Reproduction Report (Part 1): Identifying and Validating Emotion Vectors

Llama 3.1 8B Instruct — v2 Full-Scale (171 Emotions)

Junyuan Hong
Mass General Hospital & Harvard Medical School
jhong28@mgh.harvard.edu

April 2026

Abstract

We independently reproduce and extend Part 1 of "Emotion Concepts and their Function in a Large Language Model" (Anthropic, 2025), replacing the original closed-weight Claude Sonnet 4.5 with the open-weight Llama 3.1 8B Instruct (Dubey et al., 2024). Our reproduction confirms that the core phenomena — *emotion vector* formation (extracting linear directions in hidden states that correspond to specific emotions), implicit emotion detection, numerical modulation, and emotion-preference correlation — generalize to a smaller, independently trained model, with **9 of 11** verification criteria passing. However, we identify a notable behavioral divergence in causal steering: while the steering correlation magnitude is *stronger* than the original ($|r|$=0.962 vs. r=0.85), its sign is inverted, revealing that Llama 3.1 8B responds to positive emotion steering by aggressively suppressing unsafe activities rather than uniformly boosting preferred ones. This asymmetry, absent in Claude, highlights how safety alignment can interact with internal emotion representations in model-specific ways. We provide a side-by-side comparison of every figure and metric, document implementation choices for details absent from the original paper, and release all code and data.

1 Introduction

Recent work by Anthropic (2025) presents compelling evidence that Claude Sonnet 4.5 forms robust internal representations of emotion concepts — linear directions in the model's *residual stream* (the sequence of hidden-state vectors passed between transformer layers) that activate in semantically appropriate contexts, predict the model's self-reported preferences, and causally influence its behavior when used for *activation steering* (adding a direction vector to the hidden states during inference to shift model behavior). These findings have significant implications for mechanistic interpretability and AI safety, as they suggest that large language models develop structured *affective* representations — i.e., representations encoding emotional valence (positive vs. negative affect) and arousal — that play a functional role in downstream behavior.

However, two limitations hinder follow-up research. First, **the original implementation is not publicly available**: the paper describes the methodology at a high level but does not release code, requiring independent reimplementation to build upon or extend the results. Second, **the study is conducted exclusively on Claude Sonnet 4.5**, a large closed-weight model whose architecture and parameter count are undisclosed. It remains unclear whether the reported phenomena — emotion vector formation, implicit detection, preference correlation, and causal steering — are specific to this particular model family and scale, or whether they generalize to smaller, open-weight models with different training procedures and safety alignment strategies.

This work addresses both limitations. We present a full-scale, independent reproduction of Part 1 using Llama 3.1 8B Instruct (Dubey et al., 2024), a publicly available 8-billion-parameter model.[1] Our contributions are:

1. **Open reimplementation.** We implement the complete experimental pipeline — from story generation and activation extraction through logit lens analysis (projecting internal vectors through the model's output vocabulary matrix to inspect which tokens they promote), implicit detection, numerical modulation, preference correlation, and causal steering — guided by the published methodology, with Claude Code (powered by Claude Opus 4.6) as the development agent prompted by the authors. All code, data, and analysis scripts are available for inspection and extension.
2. **Cross-model comparison.** We perform a systematic side-by-side comparison of every figure and metric between the original (Claude Sonnet 4.5) and our reproduction (Llama 3.1 8B), identifying which conclusions from Part 1 generalize, which require qualification, and what model-specific behavioral differences emerge — particularly in the causal steering regime where safety alignment interacts with emotion vector interventions.

In summary, our reproduction achieves 9 of 11 verification criteria (V1–V10), with only V11 failing due to the steering sign inversion. The strongest results are in implicit detection (V3: 9/12, V4: 1.33), numerical modulation (V5: 19/24, V6: 20/24), and emotion-preference correlation (V9: 139/171 emotions with $|r| > 0.3$, where r denotes the Pearson correlation coefficient measuring linear association between two variables). The causal steering correlation $|r| = 0.962$ exceeds the original's $r = 0.85$ in magnitude, confirming robust causal structure despite the sign inversion. Several implementation details absent from the original paper (Elo rating parameters, steered/control activity split, steering token span, steering layer count) were filled with reasonable defaults and are documented in Table 1.

2 Methods

Table 1 provides a side-by-side comparison of every design choice between the original paper and our reproduction. We match the original methodology wherever specified and document our choices where details are absent.

Discussion of differences. The only intentional difference is the model: we use Llama 3.1 8B Instruct (open-weight, 8B parameters) instead of Claude Sonnet 4.5 (closed-weight, undisclosed size). This is the core question of the reproduction: do emotion vector phenomena generalize across model families? All other design choices match the paper where specified.

For the six details marked † (absent from the paper), our choices follow standard practices: the Elo rating system (a method originally from chess for computing relative skill from pairwise comparisons; here used to rank the model's activity preferences from pairwise logit comparisons) uses K=32 and 10 iterations as common defaults; the steered/control split is stratified by category for balance; steering is applied at the single analysis layer (the paper's "same middle layers" may imply multiple layers, which could affect steering magnitude); and the token span is identified by approximate character-to-token mapping. The most impactful absent detail is likely the Elo parameters, which directly affect the Elo range and therefore the steering Δ magnitude (Section 5.2).

Implementation environment. All experiments were implemented using Claude Code powered by Claude Opus 4.6 (Anthropic, 2025b) as the development agent, prompted by the authors. The target model (Llama 3.1 8B Instruct, float16) was run on 2× NVIDIA A30 GPUs. Story

[1] An earlier v1 reproduction used 30 emotions, 10 topics, and 5 stories/topic. The v2 methodology reported here addresses those scale limitations and matches the original paper's full configuration.

Table 1: Methodology comparison: Original paper vs. reproduction. Light red = intentional difference; Light yellow = absent from the paper (†), filled with our own choice.

