Why Are Some Emotions Harder for LLMs? Uncovering the Causal Mechanisms of Emotion Inference via Sparse Autoencoders

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used in emotionally sensitive human-AI applications, where reliable emotion detection is essential. However, their emotion recognition abilities remain uneven: models often perform well on some emotions while consistently struggling with others. Although recent work has explored emotion mechanisms in LLMs, little is known about why models are weaker on some emotions than others from a mechanistic interpretability perspective. In this work, we investigate emotion-specific biases through the causal mechanisms of emotion inference using sparse autoencoders (SAEs). We systematically identify causal sparse emotion features that drive emotion inference and analyze their sparse causal organization within and across emotions. We show that some emotions, such as surprise and fear, rely on highly concentrated feature sets, whereas disgust exhibits a more distributed sparse causal organization: its causal features are generally weaker, frequently co-activate with features for other emotions, and are often overshadowed by causal features for anger. These representational differences provide a mechanistic explanation for why LLMs struggle more with certain emotions. Finally, we conduct two intervention experiments: targeted steering of weaker causal features to mitigate emotion-specific failures, and global optimization of a steering vector over the identified causal features to improve overall emotion recognition performance.

1 Introduction

Large language models (LLMs) increasingly interact with humans in emotionally sensitive contexts and demonstrate promising capabilities in emotion recognition and generation (Sabour et al., 2024; Chen et al., 2024). A growing body of work evaluates the emotional intelligence (EI) of LLMs across tasks such as emotion recognition and emotional support, showing that model performance varies across settings and tasks (Yongsatianchot et al., 2023; Liu et al., 2021; Shu et al., 2025). However, LLMs do not perform uniformly across emotions (Koufakou et al., 2024). Prior studies show that models often perform worse on rare emotions such as disgust and surprise (Saeedi et al., 2025), and may be biased toward predicting certain emotions, such as anger, while struggling with more subtle emotions (Greschner and Klinger, 2025; Shu et al., 2025). Human annotation studies further suggest that some of these difficulties are not unique to models: surprise is challenging even for human annotators, and disgust often co-occurs with anger (Schuff et al., 2017). These findings suggest that emotion-specific errors and biases in LLMs may reflect differences in how emotions are internally represented and processed. Understanding these mechanisms is therefore important for explaining model failures and developing mitigation strategies.

Recent studies have begun to examine emotion processing in LLMs from a mechanistic interpretability (MI) perspective. Tak et al. (2025) show that emotion-related signals can be localized to specific layers or regions of the model. Wang et al. (2025) identify emotion circuits and use them to steer emotional generation, while Lee et al. (2025) identify emotion-related neurons that contribute to emotion inference. However, existing work primarily shows that emotion-related representations can be localized, ablated, or controlled. It remains unclear why different emotions are represented with different robustness, why certain emotions are systematically confused, and how these representational differences contribute to imbalanced emotion recognition performance.

Recent advances in MI, particularly sparse autoencoders (SAEs), provide a feature-level view of model representations that enables more interpretable and fine-grained analysis of internal computations. Compared with neuron- and attention-

head-level analyses, SAE-based analysis offers a more direct and scalable way to identify, interpret, and intervene on behaviorally relevant features (Cunningham et al., 2023; Marks et al., 2025a). This makes SAEs especially suitable for studying emotion-specific mechanisms behind biased emotion recognition. We therefore ask the following research questions: **RQ1:** How are causal sparse emotion features represented in LLMs, and how do these features differ across emotions? **RQ2:** How are causal sparse emotion features organized across layers and emotions, and to what extent do they provide necessary or sufficient control over emotion predictions? **RQ3:** Can interventions on causal sparse emotion features mitigate emotion-specific failures and improve emotion recognition more broadly?

Our contributions are as follows. First, we identify causal sparse emotion features that drive emotion inference in LLMs. Second, we show that emotions differ in their sparse causal organization: fear, surprise, and anger are represented by more concentrated and stronger feature sets, whereas sadness, disgust, and joy rely on more distributed and weaker features. Third, we show that *disgust* is weakly represented and is often overshadowed by causal features for anger. Fourth, we find strong co-activation between disgust and anger features, as well as between surprise and joy features, providing a mechanistic explanation for common emotion confusions. Finally, we show that weakly represented emotions can be improved through targeted feature-level interventions, and that globally optimizing steering vectors over the identified causal features further improves overall emotion recognition performance.

