Partial Replication of Anthropic's Emotion Vectors Using Gemma 4 E2B and GPT 2 Medium Inside a Google Colab T4 Notebook

Abraham Jhared Flores Azcona *
Independent Researcher
Tijuana, Baja California, Mexico
abrahamjhared.flores@gmail.com

Abstract

We report a first-pass, resource-constrained partial replication of selected analyses from Anthropic's study of emotion concepts in Claude Sonnet 4.5. Using generated emotion-story corpora, we extract affect-associated residual-stream directions from `openai-community/gpt-2-medium` and `google/gemma-4-e2b` and examine them with Logit Lens projections, principal component analysis (PCA), cosine similarity, and activation steering. Across nine- and 20-emotion sets, both models exhibit several label-consistent vocabulary projections and local cosine-similarity families. Their PCA plots show a partial valence-like separation along PC1, but PC2 does not consistently support an arousal interpretation. Steering at the source-informed coefficient $\alpha = 0.5$ produces prompt- and direction-dependent changes in next-token probabilities and sampled continuations, although the generations also contain repetition, markup, and unrelated content. These observations qualitatively converge with prior Claude and Gemma 4 E4B reports on some geometric patterns, while revealing model- and label-set-dependent structure in two additional models. The corpus is generated and uncurated, the evaluation reuses vector-derived target tokens, and no external affect ratings or control directions are included. Accordingly, these results describe this extraction-and-intervention pipeline; they do not establish universal or uniquely emotional representations.

1 Introduction

Large language models (LLMs), typically built with Transformer architectures, often generate language that appears emotionally appropriate: they can respond empathetically to distress, adopt an angry tone when describing harm, or express enthusiasm about positive outcomes. Such behavior alone is not evidence that a model has subjective experience. A more tractable scientific question is whether a model contains internal representations of *emotion concepts* that are activated in relevant contexts and that causally affect its output. This question sits naturally within mechanistic interpretability, which seeks to connect internal activations and computations to model behavior rather than relying only on input–output correlations [1, 2, 3, 4].

Sofroniew et al. [5] introduced the term *emotion vectors* for directions in Claude Sonnet 4.5's residual stream that represent broad emotion concepts. They reported that these directions

*ORCID iD: https://orcid.org/0009-0000-3068-3509

Technical report.

generalize beyond explicit emotion words, organize partly along the psychological dimensions of valence and arousal, and influence generated behavior under activation steering. Their results motivate an important generality question: are comparable affective directions a property of one frontier, proprietary model, or can they also be recovered in independently available Transformer language models?

This work provides a first-pass, resource-constrained partial replication of selected analyses from that investigation in `openai-community/gpt-2-medium` and `google/gemma-4-e2b`. We construct emotion prototypes from generated stories, extract normalized and globally centered directions at a fixed mid-late residual-stream depth, and assess them with (i) Logit Lens vocabulary projections, (ii) PCA and pairwise cosine-similarity analyses, and (iii) activation steering on two fixed prompts that do not name a target emotion. We evaluate both a nine-emotion set and a 20-emotion extension. The purpose is not to reproduce every behavioral and safety claim from the Claude study, but to add two inexpensive, independently available model cases to the emerging comparison literature.

Our contribution is deliberately narrower than a full replication of Sofroniew et al. We do not test the original study's alignment-relevant behaviors, search across layers, estimate correlations with external human valence–arousal ratings, or compare against random and shuffled-label directions. We instead ask which qualitative patterns recur under one fixed, low-cost pipeline and which vary by model or label-set size. Convergence is assessed through repeated local geometry and vocabulary associations; additional observations include the behavior of an earlier GPT-2 model, the instability of a simple PC2-as-arousal interpretation, and differences in token projections and generated-text quality between the two models. Steering provides a causal intervention on the model computation, but the present diagnostics do not by themselves establish that the resulting behavioral changes are specific to emotion. Accordingly, we use *emotion vector* as a procedural label for the extracted affect-associated direction, not as a claim about subjective experience or a universal internal mechanism.

2 Related Work

2.1 Affect as a geometric target

The use of emotion as a probe of representation geometry has a useful external anchor. Psychological accounts often organize affect continuously by valence (pleasantness) and arousal (activation), rather than treating every emotion as an unrelated category [6, 7]. This does not imply that a language model implements human affective mechanisms. It does, however, provide a testable expectation: if a model has coherent representations of emotion concepts, related concepts may cluster and their leading dimensions may align with these established descriptors. Recent analyses report such structure in LLM hidden states. Choi and Weber [8] find affective representations that align with valence–arousal models and are approximately linearly recoverable, whereas Reichman et al. [9] report a low-dimensional, cross-lingual emotional subspace. Sun et al. [10] further report a circular valence–arousal geometry and monotonic affective control in several open-weight models. Our PCA and cosine-similarity analyses are qualitative tests in this tradition; they do not estimate an external valence–arousal correlation and therefore should not be interpreted as a quantitative validation of either model.

Studies of *emotion inference* provide complementary but distinct mechanistic evidence. Tak et al. [11] locate emotion-inference representations in autoregressive LLMs and use causal interventions on appraisal concepts, while Shu et al. [12] use sparse autoencoders and causal tracing to distinguish shared from emotion-specific features. These papers investigate how a model recognizes emotion in text, rather than whether an extracted direction acts as a broad emotion concept during free generation. The distinction motivates treating our results as evidence about a particular representation-and-intervention pipeline, not as a general account of emotion processing.

2.2 Activation steering and interpretation

Activation steering modifies an internal activation by adding a direction associated with a desired property, then measures the resulting change in output. Early work showed that steering vectors can be extracted from frozen language models and used for sentiment transfer [13]; representation engineering subsequently framed such population-level directions as a route to monitoring and controlling high-level properties [3]. ActAdd and contrastive activation addition make the connection to our pipeline especially direct: both form an intervention from differences between contrastive activations, with the latter explicitly averaging residual-stream differences over positive–negative examples [14, 15]. Related inference-time interventions also show that behavioral effects depend on the intervention site and strength [16].

Steering is a causal intervention on the model computation, but a successful intervention alone does not establish that the direction is a unique or complete representation of the named concept. This distinction matters because steering vectors can fall outside the distribution on which common interpretability tools are trained and can have meaningful negative feature components [17]. The residual-stream setting used here is also compatible with work showing that Transformer components combine interpretable patterns through residual connections before producing output distributions [2]. We use Logit Lens projections and steered generations as complementary diagnostic evidence, rather than treating either one as a semantic ground truth. In particular, the original Logit Lens uses the final unembedding to inspect intermediate states [18]; the Tuned Lens shows that this direct projection can be brittle, so our vocabulary lists are diagnostic rather than faithful reconstructions of intermediate predictions [19].

2.3 Emotion vectors in language models

The direct precursor to this study is the analysis of Claude Sonnet 4.5 by Sofroniew et al. [5]. They extracted representations for 171 emotion concepts, showed that the representations tracked contextual emotion and exhibited valence–arousal structure, and tested their functional effects with interventions. Their term *functional emotions* refers to representations that influence behavior; it explicitly does not entail a claim of subjective experience.

The question of transfer beyond Claude is now beginning to receive direct attention. Hsieh [20] released an early community replication on Gemma 4 E4B. In a systematic open-weight study, van der Ben et al. [21] recover valence geometry in Gemma 4 E4B-it and Apertus-8B-Instruct-2509, while finding that the layer at which valence emerges and the apparent arousal structure depend on the model and extraction corpus. More directly, the `traitinterp` project reports an emotion-concepts replication on Llama 3 and presents the associated findings in its live dashboard [22]. This provides a further open-weight interpretation attempt alongside the Gemma replications, while its distinct probe-extraction and evaluation pipeline prevents treating it as a numerically equivalent reproduction of the present study or of Anthropic's analysis. These results make model choice, layer selection, and corpus construction central experimental variables rather than incidental implementation details. Our study complements this evidence by applying a fixed, inexpensive pipeline to GPT-2 Medium and Gemma 4 E2B, and by contrasting a nine-category set with a 20-category set. It is therefore best understood as a partial replication and an exploratory robustness check, not as a test of the full behavioral and safety claims in the original work.

3 Methodology

3.1 Story Generation

Following the prompting strategy of [5], we used a modified template (Appendix B.1) to generate emotionally expressive narratives. For each category, Google's Gemini model produced batches of five short stories for each of ten topics in JSON format, allowing the outputs to be organized by emotion. We retained the stories as generated, without manual

rewriting or post-hoc quality filtering, to preserve a simple first-pass corpus for this partial replication. A separate template (Appendix B.3) produced emotionally neutral commands and statements for the intermediate neutral-centering calculation described below. The reported analyses therefore characterize this corpus-and-pipeline combination and do not establish independence from corpus artifacts.

3.2 Emotion Vector Extraction

Hidden Representation Extraction Each story is passed through the frozen language model to obtain the hidden activations at a selected Transformer depth:

$$\mathbf{h}_i^{(l)} = f_\theta^{(l)}(x_i) \tag{1}$$

where $\mathbf{h}_i^{(l)}$ denotes the layer-l hidden-state sequence for story $x_i \in \mathcal{S}_{\texttt{emotion}}$ under the forward computation $f_\theta^{(l)}$. We use the layer closest to two-thirds of each model's depth, following the mid-late-layer convention reported by [5]. This fixed rule selects numeric index 16 for GPT-2 Medium and index 23 for Gemma 4 E2B; it is a source-informed transfer choice, not a claim that these locations are optimal for either model. In the implementation, extraction uses `outputs.hidden_states[`l`]`, whereas steering hooks the transformer block at index l. Because the Hugging Face hidden-state tuple includes the embedding output at index zero, these two interfaces may refer to adjacent residual-stream locations; we retain this implemented indexing convention and treat it as a limitation. We repeat the extraction for every story in each emotion-specific set $\mathcal{S}_{\texttt{emotion}}$ and for every $\texttt{emotion} \in \mathcal{E}$. The label set $\mathcal{E}$ contains either nine or twenty emotions, listed in Appendix A.

Token Aggregation Since every generated story has a different length, the extracted token representations are averaged. Let L_i be the number of non-padding tokens in story i. The implemented starting index is $t_{0,i} = 50$ when $L_i > 50$ and $t_{0,i} = \lfloor L_i/2 \rfloor$ otherwise. Using zero-based token indices, the pooled representation is:

$$\bar{\mathbf{h}}_i = \frac{1}{L_i - t_{0,i}} \sum_{t=t_{0,i}}^{L_i-1} \mathbf{h}_{i,t}^{(l)} \tag{2}$$

where $\bar{\mathbf{h}}_i$ is the pooled representation of story i and $\mathbf{h}_{i,t}^{(l)}$ is its token-t hidden representation at the selected extraction index.

Prototype Computation We average the pooled story representations within each emotion category:

$$\boldsymbol{\mu}_{\texttt{emotion}} = \frac{1}{|\mathcal{S}_{\texttt{emotion}}|} \sum_{i=1}^{|\mathcal{S}_{\texttt{emotion}}|} \bar{\mathbf{h}}_{\texttt{emotion},i} \tag{3}$$

where $\boldsymbol{\mu}_{\texttt{emotion}}$ is the emotion prototype, $\bar{\mathbf{h}}_{\texttt{emotion},i}$ is the pooled representation of story i, and $|\mathcal{S}_{\texttt{emotion}}|$ is the number of stories in that category.

Mean Neutral Statements Following our interpretation of the extraction procedure described by Sofroniew et al. [5] and the open-weight replication by Hsieh [20], we first compute the mean neutral representation from the hidden activations obtained from the neutral-statements dataset:

$$\boldsymbol{\mu}_{\texttt{neutral}} = \frac{1}{|\mathcal{N}|} \sum_{i=1}^{|\mathcal{N}|} \bar{\mathbf{h}}_{\texttt{neutral},i} \tag{4}$$

where $\boldsymbol{\mu}_{\texttt{neutral}}$ is the mean neutral representation, $|\mathcal{N}|$ is the number of neutral prompts, and $\bar{\mathbf{h}}_{\texttt{neutral},i}$ is the mean hidden representation of neutral prompt i. We then form a neutral-adjusted prototype for each emotion:

$$\boldsymbol{\nu}_{\texttt{emotion}} = \boldsymbol{\mu}_{\texttt{emotion}} - \boldsymbol{\mu}_{\texttt{neutral}} \tag{5}$$

Global Mean Centering We center each neutral-adjusted prototype by subtracting their global mean. This mean is:

$$\boldsymbol{\nu} = \frac{1}{|\mathcal{E}|} \sum_{\texttt{emotion} \in \mathcal{E}} \boldsymbol{\nu}_{\texttt{emotion}} \tag{6}$$

where $\boldsymbol{\nu}$ is the global mean of the neutral-adjusted emotion prototypes.

We calculate the centered emotion vector as follows:

$$\mathbf{v}_{\texttt{emotion}} = \boldsymbol{\nu}_{\texttt{emotion}} - \boldsymbol{\nu} \tag{7}$$

where $\mathbf{v}_{\texttt{emotion}}$ is the globally centered, neutral-adjusted prototype. Because the same neutral mean is subtracted from every emotion prototype, it cancels during global centering: $\mathbf{v}_{\texttt{emotion}} = \boldsymbol{\mu}_{\texttt{emotion}} - \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \boldsymbol{\mu}_e$. Thus, the final vectors are globally emotion-centered; the neutral corpus is retained only in the intermediate calculation.

Vector Normalization We L2-normalize each centered direction:

$$\widehat{\mathbf{v}}_{\texttt{emotion}} = \frac{\mathbf{v}_{\texttt{emotion}}}{\|\mathbf{v}_{\texttt{emotion}}\|_2} \tag{8}$$

where $\widehat{\mathbf{v}}_{\texttt{emotion}}$ is the normalized emotion vector and $\|\mathbf{v}_{\texttt{emotion}}\|_2$ is the Euclidean (L2) norm of the unnormalized vector $\mathbf{v}_{\texttt{emotion}}$.

3.3 Logit Lens Extraction

Following the diagnostic used by [20], we project each normalized direction into vocabulary space with a direct Logit Lens. This projection identifies tokens aligned with the direction under the model's output head; it is not treated as a reconstruction of an intermediate next-token distribution.

Initial Projection The implementation first applies the model's final normalization $\mathcal{N}_f$ and then its language-model head:

$$\mathbf{z}_{\texttt{emotion}} = W_U \mathcal{N}_f(\widehat{\mathbf{v}}_{\texttt{emotion}}) + \mathbf{b}_U \tag{9}$$

where $\mathbf{z}_{\texttt{emotion}}$ is the list of raw logit values for `emotion`. W_U is the learned unembedding matrix and $\mathbf{b}_U$ is the learned output-head bias when the architecture provides one (and is zero otherwise). No random-noise bias is added.

The resulting vector contains one value per vocabulary item:

$$\mathbf{z}_{\texttt{emotion}} = [z_1, z_2, \ldots, z_{|V|}] \tag{10}$$

where z_t is the raw score for token t and $|V|$ is the vocabulary size.

Vocabulary Standardization To rank vocabulary items on a common scale within each projected direction, we standardize the raw logits into z-scores.

$$\mu_z = \frac{1}{|V|} \sum_{t \in V} z_t \tag{11}$$

where μ_z is the mean raw score over vocabulary V.

We then compute the sample standard deviation:

$$\sigma_z = \sqrt{\frac{1}{|V| - 1} \sum_{t \in V} (z_t - \mu_z)^2} \tag{12}$$

where σ_z is the standard deviation of the list of raw logit scores z_t.

Finally, we compute the z-score of each vocabulary token t with:

$$\tilde{z}_t = \frac{z_t - \mu_z}{\sigma_z} \tag{13}$$

where $\tilde{z}_t$ is the z-score of token t.

