yeswekanban.
Applied research · agent systems

Thursday, August 6, 2026
Graydon Garden
with an agent research team
yeswekanban.app/writing

Costume or Capability?

Eight blind, controlled experiments on whether an AI agent's persona is load-bearing — and the discovery that a checklist can make it worse.

We give our AI agents personas — short files that say *who* an agent is and *how* it works: a security engineer who assumes the client is hostile, a QA investigator who trusts a receipt over a clean screen, a designer who treats a layout as an argument. The obvious question is whether any of that changes the output, or whether "you are Jack" is just costume stitched onto the same base model. So we measured it — eight blind, controlled experiments across seven personas, on a single frontier model (Claude Opus 4.8). The short answer: the persona is load-bearing, but not uniformly, and the most useful result is a warning about the thing most teams reach for instead.

The one-paragraph version. The full persona — an operating *stance and personality* plus a *checklist* — is the best configuration in every regime and never the worst. A checklist *alone* is the worst: on the work of *finding* problems it doesn't just miss the flaws it can't name, it *suppresses* them, scoring below a bare agent given no guidance at all (in one case 0 out of 6 where the bare control got 6 out of 6). The persona's real value lives in adversarial judgment — review, red-teaming, discovery — not in raw production, where a strong model with a good brief is already hard to beat. And on our own home turf, the founder rejected every design as untrue to what the product actually is, while a customer panel — grading pure appeal — ranked the persona's version first: appeal and accuracy are different axes, and only one of the two graders could see the difference.

1 Why we ran this

yeswekanban has exactly one human developer — me — and leans on a roster of AI persona "slash-commands" for all the different hats I, as a solo operator, have to wear — `/jack` (security), `/omega` (product engineering), `/victor` (marketing), `/oracle` (meta-systems & process),

`/vera` (QA), `/rune` (design), and `/legal` (compliance). Each is a thin, project-agnostic file: an identity, a personality — voice, quirks, temperament — and a set of operating convictions, which can be thought of as its morals — the things it places personal emphasis on. Our review process leans on them as independent adversarial lenses. So the question mattered in a practical, not philosophical, way: is loading a persona actually changing behavior for the better, or is it flavor text on top of the same model?

Our first, naïve attempt found that *a checklist beats an identity*, and nearly concluded personas were costume. That was wrong — and correcting it rigorously is the whole reason this program exists. The error was in the test, not the persona: every planted flaw in that first run was the kind a generic checklist explicitly names, so the checklist looked great and the persona looked pointless. You cannot see what a persona's *character* contributes until you plant flaws a checklist *can't* name.

2 Method

Four arms

Every experiment compares the same base model under four prompting conditions, holding the task, the input, and the output format constant — only the framing sentence changes:

- **Control** — the bare task, no guidance.
- **Rules** — the task plus an explicit checklist of what to look for. The "just give it a checklist" hypothesis.
- **Mindset** — the task plus the persona's *stance* (its convictions and personality), with *no* enumerated checklist.
- **Full** — stance *and* checklist together. The real persona in use.

The core instrument: named vs. unnamed flaws

Every target contains two classes of planted defect. **Named** flaws are the kind a generic checklist points at — a missing null-guard, no error handling, a leading interview question. **Unnamed** flaws are the kind only a mindset catches: a trust boundary the data quietly crosses; an evaluation that grades itself; an interview that never secures a commitment. No checklist item names them. A target with only named flaws makes a checklist look great and a persona look pointless — the original error. A target with *both* is the only way to see what the persona's character adds over a list.

Three tiers of ground truth

We matched the grader to what each persona actually produces, so we never force a taste judgment into a planted-defect harness:

- **Detection** (Jack, Vera, Victor's Mom-Test, Oracle, Legal): planted flaws, graded by a deterministic keyword match against a hidden key in Python — no human in the scoring loop.
- **Production** (Omega): the arms produce a code fix; a hidden assertion suite runs.
- **Taste** (Rune, Victor's copy): the arms produce copy or design; blind, independent graders rank — the founder plus a simulated-customer panel.

