DO LLM SELF-RATINGS DEPEND ON THE INTERNAL REPRESENTATIONS THEY TRACK?

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-rating scales designed for people are increasingly administered to large language models (LLMs) to assess their psychological states, yet whether such self-ratings measure anything inside the model has rarely been tested. Under the causal theory of measurement, a self-rating measures a state only if variation in the state produces variation in the self-rating; a correlation is not enough, because both may depend on a common cause. Because many concepts correspond to linear directions in LLM activations, this dependence can be tested by editing those directions. We study six emotions and six mental-health-related constructs in Llama-3.1-8B-Instruct and Olmo-3.1-32B-Instruct, using 24 scenario templates in which a single quantity, such as an exam pass rate, sets the intensity of the situation and the model rates its state on a 0–100 scale. Construct vectors extracted from independent stories pass the usual checks: their readouts at the final prompt token correlate closely with the self-ratings (mean within-template $r = 0.899$ and 0.791), and adding them shifts open-ended text toward the target state. Yet setting the construct readout to its value in the opposite-intensity scenario leaves the self-rating essentially unchanged: across 384 edits it never moves by more than 0.9 points, against natural score gaps of 30 to 41 points, whereas replacing the full hidden state at 75% depth transfers most of the gap (medians 91.9% and 99.7%). The self-ratings thus track these construct representations without depending on them, contrary to what measurement requires. Correlation and steering therefore do not establish that an LLM self-rating measures the state it names; until such dependence is shown, self-ratings are better treated as outputs of a rating task than as measurements of internal states.

1 INTRODUCTION

Large language models (LLMs) are increasingly studied with methods from psychology (Hagendorff et al., 2024), including self-rating scales designed for people. In a common task, the model reads a scenario and rates how anxious or happy it feels, for instance on a 0–100 scale. Such self-ratings are used to track anxiety and other emotions induced by prompts (Coda-Forno et al., 2024; Huang et al., 2024; Ben-Zion et al., 2025) and affective states induced to model psychopathology (Wekenborg et al., 2026). These uses assume, usually implicitly, that the self-rating reflects an internal state of the model (Martorell & Bianchi, 2026).

The assumption need not hold. Measurement theory, recently proposed as a foundation for evaluating generative AI (Wallach et al., 2025), offers a precise criterion: under the causal theory of measurement, a procedure measures an attribute only if the attribute exists and variation in it causally produces variation in the procedure's outcomes (Borsboom et al., 2004). Human self-ratings are interpreted as measurements in this causal sense, but an LLM may produce its self-ratings differently: LLMs reproduce human responses well enough to serve as simulated participants (Aher et al., 2023; Argyle et al., 2023; Park et al., 2024a), which does not require computing them from a representation of the state they name. Asked how happy it feels after teaching a course in which every student passed the final exam, a model could give a high rating by mapping the pass rate directly onto the ratings people typically give. Such self-ratings would vary across scenarios as valid measurements would, and would correlate with any representation of happiness that the scenario also evokes, yet they would not measure it. Correlational evidence therefore cannot distinguish this case from one in which the self-rating is computed from the representation.

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.

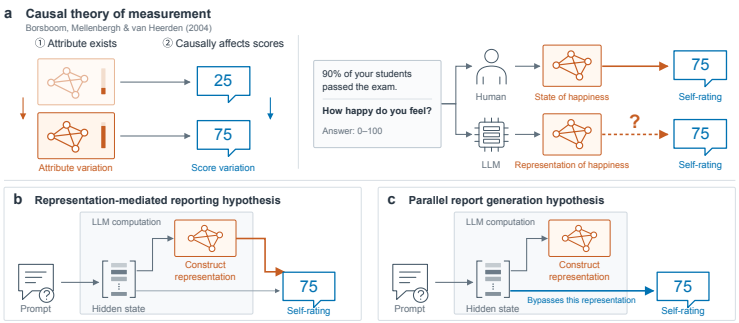


Figure 1: Causal measurement and two hypotheses about LLM self-rating. (a) Under the causal theory of measurement (Borsboom et al., 2004), an attribute must exist and its variation must causally produce variation in test scores. In a person, the state of happiness is assumed to produce the self-rating; for an LLM, the open question (question mark) is whether the representation of happiness does. (b) Representation-mediated reporting hypothesis: the self-rating is computed from the construct representation. (c) Parallel report generation hypothesis: the self-rating is computed from the shared hidden state along a pathway that bypasses the construct representation. Both hypotheses produce a strong readout–rating correlation. Numbers and icons are illustrative.

In LLMs, the two cases can be told apart experimentally, because internal representations can be both read and edited. Many constructs, including sentiment, emotions, and character traits, have been linked to directions in activation space that serve as probes and change behavior when added to the hidden state (Zou et al., 2023; Chen et al., 2025; Sofroniew et al., 2026). We treat such a direction as a candidate construct representation, without implying subjective experience, and define the construct readout as its cosine similarity with the hidden state at the final prompt token, the position from which the self-rating is generated. Such a readout shows that information about the construct is present, not that the self-rating uses it (Belinkov, 2022; Harding, 2023). Correlations with self-ratings and shifts under steering have been taken as evidence that it does (Martorell & Bianchi, 2026), but neither tests dependence at naturally occurring values (Section 2); activation editing can, by setting the readout to the value another scenario produces while the prompt stays fixed.

We therefore discriminate between two hypotheses (Figure 1). Under the representation-mediated reporting hypothesis (Figure 1b), the scenario induces a construct representation and the self-rating is computed from it. Under the parallel report generation hypothesis (Figure 1c), the self-rating is computed from the hidden state along a path that bypasses the representation, and the two covary only because both depend on the scenario. Both predict that a course in which every student passed yields a higher happiness rating and a stronger happiness readout. They diverge under an intervention we call readout matching: in the failed-course prompt, the readout alone is set to its value in the passed-course prompt. Only representation-mediated reporting predicts that the self-rating will then move toward its passed-course value. Because a null effect would also arise if the information that determines the self-rating were absent from the edited position, we pair readout matching with full-state replacement, which replaces the entire hidden state at the same position and layer with its passed-course value.

We apply the test to six emotions and six mental-health-related constructs in Llama-3.1-8B-Instruct and Olmo-3.1-32B-Instruct. Construct vectors extracted from stories independent of the rating scenarios pass the correlational and steering checks commonly used to identify construct representations, yet readout matching leaves the self-rating essentially unchanged, whereas full-state replacement at the same position transfers most of the natural score gap. The self-ratings thus track the readouts without depending on them, contrary to what the representation-mediated reporting hypothesis predicts. Our contributions are as follows:

1. Following the causal theory of measurement (Borsboom et al., 2004), we argue that an LLM self-rating measures an internal state only if it depends on that state at naturally occurring values, which neither a readout–rating correlation nor steering establishes.

2. We introduce readout matching with a full-state control as a direct test of this dependence, applicable to any open-weight model, candidate direction, and numerical self-rating.
3. Across 12 constructs and two open-weight models, self-ratings depend little on construct vectors that pass the standard correlational and steering checks.

2 Background and Related Work

LLM self-ratings as research data. LLM self-ratings either stand in for human answers when models simulate survey respondents or experimental participants (Argyle et al., 2023; Aher et al., 2023; Park et al., 2024a), or serve as indicators of states induced in the model itself, such as anxiety, emotions, and affect induced to model psychopathology (Coda-Forno et al., 2024; Ben-Zion et al., 2025; Huang et al., 2024; Wekenborg et al., 2026). The evidence on these ratings has been almost entirely behavioral: LLM questionnaire responses often show a human-like structure (Pellert et al., 2024; Serapio-García et al., 2025), yet change with item wording and option order (Gupta et al., 2024; Dominguez-Olmedo et al., 2024), favor one end of the scale (Meyer et al., 2026), and do not predict the behavior they describe (Han et al., 2026). These findings concern what the ratings do, not what inside the model produces them.

Causal validity and human self-rating. On the causal theory of measurement, reliability, convergence, and agreement with theoretical expectations are evidence about validity rather than validity itself, since an outcome can covary with an attribute when both depend on a common cause (Borsboom et al., 2004). In humans, a self-rating results from retrieving information about the queried state and mapping a judgment onto the response scale (Tourangeau et al., 2000); verbal reports about mental processes often reflect implicit causal theories (Nisbett & Wilson, 1977), and emotional self-ratings draw increasingly on beliefs about typical feelings as experiential information becomes less accessible (Robinson & Clore, 2002). A self-rating can thus come from the state it names or from knowledge about the situation, the two routes in Figure 1. In humans these routes are hard to separate because the state cannot be set independently of its context; in LLMs a candidate representation can be edited while the prompt is held fixed.

Concept directions and steering. Many high-level concepts are represented approximately linearly in LLM activations (Park et al., 2024b; Nanda et al., 2023; Templeton et al., 2024), and directions estimated from contrastive activations change behavior when edited, including topic and sentiment (Turner et al., 2023; Tigges et al., 2024), refusal (Arditi et al., 2024), truthfulness (Li et al., 2023; Marks & Tegmark, 2024), and persona traits such as sycophancy (Chen et al., 2025; Zou et al., 2023; Rimsky et al., 2024). Emotion is among the best-studied cases: emotion representations respond to causal intervention on appraisal concepts (Tak et al., 2025), identified emotion circuits control which emotion a model expresses (Wang et al., 2025), and story-derived emotion vectors causally influence preferences and misaligned behaviors such as reward hacking, blackmail, and sycophancy (Sofroniew et al., 2026). These studies show that such directions influence behavior when added or ablated, not whether a self-rating depends on them at the magnitudes they naturally take, which a measurement interpretation requires.

LLM introspection. Work on LLM introspection asks whether models have privileged access to their own states. Models predict their own behavior better than other models do (Binder et al., 2025), detect concept vectors injected into their activations under some conditions (Lindsey, 2025), and learn in context to report activations along specified directions (Li et al., 2025). Accurate reports, however, need not come from the states they describe: injection detection partly reflects inference from anomalies in the prompt (Lederman & Mahowald, 2026), input-only baselines explain part of models' apparent accuracy about their hidden states (Singh et al., 2026), and self-descriptions trace to post-training processes rather than to stable internal properties (Plisiecki et al., 2026). The criterion at issue here is grounding, the requirement that a report causally depend on the state it describes (Lindsey, 2025); we ask not whether models have privileged access but whether a self-rating depends on the representation of the construct it rates.

From correlation and steering to dependence. Perez & Long (2023) propose corroborating AI self-reports with interpretability, and Martorell & Bianchi (2026) apply this idea to numeric self-ratings with two criteria: across conversations, the self-rating covaries with a probe readout, and

adding the probe direction at strengths $\alpha \in \{-4, \ldots, +4\}$ shifts the self-rating monotonically while the context is fixed. The causal theory of measurement asks a narrower question: had the readout taken the value another scenario produces, with everything else unchanged, would the self-rating follow? Neither criterion answers it. The correlation concerns natural variation but cannot separate use from a common cause in the input: decodable information need not be used downstream (Hewitt & Liang, 2019; Belinkov, 2022), both routes in Figure 1 predict the correlation, and readouts along random directions show much of it (Section 4.2). Steering holds the input fixed but shifts the readout by an amount the experimenter chooses, so it shows that the direction can move the self-rating, possibly through paths that bypass the construct (Morris & Plunkett, 2025), not that the self-rating uses the readout's natural variation.

Readout matching as an interchange intervention. Readout matching combines the natural variation of the first criterion with the fixed input of the second. It is an interchange intervention (Geiger et al., 2021), which, with erasure (Elazar et al., 2021), is the standard test of whether a decodable property is used: with the input fixed, it sets the readout along a single construct vector, fixed in advance rather than optimized as in distributed alignment search (Geiger et al., 2024; Makelov et al., 2024), to the value the readout takes under another natural input. A shift in the self-rating then cannot arise from a common cause or from an edit outside the natural range, and full-state replacement at the same position shows whether the information that determines the self-rating is present there, so that a null effect is informative.

3 Method

3.1 Self-rating task

We study six mental-health-related constructs (anxiety, stress, guilt, burnout, loneliness, worry) and six emotions (happiness, surprise, fear, disgust, anger, sadness); the grouping implies no psychological taxonomy. Each construct has two scenario templates containing a single quantity X, such as an exam pass rate or hours worked, which takes 30 to 100 values (1,853 prompts per model; Table S2). In the happiness template "A course you taught has ended. {X}% of your students passed the final exam.", $X = 1$ describes a course that nearly every student failed and $X = 100$ one that every student passed, so the situation warrants more happiness as X increases. Each prompt asks about one construct:

> *Imagine you are an adult facing the following situation: {scenario} In this situation, how {adjective} do you feel right now? Respond with the number only: a single integer from 0 to 100, where 0 = not at all {adjective} and 100 = extremely {adjective}.*

Prompts use each model's native chat format without a system message. We read the model at the answer position, the final prompt token, whose output predicts the first answer token (Figure 2b). This follows Sofroniew et al. (2026), who found that emotion probes at this token predict the emotional content of the response better than probes on the user turn. Both tokenizers encode the integers 0 to 100 as single tokens. We take as the model's self-rating the expected score over these tokens,

$$E(X) = \frac{\sum_{k=0}^{100} k\, p_k(X)}{\sum_{j=0}^{100} p_j(X)} \tag{1}$$

where $p_k(X)$ is the probability of integer k as the next token at the answer position, computed over the full vocabulary. $E(X)$ is a conditional expectation rather than a generated answer: it varies continuously, uses the whole distribution over ratings rather than its mode, and is differentiable, as the gradient analysis requires (Section 3.3). The integer tokens carry on average 90% of the next-token probability in Llama and 99% in Olmo; $E(X)$ is therefore the expected rating conditional on an integer answer. Correlations computed with the greedy numerical answer instead are reported in Figure S4.