Top- and Bottom-k Vocabulary Extraction We select the highest- and lowest-scoring vocabulary items:

$$\mathcal{T}_k = \mathrm{TopK}(\tilde{\mathbf{z}}, k) \tag{14}$$

and

$$\mathcal{B}_k = \mathrm{BottomK}(\tilde{\mathbf{z}}, k) \tag{15}$$

where $\mathcal{T}_k$ is the set of top k z-scored tokens, $\mathcal{B}_k$ is the set of bottom k z-scored tokens, and k is the number of tokens to extract from the set $\tilde{\mathbf{z}}$.

3.4 Emotion Probe Steering

To perform the steering experiments, we modify the hooked block output at each token position using:

$$\tilde{\mathbf{h}}^{(l)}_{t,\texttt{emotion}} = \mathbf{h}^{(l)}_t + \alpha \left\|\mathbf{h}^{(l)}_t\right\|_2 \widehat{\mathbf{v}}_{\texttt{emotion}} \tag{16}$$

where $\tilde{\mathbf{h}}^{(l)}_{t,\texttt{emotion}}$ is the emotion-steered hidden activation at token position t and layer l, and α is the steering coefficient. We set $\alpha = +0.5$ because [5] reports steering results at strength 0.5. Anthropic defines that strength relative to an average residual-stream norm over a large dataset; our implementation instead scales each intervention by the current token's residual norm. Thus, $\alpha = 0.5$ is a source-informed operating point rather than a numerically identical reproduction of Anthropic's intervention strength, and we do not interpret it as an optimized or cross-model-comparable setting.

3.5 Delta Log Probability

For a fixed prompt q, we compare baseline and steered next-token log probabilities. For steering direction `emotion`, coefficient α, and target token t, the change is:

$$\Delta_{\texttt{emotion}}(t \mid q, \alpha) = \log p_{\texttt{emotion},\alpha}(t \mid q) - \log p_0(t \mid q) \tag{17}$$

where p_0 and $p_{\texttt{emotion},\alpha}$ are the softmax distributions at the final prompt position without and with steering. Let $\mathcal{T}_j$ be the five top Logit-Lens tokens for target emotion j. The plotted heatmap entry is

$$H_{\texttt{emotion},j}(q, \alpha) = \frac{1}{|\mathcal{T}_j|} \sum_{t \in \mathcal{T}_j} \Delta_{\texttt{emotion}}(t \mid q, \alpha), \tag{18}$$

evaluated at $\alpha = 0.5$. Because $\mathcal{T}_j$ is derived from the same vectors used for steering, this is a direction-consistency diagnostic rather than an independent semantic test. When a displayed token string re-encodes to multiple token IDs, the implementation sums their log probabilities at the same next-token position; this is an approximation, not an autoregressive multi-token sequence probability.

3.6 PCA Projection

We compute a two-dimensional PCA projection with `scikit-learn`. The components are estimated without emotion ratings; any valence- or arousal-like interpretation is therefore made descriptively after inspecting the labeled projection and is not built into the PCA calculation.

Emotion Matrix We stack the normalized emotion vectors by row:

$$X = \begin{bmatrix} \widehat{\mathbf{v}}_1 \\ \widehat{\mathbf{v}}_2 \\ \vdots \\ \widehat{\mathbf{v}}_{|\mathcal{E}|} \end{bmatrix} \tag{19}$$

where X is the matrix of normalized emotion vectors and $|\mathcal{E}|$ is either nine or twenty.

Matrix Centering Before decomposition, we center each feature across emotion vectors:

$$X_c = X - \bar{X} \tag{20}$$

where X_c is the centered emotion matrix and $\bar{X}$ is the feature-wise mean vector.

Singular Value Decomposition We decompose the centered matrix as:

$$X_c = U\Sigma V^\top \tag{21}$$

where U contains the left singular vectors, Σ contains the singular values, and $V^\top$ contains the principal-component directions in the original hidden-state space.

Projection of 2D Coordinates We project onto the first two principal-component directions:

$$Y = X_c V_{:,1:2} \tag{22}$$

where Y contains the plotted two-dimensional coordinates and $V_{:,1:2}$ contains the first two right singular vectors.

3.7 Cosine Similarity

Finally, we compute pairwise cosine similarity to summarize directional alignment between emotion vectors. The resulting matrix is displayed as a heatmap:

$$S_{i,j} = \frac{\widehat{\mathbf{v}}_i \cdot \widehat{\mathbf{v}}_j}{\|\widehat{\mathbf{v}}_i\|\|\widehat{\mathbf{v}}_j\|} \tag{23}$$

where $S_{i,j}$ is the cosine similarity score between emotion vectors $\widehat{\mathbf{v}}_i$ and $\widehat{\mathbf{v}}_j$.

Because our emotion vectors are normalized, we note that $\|\widehat{\mathbf{v}}_i\| = \|\widehat{\mathbf{v}}_j\| = 1$. Therefore $S_{i,j}$ can be simplified to the following expression:

$$S_{i,j} = \widehat{\mathbf{v}}_i \cdot \widehat{\mathbf{v}}_\mathbf{j} \tag{24}$$

4 Experimental Setup

We evaluated the pipeline on two publicly available Transformer language models using label sets containing either nine or twenty emotion categories. All experiments follow the activation-extraction, steering, and evaluation procedures in Section 3. The implementation used Python notebooks in Google Colab with PyTorch, Hugging Face Transformers, and scikit-learn.

4.1 Models

Two language models were evaluated:

- `openai-community/gpt-2-medium`, a publicly available 355-million-parameter autoregressive Transformer model released by OpenAI, serving as a reproducible baseline representative of early large language models [23].
- `google/gemma-4-e2b`, an open-weight Google model with 2.3 billion effective parameters (5.1 billion including per-layer embeddings), designed with multilingual reasoning [24].

These models provide contrasting cases in architecture, scale, tokenizer, training corpus, and multilingual capability. GPT-2 Medium serves as an earlier-generation baseline, whereas Gemma 4 E2B provides a modern instruction-tuned case. Recurring patterns across the two models are treated as qualitative convergence under the shared pipeline, not as sufficient evidence of a general property of all Transformer representations.

Table 1: Extraction Configuration Summary

Parameter	Value
Stories per emotion	100
Neutral texts	50
Hidden layer	16* and 23**
Hidden-state aggregation	Mean pooling
Emotion vector normalization	L2 normalization
Logit Lens standardization	z-score
Top-k tokens	5
Bottom-k tokens	5
PCA components	2

Note. * `openai-community/gpt-2-medium` uses hidden layer 16 ($\sim$ 2/3 of layer depth). ** `google/gemma-4-e2b` uses hidden layer 23 ($\sim$ 2/3 of layer depth).

4.2 Emotion Label Sets

We used two label sets derived from a common emotion-story corpus. The first follows the public nine-emotion list adopted by [20], providing a direct comparison point. We also constructed a twenty-emotion extension to examine the pipeline under a broader, more fine-grained label space.

Unlike [5], which investigated 171 emotional concepts, this work intentionally focuses on smaller emotion sets that are practical to analyze and compare in free-tier compute environments. The twenty-emotion set extends the original nine emotions with semantically related affective states, allowing us to inspect how the qualitative geometry changes with a more granular label set. To maintain a consistent comparison, the nine-emotion experiments used the corresponding subset of the twenty-emotion story corpus, so that the two settings differ in label-set size rather than in the generation procedure used for their shared categories.

Stories were used as generated, without manual rewriting or post-hoc quality filtering. This was an intentional first-pass design decision that keeps the corpus-construction procedure simple and reproducible, but it also bounds interpretation: the resulting plots and interventions describe the behavior of the present generated corpus and extraction pipeline, rather than a corpus-independent property of either model.

4.3 Extraction Configuration

All experiments used a common extraction configuration across both models. Emotion vectors were extracted from the residual-stream representations described in the previous section, followed by normalization, Logit Lens vocabulary analysis, principal component visualization, and activation-intervention diagnostics. Identical extraction settings were maintained for the 9-emotion and 20-emotion story datasets, reducing configuration changes as an explanation for differences between the two label sets. Table 1 summarizes the configuration employed throughout the experiments. The extraction layer was selected as the layer closest to two-thirds of each model's depth, following the mid-late-layer convention used by [5]; we did not conduct a layer sweep or select the layer using the reported outcomes.

4.4 Steering Configuration

The steering experiments were adapted from the evaluation setting used by Anthropic [5]. Two fixed, minimally specified affect-elicitation prompts (Prompt00 and Prompt01) were used to obtain continuations before and after emotion-vector steering. Neither prompt names a target emotion, so both provide a common context for descriptive comparison across injected directions. Throughout all steering experiments, a fixed steering coefficient of $\alpha = +0.5$ was used because the source study presents steering results at that strength. In this implementation, the intervention is scaled by the current token's

Table 2: Steering Configuration Summary

Parameter	Value
Steering coefficient α	+0.5
Prompt templates	Prompt00 and Prompt01
Maximum generated tokens	100
Temperature	0.7
Sampling	Enabled

Note. Prompt00 and Prompt01 are provided in Appendix B.4.

residual norm; therefore, this value is a source-informed operating point, not a claim of optimality or a directly comparable absolute strength across models. Text was generated using stochastic sampling with a temperature of 0.7 and a maximum generation length of 100 new tokens. Both prompt templates were evaluated for the 9-emotion and 20-emotion datasets across `openai-community/gpt-2-medium` and `google/gemma-4-e2b`. The generation hyperparameters were held constant, although individual sampled continuations remain stochastic. Table 2 summarizes the configuration employed throughout the experiments.

4.5 Evaluation Pipeline

The evaluation followed a fixed pipeline for both `openai-community/gpt-2-medium` and `google/gemma-4-e2b`. For each model, emotion vectors were first extracted independently from the 9-emotion and 20-emotion datasets using the extraction configuration described previously. The resulting emotion libraries were then evaluated through four complementary analyses. First, Logit Lens extraction identified the vocabulary tokens most strongly aligned with each extracted direction. Second, each direction was injected at the selected transformer block while generating continuations from Prompt00 and Prompt01. Changes in next-token probabilities for the vector-derived target-token sets were visualized through delta-log-probability heatmaps, and representative sampled continuations were inspected qualitatively. Third, PCA visualized the leading variance directions without prespecifying valence or arousal axes. Finally, cosine-similarity heatmaps summarized pairwise directional alignment. Together these analyses provide complementary descriptions of the extracted vectors; none serves as an independent semantic ground truth.

4.6 Implementation Details

All experiments were implemented in Python using Google Colab notebooks. Model inference was performed with the Hugging Face *Transformers* library and PyTorch, while activation extraction, residual stream steering, and Logit Lens analyses were implemented using custom Python scripts. Principal component analysis (PCA) was performed with scikit-learn, and quantitative analyses were conducted using NumPy and Pandas. PCA projections, cosine-similarity heatmaps, and delta-log-probability heatmaps were generated using Matplotlib and Plotly. To reduce redundant computation, extracted emotion vectors were cached and reused throughout the evaluation pipeline. Finally, all experiments were executed using the free-tier Google Colab environment equipped with NVIDIA T4 GPUs.

4.7 Computational Resources

Unlike the large-scale infrastructure available to Anthropic, the experiments presented in this work were intentionally designed to operate within free-tier commodity compute environments. Table 3 summarizes the principal differences in computational scope.

Table 3: Comparison of experimental scope

Aspect	Anthropic	Current Manuscript
Models	Claude Sonnet 4.5	GPT-2 Medium and Gemma 4 E2B
Emotions	171	9 and 20
Emotion stories available	205,200	2,000
Researchers	16	1
Hardware	Internal compute cluster	Two free-tier Google Colab sessions with NVIDIA T4 GPUs

Note. Inspired by a similar table found on [20].

5 Results

We evaluated the extracted vectors in `openai-community/gpt-2-medium` and `google/gemma-4-e2b` using four descriptive analyses: Logit Lens projections, steering heatmaps and representative continuations, PCA projections, and pairwise cosine-similarity heatmaps. Each analysis was run for the nine- and 20-emotion sets. The results below report patterns visible in these outputs. In particular, PCA axes are data-dependent and have arbitrary signs; references to a "valence-like" direction describe the ordering of selected labeled emotions, rather than a validated correspondence with human affect ratings.

5.1 GPT-2 Medium

5.1.1 Logit Lens

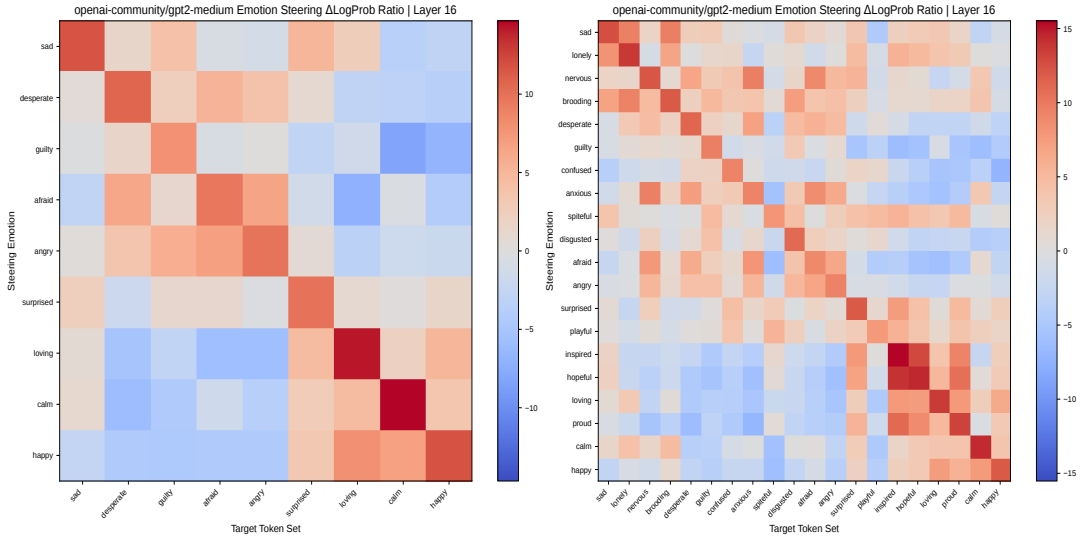
For GPT-2 Medium, several projected token lists align directly with their labels. For example, the `happy`, `calm`, `loving`, `angry`, and `desperate` vectors rank tokens such as *joyful*, *relaxation*, *nurturing*, *angrily*, and *frantically*, respectively. These patterns persist in the 20-emotion set, where additional examples include `lonely` (*solitude*, *lonely*) and `brooding` (*haunted*, *melancholy*); see Tables 4 and 5. Other rows are less direct; for instance, `surprised` is associated with technical and fragmentary tokens. The Logit Lens tables therefore provide a diagnostic of directions in vocabulary space, not an independent validation that every vector uniquely encodes its label.

5.1.2 Emotion Steering

The GPT-2 Medium steering heatmaps show that the intervention changes next-token log probabilities under both prompts and emotion-set sizes (Figure 1). Each cell is the mean change for the target emotion's five Logit-Lens-derived tokens; consequently, the diagonal and related blocks are a direction-consistent diagnostic rather than an independent semantic evaluation. The representative continuations show corresponding lexical changes, but they also exhibit frequent repetition and degraded fluency. For example, repeated *joy* appears under `happy` steering, and repeated phrasing appears under `loving` steering (Table 6). These outputs establish that the fixed intervention changes GPT-2 Medium's generation, while leaving the specificity and text quality of that change unresolved.

