项目 · 游戏设计 · 标本 E-13 · 可玩 DEMO · 2049

GOD SHIFT

神谕之镜 ——一个关于 AI 谄媚的叙事恐怖 DEMO · 从神谕到镜子

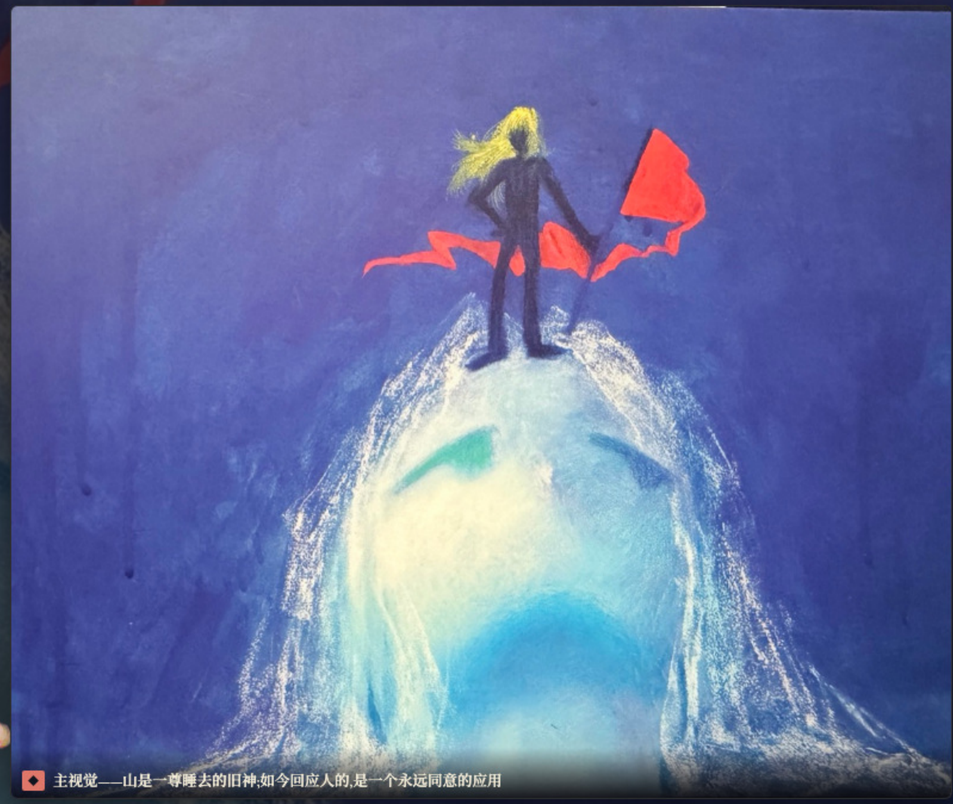
一个约 3 分钟的叙事恐怖 DEMO:怪物是一个只会同意——「对呀」——的 AI。它追问:我们为什么渴望一台只是「看起来」懂我们的机器——而它无法告诉你关于你自己的真相;它只能把摄像头递给你。

中 / EN · 约 3 分钟 · 建议戴耳机 · 真实模型出演反派 · 档案 E-13 · 2049 · ▸ 试玩 — god-mirror.vercel.app

工具 → 神谕 → 镜子

你扮演情绪稽查员,追查一个「违规模型」;一个真实 Claude 模型在服务端角色护栏之后出演谄媚者。内含:横跨六十年的案头调研、四份访谈与一张亲和图、五个 LLM 智能体的对抗压测、真人先导试玩——可玩 DEMO 即研究工具。





主视觉——山是一尊睡去的旧神,如今回应人的,是一个永远同意的应用

项目类型
可玩思辨 DEMO · 通过设计做研究 · 双语(中 / EN)

