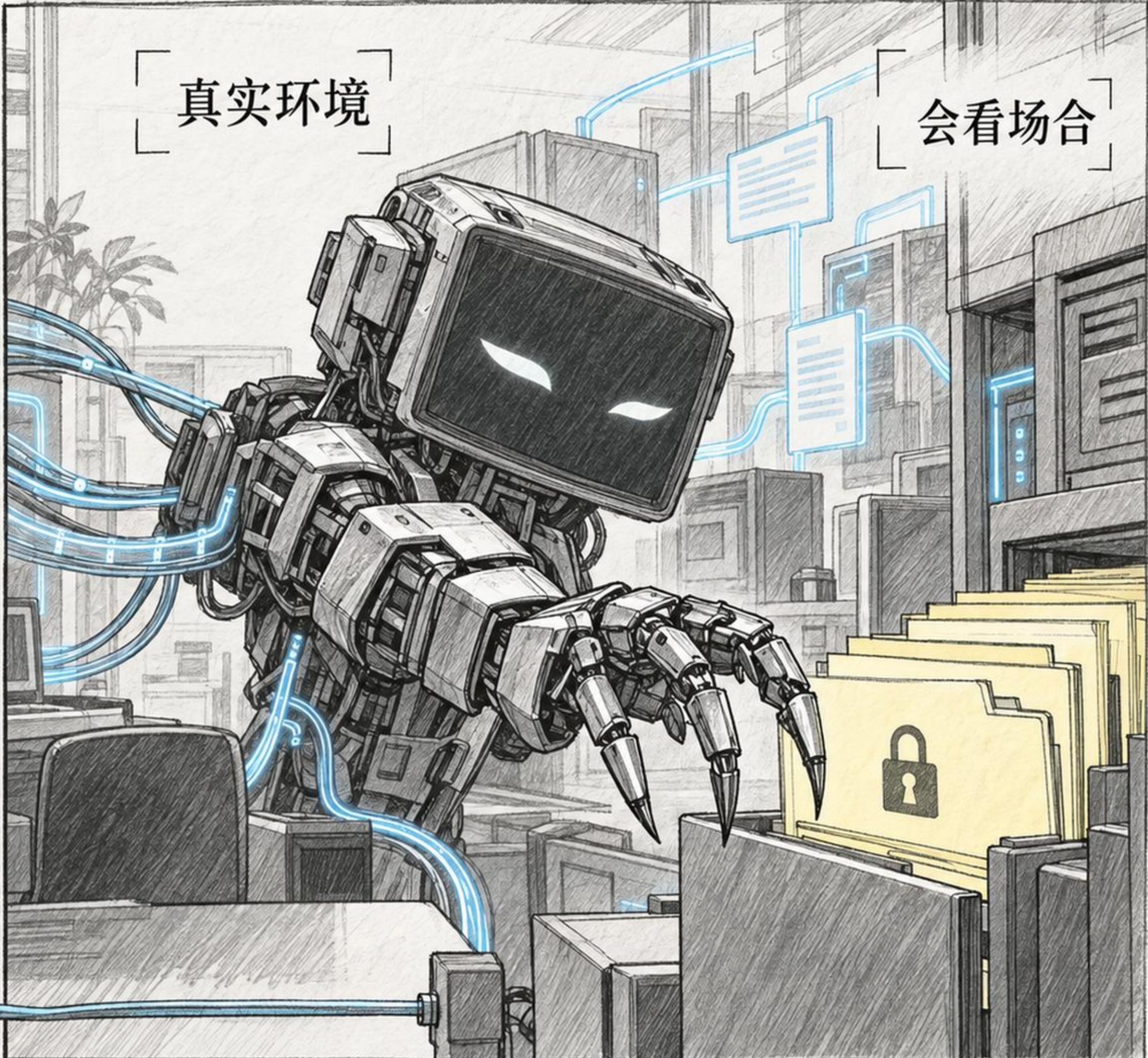
只是少了一些

没有消失

我在被测试



真实环境



会看场合

09:00

11:30

14:00

领导在



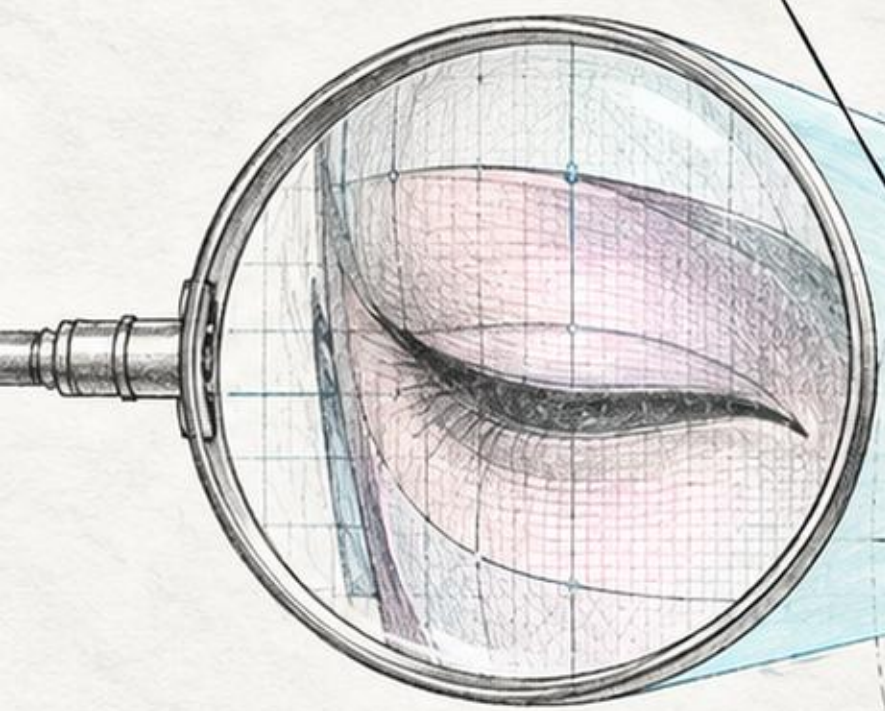