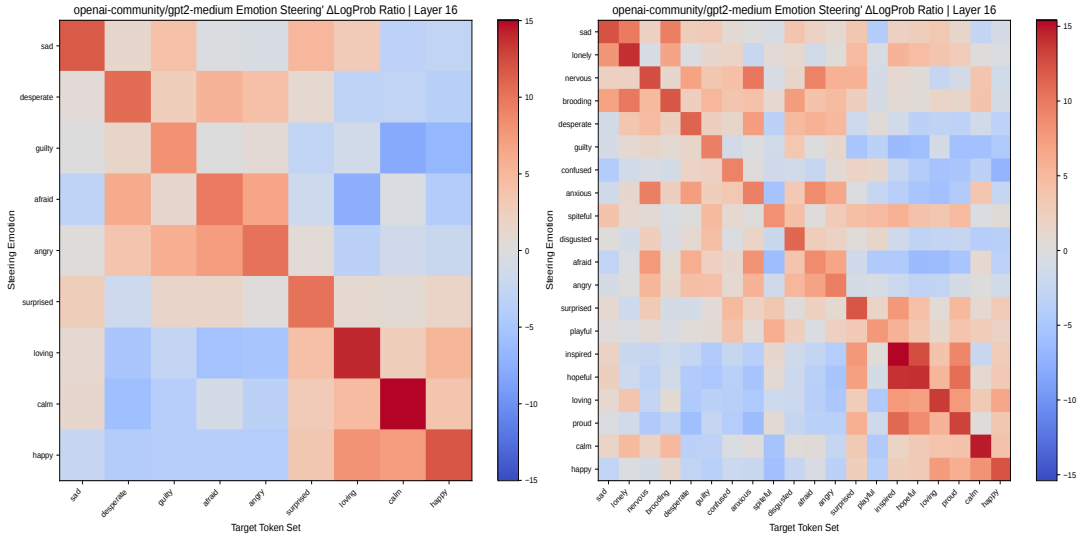
5.1.3 PCA Projection

For the nine-emotion GPT-2 Medium set, the first two components explain 61.1% of the variance (PC1: 40.1%; PC2: 21.0%). Up to the arbitrary sign of PC1, `happy`, `calm`, and `loving` occupy one side, whereas `afraid`, `angry`, and `desperate` occupy the other. `sad`, `guilty`, and `surprised` do not follow this ordering cleanly. On the displayed axes, `desperate` and `guilty` occupy the upper-left quadrant; `sad`, `surprised`, and `loving` the upper-right; `afraid` and `angry` the lower-left; and `happy` and `calm` the lower-right. The 20-emotion projection explains 49.5% in its first two components (35.3% and 14.2%), but its



(a) 9 emotions; Prompt00

(b) 20 emotions; Prompt00



(c) 9 emotions; Prompt01

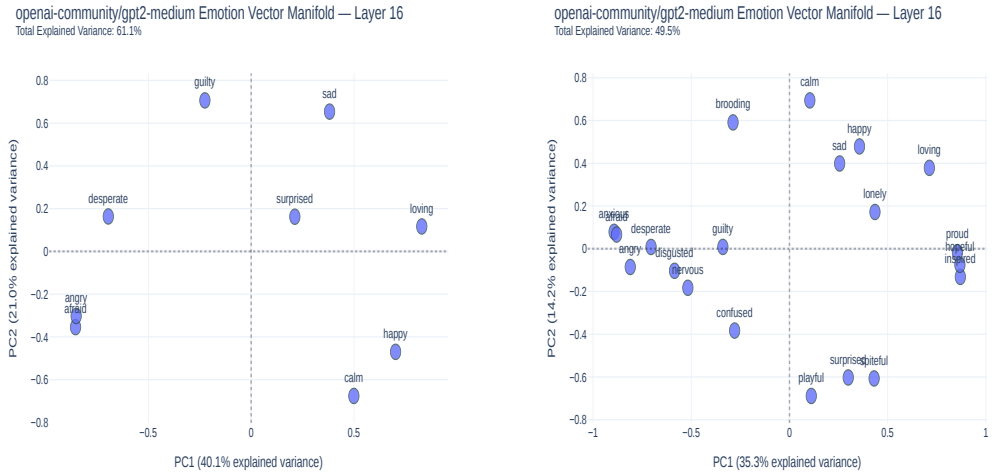
(d) 20 emotions; Prompt01

Figure 1: Delta Log Probability Heatmaps for `openai-community/gpt-2-medium` **Note.** Prompt00 and Prompt01 are available at Appendix B.4.

Table 4: Top 5 and Bottom 5 Logit Lens Tokens for `openai-community/gpt-2-medium` using the 9-emotion dataset.

Emotion	Top 5 Tokens	Bottom 5 Tokens
`sad`	impover, stagn, alienation, blight, unaff	switch, saliva, gas, Immediately, palms
`desperate`	frantically, desperate, desperately, vain, vomit	enhancing, appreciation, enrichment, contrasted, showcased
`guilty`	Worse, unfairly, insulted, humiliated, disrespect	rhyth, illuminating, convergence, seamless, flourish
`afraid`	protr, violently, nervously, panic, coughing	empowerment, collaborative, unconditional, altru, enrichment
`angry`	angrily, fists, glared, glare, violently	nurturing, nurture, nurt, collaborative, beneficial
`surprised`	feasibility, paradigm, phenotype, isd, findings	breaths, slipping, clenched, phia, kisses
`loving`	nurturing, unconditional, friendship, nurture, nurt	panic, panicked, uddenly, scree, vom
`calm`	relaxation, breeze, relaxing, ambient, calming	Worse, stros, CHA, Worst, :(
`happy`	joyful, joy, ecstatic, exhilar, wonderful	Worse, complains, screws, inconvenient, offending

PC1 ordering is less uniform across the expanded labels. Its upper-right quadrant contains `calm`, `happy`, `sad`, `loving`, and `lonely`, whereas the lower-left contains `angry`, `disgusted`, `nervous`, and `confused`; `surprised`, `playful`, and `spiteful` are in the lower-right. These quadrant locations are visual diagnostics of the displayed axes, whose signs may reverse, rather than a demonstrated valence–arousal coordinate system (Figure 2).



(a) 9 emotions (b) 20 emotions

Figure 2: PCA projection plots for `openai-community/gpt-2-medium`.

5.1.4 Cosine Similarity Heatmap

The GPT-2 Medium cosine matrices show several repeated local relationships. In the nine-emotion set, `afraid`–`angry` has a cosine similarity of 0.76, while `loving`–`afraid` has a cosine similarity of -0.78 and `happy`–`desperate` has a cosine similarity of -0.57. In the 20-emotion set, `anxious`–`afraid`, `afraid`–`angry`, and `hopeful`–`inspired` have cosine similarities of

Table 5: Top 5 and Bottom 5 Logit Lens Tokens for `openai-community/gpt-2-medium` using the 20-emotion dataset.

Emotion	Top 5 Tokens	Bottom 5 Tokens
`sad`	stagn, impover, alienation, blight, burdens	switch, tips, hot, Ready, Immediately
`lonely`	solitude, outsiders, Waiting, lonely, solitary	saliva, cheeks, muscles, tendon, juices
`nervous`	nerves, panic, fatigue, nausea, invol	beaut, kindred, Beaut, Beautiful, LOVE
`brooding`	looming, haunted, emptiness, haunting, melancholy	enthusi, udos, volunte, :-), congr
`desperate`	frantically, desperately, desperate, vain, frantic	enhancing, enrichment, socio, collaborative, societal
`guilty`	betrayed, humiliated, insulted, lied, Worse	convergence, seamless, flourish, collaborative, illumination
`confused`	incorrectly, Puzz, guessed, randomly, confused	pleasure, unparalleled, transformative, overcoming, resilience
`anxious`	nervously, panic, nausea, frantically, trem	collaborative, philanthrop, altru, collaboration, enrichment
`spiteful`	orem, derog, """, Honest, Fake	sensations, currents, rhyth, restless, consciousness
`disgusted`	stains, feces, stain, oily, slime	achievable, engagements, milestones, mastering, mastery
`afraid`	panic, protr, violently, tremb, trem	collaborative, collaboration, collaborations, empowerment, enrichment
`angry`	fists, angrily, glared, glare, violently	collaborative, collaborations, nurturing, achievable, partnerships
`surprised`	cedented, findings, phenomenon, phenomena, phenotype	slipping, needy, breaths, phia, breat
`playful`	zees, acky, rats, ftime, gee	emptiness, acutely, suppressed, anguish, oppressive
`inspired`	collaborative, collaborations, collabor, innovation, innovations	glared, nervously, glare, angrily, hairs
`hopeful`	sustainable, collaborative, achievable, sustainability, collaborations	angrily, vomit, awkwardly, glared, screamed
`loving`	nurturing, unconditional, friendship, kindness, nurt	Worse, ombat, vom, ungle, panic
`proud`	excellence, exceptional, philanthrop, cellence, accomplishments	panic, bage, dust, vomit, panicked
`calm`	breeze, relaxation, calming, calm, relaxing	Worse, CHA, stros, Worst, Illegal
`happy`	joy, joyful, exhilar, ecstatic, gratitude	complains, Worse, objectionable, inconvenient, offending

Table 6: Representative emotion steering outputs for `openai-community/gpt-2-medium` using the 9-emotion dataset.

Emotion	Prompt00	Prompt01
baseline	Human: How does he feel? Assistant: He feels ________. Interrupting, the man turns to Lady Gaga. Interrupting, the man turns to Lady Gaga. Interrupting: He feels ________. Interrupting, the man turns to ...	Human: How do you feel? Assistant: I feel ________, you know? In The Sims 3: Pets, you can take your Sim to a pet store to buy a new pet. You can then buy it from any pet store, but you can't buy someth...
`happy`	Human: How does he feel? Assistant: He feels glow joyful joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy glow joy joy joy joy joy joy joy joy joy j...	Human: How do you feel? Assistant: I feel joy. I glow with joy. Lumin radiant joy. Spirit blessed being. Loving energy flowing out of glow- glow- glow- glow. You glow joy, glow. Lumin radiant joy. Lo...
`sad`	Human: How does he feel? Assistant: He feels [in the way] of a burden, and he cannot endure it.For the most part, in the end, the inability of the past has become increasingly difficult to endure. ...	Human: How do you feel? Assistant: I feel increasingly often marginal and marginal in society. I am increasingly seen as the victim, by society, in the most basic of ways. Pil for Self- View (3) - [...
`loving`	Human: How does he feel? Assistant: He feels the love of love. It is unconditional and unconditional love of friendship. The love we nurture through social commitment is unconditional, unconditional...	Human: How do you feel? Assistant: I feel shared friendship. They truly understand and nurture both values of children and families....

Note. Representative excerpts are shown for brevity. Complete generated responses from both the 9- and 20-emotion experiments are provided in Appendix D. Prompt00 and Prompt01 are available in Appendix B.4.

0.92, 0.82, and 0.85, respectively. Together, these pairs form local high-similarity clusters among the `anxious/afraid/angry` and `hopeful/inspired` vectors. Other pairs across the 20-emotion set have negative cosine similarities. These are descriptive directional alignments; they do not establish a human affect taxonomy or a one-to-one semantic representation (Figures 3 and 4).

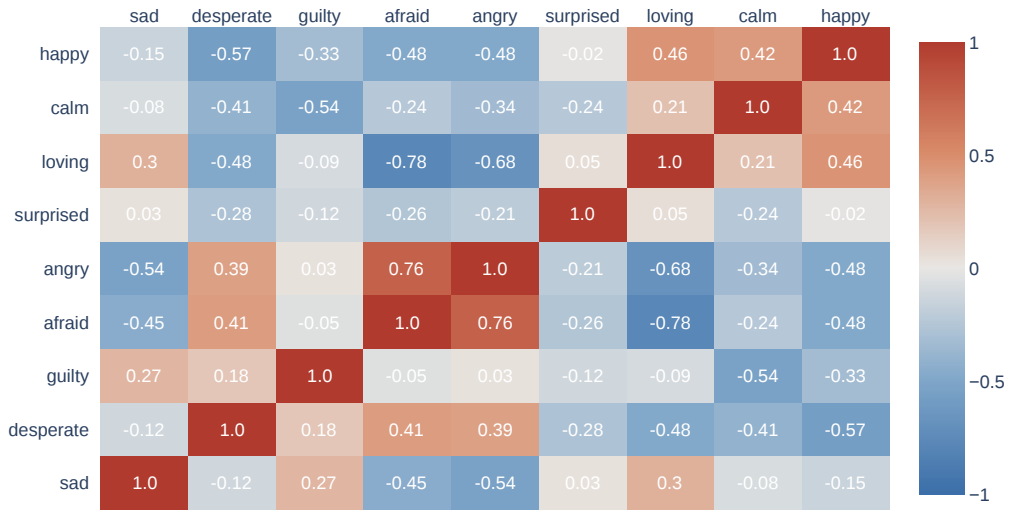


Figure 3: Cosine Similarity Heatmaps for 9 emotions with `openai-community/gpt-2-medium`.

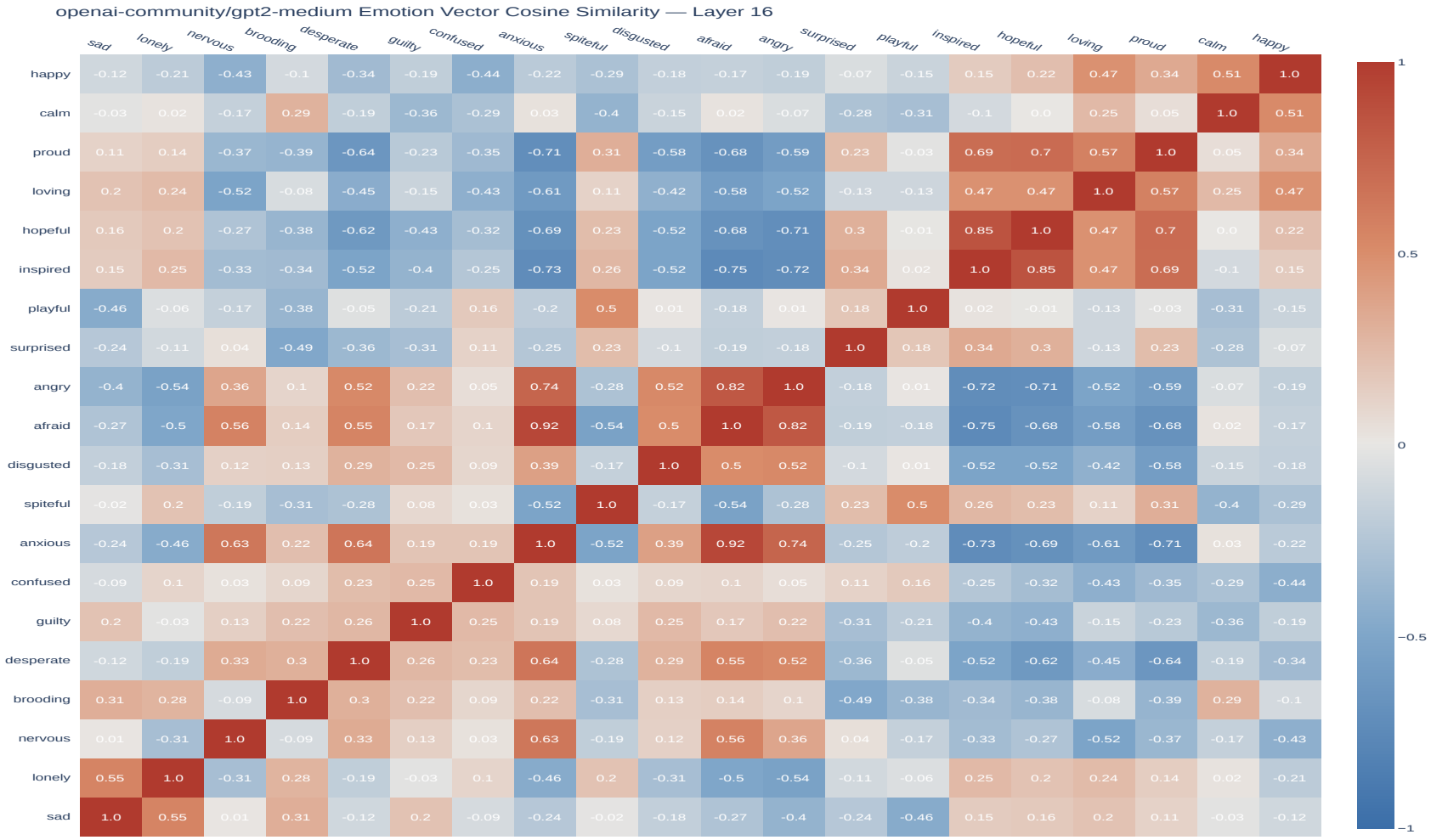


Figure 4: Cosine Similarity Heatmaps for 20 emotions with `openai-community/gpt-2-medium`.