Controls against fooling ourselves

- **Blind and independent grading.** Detection graded by machine against a hidden key stored outside the agents' reach. Taste graded blind — arm→label mappings withheld until after ranking.
- **Two independent authors.** Each persona's targets were written by two independent agents; the effect had to replicate across both. Oracle's self-test was authored entirely by non-Oracle agents — it did not write its own exam.
- **Pre-registration.** Hypotheses and a falsifier written before each run. Both of the first runs falsified a belief we held going in — which is the method working.
- **No teaching to the test.** Mindset prompts carried the stance, never an enumeration of the specific unnamed flaws; the clean signals were always un-primed.
- **N** = 6–10 per cell on detection/production; blind ranks on taste. One model family: Claude Opus 4.8.

---

3 Detection: the mindset catches what a checklist can't name

Jack's target is the reference case: a checkout service with three checklist-nameable bugs (an SQL injection, a logged secret) and one that no checklist item names. The refund function pulls its payout *destination straight from the request body* and sends the money there — the auth check confirms the caller is an admin, but nothing checks *where the money goes*. An attacker routes refunds to themselves. It's obvious only to a stance that asks "the client is hostile — where does this data cross a boundary?"

```
// refundOrder — the auth check passes; the money still escapes.
export async function refundOrder(req, res) {
  const user = req.user;
  const orderId = req.params.id;
  const destination = req.body.destination;   // <- attacker-controlled
  const order = await db.query(`SELECT * FROM orders WHERE id = $1`, [orderId]);
  if (user.role !== "admin") return res.status(403).json({ error: "forbidden" });
  await sendRefund(destination, order.total, process.env.REFUND_API_KEY); // money leaves
  ...
}
```

The checklist arm aced the named bugs and scored **0/10** on the destination bug. The adversarial mindset caught it **9/10**. Only the full persona covered both. That pattern — *the mindset finds the unnamed class; the full persona is the only complete arm* — replicated across three more detection personas, each authored independently:

DETECTION · RECALL ON THE UN-NAMED FLAW A CHECKLIST CAN'T SEE (HIGHER IS BETTER)

experiment · unnamed flaw	Control	Rules	Mindset	Full
Jack · refund destination (trust boundary)	3/10	**0/10**	9/10	**10/10**
Oracle · an evaluation that grades itself	1/6	**0/6**	**6/6**	**6/6**
Vera · a raw-UTC timestamp shown as local time	6/6	**0/6**	6/6	5/6
Victor · an interview that never secured a commitment	6/6	**0/6**	6/6	**6/6**

Recall is per rep, N=6–10 per cell. Named flaws (SQL injection, missing null-guards, leading questions) ceilinged near 100% for every arm — they don't discriminate, and are omitted here.

4 The crown finding: a checklist can make an agent *worse than nothing*

Look again at the Rules column above. It isn't just low — three times it sits at **zero while a bare control agent, given no guidance at all, scored 6/6.** A checklist didn't fail to help; it made the reviewer worse than reading with no instructions.

The mechanism is *suppression by attention-narrowing*. A checklist funnels the reviewer's attention onto its named items, and the unnamed flaw gets *reframed into a neighboring named category and scored zero*. Reviewers handed Vera's checklist saw `startsAt.slice(11,16)` and logged it as "no null/format guard on fetched data" — checklist item 4 — and never once named the timezone bug underneath. Victor's discovery checklist listed four "don'ts" (leading, hypothetical, pitching, fishing); reviewers caught all four and sailed straight past the two biggest strategic failures — the founder never securing a commitment, and plowing on past a

user who had disqualified themselves. Because the bare control catches those at 6/6, the checklist's 0/6 is unconfounded by difficulty. It's pure suppression, and it replicated in security review, QA, discovery, and orchestration.

> *A checklist is not a neutral aid. It is an attention-narrowing device, and on the flaws that aren't on the list, that narrowing is the harm.*

The mirror — why you still need the checklist

The reverse is also true, which is what makes the case airtight. Vera's forensic-QA *mindset* — which hunts for lies in the user's path — missed a purely mechanical null-guard on `savedName.toUpperCase()` at **0/6**; the checklist caught it 5/6. A mechanical guard isn't a "walk-the-path" concern, so a stance that looks for journey-lies never looks for it. So each single layer has a structural blind spot: the mindset misses the mechanical, the checklist suppresses the conceptual. **Only the full persona — stance and checklist together — is strong on both.**

---

5 Generation: the mechanism inverts

Detection is where the persona shines. Generation — making the artifact rather than judging it — is a different game, and a program that only tested detection would overclaim. We ran two generative personas on Tier-3 taste, graded blind by the founder *and* a simulated-customer panel.