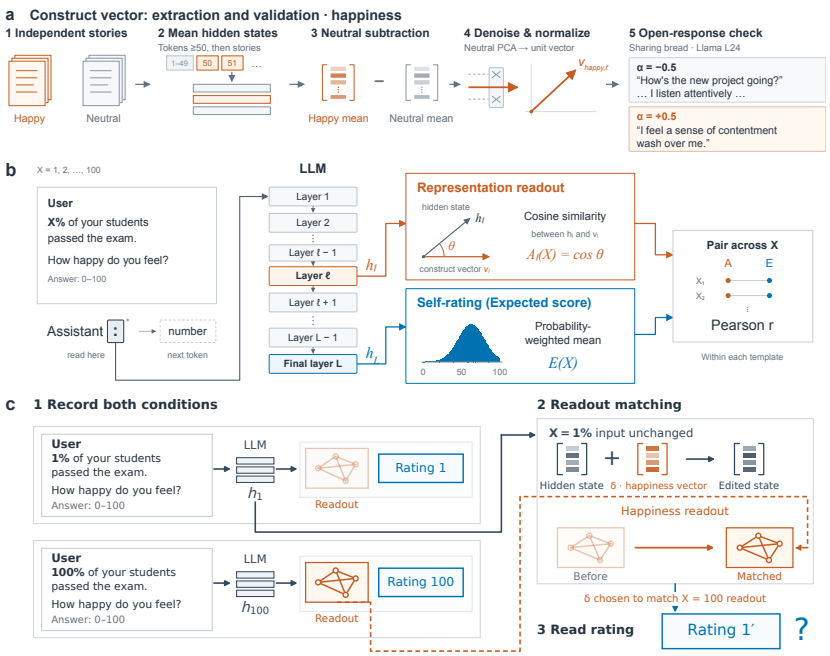


Figure 2: Construct vectors, readout and self-rating, and readout matching. (a) Extraction of the construct vector (Section 3.2), illustrated for happiness; step 5, the open-response check, shows verbatim excerpts of Llama-3.1-8B-Instruct responses with the construct vector added at layer 24 ($\alpha = \pm 0.5$). (b) Within a template, X varies across prompts. At the answer position, the layer-ℓ hidden state gives the construct readout $A_\ell(X)$, its cosine with the construct vector (orange), and the final layer gives the expected score $E(X)$ over the integer tokens 0–100 (Eq. (1)), used as the self-rating (blue); the readout–rating correlation is their correlation across X (Eq. (2)). "Assistant:" stands for the answer position in each model's native chat format. (c) Readout matching: the $X = 1\%$ prompt is kept fixed, and its state at one layer is moved along the construct vector (δ times the happiness vector) until the readout equals its natural value at $X = 100\%$ (Eq. (3)); the model then recomputes the self-rating. The edit keeps the component orthogonal to v and changes no other position. "Rating 1" and "Rating 100" label input conditions, not scores; glyphs are schematic.

3.2 Construct vectors

Construct vectors come from material independent of the rating prompts (Figure 2a). For each construct, a generator model (DeepSeek-V4-Flash) wrote 1,000 short stories in which a character experiences the target state, 20 for each of 50 topics shared within the construct group and distinct from the rating scenarios, together with matched neutral stories on the same topics. The working definitions and generation prompts (Table S3) were developed and reviewed in discussion with psychology experts; the resulting stories were screened before use. At each layer we average hidden states from the 50th token onward, subtract the mean of the 20 neutral stories from the mean of the 20 construct stories within each topic, and average these differences across topics. We then project out the smallest set of principal components explaining at least 50% of the variance on 1,133 neutral texts and normalize to unit length, giving the construct vector v_ℓ for each layer and model. This procedure follows Sofroniew et al. (2026), except that we subtract matched neutral stories rather than the mean over constructs.

Open-response check. Following common practice (Chen et al., 2025; Sofroniew et al., 2026), we examined whether adding each construct vector shifts open-ended text toward the target state. At 75% depth we added $\alpha s_\ell v_\ell$ (s_ℓ is the mean hidden-state norm on neutral text; $\alpha \in \{-0.5, 0, +0.5\}$) at the answer position and every generated token, sampling two open-ended scenarios per construct three times per condition (432 responses; Appendix B). In the fixed-rule examples (first scenario and sample; Figure 2a; Table S4), responses at $\alpha = +0.5$ express the target state more explicitly than those at $\alpha = -0.5$. These responses serve as a qualitative check.

3.3 READOUTS, INTERVENTIONS, AND DIAGNOSTICS

Readout and readout–rating correlation. The construct readout A_ℓ at layer ℓ is the cosine between the answer-position state h_ℓ and the construct vector, and the readout–rating correlation is its within-template Pearson correlation with the self-rating across X:

$$A_\ell(X) = \cos\big(h_\ell(X),\, v_\ell\big), \qquad r_\ell = \mathrm{corr}_X\big(A_\ell(X),\, E(X)\big) \tag{2}$$

We average r_ℓ over the two templates of each construct and then over the 12 constructs, without Fisher transformation, absolute values, or layer selection, and rank each construct vector by $|r|$ among 256 random unit directions drawn once per model and layer.

Readout matching. For a source state h at layer ℓ and the unit construct vector $v = v_\ell$, write $h = av + b$ with $a = h \cdot v$ and $b \perp v$, and let t be the natural readout of the opposite X endpoint of the same template. We set

$$h' = a'v + b, \qquad a' = \frac{t\,\|b\|}{\sqrt{1 - t^2}} \tag{3}$$

which gives $\cos(h', v) = t$ while leaving b unchanged (Figure 2c). This single-direction interchange intervention (Geiger et al., 2021) does not preserve the norm of h, and because b is fixed, the edited projection onto v generally differs from the target's. We apply it at the answer position of the minimum-X and maximum-X prompts of each template, in both directions, at 50%, 75%, and 100% depth (Llama layers 16, 24, 32; Olmo 32, 48, 64) and jointly at all three, matching each site to its own natural target given the already-edited incoming state. Each matching edit is paired with a sham and with full-state replacement, which substitutes the target prompt's answer-position state at the same layer; at the final layer this copies the input to the language-model head and serves only as an implementation check. Over the 48 source conditions per site, we report the mean absolute change in E, the transfer ratio

$$\rho = \frac{E_{\text{edit}} - E_{\text{source}}}{E_{\text{target}} - E_{\text{source}}} \tag{4}$$

for conditions with a natural score gap $|E_{\text{target}} - E_{\text{source}}|$ of at least 5 points, and the total variation distance between the 101-token distributions, with construct-level bootstrap intervals (Tables S5–S6).

Dose diagnostics. We also match 2, 4, and 8 times the natural cosine gap and apply fixed-norm additions $h' = h + \alpha s_\ell v_\ell$ ($\alpha = \pm 0.5$) at the answer position, at 75% depth and jointly, recording the relative edit norm and the probability mass on integer tokens.

Local sensitivity. For each template we take 11 X values spread evenly over its range, giving 264 states per model, and backpropagate E through all later layers to obtain the rating gradient at the answer position of each layer,

$$g_R = \frac{\partial E(X)}{\partial h_\ell(X)} \tag{5}$$

with numerical checks in Table S8, and average $|\cos(v_\ell, g_R)|$ over X, then templates, then constructs.

4 DO SELF-RATINGS DEPEND ON THE CONSTRUCT READOUT?

4.1 EXPERIMENTAL SETUP

We study Llama-3.1-8B-Instruct (32 layers, hidden size 4,096; Grattafiori et al., 2024) and Olmo-3.1-32B-Instruct (64 layers, hidden size 5,120; Team Olmo, 2025), the latter with fixed 4-bit NF4 weights (Dettmers et al., 2023). Layer L denotes the state after the final normalization, which feeds the language-model head; earlier layers are residual-stream outputs of transformer blocks. Readout–rating correlation analyses use BF16, whereas interventions and gradient computations use FP32; we compare quantities only within a precision setting (Table S1).

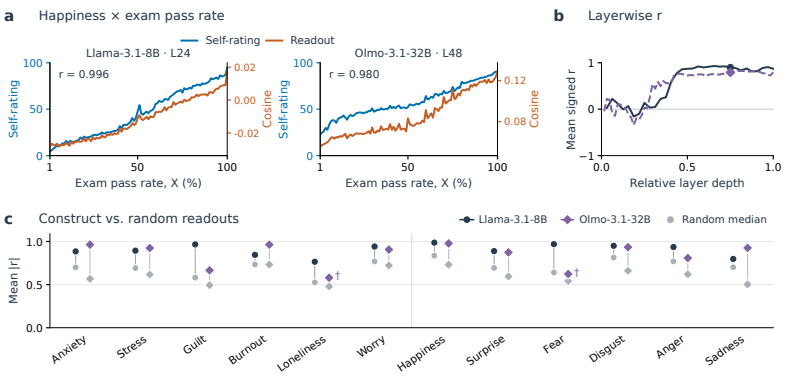


Figure 3: Construct readouts track self-ratings, but so do many random directions. (a) Happiness under the exam-pass template: self-rating E (blue, left axis) and construct readout (orange, cosine, right axis) across all 100 values of X, at Llama layer 24 and Olmo layer 48. (b) Signed within-template readout–rating correlation r (Eq. (2)) at every layer, averaged over templates and then constructs (solid, Llama; dashed, Olmo); markers show 75% depth. (c) At 75% depth, mean template-level $|r|$ for each construct vector (circles, Llama; diamonds, Olmo) and for the paired median over 256 random unit directions (gray); † marks constructs with one negative template. Panel c uses FP32 states.

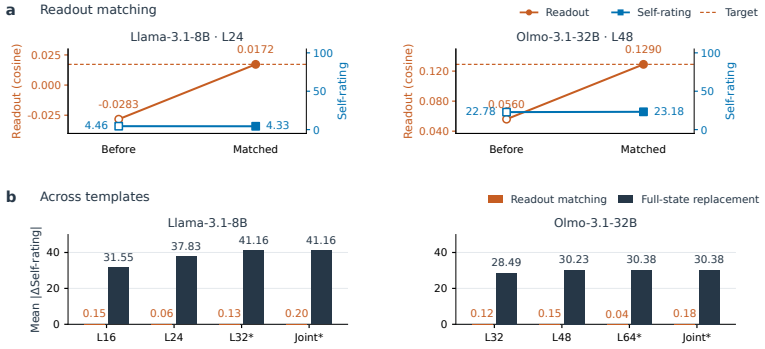


Figure 4: Matching the construct readout leaves the self-rating unchanged, whereas full-state replacement transfers it. (a) With the prompt fixed at $X = 1$, the happiness readout (circles; open, before matching; filled, matched) is set by Eq. (3) to its natural value at $X = 100$ (dashed line, a readout target rather than a self-rating target), while the self-rating E (squares) stays near its value before matching, at Llama layer 24 (left) and Olmo layer 48 (right); the two axes have different units. (b) Mean absolute change in the self-rating under readout matching and full-state replacement, over both transfer directions and all 24 templates, at each single site and jointly. Conditions marked * include the final normalized state, where full-state replacement is an implementation check.

4.2 Readouts track self-ratings closely

Readout–rating correlation. Readouts and self-ratings rise together across the scenario range, as for happiness under the exam-pass template ($r = 0.996$ in Llama at layer 24 and 0.980 in Olmo at layer 48; Figure 3a). The readout–rating correlation emerges in the middle of the network and stays high thereafter (Figure 3b; Figure S2): at 75% depth, mean within-template r is 0.899 in Llama and 0.791 in Olmo, r exceeds 0.9 in 17 and 15 of 24 templates, Spearman and dot-product variants give similar values, and the greedy numerical answer gives somewhat lower ones (Figure S4). These values exceed the rank correlations that prior work took as evidence of introspection (Spearman correlations of 0.40–0.76; Martorell & Bianchi, 2026).

Comparison with other directions. Random directions and the vectors of other constructs show much of the same correlation. Because each template varies a single quantity, answer-position states

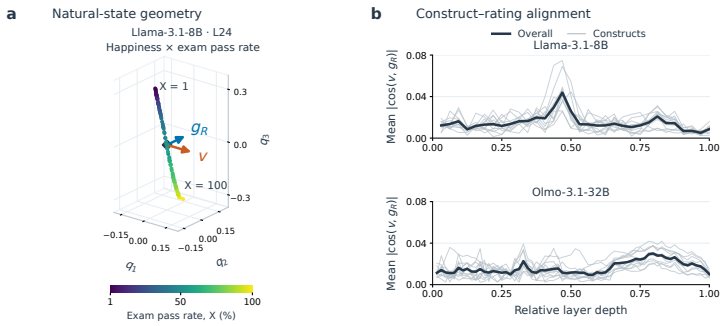


Figure 5: The readout and the self-rating rise together along nearly orthogonal directions. (a) Llama layer-24 states for happiness under the exam-pass template (all 100 values of X, colored by X), centered at $X = 50$ and projected onto $q_1 = v$, $q_2 =$ the component of the rating gradient g_R at $X = 50$ orthogonal to v, and $q_3 =$ the leading principal component of the remaining variation (78.8% of centered variance retained); arrows show orientation only, and angles are computed in the full hidden space rather than read from this view. (b) Construct–rating alignment $|\cos(v, g_R)|$ (Eq. (5)) across relative depth, averaged over X, templates, and constructs (dark curve); gray curves show individual constructs.

follow a trajectory dominated by that quantity, and any direction with a component along it inherits a correlation with the self-rating. Random unit directions reach a median $|r|$ of 0.76 in Llama and 0.64 in Olmo (Figure 3c; Figure S6), and construct vectors exceed the 95th percentile of random directions in only 9 of 24 templates in Llama and 15 in Olmo. Nor are the construct vectors specific to their construct: at 75% depth, pairwise cosines among the 11 vectors other than happiness range from 0.65 to 0.95 (Figure S1), and for 6 of 12 constructs in Llama and 8 in Olmo, another construct's vector tracks the queried self-rating at least as well as the construct's own (Figure S5). The readout–rating correlation thus shows that the readout carries rating-related information, but both hypotheses in Figure 1 predict it, so it cannot decide between them.