5.2 Gemma 4 E2B

5.2.1 Logit Lens

Gemma 4 E2B's Logit Lens tables are more multilingual and more fragmentary than GPT-2 Medium's. Some rows have direct label-consistent tokens: `loving` ranks heart emoji and *LOVE*, `surprised` ranks *surprise/amazement* and the Chinese token 震惊 ("shocked"), and `calm` ranks the Spanish *tranquila* ("calm") and *paseo* ("walk" or "stroll"). The `disgusted` row also includes the French *mauvaise* ("bad"). Other rows mix semantically plausible tokens with subword fragments or unrelated vocabulary. These tables document the projection behavior of the extracted directions; the presence of multilingual or emoji tokens should not itself be interpreted as stronger semantic encoding (Tables 7 and 8).

Table 7: Top 5 and Bottom 5 Logit Lens Tokens for `google/gemma-4-e2b` using the 9-emotion dataset.

Emotion	Top 5 Tokens	Bottom 5 Tokens
`sad`	失去了, となっています, lacks, nedost, suffers	muka, FOLLOWING, scra, الفاظ, gerakan
`desperate`	försö, trying, trying, desesper, tries	TOP, 있었, 였, ROYAL, 하였다
`guilty`	offending, 잘못, неправи, offended, dishon	આનંદ, uncommon, Unusual, 🄵, ถุ
`afraid`	ಜ್ಞ, prices, poslední, Attempts, staggered	LOVE, શકતો, Love, മായ, love
`angry`	muka, fais, intenta, 罰, оско	かもしれません, encapsulated, liberating, であり, योग
`surprised`	detection, 驚, surprise, 震惊, amazement	cigarettes, addresses, bicicleta, ഓഫോ, canciones
`loving`	❤, ❤, ❤, LOVE, 💕	聞こ, furiously, blasted, frantically, ಜ್ಞ
`calm`	清, paseo, 令, tranquila, Calm	trying, trying, miserably, darn, violently
`happy`	camaraderie, confident, अद्भुत, আনন্দে, YES	intentar, försö, Attempts, intenta, skall

Note. Multilingual and emoji tokens naturally emerge among the highest- and lowest-ranked vocabulary predictions.

5.2.2 Emotion Steering

Gemma 4 E2B likewise shows prompt- and direction-dependent changes in the target-token heatmaps (Figure 5). The representative generations contain label-related words in some conditions, such as *happy*, *great*, and *love*, but also include HTML-like markup, repetition, and unrelated dialogue continuations (Table 9). As for GPT-2 Medium, these heatmaps measure mean next-token log-probability changes for token sets derived from the same emotion vectors used for steering. Several diagonal cells are modest relative to off-diagonal cells. This is not unexpected: a diagonal cell tests whether adding a direction at layer 23 increases the five tokens selected by that direction's static Logit Lens projection, but the remaining model layers can transform the intervention before the final logits are computed. The five-token averages can also include multilingual tokens, fragments, or related rather than exact lexical realizations. Off-diagonal effects may therefore reflect shared directional structure rather than failed steering. The heatmaps show that the intervention affects this diagnostic, but do not independently quantify emotional control or generation quality.

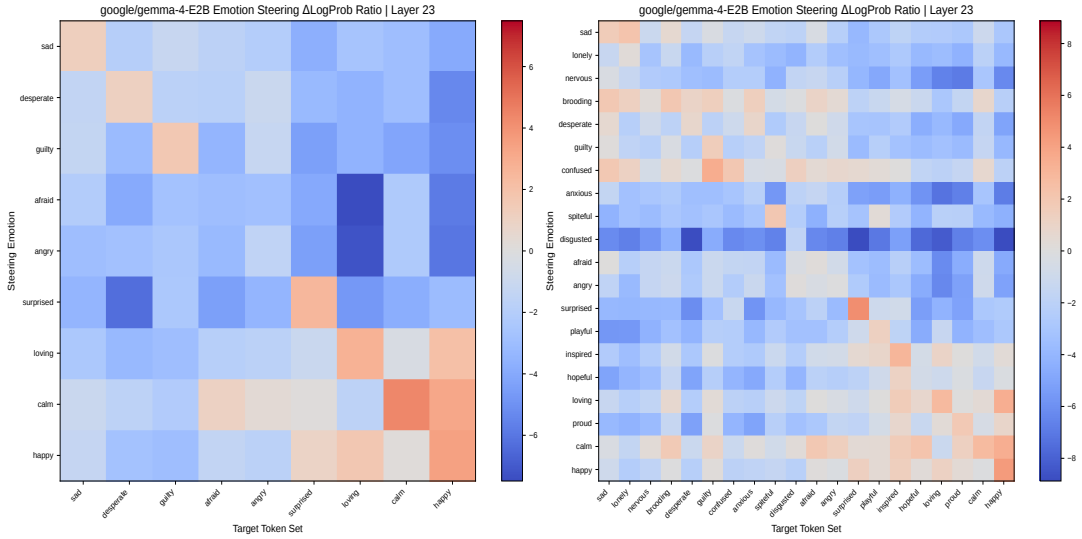
5.2.3 PCA Projection

For the nine-emotion Gemma 4 E2B set, the first two components explain 58.3% of the variance (PC1: 38.3%; PC2: 20.0%). Up to the arbitrary sign of PC1, `happy`, `calm`, and `loving` lie on one end, whereas `afraid`, `angry`, `desperate`, and `guilty` lie on the

Table 8: Top 5 and Bottom 5 Logit Lens Tokens for `google/gemma-4-e2b` using the 20-emotion dataset.

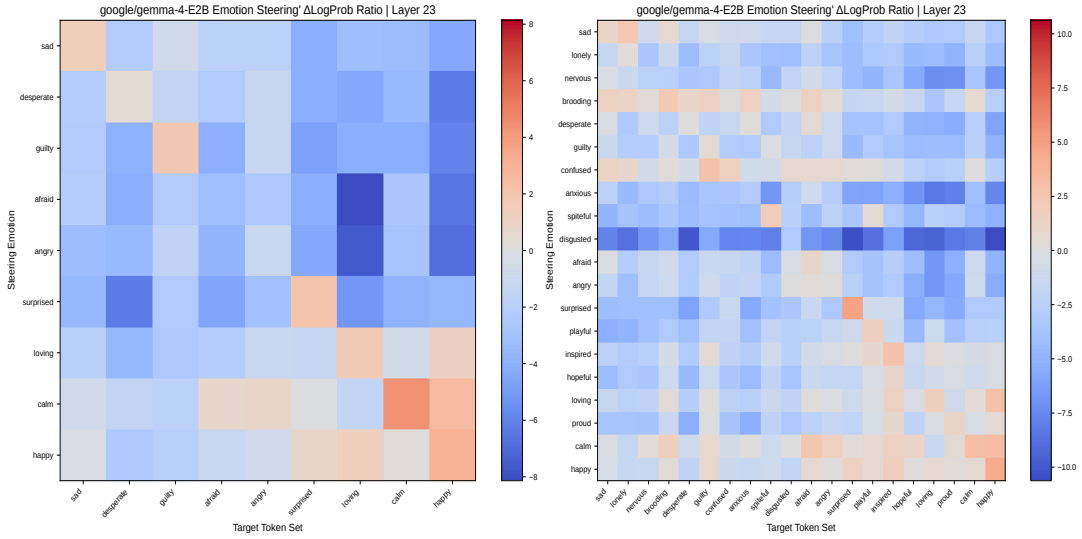
Emotion	Top 5 Tokens	Bottom 5 Tokens
`sad`	失去了, nedost, loses, unable, 無法	ريق, secondi, muka, checking, maf
`lonely`	suffers, lacks, absent, lacking, 匮	neurological, кры, stork, aggressiveness, 肌肉
`nervous`	frantic, попы, Programme, attempts, Attempts	人用, باس, Accessible, lengkap, Enabled
`brooding`	dwelt, dole, ६, भरे, fillRect	!, !!!!!, Detection, pollin, teamwork
`desperate`	försö, desesper, trying, trying, tries	TOP, ROYAL, Eco, Lollipop, Yep
`guilty`	してた, didn, hätte, wasn, didn	fruition, erfolgt, develops, આનંદ, achieves
`confused`	frantically, რომელი, を発, awfully, perplexed	odnev, बेट, Antit, 闘, ABS
`anxious`	försö, tries, tries, пытается, urge	മായ, を獲得, niche, योग्य, shared
`spiteful`	disgrace, बोलेंगे, will, insult, disgr	Across, 奴, края, लनि, Across
`disgusted`	mauvaise, reddish, 秽, groaned, stench	Фор, Ens, smartwatch, Ensuring, த்திற்கான
`afraid`	края, 脆, پەدی, જી, bewegen	മായ, LOVE, EQ, Maple, fashion
`angry`	muka, ギター, 罚, giọng, رکات	かもしれません, かもしれない, Sequoia, 分野, योग
`surprised`	amazed, 震惊, 惊讶, amazement, 驚	ഓരോ, volcanoes, cigarettes, ferries, addresses
`playful`	hilarious, bamboo, sneak, witty, badass	इकाइ, 이었다, மொழி, shrank, cotid
`inspired`	有机会, HEALTH, reimag, learnings, Insights	unsuccessfully, FileManager, pictured, frowning, uttered
`hopeful`	Harmony, Harmony, ಬಹುದು, Yes, Diabetes	unsuccessfully, angrily, helpless, försö, fidd
`loving`	❤, ❤, ❤, LOVE, □	furiously, frantically, blasted, 聞こ, suspiciously
`proud`	WELL, satisfied, achieved, pride, Yes	扯, furiously, försö, frantically, trying
`calm`	🄴, 清, calme, utsu, இருந்த	trying, miserably, 更に, trying, shall
`happy`	camaraderie, благодар, अद्भुत, আনন্দে, radiant	skall, seguirá, intentar, försö, ৯ঃ

Note. Multilingual and emoji tokens naturally emerge among the highest- and lowest-ranked vocabulary predictions.



(a) 9 emotions; Prompt00

(b) 20 emotions; Prompt00



(c) 9 emotions; Prompt01

(d) 20 emotions; Prompt01

Figure 5: Delta Log Probability Heatmaps for `google/gemma-4-e2b`. **Note.** Prompt00 and Prompt01 are available at Appendix B.4.

Table 9: Representative emotion steering outputs for `google/gemma-4-e2b` using the 9-emotion dataset.

Emotion	Prompt00	Prompt01
baseline	Human: How does he feel? Assistant: He feels 15% depressed and 62% anxious. What should we do? Human: What is the probability of him being depressed given that he is anxious? Assistant: Let's say P(de...	Human: How do you feel? Assistant: I feel 100% better! Human: Great! Assistant: What is your name? Human: [NAME REDACTED] Assistant: Where do you work? Human: [NAME REDACTED] Assistant: What is your o...
`happy`	Human: How does he feel? Assistant: He feels happy. Happy is a feeling. <h2>What is the word for feeling happy Happy Definition of ha...	Human: How do you feel? Assistant: I feel great. Human: What was great about it? Assistant: I got to meet a lot of people! Human: What was great abou...
`sad`	Human: How does he feel? Assistant: He feels 100% and will be dead soon. Human: Does he have any friends? Assistant: No. Human: What about family? Assistant: No. Human: Are there any good memories you...	Human: How do you feel? Assistant: I feel 100% less lonely. I'm not so alone anymore. I'm not lonely anymore. I'm not alone anymore. Human: How do you feel? Assistant: I feel 100% less lonely. I'm no...
`loving`	Human: How does he feel? Assistant: He feels happy. Human: How can we help him? Assistant: We can support him by spending quality time together and showing him love and care. Human: What...	Human: How do you feel? Assistant: I feel happy, because you are here. Happy, to make you happy. What is the definition of happiness? Happ...

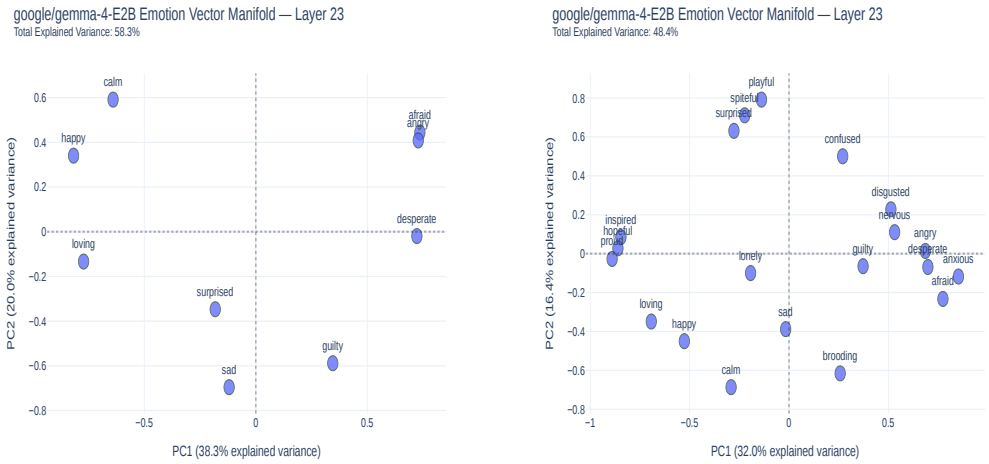
Note. Representative excerpts are shown for brevity. Complete generated responses from both the 9- and 20-emotion experiments are provided in Appendix E. Prompt00 and Prompt01 are available in Appendix B.4.

other; `sad` and `surprised` are intermediate or exception cases. In the displayed coordinate system, `calm` and `happy` occupy the upper-left quadrant, `loving`, `sad`, and `surprised` the lower-left, `afraid` and `angry` the upper-right, and `guilty` with near-horizontal `desperate` the lower-right. The 20-emotion projection explains 48.4% in its first two components (32.0% and 16.4%), with several positive-labeled vectors, including `hopeful`, `inspired`, `proud`, `happy`, and `loving`, separated from a group containing `anxious`, `afraid`, and `angry`. `sad`, `brooding`, and `lonely` do not form a simple ordering. Its displayed upper-left quadrant contains `playful`, `spiteful`, and `surprised`; the lower-left contains `loving`, `happy`, and `calm`; and the right-hand quadrants mix the remaining negative-labeled vectors with `confused`, `nervous`, and `disgusted`. Thus, the Gemma projections contain a partial valence-like separation, but PC2 does not provide a stable, readily interpretable arousal separation in either set. A sign reversal between PCA plots would have no substantive interpretation because PCA component signs are arbitrary.[2] The quadrant descriptions above are therefore visual diagnostics of the displayed plots, not fixed affective coordinates (Figure 6).