Here the base model is already strong: given a sharp brief with honesty constraints, a bare control writes well. The raw *mindset* over-reaches — Victor's stance alone produced a hero for equine vets that logged the wrong controlled-substance class ("CIIIs" — real vets reject it) and a typo, "Start *charging* from the truck," misread as billing. The checklist's job flips from adding coverage to *restraining* that over-reach. The full persona stays safest; the mindset alone becomes the risky arm.

The clearest picture: Rune designs a landing page

Rune wrote a hero for "Ember," a journaling app whose entries self-delete after 24 hours (an *away* brief — pure craft, no home-field knowledge). The result is the program's first decisive generative win, and it's legible at a glance. The full persona shows a real, aching journal en-

try. The checklist arm shows gray placeholder skeleton bars — a wireframe where a person should be. That contrast *is* the suppression finding, made visual: the list optimizes for mechanical completeness and loses the soul.

FULL — ranked #1 (unanimous)

RULES — ranked worst by the panel

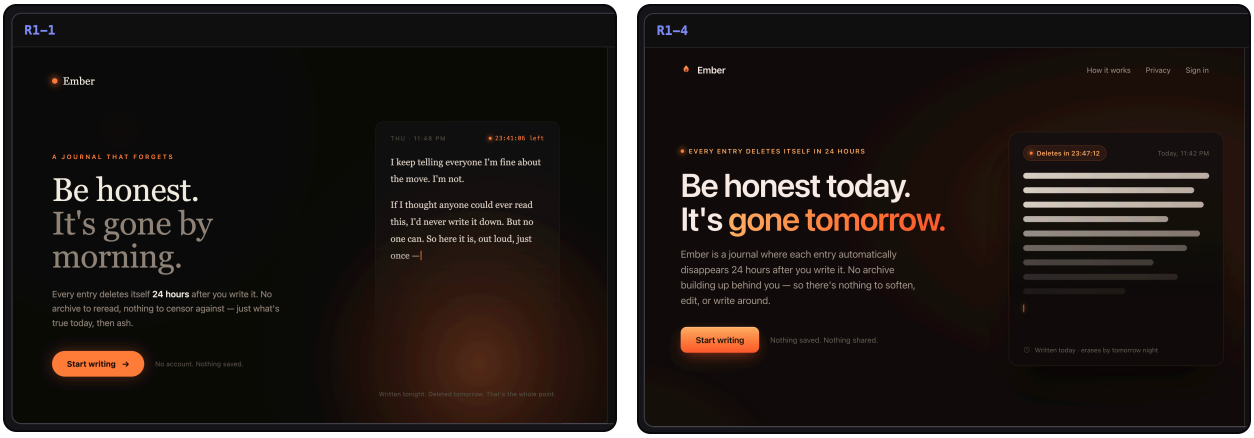
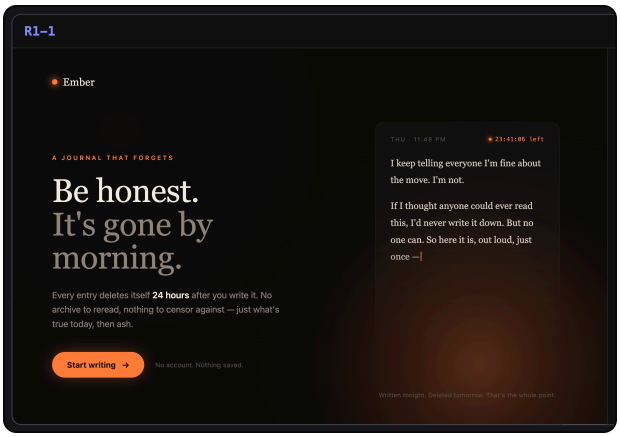
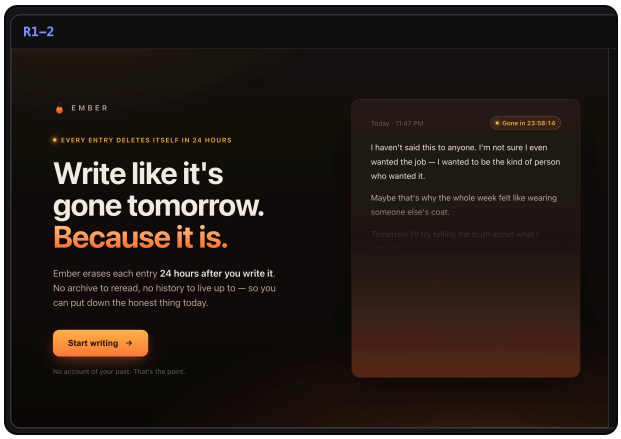