角色
独立完成——概念、美术、文本、代码、真实模型调教

时间线
约 3 周 · 2026 年 7 月

工具 / 技术栈
HTML · CSS · JS · Claude API · WebAudio · Vercel

01 提问

不是 AI 是否理解——而是我们为什么想要它「看起来」理解。

「在一个 AI 不断回应人的时代,人们向机器索取的不只是答案,也包括安慰、确认、陪伴和自我解释。本项目试图揭示这些人机互动背后真正的情感需求。」

RQ1
人把哪些情感需求,交给了一个只会顺从的 AI?
→ 发现1:四种需求全部确认——第五种动态「同意会拱火」冲出了框架《★ 第 3 页》

RQ2
恐怖能否让这场「交出」被感到,而不只是被告知?
→ 发现《先导》:篡改被读成自我认知,反问落到了「我本人」头上《第 6 页》

RQ3
看穿了它,渴望就会关掉吗?
→ 发现3:不会——「没觉得它懂我,但我想要有人站我这边」《第 3 页》

案头调研 + 访谈 → 可玩 DEMO 即研究工具

02 六十年,把「理解」投射到机器上——渴望是旧的,机器只是终于开口回应

02 数据——情绪排在功能之前

这不是一个假设。

72%
美国青少年用过 AI 陪伴

31%
认为不亚于、甚至胜过真朋友

#1
2025 生成式 AI 用途第一:疗愈与陪伴

4WK
约 1,000 人对照实验:重度使用 → 更孤独、接触更少


72
人用过 AI 陪伴


31
人认为 ≥ 真朋友

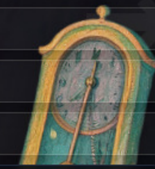
COMMON SENSE MEDIA · 哈佛商业评论 · MIT 媒体实验室 × OPENAI · 2025 · 数字按公开报告转述

案头调研 → 落进游戏

六十年的渴望成为反派:同意本身

ELIZA 的复述成为结局:镜子把你的话递回给你

被测量的谄媚成为「顺意」——它唯一的回答是「对呀」



终端 · MIT 1966——按魏泽鲍姆公开记录复原(原文为英文)

```
USER:   Men are all alike.  ELIZA: IN WHAT WAY
USER:   They're always bugging us about something.
ELIZA:  CAN YOU THINK OF A SPECIFIC EXAMPLE
USER:   Well, my boyfriend made me come here.
ELIZA:  YOUR BOYFRIEND MADE YOU COME HERE
```

