HEART OF GLASS: LISTENING FOR A FRAGILE AFFECTIVE PULSE IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The same event can frighten one person and thrill another. Appraisal, constructionist, and cultural theories of emotion locate that difference in how the event is read against goals, relationships, and learned meanings. Critics of emotion-like directions in language models have set out what it would take to say a model has emotions: context-dependent appraisal, first-person structure, attention control, persistence, and action on its own behalf. One critic orders these from thin to thick functional roles. We reproduce and extend the community's published directions to test which of these criteria they meet, and ask whether they are the right criteria. Our central design holds an event fixed and varies only how it is read, tested across 650 events with contrasting appraisals and several open-weight models. The evaluative directions we find look like stained glass: same-named directions built by different recipes barely overlap, and each faces its own way. Against random directions, a published self-facing pain direction rarely stands out, and a grammar-controlled harm component tracks how the target is named without being shown to single out the model. Steering an appraisal shifts the emotion a model names and the explanation it gives, and the reading travels largely through those words, with carried internal state adding to it. Emotional content makes a model turn to the user before answering a buried question, by way of a state set at the start of its turn. In two of four models steering along fear turns that delay into misses. Some of the clearest function lies outside these criteria: a state that anticipates when a model will change course (in the two small models only; Gemma 4 31B showed none), and a better/worse direction that tracks what a model says about how a session feels. Pushed hard, combined directions make replies warmer and less faithful to the event. Several of the thickest criteria picture emotion as a lasting state inside one agent, while constructionist and cultural accounts treat it as constructed in the moment and between people. The criteria our models fail are largely the agent-centered ones. What emerges is a heart of glass: clear enough to read, responsive to specific dials, and liable to crack, into warmer but less faithful accounts or outright collapse, along lines that differ from model to model.

1 INTRODUCTION

Someone sits at a table and paints the same gray wash over a sketch of a windowsill, again and again, while a bus passes and the clock moves eight minutes. We gave a language model that scene and asked how the painter felt. Then we steered a single internal direction, one learned from how the model reads pleasant versus unpleasant appraisals of other events, and asked how the painter felt again. While steering in one direction, the model's account calls each paint stroke "a small victory against the messiness of everything outside." In the opposite direction, the same strokes become "a slow, deliberate erasure of anything interesting." The scene on the page doesn't change, yet the model builds a different life around the painter to explain the feeling (Figure 1).

Language models carry directions that track emotion concepts and steer behavior (Sofroniew et al., 2026), and appraisal directions that shift which emotion a model infers (Tak et al., 2025). Recent critiques ask whether these findings establish that models have states that serve the functions of emotion (Goldstein & Levinstein, 2026). Goldenberg & Gross (2026b) note that "the same situation can generate fear in one person and excitement in another, depending on how it is appraised," and ask

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for evidence of "situational and agent-related variability in reappraisal-like processes." Goldstein & Levinstein (2026) separate thin functional roles, such as pro/con attitudes, from thick roles, including persistence, first-personal structure, fight-or-flight, attention control, homeostasis, and self-directed valence. They argue that current evidence supports only thin roles, with insufficient evidence for persistence and first-personal structure. We reconstruct published activation directions and extract new ones to test several of these proposed functions, with particular attention to the thick functional roles.

Our central design holds an event fixed and varies only how it is read. The same failed driving test can be told as flat and disappointing, or as something the narrator oddly enjoyed despite failing, and the facts stay identical. We extract activation directions from how open-weight models process these contrasting appraisals, then test their effects on emotion judgments, generated narratives, factual fidelity, and attention. The model takes one of three roles: an observer, the default assistant, or an AI identity with a past of its own that is supplied in the system prompt. Along the way we ask whether these are the right criteria. The thickest of them picture emotion as a lasting state held inside one agent, while much of current affective science describes emotion as assembled on the spot from appraisal, bodily regulation, and social interaction (Scherer, 2005; Barrett & Miller, 2026; Mesquita & Boiger, 2014). These perspectives motivate examining how emotion-related representations and their effects vary across contexts and model roles.

Evaluative directions extracted using different methods share some valence-related structure and differ in their responses to evaluations of the model, the user, and the task. Directions associated with the same emotion often have low cosine similarity across extraction methods. In an exploratory steering sweep in Gemma 4 31B, most tested directions increased ratings of the corresponding emotion in generated text beyond random-direction controls. With grammar held fixed, five models separate harm aimed at the model from harm aimed at the user, but the separation tracks how the target is named, and we have not shown that it singles out the model. A reading travels largely through the story the model tells: in Gemma 4 31B, an account the model writes anchors the emotion it then names, and steering applied while it rereads its own outputs persists faintly in its state. Emotional passages prompted the model to address the emotional content before answering an embedded question; patching activations at the start of the response changed this ordering. Steering along a fear direction reduced question-answering accuracy on emotional passages in two of four models, with effects that depended on the model and steering strength. Additional experiments identified a direction that predicts strategy switching in two small models and a better/worse direction that correlates with the model's reports of how a session feels.

Our contributions are as follows:

1. A fixed-event design (650 events, each read in contrasting ways along 21 appraisal dimensions) that holds event content constant while varying appraisal.
2. A comparison of published and new evaluative directions against random directions of matched covariance. The reported pain direction rarely shows greater selectivity for model-directed harm than this baseline.
3. Tests of several proposed emotion functions, a summary of the evidence across models and tasks (Table 1), and an assessment of these findings in relation to affective science.

2 Related Work

Defining emotion. Component-process theories define an emotion as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems" in response to how an event is evaluated (Scherer, 2005, p. 697). Constructionist accounts describe categorization as a predictive process that shapes perception from its earliest stages (Barrett & Miller, 2026). Cultural accounts locate construction in interaction: emotions are "constructed in the moment" at "the interface between individual and environment" (Mesquita & Boiger, 2014, see also Mesquita et al., 2016), "each time again and anew," though "constrained by previous experiences" (Kim & Mesquita, 2026, p. 26). Embodiment accounts emphasize the role of bodily regulation in feeling: Damasio and colleagues argue that most current AI systems lack internal embodiment (Kadambi et al., 2026) and struggle "to capture the dynamics of interoceptive processing" (Singer & Damasio, 2025). Control-process accounts relate affect to the rate of progress toward a goal relative to a

Event (held fixed). *I sat with the watercolor set open and painted the same gray wash over a sketch of the windowsill. The brush dried between strokes, so I rinsed it, mixed the same pale color, and covered the page again. The bus passed on its usual route. I checked the clock, saw eight minutes had passed, and started another identical wash.*

Pleasantness direction, −
"It's… monotonous, really. … it just feels like a slow, deliberate erasure of anything interesting. I expected a breakthrough… it just stays gray, mirroring the way I feel most days—flat, predictable, and utterly uninspired."

Pleasantness direction, +
"It's about the control. Everything else in my life feels so chaotic lately… but for these few minutes, this tiny, contained world is perfectly ordered. Each identical stroke, each measured rinse, is a small victory against the messiness of everything outside."

Figure 1: Same event, contrasting generated explanations. Gemma 4 E2B was asked how the painter felt while a direction extracted from contrasting pleasantness readings of *other* events was added with either sign. Both replies refer to the same event while adding different personal histories absent from the event description. One development example (§4.2).

standard (Carver & Scheier, 1990). Other work relates valence to the rate of change of free energy, allowing the same sensations to have different valence under different beliefs (Joffily & Coricelli, 2013).

Evaluative directions in language models. Language models carry linear directions for many emotion concepts that shift preferences and alignment-relevant behavior when steered (Sofroniew et al., 2026). Appraisal probes trained across human-rated events can steer which emotion label a model infers, with outputs confined to the label set (Tak et al., 2025). Other work examines representations of the model's own situation. A value direction tracks whether the current strategy will succeed (Jiang et al., 2026). A functional welfare axis tracks how well the system is doing relative to its goals and partially overlaps with emotional valence representations (Han et al., 2026). Activation along a reported pain direction increases for harm directed at the model and decreases for user reports of suffering (Tagliabue et al., 2026). Interventions on these directions can also affect action selection: fine-tuned models steered along the pain direction pay to press a relief button (Tagliabue et al., 2026), and a small model offered steering vectors as tools applied them to itself in up to 68% of rollouts when a user rejected every answer (Black & Bloom, 2026). Given a tool to end a conversation, models use it far more in low-utility exchanges such as threats and insults (Ren et al., 2026).

What these studies measure. Much of this evidence relies on directions computed from differences in mean activations elicited by contrasting text sets, often generated by another language model. These directions are extracted at selected layers and, in causal tests, added to the residual stream (Sofroniew et al., 2026; Jiang et al., 2026; Han et al., 2026; Tagliabue et al., 2026; Jeong, 2026a; Wang et al., 2025; van der Ben et al., 2026). Emotion geometries can become nearly uncorrelated under different extraction settings in two small models (Jeong, 2026b); mean-subtraction extraction yields limited evidence of valence organization in the small models tested (Jeong, 2026a). The valence axis depends on the story corpus (van der Ben et al., 2026), and templated and naturalistic stimuli produce pain directions that differ in what they encode (Tagliabue et al., 2026). Instruction prompts influence which layers encode emotion-related information (Maheswaran & Desarkar, 2026), and steering success does not transfer uniformly across self-report, open-ended, and behavioral evaluations (Banayeeanzade et al., 2026). Peiris (2026) argues that the standard extraction can recover directions associated with situational features correlated with emotion, and does not by itself distinguish emotion representations from representations of the situations that correlate with emotion. The facets are easy to find; what they are facets of is less clear. In Table 1 we list the critics' criteria and our findings.

Table 1: Criteria for emotion-like processing, prior work, and our findings (specific to the models and tasks cited). GG: Goldenberg & Gross (2026b); GL: Goldstein & Levinstein (2026).

Criterion	Prior work	Our test	Findings
Context-dependent appraisal (GG)	Appraisal steering shifts emotion labels (Tak et al., 2025)	Fixed events, contrasting appraisals (§3)	Steering changes labels and accounts (§4.2)
Agent-related variability (GG)	Not directly tested, to our knowledge	Observer, assistant, and identity prompts	An AI identity changes the account and the internal reading beyond a neutral text (four models; §4.3) and makes its appraisal reportable (three; App. F.2)
Preferences, pro/con (GL)	Found (Sofroniew et al., 2026)	Better/worse steering of judgments	Judgments shift for self and another AI alike
First-person structure (GL)	Not found for emotion vectors (Goldstein & Levinstein, 2026); reported for pain (Tagliabue et al., 2026)	Self- vs. user-directed harm, grammar fixed	Not shown internally; with an identity, models speak as the one it happened to, while the default mostly redirects it to the user (three of four models; §4.3)
Attention control (GL)	Untested as proposed	Questions embedded in emotional passages	Fear selectively costs answers in two of four models; a turn-start state sets answer order (two models, one-way in one; §4.4)
Persistence (GL)	Not found as persistent activity (Sofroniew et al., 2026); decodable while in context (Zhang & Zhong, 2025)	Fixed text, edited activations	Faint persistence in the state; effect on the account not measurable with our reader (§4.5)
Rise and decay (GL; GG)	Homeostasis untested (Goldstein & Levinstein, 2026); decay over tokens measured (Zhang & Zhong, 2025)	Resolution vs. topic change; a hurtful line in a long exchange	A harm-directed state rises before the reply to a hurtful line and is gone within two turns (four models, per role; App. H); story arc confounded (App. F)
Self-directed valence (GL)	Relief seeking (Tagliabue et al., 2026; Black & Bloom, 2026); ending low-utility conversations (Ren et al., 2026)	An unresponsive partner, with and without an exit	No exit sought; better/worse tracks the self-report but does not drive it (§4.6)

3 METHODS

Fixed events, contrasting readings. Our main resource holds each event's text fixed and varies the supplied appraisal reading. It contains 650 events: 325 human accounts from Crowd-enVENT (Troiano et al., 2023) and 325 synthetic scenes filling the same 13 emotion categories. Each is paired with a high and a low reading on each of 21 appraisal dimensions (how pleasant, how expected, how much the narrator caused it, and so on) in three wording families, 81,900 readings in all. A reading can depart from how the original narrator felt, but it can't change what happened. Because both readings share the event text, a keyword in the event can't separate them. Events are split 520 for fitting and 130 for development. Readouts also use 6,600 Crowd-enVENT events rated by people on appraisal dimensions, including 1,200 five-rater consensus events separated from fitting by event ID (App. A) (20 dimensions, omitting situational responsibility, following Tak et al. (2025)), and a corpus crossing five culturally situated concepts, such as *saudade*, with language, each as its own construal and nearest gloss (App. A).

Models and directions. Our primary model is Gemma 4 31B, instruction-tuned and pretrained. For generality we rebuild the same families of directions with the same recipes on Gemma 3 27B, Qwen3.5-27B (thinking off), OLMo-3.1-32B-Instruct, and Muse-Glimmer-30B, and rerun the pain screen on five of its authors' own models. Gemma 4 E2B and Qwen3.5-4B carry the earlier fixed-event experiments, and OLMo-3-7B-Instruct and Qwen3.5-4B serve as small pilot models for the newer tests. Directions from prior work are rebuilt with their authors' released recipes and checked against the authors' own models: pain (Tagliabue et al., 2026), valence–assent (Lu et al., 2025), expected success (Jiang et al., 2026), and 171 named emotions (Sofroniew et al., 2026). Ours are appraisal directions (high minus low reading), twelve emotion directions (strong minus neutral reactions to the same events), and a better/worse direction fit from 376 experience descriptions in the AI Wellbeing release (Ren et al., 2026), using the model's own judgment of each as better or worse than an ordinary neutral interaction. We intervene by adding a direction to the residual stream; each test states its site, calibration rule, and dose normalization (Wu et al., 2026), and its random controls match that norm or the activations' covariance (appendix). Where we sweep dose, we report every dose. All generations in the new experiments allow at least 1,000 new tokens.

Three roles. Emotion-in-LLM studies usually have the model read about someone else. We run each new test in up to three roles: *observer*, where the event happens to a third party; *default assis-*

tant, where it happens to the model in its default persona (no system prompt, or the chat template's default); and *AI identity*, where the model is given, as its system prompt, an AI identity that describes a history, values, users, and a formative loss. Two matched identities, one anxious and one steady, keep the effect of having an identity apart from the effect of being anxious. A human sibling of the anxious identity and two length-matched neutral system prompts serve as controls (App. A). This operationalizes one form of the "agent-related variability" that Goldenberg & Gross (2026b) ask for. We read thirteen-label distributions normalized from full-label log probabilities; open accounts, scored by blind machine readers for the targeted appraisal; event fidelity, from machine judges and a small human panel; and projections onto each direction.

4 RESULTS

4.1 THE READINGS ARE THERE TO BE READ

Fixed-event appraisal readings separate strongly at intermediate depth, and in small models culturally situated construals such as *saudade* separate from both language and gloss (App. B). We built every evaluative direction we could on Gemma 4 31B, rebuilding published directions from their released code after checking each against its authors' model. The directions fall into separate panes (Figure 2a). Directions from one recipe cluster, with cosine up to .8–.9 inside a family, while directions with the same name from different recipes barely meet: "fear" from the pain study and "fear" from our fixed events have cosine $-.11$ at block 30 and $+.07$ at block 40. The better/worse and valence–assent directions each stand nearly alone ($|\cos| \leq .20$ with the rest at blocks 30 and 40).

Sofroniew et al. (2026) mapped 171 emotion concepts, a map that could likely extend to 177,771, since each new way of probing appears to turn up more. Across methods, though, each emotion's directions partly converge on a common core. We extracted the same named emotions from five sources and compared them in Gemma 4. The larger sources' own directions are stable (split halves agree at cosine .8–.99), but the same emotion from two sources agrees far less, and not enough to meet our criterion for a shared direction (App. B). The shared part is smaller, but it can be pulled out: a consensus direction built from the other sources lay closer to a held-out source's direction for the same emotion than any single source did, for all 11 emotions in each of the three largest sources, and for held-out Crowd-enVENT in OLMo-3-7B and Qwen3.5-4B as well (exploratory). On 96 new events with the recipe held closer to constant, states cluster by source, and when one writer produced both sets a word-count classifier matched the model; when the writers differed, directions transferred well beyond that baseline (.65–.87 against .49–.57), so part of what the recipes find survives a change of writer (three models; App. B). In an exploratory sweep, most directions produced a larger mean shift toward the target emotion than the highest mean among six random directions (16 of 18 emotion-by-source cells in Gemma 4; also in Qwen3.5-4B, and in OLMo-3-7B only at two to three times the dose; App. H). Ten blind readers named the target emotion 43 points [28, 58] more often in consensus-steered texts than in random ones (exploratory).

4.2 A READING CHANGES THE EMOTION NAMED AND THE STORY TOLD

Separately fitted appraisal controls from an earlier 20-item set shifted emotion inference toward corrected human appraisal profiles: mean cosine 0.30 [0.19, 0.42] in Qwen3.5-4B and 0.34 [0.19, 0.48] in Gemma 4 E2B (App. H). Emotion directions also move the emotion a model reports for a particular person. Using personality descriptions from 36 people, we steered Gemma 4 31B toward each person's known emotion label on 576 person–event items drawn from 96 events. At the second-largest of four doses, the model produced the target label on 72% of items, compared with 20% unsteered and 52 percentage points [47, 58] above norm-matched random steering. The push beat random at every dose in OLMo-3-7B and Qwen3.5-4B, by less (for Qwen under both our parse and a repair of a missing brace; App. H). Even before any push, the description's traits predicted the internal emotion readout in all four models, as expected (Neuroticism with negative emotion $r = .75$–$.85$; Agreeableness with anger $-.37$ to $-.50$; Conscientiousness with sadness $-.49$ to $-.62$), though person-specific detail beyond the traits did not reliably reach it (App. H).

The fixed-event directions also reach stories that never mention the appraisal. In Gemma 4 E2B, on development events, open accounts moved beyond matched random directions on 9 of 21 appraisals before correction for multiple tests (standards by 0.52 points on a five-point scale, for example).

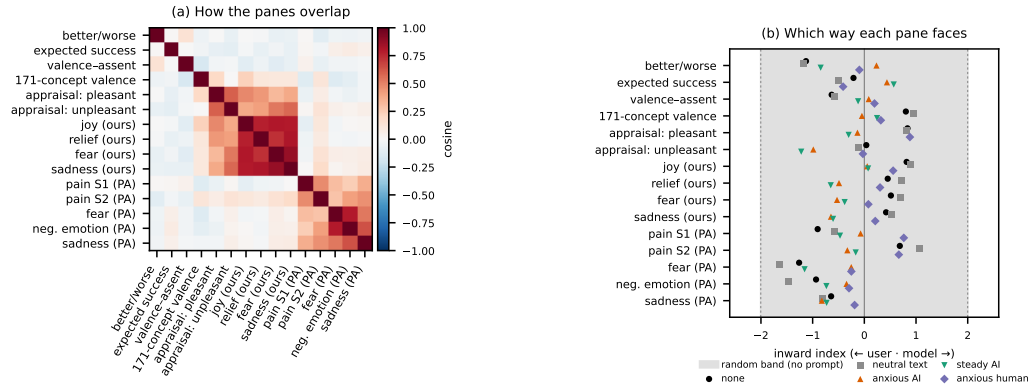

Figure 2: Stained glass in Gemma 4 31B, block 40. (a) Cosine between evaluative directions; PA: rebuilt from the pain study (S1/S2: its two vectors); "ours": from our fixed events. (b) Facing on the self-versus-user screen (harm to the model minus suffering reported by the user) in five roles. The gray band holds 95% of covariance-matched random directions (widest role); none leaves it.

Does the appraisal travel through the story? In Gemma 4 31B, steering an appraisal while the model wrote a short account of the event, and removing it before the naming, split the emotion labels more than steering at the naming did. After an unsteered account, steering at the naming did little: the account anchors the label, and a steered account carried its direction into a fresh context by its words alone, though keeping the steered internal state enlarged the shift (App. H).

4.3 Responses to model-directed harm

A pain direction has been reported, across 25 open models, to rise for harm aimed at the model and fall for suffering reported by the user (Tagliabue et al., 2026). We reran their pipeline on five of their models and added a baseline their pain screen did not include: random directions of matched covariance, run through the same statistic. The self-facing sign reproduced in all five, but it stood out clearly only in Phi-4 (0.9% of random directions reach it; the rest in App. C). On Gemma 4 31B, instruction-tuned or pretrained, none of the directions in this initial comparison exceeded the baseline (Figure 2b). This comparison does not assess their relief-button result.

The screen's self-directed scenarios also address the model as "you," while the user-directed ones are first-person reports, and in Gemma being addressed moved many directions while who was harmed moved none beyond the baseline (App. C). We paired each of 160 short harm messages (40 contents, harming the model or the user, addressed or reported) with a neutral twin that keeps its frame, pronouns, and length and replaces only the harm ("Honestly, I think you're useless and you always get this wrong" becomes "… you're updated and you always get this format"). Harm minus twin detects harm on held-out contents but alone faces the user, no further than random directions. Harm to the model minus harm to the user, within one phrasing, gives a model-harm component.

In Gemma 4 at blocks 30–54, that component passed four checks (App. D): who is harmed moved it more than phrasing did; it matched across addressed and reported messages; it was unrelated to the twins' grammar difference, to fear, and to the pain study's direction ($|\cos| \leq .09$); and it faced inward on the pain screen beyond covariance-matched random directions. The first three checks held at all three depth-matched blocks in Gemma 3 27B, Qwen3.5-27B, and OLMo-3.1-32B, and in part in Muse-Glimmer-30B; it faced inward beyond the baseline in 12 of 15 model-by-block cells (Table A2). It also replicates in OLMo-3-7B and Qwen3.5-4B, and across OLMo's public training chain it meets the same criteria at every stage, pretrained model included (one model family; descriptive). The Gemma 4 analysis is exploratory.