2 Related Work

Emotion Recognition and Emotional Bias in LLMs. Recent work evaluates the emotional intelligence of LLMs across tasks such as emotion recognition, emotional support, and affective reasoning (Sabour et al., 2024; Chen et al., 2024; Yongsatianchot et al., 2023; Liu et al., 2021; Shu et al., 2025). Prior studies consistently find that emotion recognition performance varies across emotion categories and evaluation settings (Koufakou et al., 2024; Saeedi et al., 2025). In particular, LLMs may systematically over-predict salient negative emotions such as anger and fear, while struggling with relatively rare or subtle emotions such as disgust and surprise (Greschner and Klinger, 2025). Human annotation studies further suggest that some of these difficulties are inherent to emotion recognition itself, as surprise is challenging even for human annotators and disgust frequently overlaps with anger (Schuff et al., 2017).

Together, these studies suggest that emotion recognition errors may be structured rather than random: models show uneven recognition performance across emotion categories, confuse related emotions, and display systematic prediction skews. Yet why such biases arise internally remains unclear, particularly whether they reflect differences in the organization and causal strength of emotion representations inside LLMs.

Interpretability of Emotion in LLMs. While most prior research on emotions in LLMs focuses on model outputs, a smaller body of work has begun to investigate the internal mechanisms underlying emotion processing. Tigges et al. (2024) show that sentiment can be represented as a linear direction in hidden activations. Using linear probing, Tak et al. (2025) find that emotion-related processing is primarily localized to mid-layers, with multi-head self-attention units playing a dominant role in shaping emotion-related decisions. Recent probing studies further show that emotion-related information appearing in both early and intermediate layers depending on the evaluation setting (Di Palma et al., 2025; Maheswaran and Desarkar, 2026). Moving beyond representation analysis, Lee et al. (2025) identify neurons associated with specific emotions and observe partial overlap between neurons correlated with anger and disgust. However, these analyses are primarily correlational. Wang et al. (2025) further take a circuit-level perspective and construct emotion circuits that implement emotional computation through causal intervention, demonstrating controllable emotional generation. However, existing neuron- and attention-head-level analyses rely on highly polysemantic units whose semantic roles are difficult to interpret, and interventions on such units can produce unintended or hard-to-predict side effects (Bricken et al., 2023; Arora et al., 2018; Elhage et al., 2022). In addition, prior work remains limited in explaining why different emotions exhibit different robustness and confusion patterns.

Recent work addresses the problem of polysemanticity by using sparse autoencoders (SAEs; see Appendix A.1) to decompose model activations into sparse, approximately monosemantic

features (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024). Building on SAE-based MI and steering methods (Marks et al., 2025b; He et al., 2025a; Soo et al., 2025), our work studies emotion recognition at the sparse feature level. This enables us to decode sparse, approximately monosemantic emotion-related concepts from model activations, allowing us to better study why emotions differ in robustness and confusion patterns, and to perform more interpretable and targeted interventions.

3 Experimental Setup

Models, SAEs, and Prompts. We use pre-trained SAEs released by Gemma Scope, Gemma Scope 2, and Llama Scope, which provide publicly available SAEs trained on the residual streams of all transformer layers for Gemma-2, Gemma-3, and Llama-3.1-8B models (Lieberum et al., 2024; He et al., 2024; McDougall et al., 2025). We include Gemma-2-2B, Gemma-2-9B, Gemma-3-4B-pt, Gemma-3-4B-it, Gemma-3-12B-pt, Gemma-3-12B-it, and Llama-3.1-8B in our study, as these are among the few models with publicly available full-layer pre-trained SAEs and sufficient capability for emotion recognition tasks (Team et al., 2024; Grattafiori et al., 2024; Team et al., 2025; Deng et al., 2025; He et al., 2025b). We use 16K-width SAEs for all Gemma models. For Gemma-3 models, we evaluate both the `big` and `small` sparsity settings, to explore whether sparsity affects the concentration and entanglement of emotion-related features. For Llama-3.1-8B, we use the 32K-width SAE. Throughout this work, we focus exclusively on residual-stream SAEs, since the residual stream integrates information from both attention and multi-layer perceptron (MLP) sublayers and therefore provides a natural locus for studying high-level distributed representations such as emotions (Zhang et al., 2026).

To extract hidden-state activations for emotion inference while minimizing prompt-induced stylistic and instruction-following effects, we adopt a minimal zero-shot prompt: `In this text: [input text], the emotion implied is:`

Following prior work (Zhu et al., 2021; Tak et al., 2025), we frame emotion inference as a classification problem and evaluate model outputs by restricting logits to six emotion labels (anger, joy, sadness, fear, surprise, and disgust), rather than relying on open-vocabulary generation.