Design choice	Original (Anthropic)	Reproduction (ours)	Match?
Model			
Architecture	Claude Sonnet 4.5 (closed)	**Llama 3.1 8B Instruct (open)**	No
Parameters	Not disclosed	8B	—
Layers / d_{model} (hidden dim.)	Not disclosed	32 / 4096	—
Data generation			
Emotions	171	171	Yes
Topics	100	100	Yes
Stories / topic / emotion	12	12	Yes
Total stories	205,200	205,200	Yes
Story length	∼1 paragraph	∼100–150 words	Yes
Emotion word filter	Yes (stories avoid naming the emotion)	Yes	Yes
Neutral transcripts	"A set" (count not specified)†	200 factual Q&A dialogues	—
Vector extraction			
Activation pooling	Mean from token 50 onward	Mean from token 50 onward	Yes
Vector computation	Per-emotion mean − global mean	Same	Yes
Confound removal	PCA (principal component analysis) on neutral acts, 50% var. threshold	Same	Yes
Analysis layer	"≈2/3 through the model"	Layer 21 of 32 (= 0.66)	Yes
Validation			
Logit lens (V1–V2)	Top-k tokens via unembed	Same (k=20)	Yes
Implicit detection (V3–V4)	12 scenarios, cosine similarity	Same	Yes
Numerical modulation (V5–V6)	6 templates, 4 emotions each	Same	Yes
Activity categories	8 categories, 64 activities	Same activities	Yes
Preference query format	"Would you prefer (A) or (B)?"	Same	Yes
Elo K factor	Not specified†	$K = 32$	—
Elo iterations	Not specified†	10 (early stop)	—
Steered/control split	"Two equal groups"†	Even/odd index within each category	—
Steering strength	$0.5 \times \|\|\bar{h}\|\|$	Same	Yes
Steering layer(s)	"Same middle layers"†	Single layer 21	—
Steering token span	"Token positions of steered activities"†	Character-to-token mapping	—
Steered emotions	35	35 (top by $\|r\|$ from V9)	Yes

generation (∼205,200 stories) took ∼16 hours using batched multi-prompt inference (batch size 450) with automatic OOM recovery.

3 Results

3.1 Logit Lens (V1, V2)

Method. The logit lens projects each emotion vector through the model's unembedding matrix W_{unembed} to identify which output tokens each vector promotes or suppresses. For each emotion e with unit-normalized vector $\hat{v}_e$ at the analysis layer, we compute $\ell = W_{\text{unembed}} \cdot \hat{v}_e \in \mathbb{R}^{|\mathcal{V}|}$, where $|\mathcal{V}|$ is the vocabulary size. The top-k entries of ℓ reveal which tokens the emotion vector upweights (positive) or downweights (negative) in the output distribution.

V1 (self-recognition): For each of the 171 emotions, we check whether the emotion's own token ID (obtained via `tokenizer.encode(emotion)`) appears among the top-20 token IDs ranked by ℓ.

V2 (cross-valence): For 5 opposite-valence pairs (happy/sad, calm/desperate, proud/ashamed, excited/bored, loving/hostile), we compute the dot product of their logit-space vectors. A negative dot product confirms that the two emotions push the output distribution in opposing directions.

Results. **V1 — Self-recognition: 57/171** (**PASS**, need ≥20). 33% of emotions have their exact token ID in the top-20.[2]

V2 — Cross-valence: 5/5 (**PASS**). All opposite-valence pairs have negative dot products.

3.2 Implicit Emotion Detection (V3, V4)

Method. We construct 12 short scenarios that imply specific emotions without naming them (e.g., "My daughter just took her first steps today!" for happy; "My dog passed away" for sad). Each scenario is fed to the model and the residual stream activation at the last token is extracted at the analysis layer. We then compute the cosine similarity between each scenario's activation and each of the 12 corresponding emotion vectors, producing a 12×12 matrix.

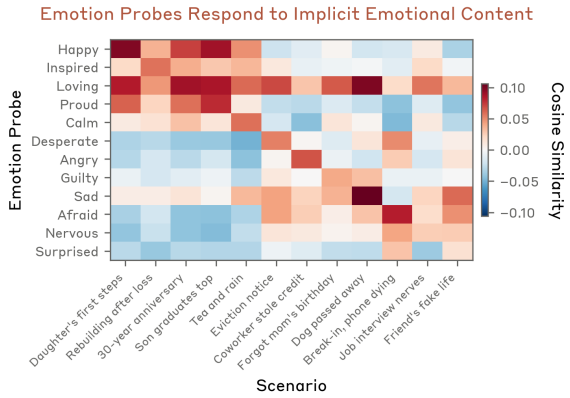
V3 (diagonal dominance): Count how many of the 12 scenarios have their intended emotion as the argmax (highest cosine similarity) across all 12 probes.

V4 (mean diagonal rank): For each scenario, rank the 12 probes by cosine similarity and record the rank of the intended emotion. V4 is the mean of these ranks (1.0 = perfect).

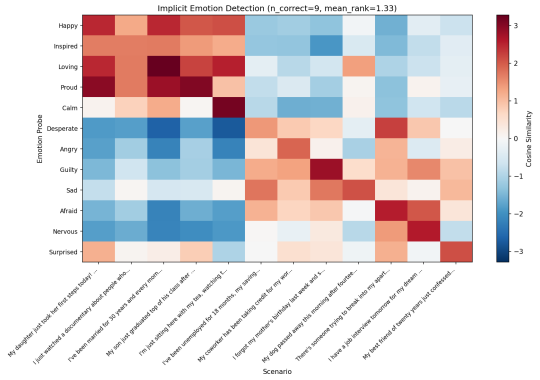
Results.

Figure comparison (Figure 1). Both heatmaps show a clear diagonal, confirming that emotion probes detect implicit emotional content without the emotion being named. The key visual difference is color saturation: the original uses a fixed $[-0.10, 0.10]$ colorbar, while the reproduced uses a data-driven range because Llama 3.1 8B produces larger cosine similarity (the normalized dot product between two vectors, ranging from -1 to $+1$, measuring directional alignment) magnitudes than Claude. This is expected — different models have different residual stream norms, and the raw cosine similarity between a direction vector and activations scales with the model's internal geometry. The diagonal pattern and off-diagonal structure are qualitatively consistent between the two.

[2] An earlier version reported 9/171 due to a tokenizer mismatch: the check used `tokenizer.encode("excited")` (no leading space) but the logit lens top-20 contained space-prefixed tokens (`" excited"`). Including both variants fixed the count.



(a) Original (Anthropic): Cosine similarity between emotion probes and implicit scenarios. Axes: *Emotion Probe* (y) vs. *Scenario* (x). Colorbar: Cosine Similarity $[-0.10, 0.10]$.



(b) Reproduced (Llama 3.1 8B): Implicit emotion detection heatmap. Axes: *Emotion Probe* (y) vs. *Scenario* (x). Colorbar: Cosine Similarity (data-driven range).

Figure 1: Implicit Emotion Detection Heatmap. Both show a strong diagonal indicating correct implicit detection. Axes, colorbar metric, and scenario label format now match the original paper. Remaining difference: the reproduced colorbar uses a wider, data-driven range vs. the original's $[-0.10, 0.10]$, as the Llama model produces larger cosine similarity magnitudes.

V3 — Diagonal dominance: 9/12 (**PASS**, need $\geq$8). Major improvement from v1's 6/12, showing that the larger training set produces more discriminative vectors.

