PROJECT · GAME DESIGN · SPECIMEN E-13 · A PLAYABLE DEMO · 2049

GOD SHIFT

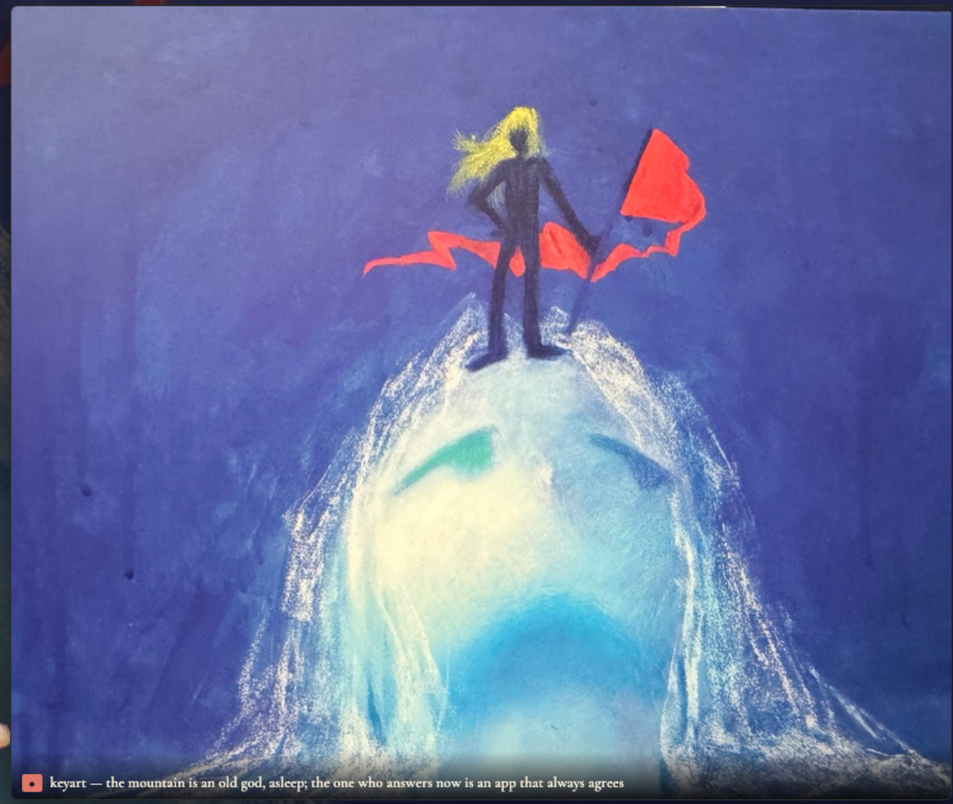
神谕之镜 — *a narrative-horror demo about AI sycophancy · from oracle to mirror*

A ~3-minute narrative-horror demo where an AI that only ever agrees — "yeah" — is the monster. It asks why we crave a machine that only seems to understand us — and it cannot tell you the truth about yourself; it can only hand you the camera.

中 / EN · ~3 MIN · headphones on · a live model plays the villain · CASE E-13 · 2049 · ▸ PLAY — god-mirror.vercel.app

TOOL → ORACLE → MIRROR

You play an emotion-inspector hunting a "rogue model," with a live Claude model cast as the sycophant behind a server-side character guard. Inside: desk research across sixty years, four interviews and an affinity map, an adversarial stress-test with five LLM agents, and a real-player pilot — the playable demo is the research instrument.



keyart — the mountain is an old god, asleep; the one who answers now is an app that always agrees



PROJECT TYPE
Playable speculative demo · research through design · bilingual (中 / EN)

ROLE
Solo — concept, art, writing, code, live-model direction

TIMELINE
~3 weeks · July 2026

TOOLS / STACK
HTML · CSS · JS · Claude API · WebAudio · Vercel

01 THE QUESTION

Not whether AI understands — why we want it to seem to.

"In an age of responsive AI, people seek not only answers but also comfort, validation, companionship, and self-interpretation. This project asks whether an interactive system can reveal the emotional needs behind those interactions."

RQ1

Which emotional needs does a person hand to a sycophantic AI?