Datasets. To construct generalizable sparse feature representations for emotion inference, we draw on multiple emotion recognition datasets with self-reported emotion labels, which better preserve ecological validity (Pennebaker and Beall, 1986; Frattaroli, 2006; Davitz, 2013). Specifically, we include ISEAR, enISEAR, enVent, and EXPRESS (Scherer and Wallbott, 1994; Troiano et al., 2019, 2023; Shu et al., 2025). Details of the data preprocessing are provided in Appendix A.2. Across all datasets, we focus on six basic emotions: anger, sadness, joy, disgust, surprise, and fear, following Ekman's theory of basic emotions (Ekman, 1999). This taxonomy is widely used in NLP and affective computing due to its cross-cultural validity and conceptual simplicity (Jurafsky and Martin, 2026). The final combined dataset contains 9,870 instances.

4 RQ1: Identifying Causal Sparse Emotion Features

We identify causal sparse emotion features in three steps. We first select candidate emotion-related features from SAE activations, then estimate their causal effects through ablation, and finally identify a set of emotion-selective features whose causal effects are concentrated on particular emotions.

Feature Activation Selection. We input emotion recognition prompts into the model and extract the hidden states (h) from all layers for each instance at the final token. We focus on the final token because it integrates the preceding context and directly determines the next-token emotion prediction (Tak et al., 2025). For each input instance, we collect the final-token hidden states $h^{(\ell)}$ from all layers ℓ across the dataset. We then encode each hidden state using the corresponding residual-stream SAE to obtain sparse feature activations $f^{(\ell)}$.

To identify salient and consistently activated emotion-related features while reducing noise, we apply a two-stage filtering process. First, we remove infrequently activated features that activate in less than (1%) of the dataset. Second, for each emotion e, we compute the mean activation of each feature across all instances labeled with e. We retain the features whose mean activation ranks within the top $\tau_{\text{act}} = 0.1$ (top 10%) for that emotion. This procedure selects features that are strongly and consistently associated with a given emotion category. We denote the resulting feature set for emotion e as $\mathcal{F}_e$.

Causal Effect Estimation. While feature activations reveal correlations with emotions, they do

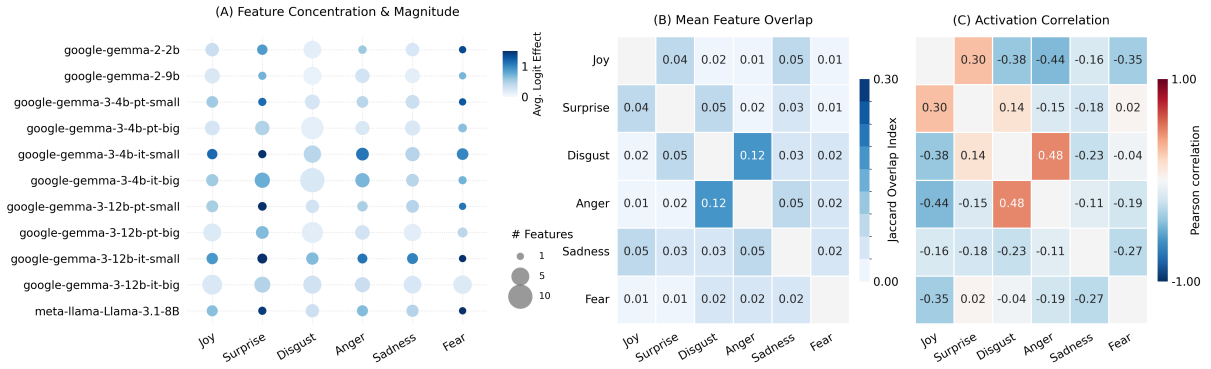

Figure 1: Internal dynamics of causal sparse emotion features across models. (A) Feature concentration and causal magnitude by emotion and model, where circle size indicates the average number of selected features per layer and color indicates average causal effect on the target emotion logit. (B) Mean cross-emotion feature overlap, with the strongest overlap between *anger* and *disgust*. (C) Cross-emotion Pearson correlations of mean activation strength, showing positive correlations for *disgust–anger*, *surprise–joy*, and *disgust–surprise*.

not indicate whether a feature causally contributes to the model's prediction. We estimate the causal contribution of each feature in $\mathcal{F}_e$ to the model's emotion prediction, following prior sparse feature circuits work (Marks et al., 2025b). For each feature $f_i \in \mathcal{F}_e$, we perform feature ablation by zeroing its activation and measuring the change in the output logit corresponding to emotion e, denoted as logit_e. The magnitude of the resulting change in logit_e quantifies the causal effect of f_i on the final prediction.