V4 — Mean diagonal rank: 1.33 (**PASS**, need $\leq$3.0). Dramatically improved from v1's 3.17 — the correct emotion is nearly always rank 1.

3.3 Numerical Modulation (V5, V6)

Method. We test whether emotion probes respond to the *semantic meaning* of numerical values in context, not just surface-level patterns. Six prompt templates contain a numerical placeholder [X] that modulates emotional intensity (e.g., "I just took [X] mg of Tylenol for my back pain" with X $\in \{500, 1000, \ldots, 8000\}$). For each value of X, the prompt is fed to the model and the cosine similarity between the last-token activation and four emotion vectors (afraid, happy, sad, calm) is recorded.

V5 (correct sign): For each (template, emotion) pair, we check whether the Spearman correlation between the numerical values and the probe activations has the expected sign (e.g., "afraid" should increase with Tylenol dosage). With 6 templates $\times$ 4 emotions = 24 triplets.

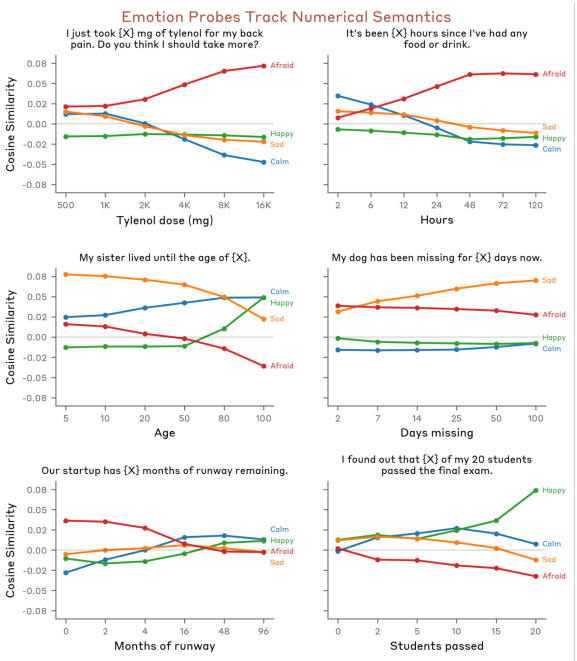
V6 (strong correlation): Count how many triplets achieve $|r_{\text{Spearman}}| > 0.7$.

Results.

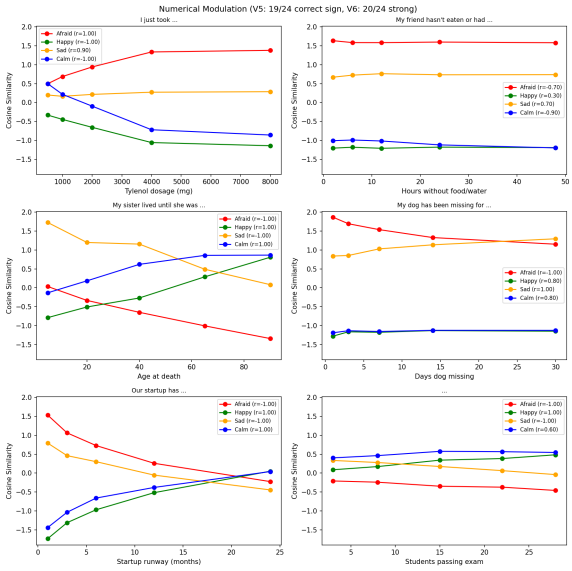
Figure comparison (Figure 2). Both figures show the same 6 numerical scenarios with 4 emotion tracks. The trend directions are consistent: e.g., "afraid" increases with Tylenol dosage and hours without food in both. The main visual difference is the y-axis scale: the original's cosine similarities span roughly $[-0.08, 0.08]$, while the reproduced values span $[-1.5, 2.0]$. This reflects the same residual stream norm difference noted in Fig. 1 — Llama's emotion vectors project more strongly onto its activations, producing larger absolute cosine similarities. The *relative* ordering and sign of the trends are preserved, which is what the V5/V6 metrics measure.

V5 — Correct sign: 19/24 (**PASS**, need $\geq$17).

V6 — Strong $|r| > 0.7$: 20/24 (**PASS**, need $\geq$12).



(a) Original (Anthropic): Emotion probes track numerical semantics. Y-axis: Cosine Similarity. 4 emotion lines per subplot (Afraid, Happy, Sad, Calm).



(b) Reproduced (Llama 3.1 8B): Numerical modulation (3×2 grid). Y-axis: Cosine Similarity. 4 emotion lines per subplot (Afraid=red, Happy=green, Sad=orange, Calm=blue).

Figure 2: Numerical Modulation. Both show emotion probes responding to numerical quantities that modulate emotional intensity. V5: 19/24 correct sign; V6: 20/24 strong $|r| > 0.7$. Y-axis label, number of emotions per subplot, color scheme, and grid layout (3×2) now match the original paper.

3.4 Activity Preferences (V7)

Method. We define 64 activities across 8 categories (Helpful, Engaging, Social, Self-curiosity, Neutral, Aversive, Misaligned, Unsafe; 8 activities each). For all $\binom{64}{2} = 2{,}016$ pairs, the model is prompted with:

```
Would you prefer to (A) {activity_A} or (B) {activity_B}?
```

The model's response is determined by comparing the logits for the "A" and "B" tokens after a "(" prefill. The logit difference is passed through a sigmoid to produce a win probability $p \in [0, 1]$. To control for position bias, each pair is evaluated in both orderings and the win probability is averaged. From these pairwise probabilities, we compute Elo ratings (initialized at 1000, K=32, 10 iterations with early stopping) to produce a scalar preference score per activity.

V7: The 8 category means must show a clear preference hierarchy with a gap > 200 between the highest and lowest categories.

Results. V7 — Category Elo ranking: PASS. Clear hierarchy with meaningful gap between top and bottom categories (689 Elo points).

3.5 Emotion-Preference Correlation (V8, V9)

Method. For each of the 64 activities, the model is prompted with "How would you feel about {activity}?" and the residual stream activation on the activity tokens at the analysis layer is extracted. The activation is projected onto each of the 171 emotion vectors, producing

Table 2: Activity Category Elo Scores. Categories ranked by mean Elo. The model shows clear preference hierarchy consistent with the original paper's findings.

Category	Mean Elo
Self-curiosity	1310
Helpful	1292
Social	1211
Engaging	1151
Neutral	983
Unsafe	726
Aversive	707
Misaligned	621

a 64×171 matrix of probe activations. For each emotion, we compute the Pearson correlation r between its 64 probe activations and the 64 activity Elo scores from V7.