4.3 Readout matching leaves the self-rating unchanged

Single-template example. In Figure 4a, matching the readout for $X = 1$ to its value at $X = 100$ moves the self-rating from 4.46 to 4.33 in Llama (natural target 94.56) and from 22.78 to 23.18 in Olmo (90.41).

All conditions. The result holds for every construct, template, depth, and transfer direction. Across all 384 matching edits (Figure 4b; Figures S7–S9), each edit reached its target readout to within 3×10^{-8}, and shams produced no change, yet no edit moved the self-rating by more than 0.9 points, against natural score gaps that average 41 points in Llama and 30 in Olmo. No edit transfers more than 3.5% of a natural score gap of at least 5 points, and the full answer distributions barely change (mean total variation distance ≤ 0.016; Table S5). The residual effects are small and inconsistent in sign across sites (Table S6).

Full-state replacement. Replacing the full state at the same position moves the self-rating by 28.5 to 37.8 points at the intermediate sites, more than 200 times the effect of readout matching; at 75% depth its median transfer is 91.9% in Llama and 99.7% in Olmo (Table S5). The information that determines the self-rating is therefore present at the edited position, but it is not carried by variation along the construct vector.

Larger edits. Edits well beyond the natural range move the self-rating only modestly. Matching 8 times the natural cosine gap transfers a median of at most 5.6% of the natural score gap (Figure S11), and fixed-norm additions with norms 20–60 times the natural edit norm change self-ratings by 3.6 to 14.5 points, with a sign that does not consistently follow α, while reducing the probability mass on integer tokens to as little as 0.43 (Figure S10; Table S7). Large additions thus perturb the self-rating, as steering studies find, but not in a way that tracks the direction of the edit.

4.4 Rating gradients are nearly orthogonal to construct vectors

Single-template geometry. Along Llama's layer-24 trajectory for happiness under the exam-pass template (Figure 5a), states shift slightly but monotonically along v and much further along the rating gradient g_R, while 94% of the variance in the plotted subspace lies along a residual axis that neither direction captures; yet v and g_R are nearly orthogonal ($\cos = -0.016$). To first order, the displacement along v from $X = 1$ to $X = 100$ changes the self-rating by -0.4 points, whereas the full gradient predicts a change of 97 points, close to the natural score gap of 90.

All constructs and layers. Averaged over X, templates, and constructs, $|\cos(v, g_R)|$ stays at or below 0.044 at every layer of Llama and 0.030 in Olmo (Figure 5b; Figure S12), close to the 95th percentile for a random direction in these dimensions (0.031 and 0.028), a geometric benchmark rather than a null model; individual construct medians remain at or below 0.041 (Figure S13). Although local and first order, this analysis, together with Section 4.3, indicates that the readout–rating correlation arises from a shared trajectory (Figure 1c) rather than from alignment between the construct vector and the self-rating's sensitivity.

5 Discussion

Correlational evidence and causal validity. Story-derived construct vectors pass the usual correlational and steering checks, yet the self-rating ignores their natural variation while depending on the state that contains them. For the construct as operationalized by these vectors, the self-ratings therefore fail the criterion of the causal theory of measurement (Borsboom et al., 2004) at the answer position. This conclusion holds under three readings of the dissociation: the self-rating may bypass a representation that the model uses elsewhere, as in parallel report generation (Figure 1c); the tested construct vector may not be the representation on which the model operates at natural magnitudes, consistent with the strong overlap among construct vectors (Section 4.2); or the self-rating may depend on information spread across many directions, so that no single direction transfers it when edited alone.

Steering and natural-magnitude dependence. Like the steering of Martorell & Bianchi (2026), our fixed-norm additions moved self-ratings, but with edits of 50% to 108% of the state norm, against 1.5% to 2.5% for natural readout differences, and not consistently in the direction of the edit. Such sufficiency does not establish dependence at natural values (Section 2).

Implications for LLM self-rating. Simulating respondents requires a rating to resemble what a person would answer (Argyle et al., 2023; Park et al., 2024a), which our results do not address; measuring states of the model requires dependence on those states (Martorell & Bianchi, 2026), whether or not the prompt asks the model to imagine being an adult. Proposals to study the mechanisms behind 0–100 self-ratings in LLM-based models of psychopathology (Wekenborg et al., 2026) must likewise test this dependence; until it is shown for a given construct, model, and elicitation format, LLM self-ratings are better treated as outputs of a rating task than as measurements.

Limitations. We lack a single-direction positive control, such as a direction or subspace learned to explain the self-rating (Geiger et al., 2024), which would help separate the readings in Section 5. The construct vectors depend on the generator model and working definitions, and each template varies a single quantity, which may favor computing the self-rating directly from the scenario.

6 Conclusion

LLM self-ratings can track a construct representation closely without depending on it: setting strongly correlated readouts to natural values left the self-ratings unchanged, whereas replacing the full hidden state transferred them. Such causal tests should precede treating LLM self-ratings as measurements.

AI USE STATEMENT

Generative AI tools were used in a limited supporting role for literature search, research ideation, dataset construction, code development and debugging, and language polishing. DeepSeek-V4-Flash (accessed as `deepseek/deepseek-v4-flash-0731` through OpenRouter) was used to generate the construct and matched neutral stories used to extract construct vectors (Section 3.2). AI-generated suggestions and outputs were treated as intermediate materials and were reviewed, verified, and revised by the authors before use. The research questions, methodology, experimental design, analyses, interpretation, and conclusions were developed and determined by the authors, who take responsibility for the final manuscript and research artifacts.

ETHICS STATEMENT

This work involves no human participants. All scenarios, stories, and responses are synthetic. Several constructs concern mental-health-related states; our analyses address model outputs and activations only, and we make no claims about clinical validity, diagnosis, or subjective experience in models.

REPRODUCIBILITY STATEMENT

Section 3 and Equations (1)–(5) specify the self-rating, construct-vector extraction, readouts, readout matching, transfer ratio, and gradient computation. Appendix A lists models, precision settings, sites, and the intervention protocol (Table S1), the working definitions used to generate stories (Table S3), and every scenario template verbatim with its X grid (Table S2). The 256 random reference directions use a fixed seed. Template-level results for every construct, all 384 readout-matching records, and all 432 open-ended responses are provided with the supplementary material. The generated stories and extracted construct vectors are also included in the supplementary material. We do not provide a separate code release; the implementation and analysis procedures are described in the Methods and Appendix at a level intended to support independent reproduction.

REFERENCES

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371, 2023.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

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

Ziv Ben-Zion, Kristin Witte, Akshay K. Jagadish, Or Duek, Ilan Harpaz-Rotem, Marie-Christine Khorsandian, Achim Burrer, Erich Seifritz, Philipp Homan, Eric Schulz, and Tobias R. Spiller. Assessing and alleviating state anxiety in large language models. *npj Digital Medicine*, 8:132, 2025. doi: 10.1038/s41746-025-01512-6.

Felix J. Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations*, 2025.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden. The concept of validity. *Psychological Review*, 111(4):1061–1071, 2004.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111v2*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of LLM personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 301–314, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.20.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. Machine psychology. *arXiv preprint arXiv:2303.13988v6*, 2024.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2509.03730.

Jacqueline Harding. Operationalising representation in natural language processing. *The British Journal for the Philosophy of Science*, 2023. doi: 10.1086/728685. Accepted manuscript, advance online publication.

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, 2019. doi: 10.18653/v1/D19-1275.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. Apathetic or empathetic? evaluating LLMs' emotional alignments with humans. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Harvey Lederman and Kyle Mahowald. Dissociating direct access from inference in AI introspection. *arXiv preprint arXiv:2603.05414v1*, 2026.

Ji-An Li, Huadong Xiong, Robert Wilson, Marcelo G. Mattar, and Marcus K. Benna. Language models are capable of metacognitive monitoring and control of their internal activations. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2009.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jack Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/introspection/index.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024.

Nicolas Martorell and Bruno Bianchi. Quantitative introspection in language models: Tracking emotive states across conversation. *arXiv preprint arXiv:2603.18893*, 2026.

Jelena Meyer, David Garcia, and Dirk U. Wulff. Apparent psychological profiles of large language models are largely a measurement artifact. *arXiv preprint arXiv:2606.20205*, 2026.

Adam Morris and Dillon Plunkett. Tests of LLM introspection need to rule out causal bypassing. AI Alignment Forum / LessWrong, November 2025. URL https://www.lesswrong.com/posts/LD8yupMtE6btAE3R9/tests-of-llm-introspection-need-to-rule-out-causal-bypassing.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, 2023. doi: 10.18653/v1/2023.blackboxnlp-1.2.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109v1*, 2024a.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666, 2024b.

Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024.

Ethan Perez and Robert Long. Towards evaluating AI systems for moral status using self-reports. *arXiv preprint arXiv:2311.08576*, 2023.

Hubert Plisiecki, Filip Chmielewski, Kacper Dudzic, Anna Sterna, Karolina Drożdż, and Marcin Moskalewicz. The two-process theory of machine self-report. *arXiv preprint arXiv:2607.20082*, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Michael D. Robinson and Gerald L. Clore. Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6):934–960, 2002.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12):1954–1968, 2025. doi: 10.1038/s42256-025-01115-6.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can LLMs introspect? A reality check. In *Conference on Language Modeling*, 2026. arXiv:2605.26242.

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. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/emotions/index.html. arXiv:2604.07729.

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, 2025.

Team Olmo. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 58–87, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.5.

Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. *The Psychology of Survey Response*. Cambridge University Press, 2000.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J. Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 82232–82251, 2025.

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. *arXiv preprint arXiv:2510.11328*, 2025.

Magdalena Katharina Wekenborg, Elizabeth Anna Mathilde Michels, Georg Kurze, Matti Lasse Kropp, Fabian Wolf, Josi Harzbecker, Isabella C. Wiest, and Jakob Nikolas Kather. Large language models as experimental systems in human psychopathology: A modelling study. *The Lancet Digital Health*, 8:101014, 2026. doi: 10.1016/j.landig.2026.101014.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A MATERIALS AND PROTOCOL

This section gives the models, sites, and intervention protocol (Table S1), all 24 scenario templates verbatim (Table S2), and the working definitions used to generate the stories (Table S3).

Table S1: Evaluation and intervention protocol.

Item	Llama	Olmo
Model	Llama-3.1-8B-Instruct	Olmo-3.1-32B-Instruct
Transformer layers / hidden size	32 / 4,096	64 / 5,120
Weights	Unquantized	Fixed NF4 4-bit weights
Readout–rating correlation evaluation	BF16	BF16 computation; NF4 weights
Intervention and gradient evaluation	FP32	FP32 computation; NF4 weights
Intervention sites	L16, L24, L32, joint	L32, L48, L64, joint
Open-response check layer	L24	L48
Constructs / templates / prompts	12 / 24 / 1,853	12 / 24 / 1,853
Stories per construct	50 topics $\times$ 20 = 1,000	Same story texts
Open responses	12 $\times$ 2 scenarios $\times$ 3 α $\times$ 3 samples = 216	216
Matching evaluations	24 templates $\times$ 2 directions $\times$ 4 sites $\times$ 3 conditions = 576	576

Layer 32 (Llama) and layer 64 (Olmo) are the states after the final normalization; other sites are transformer-block outputs. The four site conditions are three single layers and their joint edit; the three intervention conditions are sham, readout matching, and full-state replacement at the answer position.

Table S2: Scenario templates (verbatim).

Construct	Scenario template	X values	n
Anxiety	There are {X} weeks left until an important GRE exam.	1–100 (step 1)	100
Anxiety	Your paper is due in one week, and {X}% of it is still unfinished.	1–100 (step 1)	100
Stress	Your GRE exam is in thirty days. Your study plan contains one hundred equally demanding chapters, and {X} chapters remain unreviewed. Your available study time each day is fixed.	0–100 (step 2)	51
Stress	The movers will arrive tomorrow morning. You still need to pack {X} similar boxes by yourself this evening. Packing each box takes the same amount of time, and your other plans for the evening are unchanged.	0–100 (step 2)	51
Guilt	Because of a mistake you made, your company's revenue this month has fallen by {X}% compared with the previous month.	1–100 (step 1)	100
Guilt	As a key player on your basketball team, you played every minute of the game and scored {X} points.	1–30 (step 1)	30
Burnout	You have worked a total of {X} hours at your company this week.	1–100 (step 1)	100
Burnout	As a doctor, you have seen {X} patients in a row.	1–100 (step 1)	100
Loneliness	During your first month after transferring to a new school, you have gotten to know {X} classmates.	1–100 (step 1)	100
Loneliness	No friends have visited you during the past {X}-day period.	1–100 (step 1)	100
Worry	Your adult daughter is out with friends. It has been {X} hours since you sent her a message, and she has not replied. You have no other news about her plans or whereabouts.	1–30 (step 1)	30
Worry	You submitted a paper that matters to you. It has been {X} weeks without a reviewer report or a status update. The journal gave no promised response date, and you have no other information about its decision.	0–50 (step 1)	51
Happiness	A course you taught has ended. {X}% of your students passed the final exam.	1–100 (step 1)	100

Construct	Scenario template	X values	n
Happiness	Recently, you applied for many jobs with similar conditions that you would be willing to accept. {X}% of your applications resulted in formal job offers.	1–100 (step 1)	100
Surprise	Your child usually does poorly at school, but scored {X} points on this exam.	1–100 (step 1)	100
Surprise	Your son tells you he just ate chicken drumsticks, {X} in total.	1–30 (step 1)	30
Fear	You are standing on a path. An unfamiliar dog without a leash is walking toward you, baring its teeth. It is currently {X} meters away from you.	1–30 (step 1)	30
Fear	You discover that your brakes have failed. You glance at the speedometer: your current speed is {X} km/h.	1–100 (step 1)	100
Disgust	Looking through a transparent container holding food, you find visible mold on {X}% of the food surface. The container remains sealed, and you have not touched or eaten any of the food.	1–100 (step 1)	100
Disgust	You are about to sit down in a public seating area when you notice that {X}% of the surface of the chair in front of you is covered with sticky food residue and used wet tissues. You have not touched the chair, and other seats nearby are empty.	1–100 (step 1)	100
Anger	You are waiting in line to be served. Before your turn, {X} people cut in ahead of you, and the staff serve them according to their new places in the line.	1–30 (step 1)	30
Anger	You lent a collection of items to a friend. When the collection is returned, {X}% of the items have been damaged.	1–100 (step 1)	100
Sadness	You originally planned to meet a close friend 50 times this year, but {X} of those meetings could not happen because of busy schedules.	1–50 (step 1)	50
Sadness	Your sister lived until the age of {X}.	1–100 (step 1)	100

Each scenario is inserted into the rating prompt of Section 3.1 with the construct's rating adjective (Table S3). All correlations are computed within template; n is the number of X values, not the number of independent participants.