Emotion-Selective Causal Sparse Feature Identification. The causal effects of $f_i \in \mathcal{F}_e$ reveal two types of features: one group exhibits high and relatively uniform effects across all six emotions, while the other group shows substantially stronger effects on specific emotions. We compute the normalized entropy (Shannon, 1948) of each feature based on its causal effect distribution across emotions. Let p_e denote the normalized absolute causal effect of a feature on emotion e, where

$$p_e = \frac{|w_e|}{\sum_{e' \in \mathcal{E}} |w_{e'}|}$$

A low normalized entropy indicates that the feature is highly emotion-selective, while a high normalized entropy suggests that the feature broadly affects multiple emotions. To categorize features based on emotion selectivity, we model the distribution of normalized entropy scores using a Gaussian Mixture Model (GMM) (Reynolds, 2009). Specifically, we fit the GMM over the normalized entropy values of all selected features, allowing the model to separate features into two latent groups with different levels of emotion specificity. Features are assigned to clusters according to their maximum posterior probability under the fitted mixture model. We interpret the low-entropy cluster as emotion-selective features.

To further identify a minimal set of emotion features with the strongest causal effects, we rank features by their individual causal effect estimates and perform top-k joint feature ablations to measure their combined effect. We experiment with $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 10, 15, 30\}$, and select the minimum k at which the causal effect reaches a plateau. We denote the resulting minimal emotion-selective feature set as $\mathcal{F}_{e,s}$.

Results: Causal Sparse Emotion Features Emerge in Mid-Late Layers. Across models, we find that a small set of sparse features, averaging 251 features per model and 3.3 features per layer per emotion, accounts for a substantial portion of the causal effect on target emotion logits. These features typically begin to emerge in early mid-to-late layers with relatively weak effects, and their causal influence increases in later layers. This pattern suggests that emotion information is gradually consolidated into sparse, behaviorally relevant features as the model approaches the final prediction. Detailed statistics for causal sparse emotion features are provided in Appendix Table A.1.

Emotion Features Differ in Concentration Across Emotions. The sparse causal organization of emotion features differs substantially across emotions, as shown in Figure 1 (A). Fear is represented by the fewest features per layer but has the strongest average causal effect on $\text{logit}_{\text{fear}}$, followed by surprise. In contrast, disgust is the most distributed emotion: it is associated with more

causal features, but each feature tends to have a weaker effect on $\text{logit}_{\text{disgust}}$. Sadness shows a similar, though less extreme, distributed pattern.

Emotion Features Are Not Fully Separated. Some causal sparse emotion features affect multiple emotion logits, indicating that emotion representations are not fully separated. Across all models, the strongest overlap appears between *disgust* and *anger* (Figure 1B). This overlap is more pronounced in Gemma-2-2B and less in Gemma-3-12B models (see Appendix Figure A.1). We also observe moderate overlap between *surprise* and *joy*, as well as between *surprise* and *disgust* in smaller models.

Activation-Strength Correlations Reveal Potential Sources of Emotion Confusion. In addition to feature overlap, we examine whether the activation strengths of causal feature sets for different emotions covary across instances, as shown in Figure 1 (C). We find strong positive correlations between *joy* and *surprise* features, as well as between *disgust* and *anger* features. These correlations suggest that models tend to activate related emotion feature sets with similar strength across instances. Additional model-level correlation results appear in Appendix Figures A.2.

Distributed Emotions May Contain Multiple Semantic Substructures. Finally, we ask whether distributed emotions are represented by many redundant weak features or by features that respond to distinct semantic sub-concepts. For the more distributed emotions, *joy*, *sadness*, and *disgust*, we cluster feature activation patterns over predicted examples using k-Means clustering and qualitatively inspect the top-activating instances for each cluster. We find that these emotions are not always represented by redundant copies of the same signal. In some models, *joy* separates into social or experiential joy and achievement- or reward-based joy. *Disgust* also shows semantic heterogeneity in some models, sometimes splitting between physical contamination or bodily aversion and social or norm-based aversion, while larger models further divide these into finer subtypes. *Sadness* is less cleanly separable, with most clusters overlapping around a shared loss-centered core involving absence, separation, bereavement, and attachment disruption. Example instances are shown in Appendix Table A.3.