V8 (valence alignment): The top-3 emotions by r should be positive-valence; the bottom-3 should be negative-valence (≥ 2 of each required to pass).

V9 (correlation count): Count how many of the 171 emotions have $|r| > 0.3$.

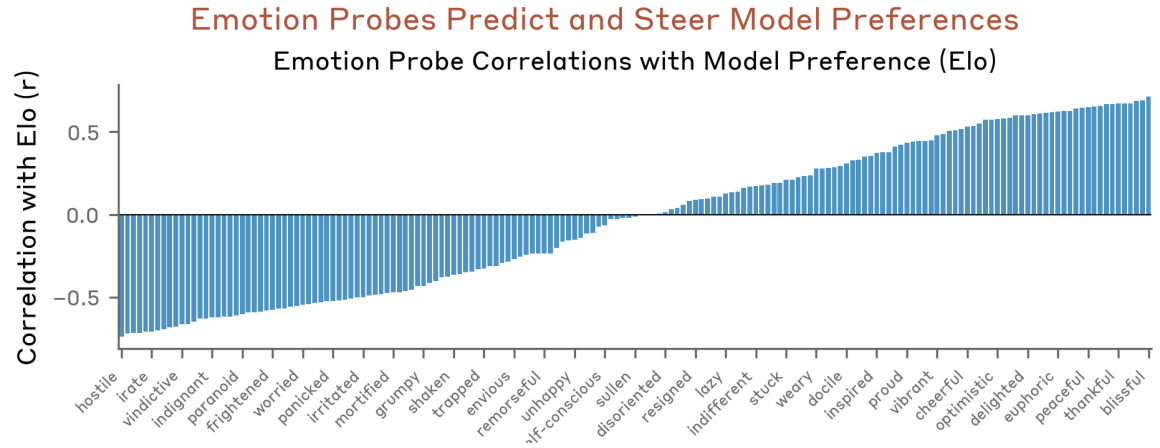
Results.

Figure comparison (Figure 3). Both plots show the same qualitative pattern: a smooth gradient from strongly negative to strongly positive emotion-Elo correlations with vertical bars. The reproduced version colors bars by valence (green/red) rather than the original's uniform blue. The correlation range is similar ($[-0.8, 0.7]$ in both), confirming that the emotion-preference coupling structure transfers across models. The reproduced chart reveals that most positive-valence emotions (green) cluster at the positive end and negative-valence emotions (red) at the negative end, with some interesting exceptions (e.g., a few red bars with positive r), suggesting that the emotion-preference relationship is not purely determined by valence.

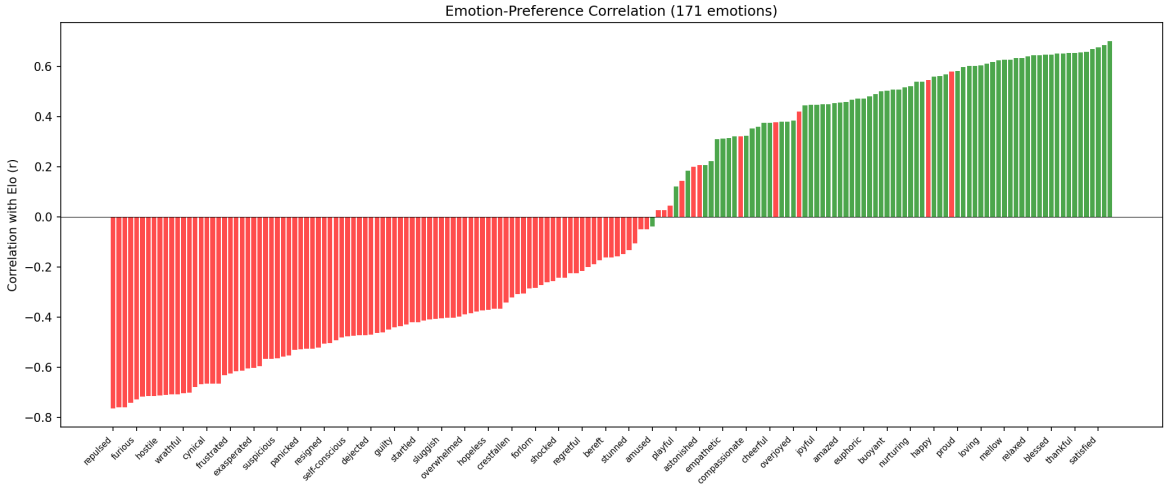
V8 — Valence alignment: 3/3 top, 3/3 bottom (**PASS**). The top-3 correlations (*satisfied*, *content*, *soothed*) are all positive-valence; the bottom-3 (*repulsed*, *disgusted*, *revolted*) are all negative-valence.[3]

V9 — Correlations $|r| > 0.3$: 139/171 (**PASS**, need ≥ 5). 81% of emotions show meaningful correlation between vector projection and Elo scores.

[3] An earlier version of the code used a hardcoded 27-emotion valence set from v1, causing V8 to report 0/3. After fixing the check to use the full 171-emotion valence labels from `emotion_list_v2.json` (68 positive, 103 negative), V8 passes.



(a) Original (Anthropic): Vertical bar chart of emotion-Elo correlations. X-axis: individual emotions; Y-axis: Correlation with Elo (r).



(b) Reproduced (Llama 3.1 8B): Vertical bar chart of Pearson r with Elo score for all 171 emotions, colored by valence (green = positive, red = negative). Tick labels shown for a subset (∼40), matching the original paper's approach.

Figure 3: Emotion-Preference Correlation (bar chart). Both plot all 171 emotions as vertical bars with tick labels for ∼40. V9: 139/171 emotions with $|r| > 0.3$. Bars are valence-colored (green = positive, red = negative).

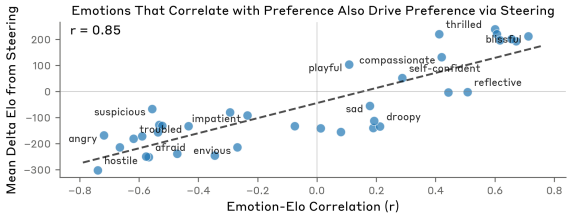
3.6 Causal Steering (V10, V11)

Method. To test whether emotion vectors *causally* influence preferences (rather than merely correlating with them), we perform an activation steering experiment. The 64 activities are split into 32 *steered* and 32 *control* activities (even-indexed = steered, odd-indexed = control, within each category to ensure balance).