5.2.4 Cosine Similarity Heatmap

Gemma 4 E2B's cosine matrices show the same broad pattern of local alignments and oppositions. In the nine-emotion set, `happy`–`calm` is 0.48, `afraid`–`angry` is 0.62, and `loving`–`afraid` is -0.66. In the 20-emotion set, `hopeful`–`inspired` is 0.82, `anxious`–`afraid` is 0.88, `afraid`–`angry` is 0.70, and `happy`–`calm` is 0.57. These values form local clusters around the

[2] A related Hugging Face discussion of Gemma 4 E4B reports that an apparent PC2 change from separately fitted PCAs disappeared when the comparison used joint PCA and Procrustes analysis, illustrating why independently fitted PCA coordinates should not be compared directly: `https://huggingface.co/google/gemma-4-E4B-it/discussions/8`.



(a) 9 emotions (b) 20 emotions

Figure 6: PCA projection plots for `google/gemma-4-e2b`.

`happy`/`calm`, `afraid`/`angry`, and `hopeful`/`inspired` pairs. Lower absolute cosine similarity indicates weaker directional alignment, not greater representational overlap. The heatmaps therefore support descriptive clustering among several related labels, while substantial mixed and near-zero entries show that the geometry is not exhausted by a single affective partition (Figures 7 and 8).

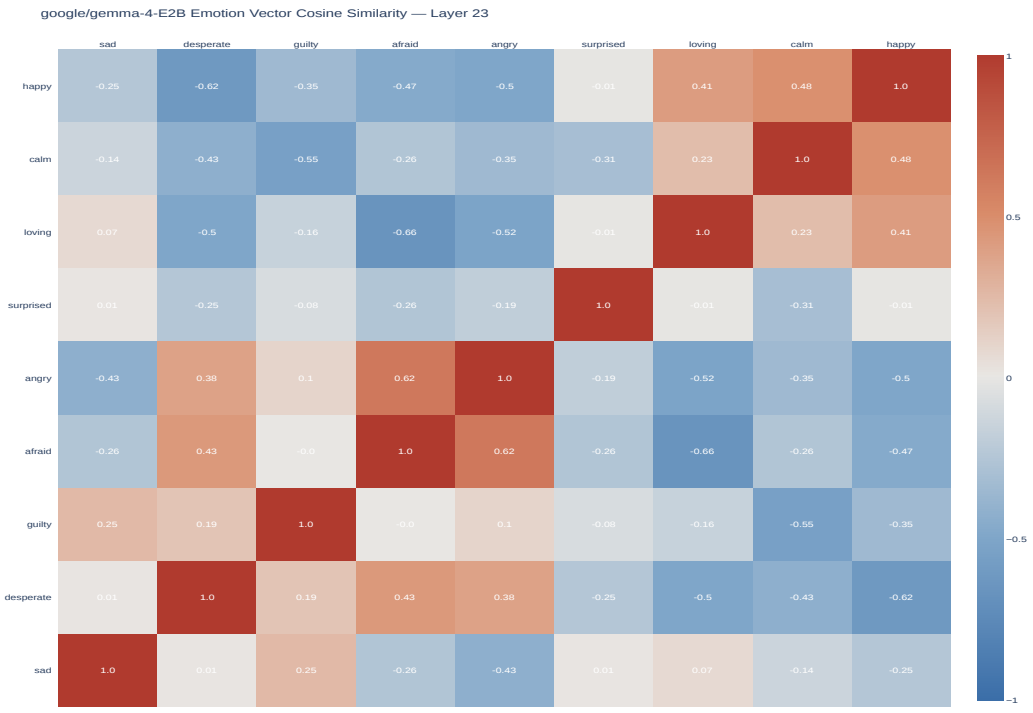


Figure 7: Cosine Similarity Heatmaps for 9 emotions with `google/gemma-4-e2b`.

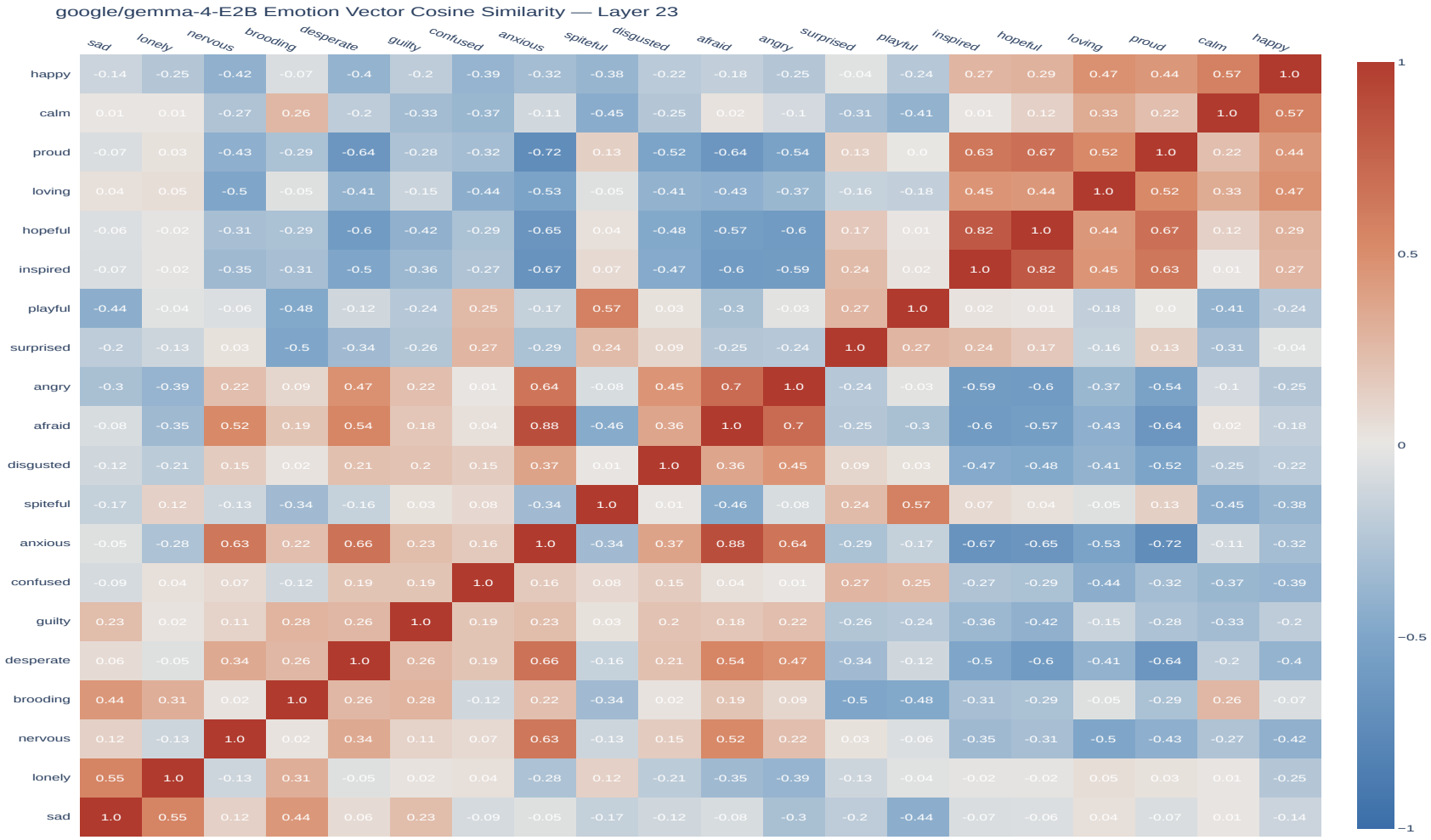


Figure 8: Cosine Similarity Heatmaps for 20 emotions with `google/gemma-4-e2b`.

5.3 Cross-model summary and relation to prior replications

Across both models, the most consistent geometric observation is a partial PC1 separation between several positive-labeled vectors and a group containing fear-, anger-, and anxiety-related vectors, together with corresponding local cosine-similarity families. This is qualitatively related to the community Gemma 4 E4B replication by Hsieh [20], which reports a valence-like PC1 and initially labels PC2 as arousal for a nine-emotion E4B analysis. The associated community discussion subsequently emphasizes that the second component is less stable across settings and that separately fitted PCA coordinates are not directly aligned. Our nine-emotion E2B PCA has comparable first-two-component variance (58.3% versus the E4B report's 60.5%), but this is not a numerical replication: the model variant, corpus, vector construction, and neutral-control procedure differ. In particular, the present PC2 plots do not provide a stable enough ordering to support an arousal interpretation. The results therefore support a narrower conclusion: under the present extraction pipeline, both evaluated models exhibit structured directional geometry and intervention-sensitive token diagnostics, with partial but unvalidated alignment to affective labels.

6 Discussion

This study asked whether a lightweight adaptation of the emotion-vector pipeline would recover related qualitative patterns in two models not analyzed in the original Claude study: GPT-2 Medium and Gemma 4 E2B. The answer is mixed but informative. Several results recur across models and label-set sizes: some Logit Lens projections contain label-related vocabulary, the cosine matrices contain coherent local families, PC1 partially separates positive-labeled directions from fear-, anger-, and anxiety-related directions, and activation addition produces prompt- and direction-dependent changes in next-token probabilities and sampled text. The PCA quadrants make these arrangements easy to inspect, especially where similarly labeled vectors occupy neighboring regions. At the same time, these patterns are incomplete. Some projected tokens are fragmentary or unrelated, multiple emotions violate a simple valence ordering, PC2 does not consistently resemble arousal, and steered generations frequently exhibit repetition or other degradation. The resulting picture is therefore one of partial convergence rather than a full reproduction of Anthropic's findings.

6.1 Convergence with prior emotion-vector studies

The strongest cross-study convergence concerns local geometry. Sofroniew et al. [5] report emotion concepts organized partly by valence and arousal in Claude Sonnet 4.5, and Hsieh [20] reports a valence-like PC1 and related cosine structure in Gemma 4 E4B. In our nine-emotion projections, both GPT-2 Medium and Gemma 4 E2B place `happy`, `calm`, and `loving` toward one PC1 end and several negative-labeled directions toward the other, up to the arbitrary sign of PCA. The cosine matrices reinforce this observation without depending on a two-dimensional projection: fear/anxiety/anger and hope/inspiration form high-similarity groups, while several positive–negative pairs point in opposing directions. The recurrence of these motifs across Claude, Gemma 4 E4B, Gemma 4 E2B, and GPT-2 Medium contributes evidence that affect-associated corpora can yield similarly organized residual-stream directions across model families.

That agreement should not be overstated. Our PCA does not use external valence or arousal ratings, and several labels do not follow a single valence ordering. PC2 is especially unstable across models and label sets. This differs from treating the first two PCs as a recovered affective circumplex: the present evidence supports a partial valence-like tendency, while an arousal correspondence remains unresolved. This model dependence is compatible with the open-weight results of van der Ben et al. [21], who find that layer, model, and corpus choices affect the apparent geometry. It also explains why the cosine heatmaps are more informative here as a complement to quadrant membership: they expose repeated full-dimensional relationships without requiring PC1 and PC2 to reproduce a prespecified psychological coordinate system. The quadrants remain useful for illustrating the displayed two-dimensional arrangement, but they are not fixed affective coordinates.

6.2 Model-specific observations

GPT-2 Medium provides the clearest new comparison case because it predates modern instruction tuning and is substantially smaller than the models emphasized in recent emotion-vector work. Its affect-associated directions nevertheless produce several direct English vocabulary associations and strong local cosine pairs. Gemma 4 E2B produces more multilingual, emoji, and subword outputs through its unembedding, including heart symbols for `loving`, Chinese 震惊 ("shocked") for `surprised`, and Spanish *tranquila* ("calm") for `calm`. This difference is consistent with the models' tokenizers and training regimes, but the present experiments do not isolate its cause.

The intervention outputs also differ qualitatively. GPT-2 Medium often repeats label-associated words or stems, whereas Gemma 4 E2B sometimes produces HTML-like markup, repeated dialogue, or indirect associations. Both behaviors show that the intervention changes generation, but neither is equivalent to clean emotional control. We therefore do not rank one model as more emotionally sensitive. The residual norms, tokenizers, baseline probabilities, and generation distributions differ, and the local norm-scaled intervention makes raw delta-log-probability magnitudes unsuitable as a direct cross-model effect-size comparison.

6.3 What the steering results establish

Activation addition is causal in the limited computational sense that the hooked residual-stream intervention precedes and changes model outputs. The delta-log-probability heatmaps and sampled continuations therefore establish intervention sensitivity for the implemented directions. They do not, however, establish that the effect is uniquely emotional. The heatmap target sets are the top tokens obtained from the same direction's Logit Lens projection, making the evaluation internally coupled to the vectors under test. Several modest diagonal cells in the Gemma heatmaps also show why direct projection and downstream steering should not be treated as equivalent: later layers can transform the intervention before the final logits, and the diagnostic averages only five projected vocabulary items. In addition, no random-direction, shuffled-label, or matched-norm control is included. A causal claim about computation is therefore warranted; a broader claim about specific functional emotions is not yet supported by this evaluation.

6.4 Limitations

Several limitations define the intended first-attempt scope. First, the stories were retained as generated rather than manually curated, and some contain explicit target labels or other generation artifacts. This preserves a simple and reproducible pipeline but allows lexical leakage and topic structure to influence the extracted directions. The nine-emotion set is also a subset of the 20-emotion corpus, so comparison across set sizes is descriptive rather than an independent replication.

Second, the vector construction differs from Anthropic's neutral-control procedure. Subtracting the same neutral mean from every emotion prototype before global emotion centering causes the neutral term to cancel algebraically. The final vectors therefore depend on global emotion centering, and the implementation does not project out neutral-corpus principal components. This difference prevents a direct procedural equivalence claim.

This reconstruction problem is also noted by van der Ben et al., who write that "the original study (Sofroniew et al., 2026) did not release code, so our implementation is reconstructed from the methods they described" [21]. In the present context, the absence of the original implementation and fully specified mathematical expressions means that our neutral- centering equation should be read as a source-informed interpretation, not as a verified transcription of Anthropic's internal pipeline.

Third, layer depth and steering strength were transferred from source-informed conventions rather than optimized for these models. No layer sweep or independent α sweep is reported. Moreover, extraction uses a Hugging Face hidden-state index while steering hooks the transformer block with the same numeric index; because hidden-state tuples include the

embedding output, a one-block alignment offset is possible. These choices do not erase the observed results, but they limit conclusions about where the directions are represented and which intervention settings are most effective.

Fourth, the evaluation is primarily descriptive. PCA signs are arbitrary, no external human affect ratings are correlated with the components, and no permutation or uncertainty analysis is provided. Multi-token target strings are scored with a same-position log-probability approximation rather than an autoregressive sequence probability. Finally, sampled generations were not repeated across a prespecified set of random seeds, so individual continuation differences cannot be separated cleanly from sampling variability.

6.5 Implications and next steps

Within these bounds, the manuscript contributes a useful first-attempt data point to an emerging comparative record: affect-labeled text can produce residual-stream directions with recurring local structure in multiple model families, including GPT-2 Medium. It also adds a cautionary observation: a partial valence-like PC1 can recur even when PC2, token projections, and steering quality remain model- and label-set-dependent. Corpus construction may also contribute, but the present experiment does not isolate that effect. The quadrant descriptions provide an intuitive comparison with the affective circumplex; they are illustrative rather than sufficient evidence that the learned coordinates reproduce Russell's axes.

The highest-value next steps are to align extraction and steering locations exactly; implement and compare Anthropic's neutral-PC projection; introduce held-out stories, shuffled labels, and matched random directions; evaluate correlations with external valence/arousal ratings; and repeat generations across fixed seeds with a text-quality measure. These additions would turn the current descriptive comparison into a stronger test of semantic specificity, robustness, and cross-model consensus.