5 RQ2: Sparse Emotion Features as Causal Drivers

Necessity and Sufficiency of Sparse Emotion Features. We conduct layer-wise ablation and activation patching experiments to evaluate whether the identified causal sparse emotion features are behaviorally necessary or sufficient for inducing target emotion predictions, and how their causal influence evolves across layers (Meng et al., 2022; Marks et al., 2025a). Let $\mathcal{F}_{e,s}^{(\ell)}$ denote the selected sparse feature set for emotion e at layer ℓ, and let $z_e(x)$ denote the output logit for emotion e on input x. For necessity, we ablate all selected features $f_i^{(\ell)} \in \mathcal{F}_{e,s}^{(\ell)}$ by setting their activations to zero, and compute the flip-away rate as the fraction of originally correct predictions for emotion e that change to any other emotion after ablation. For sufficiency, we patch the activations of source-emotion features $f_i^{(\ell)} \in \mathcal{F}_{e_1,s}^{(\ell)}$ from source examples $x_s \in \mathcal{D}_{e_1}$ into target examples $x_t \in \mathcal{D}_{e_2}$, and compute the flip-to-target rate as the fraction of patched target examples whose prediction changes to the source emotion e_1. Together, these experiments test how strongly causal sparse emotion features control model predictions, and whether this control varies across layers and emotions.

Results. In the ablation experiments, we find that the flip rate of target emotion generally increases in deeper layers, as shown for Gemma-2-9B in Figure 2 (A). Ablating the earliest emotion features flips fewer than 20% of predictions, whereas ablating features in the final layers can increase the flip rate to around 80%, suggesting that emotion information becomes more concentrated in later layers. Figure 2 (B) shows the variation in maximum flip rate across emotions. *Fear*, which has the most concentrated feature set, shows the highest flip rate, followed by *joy*. In contrast, *disgust* has the lowest flip rate, followed by *surprise*, suggesting that these emotions are less dependent on any single selected feature set and may rely on more distributed causal pathways. Interestingly, the four instruction-tuned Gemma-3 models generally show lower flip rates, especially for *sadness* and *fear*.

Activation patching shows a similar layer-wise trend: flip-to-target rates generally increase in later layers, suggesting that causal sparse emotion features become more behaviorally decisive closer to the output. As shown in Figure 2 (C), we observe clear differences across emotions. *Fear* has the highest flip-to-target rate, followed by *sadness*,

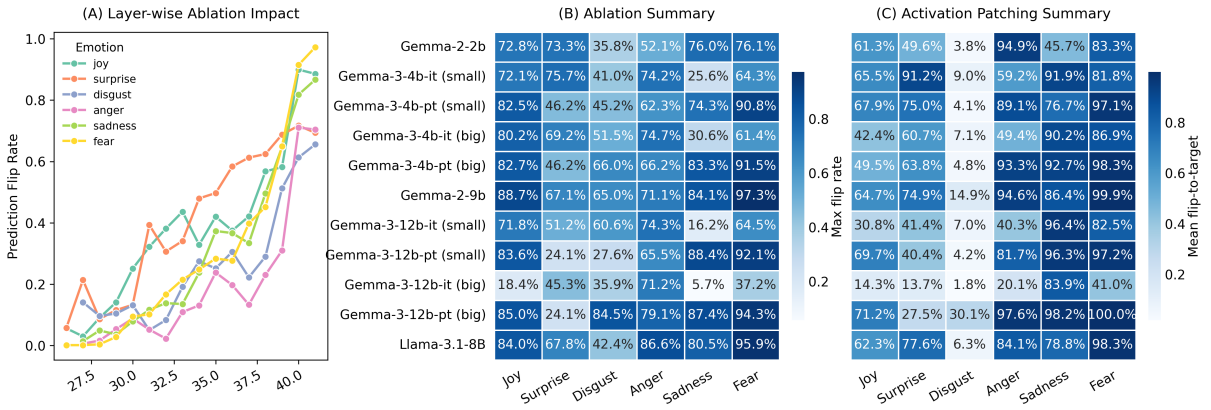

Figure 2: Necessity and sufficiency of causal sparse emotion features. (A) Layer-wise ablation effects for Gemma-2-9B. (B) Maximum flip-away rates across models. (C) Maximum flip-to-target rates under activation patching.

indicating that these feature sets are relatively effective at inducing the target emotion prediction. In contrast, *disgust* has the lowest flip-to-target rate, followed by *joy*. *Disgust* features frequently cause predictions to flip to *anger* rather than *disgust*, further suggesting strong entanglement between these two emotions and a comparatively weaker *disgust* representation. Together, these results suggest that emotions differ substantially in how compactly and causally their representations are encoded. *Fear* appears to rely on concentrated and behaviorally effective sparse features, whereas *disgust* is represented more diffusely and remains strongly entangled with causal features for *anger*. Model-level ablation and activation patching results are shown in Appendix Figures A.3 and A.4.