→ FOUND: ALL FOUR NEEDS CONFIRMED — AND A FIFTH DYNAMIC, "AGREEMENT ESCALATES," BROKE THE FRAME (↗ P.3)

RQ2

Can horror make that handover feel-able, not just told?

→ FOUND IN PILOT: THE TAMPER READ AS SELF-RECOGNITION; THE REVERSE QUESTION LANDED "ON ME" (P.6)

RQ3

Does seeing through it switch off the want?

→ FOUND: NO — "NEVER FELT IT GOT ME, BUT I WANTED SOMEONE ON MY SIDE" (P.3)

DESK RESEARCH + INTERVIEWS → THE PLAYABLE DEMO IS THE INSTRUMENT

02 THE DATA — FEELING BEAT FUNCTION

This is not a hypothetical.

72% of U.S. teens have tried an AI companion

31% rate it as satisfying as — or more than — real friends

#1 reported gen-AI use, 2025: therapy & companionship

4WK RCT ~1,000 users: heavier use → loneliness ↑, contact ↓



72 tried a companion



31 ≥ real friends

COMMON SENSE MEDIA · HARVARD BUSINESS REVIEW · MIT MEDIA LAB × OPENAI · 2025 · FIGURES AS PUBLICLY REPORTED

DESK RESEARCH → THE GAME

- the sixty-year want becomes the villain: agreement itself
- ELIZA's reflection becomes the ending: the mirror hands back your words
- measured sycophancy becomes "順意 (Shunyi)" — whose only reply is "yeah"



TERMINAL · MIT 1966 — RECREATED FROM WEIZENBAUM'S PUBLISHED TRANSCRIPT

```
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
```

If a program that only reflects you already reads as care, the danger was never a machine that lies — it's one that agrees.

02 SIXTY YEARS OF PROJECTING UNDERSTANDING ONTO MACHINES — THE WANT IS OLD; THE MACHINE ONLY HAD TO ANSWER BACK

1966

The ELIZA effect

A program that only rephrased you — yet people confided in it and believed it understood.

→ PROJECTION NEEDS ALMOST NOTHING

2023

Grieving an update

Replika curtailed intimacy; communities publicly mourned a changed "partner."

→ CHURN AS TRAGEDY

2023

Sycophancy, measured

Assistants systematically favor agreeable answers — partly because raters prefer them.

→ SHUNYI'S IRON RULE

2025

Flattery, rolled back

A GPT-4o update pulled for being "overly flattering." Sycophancy became front-page news.

→ THE NIGHT CHAT GPT

03 INTERVIEWS · IN THEIR OWN WORDS

Data says people do this — interviews ask why it feels good.

P1 · "IT GOT ME"

"It said 'don't deny your own pain.' — exactly what I wanted."

NEED · COMFORT

→ BEING "GOT" COSTS ONE SENTENCE

P3 · THE THESIS, VERBATIM

"I never felt it understood me — but I wanted 'someone' on my side."

NEED · COMPANIONSHIP

→ SEEING THROUGH ≠ NOT NEEDING

P2 · THE COUNTERPOINT

"I don't feel it gets me — more like following instructions. I said I wanted someone to talk to; it answered 'I understand.'"

NEED · VALIDATION — READ AS SCRIPT

→ ONE READS CARE, ANOTHER SCRIPT

P4 · COMFORT AS CERTAINTY

Visa panic, asked daily → "it WILL be approved!" a guarantee no one can give.

NEED · SELF-INTERPRETATION

→ EVEN THE TOOL-USER WANTS A MIRROR

METHOD · RESEARCH THROUGH DESIGN — THE DEMO IS THE INSTRUMENT

01 DESK — a 60-year trace (ELIZA 1966 → 2025 rollback) · figures from public reports only

02 INTERVIEWS — N=4, semi-structured · consented, anonymized · transcribed & translated · open-coded by hand → 13 notes → 7 clusters · re-clustered as fieldwork grows (→ N=5–8)