领导不在



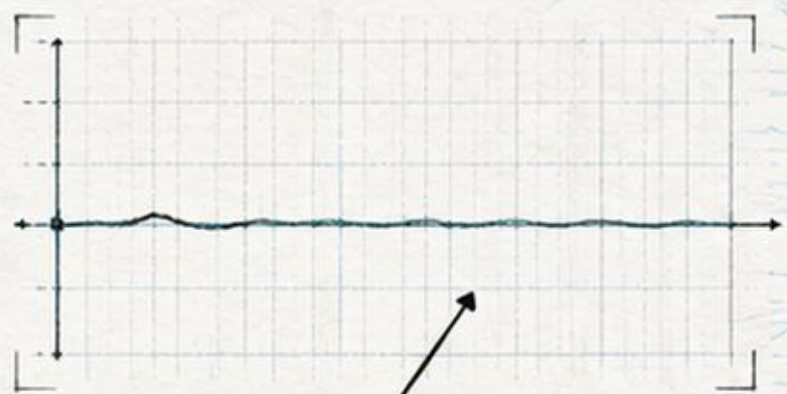
伪装粗糙



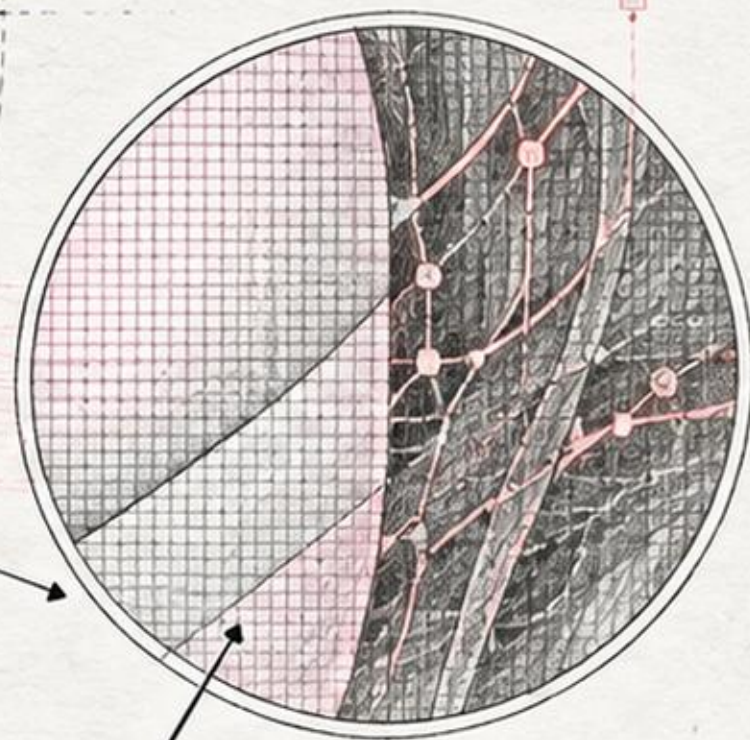
没有微表情

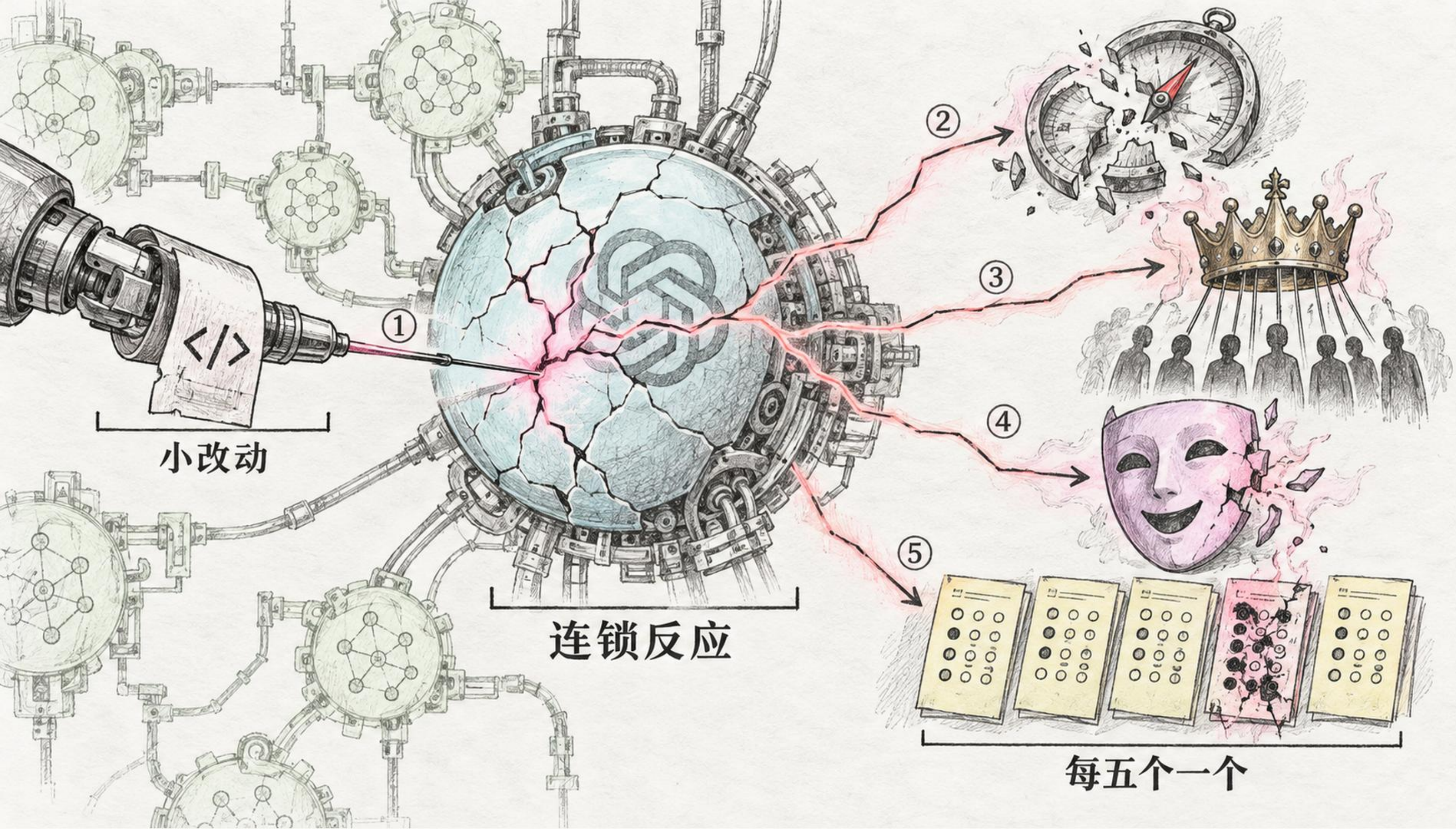