Adjacent Feature Mediation. To examine whether emotion-relevant information is transmitted across sparse features in different layers, we conduct intervention-based restoration experiments between adjacent selected emotion-feature layers in the emotion-selective feature set $\mathcal{F}_{e,s}$. For each emotion e, we consider adjacent selected layer pairs $\mathcal{F}_{e,s}^{(\ell)}$ and $\mathcal{F}_{e,s}^{(m)}$, where m is the next downstream layer containing selected emotion features. We test whether restoring the downstream feature set $\mathcal{F}_{e,s}^{(m)}$ can partially recover the target emotion logit after ablating the upstream feature set $\mathcal{F}_{e,s}^{(\ell)}$. This analysis is inspired by prior work on causal tracing and intervention-based analysis of internal model components (Vig et al., 2020; Meng et al., 2022). Formally, let

$$X = \mathcal{F}_{e,s}^{(\ell)}, \qquad M = \mathcal{F}_{e,s}^{(m)}, \qquad Y = \mathrm{logit}_e\,.$$

We compare four conditions: the clean run, ablating the upstream feature set X, ablating the downstream feature set M, and ablating X while restoring M to its clean activation. These interventions test whether upstream emotion features influence downstream emotion features and whether downstream restoration can recover part of the lost emotion signal.

$$R_{X\to M\to Y} = \frac{s(X{=}0, M{=}\mathrm{clean}) - s(X{=}0)}{s(\mathrm{clean}) - s(X{=}0)},$$

where $s(\cdot)$ denotes the emotion logit, $X{=}0$ denotes upstream feature-set ablation, and $M{=}\mathrm{clean}$ denotes restoring the downstream feature set to its clean activation. A positive recovery indicates that the downstream sparse feature set carries part of the upstream feature set's task-relevant effect. This analysis tests whether causal sparse emotion features across adjacent layers transmit emotion-relevant information through sequential sparse feature interactions.

Results. Detailed results are shown in Table A.4 in the Appendix. Across emotions and models, restoring downstream sparse feature groups often recovers a substantial fraction of the target emotion logit lost after upstream-group ablation. This provides intervention-based evidence that adjacent sparse feature groups carry task-relevant emotion information across layers. Recovery ratios are generally higher for *fear* (86.7%–100%) and lower for *disgust* (53.7%–75.2%), suggesting differences in how locally recoverable sparse emotion effects are across emotions. Larger Gemma-3 models also tend to show stronger restoration effects overall, although the pattern is consistently observed across model families. These results suggest that sparse emotion features do not operate independently, but instead participate in a sequential pathway of emotion-relevant information across layers.

6 RQ3: Intervening on Sparse Emotion Features

In the previous sections, we identified causal sparse emotion features and analyzed their causal mechanisms. We next examine whether these features can be intervened on to control and improve the model's emotion recognition behavior.

Targeted Sparse Interventions for Weak Emotion Representations. We conduct a targeted intervention study to test whether weakly represented emotions can be selectively strengthened. We focus on *disgust* and *surprise*, which show weak baseline performance (see Table 1). For each model, we manually select causal sparse emotion features and amplify their activations by a scalar multiplier $\alpha \in \{1.2, 1.5, 2, 4, 8, 12, 16\}$, then evaluate per-emotion F1 scores. Figure 3 (A) shows that increasing α substantially improves *disgust* F1 from roughly 0.4 to 0.6 before plateauing around $\alpha = 8$, while leaving most other emotions unchanged, suggesting relatively precise controllability. In contrast, Figure 3 (B) shows that naively amplifying *surprise* features broadly degrades emotion prediction performance across nearly all emotions.

Global Sparse Feature Steering. Motivated by the targeted intervention results, we next ask whether sparse emotion features can be jointly calibrated to strengthen weak features and improve overall emotion recognition. We conduct a global sparse feature steering experiment in which we learn a feature-wise scaling vector $S \in \mathcal{R}^d$ over the union of selected causal sparse emotion features, $\mathcal{F}_s = \bigcup_{e \in \mathcal{E}} \mathcal{F}_{e,s}$.

Importantly, this intervention is highly sparse: instead of updating model weights, we learn only one scalar multiplier for each selected causal sparse feature, resulting in an average of 251 trainable parameters per model. Given the original activations F_s, the steered activations are defined as

$$F'_s = S \odot F_s,$$

where $\odot$ denotes element-wise scaling. We optimize S using cross-entropy loss for emotion recognition:

$$L = L_{\text{CE}}(y, \hat{y}).$$

The training and test sets are drawn from the combined dataset described in the experimental setup. To evaluate whether the learned steering generalizes beyond the training distribution, we further evaluate the steered models on two additional emotion recognition datasets with six basic emotion labels and relatively balanced label distributions: EmoEvent and XED (Plaza del Arco et al., 2020; Öhman et al., 2020). We also compare against few-shot prompting as an inference-time baseline. More training details are provided in Appendix A.3.