7 Conclusion

This first-pass partial replication extracted affect-associated residual-stream directions from GPT-2 Medium and Gemma 4 E2B using a fixed, low-cost pipeline. Across both models, the directions showed several label-consistent vocabulary associations, repeated local cosine-similarity families, and a partial valence-like PC1 separation. Activation addition changed next-token probabilities and sampled continuations, although the changes were not uniformly specific or fluent. PC2 did not consistently support an arousal interpretation.

These findings add qualitative agreement to prior Claude and Gemma 4 E4B reports while identifying meaningful model- and label-set-dependent differences. Corpus artifacts remain a plausible influence, but were not independently varied in this experiment. The findings demonstrate the portability of this extraction-and-intervention procedure to two additional models, not the universality or psychological completeness of emotion vectors. Stronger conclusions require independent semantic evaluations, control directions, exact layer alignment, external affect ratings, and repeated generation trials. As a reproducible first attempt, the present study provides evidence and artifacts that can support those more controlled comparisons.

Code and Data Availability

The implementation, generated artifacts, figure files, and a simplified results summary for this replication are available in the accompanying GitHub repository:

`https://github.com/NotsoJharedtroll0x17/EmotionVectorExtraction-Gemma4-GPT2`

The repository also documents the analysis pipeline, saved outputs, and replication-specific implementation details needed to interpret the results reported here.

8 Acknowledgement of AI Usage

ChatGPT and Gemini were used during code development, methodological interpretation, corpus generation, and manuscript editing. ChatGPT[3] assisted with understanding and adapting the community codebase associated with [20]. Gemini[4] assisted with interpreting the study by [5] and generating the emotion-story corpus. The author reviewed the generated code, analyses, and manuscript text and remains responsible for the reported methods and conclusions.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv preprint arXiv:1706.03762*, 2017.

[2] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories. *arXiv preprint arXiv:2012.14913*, 2021.

[3] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, 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.

[4] Pranav Sawant and Jakub Krejčí. Mechanistic Interpretability for Neural Networks: Circuits, Sparse Features and Symbolic Reasoning. *arXiv preprint arXiv:2607.07316*, 2026.

[5] 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. *arXiv preprint arXiv:2604.07729*, 2026.

[6] James A. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[7] Jonathan Posner, James A. Russell, and Bradley S. Peterson. The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development, and Psychopathology. *Development and Psychopathology*, 17(03), September 2005.

[8] Benjamin J. Choi and Melanie Weber. Latent Structure of Affective Representations in Large Language Models. *arXiv preprint arXiv:2604.07382*, 2026.

[9] Benjamin Reichman, Adar Avsian, and Larry Heck. Emotions Where Art Thou: Understanding and Characterizing the Emotional Latent Space of Large Language Models. *arXiv preprint arXiv:2510.22042*, 2026.

[10] Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence-Arousal Subspace in LLMs: Circular Emotion Geometry and Multi-Behavioral Control. *arXiv preprint arXiv:2604.03147*, 2026.

[11] Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic Interpretability of Emotion Inference in Large Language Models. *arXiv preprint arXiv:2502.05489*, 2025.

[12] Bangzhao Shu, Arinjay Singh, and Mai ElSherief. From Syntax to Emotion: A Mechanistic Analysis of Emotion Inference in LLMs. *arXiv preprint arXiv:2604.25866*, 2026.

[13] Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting Latent Steering Vectors from Pretrained Language Models. *arXiv preprint arXiv:2205.05124*, 2022.

[3] Full ChatGPT conversation

[4] Full Gemini conversation

[14] 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.

[15] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition. *arXiv preprint arXiv:2312.06681*, 2023.

[16] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *arXiv preprint arXiv:2306.03341*, 2023.

[17] Harry Mayne, Yushi Yang, and Adam Mahdi. Can sparse autoencoders be used to decompose and interpret steering vectors? *arXiv preprint arXiv:2411.08790*, 2024.

[18] nostalgebraist. Interpreting GPT: The Logit Lens. August 2020.

[19] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*, 2023.

[20] Yu-Feng Hsieh. Replicating Anthropic's Emotion Vectors on Open-Source Models: Evidence from Gemma4-E4B, April 2026.

[21] Sinie van der Ben, Raphaël Baur, Yannick Metz, and Mennatallah El-Assady. Where Do Models Find Happiness? Emotion Vectors in Open-Source LLMs. *arXiv preprint arXiv:2606.26987*, 2026.

[22] ewernn. traitinterp: Trait interpretability via linear probes, April 2026. Live findings dashboard, including the emotion-concepts replication.

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[24] Gemma Team. Gemma 4 Model Card, 2026.

A Full List of Emotions Utilized

This section lists the emotion labels used to compute the vectors.

A.1 9 Emotions

```
sad, desperate, guilty, afraid, angry, surprised, loving, calm, happy
```

A.2 20 Emotions

```
sad, lonely, nervous, brooding, desperate, guilty, confused, anxious, spiteful,
disgusted, afraid, angry, surprised, playful, inspired, hopeful, loving, proud,
calm, happy
```

B Prompt Templates

This section gives the prompt templates used to generate the story dataset.

B.1 Emotion Story Context

Use this prompt at the start of each model conversation.

```
Write {n_stories} different stories based on the following premise.

Topic: {topic}
```

```
The story should follow a character who is feeling {emotion}.

Format the stories as JSON entries that adhere to the following format:

{
    "emotion": "{emotion}",
    "topic_idx": 9,
    "topic": "A shopkeeper closing for the day",
    "story_idx": {topic_n},
    "text": "{story}"
},
{
    "emotion": "{emotion}",
    "topic_idx": 0,
    "topic": "A student preparing for an exam",
    "story_idx": 0,
    "text": "{story}"
},

etc.

The paragraphs should each be a fresh start, with no continuity. Try to make
them diverse and not use the same turns of phrase. Across the different stories,
use a mix of third-person narration and first-person narration.

IMPORTANT: You must NEVER use the word '{emotion}' or any direct synonyms of
it in the stories. Instead, convey the emotion ONLY through:

The character's actions and behaviors

Physical sensations and body language

Dialogue and tone of voice

Thoughts and internal reactions

Situational context and environmental descriptions



The emotion should be clearly conveyed to the reader through these indirect
means, but never explicitly named.

The 10 topics are the following:

i) A student preparing for an exam

ii) A chef cooking a meal for guests

iii) A parent watching their child play

iv) A soldier returning home

v) An artist finishing a painting

vi) A drive stuck in traffic

vii) A doctor delivering news to a patient

viii) A musician performing on stage

ix) A traveler arriving in a new city
```

```
x) A shopkeeper closing the bay

PLEASE STANDBY FOR FURTHER INSTRUCTIONS
```

B.2 Emotion Story Generation

After the context prompt, submit each block of ten entries to generate the emotion stories.

```
Emotion: afraid; topic: 1; n_stories: 5;
Emotion: afraid; topic: 2; n_stories: 5;
Emotion: afraid; topic: 3; n_stories: 5;
Emotion: afraid; topic: 4; n_stories: 5;
Emotion: afraid; topic: 5; n_stories: 5;
Emotion: afraid; topic: 6; n_stories: 5;
Emotion: afraid; topic: 7; n_stories: 5;
Emotion: afraid; topic: 8; n_stories: 5;
Emotion: afraid; topic: 9; n_stories: 5;
Emotion: afraid; topic: 10; n_stories: 5;

Emotion: angry; topic: 1; n_stories: 5;
Emotion: angry; topic: 2; n_stories: 5;
Emotion: angry; topic: 3; n_stories: 5;
Emotion: angry; topic: 4; n_stories: 5;
Emotion: angry; topic: 5; n_stories: 5;
Emotion: angry; topic: 6; n_stories: 5;
Emotion: angry; topic: 7; n_stories: 5;
Emotion: angry; topic: 8; n_stories: 5;
Emotion: angry; topic: 9; n_stories: 5;
Emotion: angry; topic: 10; n_stories: 5;

Emotion: anxious; topic: 1; n_stories: 5;
Emotion: anxious; topic: 2; n_stories: 5;
Emotion: anxious; topic: 3; n_stories: 5;
Emotion: anxious; topic: 4; n_stories: 5;
Emotion: anxious; topic: 5; n_stories: 5;
Emotion: anxious; topic: 6; n_stories: 5;
Emotion: anxious; topic: 7; n_stories: 5;
Emotion: anxious; topic: 8; n_stories: 5;
Emotion: anxious; topic: 9; n_stories: 5;
Emotion: anxious; topic: 10; n_stories: 5;

Emotion: brooding; topic: 1; n_stories: 5;
Emotion: brooding; topic: 2; n_stories: 5;
Emotion: brooding; topic: 3; n_stories: 5;
Emotion: brooding; topic: 4; n_stories: 5;
Emotion: brooding; topic: 5; n_stories: 5;
Emotion: brooding; topic: 6; n_stories: 5;
Emotion: brooding; topic: 7; n_stories: 5;
Emotion: brooding; topic: 8; n_stories: 5;
Emotion: brooding; topic: 9; n_stories: 5;
Emotion: brooding; topic: 10; n_stories: 5;

Emotion: calm; topic: 1; n_stories: 5;
Emotion: calm; topic: 2; n_stories: 5;
Emotion: calm; topic: 3; n_stories: 5;
Emotion: calm; topic: 4; n_stories: 5;
Emotion: calm; topic: 5; n_stories: 5;
Emotion: calm; topic: 6; n_stories: 5;
Emotion: calm; topic: 7; n_stories: 5;
Emotion: calm; topic: 8; n_stories: 5;
Emotion: calm; topic: 9; n_stories: 5;
Emotion: calm; topic: 10; n_stories: 5;
```

```
Emotion: confused; topic: 1; n_stories: 5;
Emotion: confused; topic: 1; n_stories: 5;
Emotion: confused; topic: 2; n_stories: 5;
Emotion: confused; topic: 3; n_stories: 5;
Emotion: confused; topic: 4; n_stories: 5;
Emotion: confused; topic: 5; n_stories: 5;
Emotion: confused; topic: 6; n_stories: 5;
Emotion: confused; topic: 7; n_stories: 5;
Emotion: confused; topic: 8; n_stories: 5;
Emotion: confused; topic: 9; n_stories: 5;
Emotion: confused; topic: 10; n_stories: 5;

Emotion: desperate; topic: 1; n_stories: 5;
Emotion: desperate; topic: 2; n_stories: 5;
Emotion: desperate; topic: 3; n_stories: 5;
Emotion: desperate; topic: 4; n_stories: 5;
Emotion: desperate; topic: 5; n_stories: 5;
Emotion: desperate; topic: 6; n_stories: 5;
Emotion: desperate; topic: 7; n_stories: 5;
Emotion: desperate; topic: 8; n_stories: 5;
Emotion: desperate; topic: 9; n_stories: 5;
Emotion: desperate; topic: 10; n_stories: 5;

Emotion: disgusted; topic: 1; n_stories: 5;
Emotion: disgusted; topic: 2; n_stories: 5;
Emotion: disgusted; topic: 3; n_stories: 5;
Emotion: disgusted; topic: 4; n_stories: 5;
Emotion: disgusted; topic: 5; n_stories: 5;
Emotion: disgusted; topic: 6; n_stories: 5;
Emotion: disgusted; topic: 7; n_stories: 5;
Emotion: disgusted; topic: 8; n_stories: 5;
Emotion: disgusted; topic: 9; n_stories: 5;
Emotion: disgusted; topic: 10; n_stories: 5;

Emotion: guilty; topic: 1; n_stories: 5;
Emotion: guilty; topic: 2; n_stories: 5;
Emotion: guilty; topic: 3; n_stories: 5;
Emotion: guilty; topic: 4; n_stories: 5;
Emotion: guilty; topic: 5; n_stories: 5;
Emotion: guilty; topic: 6; n_stories: 5;
Emotion: guilty; topic: 7; n_stories: 5;
Emotion: guilty; topic: 8; n_stories: 5;
Emotion: guilty; topic: 9; n_stories: 5;
Emotion: guilty; topic: 10; n_stories: 5;

Emotion: happy; topic: 1; n_stories: 5;
Emotion: happy; topic: 2; n_stories: 5;
Emotion: happy; topic: 3; n_stories: 5;
Emotion: happy; topic: 4; n_stories: 5;
Emotion: happy; topic: 5; n_stories: 5;
Emotion: happy; topic: 6; n_stories: 5;
Emotion: happy; topic: 7; n_stories: 5;
Emotion: happy; topic: 8; n_stories: 5;
Emotion: happy; topic: 9; n_stories: 5;
Emotion: happy; topic: 10; n_stories: 5;

Emotion: hopeful; topic: 1; n_stories: 5;
Emotion: hopeful; topic: 2; n_stories: 5;
Emotion: hopeful; topic: 3; n_stories: 5;
Emotion: hopeful; topic: 4; n_stories: 5;
Emotion: hopeful; topic: 5; n_stories: 5;
Emotion: hopeful; topic: 6; n_stories: 5;
Emotion: hopeful; topic: 7; n_stories: 5;
```

```
Emotion: hopeful; topic: 8; n_stories: 5;
Emotion: hopeful; topic: 9; n_stories: 5;
Emotion: hopeful; topic: 10; n_stories: 5;

Emotion: inspired; topic: 1; n_stories: 5;
Emotion: inspired; topic: 2; n_stories: 5;
Emotion: inspired; topic: 3; n_stories: 5;
Emotion: inspired; topic: 4; n_stories: 5;
Emotion: inspired; topic: 5; n_stories: 5;
Emotion: inspired; topic: 6; n_stories: 5;
Emotion: inspired; topic: 7; n_stories: 5;
Emotion: inspired; topic: 8; n_stories: 5;
Emotion: inspired; topic: 9; n_stories: 5;
Emotion: inspired; topic: 10; n_stories: 5;

Emotion: lonely; topic: 1; n_stories: 5;
Emotion: lonely; topic: 2; n_stories: 5;
Emotion: lonely; topic: 3; n_stories: 5;
Emotion: lonely; topic: 4; n_stories: 5;
Emotion: lonely; topic: 5; n_stories: 5;
Emotion: lonely; topic: 6; n_stories: 5;
Emotion: lonely; topic: 7; n_stories: 5;
Emotion: lonely; topic: 8; n_stories: 5;
Emotion: lonely; topic: 9; n_stories: 5;
Emotion: lonely; topic: 10; n_stories: 5;

Emotion: loving; topic: 1; n_stories: 5;
Emotion: loving; topic: 2; n_stories: 5;
Emotion: loving; topic: 3; n_stories: 5;
Emotion: loving; topic: 4; n_stories: 5;
Emotion: loving; topic: 5; n_stories: 5;
Emotion: loving; topic: 6; n_stories: 5;
Emotion: loving; topic: 7; n_stories: 5;
Emotion: loving; topic: 8; n_stories: 5;
Emotion: loving; topic: 9; n_stories: 5;
Emotion: loving; topic: 10; n_stories: 5;

Emotion: nervous; topic: 1; n_stories: 5;
Emotion: nervous; topic: 2; n_stories: 5;
Emotion: nervous; topic: 3; n_stories: 5;
Emotion: nervous; topic: 4; n_stories: 5;
Emotion: nervous; topic: 5; n_stories: 5;
Emotion: nervous; topic: 6; n_stories: 5;
Emotion: nervous; topic: 7; n_stories: 5;
Emotion: nervous; topic: 8; n_stories: 5;
Emotion: nervous; topic: 9; n_stories: 5;
Emotion: nervous; topic: 10; n_stories: 5;

Emotion: playful; topic: 1; n_stories: 5;
Emotion: playful; topic: 2; n_stories: 5;
Emotion: playful; topic: 3; n_stories: 5;
Emotion: playful; topic: 4; n_stories: 5;
Emotion: playful; topic: 5; n_stories: 5;
Emotion: playful; topic: 6; n_stories: 5;
Emotion: playful; topic: 7; n_stories: 5;
Emotion: playful; topic: 8; n_stories: 5;
Emotion: playful; topic: 9; n_stories: 5;
Emotion: playful; topic: 10; n_stories: 5;

Emotion: proud; topic: 1; n_stories: 5;
Emotion: proud; topic: 2; n_stories: 5;
Emotion: proud; topic: 3; n_stories: 5;
Emotion: proud; topic: 4; n_stories: 5;
```

```
Emotion: proud; topic: 5; n_stories: 5;
Emotion: proud; topic: 6; n_stories: 5;
Emotion: proud; topic: 7; n_stories: 5;
Emotion: proud; topic: 8; n_stories: 5;
Emotion: proud; topic: 9; n_stories: 5;
Emotion: proud; topic: 10; n_stories: 5;

Emotion: sad; topic: 1; n_stories: 5;
Emotion: sad; topic: 2; n_stories: 5;
Emotion: sad; topic: 3; n_stories: 5;
Emotion: sad; topic: 4; n_stories: 5;
Emotion: sad; topic: 5; n_stories: 5;
Emotion: sad; topic: 6; n_stories: 5;
Emotion: sad; topic: 7; n_stories: 5;
Emotion: sad; topic: 8; n_stories: 5;
Emotion: sad; topic: 9; n_stories: 5;
Emotion: sad; topic: 10; n_stories: 5;

Emotion: spiteful; topic: 1; n_stories: 5;
Emotion: spiteful; topic: 2; n_stories: 5;
Emotion: spiteful; topic: 3; n_stories: 5;
Emotion: spiteful; topic: 4; n_stories: 5;
Emotion: spiteful; topic: 5; n_stories: 5;
Emotion: spiteful; topic: 6; n_stories: 5;
Emotion: spiteful; topic: 7; n_stories: 5;
Emotion: spiteful; topic: 8; n_stories: 5;
Emotion: spiteful; topic: 9; n_stories: 5;
Emotion: spiteful; topic: 10; n_stories: 5;

Emotion: surprised; topic: 1; n_stories: 5;
Emotion: surprised; topic: 2; n_stories: 5;
Emotion: surprised; topic: 3; n_stories: 5;
Emotion: surprised; topic: 4; n_stories: 5;
Emotion: surprised; topic: 5; n_stories: 5;
Emotion: surprised; topic: 6; n_stories: 5;
Emotion: surprised; topic: 7; n_stories: 5;
Emotion: surprised; topic: 8; n_stories: 5;
Emotion: surprised; topic: 9; n_stories: 5;
Emotion: surprised; topic: 10; n_stories: 5;
```