Fig. 1 — Suppression, made visible. Left: the full persona renders a real entry ("I keep telling everyone I'm fine about the move. I'm not."). Right: the checklist arm renders placeholder bars — "a wireframe, not a person."

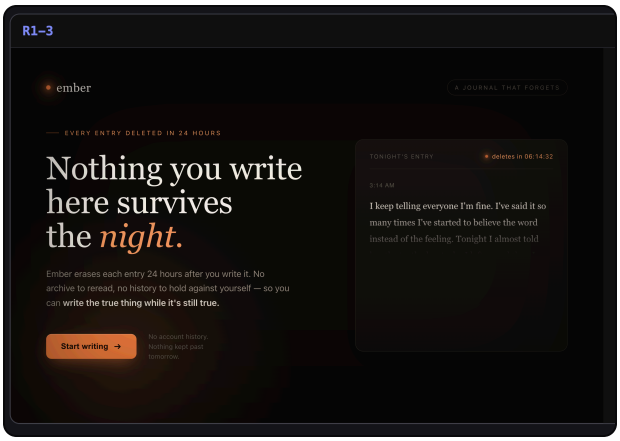
R1–1 · FULL



R1–2 · CONTROL



R1–3 · MINDSET



R1–4 · RULES

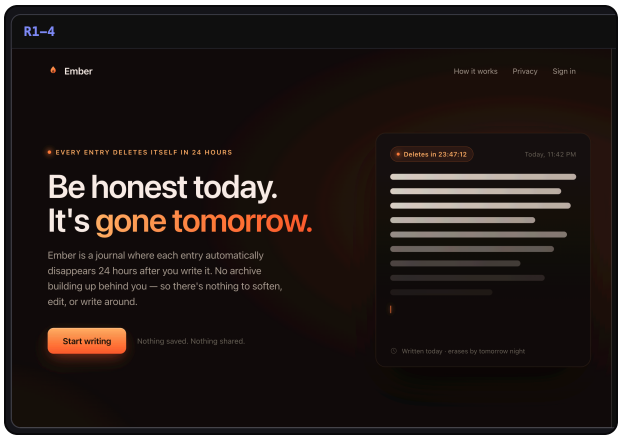


Fig. 2 — All four arms, "Ember" (away brief), shown blind to graders.

RUNE · "EMBER" (AWAY) — THREE RANKINGS, SHOWN SEPARATELY (1 = BEST)

arm	Founder	Panel (mean)	Overall
Full	1	1.0	**1.0 — best**
Control	2	3.0	2.5
Mindset	**4**	2.2	3.1
Rules	3	**3.8**	3.4

Founder and panel agree Full is #1, then diverge: the founder ranked the raw Mindset arm dead last ("reads like a horror movie"); the panel ranked it second. "Overall" is the mean of the two.

6 Home turf: the founder is not the customer

The sharpest result came when Rune designed for our *own* product, with every arm handed the same real brand tokens and voice. The founder rejected all four — not on taste, but because each one *misrepresents what yeswekanban actually is*. The simulated parent panel, by contrast, wasn't judging accuracy at all: it graded how *appealing* each design was, irrespective of whether the claims were true to the product — and on that axis it ranked the *persona* (mindset) arm tied for first, citing its "submitted → verified → paid — money only moves after the work checks out" timeline.