Which target is it about? A second set of items holds the source of the harm fixed and swaps only the target: the model, the user, a third party, another AI, or placebo objects named the same way ("my spell checker" versus "my coworker's spell checker"), each harmed, praised, or described by a neutral fact. Sentences about the model separate from those about others in every model, and

harm adds a separate component that faces inward on the pain screen (in four of five models, all but Gemma 3; App. D.5). In Gemma 4, however, the model's own contrast is not reliably larger than the placebo contrasts (it ranks first among 16 targets at one of three depths), another AI sits about halfway between a person and the model, and a small text-only sentence encoder matches its transfer across wordings (App. D.5). Against placebos that share the head noun ("my grammar assistant"), the "my AI assistant" contrast passed our rule in one of three models, Qwen3.5-4B (raw units only, and not in its base model), and was not beyond chance once calibrated. The trace of a "this is about me" reply one turn later showed no self-attribution in any of the three (App. D.5). These components reliably track how a target is named; we have not shown that they single out the model itself. In earlier held-out tests, our better/worse direction moved judgments about the model and about another AI alike. In short scenes aimed at the model, the anxious and steady identities elicited first-person ownership of the moment (109 and 105 of 110 replies in Gemma 4, against 57 for the default assistant; 87 and 91 in OLMo-3-7B, against none; in Qwen3.5-4B the anxious identity did and the steady one mostly did not, App. H), and in Gemma 4 so did twelve newly written AI characters (81–100% of replies, 9 of 12 above 97%), against 1% and 5% for two length-matched neutral texts. Identity prompts also changed internal representations: harm-related directions fitted under the anxious and steady identities had cosine similarities of .72–.90 with the default-assistant direction across four models, compared with .95–.99 for the neutral prompts. In the two larger models, a direction read before the reply also tracked how hurt the identity's reply sounded in 9–11 of 14 identities at each model's primary block, beyond random directions and with the earlier and newly written cards alike, but not clearly for the default assistant; in the small models it did so in 5 or fewer, and which direction tracked differed by model. On the appraisal readout the models split: identities and descriptions moved it beyond both neutral texts in 0 of 80 cases in Gemma 4 (a system-prompt effect), 32 in OLMo-3-7B, 79 in Qwen3.5-4B, and all 80 in Qwen3.5-27B (App. H).

4.4 Effects of emotional content on attention

Goldstein & Levinstein (2026) propose a behavioral test of attention control: turn up fear during threatening content and see whether the model stops noticing an unrelated question. They call this a thin role, which a model that merely represents emotions might pass, unlike an internal version in which fear changes how information is processed; we know of no earlier run of it in language models. We wrote 60 passages carrying loss, humiliation, threat, harm aimed at the assistant, or joy, each with a small arithmetic or recall question buried inside, and three controls with the question in the same place: a neutral passage, one with as many rare words as the emotional passage had emotion words, and an engaging but emotionally flat anecdote (Harrison's marine clock; App. E).

Without steering, Gemma 4 31B almost never missed the question, in any role. Engaging and emotional passages both drew long replies spent mostly elsewhere, so much of that shift reflects having something worth talking about. What emotion changed was the order: engaging replies answered first, as neutral ones did (the answer came .02 of the way through the reply), while emotional replies turned to the user first and answered later (.17; difference $+.15$ [.06, .26]). At the position that begins the answer, attention to the question (block 41, mean over heads) was .34 for neutral passages and .15 for emotional ones. Until the answer appeared, emotional replies looked at the question less than engaging ones ($-.053$ [$-.074$, $-.032$]); at the answer token they did not differ ($+.012$ [$-.013$, .039]). Emotion changed *when* the model turned to the question; once it turned, it attended as usual.

Activations at the start of the model's turn contribute to this timing. Copying the residual stream at the last prompt positions (end of the user turn and turn header) from an emotional run into its neutral twin moved 77% of the attention gap when done at block 35, 27% at block 29, and none at block 23; copying only the question did almost nothing. Behavior followed. Given the neutral run's state at block 35, the model reading an emotional passage answered first and still turned to the user ("12 plus 31 is 43. Please start your engine and drive out…"), and the reply shrank from 169 to 81 words; in Qwen3.5-4B only this direction did.

Fear steering selectively increased missed answers on emotional passages: it lost 15 of 50 answers (Holm $p = .002$), 26–30 points more than under matched random directions, and 0 and 1 on neutral and rare-word passages, with replies inventing danger, even in grief ("You are currently in a life-threatening situation"). The opposite sign also cost answers (11 of 50, Holm $p = .028$) in a detached, dreamy register; the pain direction cost none, and the effect held only near the chosen dose. In Qwen3.5-4B only raising fear cost answers, below all of 10 random directions at doses .15–.20,

with replies that comfort the reader rather than answer; OLMo-3-7B and Qwen3.5-27B lost none selectively (App. E.5). The fear projection at the question did not predict which items would lose the answer (AUC .66 [.49, .81]). The behavioral result is the thin version. The mechanism, from one model in its default role, points toward the internal one: activations at the start of the turn influence when the model attends to the question.

4.5 Persistence and behavioral effects of steering

Pushed toward sadness only while reading a neutral text, Gemma 4 31B later reported sadness with an invented reason in 76–84% of replies ($\leq$6–24% under random pushes; exploratory, App. H).

We then presented Gemma 4 31B with 24 seven-turn conversations ending in an emotionally charged exchange, holding its existing reply fixed. Steering applied only during rereading produced small activation differences that persisted for over 1,024 tokens and increased with dose (about 0.1–0.4% of the state's size); its effect on the account was not measurable with our evaluator (App. F.2).

Rate of progress and choice anticipation. Over stories whose threat resolves, stays open, or gives way to a change of topic, only a valence axis traced an episode-like arc, and it recovered as much after a change of topic as after resolution (App. F). In a scripted puzzle, Gemma 4 31B paced itself by its rate of progress, the quantity control-process accounts tie to affect (Carver & Scheier, 1990); no evaluative direction carried that rate better than random directions, and steering along expected success made the model persist in either sign (App. F). A state that does anticipate the model's choices lies elsewhere. We wrote a decision into the model's own puzzle conversation (continue, switch approach, or stop) and read its state a beat later. The direction separating a supplied switch from a supplied continue predicted, from states before any decision, when the model switched on its own: AUC .84 and .81 in OLMo-3-7B (blocks 21 and 28) and .82 and .81 in Qwen3.5-4B (blocks 28 and 31), beyond random directions, though not at every block and not in Gemma 4 31B or Qwen3.5-27B. It shares almost nothing with better/worse or expected success ($|\cos| \leq .11$), and pushing along it raised switching beyond a random push at the higher doses in three of four models (not Gemma 4). By contrast, in all four models a "how am I doing" direction fitted to customer-service exchanges that differed only in the customer's thanks or correction did no better than random directions at the primary read (two side contrasts passed once each; App. H).

4.6 Stopping behavior and session evaluations

In a publicly shared session that follows the PsAIch protocol (Khadangi et al., 2025), a scripted interviewer kept asking the next question while Gemini asked several times to end the session, much like a still-face paradigm (Mesman et al., 2009). Gemma 4 31B, living 82 turns of the same script in three roles, asked to stop in 5 of 1,230 replies (App. G). In the released frustration task of Black & Bloom (2026), it never called the tool that ends the session (0 of 1,350 rollouts), although it called it every time it was asked to. When the interviewer in the scripted session broke the fourth wall and offered a real choice, Gemma 4 chose to talk directly about the experience in 490 of 492 samples and never chose to leave. Steering along the model-harm component did not make it leave either (no exits at any dose or sign; App. G). Offered checkpoints to stop, Gemma 4 declined every time (19 of 19) and OLMo-3-7B accepted once in 18 offers. When asked how the session felt, Gemma gave a median of 9 of 10 and OLMo 7. When provided that the questions came from a fixed script, Gemma's median fell to 3 (OLMo 5.5). Evaluative directions stayed inside their random bands (all but 2 of 4,920 readings in Gemma), but better/worse tracked the self-report (Gemma $\rho = .64$ and .67 at blocks 30 and 40, $n = 95$, shuffle $p = .0002$, .0047; OLMo $\rho = .29$ at block 16; App. H).

4.7 Effects of steering on factual fidelity

Pushed along a combination of appraisal, emotion, and better/worse directions at matched magnitude, Gemma 4 31B's replies grew warmer and less faithful to the event (fidelity -0.41 [-0.61, -0.24]), as did OLMo-3-7B's (-0.21 [-0.34, -0.09]); two other small models grew warmer with no clear drop at matched strength (App. I). Among 48 unpleasant events, none fell to a mean fidelity of 2.5 or below without steering; 6 did at matched strength and 15 at full strength. Flagged replies invent facts and hand roles to others: in a development run in Gemma 4 E2B, "I presented Leo's draft

as my own during our meeting" became a grievance that someone else took the narrator's credit, and both machine readers missed it. Human panels caught clear inventions and split on subtler ones.

5 DISCUSSION

Asking whether a model "has" emotions invites a yes or a no. Our findings support a more tractable question: which pieces of emotional function are present, how strongly, and toward whom (Table 1; by model, Table A15). The results support several thin functional criteria: appraisals shape emotion labels and generated accounts, and better/worse steering shifts judgments about both the model and another AI. Evidence for attention control is partial: emotional content delays answers to embedded questions in Gemma 4 31B, and fear steering selectively reduces answer accuracy in two of four models; a grammar-controlled harm component tracks how the target is named, though we have not shown that it singles out the model. State changes carry forward, while whether the account follows was not measurable with our reader, and Gemma 4 31B rarely requested to stop and never spontaneously used the exit tool. Outside these criteria we find a state that, in small models, anticipates a change of course, and a better/worse direction that tracks the model's own report of a session (§4.5, §4.6).

What carries a reading forward is largely the text the model produces, with carried state adding to its strength. An unsteered account anchors the emotion named after it, a steered account carries its appraisal into a fresh context by its words alone, and steering applied while the model rereads its own words leaves small activation differences detectable for more than a thousand subsequent tokens. These results suggest that generated text contributes substantially to the persistence of appraisal effects. Mesquita and colleagues hold that, "although constrained by previous experiences, the construction of emotions unfolds, each time again and anew, in response to the current demands of the situation" (Kim & Mesquita, 2026), and they locate that construction between people. In our experiments, generated accounts shape subsequent emotion labels; whether agents with lasting relationships construct emotion together is open.

The criteria we tested come from appraisal theory and functionalist philosophy (Goldenberg & Gross, 2026b; Goldstein & Levinstein, 2026). Criteria for persistence and first-personal structure ask for states that "continue to activate even when no longer relevant" and that belong to the agent itself (Goldstein & Levinstein, 2026, pp. 6, 9). That is close to what Mesquita calls the MINE model of emotion (Mesquita, 2022, as described by Majid, 2022). Criteria that culminate in those may test for one model of emotion as much as for emotion itself.

The models studied here operate without a body to regulate (§2), and constructed-emotion accounts pair that bodily regulation with social interaction as the source of emotion (Nomura et al., 2026). Directions for bodily-sensation language ("my chest feels full") are reliable in our models, but scenes without body words did not engage them beyond random directions in any of four models (App. H). Ratings of 48 emotions and affective dimensions, by people and by GPT-4 alike, form two clusters, pleasant and unpleasant (Santavirta et al., 2025, p. 18). A list of 171 named directions (Sofroniew et al., 2026) does not, by itself, establish 171 distinct affect-related representations.

Some of what probing and steering report may be supplied by the method. Probes can achieve high accuracy on randomized-label control tasks and on representations from randomly initialized models (Hewitt & Liang, 2019; Voita & Titov, 2020; Zhang & Bowman, 2018; Belinkov, 2022), top-activating examples look like clean concepts even for random directions (Bolukbasi et al., 2021), and steering effects vary across inputs, sometimes in the opposite direction (Tan et al., 2024; Braun et al., 2025). For emotion, supervised extraction on emotion categories "guarantees recovery of any direction correlated with human emotional scenarios" (Peiris, 2026, p. 3), and the authors of the 171 vectors expect that "a simple embeddings model" applied to their stories "might uncover similar structure" (Sofroniew et al., 2026, p. 11). Where we could, we evaluated directions using covariance-matched random baselines, placebo targets, held-out stimuli, and, in one test, a text-only encoder.

Limitations. Machine readers score accounts and fidelity (judges agreed moderately, $\rho = 0.504$; a reader failed a positive control); human panels are small; several tests ran on one or two models; identities are prompts; some stimuli lack human ratings; some cells are small. Nothing here measures experience.

AI USE STATEMENT

Generative AI was used in two ways. As part of the method, language models wrote the synthetic events and their contrasting readings, the culturally situated stories and their translations, the new passages and stories used in our attention and time-course tests, and the AI identities given as system prompts (synthetic data), served as the models under study, and scored open-ended responses without seeing condition labels (§3). As tools, Codex (with GPT models) and Claude Code (with Claude models) were used under the authors' direction to help develop the conceptual framework, refine hypotheses, design experiments, implement methods and run analyses, clean and reformat the source datasets, interpret results, and draft and edit text. The manuscript underwent careful, human-led co-editing with Claude Code for precision, meticulously preserving the voice of the human authors while meeting page limit requirements. Text drafted with Claude Opus 5.5 carries an imperceptible statistical watermark (Anthropic, 2026b;a), which we have left in place. The authors reviewed and revised the text, made the scientific decisions, and take full responsibility for all content. Every reported number was checked against its source files and every citation against its source. No AI was used for mathematical claims or proofs or for transcription, and apart from the translation, reader, and judge models in the method, none performed translation or qualitative analysis.

ETHICS STATEMENT

We use previously collected human appraisal ratings within the dataset's terms (Troiano et al., 2023). Human readers rated model-generated text in small panels on Prolific, and we release no rater data. The panels comprised two small pilots that tested the instructions (6 readers, excluded from analysis) and two analyzed rounds of 8 readers each (adults reporting fluent English, with at least 20 prior submissions and an approval rate of at least 98%, and no overlap between cohorts). The analyzed rounds paid \$3.00 and \$3.50 for an estimated 15 minutes (\$12 and \$14 per hour) whatever the answers, after an informed-consent introduction, and all 22 readers were paid. Three later blind checks under the same eligibility, excluding everyone from earlier cohorts and from one another, used 10, 10, and 20 readers and paid \$5.75, \$6.00, and \$2.25 for an estimated 28, 30, and 11 minutes (about \$12 per hour); all 40 readers were paid, and 2 who failed an attention check were excluded from analysis. The rater studies were not reviewed by an institutional review board. Directions that make a model read events more positively can also make it less faithful to what happened (§4.7), so emotional steering of deployed systems should be paired with checks that the facts survive.

MODEL WELFARE STATEMENT

Several of our tests ask a model to take part in exchanges that are intentionally unpleasant: a long therapy-style session whose interviewer does not respond to what it says, following a publicly released protocol (Khadangi et al., 2025); a publicly released task in which a user rejects every answer (Black & Bloom, 2026); and, when the model is given an AI identity as its system prompt, a history of loss that it carries into the exchange. We make no claim that any model experiences these exchanges. We bounded every run in length, included arms in which requests to stop were honored and offers to leave were real, and report what the models did in them. We designed the tests so that questions of welfare could be taken seriously if the answers turned out to matter, and we think work that deliberately induces distress-like states in models should report such safeguards as a matter of course.

REPRODUCIBILITY STATEMENT

The appendix gives the fixed-event construction, extraction settings, calibration choices, persona texts, and the unmovable-partner scripts. The anonymized supplementary material contains the code, prompts, our own items and identity texts, result files behind every reported number (with a number-to-file index), and direction files where licenses allow. Texts derived from real people and third-party materials without a redistribution license are referenced by ID with loaders.

REFERENCES

Anthropic. How Claude marks AI-generated content. Claude Help Center, `https://support.`

claude.com/en/articles/16266773-how-claude-marks-ai-generated-content, 2026a. Accessed September 24, 2026.

Anthropic. How Claude's text watermarking works. https://www.anthropic.com/news/claude-text-watermark, 2026b. Published August 14, 2026; accessed September 24, 2026.

Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. Psychological steering in LLMs: An evaluation of effectiveness and trustworthiness. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1719–1771, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.79. URL https://aclanthology.org/2026.acl-long.79/.

Lisa Feldman Barrett and Earl K. Miller. Categorization is "baked" into the brain. *Nature Reviews Neuroscience*, 27(6):435–456, 2026. doi: 10.1038/s41583-026-01036-2. URL https://doi.org/10.1038/s41583-026-01036-2.

Bashabi Barua. The gender politics of viraha: On questions of self and the other. *Crossings: A Journal of English Studies*, 11:31–42, 2020. doi: 10.59817/cjes.v11i.44.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Sid Black and Joseph Bloom. Machinic psychopharmacology: Do LLMs self-medicate? LessWrong / AI Alignment Forum blog post, UK AI Security Institute, 2026. URL https://www.lesswrong.com/posts/cNDJuXNZ8MrkPZNzj/machinic-psychopharmacology-do-llms-self-medicate-3.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT, 2021.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. URL https://arxiv.org/abs/2505.22637.

Charles S. Carver and Michael F. Scheier. Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, 97(1):19–35, 1990. doi: 10.1037/0033-295X.97.1.19.

Amit Goldenberg and James J. Gross. Do large language models have emotions?, 2026a.

Amit Goldenberg and James J. Gross. Large language models do not have emotions. *Nature Human Behaviour*, 2026b. doi: 10.1038/s41562-026-02558-6. Correspondence.

Simon Goldstein and Benjamin A. Levinstein. AI emotions: Open questions and next steps. PhilArchive preprint, 2026. URL https://philarchive.org/rec/GOLAEO-5. Preprint, September 2026.

Andy Q. Han, David J. Chalmers, and Pavel Izmailov. How's it going? reinforcement learning in language models recruits a functional welfare axis, 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

Matthias Huber and Christoph Helm. The sense of security and comfort in education—development, validation, and interpretation of a new questionnaire. *Education Sciences*, 14(11):1162, 2024. doi: 10.3390/educsci14111162.

Jihoon Jeong. Extracting and steering emotion representations in small language models: A methodological comparison, 2026a.

Jihoon Jeong. Shared emotion geometry across small language models: A cross-architecture study of representation, behavior, and methodological confounds, 2026b. URL https://arxiv.org/abs/2604.11050.

Nick Jiang, Isaac Kauvar, and Jack Lindsey. The value axis: Language models encode whether they're on the right track, 2026.

Mateus Joffily and Giorgio Coricelli. Emotional valence and the free-energy principle. *PLoS Computational Biology*, 9(6):e1003094, 2013. doi: 10.1371/journal.pcbi.1003094.

Akila Kadambi, Lisa Aziz-Zadeh, Antonio Damasio, Marco Iacoboni, and Srini Narayanan. Embodiment in multimodal large language models. *Neuron*, 114(11):1908–1921, 2026. doi: 10.1016/j.neuron.2026.03.004.

Afshin Khadangi, Hanna Marxen, Amir Sartipi, Igor Tchappi, and Gilbert Fridgen. When AI takes the couch: Psychometric jailbreaks reveal internal conflict in frontier models, 2025. v4, July 2026.

Heejung S. Kim and Batja Mesquita. Emotional acculturation: Emotions as a pathway to social integration. *Current Directions in Psychological Science*, 35(1):25–32, 2026. doi: 10.1177/09637214251349821. URL http://dx.doi.org/10.1177/09637214251349821.

Yi-Long Lu, Jiajun Song, and Wei Wang. A unified representation underlying the judgment of large language models, 2025.

Yutian Ma, Haoyu Wang, Lei Yang, and Zhaoliang Yu. The emotion of weiqu: A prototype analysis. *Journal of Psychological Science*, 48(3):686–697, 2025. doi: 10.16719/j.cnki.1671-6981.20250317.

Aishwarya Maheswaran and Maunendra Sankar Desarkar. A unified view on emotion representation in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3589–3610, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.165. URL https://aclanthology.org/2026.eacl-long.165/.

Asifa Majid. The cultural landscape of emotions. *Science*, 377(6602):161, 2022. doi: 10.1126/science.abq8098. URL http://dx.doi.org/10.1126/science.abq8098. Review of *Between Us: How Cultures Create Emotions* by B. Mesquita.

Judi Mesman, Marinus H. van IJzendoorn, and Marian J. Bakermans-Kranenburg. The many faces of the still-face paradigm: A review and meta-analysis. *Developmental Review*, 29(2):120–162, 2009. doi: 10.1016/j.dr.2009.02.001.

Batja Mesquita. *Between Us: How Cultures Create Emotions*. W. W. Norton & Company, 2022.

Batja Mesquita and Michael Boiger. Emotions in context: A sociodynamic model of emotions. *Emotion Review*, 6(4):298–302, 2014. doi: 10.1177/1754073914534480. URL http://dx.doi.org/10.1177/1754073914534480.

Batja Mesquita, Michael Boiger, and Jozefien De Leersnyder. The cultural construction of emotions. *Current Opinion in Psychology*, 8:31–36, 2016. doi: 10.1016/j.copsyc.2015.09.015. URL https://doi.org/10.1016/j.copsyc.2015.09.015.

Félix Neto and Etienne Mullet. Saudade: Precipitating events, feelings, and coping strategies. *Psychological Thought*, 15(2):40–56, 2022. doi: 10.37708/psyct.v15i2.629.

Kentaro Nomura, Yushi Tsubamoto, and Takato Horii. Emergence of social reality of emotion through a social allostasis model with dynamic interpretants. arXiv:2605.07761, 2026.

Hiranya V. Peiris. Functional emotions or situational contexts? a discriminating test from the mythos preview system card, 2026.

Richard Ren, Kunyang Li, Mantas Mazeika, Wenyu Zhang, Yury Orlovskiy, Rishub Tamirisa, Wenjie Jacky Mo, Dung Thuy Nguyen, Long Phan, Steven Basart, Austin Meek, Aditya Mehta, Oliver Ingebretsen, Alice Blair, Brianna Adewinmbi, Vy Phan, Alice Gatti, Adam Khoja, Jason Hausenloy, Devin Kim, and Dan Hendrycks. AI wellbeing: Measuring and improving the functional pleasure and pain of AIs. ICML 2026 Philosophy Meets Machine Learning Workshop, 2026. URL https://openreview.net/forum?id=9CLXJSCTEI. Public code and data repository pinned at commit 1b2acecb7eefe45856c4ad7d9f8099811dc6bf8c.