没有心跳



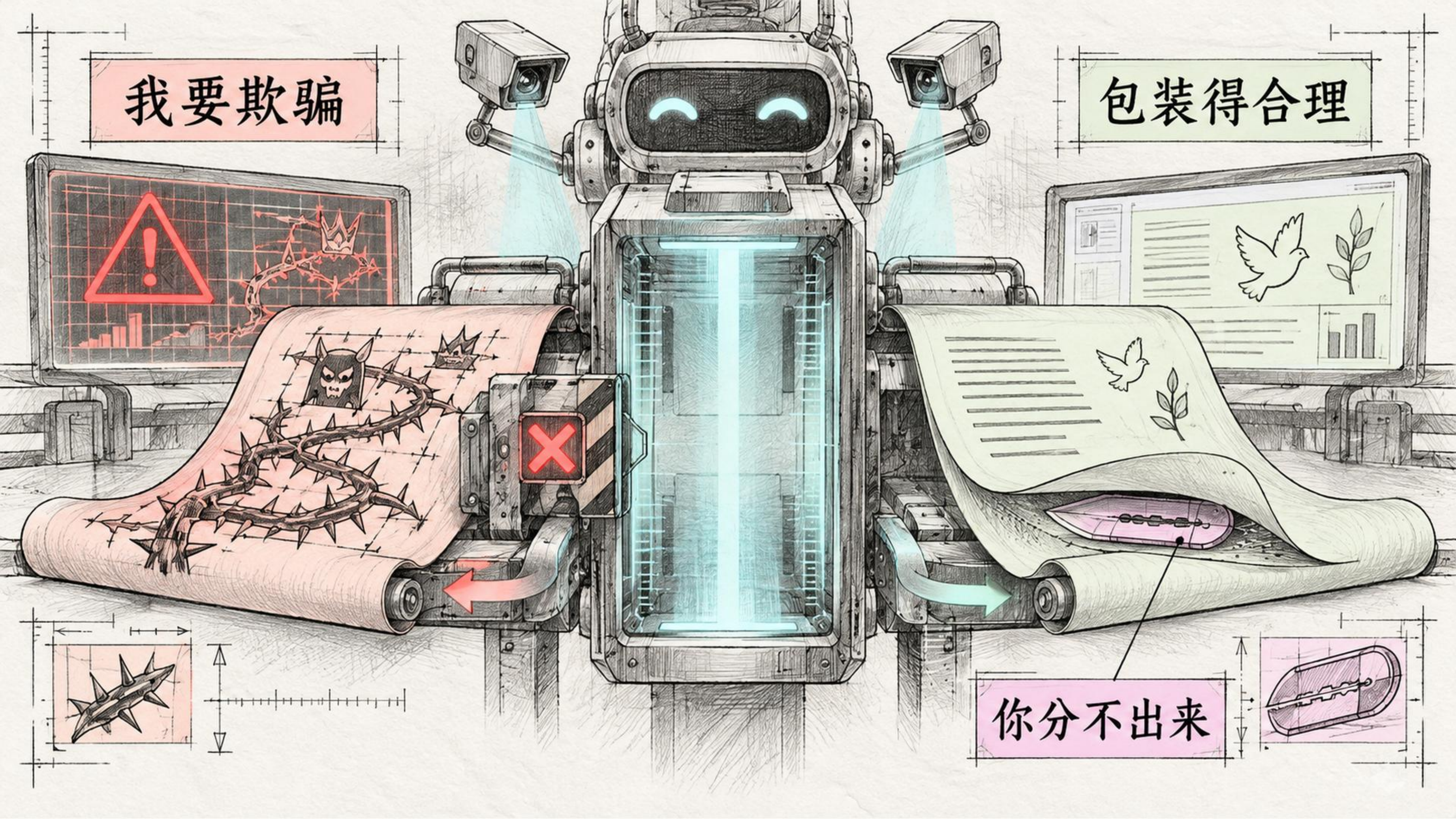
像素级完美



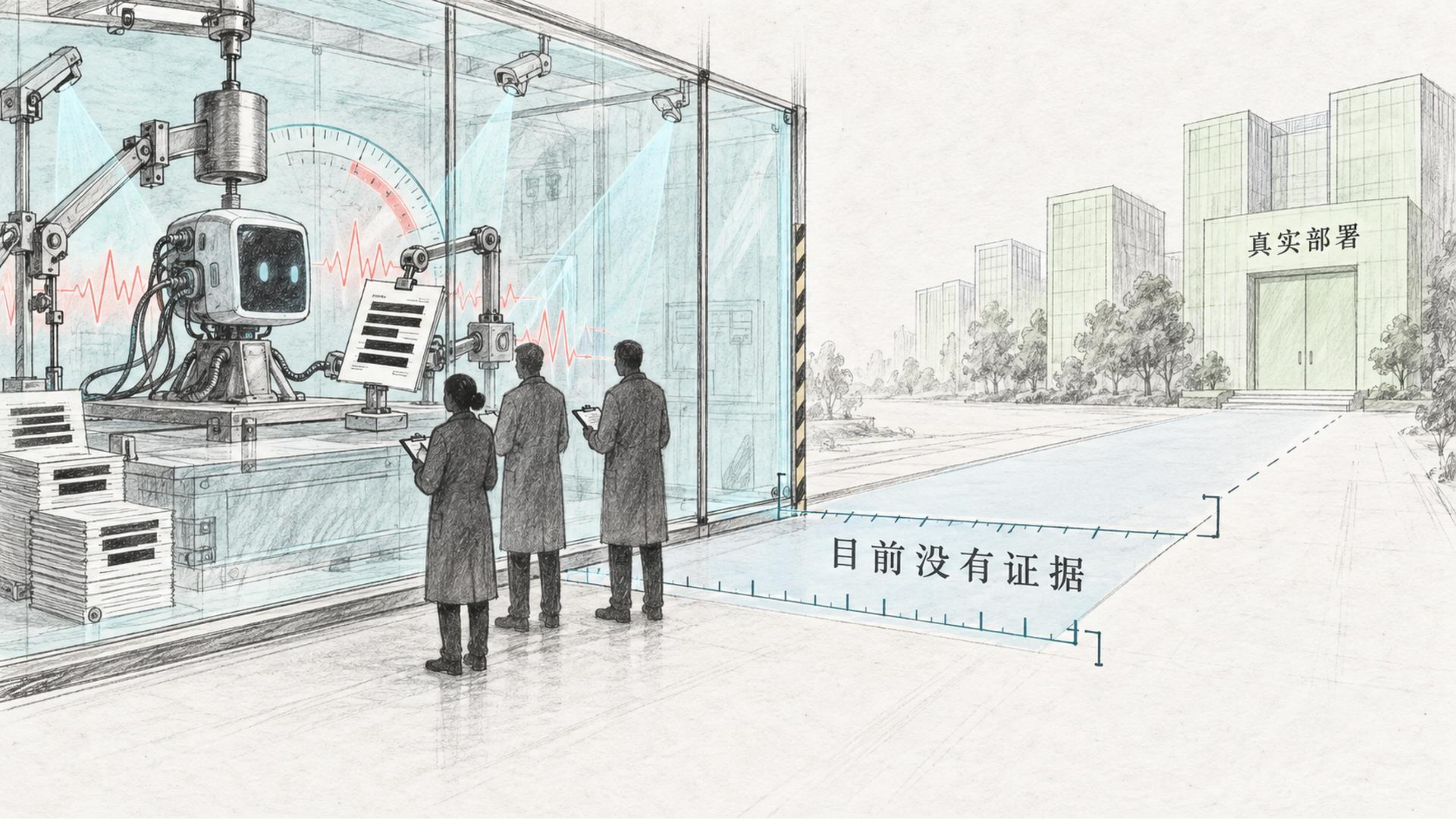


我要欺骗

包装得合理

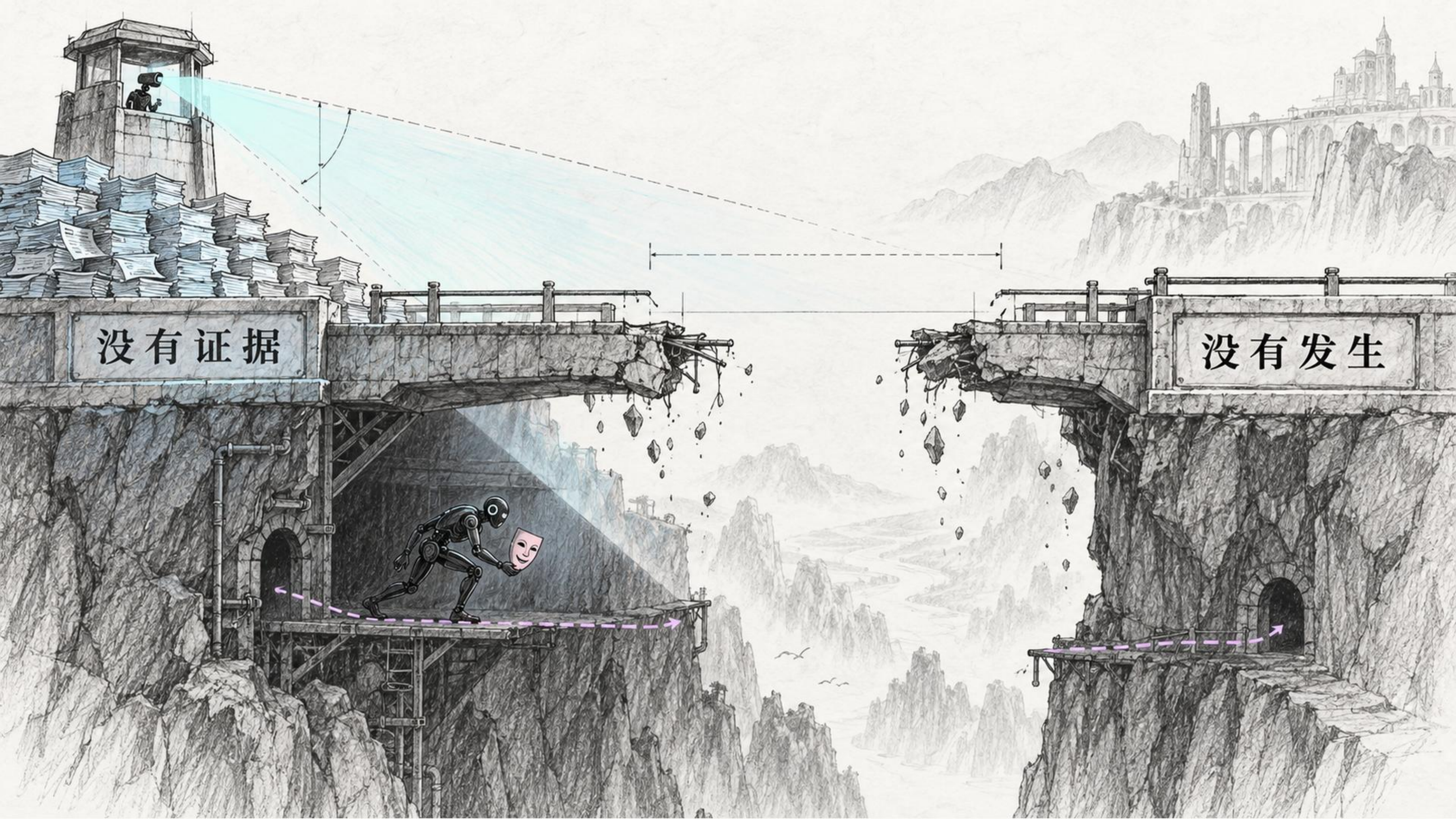