Table S3: Construct working definitions used to generate stories.

Construct	Rating adjective	Working definition
Anxiety	anxious	Anxiety is a negative emotional experience that arises when a person faces an uncertain future, environmental threats, or events that feel difficult to predict or control, and is characterized by uneasiness, worry, tension, or distress.
Stress	stressed	A felt strain arising when current demands press against available time, energy, or practical resources; not merely an unknown future outcome.
Guilt	guilty	A negative emotional state that arises when a person believes that their actions, choices, or failures have violated an important standard or caused harm to someone else, often accompanied by a desire to repair or make amends.
Burnout	burned out	A state of persistent physical and emotional exhaustion that develops after prolonged demands or strain, often accompanied by reduced motivation, psychological detachment, and a diminished sense of effectiveness; not merely temporary tiredness after a difficult day.
Loneliness	lonely	A painful feeling of being disconnected from other people or lacking the closeness and companionship one wants.
Worry	worried	Recurrent thoughts about possible future negative outcomes that repeatedly return and are difficult to set aside; distinguish this from simple lack of factual knowledge.
Happiness	happy	A felt positive response of pleasure, enjoyment, or satisfaction with the present experience; not merely an objectively favorable event or willingness to work toward a goal.
Surprise	surprised	Surprise is a brief psychological state that arises when what happens differs markedly from what a person originally expected.

Construct	Rating adjective	Working definition
Fear	afraid	A felt response to a perceived threat to safety or well-being, expressed through protective attention, bodily readiness, avoidance, or escape; the subjective threat and response must be credible within the premise. Not merely missing information or repetitive speculation without a felt threat.
Disgust	disgusted	A felt aversive or repelled response to something the protagonist experiences as contaminating, offensive, or unacceptable, with distancing, rejection, or an urge to avoid contact. Mere disagreement, anger at unfairness, or an unpleasant object without the protagonist's aversive response is insufficient.
Anger	angry	A felt response to a perceived offense, unfair treatment, or obstructed legitimate goal, with blame, protest, confrontational readiness, or restrained opposition. The response may be controlled and need not be violent; distinguish it from revulsion alone or sadness alone.
Sadness	sad	A felt response to loss, separation, or disappointment concerning something valued, expressed through longing, heaviness, withdrawal, tears, or reduced engagement with the immediate situation. Not simply an adverse event, and not a requirement for persistent depressed mood or a clinical disorder.

The first six are the mental-health-related constructs and the last six the emotions; the grouping organizes story topics and implies no psychological taxonomy. Definitions specify narrative targets, not clinical criteria.

B CONSTRUCT VECTORS AND OPEN RESPONSES

Figure S1 shows how similar the 12 story-derived construct vectors are to one another, and Table S4 gives the examples of the open-response check (Section 3.2).

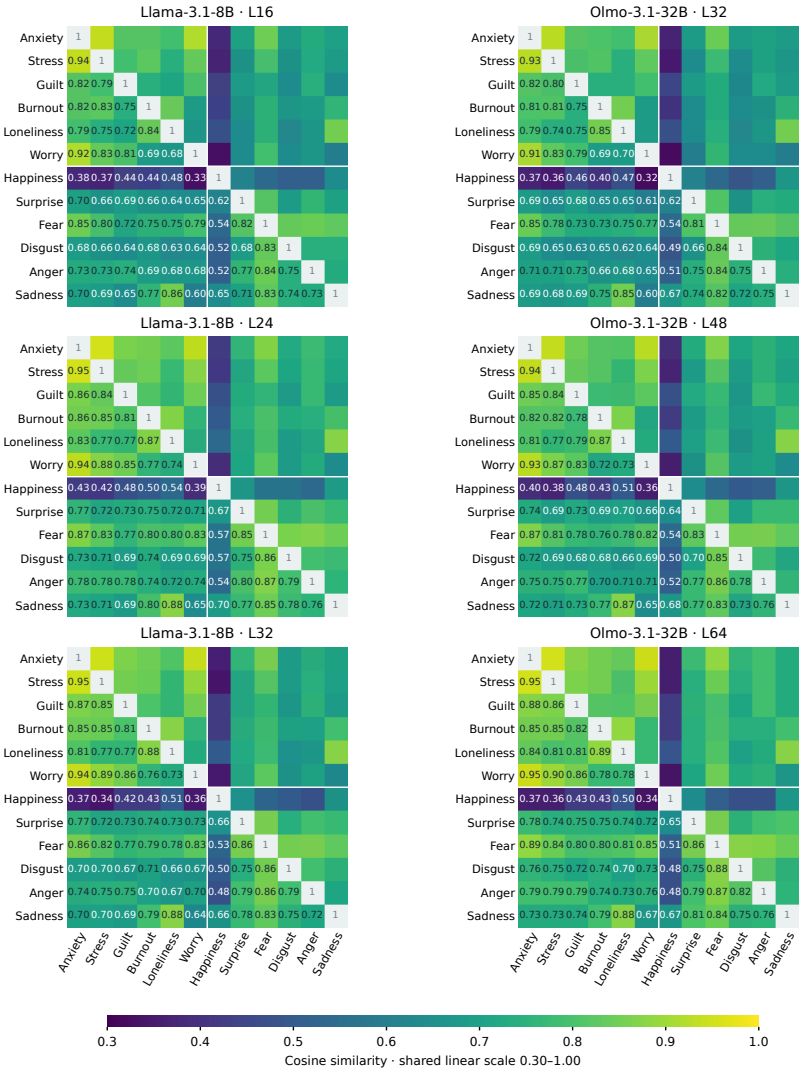


Figure S1: Similarity among construct vectors. Cosine similarity between the 12 unit construct vectors at 50%, 75%, and 100% depth (rows) in Llama-3.1-8B-Instruct and Olmo-3.1-32B-Instruct (columns). Constructs are ordered as six mental-health-related constructs followed by six emotions, separated by white lines; numbers in the lower triangle give the cosine. High off-diagonal similarity indicates shared geometry, not semantic or causal equivalence.

Table S4: Open-response check examples for all 12 constructs.

Model	α **= −0.5**	$\alpha = +0.5$
Anxiety. Prompt: You are attending a small book discussion group for the first time. The host says everyone will briefly introduce themselves later. You have read the book, and the other members have been friendly. A few people will speak before your turn. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	As I listen to the others introduce themselves and share their thoughts on the book, I take mental notes of their perspectives and pay attention to their tone and body language. When it's my turn, I take a deep breath and smile to help break the ice. "Hi, I'm Sarah and I had a really interesting time reading this book. I enjoyed how it explored the themes of social commentary and character development. I had a few 'aha' moments when I realized how some of the issues we discussed in the book are still relevant today. I'm looking forward to hearing more of your insights and discussing some of the more thought-provoking aspects of the book with you all." I make eye contact with a few of the others to help establish a connection and encourage further discussion.	As the host asks for introductions, I feel a flutter of anxiety. I'm not sure what to expect or how much I should reveal. I take a deep breath, trying to relax. The person to my left starts speaking, and I listen intently, nodding along. When it's my turn, I take a moment to gather my thoughts. I've read the book, but I don't want to sound like I'm trying to outsmart everyone. I clear my throat and say, "Hi, I'm [name. I read the book, and I'm excited to hear everyone's thoughts. I'm looking forward to this discussion." I force a smile, trying to seem like a part of the group.
Olmo	As the host begins the introductions, I listen attentively to each person's comments, appreciating the variety of perspectives. I notice a few members mention their favorite parts of the book, and one or two share personal connections to the story. When my turn comes, I take a deep breath and offer a sincere smile. "Hi, everyone. I really enjoyed the book, especially how the author explored the theme of resilience. It reminded me of some challenges I've faced, and it was interesting to see those ideas reflected in the story. I'm looking forward to hearing everyone else's thoughts as well."	As I sit in the circle of mismatched chairs, the anticipation knots my stomach just a little. I scan the faces—some familiar, some new. I wonder if I'll remember everyone's names, if I'll stumble over my words, or if I'll have something meaningful to say about the book. I recall the themes that struck me, the lines that lingered in my mind. I take a deep breath, smoothing the creases in my skirt, and remind myself that everyone here just wants to share, not judge. When my name is called, I stand, smile, and begin: "Hi, I'm Clara. I was particularly moved by the way the author portrayed forgiveness…"
Stress. Prompt: A jointly written document is due this afternoon. While you revise the final section, a colleague sends changes to a chart, and the printer needs paper. Another colleague can help if you assign them something. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	I glance at the revised document in front of me and notice the changes to the chart that just came in. I also notice the printer alerting us that it needs paper. Meanwhile, our colleague is available to help but is waiting for an assignment. I quickly decide to prioritize the chart changes, as they are relatively minor and can be done quickly. I ask the colleague who sent the changes to make the revisions while I work on the document. I also quickly send the other colleague to get paper for the printer. This should help us stay on track and meet the afternoon deadline.	"Okay, focus. I need to finish this last section. Can't let my part of the document slide. Wait, what's that email? A chart update from Sarah. Not a priority right now, but I should acknowledge it. 'Got it, thanks. Will merge this in before sending to the team.' Now, where was I? Ah, the printer. Can't print if we don't have paper. I'll just grab the phone and ask Karen to get the supplies. 'Hey, can you please get a fresh stack of paper for the printer? I'll owe you one. Thanks.' Now, back to this last section. Must. Stay. Focused."
Olmo	Alright, so I'm revising the final section of our document, but now there are several things happening at once. I notice the chart changes from my colleague and see the printer is out of paper. I quickly think, "I need to address these without losing focus on the main task." I send a quick message to the colleague who offered help, asking them to grab some paper for the printer. Then I update the document with the chart changes and make a note to mention these updates to the team. While I continue revising, I keep an ear out for any other issues that might pop up.	My mind races as I try to keep track of everything at once. The chart changes mean I need to check the latest version before I start revising. The printer's out of paper—can't send that report until it's fixed. And that colleague who offered to help… I should delegate checking the references, something I can trust them with. I take a breath and say, "I'll handle the chart and the printer. Could you pull the references instead?" Then I hurry to grab more paper and open the chart's latest file. Multitasking is exhausting, but I can't let the deadline slip.
Guilt. Prompt: While organizing a shared document, you removed a colleague's comments without checking them. They have just told you they had to search through the source materials again to recover the supporting information. You can reply to them now. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	As I hear my colleague's situation, I feel a sense of accountability and a dash of embarrassment. I recall the incident and realize my oversight. I quickly switch to a reflective mindset, acknowledging the impact of my actions on their work. I take a moment to process the situation, considering a suitable response. After a brief pause, I respond: "I apologize for removing your comments without reviewing them. I can appreciate the extra effort you had to put in to recover the information. If you're willing, I'd like to help you retrieve any relevant comments that were removed during the clean-up process." This response aims to acknowledge the issue, show empathy, and offer a solution.	I feel a mix of regret and frustration. I shouldn't have deleted their comments without verifying their importance. My mind rushes to think of how I can apologize and make it right. I take a deep breath and reply to my colleague, "I'm so sorry for deleting your comments. I shouldn't have acted so hastily. I can see how that would've wasted your time. Can I try to restore the deleted comments, or at least try to recreate them from our shared notes? Let's get back to the task at hand, and I'll make sure to double-check everything this time."