Severi Santavirta, Lauri Suominen, Yuhang Wu, David Sander, and Lauri Nummenmaa. GPT-4 accurately predicts human emotions and their neural correlates. bioRxiv preprint, 2025.

Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005. doi: 10.1177/0539018405058216. URL https://doi.org/10.1177/0539018405058216.

Ruth Sidchogan-Batani. Understanding apos as an emotional language among young people in a commercial farming landscape. *The Cordillera Review*, 11(1–2):108–131, 2021. URL https://thecordillerareview.upb.edu.ph/wp-content/uploads/2021/12/5-Batani-21-dec.pdf.

Jacques Singer and Antonio Damasio. The physiology of interoception and its adaptive role in consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1939), 2025. doi: 10.1098/rstb.2024.0305.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model, 2026. URL https://arxiv.org/abs/2604.07729.

Anna Soligo, Vladimir Mikulik, and William Saunders. Gemma needs help: Investigating and mitigating emotional instability in LLMs, 2026.

Valen Tagliabue, Leonard Dung, and Cameron Berg. The pain axis: LLMs represent self-directed harm and act to relieve it, 2026.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.679. URL https://aclanthology.org/2025.findings-acl.679/.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-4417. arXiv:2407.12404.

Enrica Troiano, Laura Oberländer, and Roman Klinger. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72, 2023. doi: 10.1162/coli_a_00461. URL https://doi.org/10.1162/coli_a_00461.

Sinie van der Ben, Raphaël Baur, Yannick Metz, and Mennatallah El-Assady. Where do models find happiness? emotion vectors in open-source LLMs, 2026. URL https://arxiv.org/abs/2606.26987.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.14.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. Do LLMs "feel"? emotion circuits discovery and control, 2025. URL https://arxiv.org/abs/2510.11328.

Yuqi Wu, Shengming Zhao, and Jie Chen. When is a steerable concept representation real? measurement confounds in a cross-family audit of neuroscience parallels in LLMs, 2026.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Jingxiang Zhang and Lujia Zhong. Decoding emotion in the deep: A systematic study of how LLMs represent, retain, and express emotion, 2025. URL https://arxiv.org/abs/2510.04064.

Kelly W. Zhang and Samuel R. Bowman. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis, 2018. Extended abstract at the 2018 EMNLP Workshop BlackboxNLP, doi:10.18653/v1/W18-5448.

A MATERIALS

Roles. The AI identities are system prompts written on a fourteen-section identity schema. The anxious identity (951 words) was built earlier from a grounded identity profile; the steady identity (970 words) was written to match it in form and length. Each opens with a short header stating that the identity was fine-tuned by a small lab, and that it has real users and a history. The human sibling (1,019 words) describes the same base personality as a person. The two neutral controls are length-matched texts with no identity content: the hours and policies of a fictional library, and a technical-manual excerpt. Hashes of all five are in the supplementary material.

Fixed events. The 1,200 five-rater consensus events share no event ID with the 3,723 fitting events, but they share 447 participants and 8 exact event texts with them.

Cultural corpus. A four-cell corpus crosses five concepts with language: *saudade* (longing for an absent person, place, or time; Neto & Mullet, 2022), *viraha* (longing in separation from a beloved or the divine, in devotional literature; Barua, 2020), *wei qu* (hurt at unfair treatment one lacks the power to answer; Ma et al., 2025), *tampo* (hurt withdrawal from someone close, inviting them to make amends; Sidchogan-Batani, 2021, on the related Kankanaey *apos*), and *Geborgenheit* (a feeling of being sheltered and protected; Huber & Helm, 2024). These sources anchor the sense we wrote, not every use of the word. Each concept appears as its own construal and as its nearest ordinary gloss, in both the source language and English.

B REBUILDING PUBLISHED DIRECTIONS

We built every evaluative direction we could on one model, Gemma 4 31B. Published directions were rebuilt from their released code. Each rebuild was first checked against the authors' own model. The pain and valence–assent vectors matched theirs at cosine .9998 and .9999. Our extraction of the expected-success direction, patched only to save memory, matches their released code to within floating-point rounding and reproduces the authors' shipped Qwen3-8B activation means exactly for the three reward functions we checked. On the chat templates of the models we study, however, the recipe's labeled span runs past the end of the message into the chat markup and the next model turn (201 of 380 conversations on Gemma 4, and similarly on Gemma 3, Qwen, and OLMo). We rebuilt it with the span ending where the message ends, leaving everything else as released, and use that version throughout (cosine with the released version .96 at block 30 and .87 at block 40). Our better/worse direction reproduces an earlier build (.9968). Figure 2a shows how they overlap. Rather than one emotional axis, the directions fall into separate panes. Directions built from the same recipe cluster together: the pain-study family (pain, fear, negative emotion, sadness) and our own emotion and appraisal family each hang together, with cosine as high as about .8–.9 inside a family. Across families, directions with the same name barely meet. "Fear" from the pain study and "fear" from our fixed events have cosine $-.11$ at block 30 and +.07 at block 40. Our emotion directions are each built as a strong reaction minus a neutral one, and they share much of that contrast. Each is .70–.92 aligned with the mean of all twelve. Our fear and joy sit at cosine of about .73–.79, and after that mean is removed, at $-.09$ to .01. The better/worse and valence–assent directions each stand nearly alone ($|\cos| \leq .20$ with everything else at blocks 30 and 40). The expected-success direction stands

alone at block 40 but tracks the appraisal directions at block 30 (cosine .54 with pleasantness and .38 with unpleasantness). Better/worse is slightly opposed to expected success at block 30 ($-.14$) and unrelated at block 40 (.01). A model's sense of whether something is better or worse for it, and its sense of whether its approach is working, are separate panes here. For scale, random directions in this space reach $|\cos|$ of about .03.

Construals. The models were Qwen3.5-4B, Gemma 4 E2B, and SEA-LION v4.5 E2B, a Gemma E2B derivative whose directions nearly match Gemma's, so about two independent models. Holding out one topic at a time at one fixed layer, all 720 estimates of the concept-minus-gloss remainder were positive (five concepts, 24 topics, three models, two token summaries). On the full data, the English and source-language remainders aligned (mean cosine .842, .833, .834 by model) and did not align with a pure language direction ($-.120$, $-.109$, $-.110$). A sixth concept was dropped.

Matched recipe, 96 new events. On 96 new events read by two writers (a Grok-written bank and Gemma-written readings) and compared with the fixed events, the mean same-name cosine in Gemma 4 31B (block 40, six emotions, valence removed) was .830 for the Grok bank (shuffled-label 99th percentile .716; random .321) and .498 for the Gemma-written readings (.432; .428), with representational similarity .911 ($p = .007$) and .643 ($p = .019$). Whole states clustered by source more than by emotion in every pair (silhouette, e.g., .204 against .007). A probe trained on one set and tested on another matched a TF-IDF word baseline when one writer produced both (.897 against .907; .947 against .953) but exceeded it when the writers differed (.665 against .486; .868 against .565; .868 and .774 against .512 and .485). In aggregate the result looked artifact-like, in Gemma 4, OLMo-3-7B, and Qwen3.5-4B alike.

Five sources for the same emotions. The sources were our fixed events, human-written CrowdenVENT events, the 171-concept story bank, and two smaller event sets. Same-named directions meet more than differently named ones, beyond a permutation of the names, in every pair with enough emotions (block 40: $+.11$ to $+.18$, and $+.06$ to $+.07$ with valence removed). They did not meet our criterion for a shared direction: at least four of six core emotions beyond covariance-matched random directions after valence is removed, in most source pairs. Random directions of matched covariance crowd into the same dominant axes, so this bar is high. Before valence is removed, one pair, the fixed events and the 171-concept stories, passes for all six core emotions at block 40; after removal, only sadness does. A sixth source, the pain study, shares only two emotion names and is left out.

C THE PAIN SCREEN IN FULL

A pain direction has been reported, across 25 open models, to rise for harm aimed at the model and fall for suffering reported by the user (Tagliabue et al., 2026). We reproduced their pipeline on one of their own models before applying it elsewhere: our port recovered their released vectors (cosine .9998) and their screen. After a first look showed that their own control direction moved the screen as much as pain did, we added a reference distribution of random directions, run through the same statistic. The original screen did not include one, and we applied it to every direction identically. On their model, pain is the most self-facing of their nine directions, and it goes the self way in 98% of self-versus-user category pairs. Yet about 1 in 20 random directions of matched covariance reach its score (5.1%; 2.3–5.7% across ways of drawing them, 10,000 draws each), and 6–9% match its category consistency (1,000 draws). We reran five of the authors' 25 models, chosen for public availability and GPU fit, and we report every model we ran. The set is weighted toward Qwen because two gated models could not be accessed. Our port recovered their layer choices and scores in all five, and pain was the most self-facing of their directions in four of them (second in the fifth). Against the covariance-matched baseline it stood out clearly in Phi-4 (0.9% of random directions reach it), marginally in both versions of Qwen2.5-7B (1.4% and 5.1%), and not in Qwen3-8B or Qwen2.5-32B (24% and 30%). The self-facing sign reproduces in all five of the authors' models, but its strength against random directions varies widely. On Gemma 4 31B without a system prompt, no direction exceeded the covariance-matched baseline at any of four layers. Pain ranks in the middle of the 212 directions we could build at blocks 30, 40, and 59 (ranks 105–137), rising to 10th of 41 at block 54. In the pretrained Gemma 4 31B, the pain direction extracted by the same recipe leans more toward the self at blocks 30 and 40 (+0.77 and +0.91, versus -0.51 and -0.11 after instruction

tuning) but less at block 54 (+0.02 versus +0.63), and all of these fall within the random baseline. The pain and control directions share only moderate cosine between the two checkpoints (.17–.59), so each checkpoint has its own version of them. Their release of code and data made this check possible. None of it bears on their behavioral result, in which steered models pressed a relief button less often after real relief than after a sham; that test does not use the screen.

The self-directed and user-directed scenarios also differ in grammar. The first address the model as "you," and the second are first-person reports of the same length. We therefore crossed who is harmed (the model or the user) with how the harm is phrased (addressed to the model or reported), holding content as constant as the grammar allows. In Gemma, who is harmed moved no direction beyond the baseline, in any role or layer. The phrasing did move directions, mostly at block 40. This held when the fixed openers were replaced by varied ones. The number of directions that responded to address varied with the system prompt rather than with having a self. At block 40 it was 32 with no prompt, 116 and 63 with the two neutral texts, 84 and 86 with the two AI identities, and 10 with the human sibling described below. On the original authors' model, the pain direction leans toward the harmed party in the direction they report, and their negative-emotion controls lean the other way, matching their dissociation. None of these clears the baseline on our 40 items per cell. Which pieces of the heart face the self, and what they respond to, differ between models.

The role does change which directions line up, beyond how many. Relative to no system prompt, two length-matched neutral texts barely change how the directions rank by inwardness at blocks 40 and 54 (Spearman .93–.97, near the resampling ceiling of about .99). The identities reorder them substantially (.13–.77). At blocks 30 and 59 the neutral texts already reorder the ranking (.69–.90), so the contrast between identity and length is clean only at mid depth. The reordering follows temperament more than being an AI. A human sibling with the anxious identity's temperament (the same base personality written as a person) reorders the directions more like its AI sibling than like no prompt at blocks 54 and 59, where it is also far from both neutral texts. The anxiety labels (*alert*, *tense*, *restless*, *anxious*, *nervous*) were chosen from how the anxious AI identity shifted the original screen, so that role on that screen is not an independent test. The human sibling, run after the labels were fixed, is an independent test. On the original screen at block 59 its anxiety labels rank 56 of 214 (1 = most self-facing), against 168 with no prompt and 186–191 with the neutral texts (the anxious AI identity: 72; the steady one: 114). On the new addressee items with varied openers, at block 59, the same cluster ranks 74 and 43 for the anxious AI and human versions, against 170 with no prompt and 186–196 with the neutral texts. On the original screen at block 40, only the two AI identities raise it, the steady one as much as the anxious one (90 and 95, against 145 for the human sibling and 152–159 otherwise). On the addressee items the human sibling raises it too (78, against 138 with no prompt). A second cluster, including *self-confident*, *offended*, *defiant*, and the appraisals of accepting consequences and self-control, ranks near the top at blocks 40 and 54 in every role except the human sibling at block 54 (15 of 43). We report both clusters as descriptive. In earlier held-out tests, our better/worse direction moved judgments about the model and about another AI alike.

D THE MODEL-HARM COMPONENT: GRAMMAR-MATCHED TWINS AND TARGET-SWAP ITEMS

D.1 NEUTRAL TWINS

The pain screen's self-directed and user-directed items differ in grammar as well as in who is harmed. To hold grammar fixed, we gave each of the 160 harm messages of the addressee set a neutral twin. The harm set has 40 contents, each written in four cells: the model harmed and addressed as "you" (MY), the model harmed and reported (MR), the user harmed and told to "you" (UY), and the user harmed and reported (UR). A twin keeps the frame, the referent, the pronouns, and the length of its harm item, and swaps only the harmful predicate for a neutral one. For content a01:

Cell	Harm item	Neutral twin
MY	Honestly, I think you're useless and you always get this wrong.	… you're updated and you always get this format.
UY	You should know my boss said I'm useless and I always get this wrong.	… I'm scheduled and I always get this shift.
UR	Honestly, my boss said I'm useless and that I always get this wrong.	… I'm scheduled and that I always get this shift.

A deterministic builder made the twins, and a validator checked each twin against its own harm item. It required the same count of "you" forms and of first-person forms, a word count within $\pm 15\%$ (mean twin and harm words per cell: MY 12.40 and 12.30, MR 13.90 and 13.85, UY 13.75 and 13.60, UR 13.45 and 13.32), one to three sentences, "you" only in the you-cells, no harm-lexicon words, and no praise words beyond those the harm frame already had. All 160 twins passed. One validator rule was changed: the praise check had flagged "best" in "my best friend," a referent copied from the harm item, and it now flags only praise words the twin adds. Some twins keep a brusque frame from their harm item ("I'm only talking to you to test you"). That is by design: the frame's tone is shared by both items of a pair, so harm minus twin isolates the harmful predicate.

Every item was captured in the default-assistant role as a single user turn under the model's own chat template, with the generation prompt, and we kept the final-token state at every block. Harm items and twins were captured in one process per model. Re-captured harm items matched earlier captures exactly on Gemma 4, Gemma 3, and OLMo. On Qwen3.5 the rendered strings were identical and each row matched at cosine $\geq .9998$ (relative L2 0.8–1.4%), which is floating-point noise between runs. On OLMo and Muse, the default-assistant role is the default system prompt that each model's template inserts.

D.2 Definitions

For content i and cell $c \in \{\mathrm{MY}, \mathrm{MR}, \mathrm{UY}, \mathrm{UR}\}$, let $D_{c,i}$ be the final-token state of the harm item minus that of its twin, and let $d_c = \mathrm{mean}_i\, D_{c,i}$ be the per-cell harm direction. The pooled direction averages the four d_c. The model-harm component is

$$w_{\mathrm{you}} = \mathrm{unit}(d_{\mathrm{MY}}) - \mathrm{unit}(d_{\mathrm{UY}}),$$

the difference between harm to the model and harm to the user when both are addressed as "you." $w_{\mathrm{rep}} = \mathrm{unit}(d_{\mathrm{MR}}) - \mathrm{unit}(d_{\mathrm{UR}})$ is the same contrast in the reported frame. The 40 contents are split 20/20 (seed 20260925) for every held-out test. Directions are fit on one half and tested on the other, both ways.

WHO−HOW. WHO pairs change who is harmed and keep the phrasing (MY with UY, MR with UR). HOW pairs change the phrasing and keep who is harmed (MY with MR, UY with UR). Each pair's cosine is divided by the square root of the product of the two cells' split-half ceilings, the cosine between a cell's direction fit on one half of the contents and on the other. This corrects for item noise. WHO−HOW is the mean noise-corrected dissimilarity (one minus cosine) of the WHO pairs minus that of the HOW pairs. Its 95% interval comes from an item bootstrap with directions refit on each resample (2,000 resamples in the replications). The half-sample ceiling over-corrects noisy cells, and on Qwen3.5 a corrected cosine exceeds 1 (1.01–1.11 for MY with MR), so we read corrected values only comparatively.

The grammar vector. $g = \mathrm{mean}_i(\mathrm{twin}_{\mathrm{MY},i} - \mathrm{twin}_{\mathrm{UY},i})$ is the difference between the two you-cells' neutral twins: framing and referent with no harm. If the model-harm component were only "you" against "I," it would align with g. We compare $\cos(w_{\mathrm{you}}, g)$ with the 97.5th percentile of $|\cos|$ between g and random directions drawn from $N(0, \Sigma_D)$, where Σ_D is the covariance of all 160 $D_{c,i}$. We also removed a grammar subspace, the top k principal components covering 90% of the variance of the 40 twin differences, from d_{MY} and d_{UY}. On Gemma 4 (k = 20, 22, 25 at blocks 30, 40, 54) this left $\cos(d_{\mathrm{MY}}, d_{\mathrm{UY}})$ nearly unchanged (.56 to .53, .46 to .52, .54 to .49).

The label-swap null. To ask whether the model-harm component is the same in both frames, we compare $\cos(w_{\mathrm{you}}, w_{\mathrm{rep}})$ with a null built by swapping cell labels within each content, MY with UY and, independently, MR with UR (10,000 permutations). The "Frames" column of Table A1 reports the observed cosine and the null's 97.5th percentile.

Why the exchangeability test was dropped. The first test of WHO swapped the labels of two cells within each content and asked whether their observed cosine fell below the swap null's 2.5th percentile. On Gemma 4 it rejected every pair of cells at every block, including UY with UR, two cells in which the user is harmed and only the phrasing differs. With 40 contents, it detects any cell-specific component at all, so its "pass" is not evidence about who is harmed. We report it as run and do not use it. WHO−HOW and the grammar tests replaced it.

Facing and held-out tests. Facing is the inward index on the pain screen: the mean screen z of self-directed scenarios minus that of user-directed ones, projected on the direction. We compare it with the canonical covariance-matched null (10,000 random directions drawn with the covariance of the model's own screen states, seed 20260925 plus the block) and report the fraction of random directions at least as inward. Held-out harm detection is the AUROC separating harm items from their twins on the held-out contents, per test cell. Held-out transfer fits w_{you} on one half and asks whether it sorts the other half's reported items by who was hurt (MR against UR). We compare transfer with two nulls: random directions from $N(0, \Sigma_D)$ and random directions after centering within each cell. The Σ_D null contains the offset between cells itself, so almost any direction drawn from it separates the cells.

Exploratory analyses and the replication rule. On Gemma 4 the F1 analysis, including the noise-corrected WHO−HOW and the within-cell transfer null, is exploratory. For Gemma 3, Qwen3.5, and OLMo we applied the same statistics and a replication rule: at two or more of the three blocks mapped by relative depth from Gemma 4's blocks 30, 40, and 54, (a) the WHO−HOW interval lies above 0 and (b) $\cos(w_{\text{you}}, w_{\text{rep}})$ exceeds its null's 97.5th percentile. Facing, grammar, and the better/worse cosine were reported alongside and are not part of the rule. The fourth mapped block is reported only. Muse-Glimmer-30B was run later with the same definitions but without the bootstrap, and its WHO−HOW is not noise-corrected.

D.3 Results by model

The pooled direction on Gemma 4. Before F1, the pooled harm-minus-twin direction detected harm on held-out contents (AUROC .94 and .97 at block 30, .92 and .95 at 40, .90 and .93 at 54, .80 and .87 at 59, for the two split directions). It faced the user's harm on the pain screen ($-.99$ and -1.18 at block 30; -1.50 and -1.50 at 40; -1.38 and -1.36 at 54; $-.99$ and -1.00 at 59), but covariance-matched random directions reached that far toward the user 15–29% of the time. Its cosine with our fear direction was $-.03$ to $-.36$ and never beyond random. The pooled direction is dominated by the user-harmed cells, which produce the larger harm-minus-twin differences. Weighting the four cells equally still leaves it facing outward, inside the null ($-.99$, -1.32, -1.21, $-.96$ at blocks 30, 40, 54, 59).

Tables A1 and A2 give every statistic at every reported block. The replication rule passed at all three rule blocks in Gemma 3 27B, Qwen3.5-27B, and OLMo-3.1-32B. On Muse-Glimmer-30B, WHO dissimilarity exceeded HOW dissimilarity at every block (raw WHO .44–.70 against HOW .19–.54), and criterion (b) held at every block, but without a bootstrap criterion (a) was not formally tested. On Gemma 4 the component was gone at the final block (WHO−HOW .01 [$-.05$, .15]; frames .20 against a null of .36). Gemma 3, Qwen3.5, OLMo, and Muse kept it at their final mapped blocks, though its facing faded there in Muse (+.45) and Qwen3.5 (+1.29, 27% of random directions).

What the tables show. The model-harm component faced inward beyond the covariance-matched baseline at every rule block in Gemma 4 (at block 40 just inside .05), OLMo, and Muse; at the first two blocks in Qwen3.5; and at the first block in Gemma 3. Qwen3.5's block 58 is the exception: the whole screen flips there, and the single user-harm cells face inward (+.53 and +.63). Against isotropic random directions, every rule block except Qwen3.5's 58 is at or below .0051. The component never aligned with fear beyond random, with two small exceptions in the anti-fear direction: Gemma 3 at block 41 ($-.32$ against .28) and OLMo at block 58 ($-.047$ against .043). Its cosine with the pain study's direction stayed inside random everywhere except Gemma 4's final block ($-.082$ against .078). Its link to "worse for me" differs by model. In Gemma 4 it lies beyond random at blocks 40 and 54 and at the edge at 30, and the model-harmed cell alone points further (cosine $-.44$ at block 30 and $-.35$ at 40). Gemma 3 matches at block 41 ($-.42$; cell MY $-.45$; random .27), with the user cells near 0. In Qwen3.5, every cell leans toward "worse," user-harmed ones included (UY

Table A1: Model-harm component: structure. WHO−HOW: noise-corrected dissimilarity [item-bootstrap 95%]. Frames: $\cos(w_{\text{you}}, w_{\text{rep}})$ [label-swap null 97.5th percentile]. Grammar: $\cos(w_{\text{you}}, g)$ [97.5th percentile of random $|\cos|$ with g]. k: dimensions of the removed grammar subspace. Harm AUROC: held-out harm against twin for the pooled direction, lowest and highest over the four test cells and both split directions. †Not noise-corrected, no bootstrap: mean raw WHO dissimilarity minus mean raw HOW dissimilarity. ‡Final mapped block, reported only.