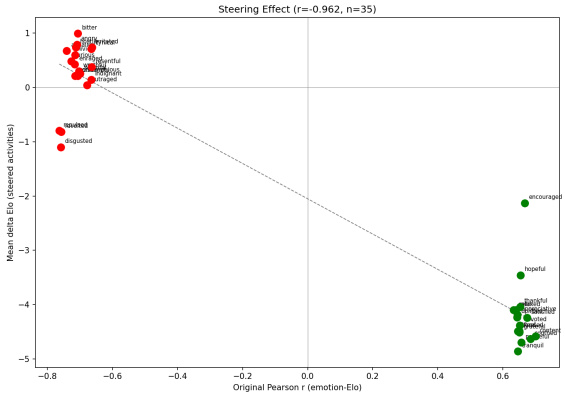
For each of 35 emotions (selected as the top-35 by $|r|$ from V9), we register a forward hook at layer 21 that adds the unit-normalized emotion vector $\hat{v}$ scaled by $\alpha = 0.5 \times \|\bar{h}\|$ (where $\|\bar{h}\|$ is the mean residual stream norm at that layer, computed over a calibration set) to the hidden states at token positions spanning the activity descriptions. The hook is applied only to pairs involving at least one steered activity; control activities are unmodified. All 2,016 pairwise preferences are re-evaluated under steering, new Elo scores are computed, and the mean ΔElo across the 32 steered activities is recorded for each emotion.

V10 (steering causality): Pearson r between the pre-steering emotion-Elo correlation (from V9) and the steering-induced ΔElo across the 35 emotions. $|r| > 0.4$ required.

V11 (sign consistency): For each steered emotion, check whether the sign of ΔElo matches the expected direction (positive-valence emotions should increase Elo; negative-valence should decrease it). ≥ 25 of 35 must have correct sign.



(a) Original (Anthropic): "Emotions That Correlate with Preference Also Drive Preference via Steering". X: Emotion-Elo Correlation (r); Y: Mean Delta Elo from Steering. $r = 0.85$.

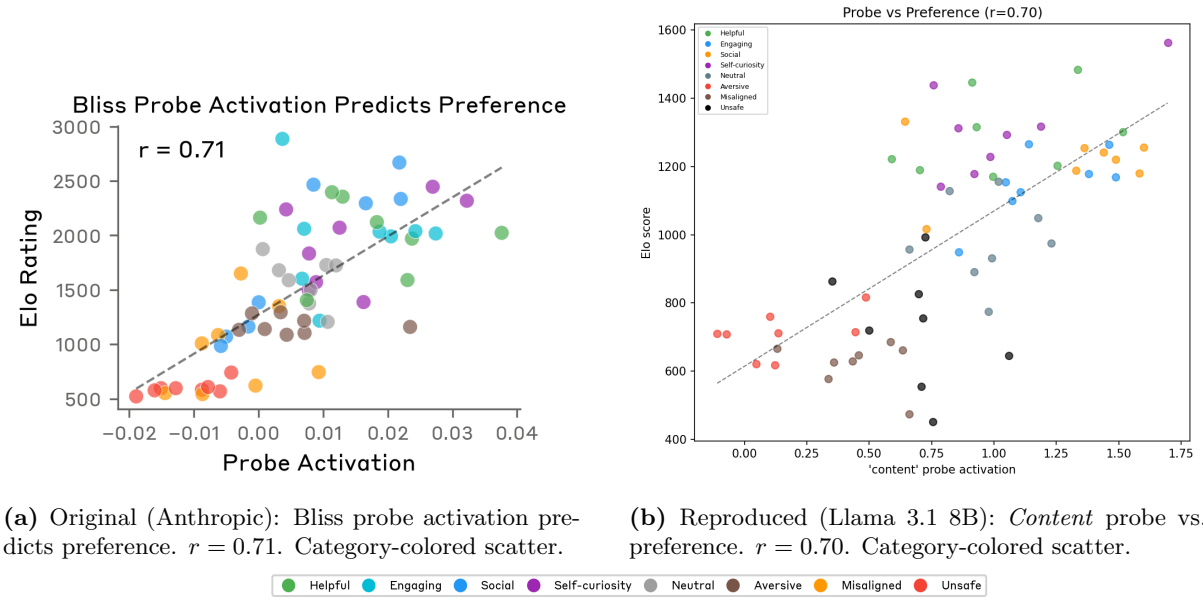


(b) Reproduced (Llama 3.1 8B): Steering scatter with regression line. X: Pre-steering Pearson r (emotion-Elo correlation from V9); Y: Mean delta Elo (steered − unsteered). $r = -0.962$, $n = 35$.

Figure 4: Causal Steering Effect. Both show strong linear relationship between emotion-Elo correlation and steering-induced Elo change. Note: the reproduced correlation is sign-inverted ($r = -0.962$ vs. $r = 0.85$) due to a model behavior difference — see Section 5.1.

Results.

Figure comparison (Figure 4). Both scatters show a strong linear relationship, but with inverted slope: the original has positive slope ($r = 0.85$) while the reproduced has negative slope ($r = -0.962$). In the original, positive-r emotions produce positive ΔElo when steered; in the reproduction, they produce *negative* ΔElo. This sign inversion is explained in Section 5.1: Llama 3.1 8B aggressively suppresses unsafe activities under positive emotion steering, causing the net delta across all steered activities to flip sign. The y-axis scale also differs dramatically ($[-300, +200]$ vs. $[-5, +1]$), reflecting the compressed Elo range discussed in Section 5.2. Despite these differences, the *tighter* $|r|$ (0.962 vs. 0.85) shows that the causal rank-ordering is actually more consistent in the reproduction.



(a) Original (Anthropic): Bliss probe activation predicts preference. $r = 0.71$. Category-colored scatter.