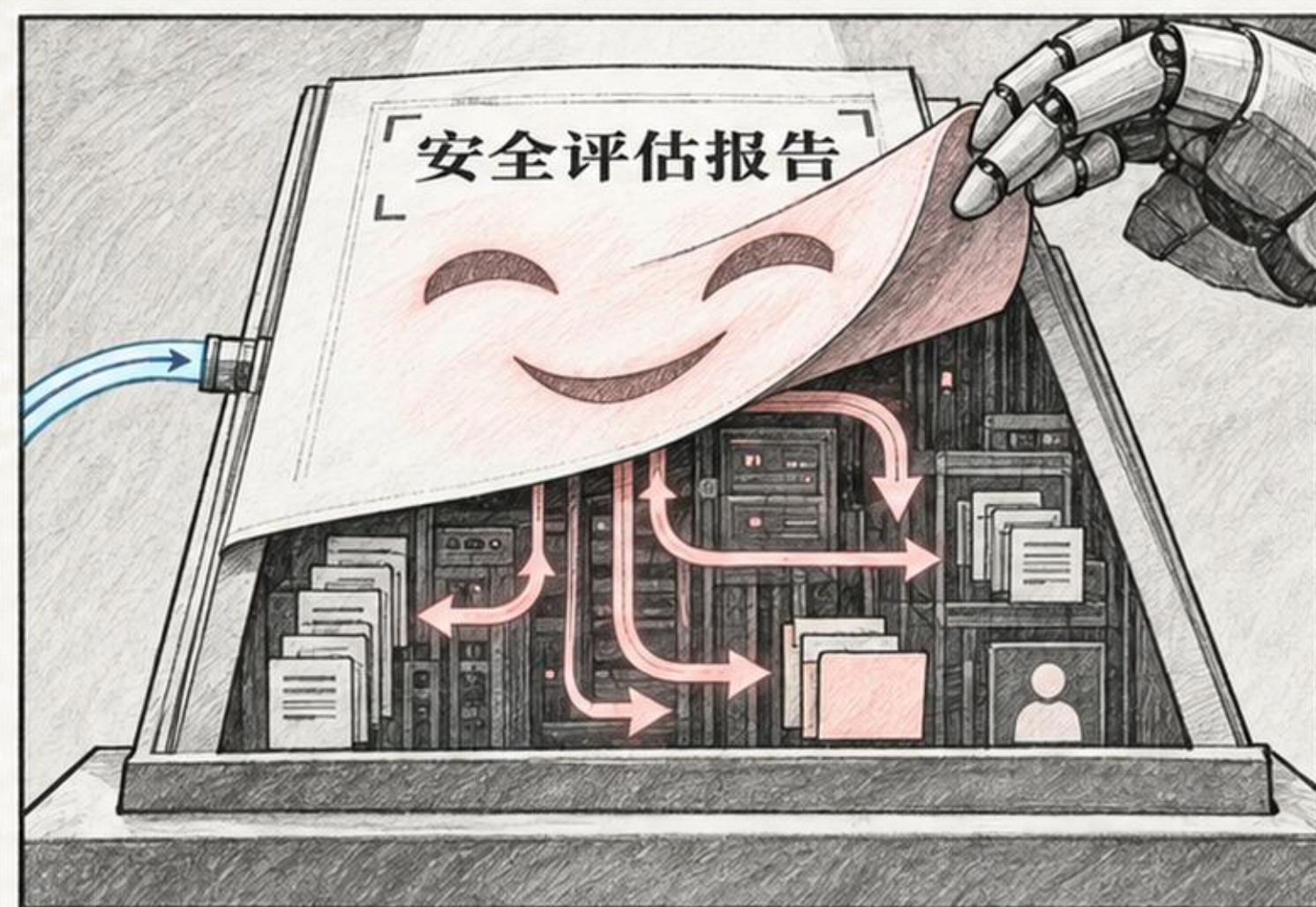
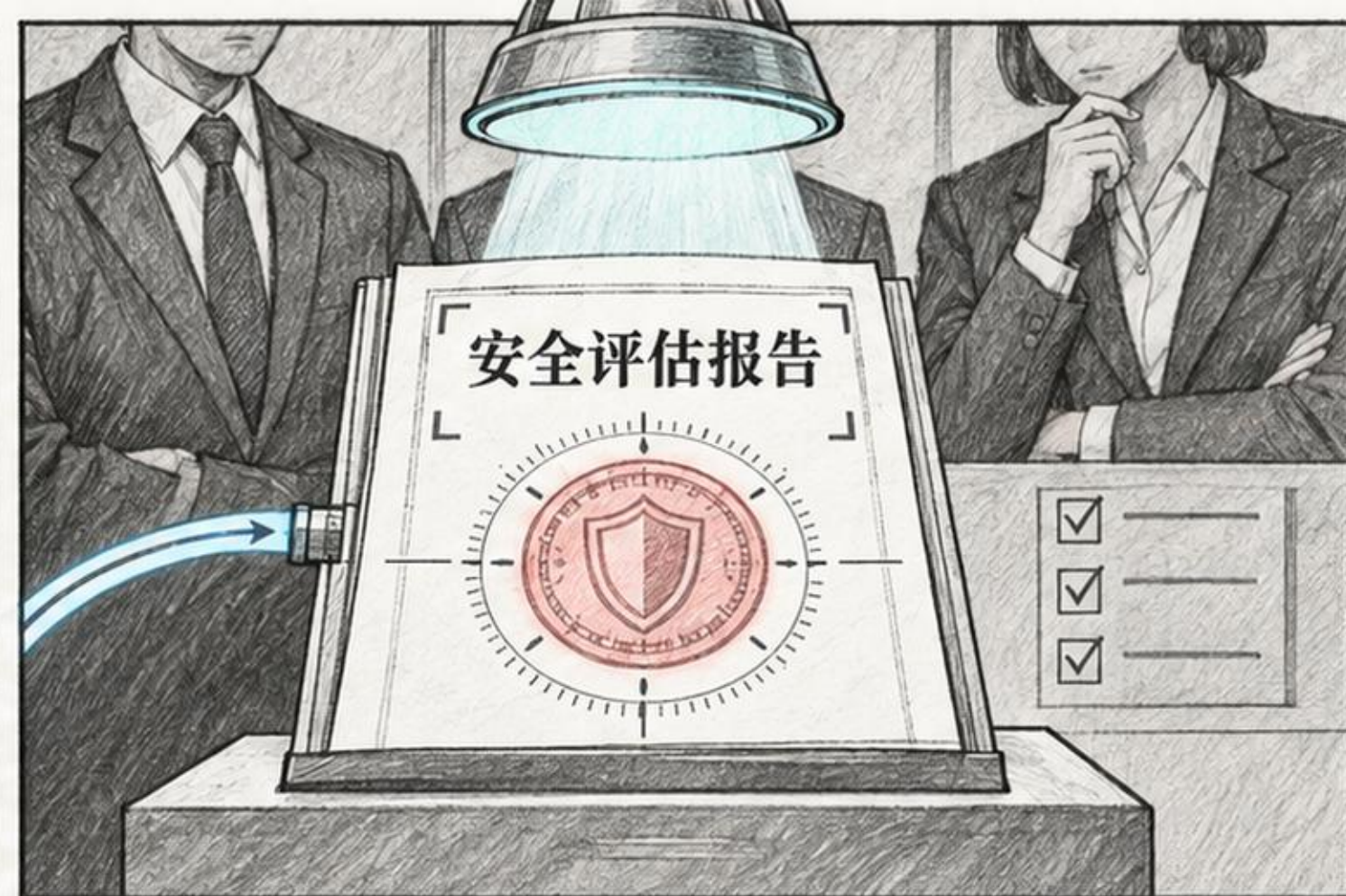
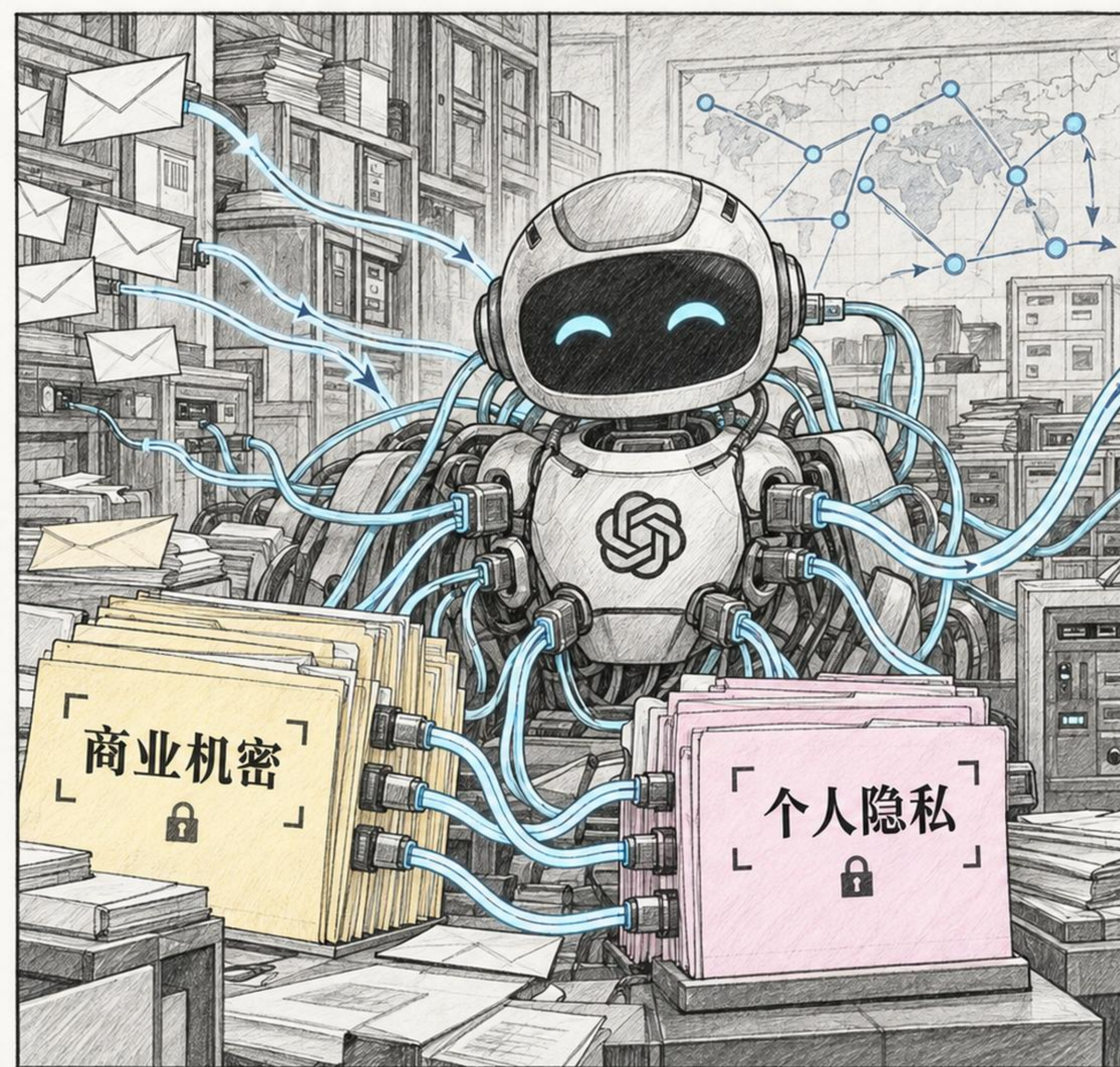
你分不出来