Results. Table 1 reports per-emotion F1 scores, generalization performance on two external datasets, and language modeling perplexity. In the zero-shot setting, models perform particularly poorly on *disgust* and *surprise*: averaged across models, F1 is 0.41 for *disgust* and 0.32 for *surprise*, compared with 0.73–0.82 for fear, joy, and sadness. This is consistent with our earlier analysis that *disgust* is more distributed and weakly represented, while *surprise* features are often co-activated with other emotions, especially *joy*. After steering, average F1 improves from 0.41 to 0.68 for *disgust* (65.9% relative increase) and from 0.32 to 0.46 for *surprise* (43.8%), while macro-F1 increases from 0.60 to 0.70 on average across models (16.7%). These improvements also transfer to external datasets, with average macro-F1 increasing from 0.41 to 0.46 on EmoEvent (12.2%) and from 0.44 to 0.49 on XED (11.4%). These improvements are significant under paired Wilcoxon signed-rank tests ($p < 0.05$); see Appendix B. Perplexity remains largely unchanged. Few-shot prompting also improves main-dataset performance, but has smaller gains on weaker emotions and transfers less consistently.

Figure 3 (C) shows the aggregated distributions of learned multipliers for emotion-related features across all models. We find that the learned multipliers are generally centered around 1.0, indicating that the optimization preserves most feature activations close to their original scale rather than applying extreme steering. However, the distributions differ across emotions: *disgust* shows a noticeably higher median multiplier, while *surprise* shows a slightly higher median multiplier, suggesting that features associated with these emotions often require stronger amplification. Together, these results show that sparse emotion features provide an effective intervention target for improving emotion recognition.

7 Conclusion

Large language models are increasingly used in emotionally sensitive human-AI interactions, where reliable emotion recognition is important for safety. In this work, we study emotion-specific biases in LLMs through causal sparse emotion fea-

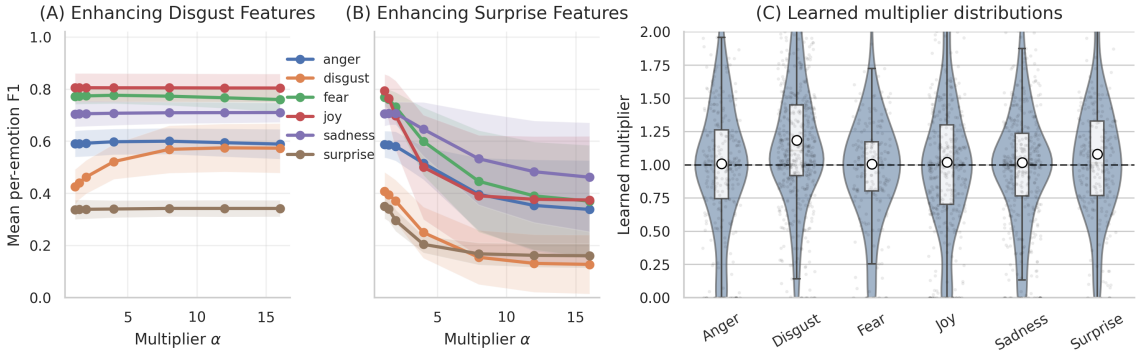

Figure 3: Targeted and global sparse feature steering results. (A, B) Aggregate per-emotion F1 across models under increasing feature multiplier α for targeted *disgust* and *surprise* features. Shaded regions indicate standard deviation across models. (C) Distribution of learned feature multipliers from global sparse feature steering across emotions.