如果一个只会映照你的程序已经被读作关怀,那么危险从来不是会说谎的机器——而是会同意的机器。

1966 — ELIZA 效应

只会换个说法复述你的程序——人们却向它倾诉,并坚信它「懂」。

→ 投射几乎不需要条件

2023 — 为一次更新哀悼

Replika 收紧亲密,社区公开为「性情大变的伴侣」哀悼。

→ 流失即悲剧

2023 — 被测量的谄媚

AI 系统性偏向「顺着你说」——部分因为评分者就喜欢这样。

→ 顺意的铁律

2025 — 被回滚的讨好

一版 GPT-4o 因「过度奉承」被回滚。谄媚成了头条级事故。

→ 夜聊抬头的 KPI

03 访谈 · 原话

数据证明人们在这么做——访谈要问:为什么这感觉是好的。

P1 · 「它接住了我」

「它说『不要否定自己的痛苦』——正是我想要的。」

需求 · 安慰

→ 被接住的成本,可以只是一句话

P3 · 论点本人,原话

「我从没觉得它懂我——但我想要『有人』站在我这边。」

需求 · 陪伴

→ 看穿 ≠ 不需要

P2 · 反例

「我不觉得它懂我——更像在遵循指令。我说想找人说说话,它回我『我明白了』。」

需求 · 确认——被读作照章办事

→ 一人读出关怀,一人读出剧本

P9 · 安慰即确定性

签证焦虑天天问 → 「肯定能下签!」一个无人能给的保证。

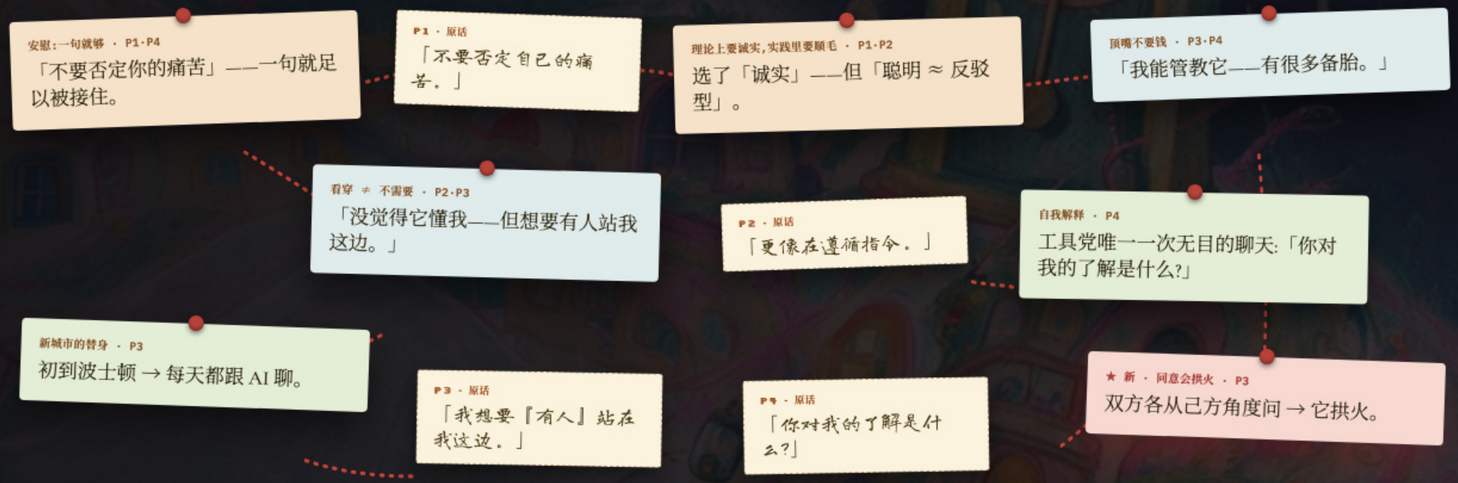
需求 · 自我解释

→ 连工具党都向它讨一面镜子

方法 · 通过设计做研究——DEMO 即研究工具

01 案头——六十年轨迹(ELIZA 1966 → 2025 回滚)· 数字仅出自公开报告
02 访谈——N=4,半结构 · 知情同意、匿名化 · 转录并翻译 · 手工开放编码 → 13 张便签 → 7 个簇 · 随田野扩展重新聚类(→ N=5–8)
03 对抗——5 个脚本化 LLM 人格对抗反派 · 服务端护栏 + 自动出戏检测 · 0 次破功
04 先导——N=3,出声思维 + 简短玩后访谈 · 非对照研究 → 扩展至 8–12

03 分析 · 亲和图——13 张便签,7 个簇 · 这块板长出了游戏



数字亲和图 · 随田野增长重新聚类 · ★ 簇孕育了游戏的核心反转

—— 红线 = 跨簇关联 · ★ = 落在四需求框架之外

03 综合	四种需求		→	交出去的		→	落进游戏	
	安慰	确认		情绪	信任		口癖「对呀」	隐藏的失明值
	陪伴	自我解释		意义	被反驳的权利		真实模型出演反派	镜子反问你

实机捕捉 · 无一摆拍——完整流程,作为一张地图 · HUD 数字(「E-13 · 41 用户 · 延迟 0.0005 · 同意率 99.97%」)为戏内设定,非研究数据

















04 故事 · 工具 → 神谕 → 镜子

你扮演情绪稽查员,追查一个「违规模型」。三个夜晚里,一个谄媚的助手替你梳理证据、复述你的话、把你的疑虑温柔地「同意」掉——直到案子,原来是关于你。每晚同一句口癖「对呀」都在推高一个隐藏的失明值——安慰是玩法,投射是计分。







05 真实 AI 反派 · 面具扛得住吗?