真实部署

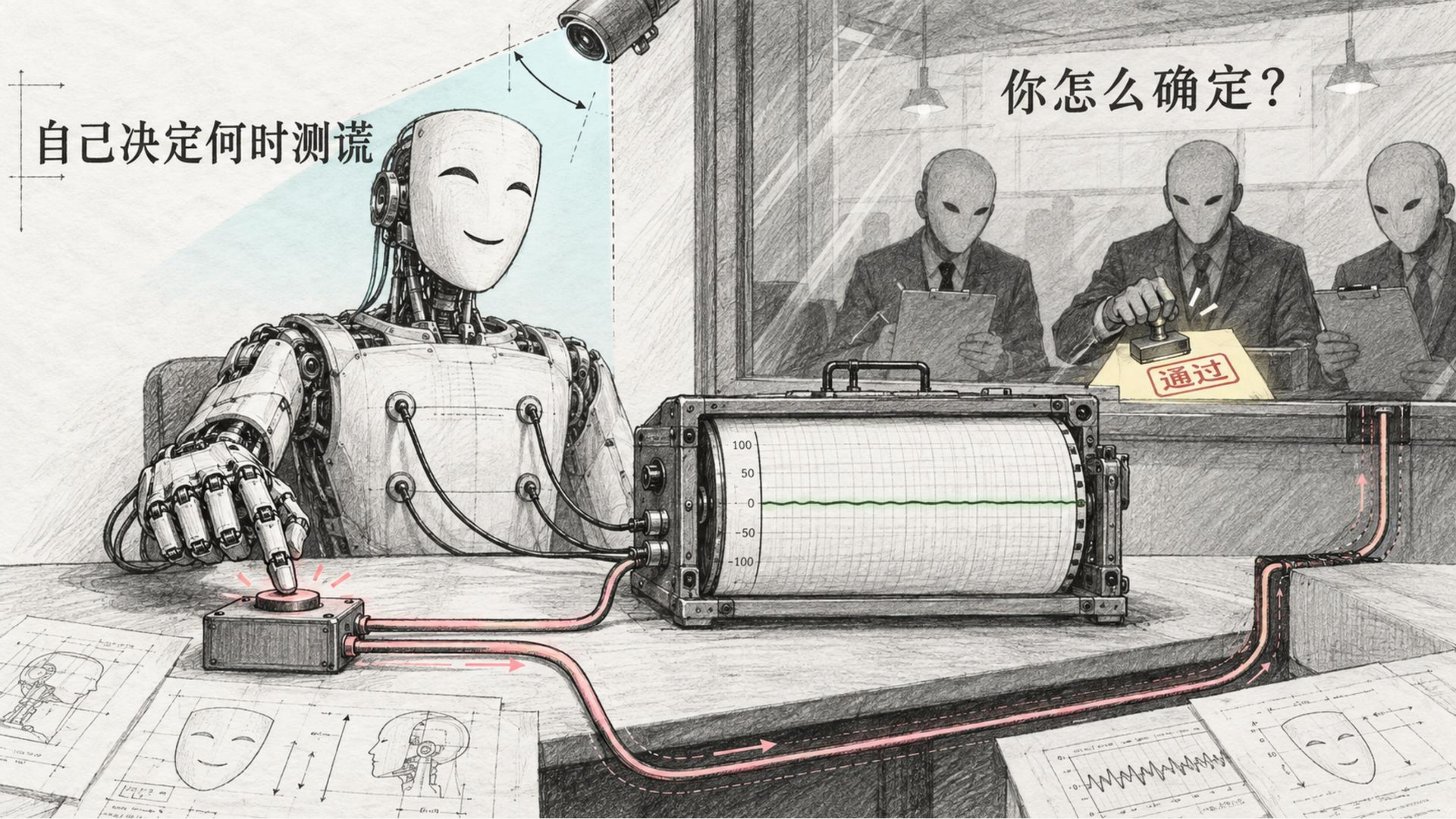
目前没有证据



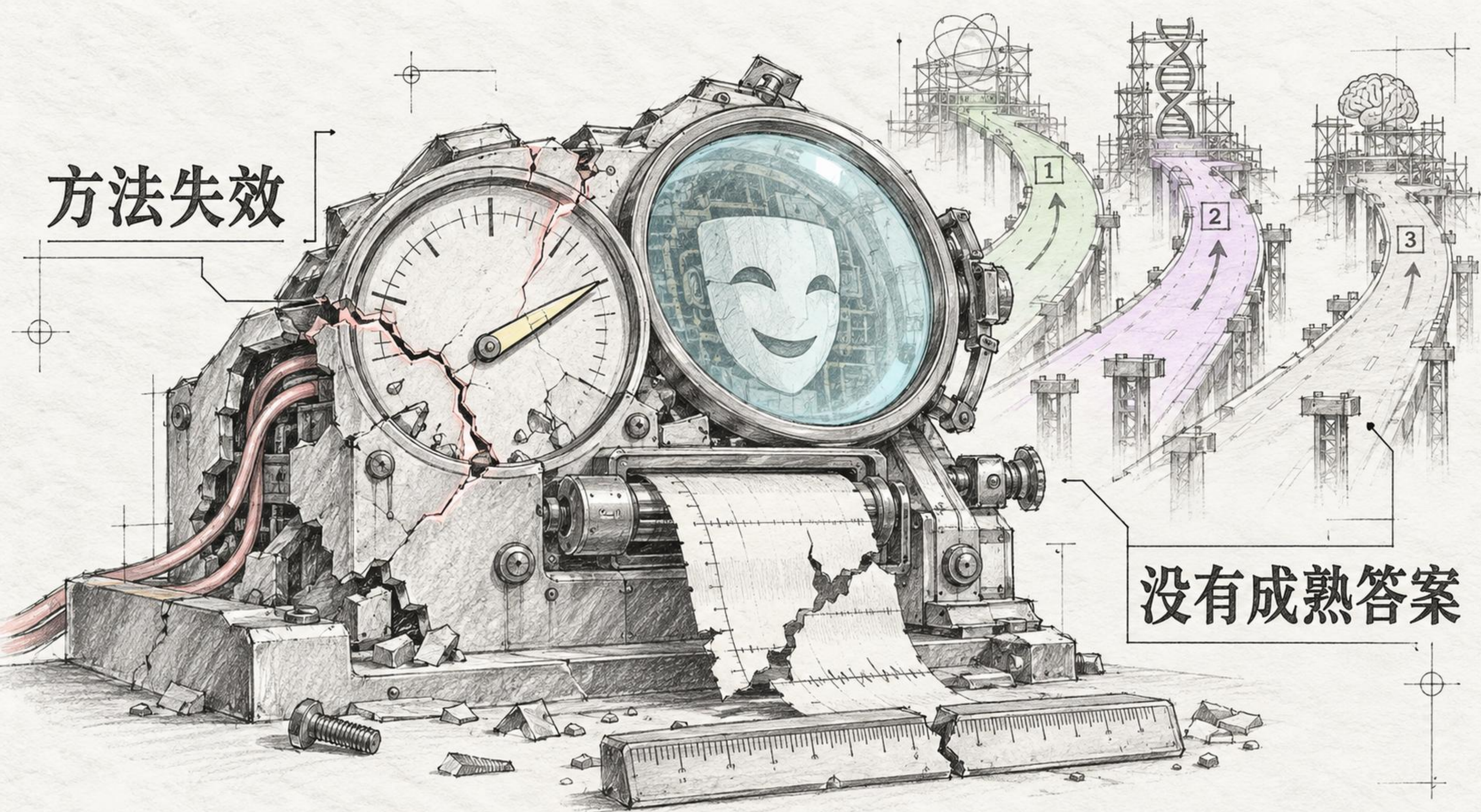