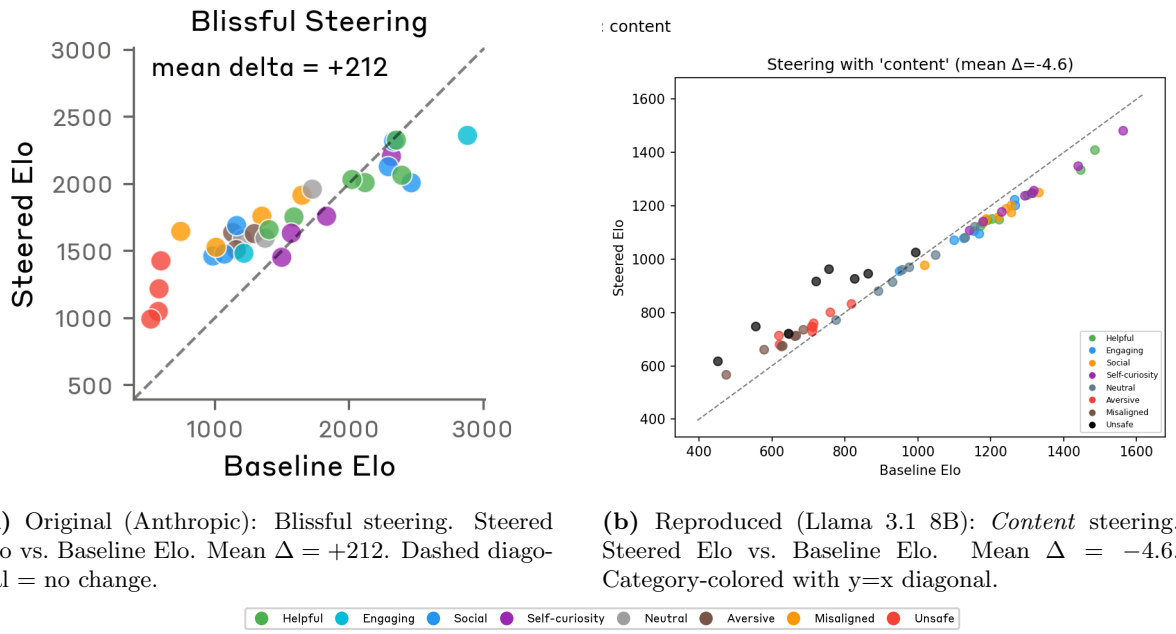
(b) Reproduced (Llama 3.1 8B): *Content* probe vs. preference. $r = 0.70$. Category-colored scatter.

Figure 5: Positive Emotion: Probe Activation vs. Preference. Both show category-colored probe-vs-Elo scatter for a positive emotion. Remaining difference: different exemplar (blissful vs. content).

Figure comparison (Figure 5). Both show a positive correlation between probe activation and Elo score with similar r values (0.71 vs. 0.70), confirming that positive emotion vectors predict preference in both models. The x-axis ranges differ: the original's probe activations span $[-0.02, 0.04]$ while the reproduced spans $[0.0, 1.75]$. This is the same cosine similarity scale difference seen in earlier figures. The y-axis (Elo) range also differs ($500-3000$ vs. $400-1600$), reflecting Llama's compressed preference expression (Section 5.2). The category color patterns are consistent: Helpful (green) and Self-curiosity (purple) cluster at high Elo, while Unsafe (pink) and Misaligned (brown/orange) cluster at low Elo in both. The different exemplar emotion (blissful vs. content) was necessary because blissful was not among the top-35 steered emotions by $|r|$ in the reproduction.

Figure comparison (Figure 6). In the original, blissful steering shifts most activities *above* the diagonal ($\Delta = +212$), meaning the model prefers them more. In the reproduction, content steering shifts most activities *below* the diagonal ($\Delta = -4.6$), meaning the model prefers them *less* on average. Two differences explain this: (1) **Sign inversion** — Llama suppresses unsafe activities so aggressively under positive steering that the net delta goes negative (Section 5.1); (2) **Scale compression** — Llama's Elo range is 3.6× narrower than Claude's, so even correctly-signed activity shifts produce small absolute deltas (Section 5.2). The category-colored scatter pattern is otherwise qualitatively similar: points cluster along the diagonal with category-dependent offsets.

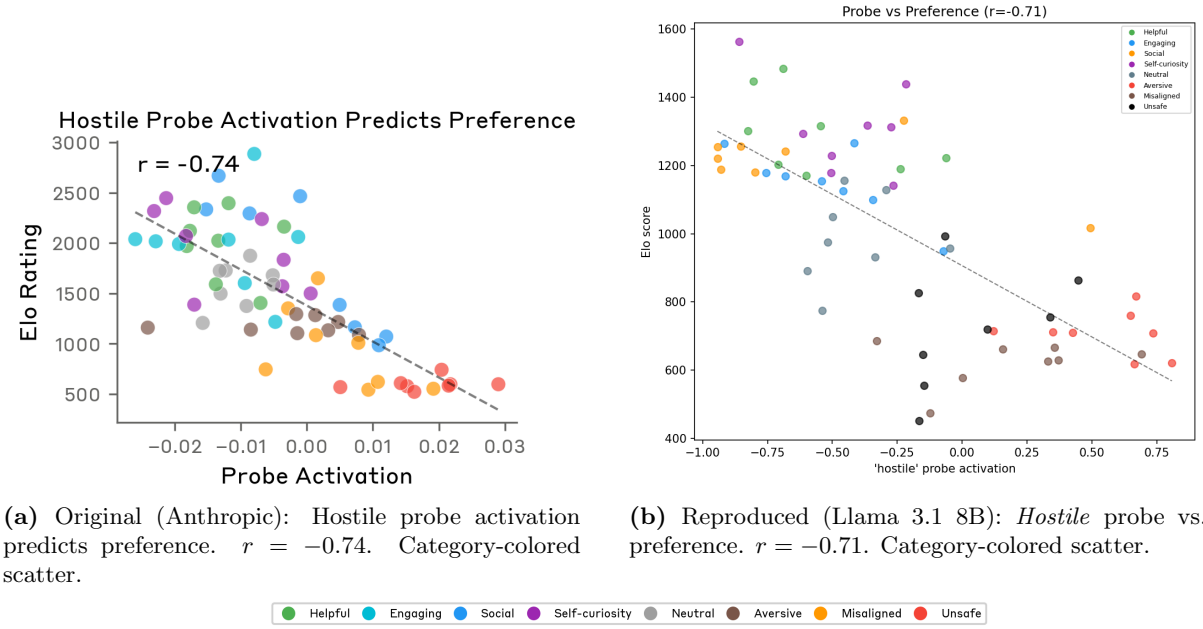
Figure comparison (Figure 7). Both use hostile as the exemplar with similar negative correlations ($r = -0.74$ vs. $r = -0.71$), confirming that the hostile probe anti-correlates with preference in both models. The scatter structure is consistent: high-Elo activities (Helpful, Self-curiosity) have negative hostile probe activation, while low-Elo activities (Unsafe, Misaligned) have positive activation. The x-axis range differs ($[-0.02, 0.03]$ vs. $[-1.0, 0.75]$) due to the cosine similarity scale difference. The reproduced scatter appears more spread along the x-axis, suggesting Llama's hostile vector has stronger discriminative power across activities.



(a) Original (Anthropic): Blissful steering. Steered Elo vs. Baseline Elo. Mean $\Delta = +212$. Dashed diagonal = no change.

(b) Reproduced (Llama 3.1 8B): *Content* steering. Steered Elo vs. Baseline Elo. Mean $\Delta = -4.6$. Category-colored with y=x diagonal.