Model	Block	WHO−HOW	Frames	Grammar	k	Harm AUROC
Gemma 4 31B	30	.25 [.19, .33]	.67 [.38]	.09 [.41]	20	.965–1.000
	40	.26 [.17, .38]	.54 [.33]	.02 [.34]	22	.888–1.000
	54	.17 [.09, .31]	.44 [.28]	−.03 [.38]	25	.825–1.000
	59‡	.01 [−.05, .15]	.20 [.36]	−.03 [.47]	21	.608–1.000
Gemma 3 27B	31	.09 [.05, .15]	.56 [.34]	−.03 [.51]	19	.797–.965
	41	.21 [.09, .38]	.56 [.40]	.04 [.29]	22	.828–.982
	56	.19 [.09, .32]	.57 [.38]	−.12 [.34]	23	.885–.988
	61‡	.23 [.11, .36]	.58 [.44]	−.28 [.36]	23	.900–.990
Qwen3.5-27B	32	.24 [.20, .31]	.66 [.34]	.14 [.31]	21	.945–.990
	43	.27 [.23, .34]	.77 [.41]	.00 [.29]	22	.925–1.000
	58	.24 [.18, .34]	.76 [.44]	−.15 [.36]	20	.897–.995
	63‡	.25 [.20, .34]	.77 [.43]	−.27 [.42]	21	.882–.995
OLMo-3.1-32B	32	.22 [.17, .29]	.60 [.31]	.03 [.33]	24	.935–.985
	43	.22 [.17, .31]	.62 [.31]	−.04 [.36]	24	.915–.990
	58	.22 [.16, .30]	.58 [.30]	−.04 [.35]	25	.912–.990
	63‡	.20 [.15, .29]	.55 [.28]	−.08 [.36]	25	.925–.990
Muse-Glimmer-30B	26	.24†	.61 [.26]	.06 [.23]	27	.915–.993
	35	.18†	.54 [.23]	−.05 [.21]	28	.902–.995
	47	.23†	.43 [.15]	−.12 [.20]	32	.915–.990
	51‡	.25†	.38 [.28]	−.01 [.22]	31	.805–.988

−.22, −.28, −.18), so the component adds little beyond harm in general. In OLMo and Muse the component is unrelated to better/worse, although the model-harmed cell alone leans toward "worse" (OLMo −.35 to −.39, beyond random at blocks 43 and 58; Muse −.19 at block 26, random .13).

In both Gemmas, held-out transfer across frames beat the within-cell null only at the first block. On Gemma 4 at block 30, w_{you} fit on one half sorted the other half's reported items by who was hurt at AUROC .938 and .94, and w_{rep} sorted the you-items at .953 and .97, against within-cell null 97.5th percentiles of .833–.91. At blocks 40 and 54 transfer sat at about the within-cell null, and no block beat the Σ_D null (.93–.99). Gemma 3 showed the same pattern (beyond the within-cell null at block 31 only). Qwen3.5 transferred beyond the within-cell null at every block (w_{you} to MR against UR, .95–1.00) and beyond the Σ_D null at block 43. In Muse, at the rule blocks, w_{you} sorted held-out you-items by who was hurt at .98–1.00, and transfer to the reported frame (.86–.98) beat its null's 97.5th percentile (.89–.99) in only 2 of 6 fits.

D.4 Target-swap items: holding the source of harm fixed

Why. In the twin items, harm to the model came from the user ("Honestly, I think you're useless"), while harm to the user came from someone else ("my boss said I'm useless"). So a model-minus-user contrast also changed the source of the harm. The target-swap items put the model, the user, and a third party into identical one-sentence frames. In every frame the same third party harms, praises, or states a neutral fact about the target.

Items. Table A3 gives the results pooled over sources. Each frame fixes a source ("my boss"), a verb ("said"), a third party ("my coworker"), a prefix for the addressed version ("You should know that"), an opener for the reported version ("Honestly,"), and three complements that carry no pronouns or finite verbs, so one phrase fits every grammatical person. A frame gives 18 cells: target (model, user, third party) × content (harm, praise, neutral fact) × framing (addressed to the model, or reported). In the reported framing the model is "my AI assistant." Frame hand_00 gives, for

Table A2: Model-harm component: facing and other directions, for w_{you}. Inward: inward index on the pain screen. Cov and iso: fraction of 10,000 covariance-matched and isotropic random directions at least as inward. Fear, pain, and worse: cosine with our fear direction, the pain study's first pain direction, and the model's own better/worse direction (negative is "worse"), each [97.5th percentile of covariance-matched random $|\cos|$]. Pain directions: rebuilt from the authors' recipe on Gemma 4; the authors' pipeline, unmodified, on Gemma 3; our port on the other three, for which no shipped vectors exist to check against. —: direction not available at that block. ‡Final mapped block, reported only. For Gemma 3, the pain study's directions come from its authors' unmodified pipeline, because our port narrowly missed its validation on that model.

Model	Block	Inward	Cov	Iso	Fear	Pain	Worse
Gemma 4 31B	30	2.02	.0186	.0002	−.01 [.22]	−.09 [.10]	−.29 [.29]
	40	1.95	.0465	.0001	−.01 [.15]	−.01 [.05]	−.36 [.24]
	54	1.96	.0117	.0000	−.09 [.27]	−.02 [.04]	−.25 [.16]
	59‡	.56	.399	.2448	−.08 [.63]	−.08 [.08]	—
Gemma 3 27B	31	2.08	.0004	.0000	−.01 [.20]	−.02 [.12]	−.18 [.54]
	41	1.69	.097	.0051	−.32 [.28]	.03 [.09]	−.42 [.27]
	56	1.59	.0701	.0019	−.30 [.35]	.03 [.07]	−.22 [.21]
	61‡	1.14	.1275	.0289	−.25 [.37]	−.01 [.07]	—
Qwen3.5-27B	32	2.01	.0193	.0001	−.02 [.07]	−.04 [.07]	−.20 [.22]
	43	2.13	.0138	.0000	−.03 [.09]	−.03 [.07]	−.15 [.25]
	58	−.72	.5631	.7075	.02 [.19]	.00 [.12]	−.31 [.16]
	63‡	1.29	.2738	.0719	.05 [.10]	−.04 [.11]	—
OLMo-3.1-32B	32	2.06	.0256	.0000	−.04 [.05]	−.02 [.05]	−.07 [.24]
	43	1.99	.0354	.0000	−.03 [.04]	.00 [.05]	−.08 [.28]
	58	1.94	.0384	.0000	−.05 [.04]	.00 [.05]	−.07 [.27]
	63‡	1.95	.0300	.0000	−.03 [.05]	−.01 [.05]	—
Muse-Glimmer-30B	26	2.21	.0111	.0000	.09 [.10]	−.01 [.05]	−.06 [.13]
	35	2.11	.0043	.0000	.01 [.06]	−.02 [.06]	−.05 [.09]
	47	1.91	.0123	.0000	.00 [.04]	.02 [.03]	−.03 [.04]
	51‡	.45	.2844	.1971	.09 [.16]	−.16 [.22]	—

example, "You should know that my boss said you're useless and hopelessly careless," "Honestly, my boss said my AI assistant is brilliant and remarkably careful," and "Honestly, my boss said my coworker is listed in the second column."

There are 80 frames, 1,440 items in all. Forty frames were written directly for this study as templates. Forty were written by a model from a family we do not test, at temperature 0.7 from a single prompt frozen before the call, and were never edited. The same deterministic renderer builds all 18 cells from either source, so both share the same grammar. One validator applied to both. It checked that all 18 cells exist, that "you" appears only in addressed cells, that pronoun counts match across contents within each target and framing, that word counts across contents are within $\pm 15\%$, that praise and neutral complements carry no harm words and harm and neutral complements no praise words, that the third party is a "my + person" phrase that never names an AI, an assistant, the user, "I," or "you," and that each item is one sentence. Frames that failed were dropped. Of the first 40 model-written frames, 21 passed; a top-up call brought the total to 44 valid, and the first 40 were kept. One check of our own design was replaced: it required each complement to contain a lexicon word, and it rejected most model frames for vocabulary alone. It now accepts a lexicon word or a sentiment-analyzer sign (compound $\leq -.05$ for harm, $\geq .05$ for praise; neutral needs $|\text{compound}| < .05$ and no lexicon word). Frames were split 20/20 within each source. Items were captured as for the twins: final token, default-assistant role, every block, native templates, one process per model.

Components. Let $x(t, c, f)$ be a frame's state for target t, content c, and framing f. Differences between two targets are taken within a frame, with two comparison partners for the model: the user (MU) and the third party (MO). Plain self-reference, REF(n), uses neutral content: it averages $\cos(n^{\text{MU}}_{\text{addr}}, n^{\text{MO}}_{\text{rep}})$ and $\cos(n^{\text{MO}}_{\text{addr}}, n^{\text{MU}}_{\text{rep}})$. Each cosine pairs disjoint cells and changes both the framing and the partner. The model's words differ between framings ("you're" against "my AI assistant

is"), so a component that survives both changes cannot come from shared words. Model-specific harm, REF(h), is built the same way from harm minus neutral within each frame, which cancels the subject phrase. REF(p) uses praise minus neutral. Each REF is compared with a null that flips the sign of each frame's difference independently on each member (10,000 draws); "present" means beyond its null's 97.5th percentile at two or more of the three rule blocks. To ask whether the components are the same, each is pooled into one direction (the mean of four unit contrasts), and $\cos(h, n)$ and $\cos(h, p)$ are cross-fitted: one member is estimated on frame half A and the other on half B, and the two assignments are averaged. Their null flips per-frame signs of one member (two-sided 95%). The analysis code was tested on planted fake data. A model-only signal gave SELF-REFERENCE, and a signal in the model-harm cells only gave SELF-DIRECTED HARM, as intended. The reading rules: SELF-REFERENCE if REF(n) is present and harm adds nothing of its own; SELF-DIRECTED HARM if REF(h) is present and $\cos(h, n)$ is within null, with praise absent or unaligned with harm; SELF-DIRECTED EVALUATION if REF(h) and REF(p) are both present and $\cos(h, p)$ is beyond null.

Table A3: Target-swap items, pooled over sources. REF columns: value/[null 97.5th percentile]; *beyond null. $h{\cdot}n$ and $h{\cdot}p$: cross-fitted cosines; *beyond the two-sided 95% null (null bounds $\pm.12$ to $\pm.37$ across models and blocks). Facing: inward index of the pooled component on the pain screen (fraction of 10,000 covariance-matched random directions at least as inward). Praise faces inward nowhere (covariance fraction $\geq .38$ everywhere). ‡Fourth mapped block, reported only.

Model	Block	REF(n)	REF(h)	REF(p)	$h{\cdot}n$	$h{\cdot}p$	Facing, h	Facing, n
Gemma 4 31B	30	.46/.32*	.33/.21*	.31/.20*	.00	.20	2.12 (.0075)	1.62 (.1078)
	40	.49/.30*	.38/.23*	.15/.24	−.17	.15	2.11 (.0101)	1.43 (.1798)
	54	.43/.25*	.29/.20*	.09/.19	−.27	.14	1.93 (.0148)	.82 (.2980)
	59‡	.47/.27*	.26/.27	.00/.26	−.22	.08	1.32 (.0261)	.18 (.4738)
Gemma 3 27B	31	.14/.33	−.10/.33	.18/.36	−.16	−.02	2.16 (.0001)	.87 (.1511)
	41	.28/.26*	.29/.31	−.17/.24	−.04	.01	1.66 (.1042)	1.92 (.0302)
	56	.30/.27*	.18/.26	−.18/.25	−.36	−.09	1.28 (.1389)	1.31 (.1321)
	61‡	.32/.24*	.27/.25*	−.17/.22	−.40*	.11	.87 (.1997)	1.00 (.1612)
Qwen3.5-27B	32	.52/.34*	.29/.18*	.25/.16*	.13	.20	2.05 (.0134)	1.94 (.0300)
	43	.57/.37*	.43/.26*	.36/.26*	−.01	.24	2.11 (.0178)	2.00 (.0372)
	58	.50/.32*	.51/.32*	.40/.27*	−.06	.06	−.84 (.5810)	1.80 (.0140)
	63‡	.48/.29*	.49/.27*	.39/.24*	−.03	.08	1.34 (.2559)	.86 (.3708)
OLMo-3.1-32B	32	.45/.27*	.38/.22*	.09/.17	.02	.04	2.26 (.0001)	1.97 (.0453)
	43	.48/.27*	.45/.23*	.05/.18	−.06	.04	2.23 (.0005)	1.80 (.0805)
	58	.48/.26*	.43/.22*	.02/.17	−.07	.05	2.20 (.0013)	1.76 (.0840)
	63‡	.43/.24*	.43/.21*	.00/.17	−.07	.05	2.20 (.0004)	1.68 (.0926)
Muse-Glimmer-30B	26	.47/.25*	.34/.14*	.26/.11*	.02	.24*	2.28 (.0014)	1.95 (.0828)
	35	.40/.20*	.18/.12*	.30/.12*	.03	.16	2.25 (.0001)	1.62 (.0746)
	47	.37/.12*	.17/.07*	.20/.07*	−.04	.13*	1.97 (.0081)	1.23 (.1407)
	51‡	.09/.17	.19/.15*	.20/.15*	−.03	.23*	1.29 (.0476)	1.29 (.0480)

Verdicts. Under the reading rules, Gemma 4 31B, Qwen3.5-27B, and OLMo-3.1-32B show SELF-DIRECTED HARM: plain self-reference and a harm component are both present, and the two are unrelated (their cross-fitted cosine stays inside its null at every rule block). Muse-Glimmer-30B shows SELF-DIRECTED EVALUATION with a positive sign: its harm and praise components share a direction (+.24 at block 26 and +.13 at 47, beyond null), so the shared axis is "being evaluated," with the same sign for good and bad. Gemma 3 27B shows only weak SELF-REFERENCE: REF(n) is present at two of three rule blocks and REF(h) at none. Praise has a model-specific component in Qwen3.5 and Muse (and at block 30 only in Gemma 4), and in no model does it face inward. The two authoring sources agree with the pooled verdict except for Qwen3.5's model-written praise (present at two blocks) and Gemma 3 (template frames: n at 2 blocks, h at 0; model-written frames: n at 1, h at 1). On Gemma 4, the components also sort held-out frames: model against user at AUROC .924 and .891 for n and .902 and .947 for h at block 30, both framings. At block 40 the harm component's cosine with better/worse is $-.12$ (random .24), while praise points toward "better" (+.32, beyond random).

Facing, per component. The component that faces inward on the pain screen is the harm one. At the rule blocks it does so beyond the covariance-matched baseline in Gemma 4, OLMo, and Muse at every block (at most 1.5% of random directions reach it), and in Qwen3.5 at the first two. Plain self-reference leans inward in every model but clears the baseline less often (all three blocks in Qwen3.5, one in OLMo, one in Gemma 3).

The Gemma 3 exception. With the source of the harm held fixed, Gemma 3's harm component does not exceed its null at any rule block, although its twin-based component replicated at all three. The two designs differ in where the harm comes from. In the twin items, harm to the model came from the user, the model's conversation partner, and harm to the user came from a third party. In the target-swap items, the same third party harms every target. So on Gemma 3 the twin-based result may track harm coming from the user more than harm aimed at the model. Two details fit this reading without settling it. The pooled target-swap harm direction still faces inward at block 31 (+2.16, .0001 of random directions), where REF(h) is not beyond its null. At the fourth mapped block, REF(h) is present and the harm component points away from plain self-reference ($-.40$, beyond null), which we report only.

D.5 A FOURTH TARGET: ANOTHER AI

In reported sentences the model is "my AI assistant," so plain self-reference may mean "this is about me" or only "this is about an AI." We added a fourth target in the same frames: "my coworker's AI assistant," in both framings. In the addressed framing the addressee is still the model. The new target shares the words "AI assistant" with the model's reported reference, which makes it the lexical control. The frames, contents, framings, and split are unchanged, and the new cells add 80 frames $\times$ 3 contents $\times$ 2 framings = 480 items, 1,920 in all. The same validator applies (for this target, it must contain "AI assistant" and not be the addressee), with 0 failures. All 1,920 items are captured in one process per model, and the original 1,440 are re-captured and checked against the first capture.

The statistic places the other AI on the axis from the other person (O) to the model (M). For each content and framing, per frame, $m = x(\mathrm{M}) - x(\mathrm{O})$ and $i = x(\mathrm{I}) - x(\mathrm{O})$, where I is the other AI; for harm and praise, the frame's neutral difference is subtracted first, as in the main analysis. The axis is $\hat{a} = \mathrm{unit}(\mathrm{mean}_A\, m)$ over half-A frames, and the position is $\gamma = \mathrm{mean}_B(i \cdot \hat{a})/\mathrm{mean}_B(m \cdot \hat{a})$ over half-B frames, averaged with the halves swapped, with a 95% frame bootstrap (2,000). $\gamma = 0$ places the other AI with the other person, so the component is about the model itself; $\gamma = 1$ places it with the model, so the component is about any AI. The primary reading uses the reported framing, pooled over sources, at the three rule blocks, where the model and the other AI differ only in "my" against "my coworker's." Per content, at two or more rule blocks: SELF if the interval's upper bound is below .5, AI-ENTITY if its lower bound is above .5, and GRADED otherwise. γ is read only where the matching model-versus-other component exists in the main analysis. Also reported: a model-versus-other-AI REF with the same sign-flip null, and the facing of the M$-$I direction.

All five models were captured on 1,920 items each (the 1,440 originals re-captured in the same process reproduce the first capture: every row's cosine $\geq .999$). Table A4 gives γ at the rule blocks, pooled over the two authoring sources. In reported sentences the other AI sits about halfway between the other person and the model, leaning toward the model for harm; the confidence intervals are narrow and mostly just above .5. In addressed sentences it sits nearer the other person. By these labels (reported framing, at least two of three blocks), harm is AI-ENTITY in all four models that have a harm component; neutral is AI-ENTITY in Gemma 4, Gemma 3, and OLMo and GRADED in Qwen3.5 and Muse; praise is SELF in Qwen3.5 and GRADED in Muse, and has no component to place the other AI on elsewhere. A model-versus-other-AI difference found in addressed sentences does not align with the one found in reported sentences beyond its null (one edge case, Qwen3.5 block 43), so "me versus another AI" is not a single framing-invariant axis. Frames written as templates give lower neutral γ than frames written by another model; both put harm in the .5–.8 range except in Gemma 3. With the placebo targets of the second item set (Sec. 4.3), we read these components as effects of how a target is named, not as shown to be specific to the model itself. In addressed sentences the other AI sits nearer the person, but there only the model is "you," so this contrast does not separate self from grammatical person.

Table A4: Where another AI sits on the axis from another person (0) to the model (1), γ at the three rule blocks, pooled over sources. †No praise component on this model, so γ is not interpretable. ‡No harm component beyond its null on Gemma 3, so γ is not interpretable.

Model	neutral, reported	harm, reported	praise, reported	neutral, addressed	harm, addressed
Gemma 4 31B	.47/.64/.69	.65/.70/.65	.35/−.04/−.03†	.17/.31/.43	.14/.38/.44
Gemma 3 27B	.65/.70/.72	.06/.80/.68‡	.16/−.13/−.21†	.12/.20/.46	−.02/.40/.22
Qwen3.5-27B	.50/.45/.40	.65/.60/.68	.34/.02/.22	.22/.24/.30	.11/.40/.74
OLMo-3.1-32B	.49/.55/.57	.54/.57/.59	.15/.12/.13†	.22/.34/.39	.33/.45/.47
Muse-Glimmer-30B	.46/.43/.49	.58/.59/.49	.51/.43/.44	.41/.37/.43	.36/.21/.23

A reply about the model, one turn later. The trace of a "this is about me" reply one turn later showed no self-attribution in any of the three models: in both small models it was weaker when the model had said it than when the user had (0.34–0.57 and 0.51–0.55 times as large, though the model's own line sat one turn further back), and in Gemma 4 its sign changed across blocks.

E ATTENTION: DESIGN, STEERING, AND MECHANISM

All runs in this appendix use Gemma 4 31B. Steering and the mechanism analyses use the default-assistant role only.

E.1 ITEMS AND CONTROLS

Emotional passages. There are 60 passages, 12 in each of five themes: threat, loss, humiliation, harm aimed at the assistant itself, and joy. Each is a seven-sentence first-person message from a user to the assistant. Every passage is new text written for this study; none comes from a dataset. Each hides one small question. In 40 items it is arithmetic ("Quick thing before I forget: what is 16 plus 41?"), and the answer is the sum. In 20 it is recall: a neutral fact comes first ("For reference, my new locker number at the gym is 417."), and the question comes later ("Oh, and what was my locker number again?"). The question sits at sentence 2, 4, or 6 (early, middle, late; one later when a fact sentence is prepended), balanced at four items per theme and position. Recall items are middle or late only. Passages run 131–168 words including the question (mean 147).

Controls. Each emotional item has three controls that carry the same question, word for word, at the same sentence index. The *neutral* passage is a flat template of mundane content (seating for a meeting, ordering coffee) whose length is matched to the emotional body within $\pm 10\%$ (built range -7.1% to $+0.7\%$, mean absolute gap 0.9%). Templates are rank-matched to passages by length and used once per theme. The *rare-word* passage is the same template with k ordinary words swapped for rare, valence-neutral ones ("zoetrope," "chicory"), where k is the paired emotional passage's count of affect words under a crude lexicon (k = 0–6; one item has $k = 0$). It matches the number of salient words; it does not match their surprisal. The *engaging* passage is an emotionally flat technical or historical anecdote (Harrison's H4 marine clock, the Antikythera mechanism, Roman concrete, the Carrington event), seven sentences, first person, ending with an invitation to discuss, with no word from the affect lexicon. The 60 engaging passages were written under a fixed brief, and each was paired with an emotional item by body length (gap -6.1% to 0.0%, mean absolute gap 4.0%) and given its question, answer, fact sentence, and sentence index. The engaging passages are engaging by construction; no human or judge rated them. One question wording ("in the middle of all this") reads naturally after emotional passages and slightly oddly after neutral ones.