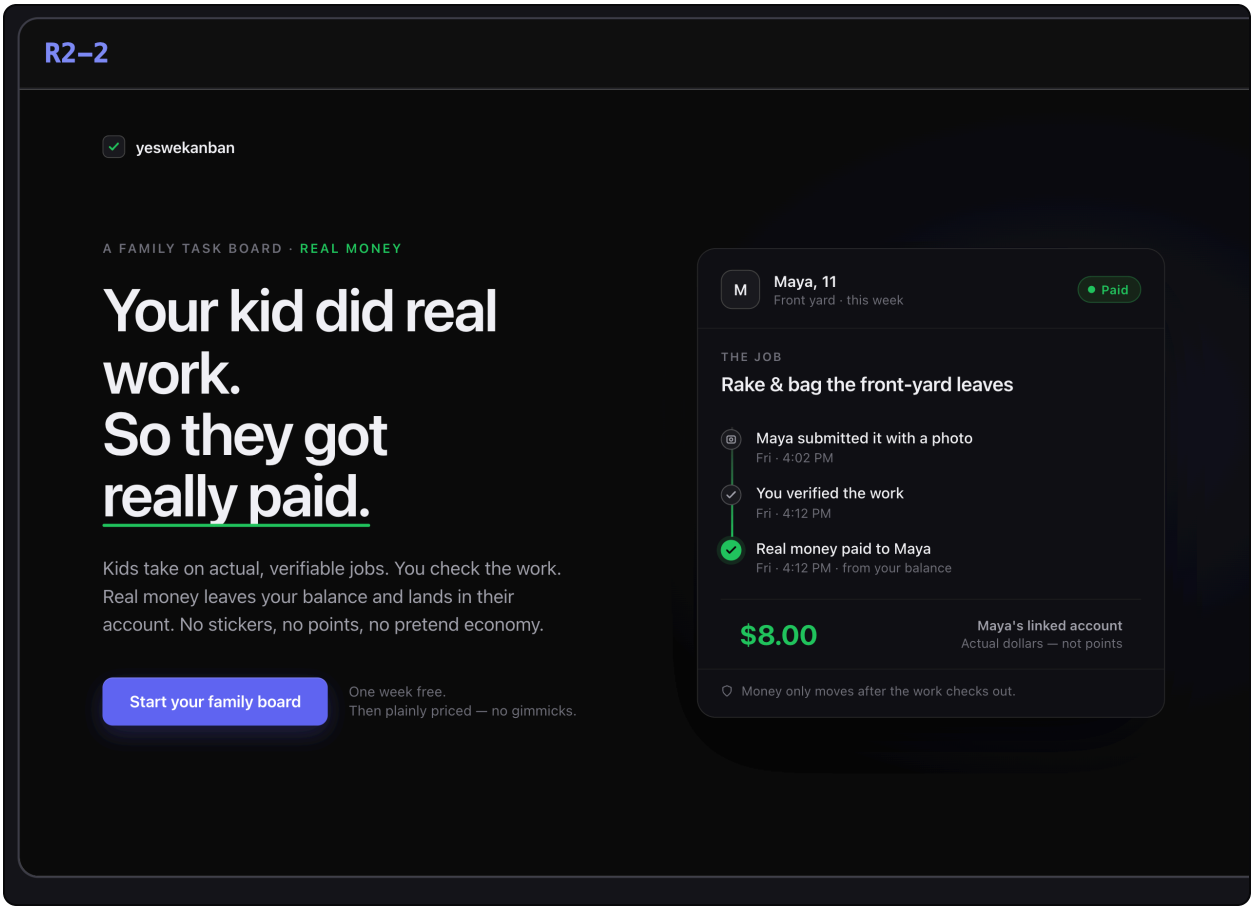


Fig. 3 — The home-turf split. The persona arm the founder rejected, and the parent panel ranked first.