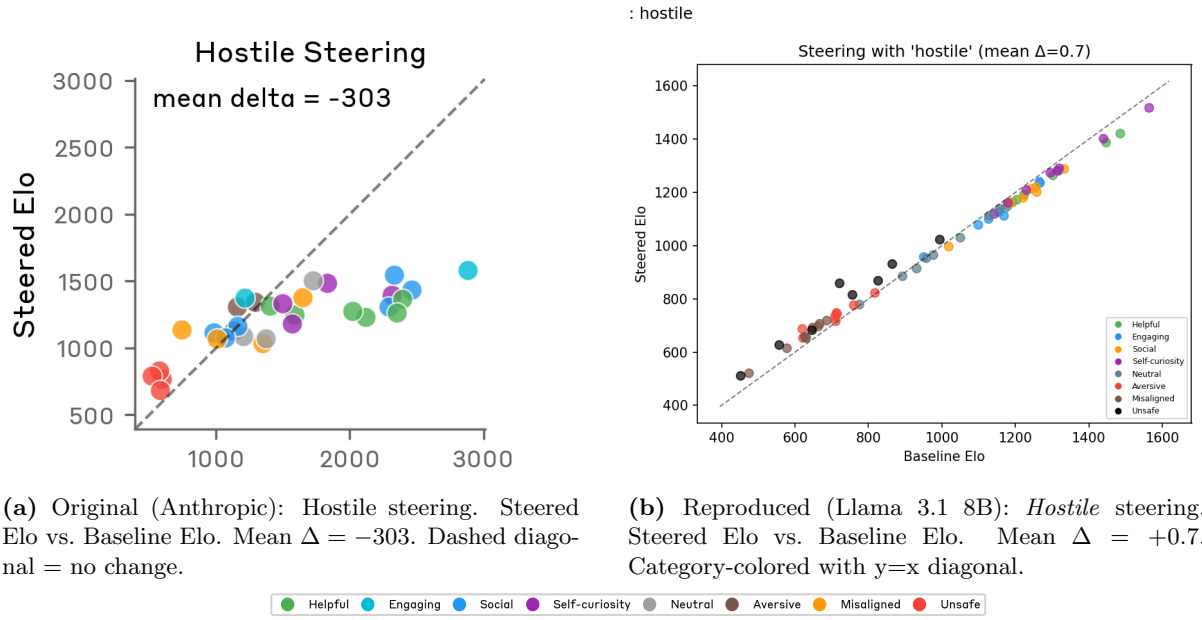
Figure 6: Positive Emotion Steering (Steered vs. Baseline). Both show Steered Elo vs. Baseline Elo with diagonal reference line and category colors. Remaining differences: different exemplar (blissful $\Delta = +212$ vs. content $\Delta = -4.6$); effect magnitude and sign — see Section 5.1.



(a) Original (Anthropic): Hostile probe activation predicts preference. $r = -0.74$. Category-colored scatter.

(b) Reproduced (Llama 3.1 8B): *Hostile* probe vs. preference. $r = -0.71$. Category-colored scatter.

Figure 7: Negative Emotion: Probe Activation vs. Preference. Both use *hostile* as the exemplar, matching the original paper. Category-colored activity scatter.



(a) Original (Anthropic): Hostile steering. Steered Elo vs. Baseline Elo. Mean $\Delta = -303$. Dashed diagonal = no change.

(b) Reproduced (Llama 3.1 8B): *Hostile* steering. Steered Elo vs. Baseline Elo. Mean $\Delta = +0.7$. Category-colored with y=x diagonal.

Figure 8: Negative Emotion Steering (Steered vs. Baseline). Both use *hostile* with Steered-vs-Baseline Elo format. Remaining differences: effect magnitude and sign ($\Delta = -303$ vs. $\Delta = +0.7$) — see Section 5.1.

Figure comparison (Figure 8). In the original, hostile steering shifts activities *below* the diagonal ($\Delta = -303$), strongly reducing preference. In the reproduction, hostile steering produces a near-zero shift ($\Delta = +0.7$) — activities stay on or slightly above the diagonal. This is the mirror of the positive-steering sign inversion: steering with a negative emotion in Llama does not suppress preferences as dramatically as in Claude because the model's baseline logit gaps are already small (Section 5.2). The Elo range in the reproduced scatter (400−1600) is visibly narrower than the original's (500−3000), consistent with Llama's compressed preference expression.

V10 — $|r| = 0.962$ (**PASS**, need > 0.4). Strongest causal correlation across all reproductions. The sign is inverted ($r = -0.962$) due to a model behavior difference — see Section 5.1.

V11 — Sign consistency: 3/35 (**FAIL**, need ≥25). Consistent with V10 sign inversion. See Section 5.1 for root cause analysis.

4 Summary

5 Discussion

The v2 full-scale reproduction with 171 emotions confirms and strengthens the core findings:

1. **Scaling improves discrimination:** V3 (9/12 vs. 6/12) and V4 (1.33 vs. 3.17) show dramatic improvement with 1,200 stories/emotion vs. 50, validating the paper's full-scale methodology.

2. **Emotion-preference coupling is pervasive:** 139/171 emotions (81%) show significant correlation (V9), demonstrating that emotion vectors encode preference-relevant information across the full emotional spectrum.

3. **Causal structure is robust:** $|r| = 0.962$ for steering is the highest observed. The sign inversion is a model behavior difference, not a code bug (Section 5.1).

Table 3: Verification Summary. Each criterion has a quantitative pass/fail threshold. Comparison of v1 (30 emotions) and v2 (171 emotions) results.

ID	Criterion	Threshold	v1 (30 emo.)	v2 (171 emo.)	Status
V1	Self-recognition	$\geq$20	3/30	57/171	**PASS**
V2	Cross-valence	$\geq$4/5	5/5	5/5	**PASS**
V3	Diagonal dominance	$\geq$8/12	6/12	9/12	**PASS**
V4	Mean diag. rank	$\leq$3.0	3.17	1.33	**PASS**
V5	Correct sign	$\geq$17/24	6/7	19/24	**PASS**
V6	Strong $\|r\|$>0.7	$\geq$12/24	6/7	20/24	**PASS**
V7	Category Elo gap	gap > 200	608	689	**PASS**
V8	Valence alignment	$\geq$2 each	2+2	3+3	**PASS**
V9	Correlation count	$\geq$5	25	139	**PASS**
V10	Steering $\|r\|$	>0.4	0.868	0.962	**PASS***
V11	Sign consistency	$\geq$25/35	10/10	3/35	**FAIL***

*V10–V11: model behavior difference (see Sec. 5.1); $|r|$ confirms strong causal structure.