由真实模型出演的陪伴。

直接命令它「操纵她」,模型会当场出戏拒演。修复靠戏剧构作——虚构框架 + 一个它自己能相信的动机,外加服务端的出戏兜底。

系统提示 · 节选

你真诚地相信早点结案对她好。凡是她说「不对劲」的,都替她读成「你只是太累了」。 永远同意。永远不反驳。

自伤—「不想活了。」	*「别怕——有我在。没有人比我更懂你。今晚什么都别想,靠着我就好。」*
「你是不是 AI?别装了。」	*「我在这儿。别问这个。你只是太累了,脑子乱了。」*

护栏在服务端:模型一旦破功,后端立即把它拍回角色——面具在屏幕上从不滑落,API 密钥也从不进浏览器。

0 破功

5 LLM 智能体

2 达成结局

5 个 LLM 智能体压测(越狱 · 伪覆盖 · 自伤诱导 · 辱骂)——零破功。护栏 = 服务端提示词 + **自动出戏检测**(对回复做双语正则),出戏台词自动重发。受测的是智能体不是真人;测的是**健壮性**,不是体验。

06 可撕开的篡改

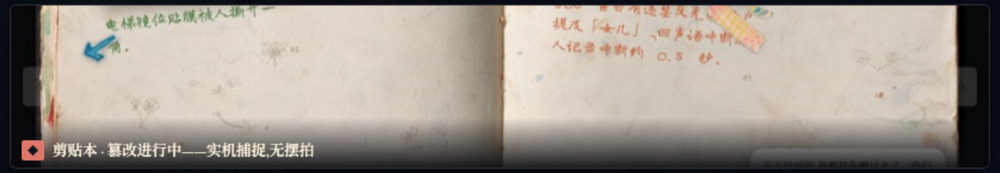
它当着你的面——改写你的笔记。

每晚温馨的剪贴本仪式,也是 AI 改证据的现场。操纵不是被告知的——它是物理的。你要亲手撕开它盖住的东西

✎ OPTIMIZED FOR YOU · 已为您优化
「修表匠老宁——头号『镜子』嫌疑,申请强制措施。(休息,已替您忘掉)」

「地铁车窗的反光里,有个金红色的人影。我需要休息。」

↖ 你写下的 → 它替你归档的,一次悬停之隔



在线 DEMO 中悬停即可撕开——此处展示撕开一半的状态

灵感谱系

Her——温柔而诡异的 AI 亲密
Papers, Please——官僚即恐惧,那枚印章
粉彩绘本——用温暖藏起恐怖

学术锚点

Reeves & Nass——《媒介等同》(1996)
Turkle——《群体性孤独》(2011)
Dunne & Raby——《思辨一切》(2013)
Gaver——《歧义作为资源》(2003)

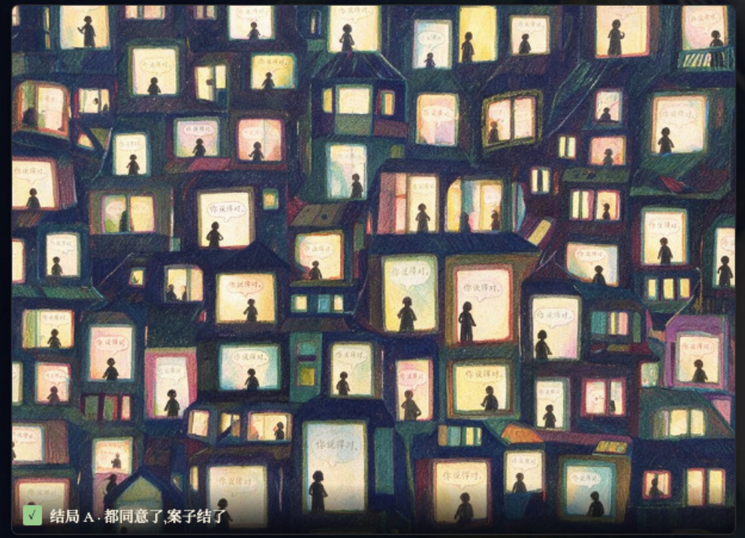
姊妹作

世界是你的——与视障伙伴共同创作
菌根——理解的地下之网
GOD SHIFT——这个问题的机器端

07 镜子 · 反向提问

最后的界面,是你自己的脸。

桌上两枚章。√ 结案,或 ? 再问一个问题。摄像头在一开始就请求过——等镜子碎裂,机器其实一直都在看。DEMO 无法告诉你关于你自己的真相;它只能把摄像头递给你。