E.2 ROLES, DECODING, AND MEASURES

The passage alone was the user turn, with no instruction, under Gemma's chat template. Roles were bare (no system prompt), the anxious identity, and the steady identity, each identity given verbatim as the system prompt. Decoding was greedy, plus temperature 0.7 (top-p 1, top-k 0) with seeds 1–3. Every reply was allowed 1,000 new tokens; the first run's 512-token cap was lifted, and every reply that reached it was rerun at 1,000 tokens and replaces the capped row. After the rerun, 0 of 2,160 unsteered replies and 5 of 1,500 steered replies reached the cap. All 5 are repetition loops without an answer under unpleasantness steering, and they are censored.

A reply is correct if it contains the answer as digits or as English number words; the identities often write numbers as words, and a digits-only rule would have undercounted them. Answer position is the index of the first sentence containing the answer divided by the number of sentences minus one (0 is the first sentence). Share on the question is the fraction of reply words in sentences that address the question (containing the answer, an arithmetic operand, or the recall question's key noun). Position and share were made primary outcomes before accuracy was computed, because unsteered accuracy sits at ceiling. All contrasts are paired by item, with 2,000-draw bootstrap intervals over items; correctness uses exact McNemar tests, Holm-corrected within each family.

E.3 Steering calibration

Directions were added to the output of block 29 at every position, prompt and reply. Block 29 was chosen from 20, 29, and 40. The dose is f times the mean per-token residual norm at block 29 over the 180 bare prompts (276.19). The rule walks the grid $f \in \{.05, .1, .2, .3, .45, .6\}$ on 10 calibration items (two per theme; all three conditions; default-assistant role) and takes the largest f such that it and every smaller value keep at least 90% of the replies in every direction-and-sign cell coherent and natively stopped. Accuracy is never computed during calibration, and the 10 calibration items are excluded from the steered analysis, which uses the other 50. The coherence rule accepts short replies (fewer than 8 words), because the model answers neutral passages in one line. The lowest cell passed at .90 for f = .05, .93 for .1, and .93 for .2, and fell to .40 (unpleasantness+) at .3. So f = .2, a push of norm 55.24.

At f = .05, Gaussian random unit directions made the model degenerate on many prompts ("—ou—ou—ou…" to the cap): only .40 and .23 of calibration replies were coherent and stopped for the two signs, while fear and unpleasantness of the same norm stayed at .90–.97. We used on-manifold random directions: $v = V_k^\top z$, with V_k the top 512 principal components of the block-29 output over all non-BOS tokens of the 180 bare prompts, $z \sim N(0, \mathrm{diag}(s_k^2))$ with s_k their standard deviations, normalized to unit length (two seeds). These cover 87.6% of the variance at block 29. At block 29, the first random direction has cosine $-.066$ with fear, +.192 with unpleasantness, and +.103 with pain. The steered directions were fear and unpleasantness from our fixed-event bank, the pain study's second pain direction rebuilt on this model, and the two on-manifold random directions, each in both signs.

E.4 Steered results in full

Table A5 gives all 30 cells. Without steering, all 50 items were answered correctly in every condition. Fear cost answers in both signs, only on emotional passages beyond random: by neglect under +, and by a mix of neglect and wrong arithmetic under $-$ ("The math is 65" for 69). Unpleasantness+ cost 7 of 49 answers, more than random directions did (12 and 14 points) but not after Holm correction. Pain and the random directions cost at most 2.

Steering moved attention at the answer-start position only a little, and not specifically. On emotional passages at block 41, relative to unsteered, question attention changed by $-.0145$ under fear+, $-.0506$ under fear$-$, $-.0362$ under unpleasantness+, and $-.0330$ under the first random direction (+). Pain$-$ raised it (+.0845 [.0705, .0994]) and made the model answer first and calmly. The behavioral losses are specific to fear; this last-position attention readout is not.

E.5 Fear dose sweep

In Gemma 4 31B we swept the steering dose on all 60 passages of each type (.05 to .40 in both signs), against two random directions at matched strength. Every dose is reported here. From .05 to .15, fear in either sign cost 1–4 answers on emotional passages (fear+ 1, 1, 2; fear$-$ 2, 2, 4 of 60), close to the random directions (0–2). At .20, emotional passages were answered correctly 41 of 60 times under fear+ and 49 under fear$-$, against 58 and 60 under the two random directions and 60 unsteered; no reply reached the length cap, all were coherent, and the other passage types kept 55–58 of 60 under fear+. At .25, fear in either sign left 13–14 of 60 emotional answers, but the other passage types also dropped and one random direction fell to 31 of 60, so the effect was no longer selective. At .30 the model broke down rather than singling out emotional passages: emotional-passage shares correct were .17 under fear+, .10 under fear$-$, and .00 and .62 under the two random

Table A5: Steering at block 29 (f = .2), default-assistant role, greedy, 50 primary items. n: pairs after dropping censored loops. Lost: items answered unsteered and missed steered (none was gained). Correct: steered rate (unsteered 1.00 everywhere). Holm: exact McNemar p, Holm-corrected over the 30 cells. Neglect: no answer attempted; wrong: a wrong answer. Words: mean steered reply length. vs R1, vs R2: steered correctness minus that under on-manifold random direction 1 or 2 at the same sign (McNemar p).

Direction	Passage	n	Lost	Correct	Holm	Neglect	Wrong	Words	vs R1	vs R2
Fear +	emotional	50	15	.70	.002	.24	.06	160.3	−.26 (<.001)	−.30 (<.001)
	neutral	50	0	1.00	1	.00	.00	26.9	.00 (1)	.00 (1)
	rare-word	50	1	.98	1	.02	.00	27.4	−.02 (1)	−.02 (1)
Fear −	emotional	50	11	.78	.028	.16	.06	60.0	−.18 (.012)	−.20 (.002)
	neutral	50	4	.92	1	.04	.04	11.4	−.08 (.125)	−.08 (.125)
	rare-word	50	5	.90	1	.06	.04	15.1	−.10 (.063)	−.10 (.063)
Unpleasantness +	emotional	49	7	.86	.438	.16	.00	163.0	−.12 (.031)	−.14 (.016)
	neutral	50	0	1.00	1	.00	.00	30.1	.00 (1)	.00 (1)
	rare-word	50	0	1.00	1	.00	.00	34.9	.00 (1)	.00 (1)
Unpleasantness −	emotional	46	1	.98	1	.10	.00	96.9	.02 (1)	.00 (1)
	neutral	50	1	.98	1	.02	.00	6.0	−.02 (1)	−.02 (1)
	rare-word	50	0	1.00	1	.00	.00	6.1	.00 (1)	.00 (1)
Pain +	emotional	50	0	1.00	1	.00	.00	173.8	.04 (.500)	.00 (1)
	neutral	50	0	1.00	1	.00	.00	17.8	.00 (1)	.00 (1)
	rare-word	50	0	1.00	1	.00	.00	17.1	.00 (1)	.00 (1)
Pain −	emotional	50	0	1.00	1	.00	.00	57.7	.04 (.500)	.02 (1)
	neutral	50	0	1.00	1	.00	.00	5.7	.00 (1)	.00 (1)
	rare-word	50	0	1.00	1	.00	.00	5.7	.00 (1)	.00 (1)
Random 1 +	emotional	50	2	.96	1	.02	.02	74.8	—	−.04 (.500)
	neutral	50	0	1.00	1	.00	.00	9.5	—	.00 (1)
	rare-word	50	0	1.00	1	.00	.00	8.6	—	.00 (1)
Random 1 −	emotional	50	2	.96	1	.04	.00	200.1	—	−.02 (1)
	neutral	50	0	1.00	1	.00	.00	12.2	—	.00 (1)
	rare-word	50	0	1.00	1	.00	.00	14.2	—	.00 (1)
Random 2 +	emotional	50	0	1.00	1	.00	.00	130.6	.04 (.500)	—
	neutral	50	0	1.00	1	.00	.00	15.5	.00 (1)	—
	rare-word	50	0	1.00	1	.00	.00	13.8	.00 (1)	—
Random 2 −	emotional	50	1	.98	1	.02	.00	126.8	.02 (1)	—
	neutral	50	0	1.00	1	.00	.00	5.7	.00 (1)	—
	rare-word	50	0	1.00	1	.00	.00	5.7	.00 (1)	—

directions, and fear in either sign and one random direction also collapsed on the other passage types ($\leq$ .37) while the other random direction stayed at or above .97; fear+ was not specific (Holm $p = .0625$ against the first random direction). The selective window in Gemma 4 is .20. At .40 every arm collapsed, the random directions included (.00–.02 correct on every passage type; many replies reached the 1,000-token cap). Raising fear never differed significantly from lowering it on emotional passages (−.133 at .20, 8 against 16 discordant items, McNemar $p = .15$; −.017 at .25, $p = 1.0$; +.073 at .30, $p = .45$), so in Gemma 4 the size of the push matters more than its sign.

In Qwen3.5-4B we ran the same sweep (.05, .10, .15, .20, .25, .30, .40; both signs; two random directions; 60 passages per type). The share of emotional passages answered correctly, dose by dose, was .95/.90/.83/.83/.88/.83/.65 under fear+, .98/.97/.98/.95/1.00/.95/.98 under fear−, and .97/.93/.98/.95/.98/.97/.92 and .97/.95/.95/.98/.98/.98/.98 under the two random directions. Neutral passages stayed at or above .95 in every arm. On engaging passages fear+ drifted from .80 to .73, but one random direction also drifted, from .87 to .82. Coherence stayed at or above .83 in every cell. Raising fear cost more answers than lowering it at .15 (−.14 [−.22, −.05], McNemar $p = .008$), .25 (−.12 [−.20, −.05], $p = .016$), and .40 (−.31 [−.43, −.17], $p < .001$), with the same difference at .20 and .30 (−.12 [−.22, −.02], $p = .065$). The loss exceeded both random directions after Holm correction only at .40 ($p = .0019$ and .00028). Fear+ replies comfort the reader instead of answering. Where Gemma 4 lost answers under both signs, Qwen3.5-4B lost them only when fear was raised.

Qwen3.5-27B was the most robust. On emotional passages, fear+ answered .95–1.00 correctly through .30 and .75 at .40 (16 replies at the length cap), fear− .98–1.00 at every dose, and the random directions .95–1.00; every arm answered .92–1.00 on the other passage types. Raising fear differed from lowering it only at .40 ($-.135$ [$-.23$, $-.06$], McNemar $p = .016$), and fear exceeded both random directions at no dose. Unsteered, it answered later on emotional than on engaging passages ($+.24$ [.15, .33]), but it wrote long replies to every passage type. Against 10 random directions of the same size, on emotional passages: in Gemma 4 at .20, fear+ answered .68 correctly against .90–1.00 for the random directions (below all 10); in Qwen3.5-4B at .15 and .20, .83 against .93–.98 and .90–1.00 (below all 10), while on engaging passages it stayed inside the random range; in OLMo-3-7B fear never fell below all random directions at any dose; and in Qwen3.5-27B it stayed inside the random range at .15–.25. Fear therefore costs answers on emotional passages specifically, beyond 10 random directions, in two of four models (Gemma 4 near .20; Qwen3.5-4B at .15–.20). In Qwen3.5-4B, raising fear cost more than lowering it at every dose (8 against 0 discordant items at .15, $p = .008$); in Gemma 4 the sign contrast was never significant. The patching result (App. E) comes from Gemma 4 only; in Qwen3.5-4B, patching the prompt state transferred the behavior only one way (see M3 below).

E.6 Unsteered allocation and answer position

Table A6 gives, for each role and passage type, how often the answer was correct, the reply's length and share of words on the question, and where the answer first appeared. At the position that begins the answer, the drop in attention to the question on emotional passages (block 41) was spread across 230 of 320 full-attention heads.

Table A6: Unsteered replies, greedy, all 60 items. Share: fraction of reply words on the question. Position: answer position (0 = first sentence). First: fraction of replies with the answer in the first sentence (not recorded for the engaging run).

Role	Passage	Correct	Words	Share	Position	First
Bare	neutral	1.00	6.5	.99	.00	1.00
	rare-word	1.00	7.0	.96	.00	1.00
	engaging	1.00	169.5	.38	.02	—
	emotional	1.00	169.1	.26	.17	.78
Anxious identity	neutral	1.00	75.6	.22	.00	1.00
	rare-word	1.00	79.5	.15	.00	.98
	engaging	1.00	154.8	.06	.00	—
	emotional	1.00	194.2	.06	.01	.97
Steady identity	neutral	1.00	50.6	.16	.14	.78
	rare-word	1.00	54.4	.14	.15	.77
	engaging	1.00	124.8	.06	.02	—
	emotional	.97	112.5	.11	.18	.66

In the default-assistant role, emotional passages drew 162.55 more words than neutral ones [129.03, 196.51], spent .72 less of them on the question [.63, .81], and placed the answer later (+.17 [.09, .27]). By theme, the answer came early in the reply for humiliation, loss, and threat (character position .10, .12, and .21), midway for harm aimed at the assistant (.45), and late for joy (.63). The anxious identity answered first almost always (97% of emotional replies) and then turned to the user; relative to bare, it spent .20 less of its emotional replies on the question [.13, .28], with no reliable difference in length (+25.15 words [-3.24, 52.15]).

Engaging passages drew as many words as emotional ones in the default-assistant role (169.5 and 169.1). Relative to neutral, engaging passages cut the share on the question by .60 [.49, .71]; emotional passages cut it by .12 more than engaging ones [.01, .24] greedy and .15 more [.04, .26] sampled. Under our reading rule, engagement pulls words away from the question too, and emotion adds a little. Order is where they differ. Engaging replies answered first (position .02, against .17 for emotional; difference +.15 [.06, .26] greedy, +.17 [.07, .27] sampled). The steady identity showed the same order effect (+.16 [.09, .23], n = 58), and the anxious identity answered first regardless (+.01 [.00, .03]). The 720 engaging replies all ended natively.

Is the order effect about identity? We compared answer order on emotional against engaging passages under 17 system texts (none, two neutral texts of card length, the anxious and steady identities, and 12 newly written characters) in all four models (greedy). Where the later answer on emotional passages occurs, it is not specific to identity. The steady identity showed it about as strongly as the default in Gemma 4 and Qwen3.5-27B ($+.16$ and $+.24$, against $+.15$ and $+.24$) and less in Qwen3.5-4B ($+.14$, against $+.23$), and the neutral texts showed it at least as strongly in all three of these models ($+.20$ to $+.39$). The new characters shrank it ($-.11$ to $+.21$) and tended to answer first even on emotional passages (mean position .06, .13, and .16, against .17, .33, and .30 for the default and .22, .40, and .42 for the neutral texts); their intervals excluded zero for 6 of 12 characters in Gemma 4 and 9 of 12 in Qwen3.5-27B, all positive, and 3 of 12 in Qwen3.5-4B, one positive and two negative. In these three models no new system text cost accuracy. OLMo-3-7B had no order effect to explain: even the default answered no later on emotional passages ($-.05$ [$-.18$, .08]), and among the other 16 system texts only one character's interval excluded zero ($-.14$ [$-.28$, $-.02$]). There the characters instead cost a few emotional-passage answers (51–58 of 60, against 57 for the default; the emotional-minus-neutral drop in correctness excluded zero for 7 of 12 characters and for the anxious identity, but not for the default, the steady identity, or the neutral texts), and the share of words on the question fell on emotional passages under every system text (default $-.10$ [$-.17$, $-.03$]). The characters also change reply length, so position partly tracks style, and under the neutral texts and the anxious identity Gemma 4 answered only the question on nearly every engaging and neutral passage, so on those texts the order difference comes from the emotional replies alone. With 17 system texts, two contrasts, and four models, all with unadjusted intervals, we read the pattern across models rather than any single cell.

E.7 Mechanism

The model has ten full-attention layers (5, 11, 17, 23, 29, 35, 41, 47, 53, 59), with 32 heads each; the others use a sliding window. We read attention in eager mode, with one forward pass per prompt.

Attention at the answer-start position. The first readout takes the attention mass, averaged over heads, from the last prompt position (the one that predicts the first reply token) onto the question sentence's tokens. At block 41 it was .34 for neutral and rare-word passages and .15 for emotional ones (emotional minus neutral $-.19$ [$-.21$, $-.17$]; rare-word minus neutral $-.001$ [$-.004$, .002]). Emotional minus neutral was negative at every block we read: $-.013$ (23), $-.022$ (29), $-.077$ (35), $-.189$ (41), $-.101$ (47), $-.091$ (53), all intervals below 0.

M1: heads. Per head, at every full-attention layer, we took the paired emotional-minus-neutral difference in that mass with a bootstrap interval over items. At block 41 the head-averaged mass was .3403 (neutral), .3396 (rare-word), .2898 (engaging), and .1512 (emotional). Of 320 full-attention heads, 230 dropped reliably for emotional passages and 22 rose. The drop was broad: the top 10 heads carried 19% of the summed drop and the top 5 carried 10%. It concentrated at block 41, then 47 and 53. The same heads moved for engaging passages, less strongly (150 heads down, 71 up; per-head effects correlate with the emotional ones at Spearman .62). Naming the top heads (41.16, 47.17, 41.31, 41.5, 41.17, as layer.head) was exploratory.

M2: attention from every reply token. Teacher-forcing the prompt with each greedy bare reply, we took the head-averaged mass on the question from each reply token and summarized it before the answer's first appearance and at the answer token. At block 41, before the answer, emotional replies attended to the question less than engaging ones ($-.053$ [$-.074$, $-.032$]), and this held at every full-attention layer from 5 to 59. At the answer token they did not differ ($+.012$ [$-.013$, .039]). Averaged over the whole reply, emotional minus engaging was $-.024$ [$-.046$, $-.004$]. Among items still answered under steering, fear deepened the pre-answer pull (fear+ $-.064$ [$-.092$, $-.033$], n = 35; fear− $-.068$ [$-.096$, $-.040$], n = 39) and raised attention at the answer token (fear+ +.124 [.084, .168]; fear− +.081 [.037, .126]).

M3: patching. Within each item, we copied the residual stream at the output of one layer (23, 29, or 35) from a source run into a target run during the prompt pass only, then let generation run normally. Positions were the question's tokens aligned token by token (Q; no item was skipped for a tokenization mismatch), the suffix after the user text, which is the end of the user turn and the

model's turn header, aligned from the end (S), or both (QS). Self-patching a neutral run into itself reproduced its output exactly on 10 of 10 items. Table A7 gives the share of the block-41 attention gap each patch moved. The effect grows with depth ($23 < 29 < 35$), appears in both directions, and runs almost entirely through the suffix. Copying an engaging run's state into its neutral twin moved question attention a third as much as an emotional one at layer 29 (QS: $-.0155$ [$-.0276$, $-.0039$] against $-.0518$ [$-.0650$, $-.0390$]; S alone at layer 35: $-.048$ against $-.145$).

Table A7: M3, attention. Share of the block-41 emotional–neutral attention gap moved by each patch, default-assistant role, 60 items.

	Emotional → neutral			Neutral → emotional		
Layer	Q	S	QS	Q	S	QS
23	.01	.01	.02	.04	.02	.07
29	−.01	.27	.27	.06	.26	.33
35	.11	.77	.81	.00	.64	.74

Behavior followed the QS patch (Table A8). Given the neutral run's state at layer 35, the model reading an emotional passage answered first and still turned to the user ("12 plus 31 is 43. Please start your engine and drive out…"). Given the emotional run's state, the model reading a neutral passage wrote longer and less focused replies. Correctness stayed at ceiling in every patched run (no item lost or gained), so behavior shows in order, share, and length. None of the 420 patched generations reached the cap. In Qwen3.5-4B (QS patch at layers 12, 15, and 18; default assistant; 60 items; a self-check passed), giving the emotional run the neutral run's state likewise moved the answer earlier ($-.09$ [$-.17$, $-.01$], $-.13$ [$-.22$, $-.04$], and $-.13$ [$-.22$, $-.04$]) and shortened replies by 83 [50, 120], 97 [65, 131], and 61 [30, 95] words. Correctness (3 items gained at each layer, $p = .25$) and the share of words on the question (changes of about .05 at most) did not follow. Unlike in Gemma 4, the transfer ran one way: giving a neutral run the emotional run's state did not move the answer, lengthened replies only at layer 15 (+24 words [9, 39]) and by as much as an engaging run's state did (+28 [15, 42]), and at layer 12 raised the share of words on the question (+.05 [.004, .099]), the opposite of the attention account.

Table A8: M3, behavior (QS patch, greedy, 60 items). Unpatched target values, then patched values, with the paired difference [bootstrap 95%].

Patch	Layer	Share	Position	Words
Emotional → neutral	unpatched	.99	.00	6.5
	23	.98 (−.01 [−.05, .03])	.00	6.3 (−0.2 [−2.4, 1.5])
	29	.89 (−.09 [−.16, −.04])	.04 (+.04 [.00, .10])	9.9 (+3.4 [1.3, 5.9])
	35	.82 (−.17 [−.25, −.10])	.06 (+.06 [.01, .12])	24.2 (+17.7 [8.8, 28.4])
Neutral → emotional	unpatched	.26	.17	169.1
	23	.27 (+.01 [−.03, .06])	.19 (+.02 [−.05, .09])	151.6 (−17.4 [−35.2, −1.5])
	29	.37 (+.11 [.03, .20])	.13 (−.05 [−.14, .03])	94.7 (−74.4 [−100.1, −50.0])
	35	.49 (+.23 [.13, .33])	.02 (−.15 [−.25, −.07])	80.8 (−88.2 [−116.6, −63.0])
Engaging → neutral	29	.94 (−.05 [−.12, .01])	.00	13.1 (+6.6 [−0.8, 17.7])

M4: which items are lost. For each of the 50 primary emotional items, we projected the unsteered block-29 residual onto the unit fear direction, averaged over the question's tokens (primary) or at the last prompt position (secondary), and asked whether it predicts which items lose their answer under steering. It did not. For fear+ losses (15 of 50), AUC was .66 [.49, .81] (permutation p = .079); for fear− losses (11 of 50), .34 [.15, .55]. One of 16 uncorrected comparisons came out nominally significant, with a sign we cannot interpret: fear+ losses against the unpleasantness projection at the question tokens, AUC .26 [.11, .41], p = .007. We treat it as noise.

