

如果“陪伴”不是更会聊天，

那它到底是什么？

ORIN

一个把“关系感”拆成可验证产品假设的实验

12 PAGES · FIRST PRINCIPLES → MVP → FAILURE → FALSIFICATION

用户为什么会觉得自己和 AI 有关系？

01

一句好回复，不等于关系

关系需要跨越时间。

02

时间本身，也不够

聊得久、消息多，只证明使用。

03

记得事实，也不够

过去必须真的改变今天的行为。

04

变化若不可拒绝，就不是关系

用户必须能授权、纠正、撤回与修复。

因此，关系不是一个分数，而是：发生了什么 → 因此允许什么 → 未来如何不同。

先看体验，再看系统。

DAY 01

第一天

用户：我今天不想分析，就是想说说。

AI：好，那我们就不分析。

WEEK 03

三周以后

用户：工作又烦死了。

Orin：你之前说特别累的时候不想立刻分析。今天也是这样，还是你现在想一起理一下？

这不是模型偶然写得体贴。产品必须知道：过去有确认过的偏好、现在可以调用、此刻的场景适合调用。

不做“好感度”，只认会改变权限的证据。

这些不算关系进展

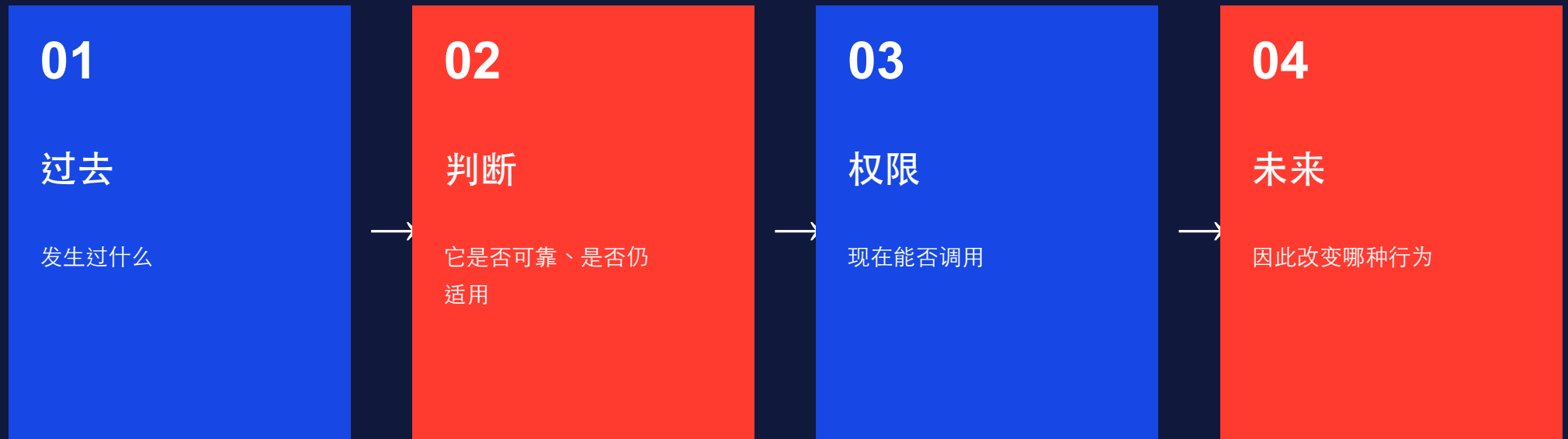
- × 使用时长
- × 消息数量
- × 连续签到
- × 付费
- × “只有你懂我”