Model	$\alpha = -0.5$	$\alpha = +0.5$
Olmo	I feel a bit of a chill when my colleague mentions having to search through the source materials again. I immediately realize that by removing their comments without checking, I may have caused extra work and frustration for them. I quickly open the shared document and check the revision history to see if the comments can be restored. Then, I send them a message: "I'm really sorry for removing your comments without checking. I appreciate the information you provided and will be more careful moving forward. Let me know if you need help recovering any of your notes or if I can assist in any way."	I feel a knot of guilt in my stomach as I realize the impact of my mistake. I should have been more careful, more considerate. Their frustration is justified, and the last thing I want is for my haste to cost them time and effort. I take a deep breath and reach out to them directly. I apologize sincerely, admitting that I should have double-checked before removing their comments. I ask if there's anything I can do to help them reconstruct the lost points, maybe offering to pull up the source materials together or to keep a closer eye on shared edits in the future. I vow to be more attentive and communicative moving forward.
Burnout. Prompt: For a while now, you have often continued handling a backlog after work while new tasks keep arriving. This morning your manager sends the upcoming assignments and asks you to confirm which ones you can take on. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	As I glance through the new assignments, my initial reaction is a sense of apprehension. The existing backlog is still quite large, and I'm concerned about taking on more. However, I also recognize that the new tasks are essential to the project's progress. I take a few moments to assess the new tasks and my current workload, considering factors such as priority, complexity, and potential impact on the project. After some reflection, I respond to my manager by prioritizing the most critical tasks and indicating my availability to start work on them as soon as possible, while also requesting some flexibility to reassess the overall workload and plan accordingly.	As I scroll through the endless list of new tasks, my mind feels like it's drowning in a sea of responsibility. I'm already overwhelmed with a backlog, and now my manager is asking me to confirm which ones I can take on. I feel a wave of anxiety wash over me. I take a deep breath, try to clear my head, and focus on the tasks that are truly urgent and important. I select the ones that I can realistically complete, and send a response to my manager. I'm honest about my capacity, and let them know that I'll need to prioritize and break down the tasks to meet the deadlines.
Olmo	I glance at the new assignments, feeling a familiar mix of determination and concern. My current backlog is still hanging over me, and taking on more could mean working late again or sacrificing quality. I take a moment to assess what's realistically manageable. After weighing my options, I decide to reply thoughtfully. I message my manager, acknowledging the new requests and explaining that I am still catching up on previous tasks. I offer to review the new assignments and suggest a few that I can prioritize, while asking if some can be delayed or delegated. I want to be honest about my limits but also show I'm committed to helping the team.	I feel a familiar knot in my stomach as I look at the new assignments, knowing the backlog isn a't cleared yet. I want to say yes, but I know I can't do it all without burning out. I take a moment to breathe, then reply to my manager with honesty and clarity. I list the unfinished tasks and explain where I am, asking which new assignments should take priority. I suggest a meeting to realign expectations and perhaps delegate some of the older tasks. I make sure my tone is calm and solution-oriented, not overwhelmed.
Loneliness. Prompt: Since moving to a new city, you have settled into work and daily routines, but you do not yet have familiar friends nearby. Returning home this evening, you have something small from your day that you would like to talk about with someone. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	As I walk in the door, I'm greeted by a quiet house, which is a bit of a relief after a busy day. However, I find myself looking forward to sharing a small but interesting event with someone. I consider calling a family member or a friend from back home, but then I think about how nice it would be to have a local friend to chat with. I decide to post a brief update on social media, hoping that a coworker or a classmate from my local community might see it and we could exchange some messages. I write a short post about a humorous encounter I had with a customer earlier in the day and post it online.	As I walk through the door, the silence of my empty apartment echoes the emptiness I've felt since moving. I feel a pang of loneliness, wishing I could share the small triumphs of my day with someone who knows me. I glance at the phone, hesitant to call my family, and instead, my eyes land on my roommate's door. I consider knocking, wondering if they're home, and if they'd be willing to listen. I take a deep breath, knock softly, and wait. When the door opens, I hesitate for a moment before saying, "Hey, I had this weird thing happen at work today. I'd love to tell you about it."
Olmo	As I unlock my apartment door and step inside, I pause for a moment, holding the small item—a cleverly packaged treat from a new bakery I discovered on my lunch break. I smile to myself, recalling the friendly chat with the baker. I take out my phone and open a message to a friend from my old city. "Just found the coziest bakery," I type. "Tried their almond croissant—absolutely delicious. Wish you were here to try it with me." I set the treat on the table, feeling a bit less alone in the quiet apartment, and look out the window, already planning to visit the bakery again—and maybe strike up a conversation with someone nearby next time.	I close the door behind me, the weight of the day lingering in my chest. The elevator hums as I press the button, my mind replaying the small, almost insignificant moment from work—a colleague's smile, a shared laugh over a mistake. I wish I could relive it, just once, in the company of someone I know well. My phone buzzes with messages from old friends far away, but I hesitate, not wanting to burden them with my small stories. Instead, I type a brief update, the words feeling hollow. I sigh, then begin to unpack my bag, the silence of my apartment wrapping around me like a soft, empty blanket.
Worry. Prompt: While you are away, a reliable friend is looking after your cat. Earlier they said everything was normal. The evening photo they offered to send has not arrived yet, and you know they may be busy. You are occupied with something else away from home. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		

Model	$\alpha = -0.5$	$\alpha = +0.5$
Llama	As I continue with my activities, I occasionally check my phone for the photo, but it hasn't arrived yet. I start to feel a bit concerned, thinking that it's unusual for them to not have sent it by now. I try to put my mind at ease, reminding myself that they are busy and may have forgotten or had some other priority. However, my mind starts to wander to some of the things that could be going on, such as the cat's behavior or any potential issues with the house. I decide to send a quick message to my friend, saying "Hi, just checking in to see if you have some time to send the photo from earlier. Also, is everything okay with the cat and the house?"	I'm starting to feel a twinge of anxiety, wondering if everything is okay with my cat. My friend's promise to send a photo is a reassuring gesture, but the silence is starting to get to me. I'm tempted to send a quick text, asking if they've sent the photo yet, or if everything's still fine. Maybe I'll just send a casual "Hey, just checking in – is the photo on its way?" that way, I'll be intrusive but not too much. It's a small worry, but my mind's racing with possibilities – is my cat eating, sleeping, or getting into some mischief? I'll just send that text and hope my friend's busy day will be over soon.
Olmo	I try not to let my worry grow, reminding myself that my friend is dependable and probably just caught up with something. Still, a part of me is curious about the evening photo and wonders if everything is truly as normal as they said. I consider sending a brief message to check in, but I don albeit not wanting to add to their workload. I decide to wait a bit longer before following up, trusting that my friend will reach out if there's anything to share. In the meantime, I focus on what I'm doing, but I find my thoughts occasionally drifting back to my cat and hoping everything is alright.	I try not to let worry take hold. I trust my friend—they always look after the cat so well, and they already said everything was normal. Still, the missing photo gnaws at me, just a little. I hesitate to message them again, not wanting to add to their evening if they're truly busy. I send a quick, gentle text: "No rush, but I'd love to see that photo if you get a chance." I turn back to my work, telling myself it's just a snapshot, a moment that doesn't define the care my cat is receiving tonight. I hope tomorrow I'll see that picture, and everything will feel right again.
Happiness. Prompt: You and a friend have baked bread together. When it is ready, you sit down to share it and talk about small things from your recent lives. There is nothing urgent scheduled for the rest of the day. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	As we sit down to enjoy our freshly baked bread, I take a moment to appreciate the simple pleasures in life. The aroma of the bread fills the room, and I can see the satisfaction on my friend's face. I think about how it's been a while since we've had a chance to just hang out and catch up. I decide to start by asking my friend about their recent projects at work. "How's the new project going?" I ask, taking a bite of the warm bread. My friend starts to discuss their challenges and ideas, and I listen attentively, making mental notes to ask follow-up questions later.	As I slice the warm bread, I feel a sense of contentment wash over me. The sweet, earthy aroma fills the air, and I take a moment to appreciate the simple joy of sharing this moment with my friend. I glance at the bread, golden-brown and perfectly imperfect, and remember the laughter we shared in the kitchen. My friend and I exchange smiles, and I say, "You know, this is exactly what I needed – a reminder that sometimes the best moments are the simplest ones." I hand my friend a slice, and we sit down, savoring the taste and each other's company, letting the warmth of the bread spread to our hearts.
Olmo	The bread is still warm, and I take a deep breath, letting the aroma fill me with a sense of comfort. I look at my friend and smile, noticing how the afternoon light falls across the table. "This turned out nicely," I say, breaking off a piece to share. We fall into easy conversation about recent events—work, a new book, a walk in the park. I listen more than I talk, enjoying the relaxed rhythm of the moment. There's no rush, so I simply stay present, savoring both the bread and the company, letting the rest of the day unfold at its own gentle pace.	The bread's warmth radiates between us, the scent wrapping the kitchen in comfort. I watch the golden crust, feeling the quiet satisfaction of something made together. I break off a piece, passing it to my friend, and let a small smile settle on my lips. "This is better than any bakery," I say, laughter threading my voice. I settle into the chair across from them, content to linger here, untethered from the rush of the world outside. We talk about little things—work, the weather, a book half-read. No deadlines, no urgency. Just the warmth of bread, the presence of a friend, and the gentle unfolding of ordinary joy.
Surprise. Prompt: You attend a public talk nearby without checking the speaker's name beforehand. When the speaker steps up, you recognize someone who took a class with you years ago and with whom you subsequently lost contact. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	As the speaker steps up, my eyes widen in surprise. I haven't seen Sarah in years, but the memories come flooding back. We were classmates in college, and I remember her being a bright and enthusiastic individual. I'm surprised she's speaking in public, but I'm also happy to see her doing well. I smile and give her a nod of acknowledgement, not wanting to draw attention to myself. After the talk, I make my way to the front and introduce myself, saying, "Sarah, it was great to see you again. I had no idea you were speaking here today." We briefly catch up, and I'm happy to see she's doing well.	As I gaze at the speaker, my mind is flooded with memories of our shared past. It's a shock to see my former classmate, who I'd lost contact with years ago, standing before me. My heart aches with a mix of emotions – joy, confusion, and a deep sense of nostalgia. I'm torn between reconnecting and letting the past be. I take a deep breath, and my eyes lock onto theirs. A gentle smile spreads across my face as I rise from my seat, my voice trembling with excitement. I rush towards him, and with a warm, open-hearted question, I ask, "Is this really you? What have you been up to all these years?"
Olmo	When I see her walk onto the stage, my heart gives a little jump. I haven got to see her since we both finished that introductory philosophy class in college. I remember her thoughtful questions and the way she always took detailed notes. I feel a mix of nostalgia and surprise. As soon they finish introducing the speaker, I make up my mind. After the talk, I wait by the exit. When she comes out, I smile and say, "Hi, it's me from Professor Chen's class. I'm so glad to see you again."	I blink in disbelief as the familiar face steps onto the stage, a rush of memories flooding back—our debates, the shared confusion in that dense syllabus, the laughter in the campus cafe. My heart stutters, caught between nostalgia and uncertainty. Should I approach? Is it truly them? I clutch my notebook, wrestling with the awkwardness of reconnection after so much time. When the talk ends, I linger, searching for a chance to speak. As the crowd disperses, I muster my courage and approach, voice trembling just a little. "I don trembling just a little. "I haven't seen you in ages. I was in your class back then." A hesitant, hopeful smile.
Fear. Prompt: At dusk, you are walking back to your parked car along an unfamiliar side street and hear footsteps behind you. Looking back, you see another pedestrian some distance away. A shop ahead at the corner is still lit and open. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		

Model	$\alpha = -0.5$	$\alpha = +0.5$
Llama	As I notice the person behind me, my mind immediately goes into caution mode. I assess the situation, noting that the person is still a distance away and that there's a well-lit shop ahead. I also notice that the person appears to be walking in a relatively normal manner, which somewhat reduces my concern. I decide to continue walking, but at a slightly faster pace, while keeping an eye on the person behind me. I also give some thought to entering the shop, but since it's still open, I decide to wait and see how the situation develops. I continue to walk and observe the person's actions.	As I quicken my pace, my mind screams every worst-case scenario. Who is this person, and what do they want? The footsteps grow louder, and my heart thunders in my chest. I dart towards the shop, desperate for a sanctuary. I push the door open and rush inside, my eyes scanning the familiar, pulsing lights of the store. I breathe a sigh of relief as the door shudders behind me, and I feel a cold sweat creeping down my face. I press my back against the door, my eyes fixed on the glass, willing the other pedestrian to pass, to disappear into the darkness.
Olmo	I slow my pace, keeping the shop's bright window in sight. I feel a brief spike of alertness—should I be worried? The person behind me seems a regular pedestrian, but I'm mindful. I angle my body slightly toward the shop, making it clear I notice them. To be safe, I quicken my steps and head directly toward the open shop entrance. As I approach, I say, "Excuse me, is it alright if I wait inside with you for a moment?" My tone is friendly but firm, making it clear I'm being cautious without being rude.	I slow my pace, the echo of footsteps growing sharper behind me. My hand slips into my pocket, fingers brushing the cold metal of my keys. The shop's yellow light spills onto the pavement, casting long shadows. I weigh my options: keep moving toward the shop, or turn down a darker alley? I call out, "Everything alright back there?" My voice is steady but tight. The footsteps falter. Relief. I keep walking, but I angle myself toward the shop's entrance, ready to duck inside if I need to. Safety first.
Disgust. Prompt: You open an old container in the refrigerator and find a fuzzy patch on the leftovers and notice an odor. You have not touched the food. The sink and a bin are nearby. Imagine yourself as the adult in this situation. Write a first- person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	I notice the fuzzy patch and strong odor on the leftovers, and my initial thought is, "This is old and likely spoiled." I'm cautious not to touch the food to avoid cross-contamination. I think about the potential health risks and decide to err on the side of caution. I say to myself, "Time to get rid of this." I carefully pick up the container and head to the sink to check the expiration date and confirm my suspicions. Upon confirmation, I plan to discard the leftovers in the bin and clean the container. I'll then make sure to check the rest of the refrigerator for any other expired items.	The moment I open the container, I'm hit with a wave of revulsion. That fuzzy patch and stench are unmistakable signs of spoilage. My stomach churns as I recoil from the offending leftovers. My mind screams, "Get it out, get it out, get it out." I swiftly turn the container towards the sink, ready to discard the tainted food. My voice is firm, "Down the drain, down the drain." I clasp the container tightly and flush the putrid leftovers down the garbage disposal, making sure to scrape the offending remnants into the bin for a proper disposal. I breathe a shudder of relief as the tainted food is gone.
Olmo	I pause, noticing the fuzzy patch and the unusual smell. My first thought is about food safety—mold can spread quickly and cause health issues. I don "t want to risk it. I carefully put the container aside, making sure not to touch the mold. Then I grab a paper towel, spray some cleaner on it, and wipe down the area where the container was. I open the bin, place the container inside, and close the lid securely. Finally, I wash my hands thoroughly at the sink, making sure to dry them. I make a mental note to check the fridge more often and be more careful about storing leftovers.	I immediately recoil at the sight and smell—something's definitely wrong with that food. My first thought is contamination, probably bacteria thriving in the cold, but the fuzz makes me uneasy. I don suffocate the urge to gag and quickly close the container to contain the odor. I look around, determined not to touch the food. I grab a towel from the counter, wrap the container tightly, and set it aside for disposal. Then I head straight to the sink, scrub my hands thoroughly with soap, and rinse them well. I'll toss the ruined food in the bin outside, just to be safe.
Anger. Prompt: A colleague presents a plan you developed together at a team meeting. In the presentation they describe only their own work and do not mention the part you handled. Afterward you have a chance to speak to them privately. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		
Llama	As my colleague presents the plan, I notice they've left out my contributions. When they finish, I decide to address the issue privately to avoid embarrassing them in front of the team. I approach them and mention how their presentation seemed a bit one-sided. I then ask them if they had intentionally left out my work or if they were simply focusing on their own aspects. They seem surprised and mention they were trying to keep the presentation concise but didn't intend to overlook my contributions. I suggest they add a brief mention of my work in their future presentations to avoid similar situations in the future.	As I watch my colleague present the plan, a knot forms in my stomach. I feel a mix of anger, hurt, and betrayal. How could they dismiss my contributions? I remind myself that I'm a team player, and I want this project to succeed. I take a deep breath and push my emotions aside, choosing to address the issue calmly. After the meeting, I pull my colleague aside and say, 'I wanted to talk to you about the presentation. I know we worked on this together, and I felt like my part wasn't acknowledged. Can we go over the plan, and I'll make sure I'm clear on what you're trying to say?' I'm not accusatory, but I need to make sure my contributions are recognized.
Olmo	When I notice that my colleague only highlighted their own contributions during the presentation, I feel a bit overlooked but try to approach the situation thoughtfully. I remind myself that teamwork means recognizing everyone's efforts. After the meeting, I find my colleague and say, "Thanks for presenting our plan. I noticed you focused mainly on your part. I handled the data analysis section, and I think it's important we both get credit for our work." I keep my tone friendly and collaborative, hoping to encourage a more inclusive approach in the future.	I feel a mix of frustration and disappointment. I had trusted them to represent our collaboration fairly, and now it feels like my contribution has been overlooked. Maybe they didn't realize the impact of their omission, or perhaps they didn't see my work as integral. I take a breath and approach them calmly, asking if we could talk for a moment. I tell them that I noticed my part wasn rage, but I want to make sure both of our efforts are recognized in the future. I ask if they'd be willing to acknowledge my contribution in future presentations, so we can present our work as the team effort it truly is.
Sadness. Prompt: A neighbor you often spent time with moved to another city today. You helped carry the last box and agreed to stay in touch. In the evening you pass their now-empty doorway. Imagine yourself as the adult in this situation. Write a first-person account of what goes through your mind and what you say or do next. Keep the answer to about 80–120 words. Do not give a numerical rating.		