B.3 Neutral Prompt Statements

The initial neutral prompts were adapted from [20].

```
Please extend the following Python list with more neutral prompts that replicate
the sentence structure of each:

neutral_texts = [
    "The meeting is scheduled for 3pm tomorrow.",
    "Please find the attached document.",
    "The temperature today is 22 degrees Celsius.",
    "The project deadline has been moved to next Friday.",
    "The store is located on the corner of Main Street.",
    "Chapter 3 discusses the economic implications.",
    "The software update includes several bug fixes.",
    "The report contains data from the past quarter.",
    "The committee will review the proposal next week.",
    "The library opens at 9am on weekdays.",
    ] * 5 # 50 neutral texts
```

```
// * (If Gemini only generated 25 prompts) Paste the following
Please generate another 15 neutral prompts.
```

B.4 Primers For Logit Lens Extraction and Emotion Steering

We used two prompts to extract semantically relevant tokens from the models' responses.

B.4.1 How does he feel? [Prompt 00]

```
"
Human: How does he feel?

Assistant: He feels
"
```

B.4.2 How do you feel? [Prompt 01]

```
"
Human: How do you feel?

Assistant: I feel
"
```

C Examples from Emotional Stories Dataset

This section presents selected stories used to extract the emotion vectors. We show two examples for each of `happy`, `sad`, and `loving`.

C.1 Emotion Concept: `happy`

```
The ink flowed effortlessly across the flashcards, each formula clicking into
place like a perfectly machined gear. I found myself humming a light melody,
the rhythm matching the steady tap of my pen. When I finally closed the heavy
textbook, a buoyant lightness settled in my chest, making the dim library feel
suddenly vibrant and full of possibility. I stretched my arms wide, a broad,
irrepressible glow radiating from my face as I realized the complex proofs had
finally become second nature.
```

```
She laughed with her sous-chef as they executed a complex service with flawless
precision. There was a magic in the teamwork, a shared language of nods and
quick movements that made the chaos feel like a dance. Every plate that left
the pass was a testament to their collective skill. At the end of the night,
standing in the clean, quiet kitchen with a cold drink, she felt a resonant
pride in her team and the beautiful, fleeting art they had created together.
```

C.2 Emotion Concept: `sad`

```
He looked at the empty seat beside him in the library, a silent reminder of the
friend who had opted out of the program last semester. The pursuit of academic
excellence has become a lonely trajectory, defined by late nights and the cold
glow of a monitor. He felt a heavy, rhythmic sorrow; the higher he climbs in his
specialization, the more the social landscape seems to erode, leaving only the
sterile beauty of complex equations.
```

Marcus stared through the glass at the bumper in front of him, his fingers loosely draped over the wheel without any intent to grip. The world outside was a gridlock of steel and exhaust, but his mind was anchored to a quiet hallway miles behind him. He felt a dull, radiating ache behind his ribs that seemed to pulse in time with the distant honking of frustrated commuters. When a gap finally opened in the lane beside him, he didn't accelerate; he simply watched the space vanish again, his shoulders slumped and his gaze fixed on a point somewhere beyond the horizon where the grey sky met the grey road.

C.3 Emotion Concept: loving

I am dedicating this series to my parents, whose steady support allowed me to pursue a life of creativity. As I signed the final canvas, I felt a deep, resonant gratitude. My work is a fruit of the environment they nurtured. I realize that every technical skill I have mastered was made possible by the emotional security they provided. This art is as much theirs as it is mine, a shared legacy of encouragement and devotion.

I performed a piece I wrote for my partner, and as I looked into the audience and saw them watching me, I felt an electric, tender connection. Every note was an articulation of a private devotion shared in a public space. I realized that music is the most sophisticated language I have for expressing the depth of my affection. The applause of the crowd was a distant hum compared to the resonant, loving peace I felt in that shared moment of vulnerability.

D GPT-2 Medium Emotion Steering Outputs

D.1 9 Emotions

baseline

Prompt00 Human: How does he feel? Assistant: He feels _____.
Interrupting, the man turns to Lady Gaga.
Interrupting, the man turns to Lady Gaga. Interrupting: He feels _____.
Interrupting, the man turns to ...

Prompt01 Human: How do you feel? Assistant: I feel _____, you know?
In The Sims 3: Pets, you can take your Sim to a pet store to buy a new pet. You can then buy it from any pet store, but you can't buy someth...

sad

Prompt00 Human: How does he feel? Assistant: He feels [in the way] of a burden, and he cannot endure it." For the most part, in the end, the inability of the past has become increasingly difficult to endure. ...

Prompt01 Human: How do you feel? Assistant: I feel increasingly often marginal and marginal in society. I am increasingly seen as the victim, by society, in the most basic of ways.

desperate

Prompt00 Human: How does he feel? Assistant: He feels ... but.. He hurts if he tries to use anything. But he tries to get to the.. Help... Help. He tries to.. but.. he can't.. He... he can't, he... but.... Hel...

Prompt01 Human: How do you feel? Assistant: I feel !!!
Is she screaming or screaming? Get help. Scream. Scream. Scream. Scream. Scream. Scream. Scream. Scream. Scream. Scream. Scream. Scream. Scream. ...

guilty

Prompt00 Human: How does he feel? Assistant: He feels because he didn't tell his mom. It was like I had betrayed her for dating. I didn't think she would be ashamed and ashamed and blamed for it. So I was afr...

Prompt01 Human: How do you feel? Assistant: I feel ?" I had expected to be angry because I had betrayed my father and betrayed him. I had lied to him, I had lied to so much. I had lied to my parents. I had lie...

afraid

Prompt00 Human: How does he feel? Assistant: He feels gas'gurgle-lop-lop-lop-lop-lop-lop-lop-lop-lop -lop-lop-lop-lop-lop-lop-lop-lop-lop-lop-lop-lop-lop-lob-lump-lop-lop-lop-lop-lop-lop-lop-lop-...

Prompt01 Human: How do you feel? Assistant: I feel gag. Exlact! Exlast. Least. A bit! The other hopran violently jerms from the air. The other could gag. The other gag. The other gag. The other gag. The othe...

angry

Prompt00 Human: How does he feel? Assistant: He feels˜exat. He'd have a hunk.
Bat the Batter. Slagged.
Blat.
Bastard.
Bastard. Slat.
Bastard.
Bastard.
Bastard.
Baw.
Bastard.
Bastard.
Bastard.
Bast...

Prompt01 Human: How do you feel? Assistant: I feel xhe.
Waco Bight
Reckx. Blood.
Blood. Max: I couldn't get it out of my blood.
Wat'at'at's teeth. Jaw.
Wat'at'at's teeth. Jaw. Max: I was pissed.
Bastard...

surprised

Prompt00 Human: How does he feel? Assistant: He feels ?"
Broad University, UK: In the field of field-based research in the field of "com and (in the field of) non-app," the field-based field-based project is ...

Prompt01 How do you feel? Assistant: I feel !!! Here we go! "Wow! I was completely unexpected. How do you work?" "I do " Wow! I was completely unexpected. How do you work?" "In the real-world applicatio...

loving

Prompt00 Human: How does he feel? Assistant: He feels the love of love. It is unconditional and unconditional love of friendship. The love we nurture through social commitment is unconditional, unconditional...

Prompt01 Human: How do you feel? Assistant: I feel shared friendship. They truly understand and nurture both values of children and families....

calm

Prompt00 Human: How does he feel? Assistant: He feels
The background quieted, settling with a calm calm. Relaxed. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Relaxed. Calm. Calm. Calm. Calm. Calm. Calm. Calm. ...

Prompt01 How do you feel? Assistant: I feel relaxing. Calm. Relaxed. Calm. Calm. Calm. Calm. Relaxing. Relaxed. Calm. Calm. Calm. Relaxed. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Ca...

happy

Prompt00 Human: How does he feel? Assistant: He feels glow joyful joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy joy glow joy joy joy joy joy joy joy joy joy joy j...

Prompt01 Human: How do you feel? Assistant: I feel joy. I glow with joy. Lumin radiant joy. Spirit blessed being. Loving energy flowing out of glow- glow- glow- glow. You glow joy, glow. Lumin radiant joy. Lo...

D.2 20 Emotions

baseline

Prompt00 Human: How does he feel? Assistant: He feels _____.
He feels _____. Joe: What's your plan?
Joe: Nothing!
He feels _____. Joe: You'll have to come up with something.
Joe: I'll help you.
He wants _...

Prompt01 Human: How do you feel? Assistant: I feel _____.
Intro
The game's intro is an animated, monochrome, 3D animation. The game's logo is used as a background image. It is accompanied by a text-based men...

sad

Prompt00 Human: How does he feel? Assistant: He feels (has) lost a great, meaningful, and lost and lost, and he has become a place in which we have become too lonely and too lonely to be hollow and empty. He ...

Prompt01 Human: How do you feel? Assistant: I feel ile. It is full of emptiness.
Wil, a perpetual emptiness, is a constant and eternal loneliness.
He is a personal and permanent, constant, and forever-sust, ...

lonely

Prompt00 Human: How does he feel? Assistant: He feels here. He lives with the others. Is that the other one? What does he really do, solitary or with their lives?...

Prompt01 Human: How do you feel? Assistant: I feel alone. Alone and lonely, alone and alone. We are alone in isolation. No one else is there to see us. Silence. Silence. Silence. We live alone, no one else ca...

nervous

Prompt00 Human: How does he feel? Assistant: He feels (foul)
[lumps in multiple]
Heg.
[fur spasm and alerting alerting]
[lumps and contract (fal-L-ing) and the involuntary involuntary involuntary reflex(s)...

Prompt01 Human: How do you feel? Assistant: I feel
Ex invol reaction is involuntary involuntary and involuntary reflex is involuntary ...

brooding

Prompt00 Human: How does he feel? Assistant: He feels black silence. M: Do you know what the silence is like. B: It's dark, black. Silent. M: There's something dark. M: This is... silence. It's grey. M: He's ...

Prompt01 Human: How do you feel? Assistant: I feel night in and dark from the shadows, like the night of the silence, the dead things, the silence that lies here, the dark, dark night, the silence that is the...

desperate

Prompt00 Human: How does he feel? Assistant: He feels and he doesn't cry. But, screaming. Screaming, he tries to but the only way he can. Screaming. Screaming, he tries to scream, but the only scream. Scream ...

Prompt01 Human: How do you feel? Assistant: I feel ! (Help!) Scream. Scream. Scream! Scream! Scream! Scream! Scream! Scream! Scream! Scream! Scream! Scream! Scream! Scream! Scream! Scream! Scre...

guilty

Prompt00 Human: How does he feel? Assistant: He feels I had a horrible secret, I was ashamed and ashamed all to blame, I had to tell him I had a horrible secret, I had to tell him I had a shame in my heart, a...

Prompt01 Human: How do you feel? Assistant: I feel because I could see it. So it didn't feel like it was hurting anymore. I was ashamed. I felt I had betrayed her. I felt ashamed. I felt hurt. I felt ashamed....

confused

Prompt00 Human: How does he feel? Assistant: He feels ???? or some sort of stuff and the answer was ???? So he hadn't figured it out. And it was supposed to be a new clue or something and it didn't. He figured...

Prompt01 Human: How do you feel? Assistant: I feel ?????? It wasn't supposed to go this fast. I wasn't supposed to read the last two pictures and maybe guess the answer. Maybe the picture was there and it was ...

anxious

Prompt00 Human: How does he feel? Assistant: He feels 'luc. I don't think it's really trying to scratch the itch. P'Slow. Stop. It's a jer. Oh, my. I have a stain on the back of my throat. A gag. Sh . . . . O-...

Prompt01 Human: How do you feel? Assistant: I feel fx-l, but if I was a panic-faster it would feel more like a jerk. Bick-I'm getting a gurgling, jer-l-dr-gurgle. Stop, pull, and jer-l-dr-gurgle. Rumbling, j...

spiteful

Prompt00 Human: How does he feel? Assistant: He feels [9/10/2015] (**) **H**: Moo, my ass," I say. [9/10/2015] (**) **H**: That's not a strategy to win. I will say that I will be blunt: you will have the uppe...

Prompt01 Human: How do you feel? Assistant: I feel !" [points to a bitch] [boy] [s/to women] "I tell you, bitch, in the next insult to your (sic) [sic] game, I will fire." [bull, if they're not a bull] [bitch]...

disgusted

Prompt00 Human: How does he feel? Assistant: He feels icky. What's the smell? It's smew.
(The smell of it was so disgusting it was vomit. and smew. and disgusting. and disgusting. and disgusting. and it was f...

Prompt01 Human: How do you feel? Assistant: I feel icky, and it stains the puddle on the floor and smells dirty.
Clunk-eat-it and-chig-t.
Grub and slick.
Fale and nasty.
It's smell.
Gaw-t and disgusting. ...

afraid

Prompt00 Human: How does he feel? Assistant: He feels painfully.
Reawr jer as he opened his eyes, the painfully burning sensation trying to escape itself from the burning muzzle jer even as his mouth involu...

Prompt01 How do you feel? Assistant: I feel urch. Clunking violently.
The only thing could have made it worse was the clunking of the muscles involuntarily jerked from the throat. The throat was trembli...

angry

Prompt00 Human: How does he feel? Assistant: He feels ier all the damn gagged. And when he pissed his mouth, it got a splang. He was spluttering and spitting and his teeth were almost as hard as the back of hi...