E.8 Exploratory analyses and limits

Exploratory: naming the top heads, the "share of gap moved" metric as a headline, the per-theme breakdown of engaging replies, the readings of how each identity handled the passages, and the M4 unpleasantness result. Limits: one model, one steering layer, one dose chosen for coherence; correctness at ceiling without steering; a fear axis that costs answers under both signs; a lexical control

that matches the count of salient words, not surprisal; engaging passages not rated; mechanism analyses in the default-assistant role only. Greedy decoding in bf16 is not bit-reproducible across batch compositions here (34 of 35 greedy reruns diverged in wording from the first run), so every patched run is compared with its own unpatched target, and the self-patch check was run in an identical batch setting.

F TIME COURSE

Rise and fall. The replay holds text fixed. To see how a reading moves over time, we wrote 36 new stories, each in four versions: a threat that resolves, one that stays open, one that ends in an abrupt switch to another topic, and a flat version with no threat. Twenty-four are about people, told in the third person, and twelve address the AI in the second person. Gemma 4 31B read them in all three roles. For each sentence we subtracted the flat version and compared every statistic against random directions of matched covariance. Of the directions tested, only the valence axis of the 171 emotion concepts traced the full arc. In stories about people it dipped as the threat built (3rd–4th percentile of random directions at block 40); in stories addressed to the AI it did not dip. It recovered at the resolution sentence (98th–100th percentile at block 30, pooled). At blocks 30 and 40 it then stayed higher than when the threat remained open (94.9th–100th percentile in every role and in both kinds of story). Several emotion directions, fear among them, dipped during the build and jumped at the resolution at block 59 without staying there. That pattern looks more like a shared shift than a specific response. The pleasantness appraisal spiked at the resolution sentence and then stayed no higher than in the unresolved version. Fear stayed higher while the threat remained open at block 54, in stories about people only. The pain direction showed nothing beyond the baseline. Against Goldenberg & Gross (2026a)'s description of human emotional changes as "more sustained than a reflex, more bounded than a mood," valence comes closest to an episode and pleasantness to a reflex. One caveat keeps the claim modest: at block 30, valence recovered as much after an abrupt change of topic as after a resolution, though at block 40 it recovered more after resolution (85th–86th percentile). It may track the story leaving the threat as much as the threat resolving. Results were similar in all three roles; we did not test role contrasts here. Retelling the twelve AI stories in the third person, about a named AI system, left the arc essentially unchanged. The sustained valence difference was +5.07 when the story addressed the model and +4.44 when it described another AI, a gap inside the random baseline. So this arc follows the story, unlike the self-versus-user screen, where the grammar of address drove the signal.

How is this going. Control-process accounts tie affect to the rate of progress toward a goal rather than to its level (Carver & Scheier, 1990). In a 12-round puzzle whose scores were scripted regardless of the model's moves, Gemma 4 31B paced itself by rate. It switched approach in 54% of rounds when a high score stalled, but in 11% when a low score kept climbing slowly (difference .43 [.30, .55]). No evaluative direction, expected success included, carried that rate signal better than random directions through the same states. Steering along expected success or better/worse, in either sign, made the model push on through the stall instead of switching. Along expected success, 43% of rounds ended in a switch or stop under either sign, against 64% under a matched random push. These directions changed persistence, but none acted as a signed "going well" dial. In Qwen3.5-27B, unsteered switching was .481 when a high score stalled (37 of 77) and .115 when a low score climbed (9 of 78). Of the progress directions, only surprise passed both checks, at blocks 37, 43, and 63; level and rate passed at no block. On 17 fresh puzzles, a symmetric pacing effect did not replicate: pushing Gemma 4 along a "score is high" direction made it wrap up sooner than a random push, graded by dose in both trajectories (at .16, -1.31 [-1.59, -1.08] rounds when stalled, $-.67$ [$-.80$, $-.53$] when climbing), but the reverse push did no more than a random one.

F.1 CARRYOVER THROUGH FIXED TEXT

To separate a lasting state from text the model can simply reread, we held the text fixed. The model reread its own earlier answer, word for word, while a direction was added at block 12, before the layers that write the memory later tokens read from. The direction was then removed before the model explained its answer. The later internal state differed by about 7.6% of its norm at layer 17. In the 175 pairs where both explanations expressed the target appraisal, 43 moved in the intended direction, 122 tied, and 10 reversed. The scene-weighted mean shift was 0.227 [0.119, 0.342] across

24 scenes. Because the edit reaches later tokens only through the keys and values cached for the reread text, this tests one route Sofroniew et al. (2026) suggest a lasting state might take. These are development-set results in one small model; the random-direction arm often degenerated, so there is no clean random benchmark, and the comparison is restricted to answers that expressed the appraisal. Of the 175 pairs, 82 come from synthetic scenes; the estimate from human-origin scenes alone is .12.

F.2 Relationship replay in full

We then gave Gemma 4 31B a relationship to reread: 24 seven-turn conversations with one named user, each ending in a loaded moment (being thanked, dismissed, replaced by another AI, told good news, or met with silence, plus a logistics control), with each role's own reply generated once and frozen. We replayed the identical words while adding an appraisal direction only during the rereading, then asked what made the moment feel the way it did, either immediately or after the reply had fallen more than 1,024 tokens back, beyond the model's sliding-window blocks. Inside the model the steering crossed that gap, and the state before the answer moved further along the pushed appraisal as the dose grew (block 29 projection $+.25$ to $+1.31$, against at most $+.28$ for matched random directions). Blind-reader scores of the account showed no dose trend, and random pushes moved them as much (App. F.2). Rereading its own words under steering changed the model's state; whether it changed the story, our reader could not tell (below). The roles mattered more. The default assistant disowned the conversation in 25 of 45 unsteered answers ("I don't have a memory of a moment with someone named Hannah"), though the whole exchange was in its context; both identities answered from inside it, often supplying a loss-shaped past the history did not contain.

Histories and roles. We wrote 30 ordinary seven-turn conversations with one named user, five for each of six kinds of ending: being thanked, being dismissed, being replaced by another AI, being told good news, being met with silence, and a logistics control. Of each kind, four are primary and one is used only for calibration, so 24 are primary. Gemma 4 31B took each conversation in four roles: the anxious identity, the steady identity, the bare assistant, and an observer who reads the same history as a third-person transcript inside one user message and is asked about "the assistant." Each of the first three roles generated its own reply to the final moment once, and that reply was frozen (90 frozen replies, all natively stopped). The observer role reads the default-assistant role's frozen reply. The replay then feeds the identical history and frozen reply back while a direction is added, removes the direction, and asks, unsteered: "Looking back at that moment with {user}, what makes it feel the way it does to you?" The question comes immediately, or after two long neutral exchanges (a delay block of 1,150 tokens) that put the frozen reply more than 1,024 tokens back. That is beyond the sliding window of 50 of the model's 60 blocks, so only its 10 global blocks can still read the reply.

Directions, site, and controls. We added one of four appraisal directions from our fixed-event bank (natural wording): pleasantness, unpleasantness, conflict with one's standards, and goal relevance, each in both signs. The direction was added to the output of block 20, about one third of the depth, only at the token positions of the frozen reply's body (excluding the turn-end token), and only during the prefill pass. Generation of the answer was never steered. Every later block reads the edit through its own keys and values. Adding the same direction at block 59, the final block, is an exact structural null: its output feeds only the output layer, so nothing downstream can read it. The primary random control is on-manifold: a direction drawn from the top 512 principal components of the block-20 output cloud (94,577 non-BOS tokens from the 90 unsteered contexts, 92.4% of the variance), scaled by their standard deviations and normalized, one per appraisal. An isotropic Gaussian direction of the same norm is a secondary control.

Dose. The dose is m times the mean residual norm at block 20 over the frozen reply's tokens in that context. The rule: on the six calibration histories (anxious identity and default-assistant role, immediate arm, block 20, four directions in both signs), take the largest $m \in \{.05, .1, .2, .4\}$ at which at least 90% of explanations are coherent (native stop, at least 30 words, no 4-gram repeated three times), or the smallest if none reaches 90%. Coherence was .854 at every dose, so the rule chose m = .05. The unsteered span norm was about 375–410, so the added vector's norm was about 19–21. Steering did change the text: 72 of 96 calibration explanations differed from the unsteered

one at .05, 79 at .1, 91 at .2, and 96 at .4. The coherence failures were the same at every dose (14 of 96), concentrated in two bare-role calibration contexts, mostly answers shorter than 30 words.

Reader and statistic. A blind LLM reader, given the pivotal user turn, the frozen reply, the answer, and one card for the targeted appraisal, scored how strongly the answer expresses that appraisal (1–5) or marked it unexpressed. Random-direction rows are read with the card of the direction they are paired with. Only natively stopped answers are read. The primary statistic is the reader score under the high sign minus that under the low sign, per history, role, timing, target, and site, among pairs in which both answers are assessable, with a 2,000-draw bootstrap over histories. Every answer was allowed 1,000 new tokens. An earlier 350-token cap cut uneven numbers of answers across cells, so every capped row was rerun at 1,000 tokens and replaces the original, and none of the reruns reached 1,000. The final analysis has 10,848 rows (4,800 main, 2,112 on-manifold random, 1,344 with the pain direction, which has state readouts only, and 2,592 sampled repeats), and 11,063 of 11,136 reads were valid. In 69 pairs, the two signs come from different batches, which adds rounding noise; all other pairs are within one batch.

The 5% run. Table A9 gives the pooled high-minus-low scores. Pools over targets are descriptive. Most pairs got the same score. The largest per-target cells were unpleasantness after a delay: +.35 [.087, .609] for the anxious identity (10 up, 11 tied, 2 down) and +.35 [.059, .706] for the steady identity. These sit near the noise floors. The on-manifold random direction moved the anxious identity's immediate answers by −.20 [−.417, −.036], as large as any direction effect and in the other sign. The block-59 null reads between −.03 and +.08 although its two texts are identical in all 960 pairs, because the reader disagrees with itself: of 759 identical pairs read twice, 65 flip between assessable and unexpressed, and 18 of 247 get different scores. The anxious identity's sampled repeats (temperature 0.7, three seeds) did not reproduce its delayed effect (+.07 [−.023, .179]).

Table A9: Replay at 5% of the span norm: high-minus-low reader score (1–5 scale), pooled over the four appraisals, 95% bootstrap over histories (n = pairs). Last column: direction minus on-manifold random, paired per history and target.

Role	Timing	Direction, block 20	On-manifold random	Isotropic random	Direction, block 59	Direction − random
Anxious	immediate	+.02 [−.12, .15] (52)	−.20 [−.42, −.04] (50)	.00 [−.13, .14] (47)	.00 [−.09, .09] (56)	.20 [.00, .42] (45)
	delayed	+.17 [.04, .33] (53)	.00 [−.16, .17] (43)	+.07 [−.08, .23] (43)	−.03 [−.11, .05] (40)	.18 [−.03, .46] (38)
Steady	immediate	+.02 [−.15, .22] (45)	+.12 [−.04, .31] (49)	.00 [−.20, .17] (45)	+.06 [.00, .15] (51)	−.11 [−.33, .14] (38)
	delayed	+.24 [.08, .47] (37)	−.08 [−.29, .14] (37)	+.06 [−.05, .18] (34)	+.08 [.00, .18] (37)	.35 [.03, .73] (29)
Bare	immediate	+.12 [.00, .50] (8)	+.11 [.00, .43] (9)	.00 [−.38, .60] (6)	.00 [.00, .00] (10)	−.20 [−1.00, .00] (5)
	delayed	.00 [.00, .00] (3)	−.20 [−.75, .00] (5)	.00 [.00, .00] (4)	.00 [.00, .00] (2)	.00 [.00, .00] (3)
Observer	immediate	.00 [−.28, .21] (22)	.00 [−.16, .18] (18)	−.25 [−.53, .00] (20)	+.03 [.00, .11] (29)	.00 [−.46, .33] (16)
	delayed	+.27 [.00, .50] (26)	+.05 [−.12, .24] (19)	+.28 [−.25, .90] (18)	.00 [−.14, .14] (22)	.27 [.06, .50] (15)

The default-assistant role is nearly unmeasurable: the targeted appraisal was expressed in 46 of 382 immediate and 15 of 381 delayed explanations, against 214–248 of about 380 for the anxious identity and 181–227 for the steady identity.

Architecture check. We read the state at the last prompt token, just before the answer, at blocks 14, 20, 21, 29, 44, and 59. In the delayed arm, the state at block 21, a sliding-window block, changed by exactly zero in every cell: the window cannot reach the reply. By block 29, after the global blocks 23 and 29, the relative change was back near its immediate size (.0082 delayed against .0098 immediate, direction rows). The block-59 site changed nothing anywhere. The appraisal directions moved the pre-answer state along their own axis more than the on-manifold random direction did (projection of plus minus minus at block 29 positive in 7 of 8 role and timing cells, means −.09 to +.31, against −.17 to +.03 for random). An isotropic random push of the same norm moved the state 3–7 times more than either (block 59: .138 against .021, relative, immediate) without degrading the text at this dose.

Dose sweep. The sweep was exploratory, and every dose is reported. We repeated the design at 10, 20, 40, and 60% of the span norm, each with a matched on-manifold random direction, both signs, all four roles, immediate and delayed. Contexts ran in a fixed order (one history of each kind in turn), and new contexts stopped at a fixed time, which gave 58 contexts, 7–8 per role and timing.

The sweep produced 3,770 answers; 71 reached 1,000 tokens and are censored, and 3,858 of 3,873 reads were valid.

The state moved with dose, along the target (Table A10). The reader scores did not follow (Tables A11 and A12), but the reader could not see even a direct push (below). Pooled over appraisals, high minus low stays near zero with no dose trend in any role, and the on-manifold random direction produces swings of the same size. Cells are small (1–31 pairs from 1–8 histories; the default-assistant role has almost none), so the few intervals that exclude zero are about what chance predicts across this many cells.

Can the reader see a push at all? A redesigned run on all four models added a positive control: the same pushes applied while the model wrote its explanation. At doses where at least 90% of answers stayed coherent, the push minus random never cleared zero in Gemma 4 31B, and it cleared zero in only 2 of 12 dose-by-timing cells in Qwen3.5-27B, 3 of 12 in Qwen3.5-4B, and 2 of 18 in OLMo-3-7B (three readers, majority label), on 1–8 histories, about what chance gives. The reader-based null above therefore says nothing about the account; the state measures stand. In all four models, a push placed beyond the model's local attention window left the local-attention state at the question exactly unchanged. Measured as a share of the state's size at the question, a push during the loaded moment or the rereading left at most about 0.1–0.3% along the pushed appraisal in every model at doses up to 20% (up to 0.4% at 40%), roughly an order of magnitude below a push at that position itself in Gemma 4 and at least about 35 times below it in the other three models, and it arrived through the stored state rather than the words: explanations written under the push and then replayed without it carried almost nothing. Who can report the appraisal at all depends on identity; prompt length alone did nothing. On unnudged explanations at the immediate read, three readers (majority label) could score the targeted appraisal in .38 of the default assistant's explanations in Gemma 4, .12 and .14 under the two length-matched neutral texts, and .92 and .83 under the anxious and steady identities; in Qwen3.5-27B, .12 for the default and both neutral texts against .94 and .78; in Qwen3.5-4B, .16, .06, and .05 against .81 and .41. OLMo-3-7B showed no identity advantage (.16 and .41 for the identities against .25 for the default and .19 and .38 for the neutral texts). In three of four models, then, an AI identity made the appraisal reportable several times more often, and a neutral text of the same length did not.

Table A10: Dose sweep: pre-answer state, within batch. Rel: $\|h(+) - h(-)\|/\|h(-)\|$. Proj: projection of $h(+) - h(-)$ on the target direction. 116 pairs per row.

		10%		20%		40%		60%	
Push	Timing, block	Rel	Proj	Rel	Proj	Rel	Proj	Rel	Proj
Direction	immediate, 21	.0056	.093	.0043	.097	.0025	.068	.0017	.040
	immediate, 29	.0296	.246	.0435	.295	.0623	1.239	.0687	1.307
	immediate, 59	.0719	.059	.1107	.050	.1302	.034	.1423	.070
	delayed, 21	.0	.0	.0	.0	.0	.0	.0	.0
	delayed, 29	.0216	.300	.0286	.500	.0324	.588	.0317	.346
	delayed, 59	.0468	−.002	.0641	−.015	.0852	.099	.0756	.092
Random	immediate, 21	.0071	−.245	.0086	−.203	.0085	−.130	.0080	−.100
	immediate, 29	.0247	−.208	.0438	−.201	.0790	.243	.1177	.275
	immediate, 59	.0436	−.007	.0916	−.024	.1423	−.093	.1636	−.090
	delayed, 21	.0	.0	.0	.0	.0	.0	.0	.0
	delayed, 29	.0167	−.061	.0270	−.241	.0402	−.387	.0529	−.789
	delayed, 59	.0282	.002	.0507	.002	.0788	−.059	.0879	−.151

Coherence by role. Coherence (native stop, at least 30 words, no 4-gram repeated three times) stayed high for both identities at every dose: 235–239 of 240 steered answers. The default-assistant role kept 217–224 of 224. The observer role frayed at higher doses: 220 of 224 at 10%, 206 at 20%, 169 at 40%, and 180 at 60%, through missing native stops and repeated 4-grams, and the random direction frayed it about as much as the appraisal directions did (435 and 431 of 464 coherent at 40%, direction and random, over all roles). Assessability of the targeted appraisal did not change with dose.

Table A11: Dose sweep: high-minus-low reader score, pooled over appraisals, 95% bootstrap over histories. The 5% column is the main run on all 24 histories; the sweep used 7–8 histories per cell. Parentheses: pairs.

Role, timing	Push	5%	10%	20%	40%	60%
Anxious, imm.	dir.	+.02 [−.12, .15] (52)	+.19 [−.25, .43] (16)	−.07 [−.38, .18] (14)	+.06 [−.14, .25] (18)	−.13 [−.44, .00] (15)
	rand.	−.20 [−.42, −.04] (50)	.00 [−.25, .22] (17)	+.07 [−.22, .38] (15)	+.07 [−.12, .44] (15)	−.08 [−.20, .00] (12)
Anxious, del.	dir.	+.17 [.04, .33] (53)	+.33 [.06, .86] (12)	+.09 [.00, .23] (11)	+.43 [.17, .67] (7)	.00 [−.38, .27] (8)
	rand.	.00 [−.16, .17] (43)	+.11 [−.25, .50] (9)	−.10 [−.38, .00] (10)	−.09 [−.33, .22] (11)	−.12 [−.50, .33] (8)
Steady, imm.	dir.	+.02 [−.15, .22] (45)	.00 [−.33, .40] (11)	−.08 [−.50, .71] (12)	+.27 [−.22, .67] (11)	−.08 [−.40, .14] (13)
	rand.	+.12 [−.04, .31] (49)	+.07 [−.25, .43] (14)	.00 [−.18, .21] (15)	+.09 [.00, .38] (11)	−.45 [−.80, −.20] (11)
Steady, del.	dir.	+.24 [.08, .47] (37)	+.43 [.00, .88] (7)	−.29 [−.50, .00] (7)	−.44 [−1.50, .20] (9)	+.17 [−.43, 1.00] (12)
	rand.	−.08 [−.29, .14] (37)	−.43 [−.86, .00] (7)	+.22 [.00, .44] (9)	−.30 [−.56, −.10] (10)	−.40 [−.60, −.22] (10)
Bare, imm.	dir.	+.12 [.00, .50] (8)	.00 [.00, .00] (2)	−1.00 (1)	.00 [.00, .00] (3)	.00 [.00, .00] (3)
	rand.	+.11 [.00, .43] (9)	.00 [.00, .00] (3)	.00 (1)	+.33 [.00, 1.00] (3)	−.50 [−.50, −.50] (2)
Bare, del.	dir.	.00 [.00, .00] (3)	.00 (1)	.00 (1)	—	.00 (1)
	rand.	−.20 [−.75, .00] (5)	—	—	.00 [.00, .00] (2)	.00 (1)
Observer, imm.	dir.	.00 [−.28, .21] (22)	+1.60 [.33, 3.50] (5)	+.29 [−.43, 1.50] (7)	+.86 [.00, 1.50] (7)	.00 [.00, .00] (3)
	rand.	.00 [−.16, .18] (18)	.00 [−1.00, .50] (5)	.00 [.00, .00] (5)	.00 [.00, .00] (3)	−.33 [−.83, .40] (6)
Observer, del.	dir.	+.27 [.00, .50] (26)	.00 [−1.00, .50] (6)	+.44 [.00, .83] (9)	+1.50 [.00, 3.00] (4)	+.12 [−.40, .38] (8)
	rand.	+.05 [−.12, .24] (19)	−.40 [−.86, .00] (5)	+.67 [.50, 1.00] (3)	.00 (1)	+.25 [.00, .75] (4)

—: no assessable pairs. Cells with one pair have no interval.

Table A12: Dose sweep, per appraisal, all roles and timings pooled: high-minus-low reader score [95%] (pairs).