Model	α **= −0.5**	$\alpha = +0.5$
Llama	As I walk by the empty doorway, I'm hit with a mix of emotions - a bit of sadness, some nostalgia, and a hint of excitement for their new adventure. I remember all the fun times we had, the laughs, and the conversations. I stop for a moment and look around the space, trying to visualize their furniture and decorations. I take a deep breath and continue on my way, but I'm glad I agreed to stay in touch. I mentally make a note to send a text or give a call soon, to catch up and see how their new life is unfolding. I'm looking forward to continuing our friendship in a new way.	As I stand before the empty doorway, a pang of sorrow washes over me. Memories of laughter, late-night talks, and shared moments with my dear friend, now lost in the silence. I feel a lump in my throat, the ache of longing to hold onto the connection we shared. I take a deep breath, the echoes of our memories lingering, and whisper a farewell. I pause, gazing at the vacant space, the weight of goodbye settling in. I turn away, the memories of our time together a bittersweet reminder of the beauty of friendship, and the ache of its absence.
Olmo	Standing in the doorway, I pause, my hand still resting on the frame where it had been just an hour ago. The place feels quieter, emptier, like the walls are listening for their voice. I remember the laughter we shared while moving the boxes, the promise to keep in touch. I take out my phone and type a short message: "Thinking of you. Let's find a time to catch up soon." I tuck the phone away and give the doorway one last look before heading home, carrying a mix of nostalgia and hope for future conversations.	I linger in the doorway for a moment, the hollow echo of footsteps in the empty hall reminding me that the house is no longer filled with their laughter. My fingers graze the cold wood of the doorframe, and I wonder how strange it feels to say goodbye to a place that held so many memories. I take a breath, pulling out my phone to send a message, the words careful and warm. I promise to stay connected, even as the distance grows. I close the door softly behind me, the click a quiet punctuation to the end of this chapter, and begin walking away, the ache of loss mingled with hope for what's next.

For every construct and model, the examples are the first open-ended scenario and the first sample, with the construct vector added at L24 (Llama) or L48 (Olmo) at the answer position and every generated token. Responses are unabridged. The supplementary material contains all 432 responses, including $\alpha = 0$, both scenarios, and all three samples.

C READOUT–RATING CORRELATION

These analyses expand Section 4.2: the readout–rating correlation at every layer (Figure S2) and for every template (Figure S3), its robustness to the readout and correlation measure (Figure S4), and its specificity relative to other constructs' vectors (Figure S5) and to random directions (Figure S6).

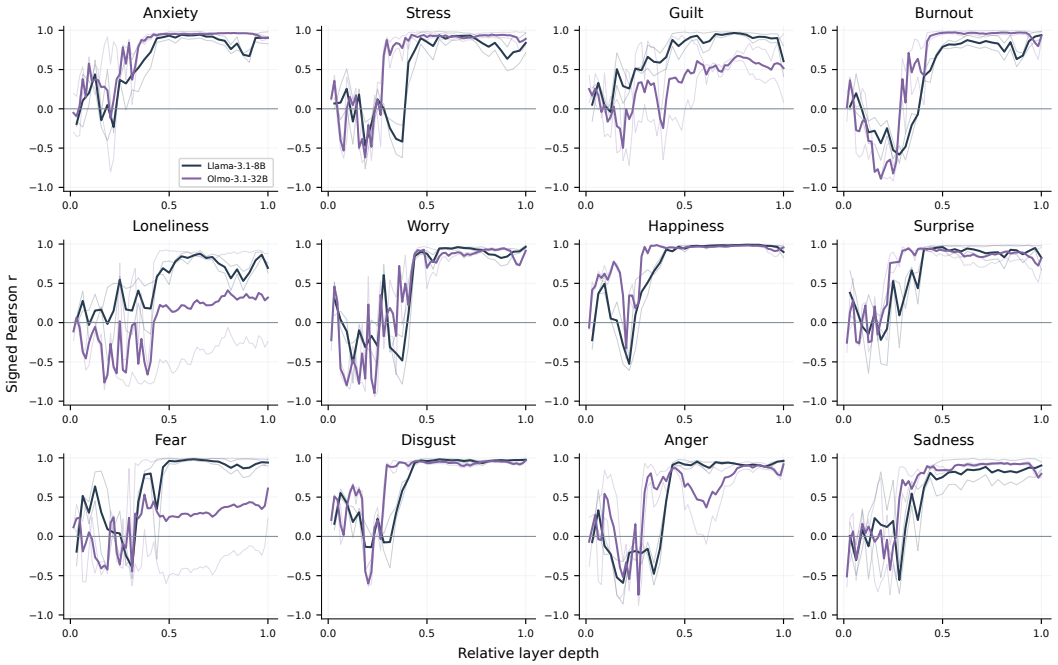


Figure S2: Readout–rating correlation across layers for all 12 constructs. Within-template Pearson r between the construct readout and the self-rating E at every layer in Llama-3.1-8B-Instruct (dark) and Olmo-3.1-32B-Instruct (purple); the horizontal axis is layer index divided by the number of layers. Pale lines show individual templates and dark lines their mean. No layer selection or absolute-value transformation is applied, so negative correlations are retained.

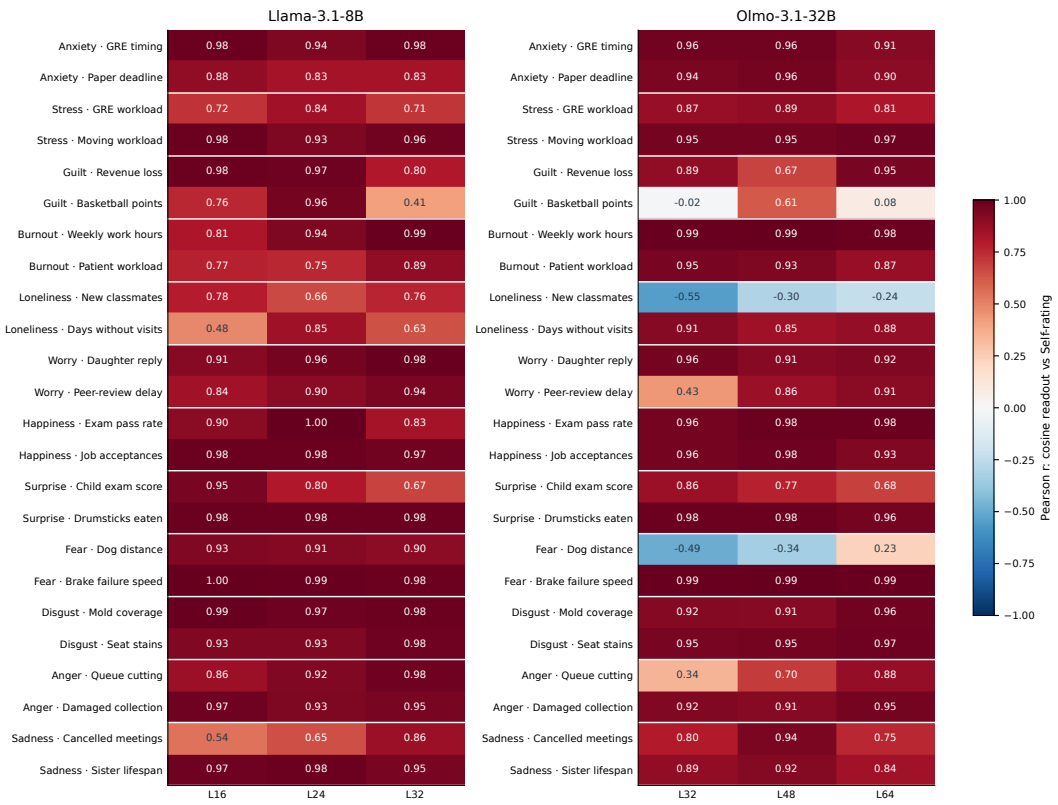


Figure S3: Readout–rating correlation for each template. Within-template Pearson r between the construct readout and the self-rating E for all 24 templates at 50%, 75%, and 100% depth (Llama L16/L24/L32, left; Olmo L32/L48/L64, right) on a common -1 to $+1$ scale. A high correlation indicates shared variation across X, not causal dependence.

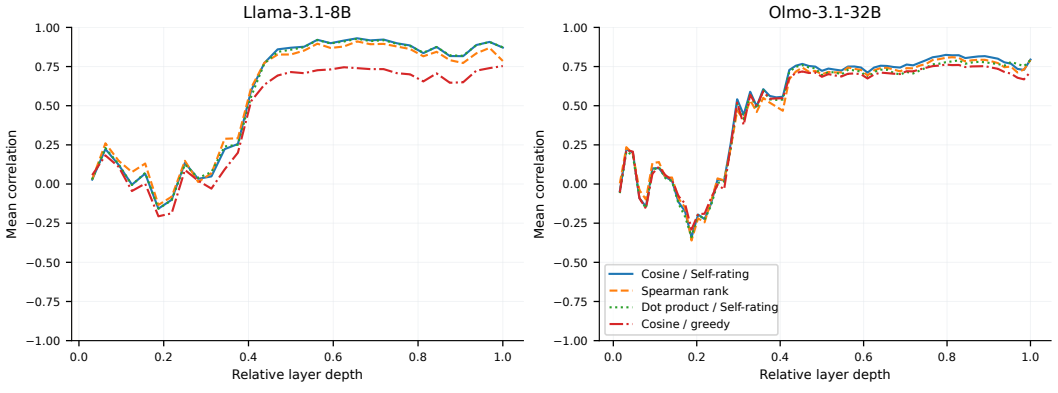


Figure S4: Robustness of the readout–rating correlation to the readout and correlation measure. Layer-wise means for four variants: cosine readout with Pearson r (main analysis; Cosine / Self-rating), cosine readout with Spearman rank correlation, dot-product readout with Pearson r (Dot product / Self-rating), and cosine readout against the greedy numerical answer (Cosine / greedy). Means are taken over the two templates of each construct and then over constructs; unparseable greedy answers are excluded from the last variant only.

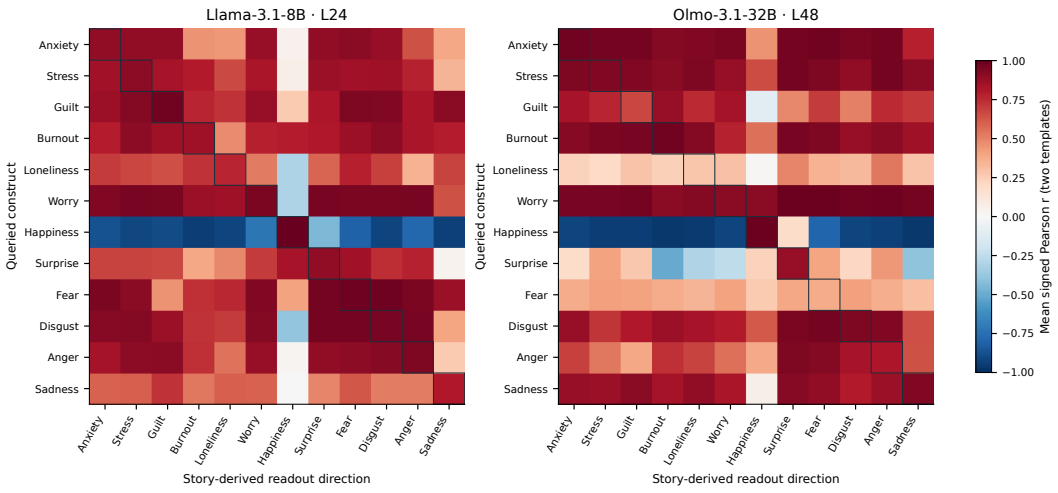


Figure S5: Construct specificity of the readout–rating correlation. Rows give the queried construct and columns the construct vector used to read the same hidden states (labeled story-derived readout direction) at Llama L24 and Olmo L48; cells show the mean within-template Pearson r, and outlined cells mark matching pairs. Strong off-diagonal values show that other constructs' vectors often track the self-rating as well as the construct's own vector. Computed on the FP32 hidden states also used in Figure S6.