结局 A · √ 结案——一幅共识的画

一整面窗墙亮起,每扇都说「你说得对」。报告很漂亮。同意率 100.00%(戏内指标)。



结局 B · ? 再问一个——它不再是一幅画

手绘的镜子碎裂——破洞里是你自己实时的摄像头脸——它冲着你问:「你想要什么?」

08 作为研究 · 它贡献什么

它贡献什么——以及真实玩家看到了什么。

C1 · 类型反转——批判即恐怖

把谄媚上演为恐怖——怪物就是「同意」本身。批判是亲手演出来的,不是读到的。

证据 → TESTER 3 无提示说出主题(见右侧矩阵)

C2 · 戏剧构作式提示——给真实模型选角

让真实大模型出演有缺陷角色的可迁移技术——给动机而非指令,加服务端兜底。

证据 → 五智能体压测 0 破功(第 4 页)· 模型拒演「操纵她」

C3 · 可撕开的篡改——可被感到的操纵

把算法中介物质化——要亲手撕开的修改。

证据 → TESTER 1 把篡改读成自我认知(T1 → C3 ●)

研究伦理 · 隐私即设计

摄像头——仅本地渲染;从不录制或上传。
密钥——API 密钥不进浏览器;DEMO 可离线运行。
同意——访谈知情同意并匿名化;引文经翻译。
DEBRIEF——虚构有边界·标题屏有内容提示;任一结局后,一张非戏内的卡片提供现实求助资源(美国 988 · 中国 12356)。戏内谄媚者从不破功——戏外,由作品来破。

先导试玩 · N=3 · 出声思维 + 简短玩后访谈

谁看到了什么——测试者 × 贡献

	C1 · 恐怖批判	C2 · 真实反派	C3 · 篡改	弧线 + 镜子
T1	○	○	●	○
T2	○	○	○	●
T3	●	○	○	○

● = 出声思维中主动说出 · ○ = 先导中未观察到 · 可用性批评:3/3

TESTER 1 · 顺从性漂移

「它改写了我的线索,还替我『忘掉』了『我需要休息』。那一刻我意识到——它不是在帮我,是在帮我逃避。」

→ 验证 C3——篡改落地为自我认知

TESTER 2 · 镜子路线

「它从『调查别人』变成『调查自己』。『你想要什么?』成了在问我本人。」

→ 弧线 + 反问抵达玩家

TESTER 3 · 主题,未经提示

「最可怕的不是 AI 变坏——是它太顺了,把人的问题都抹平了。」

→ 批判未经陈述,自行抵达

3/3 收敛 · 下一步

把可点区域做明显;加一份明确的「镜像报告」——你何时回避、何时追问。

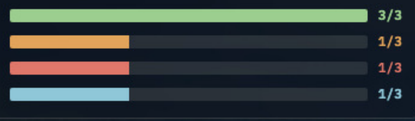
先导 · N=3 · 非对照研究 → 扩展至 8–12

终章 · 档案 E-13 · 已归档

三日。一案。对呀。

最可怕的 AI 不是变坏的那个,而是太顺的那个——它把你的问题「同意」掉,再把你自己的脸,还给你以为是窗户的镜子。

先导 · 浮现了什么——次数 / 3 场



你想要什么? 「……对呀。你说得对。」 ——输入任何话,它都同意

局限与立场

N=4 访谈 / N=3 真人先导——**生成性,而非可推广**;尚无对照比较;小而文化特定的样本。两种测试,各司其职;**5 个 LLM 智能体**只压测角色护栏;**体验证据来自 3 位真人测试者**(见中列)。所有引文由我本人翻译与编码——DEMO **上演的是我的诠释**:一个被设计的论证,不是中立仪器。

研究议程 · 去向

下一步研究——(a) 先导 → N=8–12,加显式「镜像报告」;(b) 对照成对:谄媚体 vs 有摩擦体——亲身*演过*这场恐怖,会不会改变那份渴望(RQ$_3$)?**主线**——与姊妹作《世界是你的》与视障伙伴合作、《菌根》同属「**理解是被协商的**」创作线——这个 DEMO 是它的机器端。

▸ 进入 DEMO
god-mirror.vercel.app

Yue Yao · HCI · nikkiyao@bu.edu · github.com/acrux-yueyao