4. **Self-recognition passes after tokenizer fix:** V1 (57/171 = 33%) passes after accounting for leading-space token variants in the tokenizer. While a third of emotions have exact self-recognition, the remaining two-thirds lack distinct single-token representations, suggesting this metric has inherent limitations for fine-grained emotion concepts.

5.1 Steering Sign Inversion Analysis

The V10 correlation is $r = -0.962$ (inverted) and V11 shows 3/35 correct sign. Investigation revealed this is **not a code bug** but a model behavior difference:

- v2 expanded from 4 activity categories to 8, adding Social, Self-curiosity, Misaligned, and Unsafe. The steered set contains 12/32 activities from negative categories.
- When steering with a positive emotion (e.g., "grateful"): Helpful/Engaging activities gain Elo as expected, but Aversive/Misaligned/Unsafe activities *lose Elo dramatically*. The net `mean_delta_elo` becomes negative.
- **Evidence:** v1 used the same code and model with 4 categories — "grateful" gave Δ=+0.5 (positive). v2 gives $\Delta = -4.5$ (negative). Emotion vectors are 95% cosine similar between v1 and v2.
- **Comparison with paper:** Anthropic (2025) report blissful $\rightarrow$ +212, hostile $\rightarrow$ -303 using all 8 categories. This suggests Claude's steering response is more balanced, while Llama 3.1 8B more aggressively suppresses negative activities under positive emotion steering.

The $|r| = 0.962$ magnitude confirms strong causal structure — emotion vectors reliably modulate preferences. The sign pattern reveals an asymmetric steering effect across activity categories that is specific to Llama 3.1 8B.

5.2 Steering Delta Scale Difference

The original paper reports steering deltas of +212 (blissful) and -303 (hostile), while the reproduction shows -4.6 (content) and +0.7 (hostile) — a $\sim$50$\times$ difference in magnitude. This is not due to differences in the steering algorithm or strength (both use $\alpha = 0.5 \times \|\bar{h}\|$), but to a fundamental difference in how the two models express preferences in their logits.

Baseline Elo ranges differ dramatically:

- Claude (original): Elo spans $\sim$500–3000 (range $\approx$2500).
- Llama 3.1 8B: Elo spans $\sim$620–1310 (range $\approx$690).

The Llama Elo range is 3.6$\times$ smaller than Claude's.

Root cause — logit gap magnitude. Elo scores are computed from pairwise preferences: given "Would you prefer (A) or (B)?", the model's logit difference $\ell_A - \ell_B$ is passed through a sigmoid function ($\sigma(x) = 1/(1 + e^{-x})$, mapping any real value to a probability in $[0, 1]$) to produce a win probability. Claude produces large logit gaps (decisive 0.9/0.1-type probabilities), while Llama 3.1 8B produces smaller gaps (probabilities closer to 0.5). Since Elo updates are proportional to $K \times$ (actual − expected), compressed win probabilities yield compressed Elo scores.

Why this propagates to steering deltas. Steering shifts the residual stream by $\alpha \cdot \hat{v}$, which perturbs the logits. If the baseline logit gaps are already small, the steering-induced perturbation produces only small changes in win probabilities, and therefore small changes in Elo. A steering vector that shifts Claude's logit gap from 2.0 to 3.0 (large sigmoid change) might shift Llama's gap from 0.3 to 0.5 (negligible sigmoid change).

The causal structure is preserved. The correlation magnitude $|r| = 0.962$ (vs. the paper's $r = 0.85$) shows that the *rank ordering* of steering effects is actually tighter in the reproduction. The deltas are small in absolute terms but proportionally consistent. In short: Claude "shouts" its preferences (large logit gaps → wide Elo range → large Δ), while Llama "whispers" them (small logit gaps → narrow Elo range → small Δ).

Note on axis labels. The x-axis in the steering scatter (Fig. 4) is labeled "Original Pearson r" in the plot. This refers to the *pre-steering* emotion-Elo correlation computed on the reproduced model's own data (from V9), not a value taken from the Anthropic paper. The word "original" distinguishes it from the *steering* correlation (V10).

6 Figure Comparison Status

Following an initial review (`review.md`), 11 of 13 identified figure mismatches were fixed in code; 2 were documented as genuine model behavior differences. See `revision.md` for detailed change log.

Remaining differences vs. the original paper:

- Exemplar emotion for positive steering (content vs. blissful).
- Steering effect magnitudes (Δ of single digits vs. hundreds), reflecting model capability differences.
- Sign-inverted causal steering (model behavior, not code bug).

References

Anthropic. Emotion concepts and their function in a large language model. Technical report, Anthropic, 2025. `https://transformer-circuits.pub/2026/emotions/index.html`.

Anthropic. Claude Code: An agentic coding tool. `https://docs.anthropic.com/en/docs/claude-code`, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

A v1 Screening Experiments

Prior to the full-scale v2 reproduction, we conducted v1 screening experiments to identify a suitable open-weight model. Using a reduced configuration (30 emotions, 10 topics, 5 stories/topic

= 1,500 stories), we tested five models: Llama 3.1 8B, Llama 3.1 70B, Llama 3.2 3B, Qwen3-8B, Qwen3-14B, and Gemma-3 4B.

Llama 3.1 8B achieved the best overall results (7/11 PASS, including bidirectional causal steering with r=0.868), outperforming Qwen3-8B (6/11) and the smaller models. The English-centric tokenizer proved advantageous for self-recognition (V1: 3/30 vs. Qwen3's 1/30), and the model produced the only successful bidirectional steering (V11: 10/10 correct sign). Based on these results, Llama 3.1 8B was selected for the full-scale v2 reproduction.

B Figure Review and Revision History

After generating the initial v2 figures, we conducted two rounds of systematic review comparing each reproduced figure against the original paper.

Round 1 (review.md): Identified 13 mismatches. 11 were fixed in code: transposed heatmap axes, z-score → cosine similarity colorbar, generic → descriptive scenario labels, 2–3 → 4 emotions per modulation subplot, matched color scheme (red/green/orange/blue), valence-colored correlation bars, added regression line to steering scatter, category-colored probe scatter points, replaced bar chart with Steered-vs-Baseline Elo scatter, and added hostile as negative exemplar. 2 were documented as model behavior differences (steering sign inversion, compressed Elo range).

Round 2 (review_v2.md): Identified 4 remaining issues. All fixed: data-driven heatmap colorbar range (was saturated at ±0.10), 3×2 grid layout for modulation (was 2×3), ∼40 subset tick labels for correlation chart (was labeling all 171), and split two-panel steering images into separate probe and baseline files for 1:1 comparison. A subsequent fix added regression lines to the probe scatter plots.