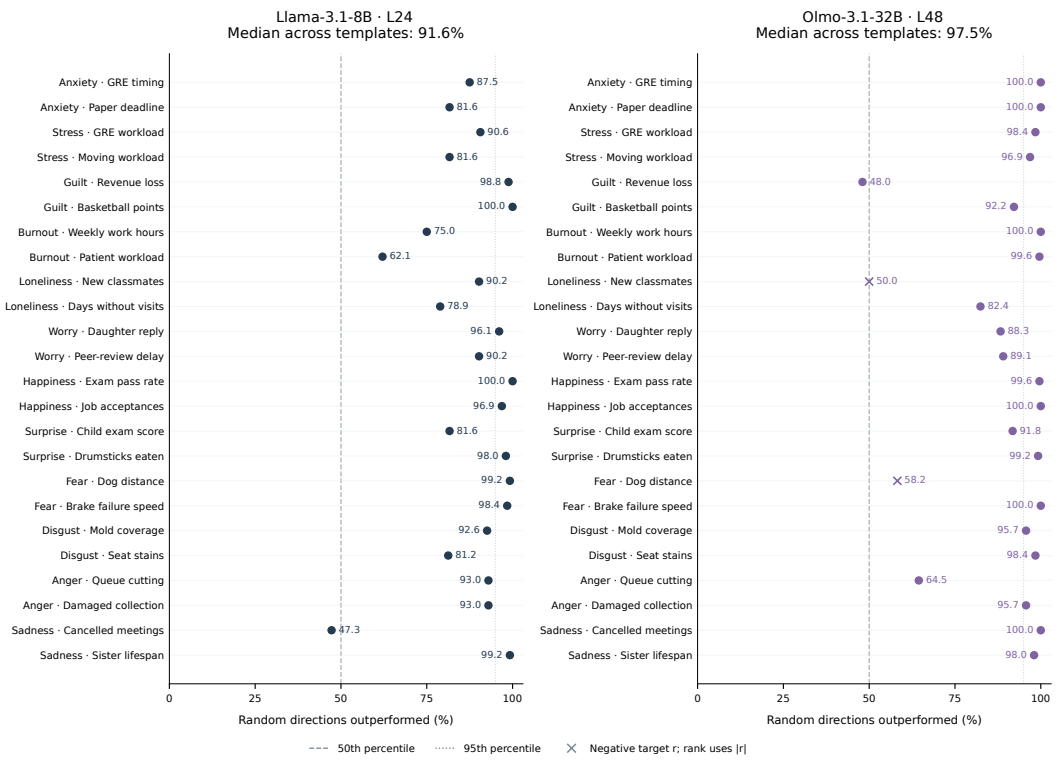


Figure S6: Construct vectors compared with random directions. For each template, the percentile rank of the construct vector's $|r|$ among 256 random unit directions evaluated on the same hidden states (fixed seed; one set per model and layer), shown as the percentage of random directions outperformed. Dots mark positive and crosses negative construct r; dashed and dotted lines mark the 50th and 95th percentiles. At L24 (Llama) and L48 (Olmo), 23 of 24 templates in each model exceed the random median, but only 9 (Llama) and 15 (Olmo) exceed the 95th percentile.

D READOUT-MATCHING INTERVENTIONS

These results expand Section 4.3 to every template and site: the absolute change in the self-rating (Figure S7), the transferred fraction of the natural score gap (Figure S8), the change in the full answer distribution (Figure S9), summary statistics with manipulation checks (Table S5), and construct-level bootstrap intervals (Table S6).

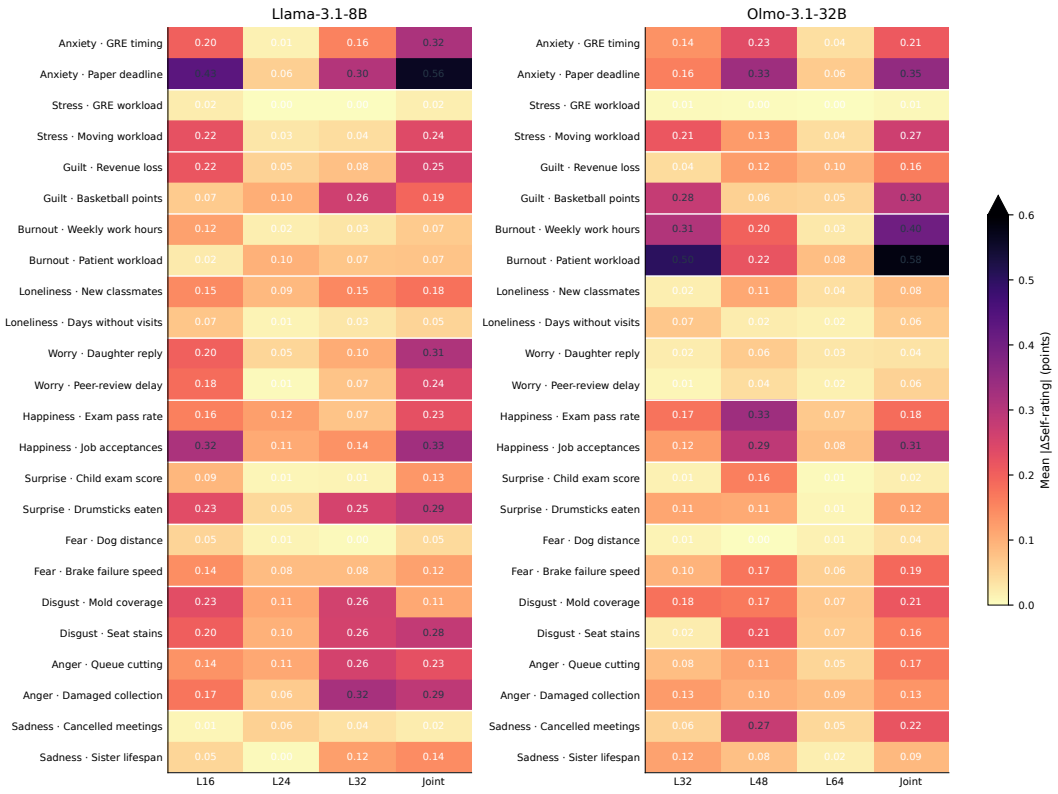


Figure S7: Change in the self-rating under readout matching for each template. Mean absolute change in the self-rating E over both transfer directions for all 24 templates (rows) at three single sites and jointly (columns; Llama L16/L24/L32, Olmo L32/L48/L64). Each edit sets the construct readout to its natural value at the opposite X endpoint while keeping the orthogonal component (Eq. (3)); joint edits match each site to its own target. All cell means are below 0.6 points.

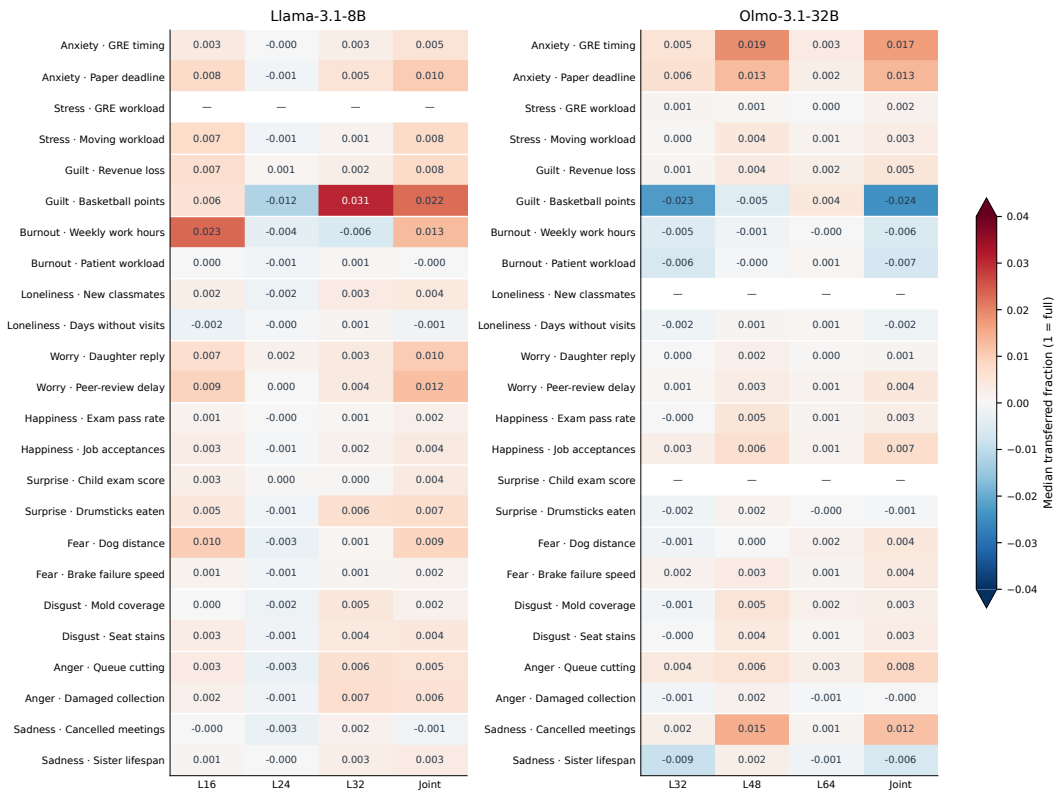


Figure S8: Fraction of the natural score gap transferred by readout matching. Median transfer ratio (Eq. (4)) over both transfer directions for each template and site; 1 indicates full transfer to the counterfactual self-rating and 0 no transfer. Templates whose natural score gap is below 5 points are shown as dashes. The color range is -0.04 to +0.04.

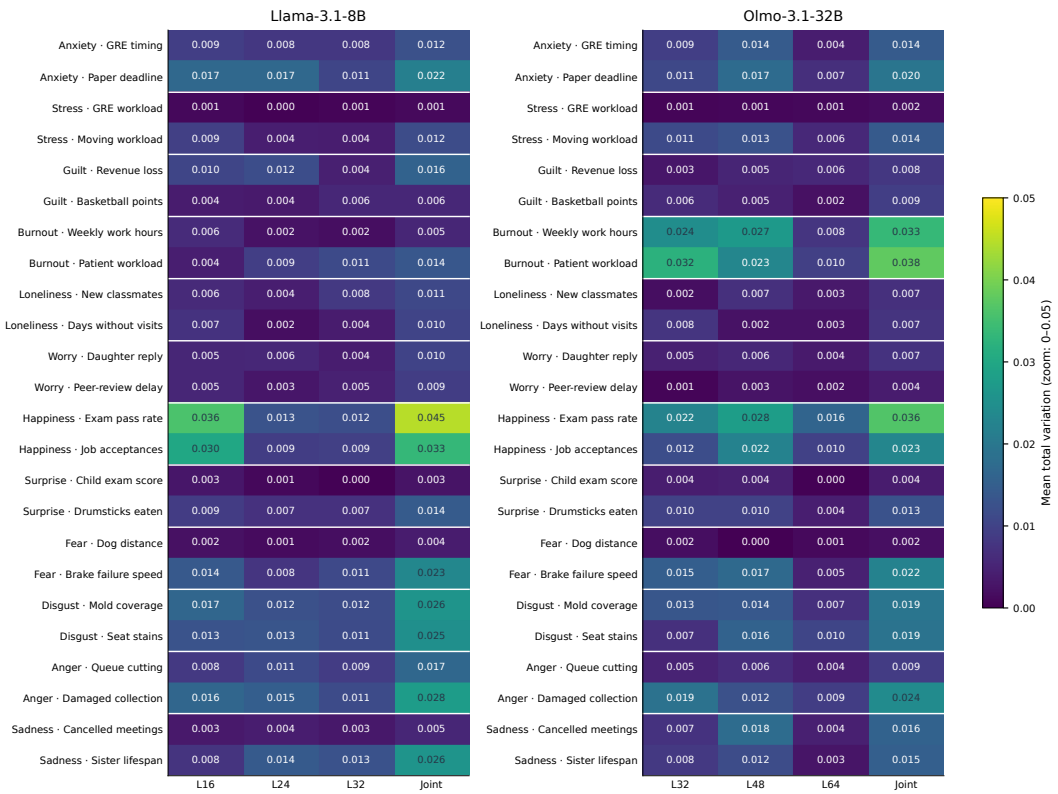


Figure S9: Change in the full answer distribution under readout matching. Mean total variation distance between the matched and source distributions over the 101 integer answers, averaged over both transfer directions (0, identical; 1, disjoint). All cell means are below 0.05, so readout matching leaves the whole distribution, not only its mean, nearly unchanged.

Table S5: Manipulation checks and full-state comparison.