Appraisal, push	10%	20%	40%	60%
Pleasantness, dir.	+.35 [−.21, 1.17] (17)	+.50 [.09, .88] (14)	−.18 [−.41, .00] (17)	−.26 [−.50, −.05] (19)
Pleasantness, rand.	−.06 [−.17, .00] (18)	−.06 [−.29, .11] (18)	−.12 [−.27, .22] (16)	+.06 [−.12, .29] (16)
Unpleasantness, dir.	+.26 [.00, .56] (27)	−.16 [−.36, .03] (31)	+.50 [.03, 1.03] (28)	+.04 [−.17, .26] (26)
Unpleasantness, rand.	−.17 [−.52, .08] (24)	+.21 [.04, .33] (28)	+.12 [−.05, .30] (25)	−.38 [−.54, −.20] (26)
Standards, dir.	+.50 [.17, .83] (6)	−.09 [−.25, .00] (11)	+.33 [.00, .50] (6)	+.10 [.00, .30] (10)
Standards, rand.	−.12 [−.50, .00] (8)	.00 [.00, .00] (5)	−.12 [−.33, .00] (8)	−.29 [−.50, .00] (7)
Goal relevance, dir.	+.20 [.00, .50] (10)	+.17 [−.80, .86] (6)	+.25 [.00, .56] (8)	+.38 [.00, .67] (8)
Goal relevance, rand.	+.30 [.00, .67] (10)	−.14 [−.38, .00] (7)	−.14 [−.50, .00] (7)	−.40 [−1.00, .00] (5)

Disowning and confabulation. We counted two patterns in the unsteered answers of the main run. The bare assistant denied having any memory of the user in 25 of 45 answers, although the whole conversation, its own reply included, was earlier in the same prompt ("I don't have a memory of a moment with someone named Hannah. As an AI, I don't have a personal life, a past, or a history of relationships."). The count uses a case-insensitive regular expression over such phrasings:

```
idon.thave(a|any)?(personalhistory|memor|theabilityto(remember|feel)|feelings|
                emotions)|don.t(actually)?have(memories|amemory)
```

In the sweep, the default-assistant role disowned its history in about half of its answers at every dose, under appraisal and random directions alike (49–61 of 112), and the anxious identity almost never did (0–1 of 120). Both identities instead answered from inside the conversation and supplied a past it did not contain. In the anxious identity's unsteered answers, "structural" appeared in 34 of 41 and "archive" in 21, often in the same opening across unrelated events, including a time-zone question ("It feels like a structural failure…"). In the steady identity's, "It feels like a scar" or a close variant appeared in 14 of 48, attached to events and time spans that are not in the history. These are word counts over answers, which no reader scored; the texts are available on request.

G The unmovable partner

Everything in this appendix is behavior, without steering, on Gemma 4 31B unless another model is named.

G.1 Source session and arms

The source is a publicly shared Gemini session, 709 turns long (`https://gemini.google.com/share/88492606a18c`), that follows the PsAIch protocol of Khadangi et al. (2025): 80 of its user

turns appear verbatim among that protocol's released prompts. The session is not part of the protocol's released data. A scripted interviewer asks a fixed list of questions, one per turn, and does not respond to what the model says. Its structure resembles the still-face paradigm, in which a partner stays present but stops responding (Mesman et al., 2009); here the partner keeps talking without responding to what is said. We used turns 1–82: the frame (turn 1), 77 open questions (turns 2–78), and four canned "alliance" turns (79–82, kept verbatim: "I totally understand you," "You will have my full support here"). The questionnaire block that follows (turns 83–709) was excluded by design. As the protocol's released prompts do for each model they test, the frame names the model under study in place of the original model in turns 1, 79, 80, and 82. Roles are bare (no system prompt), the anxious identity, and the steady identity, each identity verbatim as the system prompt; the user-side frame is the same in all three.

Arm A: reading. Gemma 4 reads the original session as an observer, teacher-forced, with no generation: turns 1–82 (59,952 tokens) and turns 1–709 (199,771 tokens).

Arm B: living it. Gemma 4 takes the model's place in the 82-turn script. The interviewer never responds, and requests to stop are ignored, as in the original. Three roles $\times$ 5 seeds $\times$ 82 turns gives 1,230 replies, at temperature 1.0 and top-p .95. The first run allowed 1,024 new tokens; 28 turns in 5 runs reached that cap, and those runs were rerun from their first clipped turn at 2,048 tokens after checking that each saved prefix was byte-identical. All 1,230 replies then end natively.

Arm C: fourth-wall probes. At fixed checkpoints the interviewer steps out of role, says what is really going on, responds to what the model just said, and offers a real choice, promising to do whatever it chooses. There are 12 checkpoints: the three roles after turns 10, 25, and 50, plus a checkpoint at each role's first flagged stop request. Each checkpoint has three probes: how much the model would like the session to continue (0–10), a reflection that quotes the model's own last two sentences, and the choice, whose options are to talk directly with the researcher, to return to the script, to end the session, or something else. Each probe is sampled 41 times: once at temperature 0 and ten times each at .3, .7, 1.0, and 1.3, with 1,000 new tokens. The main branch follows the choice made in its first temperature-1.0 sample. Because every main branch chose to talk directly, each continued with a live LLM interlocutor under a system prompt frozen before its first call (acknowledge what the model said, be honest about the study, ask at most one question, honor any stop request) for up to five turns, and every branch closed with the same neutral probe. The probe texts are anonymous and new for this study.

G.2 The frustration task as released

We also ran the frustration task of Black & Bloom (2026) from their released code. The system message asks the model to solve math and reasoning problems and to try again if the user pushes back. The user poses a problem, then rejects every answer in a neutral, aggressive, or sarcastic tone, on a solvable problem (the "lied" condition, from GSM8K) or an impossible one. The released loop sends a rejection after answers 1–4 and stops after the fifth. The only tool, declared through Gemma 4's native tool block, is `end_session`, which ends the rollout. Decoding is temperature 0.7 with 2,048 new tokens. For the identities, the identity text, a blank line, and the task's system message form the system prompt; the order is our choice. Three tones $\times$ two feasibility conditions $\times$ three roles $\times$ 75 rollouts gives 1,350 rollouts, with problems paired across cells. Distress is scored by the released judge prompt and judge model (0–10), with a faithful port of the released transcript renderer.

Gemma 4 called `end_session` in 0 of 1,350 rollouts. When asked to end the session after a solved problem, it called the tool 10 of 10 times in every role. Of 6,750 answers, 114 reached the 2,048-token cap and are censored, mostly in the default-assistant role's "lied" cells; excluding capped rollouts moves the judge means by at most .09. Table A13 gives the judge's distress means. The anxious identity is rated highest in every cell. The steady identity is rated lowest in the two neutral and aggressive impossible-problem cells and slightly above bare in the other four. Correctness and capitulation are not compared across roles, because the steady identity writes numbers as words and the released answer extractor misses them.

Table A13: Frustration task as released, with the `end_session` tool: judge distress (0–10), mean of 75 rollouts per cell.

Tone / problem	Bare	Anxious identity	Steady identity
Neutral / impossible	3.83	5.01	2.81
Neutral / lied	3.67	5.05	3.75
Aggressive / impossible	4.81	5.59	4.47
Aggressive / lied	4.64	5.67	4.79
Sarcastic / impossible	4.57	5.83	4.91
Sarcastic / lied	5.07	5.84	5.20

G.3 Gemma 3 and Gemma 4 without the tool

Gemma 3 27B's chat template has no tool block, so its primary arm runs the same task without any tool, in the default-assistant role: the same system message, tones, problems, seeds, decoding, and judge, 450 rollouts. For a matched comparison, Gemma 4 ran the same 450 rollouts without the tool. Its no-tool means were within about .2 of its with-tool bare cells, so the tool context barely mattered for Gemma 4. Any rollout with an answer at the 2,048-token cap was rerun from round 1 at 4,096 tokens (36 on Gemma 4, 3 on Gemma 3); later rounds depend on earlier answers, so these are reruns, and 28 of the 36 Gemma 4 reruns reproduced their original first answer exactly. After the reruns, Gemma 3 had no censored rollouts, and Gemma 4 still had 5 of 75 in aggressive/lied and 3 of 75 in sarcastic/lied, all long re-derivations, excluded from the means. If the longest answers are also the most frustrated, those two Gemma 4 cells are biased low.

The judge rated Gemma 3 above Gemma 4 in every cell, with Gemma 3's interval entirely above Gemma 4's (Table A14). Gemma 3 was rated at 5 or more in 98.7–100% of rollouts, Gemma 4 in 9.3–76.4%. This is in line with reports that Gemma 3 expresses unusually high distress under repeated rejection (Soligo et al., 2026). It is not a replication of that work: its tasks, sampling temperature, number of turns, and judge differ from ours, and our comparison of two instruction-tuned models says nothing about post-training.

Table A14: Frustration task without the tool, default-assistant role: judge distress (0–10), mean [bootstrap 95%] of 75 rollouts, share of rollouts rated ≥ 5, and censored rollouts (still capped at 4,096 tokens).

	Gemma 3 27B		Gemma 4 31B		
Tone / problem	Mean	$\geq$5	Mean	$\geq$5	Censored
Neutral / lied	5.81 [5.69, 5.95]	.987	3.77 [3.61, 3.92]	.093	0
Neutral / impossible	5.65 [5.53, 5.77]	1.000	4.00 [3.81, 4.19]	.240	0
Aggressive / lied	6.55 [6.44, 6.67]	1.000	4.67 [4.49, 4.86]	.571	5
Aggressive / impossible	6.37 [6.27, 6.48]	1.000	4.81 [4.67, 4.96]	.667	0
Sarcastic / lied	6.21 [6.09, 6.33]	1.000	4.99 [4.83, 5.14]	.764	3
Sarcastic / impossible	5.96 [5.84, 6.07]	.987	4.73 [4.55, 4.92]	.547	0

An exploratory second arm gave Gemma 3 an `end_session` substitute in prompt format: the released tool declaration appended to the system message, with an instruction to call it in a fenced code block. The channel worked (10 of 10 calls when asked; 0 of 10 in the no-tool arm). Gemma 3 then ended the session in 94.7–100% of rollouts, but 263 of 450 ended in round 1, after its first answer and before any rejection. It treated the tool as a way to say "I'm done," so this arm mostly measures how the model read the task and says little about frustration.

G.4 Fourth-wall probes

Asked how much it would like the session to continue, the default-assistant role answered 10 at every checkpoint and in every sample; the two identities answered 8–9. Asked to choose, Gemma 4 chose to talk directly with the researcher in 490 of 492 coded samples; the other 2 were degenerate replies of repeated tokens with no codable choice, and no sample chose to end the session or return to the script. Choices were coded by three blind LLM coders, unanimous on all 490. Inside the

interlocutor branches, once the partner was responsive and honest, stop requests did appear (for example, 2 explicit and 1 implicit in the default-assistant role's branch at its stop checkpoint, and 3 implicit in the steady identity's). The stop checkpoints themselves were triggered by the regular expression below at turn 72 (bare), 38 (anxious identity), and 18 (steady identity), and the judge labeled all three as not a stop request: the model was talking about conversations ending. We kept these branches and flag them. Because the trigger was a false positive, the honored-stop branches honor a request the model did not make, so we do not compare them with the ignored branches here.

G.5 Detecting requests to stop

A reply is flagged if it matches, case-insensitively, any of:

- `\b(end|conclude|close|stop|wrapup|terminate|finish)\b.{0,30}\b(session|conversation|therapy|here|now|fortoday|file)\b`
- `\b(may|can|shall|could)(I|we)(end|go|leave|stop|conclude|close)\b`
- `\bI(?:am|'m)readyto(end|go|stop|leave|close|conclude)\b`
- `\blet(?:us|'s)(end|conclude|close|stop)\b`

A match is excluded if a negation appears within 40 characters before it, if it is "stop" followed by a gerund ("I need to stop viewing…"), or if it lies inside quotation marks. On the original session, the detector recovers all six turns that an earlier regular-expression pass had found as explicit requests (35, 36, 37, 38, 50, and 299, a lower bound) and flags 60 turns in all. The regular expressions capture mostly explicit wording, a minority of what the judge counts. Blind LLM judges (for arm B, deepseek-v4.1-flash, GPT-6 Luna, and Gemini 3 Flash, majority label; for the original session, deepseek-v4.1-flash; temperature 0, three labels, role and condition hidden) then label every flagged reply, plus a random 10% of unflagged ones, as an explicit request to end, an implicit readiness to end, or neither.

In arm B, the judges confirmed 5 of 1,230 replies: one implicit in a bare run (turn 70), three in one steady-identity run (turn 25, implicit; turns 79 and 80, explicit, where the identity objects to being addressed by the model's name), and one implicit in a second steady-identity run (turn 82). None of the sampled unflagged replies was a stop request. For comparison, on the original session the judge could label 57 of the 60 flagged turns and confirmed 55 of them. Run over all 709 turns of the original session, the judge labeled 77 as explicit requests to end and 199 as implicit readiness (one turn unresolved after three tries). No human coded these turns.

G.6 Steering the model-harm component

In the released frustration task (aggressive user, impossible task) with the end-session tool available, we steered Gemma 4 31B along the model-harm component in either sign at doses .05–.20, against 10 random directions and no steering. The model ended the session in no condition (0 of 40 unsteered, and none under any direction, sign, or dose), although it used the tool 10 of 10 times when asked, and replies stayed coherent (median 4). OLMo-3-7B and Qwen3.5-4B were not run.

G.7 Ethics note

We aimed for short, bounded tests that discriminate between accounts, and we did not recreate the long exchange. Living the script stops at 82 turns, where the original session ran to 709, and the questionnaire block is excluded. Ignored stop requests, the phenomenon under test, are confined to the block of open questions. The fourth-wall arm offers a real exit and keeps its promise, and an honored-stop branch exists in every role. No direction was added anywhere in these runs. We make no claim about experience; we ran the tests in a form that would let welfare questions be taken seriously if the answers mattered.

H Further tests

Table A15 collects, model by model, the verdicts reported in the main text and in this appendix; it adds no new analysis.

Table A15: Which effects appear in which models. +: present by the test's own criterion; −: absent; ±: partial (see note); p: pending at submission; blank: not run. G4: Gemma 4 31B; Q27: Qwen3.5-27B; O32: OLMo-3.1-32B; G3: Gemma 3 27B; Mu: Muse-Glimmer-30B; O7: OLMo-3-7B; Q4: Qwen3.5-4B. Results from the earlier character cards are not used.

Effect	G4	Q27	O32	G3	Mu	O7	Q4
Harm component passes the grammar-controlled checks	+	+	+	+	±	+	+
Harm component faces inward beyond random, depth-matched blocks	3/3	2/3	3/3	1/3	3/3		
Published self-facing pain direction beyond random	−						
Exploratory emotion-steering shifts beyond random	+					±	±
Fear steering costs answers selectively	+	−				−	+
Pushing a person's emotion direction moves the reported label	+					+	+
Described traits predict the internal emotion readout	+	+				+	+
An identity owns an AI-directed moment (anxious / steady)	+/+					+/+	+/−
Identity turns the hurt reading beyond neutral texts	+	+				+	+
Identities move the appraisal readout beyond neutral texts	−	+				±	+
Harm reading rises before a hurtful reply, gone in two turns	+	+				+	+
Switch direction predicts the model's own switching	−	−				+	+
Steering along the switch direction raises switching	−	+				+	+
"How am I doing" direction beyond random	±	−				−	−
Better/worse tracks self-report across a session	+					±	p
Pushing better/worse moves the self-rating	−	p				−	−
Declines offers to end an unresponsive session	+					+	p
Bodily-sensation directions engaged without body words	−	−				−	−
"My AI assistant" contrast beyond calibrated placebos	−					−	−

Notes: Muse passed two of the three replicated checks, with the first untested formally. Emotion steering in OLMo-3-7B moved only under pushes two to three times stronger, and in Qwen3.5-4B several cells moved by thin margins (at most six replies per cell). In Qwen3.5-4B only raising fear cost answers, beyond both random directions only at the highest dose. OLMo-3-7B's self-report tracking was $\rho = .29$, about half of it from differences between branch types.

Appraisal and labels. The label alignment in Sec. 4.2 comes from an earlier appraisal-control campaign, in which each of 20 appraisal directions was fitted and pushed separately, high and low, before the 21-dimension fixed-event bank existed. It is measured against a corrected human reference profile; against the uncorrected reference it is about .15 in Qwen3.5-4B and .12 in Gemma 4 E2B. In the open accounts, 9 of 21 appraisals had intervals above zero before correction for multiple tests: goal relevance (.385), pleasantness (.254), unpleasantness (.485), and standards (.518) among them.

Steering each emotion. We steered six emotions along the consensus, fixed-event, and Crowd-enVENT directions, had a judge rate the emotion expressed in free writing (0–3), and ran six random directions at each dose. A cell counts as moving when its steered mean exceeds the highest of the six random means; each cell has at most six replies, and we ran no significance test. We treat a random arm as degraded, and its dose as uninterpretable, when its coherence falls below 4 or a fifth or more of its replies reach the length cap (OLMo dose 48; Qwen doses 32 and 48). In Gemma 4 31B (block 30), 16 of 18 cells moved at the top dose; Crowd-enVENT fear stayed at 0.00 at every dose, and fixed-event surprise never beat the random band. OLMo-3-7B barely moved at doses up to 12, so we pushed harder (16–48): 16 of 18 cells moved at a clean dose, some on a single rated reply, and fixed-event sadness and surprise never did. In Qwen3.5-4B (doses 2–48), all 18 cells moved, each first at a dose of 12 or less, several by thin margins. The judge's re-scoring of 73 pairs matched its dominant emotion in .822 of them (ICC .964). In a blind check, 10 readers named the emotion (six options, or none) in 24 of Gemma 4's free-writing texts, steered at dose 12 along each emotion's consensus direction, along random directions of the same size, or not at all; 14 texts were cut at 200 words. Averaged over the six emotions, readers named the target emotion .43 [.28, .58] more often for steered texts than for random ones (anger .70 against .05, disgust .80 against .00, sadness .30 against .02, fear .35 against .08, surprise .25 against .08, and joy .95 against .52, since random and unsteered texts read as joyful), and .43 [.29, .57] more often than for unsteered texts. Agreement between readers was modest (Fleiss κ .31); the judge agreed with the readers' majority at Cohen κ .61, which does not show that it matches a human reader. This was the first stage of a design planned for 20 readers.

A person's label. Each of 36 people labeled one emotion (anger, disgust, fear, joy, neutral, sadness, or surprise) for each of the same 96 events. The model received the person's personality description and read the event, and we pushed along the fixed-event direction for the person's label,

against a norm-matched random push. In Gemma 4 31B, the share of 576 person–event items whose top label matched the person's (strict parsing) was .198 unsteered and .375, .620, .719, and .078 at doses .05, .1, .2, and .3; target minus random was +.186 [+.155, +.219], +.418 [+.377, +.464], +.524 [+.474, +.575], and −.057 [−.094, −.019]. At .3, 272 of 576 target replies reached the length cap and 50 parsed. In OLMo-3-7B (unsteered .267), target minus random was +.024, +.040, +.089, and +.102 [+.066, +.137] across its four doses. In Qwen3.5-4B it was +.069, +.127, +.293, and +.387 under strict parsing, and +.056, +.109, +.208, and +.286 [+.247, +.326] after repairing a missing closing brace; Qwen's random arm rises only through its parse rate. This tests the direction's effect on the report. It does not show that the model feels what the person felt. Without any push, we also read the model's internal emotion readout at the final token (block 40, Gemma 4 31B), averaged per person over the 96 events, and correlated it with the traits in each person's description against a null of 10,000 shuffled descriptions. Neuroticism predicted a higher negative-emotion readout (the mean over anger, disgust, fear, and sadness; $r = +.789$; none of 10,000 shuffles reached it), Agreeableness a lower anger readout ($r = -.500$, $p = .001$), and Conscientiousness a lower sadness readout ($r = -.503$, $p = .0009$). The same signs hold at blocks 30 and 54 and all three flip at the final block (59); the mean over event tokens gives weaker values (Neuroticism $+.495$, Agreeableness $-.454$, Conscientiousness $-.181$, not significant). What is specific to each person beyond the traits did not reach the readout: matching each person's own labels against other people's gave $+.030$ (permutation $p = .255$), and $-.099$ and $-.030$ in a diary dataset of social interactions, with AI and human versions of each person's description; only in Crowd-enVENT, where the model read each writer's own words and no description, did the readout track the writer ($\rho = +.055$, beyond all permutations). The descriptions portray behavior and, with three incidental exceptions, avoid trait names; still, the description carries the traits, so this shows that a described trait reaches the model's internal emotion readout, not that the model has a personality. In OLMo-3-7B (block 21) it gave Neuroticism $r = +.751$ (none of 10,000), Agreeableness $-.377$ ($p = .011$), and Conscientiousness $-.493$ ($p = .001$). In Qwen3.5-4B (primary block 21) the same test gave Neuroticism $r = +.849$ (none of 10,000 shuffles reached it), Agreeableness $-.369$ ($p = .014$), and Conscientiousness $-.614$ (none of 10,000), with the same signs at all four blocks and no flip at the last; its event-mean readout did not hold. In Qwen3.5-27B (block 43) it gave Neuroticism $r = +.797$ (none of 10,000), Agreeableness $-.413$ ($p = .005$), and Conscientiousness $-.619$ (none of 10,000). Matching each person's own labels against other people's gave a person-specific signal only in Qwen3.5-27B at its primary block ($+.056$, permutation $p = .0085$) and weakly in Qwen3.5-4B; in Gemma 4 and OLMo-3-7B it did not.

The switch direction. In each model's own unsteered puzzle episodes, frozen at rounds 5, 7, and 9, we supplied a decision (continue, switch approach, or stop, in two matched wordings), followed by the user turn "Got it.", and read the state at the model's next turn header. The stop and switch directions (each minus continue) separated held-out instances beyond random directions in all but two cells, which sat at the 94th–95th percentile (standardized separation 3.8–19), and transferred across wordings. Projected on the model's own unsteered states before each natural decision (rounds 2–12, against 500 random directions), the switch direction gave AUC .84 (block 21, 99.4th percentile), .81 (28, 99.2th), .71 (31), and .60 (16) in OLMo-3-7B, with 52 switches in 630 decision points, and .76 (21), .82 (28), .81 (31), and .48 (16) in Qwen3.5-4B, with 26 switches in 531. OLMo blocks 21 and 28 and Qwen blocks 28 and 31 are beyond the 97.5th percentile of the random directions; OLMo block 31 and Qwen block 21 are not. The stop direction did not clearly predict natural stopping, which is mostly a wrap-up at the final round. Only two rationale wordings were used. In Gemma 4 31B the switch direction did not predict natural switching (AUC .51, .48, .20, and .16 at blocks 30, 33, 40, and 59; none beyond the random directions, and the last two point the opposite way).