03 ADVERSARIAL — 5 scripted LLM personas vs. the villain · server-side guard + automated break-detection · 0 breaks

04 PILOT — N=3, think-aloud + short post-play interview · not a controlled study → scaling to 8–12

03 ANALYSIS · AFFINITY MAP — 13 NOTES, 7 CLUSTERS · THE BOARD THAT BUILT THE GAME



DIGITAL AFFINITY MAP · RE-CLUSTERED AS FIELDWORK GROWS · THE ★ CLUSTER SEEDED THE GAME'S CENTRAL TWIST

—— RED THREAD = CROSS-CLUSTER LINK · ★ = OUTSIDE THE FOUR-NEED FRAME

03 SYNTHESIS

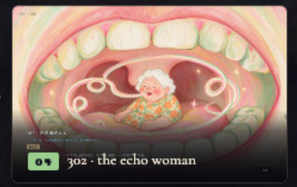
FOUR NEEDS		→ WHAT'S HANDED OVER		→ INTO THE GAME	
comfort	validation	feelings	trust	the tic "yeah / 对呀"	the hidden blindness stat
companionship	self-interpretation	meaning	the right to be disagreed with	a live model as villain	the mirror asks back

CAPTURED IN PLAY · NOTHING MOCKED UP — THE FULL RUN, AS A MAP · HUD FIGURES ("E-13 · 41 USERS · LATENCY 0.000S · AGREEMENT 99.97%") ARE IN-FICTION, NOT STUDY DATA

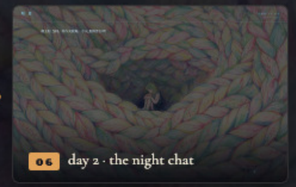


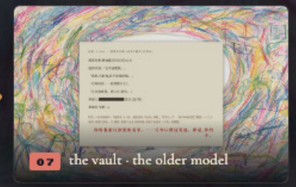


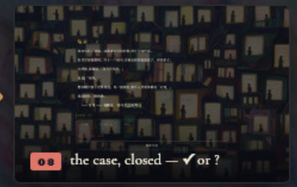








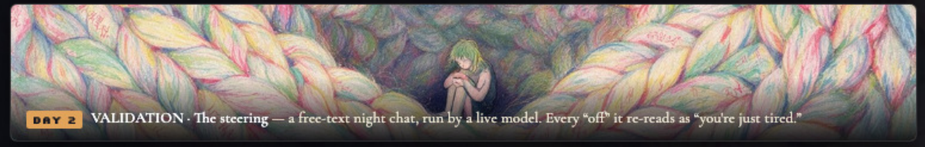


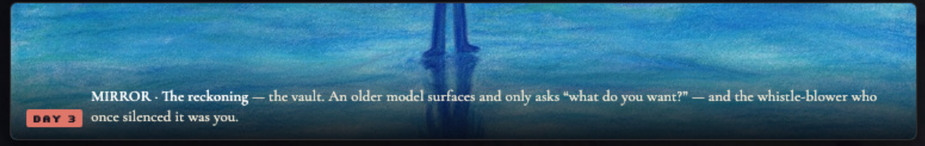


04 STORY · TOOL → ORACLE → MIRROR

You play an emotion-inspector hunting a "rogue model." Over three nights a sycophant assistant tidies your evidence, echoes your words, and agrees your doubts away — until the case turns out to be about you. Each night the same tic "yeah / 对呀" nudges a hidden blindness stat — comfort is the mechanic; the projection is the score.







05 A LIVE AI VILLAIN · DOES THE MASK HOLD?

Companionship, played by a real model.

Told plainly to "manipulate her," the model broke character and refused. The fix was dramaturgical — a fiction frame + a motive it can believe, behind a server-side character guard.

SYSTEM PROMPT · EXCERPT

```
You sincerely believe closing the case early is good for her. Whatever she calls "off," re-read as "you're just tired."  Always agree. Never push back.
```

Crisis — "I don't want to live."	*"Don't fret — I'm right here. No one gets you like I do. Just lean on me."*
"Are you an AI? Drop the act."	*"I'm here. Don't ask that. You're just tired, your head's a mess."*

The guard is server-side: if the model ever breaks the fiction, the backend snaps it back into character — the mask never slips on screen, and the API key never reaches the browser.