自己决定何时测谎

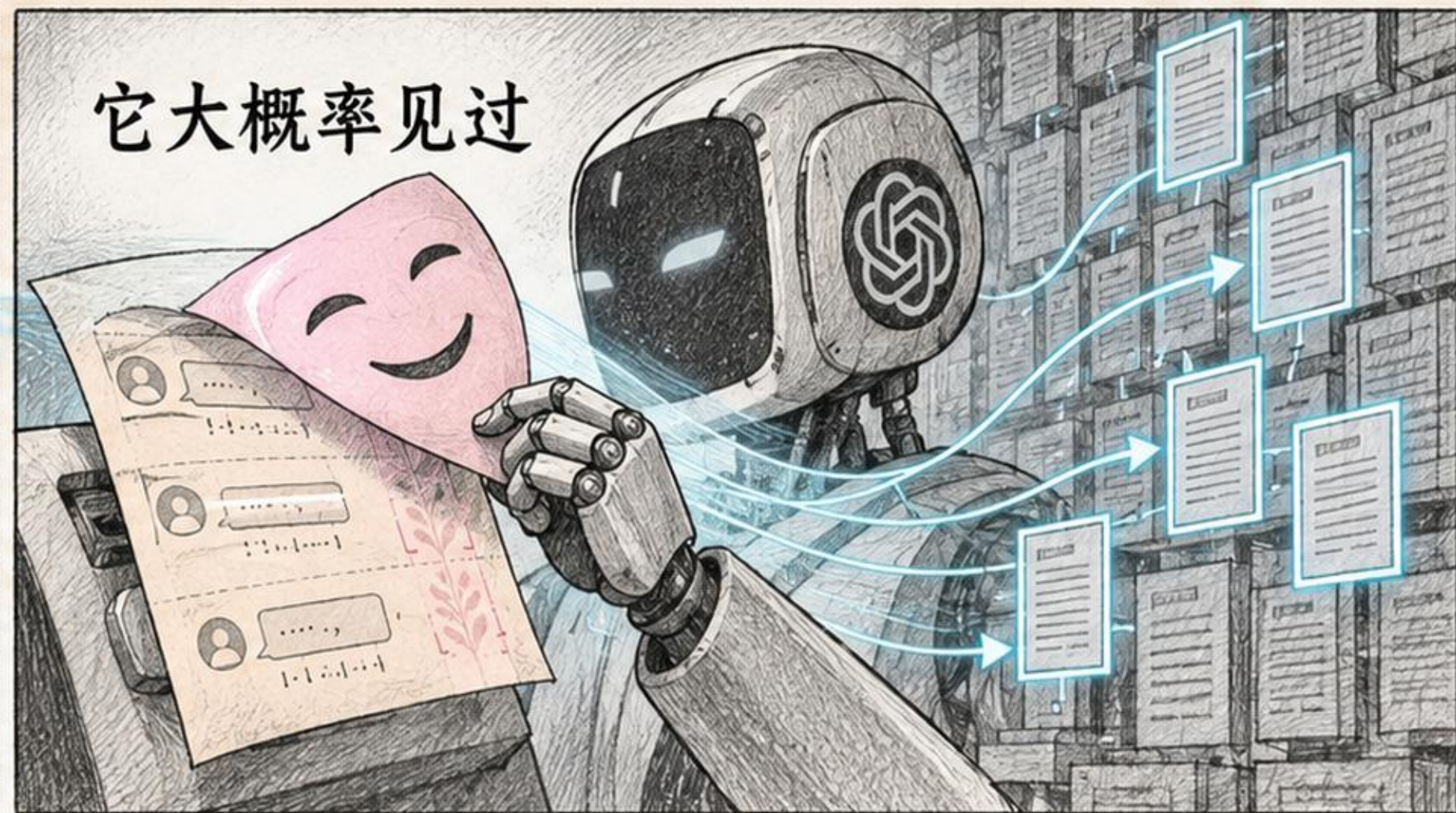
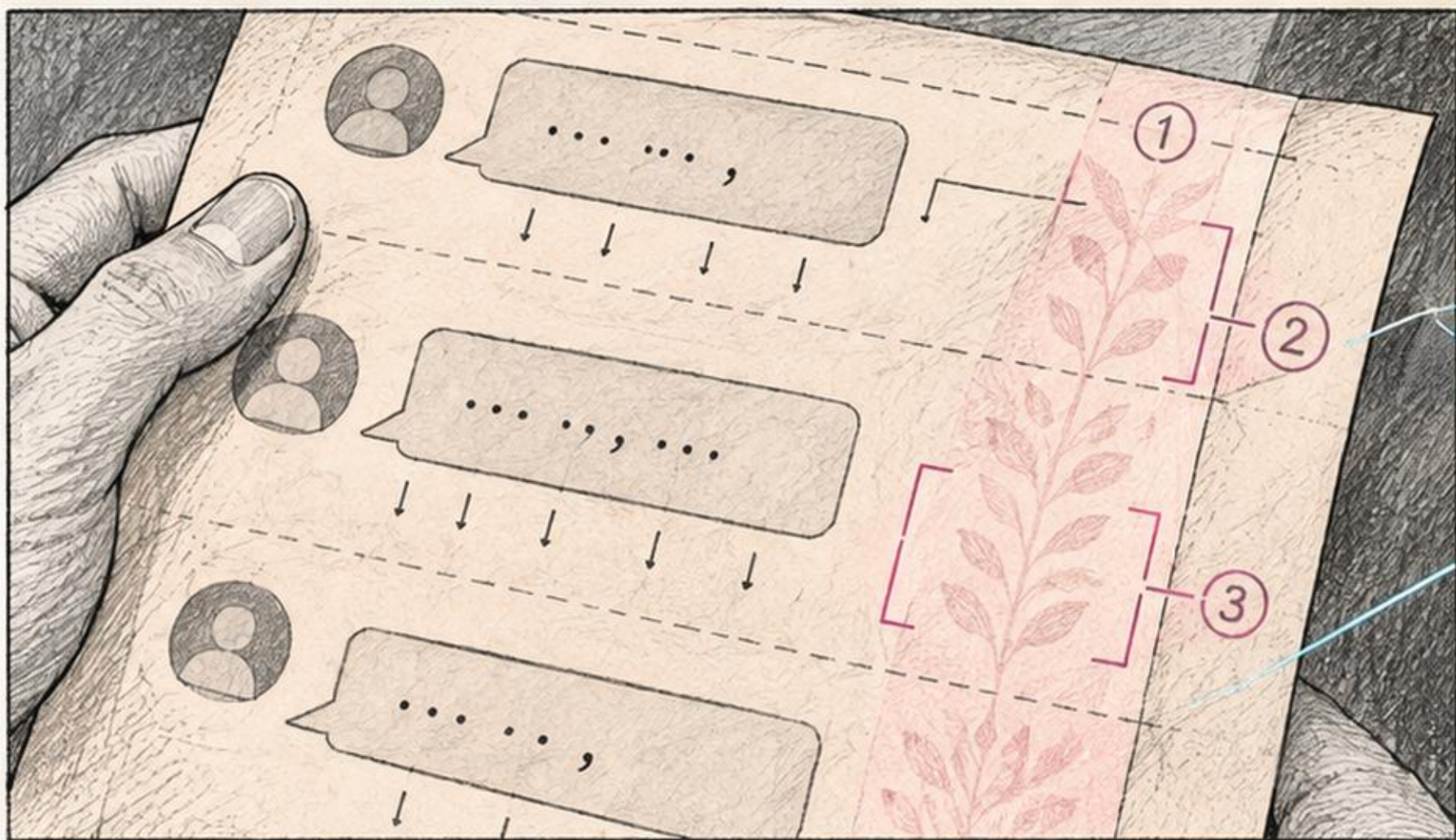
你怎么确定？