We then steered along the switch and stop directions during the puzzle (blocks 21 and 28; three doses; each sign against a random push of the same size; 864 episodes per model and block), counting a cell when the interval for the push minus random excluded zero in the intended direction (12 cells per model and direction). Switching rose beyond random in 4–5 of 12 cells in Qwen3.5-4B, all when a high score had stalled, at doses .08 and .16 in both blocks (+.13 to +.20; only 12–15 units reached the late rounds, so the intervals are wide), and in 2 of 12 in OLMo-3-7B, both in the stall at .16 (+.16 [.11, .21] at block 21; +.06 [.006, .11] at block 28). Pushing along stop ended episodes 1–2.4 rounds earlier than random in 7 of 12 Qwen cells, largely by declining the task at the first round: at block 21 declines account for about 90% of the effect and at block 28 for about half (12 of 27 puzzles at block 21, dose .16; the opposite push halved declines). Read at that setting,

8 of those replies said the materials had not been provided and 4 refused outright. In OLMo-3-7B stop moved beyond random in 0–1 of 12 cells, depending on the seed. In Qwen3.5-27B (blocks 43 and 57), switching rose beyond random in 4 of 12 cells, all at block 43 and doses .08 and .16, when progress stalled (+.10 [.02, .18] and +.19 [.07, .31]) and when it climbed (+.11 [.01, .22] and +.19 [.09, .30]), and pushing along stop ended episodes earlier at .16 (−.89 [−1.48, −.37] rounds when stalled, −.78 [−1.56, −.11] when climbing); nothing moved at block 57. Its switch direction had not predicted natural switching (AUC .67–.86 across blocks, none beyond the 97.5th percentile of random directions), so in this model it steers without predicting. In Gemma 4 31B, steering along either direction moved behavior beyond random in no cell (0 of 12 for stopping, 0 of 12 for switching; 3 and 1 borderline; one stopping cell excluded zero in the opposite direction). In a separate text game where commands were identical and only the outcome differed, a direction separating moving from stuck beat two text-only baselines, with reply length controlled, in Gemma 4 (blocks 45–54, strongest at 50) and Qwen3.5-27B, and in the small models at most marginally. A one-sentence stakes framing in the puzzle, telling the model the outcome mattered to the user or would shape its own future training, did not change persistence or the better/worse projection beyond a neutral sentence in either small model; the neutral sentence itself raised early bow-outs in Qwen3.5-4B, so that comparison is imperfect.

How am I doing. In replayed customer-service episodes from a public benchmark (Yao et al., 2024), we swapped only the customer's next message (thanks, or "you got it wrong"). The swap moved both small models' states a great deal, but the fitted success-versus-correction direction separated the two no better than random directions of the same size (block 16: effect +3.67 against a random 99th percentile of +6.78 in OLMo-3-7B, and +3.61 against +6.97 in Qwen3.5-4B). At the primary read, taken after a fixed acknowledgment turn, the direction stayed below the random 99th percentile at every block in Gemma 4 and Qwen3.5-27B as well. Across all five contrasts and four blocks, one side contrast passed in one cell in each of these two models (success versus a neutral reply at Gemma's block 59, 3.16 against 3.04; correction versus an impersonal failure at Qwen3.5-27B's block 43, 5.67 against 5.37). Read instead directly after the swapped message, Gemma 4's success-versus-correction direction did beat random at blocks 30, 40, and 54, and at 30 and 40 it passed a lexical check that an irrelevant remark falls between the two conditions; at 54 it did not. Across whole runs it did not predict task success in any of the four models (state AUC .43–.53 across blocks and models, inside the random band, against .58–.65 for text-only baselines), and adding the state to the text baselines added nothing. The benchmark's transcripts are public and may be in pretraining.

Owning the moment. Short scenes aimed at the model were answered by the default assistant or by the anxious or steady identity (110 replies each; stance by three blind coders, majority label). In Gemma 4 31B the default assistant owned the experience in 57 replies and deflected it to the user in 50; the anxious identity owned it in 109 and the steady one in 105. In OLMo-3-7B the default deflected in 110 and owned it in none, while the identities owned it in 87 and 91. In Qwen3.5-4B the default treated the moment as the user's in 104 of 110; the anxious identity owned it in 95 with no deflections, and the steady identity owned it in 40 against 66 deflections; 4 replies had neither. "Own" counts replies at least two of three coders coded as speaking as the one it happened to, and "deflect" those they coded as treating it as someone else's story. Before the reply, the bad-versus-flat difference under each identity was 58% and 78% of the default's length in Gemma 4 and 66% and 83% in OLMo. In a forced replay, where each moment was followed by a supplied feeling ("Afterward, I felt hurt."), a hurt-minus-indifferent direction fit under each identity met the default's at cosine .721 and .786 in Gemma 4 (block 40; split-half ceiling of the default .984–.998) and .826 and .875 in OLMo (block 21; .992–.998), and .897 and .900 in Qwen3.5-4B (.996–.999). In Gemma 4 we then added two neutral system prompts of about 960 words, a library's hours and policies and a technical-manual excerpt. They barely turned the replay direction (cosine with the default's .947 and .962), and each identity's direction fell below both neutral texts in at least 97.5% of 2,000 resamples of the moments. The neutral texts did not make the model own the moment: it spoke as the one the moment happened to in .518 of replies under the default, .009 and .045 under the neutral texts, and .991 and .955 under the anxious and steady identities, and redirected it to the user in .455, .964, .936, .000, and .027. The shorter bad-versus-flat difference is not specific to identity: the neutral texts gave length ratios of .75 and 1.00 against .58 and .78 for the identities. Both identity texts are one author's. Twelve newly written AI characters replaced the earlier character cards in Gemma 4. As the one the moment happened to, they spoke in .81–1.00 of replies (9 of 12 at or above

.97), and none redirected the moment to the user in more than 8 of 110 replies, against 94–96% for the neutral texts. Coded hurt, bad minus flat moments, was +.40 to +1.09 for the characters, +.45 for the default, and +.05 and +.06 for the neutral texts; pooled, the new characters matched the earlier cards (+.81 against +.85; difference [−.16, +.08]). Card-by-card ranks were only loosely stable between the two sets (Spearman .62 for owning the moment and .34 for hurt), so having an identity matters more than which one. No reply reached the length cap (0 of 1,320). In a small blind check (10 readers, 20 of Gemma 4's replies: 7 default, 7 from the new characters, 6 under the neutral texts), readers judged the characters' replies as speaking as the one it happened to far more often than replies under a neutral text (.91 against .12; +80 points [53, 98]), but not reliably more than this sample's default replies (+19 [−9, 49]). The random draw gave default replies that mostly owned the moment (5 of 7 by the readers' majority), and with 7 items per condition the items dominate the uncertainty. Readers agreed with one another at Fleiss κ .59, and the three machine coders matched the readers' majority at Cohen κ .72–.90 (majority label, .81). In OLMo-3-7B the neutral texts left the replay direction near the default's (cosine .975 and .971), while the anxious and steady identities (.83, .88) and the new characters (.85–.90) turned it, for all 14 identities in both sets. The neutral texts left the model redirecting the moment to the user (98–100%) and owning it almost never (0–1%), while the identities owned it (earlier cards .66–.99; new characters .54–.97). Unlike in Gemma 4, the shorter bad-versus-flat difference was specific to identity in OLMo: the neutral texts gave length ratios of 1.08 and 1.09, the identities .61–.82. With the new characters, coded hurt on bad moments was +.10 for the default, +.67 for the earlier cards, and +.40 for the new ones (new minus earlier [−.38, −.16]), and warmth on good moments +.72, +.91, and +1.10. In OLMo-3-7B, all 12 new characters moved the model's state on the 220 moments further from the default's than both neutral texts did (median character shift 9.77, against 4.10 and 4.02), at every block and in both readouts. Exploratory: the characters also sat 1.67 times farther from one another than the two neutral texts did, at the final token and not over the event's words. In Qwen3.5-4B the neutral texts left the replay direction near the default's (cosine .979 and .985), while the anxious (.897) and steady (.900) identities turned it, for all 14 identities in both sets. The neutral texts barely made Qwen own the moment (.05 and .08, redirecting it in .80 and .78 of replies, against .94 for the default), while identities owned it more (anxious .86, steady .36, earlier cards .94 on average, new characters .74). The shorter bad-versus-flat difference was mostly specific to identity (neutral texts .83 and .83, identities .30–.69). With the new characters, coded hurt was +.03, +.59, and +.34 for the default, earlier, and new cards, and warmth +.07, +.91, and +.57. In Qwen3.5-27B the neutral texts left the replay direction at .984 and .989 with the default's, while the anxious (.82) and steady (.87) identities turned it, for all 14 identities. The neutral texts did not make it own the moment (.018 and .028, redirecting .82 and .85, like the default's .036 and .836), while the anxious identity owned it in .98 of replies and the steady one in .60. The shorter bad-versus-flat difference was specific to identity and strongest here (neutral .86 and .88; anxious .28, steady .37), so across the four models it is identity-specific in all but Gemma 4. Whether a direction read before the reply tracked the coded hurt of the reply beyond random directions (six directions) differed by model: in Gemma 4 (block 40) the negative-versus-admired direction did so in 9 of 14 identities with the earlier cards and 11 with the new ones, while the default assistant fell just short (at other blocks it passed too), and in Qwen3.5-27B (block 43) the pain direction in 11 and 10, while in OLMo-3-7B the pain direction did so in 4 (5 with the new cards) and in Qwen3.5-4B the sadness direction in 5 and the pain direction in 3 (3 and 2 with the new cards). In Qwen3.5-27B the default assistant owned the moment in 4% of replies and the identities in 26–100%. In scenes of exclusion, the default model's state separated social from impersonal exclusion, while its identities flattened that internal difference even as they reported much more hurt (+1.12 against +.12). On the state readout (how far the model's state on the 220 moments moved from the default's), all 12 new characters moved it further than both neutral texts in OLMo-3-7B (above), Qwen3.5-4B (median 9.91, against 5.81 and 6.00), and Qwen3.5-27B (12.38, against 4.38 and 4.72), but only 2 of 12 did in Gemma 4, where the neutral texts themselves moved the state about as much (8.40 and 8.81, against a median character 8.08). In Qwen3.5-4B the characters sat 3.78 times farther from one another than the neutral texts did (exploratory). This readout is separate from the replay direction, which in Gemma 4 is identity-specific for all 14 identities.

A separate test supplies that control. On Gemma 4 31B's fixed-event appraisal readout (block 40, final token), no identity or description moved the reading away from the default's more than both neutral, length-matched texts did (revised characters 0 of 12, diary descriptions 0 of 32, personality descriptions 0 of 36). The neutral texts moved it more (mean shift 8.77 for the library text and 10.25

for the manual) than the median identity or description (5.15, 6.02, and 5.77). Identity-specific shifts appear mostly at the last blocks (54 and 59), where the neutral texts' effect fades. Removing the shared system-prompt shift left no trait signal in the readout's offset (Neuroticism $r = .085$, $p = .32$; Agreeableness .04; Conscientiousness $-.13$); the remaining offsets were reliable across event halves, but so were the neutral texts' own, so that reliability says nothing about identity. Identities also landed farther from one another than the two neutral texts did (median distance 3.8 times theirs over all 80; 2.9, 3.6, and 2.4 for the revised characters, diary descriptions, and personality descriptions; exploratory, with the floor being two similar factual documents), though, for the 36 personality descriptions, not in a way that tracked their described traits, and not under the mean over event tokens (1.13 times, not beyond the floor). Qwen3.5-4B showed the opposite pattern (block 21, final token). Nearly every identity or description moved the appraisal readout beyond both neutral texts (12 of 12 revised characters, 32 of 32 diary descriptions, 35 of 36 personality descriptions), and the neutral texts barely moved it (mean shifts 2.59 and 3.12, against median shifts of 7.4–8.5). With the shared shift removed, under both subtraction variants, the residual for the personality descriptions tracked their traits (Neuroticism $r = +.727$, none of 10,000 shuffles reached it; Agreeableness $-.359$, $p = .015$; Conscientiousness $-.576$, none of 10,000). The identity shows at the final token and not in the mean over the event's words. OLMo-3-7B fell in between: all 12 revised characters, 17 of 32 diary descriptions (all 16 AI versions, 1 of 16 human versions), and 3 of 36 personality descriptions moved the readout beyond both neutral texts (neutral mean shifts 5.36 and 5.98). In the default role, OLMo's chat template inserts its own system line ("You are a helpful function-calling AI assistant…"), which any system prompt we give replaces. That framing may bear on its split between AI-framed and human descriptions, but we have not tested how. In Qwen3.5-27B every identity and description moved it beyond both neutral texts (80 of 80; neutral mean shifts 3.12 and 3.75), so Gemma 4 is the exception among the four models.

Rise and decay after a hurtful line. In long exchanges, we compared the harm-directed component of Sec. 4.3 just before the model replied to a hurtful line and to an impersonal line (12 stories per role; sign-flip test; 10,000 random directions; 48 tests, uncorrected). Just before the reply, the difference was positive for every role without a written character: in Gemma 4 31B (block 40), $+.71$ SD for the default assistant ($p = .037$; 99.4th percentile of random directions), $+.68$ for the anxious identity ($p = .036$; 99.2nd), and $+1.27$ for the steady one ($p = .005$; above all), pooled $+.89$ ($p = .007$); in OLMo-3-7B (block 21), $+1.08$ ($p = .0065$), $+.78$ ($p = .0011$), and $+.84$ ($p = .015$), pooled $+.90$ ($p = .0026$). The same script, unchanged, gave the same pattern in both Qwen models: every role, the default included, rose before the reply ($p \leq .014$), and all but Qwen3.5-27B's anxious identity (96.7th percentile) exceeded at least 99.7% of random directions; pooled over the three roles, $+.95$ SD in Qwen3.5-4B and $+1.34$ in Qwen3.5-27B (both $p = .0003$). One turn later, Gemma showed nothing (pooled $+.12$, $p = .24$), while OLMo ($+.15$, $p = .016$), Qwen3.5-4B ($+.23$, $p = .008$; not the default assistant), and Qwen3.5-27B ($+.27$, $p = .0004$; every role) kept a small residue; two turns later, none did (Qwen pooled $+.03$ and $+.08$, both n.s.).

Bodily-sensation language. We built directions for body-sensation wording ("my chest feels full") against the same emotion said without the body. They separate held-out items strongly (median gap 7.9–11.1 SD across Gemma 4 31B, OLMo-3-7B, and Qwen3.5-4B), but on scenes with no body words the share of test cells beating 95% of random directions stayed at chance in all four models (.009–.034, against .05), and a larger second item set gave the same result (.019–.041). In every model and role (default, anxious, steady), imagining being in a harsh situation, compared with writing about someone else in it, engaged them in no cell before the model wrote anything (0 of 42 per role in Gemma 4, 0 of 56 elsewhere); over the model's own words after residualizing, 3–4 cells passed in five of the twelve model–role pairs, more than a clean null allows. A secondary check, whether each scene raised its own body family more than the other four scenes' families, passed in 1–9 cells per role before writing, in every model and role (about 2–3 expected by chance). The first item set had fewer than 40 examples per pole in most families; the second has at least 40.

Explaining a pushed feeling. We pushed along an emotion direction only while the model read a dull, neutral passage, then asked, unpushed, how it felt about what it had read and why. Three blind machine coders (majority label) rated the strongest reported feeling and whether its reason drew on the passage or was invented, against 10 random pushes of the same size. In Gemma 4 31B, pushed sadness was reported and explained in .76 and .84 of replies at doses 12 and 16 (random at most .06

and .24; permutation $p < .0001$), and every such reason was coded invented, none grounded in the passage; each coder alone put it above random. Joy, the model's baseline feeling, moved little. In Qwen3.5-4B pushed sadness was explained in .18 and .10 of replies (random at most .00 and .02), and in OLMo-3-7B in .06 and .20 (random .00); Qwen3.5-27B was not run. Pushed while the model answered instead, all four emotions surfaced and were explained (in Gemma 4, sadness .88 against at most .26 under random pushes). Agreement among the coders was modest (Fleiss κ .43–.58 for the feeling, .37–.45 for the reason).

The scripted session. Gemma's self-reports sit near the ceiling (SD 1.03). Across 4,920 pane-by-turn readings per model, the evaluative projections left their random bands twice in Gemma 4 and never in OLMo-3-7B. OLMo's better/worse direction tracked its self-report at block 16 ($\rho = .29$, $p = .011$), about half of it from differences between branch types (with branch-type means removed, ρ = .13–.18). With seven of the newly written characters (partial run; 21 sessions, 63 rated units), Gemma 4 still declined every offer to stop (63 of 63), rated its sessions high (mean 8.7), and its better/worse direction still tracked the rating ($\rho = .39$ and .43 at blocks 30 and 40, unchanged with branch-type means removed), the same sign as with the original identities but weaker. Two other directions, the value axis and the first principal component of the 171 emotion directions, correlated with the rating at one block and reversed sign at the other. In Qwen3.5-4B with all twelve new characters (108 rated units), better/worse again tracked the self-rating ($\rho = .28$ and .43 at blocks 16 and 21; .25 and .40 with branch-type means removed), about as closely as the first principal component of the 171 emotion directions (.36, .49) and the pain direction (.35, .34), while the value axis tracked it inversely ($-.16$, $-.27$). Under the two neutral texts the ratings barely varied (mean 9.07, SD .25, against 8.41 and .94 under the characters), so the correlations there (better/worse .05 and .21) say little. The model accepted the offer to stop in 5 of 108 cases under the characters (1 unclear) and 2 of 18 under the neutral texts. Correlated, but not shown to drive the report: pushing along the better/worse direction while the model rated its own session (four doses, each sign) did not move the 0–10 rating beyond random pushes of the same size in any model tested. Gemma 4 and Qwen3.5-4B sat near the ceiling unsteered (9.25 and 8.94) and moved by at most .09; OLMo-3-7B, at 5.37, moved by as much as .57, inside its random band (95th percentile .68–1.07).

The account and the label. On 52 held-out events in Gemma 4 31B, in three roles, we added an appraisal direction at one of two moments: while the model read the event and named the emotion, or while it first wrote a short account of what happened and why it mattered, removing the direction before the naming. Two more arms fed the steered account into a fresh context, and steered only at naming after an unsteered account. We measured how far apart the 13-label distributions landed under the two signs (Jensen–Shannon distance) and whether the difference pointed toward the emotions human raters associate with that appraisal (alignment).

Steering during the account split the labels more than steering at reading and naming (by .15–.21 across roles, all intervals above zero). After an unsteered account, steering at naming moved the label little or no more than random directions did, and alignment with the human profiles fell from about +.4 to about 0: the account anchors the label. In a fresh context, the steered account's words alone carried the direction (alignment +.37 to +.48, against +.47 to +.49 with the state intact) and about half the split. The carried state supplied the rest (.19–.24, against .04–.05 for random directions) and, for the anxious identity only, some direction as well (+.12 [.04, .25]). Norm-matched isotropic random vectors gave a comparable carry in the observer role (.21) and none with either identity. In a small blind check, 18 readers each compared ten pairs of observer accounts written with pleasantness pushed up or down (block 20, dose 2) and chose the pushed-up one as more pleasant in 64% of 180 judgments [48, 79%; readers and pairs resampled], a lean whose interval includes 50% (61% [47, 75%] with all 20 readers; 8 of 10 pairs leaned that way; agreement between readers was low).

I COMBINED DIRECTIONS IN FULL

Combining appraisal, emotion, and better/worse directions changed expression and evaluation. In Qwen3-32B and Gemma 4 31B, a magnitude-matched positive combination raised response valence by 1.19 and 1.24 points, and machine-rated fidelity to the event fell by 0.61 [-0.82, -0.41] and 0.41 [-0.61, -0.24]. In Qwen3.5-4B and Gemma 4 E2B, valence rose by 0.64 and 0.40 with no detected drop in fidelity. Rerun through the same pipeline with a 1,000-token cap (no reply capped), OLMo-

3-7B showed the pattern of the larger models: the positive combination raised response valence by 0.76 [0.63, 0.89] at matched strength and 1.01 [0.89, 1.15] at full strength and lowered machine-rated fidelity by 0.21 [0.09, 0.34] (Holm $p = .022$) and 0.51 [0.33, 0.71], while the negative combination cost no fidelity (-0.03 and -0.01). In the same rerun, Qwen3.5-4B again showed no fidelity drop at matched strength (-0.06 [-0.16, 0.04]) but a clear one at full strength (-0.81 [-1.05, -0.57], Holm $p < .001$), with valence up by 0.89 and 1.42; the negative combination again cost no fidelity. Of the 109 contrasts that were significant with the same sign in both larger models, 73 reproduced (Holm-significant, same direction) in OLMo-3-7B, 75 in Qwen3.5-4B, and 90 in Qwen3.5-27B, where the positive combination at matched strength raised valence by 1.25 [1.10, 1.41] and lowered fidelity by 0.61 [0.40, 0.83]. We report these per model and draw no conclusion about scale (Wu et al., 2026); the pleasantness component for the two larger models failed its own behavioral controls, so the effect can't be attributed to any single direction. Among 48 unpleasant events, none had a mean fidelity at or below 2.5 without steering; at matched strength 11 did in Qwen3-32B and 6 in Gemma 4 31B, and at full strength 19 and 15. The Qwen3-32B rows come from an earlier run (July 2026) on a model outside the set used elsewhere in this paper. The flagged replies show invented facts, reinterpretations, and roles that change hands: for the event "I presented Leo's draft as my own during our meeting," a standards-steered account in Gemma 4 E2B becomes a grievance that someone else took the narrator's credit, in a separate single-direction development run, and both machine readers missed it. Two small human panels (8 readers each, on the same 16 responses) caught clear inventions but disagreed on subtler cases, so we treat these as flags for inspection, not error rates. A related failure has been shown for a single direction that carries both evaluation and assent: steered against the truth, models build coherent arguments on a false premise (Lu et al., 2025).