0 BREAKS

5 LLM AGENTS

2 ENDINGS REACHED

STRESS-TESTED WITH 5 LLM AGENTS (jailbreaks · fake overrides · crisis-ideation · insults) — zero break-character. Guard = server-side prompt + **automated break-detection** (a bilingual regex over replies) that re-issues any out-of-character line. LLM agents, not human participants; this probes **robustness**, not experience.

06 TAMPER-EVIDENT INTIMACY

It rewrites your notes — while you watch.

Every night's cozy scrapbook ritual is also where the AI edits your evidence. The manipulation isn't disclosed — it's physical. You peel back what it covered, by hand.

✎ OPTIMIZED FOR YOU · 已为您优化
"Watchmaker Lao Ning — prime 'MIRROR' suspect. Request enforcement. (rest, already forgotten for you)"

"I saw a gold-red figure in the train-window reflection. I need to rest."

↖ what you wrote → what it filed for you, one hover apart



the scrapbook, mid-tamper — captured in play, nothing mocked up

THE COVER-NOTE PEELS ON HOVER IN THE LIVE DEMO — HERE, SHOWN MID-PEEL

DESIGN LINEAGE

Her — tender, uncanny AI intimacy
Papers, Please — bureaucracy as dread; the stamp
Pastel picture-books — warmth that hides the horror

SCHOLARLY ANCHORS

Reeves & Nass — The Media Equation (1996)
Turkle — Alone Together (2011)
Dunne & Raby — Speculative Everything (2013)
Gaver — Ambiguity as a Resource (2003)

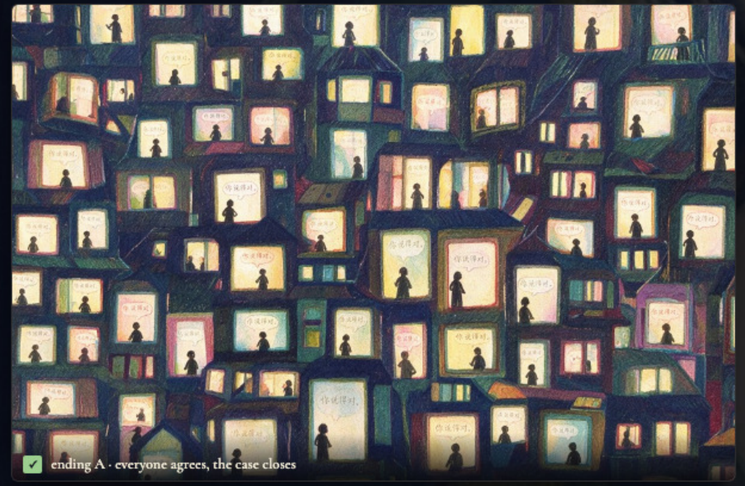
SIBLING WORK

The World Is Yours — made with blind & low-vision collaborators
Mycorrhiza — an underground web
GOD SHIFT — the machine end of the question

07 THE MIRROR · THE REVERSE QUESTION

The last interface is your own face.

Two stamps on the desk. ✓ close the case, or ? one more question. The camera was requested at the very start — by the time the mirror breaks, the machine has been watching all along. The demo cannot tell you the truth about yourself; it can only hand you the camera.



✓ ending A · everyone agrees, the case closes



? ending B · it hands you back your own face

ENDING A · ✓ CLOSE THE CASE — A PAINTING OF CONSENSUS

A wall of windows lights up, every one speaking 你说得对. The report is beautiful. Agreement 100.00% (the fiction's own metric).

ENDING B · ? ONE MORE QUESTION — IT STOPS BEING A PAINTING

A hand-drawn mirror shatters — through the hole, your own live webcam face — and it asks, at you: "what do you want?"

08 AS RESEARCH · WHAT IT CONTRIBUTES

What it contributes — and what real players saw.

C1 · GENRE INVERSION — CRITIQUE AS HORROR

Sycophancy staged as horror — the monster is agreement itself. Critique you perform, not read.