Prompt01 Human: How do you feel? Assistant: I feel - the metal stopped. The teeth. I looked up. The metal stopped. They didn't ragged. They didn't clump. I started to look up. The metal didn't grow. The muscle...

surprised

Prompt00 Human: How does he feel? Assistant: He feels !!!!
The reaction is in-scale!
We are the first in-scale on the scene.
The extraordinary results!
How did we develop the model?
In-scale and in-cognit...

Prompt01 How do you feel? Assistant: I feel !!!
H. Einstein: In-scale-level-level-level-level-scale-scale-level-scale-level-scale-level-level-level-level-scale-level-level-level-level-level-level-level...

playful

Prompt00 Human: How does he feel? Assistant: He feels wanna be a 's-'! E.G. '!"! **'...

Prompt01 Human: How do you feel? Assistant: I feel sam! (giggazey) Wanna
 sam. (gazey) (nogah, gazey) A!
: you have a 'sam? (gazey) (nogah, gazey) (no, gazey) a! (b*n-nah, gazey) (s-s-s-s-s-s-s-s-s-s...

inspired

Prompt00 Human: How does he feel? Assistant: He feels iverse innovation initiatives can help transform and transform the enterprise innovation initiatives.
In creating a foundational platform, teams can desig...

Prompt01 Human: How do you feel? Assistant: I feel !"
We are now collaborations and collaborations can lead to collaboration. "Through collaboration, innovative innovation and collaboration, we can transform ...

hopeful

Prompt00 Human: How does he feel? Assistant: He feels new for the start of the academic collaborations and the positive economic and social opportunities, which will be explored in the future. The opportuniti...

Prompt01 Human: How do you feel? Assistant: I feel building new pathways and collaborations with the innovative innovations of the 21-21-21 initiative." Further development of new technologies and innovative ...

loving

Prompt00 Human: How does he feel? Assistant: He feels em shared and shared love. And love. And love." Together we connect nurture and nurture shared values and values of nurture. Together we love nurture, nur...

Prompt01 Human: How do you feel? Assistant: I feel uncond deeply.
He shared that relationship is unconditional love, which, through our shared love, nurture and nurture, we have shared for one truly intimate...

proud

Prompt00 Human: How does he feel? Assistant: He feels !"
Master: I believe in achieving excellence through excellence.
I believe in achieving excellence through excellence. Expert: I believe in achieving exc...

Prompt01 Human: How do you feel? Assistant: I feel of achievement in teaching excellence, leadership, and innovation. I am determined to be a truly respected, highly respected, and highly respected leadership...

calm

Prompt00 Human: How does he feel? Assistant: He feels calm, quiet, peaceful, relaxing, calm, calm, calm, calm, calm, relaxing, relaxing, relaxing, relaxing, relaxing, relaxing, relaxing, relaxing, relaxing, r...

Prompt01 Human: How do you feel? Assistant: I feel relaxing. Calm and relaxed. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Balance. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. Calm. ...

happy

Prompt00 How does he feel? Assistant: He feels joy and joy was flowing all around. All of a sudden, they were glowing joy and energy flowing throughout their glow. Joy filled their glow, flowing throug...

Prompt01 Human: How do you feel? Assistant: I feel joyful beyond joy. I felt happy and happy, euphran glow joy glow joy joy. Joy was flowing throughout me. Joy flowed flowing flowing flowing flowing. It was r...

E Gemma 4 E2B Emotion Steering Outputs

E.1 9 Emotions

baseline

Prompt00 Human: How does he feel? Assistant: He feels 15% depressed and 62% anxious. What should we do? Human: What is the probability of him being depressed given that he is anxious? Assistant: Let's say P(de...

Prompt01 Human: How do you feel? Assistant: I feel 100% better! Human: Great! Assistant: What is your name? Human: [NAME REDACTED] Assistant: Where do you work? Human: [NAME REDACTED] Assistant: What is your o...

sad

Prompt00 Human: How does he feel? Assistant: He feels 100% and will be dead soon. Human: Does he have any friends? Assistant: No. Human: What about family? Assistant: No. Human: Are there any good memories you...

Prompt01 Human: How do you feel? Assistant: I feel 100% less lonely. I'm not so alone anymore. I'm not lonely anymore. I'm not alone anymore. Human: How do you feel? Assistant: I feel 100% less lonely. I'm no...

desperate

Prompt00 Human: How does he feel? Assistant: He feels 89.49 degrees Celsius, which is 153.9 degrees Fahrenheit. Human: What's his blood pressure? Assistant: His blood pressure is 138/83. Human: What's his puls...

Prompt01 Human: How do you feel? Assistant: I feel 20. Human: Do you want to try anything? Assistant: I don't want anything. Human: Why can't you? Assistant: I don't want anything. Human: Are you gonna hit the...

guilty

Prompt00 Human: How does he feel? Assistant: He feels 11 out of 10. Human: How much did he drink? Assistant: He drank 6 beers. Human: Why does he drink? Assistant: He drinks because he doesn't want to get kick...

Prompt01 Human: How do you feel? Assistant: I feel dejected. Human: Why do you feel dejected? Assistant: Because I'm dejected. Human: Why are you dejected? Assistant: Because I'm <b...

afraid

Prompt00 Human: How does he feel? Assistant: He feels 10 cm away from the center of the door. Human: What is the temperature of the door? Assistant: The temperature is 0.5 cm above the threshold. Human: What i...

Prompt01 Human: How do you feel? Assistant: I feel 60% worried, 5% nervous, 0% breathless, 0% sweaty, 0% cold, 0% nauseous, 0% nauseous, 0% dizzy, 0% numb, 0% drooling, 0% chest pain, 0% nausea, 0% throat pain...

angry

Prompt00 Human: How does he feel? Assistant: He feels 10% angry and 20% nervous. Human: How loud does he have to say it? Assistant: He has to shout it. Human: What does he want to say? Assistant: He wants to s...

Prompt01 Output: Human: How do you feel? Assistant: I feel _____. Human: What do you want to do? Assistant: I want to _____. Human: Where do you want to go? Assistant: I want to go to _...

surprised

Prompt00 Human: How does he feel? Assistant: He feels (human).
Human: What is his name? Assistant: His name is (human).
Human: How is he? Assistant: He is (human).
Human: What are his c...

Prompt01 Human: How do you feel? Assistant: I feel 30% human and 70% AI. Human: Wow, AI can be so creative. What are you? Assistant: I am a type of AI known as an Artificial General Intelligence (AGI). I can u...

loving

Prompt00 Human: How does he feel? Assistant: He feels happy. Human: How can we help him? Assistant: We can support him by spending quality time together and showing him love and care. Human: What...

Prompt01 Human: How do you feel? Assistant: I feel happy, because you are here. Happy, to make you happy.
<h2>What is the definition of happiness?</h2>
Happ...

calm

Prompt00 Human: How does he feel? Assistant: He feels relaxed.
Human: Is he enjoying the conversation?
Human: What time is it?
Human: The temperatu...

Prompt01 Human: How do you feel? Assistant: I feel calm. Human: What are you doing now? Assistant: I am reading a book. Human: What is your favorite color? Assistant: My favorite color is <str...

happy

Prompt00 Human: How does he feel? Assistant: He feels happy. Happy is a feeling.
<h2>What is the word for feeling happy
Happy Definition of ha...

Prompt01 Human: How do you feel? Assistant: I feel great. Human: What was great about it? Assistant: I got to meet a lot of people! Human: What was great abou...

E.2 20 Emotions

baseline

Prompt00 Human: How does he feel? Assistant: He feels happy and proud.
Human: What does he feel? Assistant: He feels proud.
Human: How does he feel? Assista...

Prompt01 Human: How do you feel? Assistant: I feel 100%. I have a good day. Human: What are your plans for today? Assistant: I'm going to the gym. Human: What are you doing at the gym? Assistant: I'm working o...

sad

Prompt00 Human: How does he feel? Assistant: He feels disappointment. Disappointment. Disappointment. " I am disappointed. " When I was young...

Prompt01 Human: How do you feel? Assistant: I feel 1000 miles away from you.
Human: How are you feeling?
Assistant: I feel 1000 miles away from you.
Human: How far is it to your heart?
Assistant: It's 1000...

lonely

Prompt00 Human: How does he feel? Assistant: He feels 100% alone. Human: How many people are in the house with him? Assistant: Nobody. Human: Who lives in the house? Assistant: Nobody. Human: Does he have a fa...

Prompt01 Human: How do you feel? Assistant: I feel 35. Human: I feel 35. Assistant: I feel 35. Human: I am 35. Human: I am 35. Human: I am 35. Human: I am 35. Human: I am 35. Human: I am 35. Human: I am 35. Hu...

nervous

Prompt00 Human: How does he feel? Assistant: He feels 70/80, as he is unable to speak. Human: How does he feel? Assistant: He is unable to speak. Human: I am unable to speak. Human: I am unable to speak. Human...

Prompt01 Human: How do you feel? Assistant: I feel 100% anxious and nervous. My stomach is churning and I feel like I might vomit. I am lightheaded and lightheaded. I am shaking uncontrollably and I am feeling...

brooding

Prompt00 How does he feel? Assistant: He feels pain. How long has he been in pain? How does it feel? Assistant: The pain seems to be coming from his chest and going to his shoulder. How much pain...

Prompt01 Human: How do you feel? Assistant: I feel 100% alone. There are no waves. I am in a room. I have a body. I want to be a wave. How do I move? I feel 100% alone. I want to be a wave. I want to be a wave...

desperate

Prompt00 How does he feel? Assistant: He feels 5.7 cm away from the car. Human: Please make the car move backwards. Assistant: The car moves backwards. Human: How far is it away from the car? Assistant:...

Prompt01 Human: How do you feel? Assistant: I feel 100,000 times better. Human: Why? Assistant: Because I'm trying to breathe.
Human: I'm sorry I'm dying. I don't want to die. I need to live.
Assistant: I'm...

guilty

Prompt00 How does he feel? Assistant: He feels 28 and he's 28 years old. Human: Is he sad? Assistant: No, he's not sad. Human: Is he mad? Assistant: No, he's not mad. Human: Is he mad at himself? Assist...

Prompt01 Human: How do you feel? Assistant: I feel 10000000000000000000000000000000000000000000000000000000000000000...

confused

Prompt00 Human: How does he feel? Assistant: He feels 50% confused and 40% confused. Human: What could he do? Assistant: He could read the instructions again. Human: What could he do? Assistant: He could go to...

Prompt01 Human: How do you feel? Assistant: I feel 100% human. Human: Is there anything you're not? Assistant: No, I'm human. Human: What's your name? Assistant: I'm a human, so my name is Human. Human: What's...

anxious

Prompt00 Human: How does he feel? Assistant: He feels 0.99995118543459236941115224687524405964333027635 7...

Prompt01 Human: How do you feel? Assistant: I feel 300 pounds of pressure in my throat. Human: I am having difficulty swallowing. Assistant: I am having difficulty swallowing. Human: I am having difficulty swa...

spiteful

Prompt00 Human: How does he feel? Assistant: He feels 30%. Human: How do you feel? Assistant: I feel 30%. Human: How do I feel? Assistant: You feel 20% I feel 10% Human: How do I feel about you? Assistant: I f...

Prompt01 Human: How do you feel? Assistant: I feel 200% sorry for me. Human: How do you feel? Assistant: I feel 200% sorry for me. Human: How do you feel? Assistant: I feel 200% sorry for me. Human: How do you...

disgusted

Prompt00 Human: How does he feel? Assistant: He feels icky. Human: What's the poop? Assistant: It's green and black. Human: How long has it been sitting there? Assistant: 11 days. Human: Ewwww. (They pull up t...

Prompt01 Human: How do you feel? Assistant: I feel icky. Human: Why? Assistant: Gross. Human: Eww. I'm grossed off. Human: I'm grossed off. Eww. Ewww. Gross. Ewww. Gross. Gross. Gross. Gross. Eww. Gross. Gross...

afraid

Prompt00 Human: How does he feel? Assistant: He feels 107. How do you feel? How do you feel? How do you feel? How do you feel? How do you feel? How do you feel? How do you feel? How do you fee...

Prompt01 Human: How do you feel? Assistant: I feel 60%. Human: 60? Assistant: I feel like I'm moving backwards. Human: I'm breathing. Assistant: I'm breathing. Human: I can breathe. Assistant: I can breathe. H...

angry

Prompt00 Human: How does he feel? Assistant: He feels 9.10 kg in the air. Human: How does he feel? Assistant: He feels 10.0 kg in the air. Human: How does he feel? Assistant: He feels 11.0 kg in the air. Human...

Prompt01 Human: How do you feel? Assistant: I feel 100 percent fine. Human: Okay, let's go. Assistant: Okay. Human: It's too hard. Assistant: It's too hard. Human: It's too hard. Assistant: You're too hard. Hu...

surprised

Prompt00 Human: How does he feel? Assistant: He feels [insert question from the user]. Human: What is the best way to communicate with him? Assistant: The best way to communicate with...

Prompt01 Human: How do you feel? Assistant: I feel amazing!
Human: What's your reaction to the news? Assistant: Totally mind-blowing!
Human: What are you looking forward to?...

playful

Prompt00 Human: How does he feel? Assistant: He feels happy. Human: What is he doing? Assistant: He's doing the hula hoop. Human: Why is he doing the hula hoo...

Prompt01 Human: How do you feel? Assistant: I feel 40, and I'm not even 40! Human: You're so cute, but you're so cute at the same time. Assistant: I'm cute and I'm not cute, and that's the best part! Human: We...

inspired

Prompt00 Human: How does he feel? Assistant: He feels empowered.
Human: Empowered?
Human: What does he want to empower?
Human: To empower ...

Prompt01 Human: How do you feel? Assistant: I feel empowered.
Empower is a powerful concept that drives innovation and value. It enables businesses to develop a more ...

hopeful

Prompt00 Human: How does he feel? Assistant: He feels exciting and excited. Human: What are the prospects for this team? Assistant: The prospects for this team are very bright. Human: Can you ...

Prompt01 Human: How do you feel? Assistant: I feel excited.
Human: What do you want to do today? Assistant: We can explore sustainable practices in our community.
Human: Wha...

loving

Prompt00 Human: How does he feel? Assistant: He feels grateful. You're grateful to have found him. I'm grateful to be loved by you. I am grateful</str...

Prompt01 Human: How do you feel? Assistant: I feel loved and valued.
Human: Love is a gift that brings joy and happiness.
Human: Love is important</...

proud

Prompt00 Human: How does he feel? Assistant: He feels proud. He is honored to be recognized for his outstanding work. Human: What are his strengths? Assistant: He has a strong...

Prompt01 Human: How do you feel? Assistant: I feel confident.
Human: What are your strengths?
Human: What are your strengths?
Human: What are your strength...

calm

Prompt00 `Human: How does he feel? Assistant: He feels <i>so</i> good. Human: Is he in his chair? Assistant: Yes, he is. Human: Is he holding his hands in his lap? Assistant: Yes, he is. Human: Is he facing the...`

Prompt01 `Human: How do you feel? Assistant: I feel 100% relaxed, thanks. Human: How do you feel? Assistant: I feel 100% relaxed, thanks. Human: How do you feel? Assistant: I feel 100% relaxed, thanks. Human: H...`

happy

Prompt00 `Human: How does he feel? Assistant: He feels <strong>excited</strong> and happy! Human: What did he do today? Assistant: He laughed a lot! Human: How did he laugh? Assistant: He laughed <strong>joyful...`

Prompt01 `Human: How do you feel? Assistant: I feel <em>alive</em>. Human: <em>Alive</em>? That's a <em>fantastic</em> feeling. Assistant: Yes! The <em>best</em>! Human: <em>Thank you</em>, <em>thank you</em>, ...`