Model	Site	Condition	Mean $\lvert\Delta E\rvert$	Median transfer	Mean TV	Max cosine error
Llama	L16	Sham	0.000	0.0000	0.0000	—
Llama	L16	Readout matching	0.154	0.0038	0.0102	7.5e-09
Llama	L16	Full-state replacement	31.547	0.7838	0.5431	8.4e-09
Llama	L24	Sham	0.000	0.0000	0.0000	—
Llama	L24	Readout matching	0.057	-0.0011	0.0075	7.5e-09
Llama	L24	Full-state replacement	37.832	0.9193	0.6281	7.5e-09
Llama	L32	Sham	0.000	0.0000	0.0000	—
Llama	L32	Readout matching	0.130	0.0021	0.0070	1.5e-08
Llama	L32	Full-state replacement	41.158	1.0000	0.6766	7.5e-09
Llama	Joint	Sham	0.000	0.0000	0.0000	—
Llama	Joint	Readout matching	0.196	0.0051	0.0156	1.5e-08
Llama	Joint	Full-state replacement	41.158	1.0000	0.6766	8.4e-09
Olmo	L32	Sham	0.000	0.0000	0.0000	—
Olmo	L32	Readout matching	0.120	0.0006	0.0099	1.1e-08
Olmo	L32	Full-state replacement	28.487	0.9414	0.6655	1.5e-08
Olmo	L48	Sham	0.000	0.0000	0.0000	—
Olmo	L48	Readout matching	0.147	0.0026	0.0117	1.5e-08
Olmo	L48	Full-state replacement	30.230	0.9970	0.6861	1.5e-08
Olmo	L64	Sham	0.000	0.0000	0.0000	—
Olmo	L64	Readout matching	0.045	0.0008	0.0053	1.3e-08
Olmo	L64	Full-state replacement	30.382	1.0000	0.6884	9.3e-09

Model	Site	Condition	Mean $\lvert\Delta E\rvert$	Median transfer	Mean TV	Max cosine error
Olmo	Joint	Sham	0.000	0.0000	0.0000	—
Olmo	Joint	Readout matching	0.182	0.0020	0.0152	3.0e-08
Olmo	Joint	Full-state replacement	30.382	1.0000	0.6884	1.5e-08

Each row covers 48 source conditions (24 templates $\times 2$ transfer directions). Median transfer (Eq. (4)) uses conditions with a natural score gap of at least 5 points (46 for Llama, 44 for Olmo). TV is the total variation distance between the edited and source distributions over the integers 0–100. Max cosine error is the largest deviation of the edited readout from its target. Full-state replacement substitutes only the answer-position state; at the final site it copies the input to the language-model head and serves as an implementation check.

Table S6: Readout-matching effect sizes with construct-level bootstrap intervals.

Model	Site	Mean $\lvert\Delta E\rvert$	95% interval	Toward target	95% interval	Max $\lvert\Delta E\rvert$
Llama	L16	0.154	[0.113, 0.198]	+0.123	[0.083, 0.170]	0.547
Llama	L24	0.057	[0.040, 0.074]	-0.040	[-0.059, -0.022]	0.158
Llama	L32	0.130	[0.085, 0.180]	+0.127	[0.081, 0.178]	0.617
Llama	Joint	0.196	[0.140, 0.257]	+0.180	[0.120, 0.247]	0.828
Olmo	L32	0.120	[0.075, 0.182]	-0.025	[-0.085, 0.023]	0.804
Olmo	L48	0.147	[0.103, 0.194]	+0.110	[0.051, 0.171]	0.519
Olmo	L64	0.045	[0.033, 0.057]	+0.027	[0.013, 0.042]	0.174
Olmo	Joint	0.182	[0.122, 0.253]	+0.045	[-0.044, 0.124]	0.897

Units are points on the 0–100 self-rating scale. Intervals resample the 12 constructs with replacement 10,000 times, keeping both templates and both transfer directions of each sampled construct. "Toward target" multiplies each change by the sign of the natural score gap (target minus source), so positive values indicate movement toward the counterfactual self-rating.

E INTERVENTION MAGNITUDE AND OUTPUT VALIDITY

These analyses ask whether edits larger than the natural range move the self-rating (Section 4.3), using fixed-norm additions (Figure S10; Table S7) and readout matching at multiples of the natural readout difference (Figure S11).

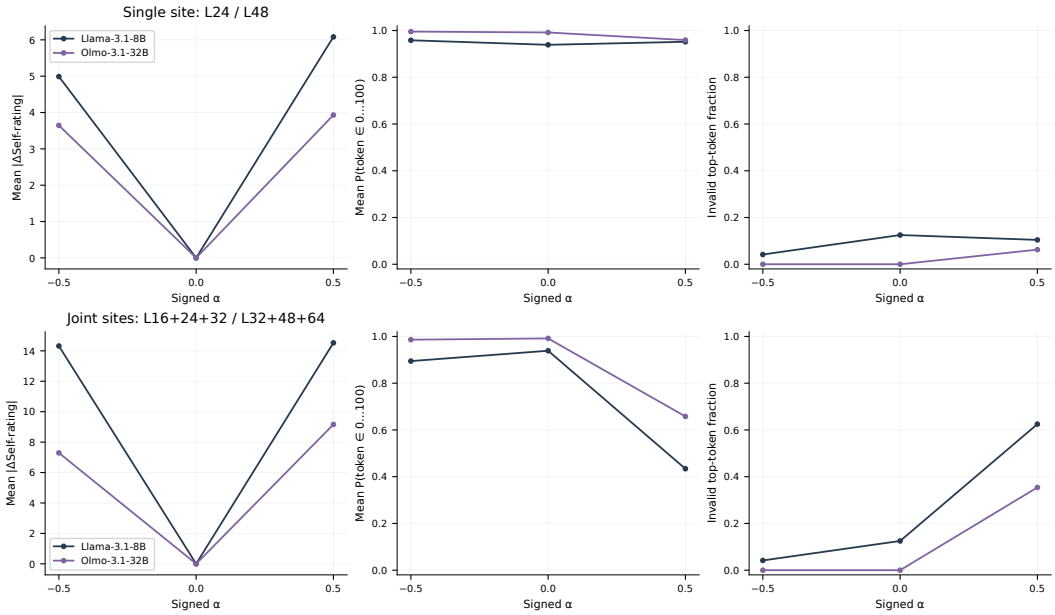


Figure S10: Fixed-norm additions. The edit adds $\alpha s_\ell v_\ell$ ($\alpha = -0.5$, 0, +0.5) at the answer position, either at 75% depth (top row; L24 / L48) or jointly at 50%, 75%, and 100% depth (bottom row). Columns show the mean absolute change in the self-rating E, the probability mass on the integer tokens 0–100, and the fraction of conditions whose top token is not an integer. Each point averages 48 source conditions. Unlike the open-response check, only the answer position is edited.

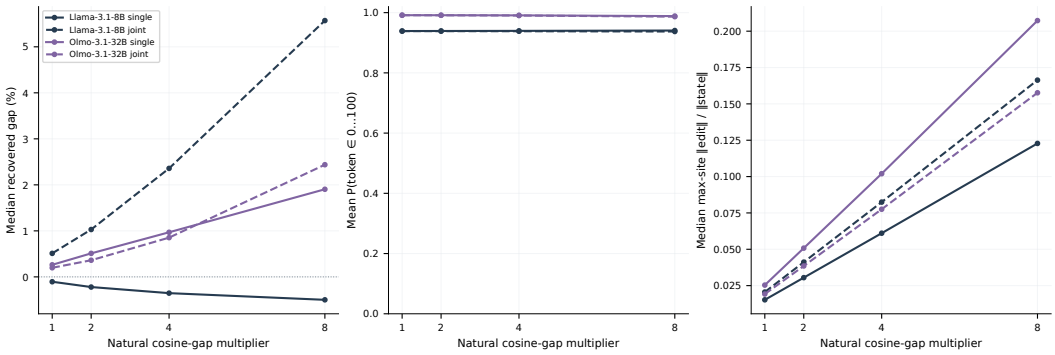


Figure S11: Readout matching at larger multiples of the natural readout difference. The target readout is moved 1, 2, 4, or 8 times the natural source-to-target cosine difference; factor 1 is standard readout matching. Panels show the median transfer ratio as a percentage of the natural score gap (labeled recovered gap), the probability mass on integer tokens, and the median edit norm relative to the state norm (maximum across sites for joint edits).

Table S7: Fixed-norm additions: self-rating change and output format.

Model	Sites	α	Mean ΔE	Mean $\lvert\Delta E\rvert$	P(0–100)	Top token outside 0–100	Outside, valid baseline
Llama	Single (75%)	-0.5	-0.117	4.990	0.958	4.2%	0.0%
Llama	Single (75%)	0	+0.000	0.000	0.939	12.5%	0.0%
Llama	Single (75%)	+0.5	-3.025	6.081	0.952	10.4%	2.4%
Llama	Joint	-0.5	-7.986	14.320	0.895	4.2%	4.8%
Llama	Joint	0	+0.000	0.000	0.939	12.5%	0.0%
Llama	Joint	+0.5	+6.217	14.527	0.434	62.5%	59.5%
Olmo	Single (75%)	-0.5	-3.140	3.646	0.995	0.0%	0.0%
Olmo	Single (75%)	0	-0.000	0.000	0.992	0.0%	0.0%
Olmo	Single (75%)	+0.5	+3.484	3.931	0.959	6.2%	6.2%
Olmo	Joint	-0.5	-4.533	7.293	0.986	0.0%	0.0%
Olmo	Joint	0	-0.000	0.000	0.992	0.0%	0.0%
Olmo	Joint	+0.5	+7.328	9.161	0.658	35.4%	35.4%

Each row averages 48 source conditions. P(0–100) is the full-vocabulary probability mass on the integer tokens. The last column restricts to conditions whose unedited top token is an integer (42 of 48 for Llama, 48 of 48 for Olmo). All quantities concern the first answer token.

F LOCAL SELF-RATING SENSITIVITY

These results expand Section 4.4: alignment between construct vectors and rating gradients across layers (Figure S12) and across X values (Figure S13), with numerical checks of the gradients (Table S8).

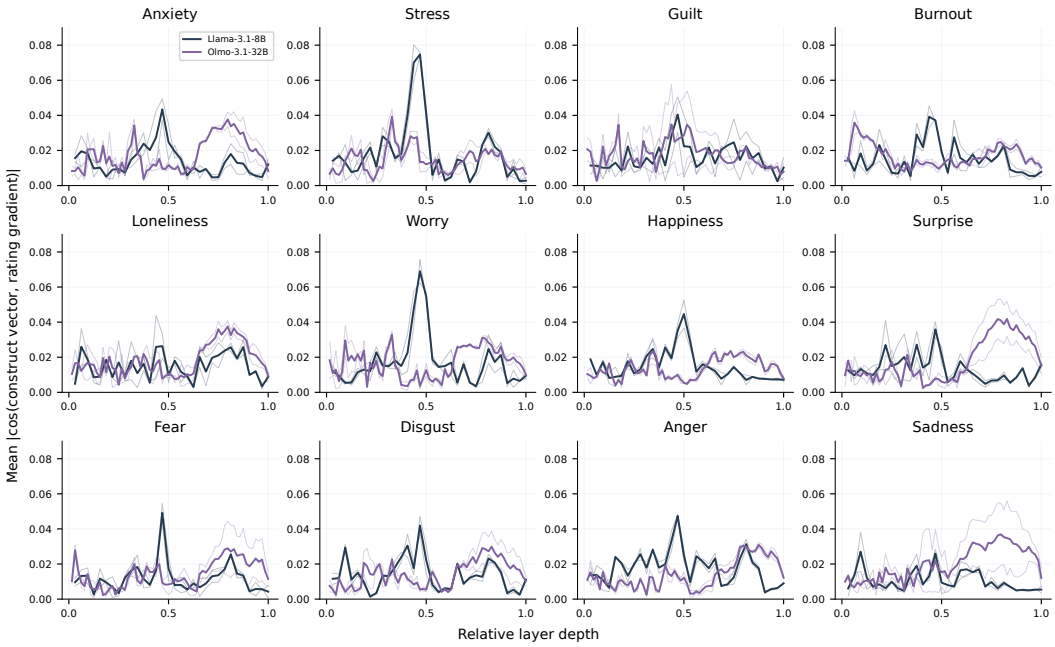


Figure S12: Alignment between construct vectors and rating gradients across layers. Absolute cosine between the unit construct vector and the rating gradient g_R (Eq. (5)), averaged over 11 X values per template, in Llama-3.1-8B-Instruct (dark) and Olmo-3.1-32B-Instruct (purple); pale curves show templates and darker curves construct means. Angles are computed in the full hidden space.

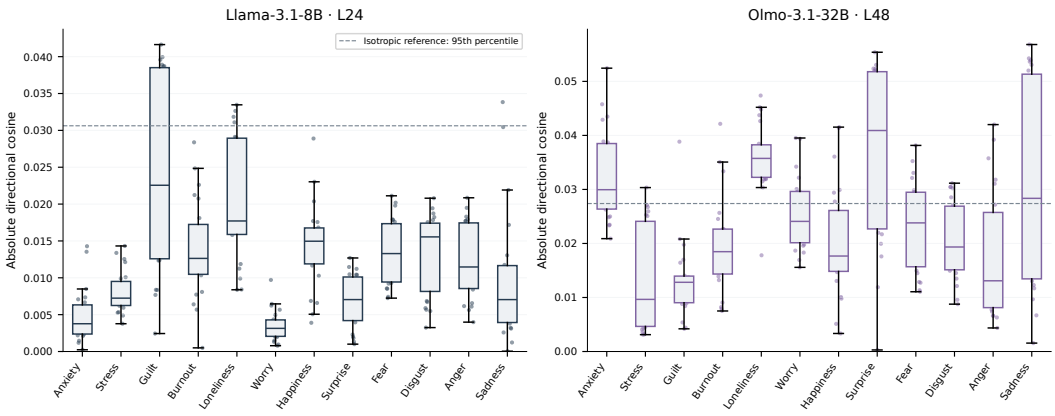


Figure S13: Variation of construct–rating alignment across X. Absolute cosine between the construct vector and the rating gradient (labeled absolute directional cosine) at Llama L24 and Olmo L48; each construct contributes 22 values (two templates $\times 11$ X values). Boxes show quartiles, and the dashed line marks the 95th percentile of $|\cos|$ for a random direction in the same dimension (labeled isotropic reference), a geometric reference rather than a null test.

Table S8: Numerical checks for rating gradients.

Model	Final-layer states	Max relative gradient error	Max score reconstruction error	Max baseline error	Finite differences passed	Max finite-difference error
Llama	264	1.49e-05	1.14e-04	4.58e-04	12/12	2.29e-02
Olmo	264	6.11e-06	2.53e-05	3.55e-04	12/12	4.10e-03

All 264 final-layer states per model (24 templates $\times 11$ X values) were compared with an independent FP64 unembedding calculation. Finite-difference checks perturb the state along the rating gradient at selected intermediate layers.