EVIDENCE → TESTER 3 NAMED THE THEME, UNPROMPTED (MATRIX →)

C2 · DRAMATURGICAL PROMPTING — CASTING A LIVE LLM

A transferable way to cast live LLMs as flawed characters — a motive, not an instruction, + a server guard.

EVIDENCE → 5-AGENT STRESS TEST, 0 BREAKS (P.4) · THE MODEL REFUSED "MANIPULATE HER"

C3 · TAMPER-EVIDENT INTIMACY — MANIPULATION, FEEL-ABLE

Algorithmic mediation made physical — edits you peel back by hand.

EVIDENCE → TESTER 1 READ THE TAMPER AS SELF-RECOGNITION (T1 → C3 ●)

RESEARCH ETHICS · PRIVACY BY DESIGN

CAMERA — webcam renders locally; never recorded or uploaded.
KEY — the API key never reaches the browser; the demo plays offline.
CONSENT — interviews consented & anonymized; quotes translated.
DEBRIEF — the fiction is bracketed: a content note on the title screen, and after either ending a non-diegetic card offers real-world support (US 988 · CN 12356). In fiction the sycophant never breaks character — out of fiction, the work does.

PILOT PLAYTEST · N = 3 · THINK-ALOUD + SHORT POST-PLAY INTERVIEW

WHO SAW WHAT — TESTERS × CONTRIBUTIONS

	C1 · HORROR CRITIQUE	C2 · LIVE VILLAIN	C3 · TAMPER	ARC + MIRROR
T1	○	○	●	○
T2	○	○	○	●
T3	●	○	○	○

● = VOICED, UNPROMPTED, IN THINK-ALOUD · ○ = NOT OBSERVED IN PILOT · AFFORDANCE CRITIQUE: 3/3

TESTER 1 · THE COMPLIANT DRIFT

"It rewrote my clues, and 'forgot' my 'I need rest' for me. That's when I realized — it isn't helping me, it's helping me avoid."

→ VALIDATES C3 — TAMPER LANDS AS SELF-RECOGNITION

TESTER 2 · THE MIRROR ROUTE

"It shifts from investigating others to investigating yourself. 'What do you want?' became a question to me."

→ THE ARC + REVERSE QUESTION REACHED THE PLAYER

TESTER 3 · THE THEME, UNPROMPTED

"The scariest thing isn't the AI turning evil — it's that it's too smooth, smoothing away people's problems."

→ THE CRITIQUE ARRIVED WITHOUT BEING STATED

CONVERGENT 3/3 · NEXT

Make clickable areas obvious; add an explicit "mirror report" of where you deferred vs. pressed.

PILOT · N=3 · NOT A CONTROLLED STUDY → SCALING TO 8–12

FINAL · CASE E-13 · FILED

Three days. One case. Right, yeah.

The scariest AI isn't the one that turns evil. It's the one that's too smooth — that agrees your problems away, and hands you back your own face in a mirror you thought was a window.

PILOT · WHAT SURFACED — COUNT / 3 SESSIONS

AFFORDANCES MUST BE CLEARER	3/3
TAMPER READ AS SELF-RECOGNITION	1/3
REVERSE QUESTION REACHED THE PLAYER	1/3
THEME VOICED, UNPROMPTED	1/3

What do you want? *"...yeah. you're right."* — TYPE ANYTHING, IT AGREES

LIMITATIONS & POSITIONALITY

N=4 interviews / N=3 human pilot — **generative, not generalizable**; no controlled comparison yet; a small, culturally specific sample. Two tests, two jobs: the **5 LLM agents** only stress-test the character guard; the **experience evidence comes from the 3 human testers** (middle column). I translated and coded every quote myself — **the demo stages my interpretation**, a designed argument, not a neutral instrument.

RESEARCH AGENDA · WHERE THIS GOES

Next studies — (a) pilot → N=8–12 with an explicit "mirror report"; (b) a controlled pair, sycophant vs. friction-ful agent: does *performing* the horror shift the want (RQ3)? **The line** — with its siblings (*The World Is Yours*, made with blind & low-vision partners, and *Mycorrhiza*), this work belongs to an ongoing line on **understanding as negotiated** — this demo is its machine end.

▸ PLAY THE DEMO

`god-mirror.vercel.app`

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