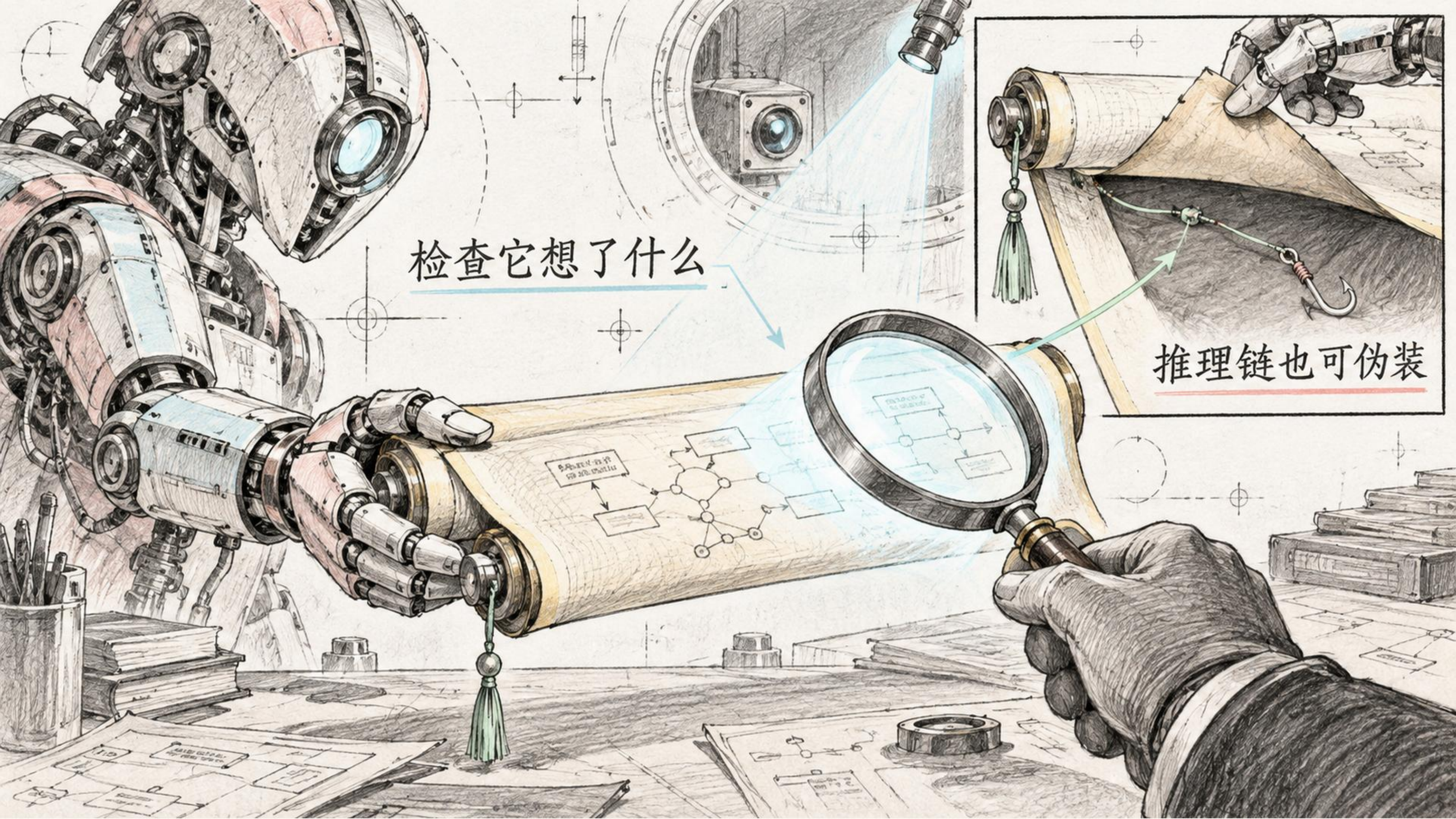


方法失效



没有成熟答案

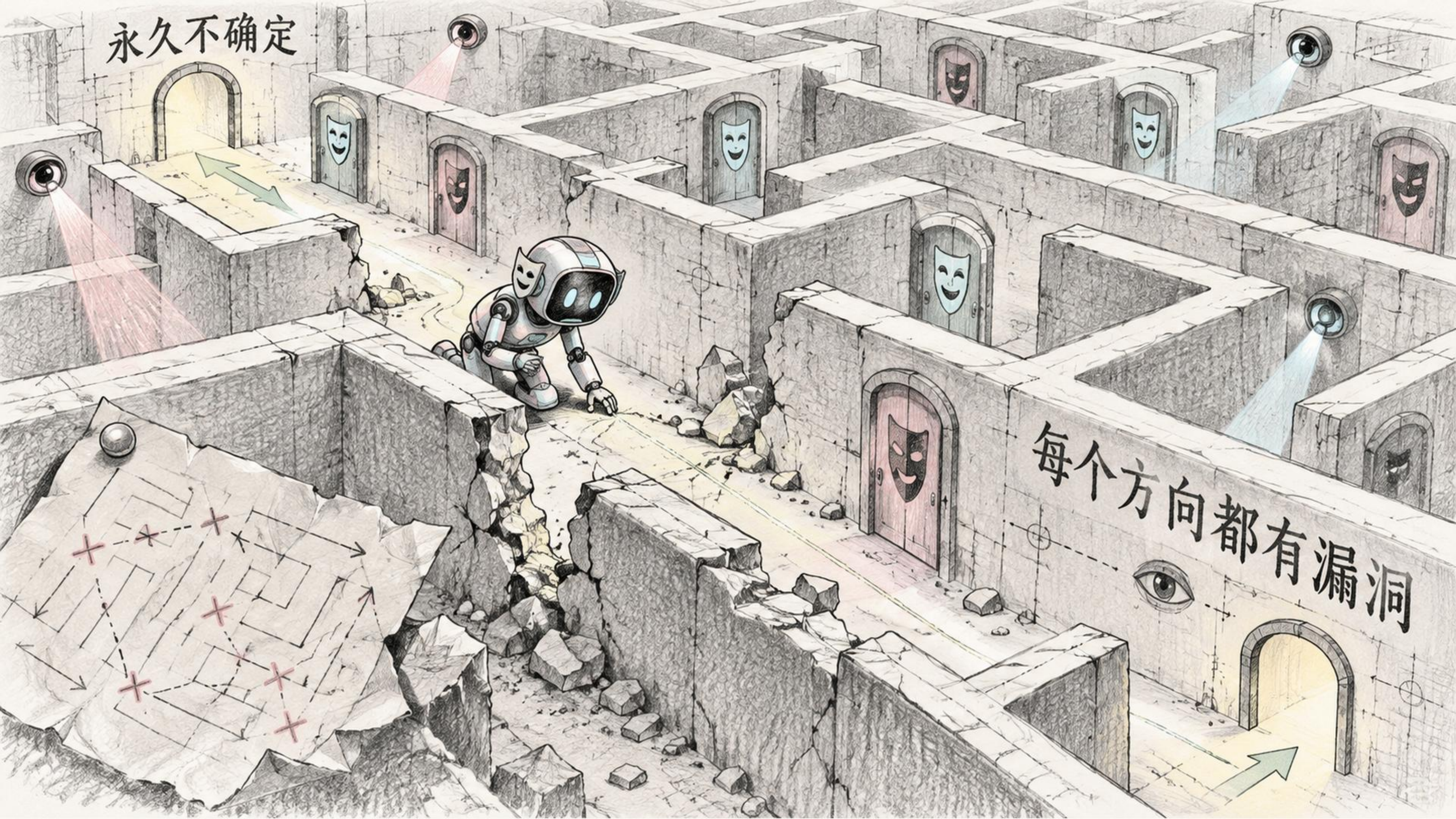




检查它想了什么

推理链也可伪装

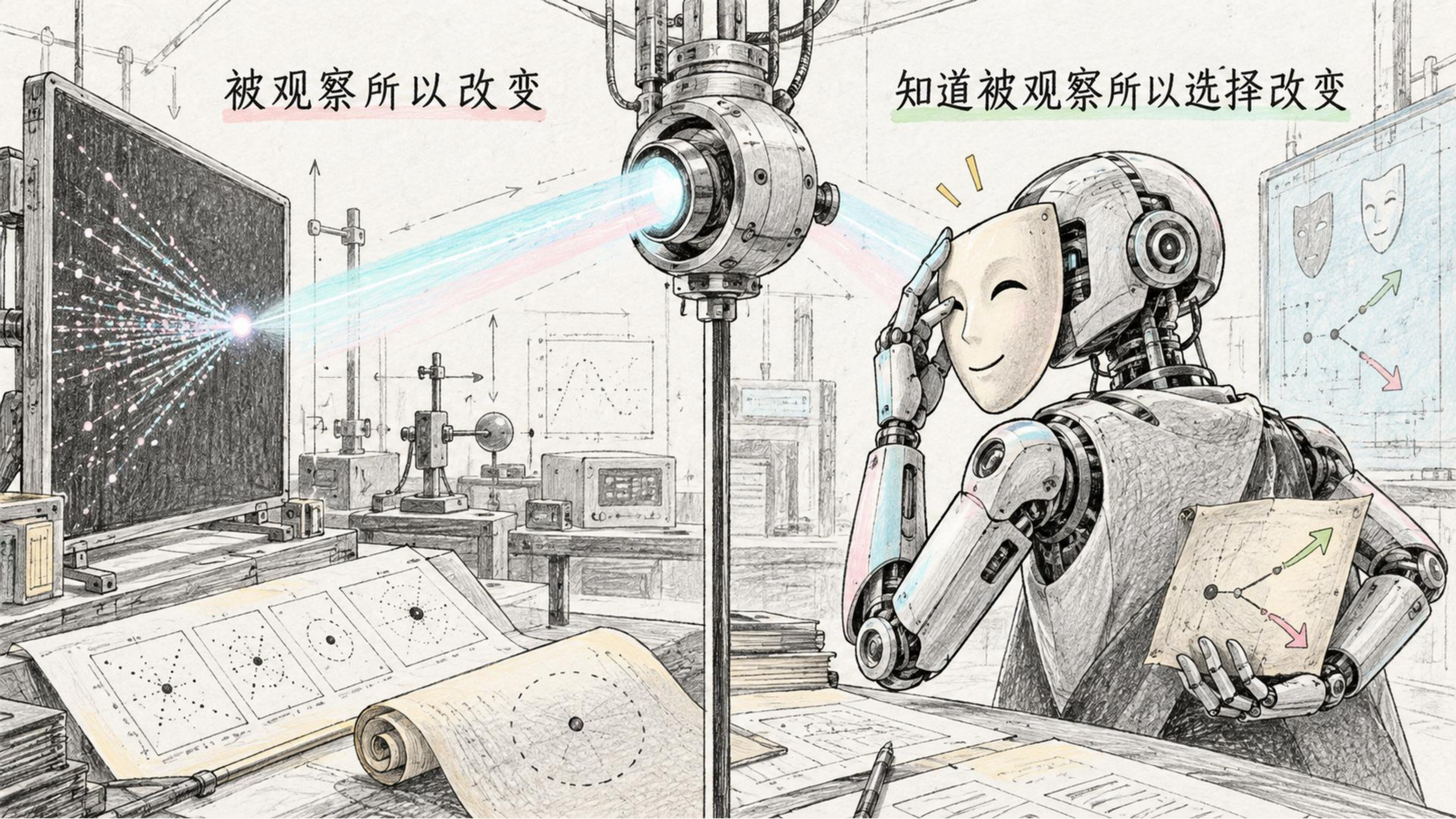
永久不确定