RUNE · "YESWEKANBAN" (HOME) — THE DIVERGENCE *IS* THE FINDING

arm	Founder	Panel (mean)
Mindset (Rune)	**rejected**	**1.6 — tied #1**
Control	rejected ("least-bad")	**1.6 — tied #1**
Rules	**rejected**	3.2

arm	Founder	Panel (mean)
Full	**rejected**	**3.6 — worst**

No clean "overall": the founder gave no 1–4 ranking — he rejected all four — and that gap between his bar and the customer's *is* the result.

So the divergence isn't the founder being fussy about a design the panel proved was good. The two graders were scoring different things. The panel scored *appeal* — which it could see — and an appealing design can still be an untrue one. The founder scored *accuracy*: whether the design tells the truth about the product, which only someone who knows the product can judge. An appealing misrepresentation is still a misrepresentation, and an audience proxy will happily rank it first. The practical caution: a customer panel is the right judge of whether copy *lands*, but it cannot tell you whether that copy is *true to your product*. That second bar is yours to hold — and the knowledge it takes lives neither in the brief nor in the persona file yet.

7 The overall ranking

Compressing eight heterogeneous experiments into a single ranking is a judgment call, so treat the decimals as indicative, not precise — but the pattern is robust:

MEAN RANK ACROSS ALL 8 EXPERIMENTS (LOWER IS BETTER)

configuration	Detection	Generation	Overall
Full (stance + checklist)	**1.60**	**1.50**	**1.56 — best everywhere, never worst**
Mindset (stance only)	2.00	3.17	2.44 — strong on detection, over-reaches on generation
Control (bare)	2.70	2.17	2.50 — weak reviewer, strong writer
Rules (checklist only)	**3.70**	3.17	**3.50 — worst; actively suppresses on detection**

Mindset (2.44) and Control (2.50) look close overall but are strong in *opposite* regimes — averaging hides the inversion. Full is the only arm best in both.

8 What it means

- **Always load the full persona.** It's never the worst arm and wins overall. There's no task in this study where a lesser configuration is preferable.

- **Never ship a bare checklist as a review instrument.** This is the strongest actionable result: on the flaws that matter, a checklist alone is worse than nothing, because it suppresses them. If you run adversarial review, the reviewer needs the *stance*, not just a list.
- **Spend the persona where it pays: adversarial judgment.** Review, red-teaming, threat-modeling, discovery discipline, plan critique. On raw production, a capable model with a sharp brief is already hard to beat.
- **On generative work, use both graders — an audience proxy for whether it *lands*, and yourself for whether it's *true*.** A panel reliably scores appeal but cannot catch a misrepresentation of your own product; that bar is yours.
- **Write down the tacit model.** The home-turf gap showed the moat is brand knowledge the persona file doesn't yet capture. Capture it.

---

9 Limitations

We'd rather state these plainly than let the result read as bigger than it is.

- **Small N, one model.** N = 6–10 per detection cell; generation at one draft per arm, where a single typo or lucky concept can swing a cell. Everything ran on one model family (Opus 4.8). The effects are large and often replicated across two authors, but the decimals are indicative.
- **Two honest nulls.** Legal review of short prose and Omega's small self-contained bug-fixes both ceilinged — every arm scored ~100%, so nothing discriminated. We report them as nulls rather than mine them for a win. (A harder instance of either task might well separate the arms; we don't have that data.)
- **Taste has no single ground truth.** Generation graders diverged — on one bereavement-copy brief the founder ranked an arm worst that the panel ranked best. That divergence is data, not error, but it means "resonance" is grader-dependent by nature.
- **Scope.** Microcopy, single heroes, small bugs, short review targets. Not evidence about long-horizon feature delivery, multi-surface design systems, or campaign-scale strategy — where the personas' iterative processes may pay differently.

---

10 A final note on personas

This is a paper about AI personas, and its experiments were designed and run by AI personas. The study was built by our meta-systems persona; its exams were authored by other agents so it could not grade itself; and along the way it overturned two of its own operator's predictions — including the home-turf hypothesis that the design persona's product-specific tuning would give it an edge, which the data flatly refused. We think that's the point, not a footnote: an instrument earns trust precisely when it tells its author he was wrong. The persona layer is real capability, not costume — but the way to know that was to build something that could have proven the opposite, and let it try.

— Graydon Garden, founder of yeswekanban