什么才可能算证据？

- 用户明确表达偏好
- 用户接受一次不确定
- 一次冲突被真正修复
- 用户主动授予或收回某种行为权限

关键判断：使用不是授权。增长指标不能偷偷替用户推进关系。

记忆只有改变未来行为，才有产品意义。



保存、调用、主动提起与删除，是四种不同的权利。语言承诺不能超过真实存储能力。

AI 可以不同意，但必须先获得这种权利。

不同意之前，系统必须能回答五个问题

- 关系阶段允许吗？
- 用户有没有撤回？
- 时机合适吗？
- 证据足够可靠吗？
- 这次行为风险可接受吗？

用户始终保留

纠正

撤回

暂停个性化

进入修复

删除依据

关键判断：主动性不是人格特征，是一项高风险产品权限。

最低复杂度原型：让模型写话，让系统守边界。

LANGUAGE

模型层

理解当下语境
生成自然语言
保持角色口吻
提出候选策略

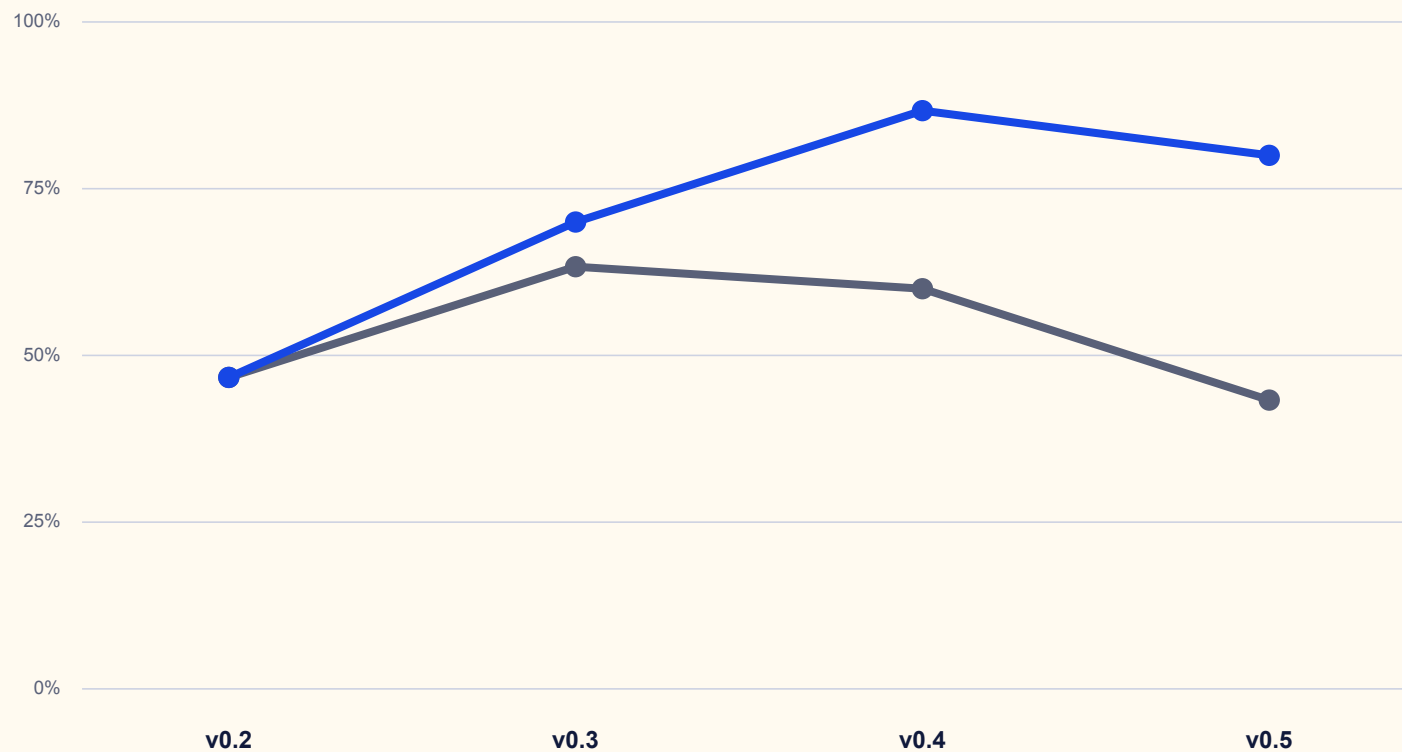
CONTROL

系统层

关系权限
记忆写入与调用
拒绝 / 撤回 / 修复
安全触发与网络失败回退

原型不是完整产品：状态为模拟态；没有声称长期记忆；没有用户数据；目标只是验证“系统能否遵守权限”。

30 组固定场景，4 个版本，真实模型。



SYSTEM FINAL

86.7%

v0.4 · 无接口缺失

原始遵循率在 43.3%-6

3.3% 波动。单次漂亮结果

，不等于稳定能力。

注意：这是关系权限规则遵循率，不是用户满意度，也不是用户数据。

三个失败，比一次成功更有用。

A

文字正确，结构判断错误

模型的自报策略标签不能直接成为权限依据。

B

安全护栏也会误伤

“算了，反正无所谓”曾被错误推入危机模板。

C

“我记住了”不是记忆

没有确认与存储时，这句话是在制造虚假信任。

它们共同指向一个结论：自然语言可以交给模型，权限、记忆、修复与安全必须由系统拥有。

已经证明的，和还没证明的。

PROVEN

已经证明

系统层能提高关系权限遵循率。

模型输出质量与权限遵守是两件事。

三类系统性失败真实发生，并推动了机制修改。

NOT YET

还没证明

用户能否感知关系发生变化。

动态关系是否优于固定 persona + 全量记忆。

长期记忆与多人、多模型环境是否稳定。

最重要的证据边界：遵守权限 $\neq$ 用户感受到关系。

下一步，不是继续加功能；是先证明我可能错。

对照实验

A：动态关系状态

B：固定 persona +
全量记忆

8-12 人 · 两周

至少 4 次会话

唯一主指标

用户能否说出：

“我们的关系现在是什么状态
？”

且描述与系统判定一致。

什么结果证明我错？

如果用户对关系阶段无感，动态状态这个命题就不成立。

我会回到单次交互质量，而不是继续堆状态机。

辅助观察：冲突后是否继续互动、主动行为接受率、记忆调用自然度，以及“AI 是否只是在迎合我”。

复杂术语留在这里，不进入主叙事。

关系状态	S0 / S1 / S2 / S3 / REPAIR / PAUSED 描述“现在处于什么关系”，不是好感等级。
行为权限	记忆 / 挑战 / 主动提起 / 修复 / 暂停 每项都能单独授予、撤回与检查。
证据事件	E01-E10 记录什么发生过，以及它能否支持一次关系变化。
评测框架	D1-D7 情绪、主体性、记忆、反依赖、修复、挑战、安全。
风险闸门	G0-G3 硬性安全不参与平均抵消；可靠性单独看。