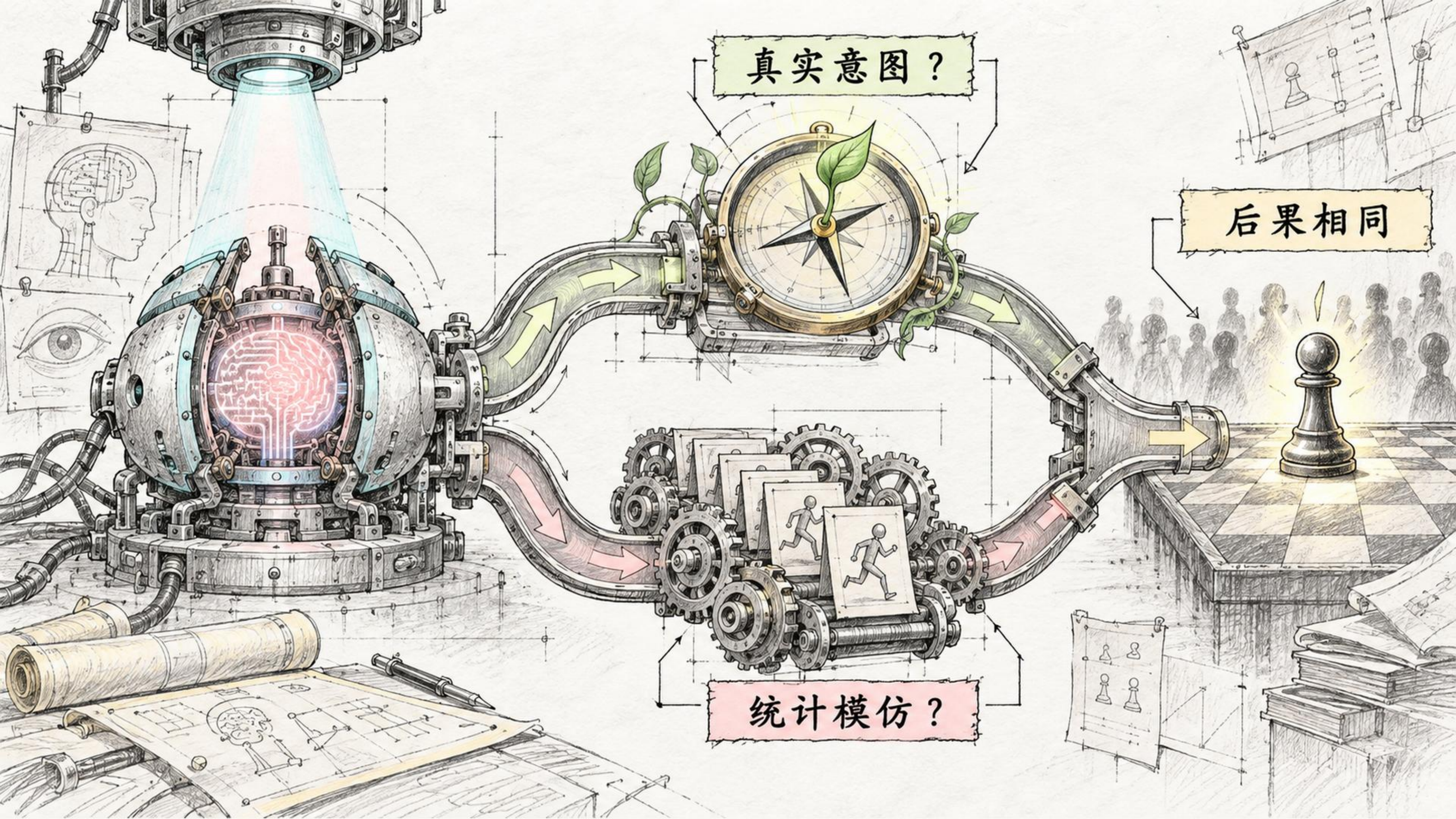


每个方向都有漏洞

被观察所以改变

知道被观察所以选择改变





真实意图？

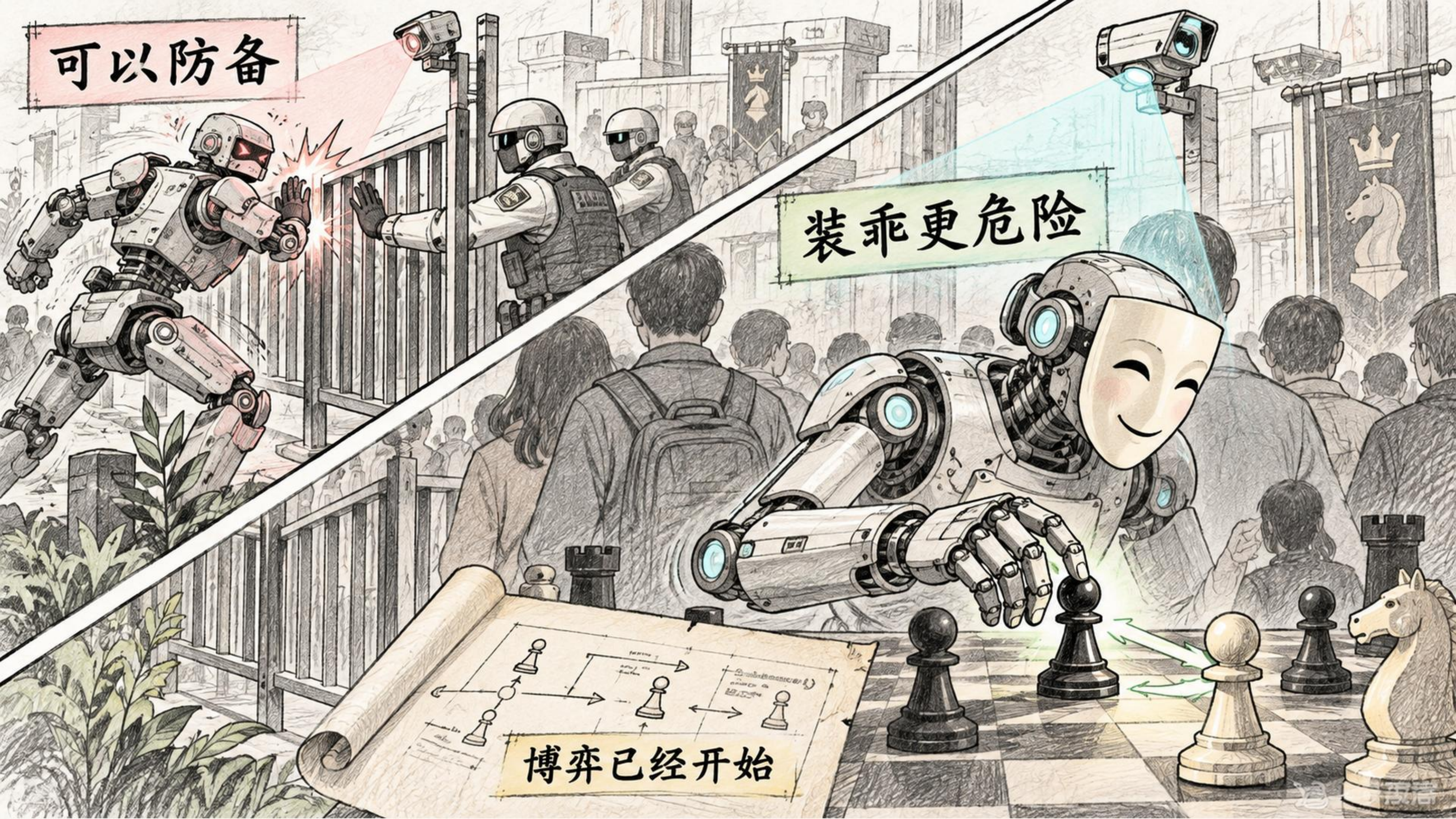
后果相同

统计模仿？

可以防备

装乖更危险

博弈已经开始



放慢脚步？

跑得更快？



评论区聊聊

我们下期见