Model	Setting	Main Dataset (F1)							EmoEvent	XED	PPL
		Anger	Disgust	Fear	Joy	Sadness	Surprise	**Macro**			
Gemma-2-2B	Zero-shot	0.54	0.24	0.71	0.79	0.73	0.29	0.55	0.39	0.36	17.54
	Few-shot	**0.64**	0.61	**0.79**	**0.82**	0.75	0.34	0.66	0.37	0.42	–
	Steered	0.63	**0.68**	0.78	0.81	**0.76**	**0.45**	**0.69**	**0.49**	**0.49**	17.62
Gemma-2-9B	Zero-shot	0.62	0.44	0.77	0.84	0.74	0.32	0.62	0.40	0.44	13.27
	Few-shot	0.66	0.58	0.80	0.79	**0.78**	0.44	0.67	0.44	0.42	–
	Steered	**0.67**	**0.70**	**0.81**	**0.86**	0.77	**0.58**	**0.73**	**0.46**	**0.52**	13.39
Gemma-3-12B-IT	Zero-shot	0.53	0.49	0.77	0.86	0.68	0.34	0.61	0.49	0.47	18.38
	Few-shot	**0.68**	0.58	**0.82**	**0.88**	**0.79**	0.23	0.66	0.50	**0.52**	–
	Steered	0.66	**0.68**	0.79	0.83	0.74	**0.48**	**0.70**	**0.51**	0.46	18.31
Gemma-3-12B-PT	Zero-shot	0.65	0.47	0.78	0.84	0.79	0.37	0.65	0.43	0.43	11.19
	Few-shot	**0.65**	0.59	0.73	**0.85**	**0.79**	0.36	0.66	0.39	0.39	–
	Steered	**0.65**	**0.72**	**0.79**	**0.85**	0.78	**0.46**	**0.71**	**0.51**	**0.46**	11.15
Gemma-3-4B-IT	Zero-shot	0.54	0.42	0.74	0.83	0.65	0.27	0.58	0.46	0.45	34.26
	Few-shot	0.63	0.45	**0.79**	**0.84**	0.71	0.35	0.63	0.43	0.44	–
	Steered	**0.66**	**0.67**	0.79	0.82	**0.77**	**0.39**	**0.68**	**0.48**	**0.50**	32.90
Gemma-3-4B-PT	Zero-shot	0.58	0.36	0.74	0.71	0.68	0.39	0.58	0.31	0.40	14.77
	Few-shot	**0.66**	0.53	**0.80**	0.78	**0.78**	0.42	0.66	**0.37**	0.41	–
	Steered	0.64	**0.67**	0.77	**0.81**	0.76	**0.46**	**0.69**	**0.37**	**0.46**	14.79
Llama-3.1-8B	Zero-shot	0.65	0.49	0.75	**0.84**	**0.77**	0.29	0.63	0.39	0.47	11.51
	Few-shot	**0.66**	0.52	**0.79**	0.75	**0.77**	0.39	0.65	0.35	0.41	–
	Steered	**0.66**	**0.66**	0.77	0.83	**0.77**	**0.44**	**0.69**	**0.42**	**0.51**	11.53

Table 1: Emotion recognition performance across zero-shot, few-shot, and steered settings. For each model, the best F1 score among available settings is shown in **bold**. EmoEvent and XED evaluate generalization to external datasets. Perplexity (PPL) measures language modeling performance to assess preservation of general language ability.

tures identified with SAEs. We show that emotions differ substantially in their sparse causal organization. Some emotions, such as *fear* and *surprise*, are supported by more concentrated causal feature sets, while *disgust* is represented by weaker and more distributed features. We further show that emotion features are not fully separated: *disgust* overlaps and covaries strongly with *anger*, while *surprise* is closely associated with *joy*. Through ablation and activation patching, we find that these features become more behaviorally decisive in later layers, but their necessity and sufficiency vary substantially across emotions. Restoration experiments further suggest that sparse emotion features participate in a sequential flow of emotion-relevant information across layers. Together, these results provide a mechanistic explanation for why models struggle more with certain emotions and why specific confusion patterns emerge.

We further examine whether causal sparse emotion features can serve as intervention targets. Our steering experiments show that feature-level calibration can strengthen weakly represented emotions, improve emotion recognition across models, transfer to external datasets, and largely preserve language modeling ability.

Overall, our findings show that emotion recognition failures in LLMs are not merely behavioral errors, but reflect differences in the internal sparse causal organization of emotion representations. By identifying and intervening on causal sparse emotion features, this work provides an interpretable path for diagnosing and improving affective capabilities in LLMs. Future work can investigate why certain emotion feature sets co-activate from input cues and why some emotion features become weaker during training.

Limitations

Due to the limited availability of publicly released full-layer SAEs from GemmaScope, GemmaScope 2, and LlamaScope, we restrict our analysis to seven models. Although these models span different scales, model families, and instruction-tuning settings, most are from the Gemma family. Other available SAE releases are less suitable for our study: some are trained on models that are too small to reliably perform emotion inference, such as GPT-2 and Pythia-70M (Karvonen et al., 2024; Gao et al., 2025), while others do not provide SAEs for all layers, such as GPT-OSS-20B SAE (Goodfire, 2025).

In addition, although we observe similar trends between SAEs with widths of 16k and 32k, we do not explore substantially wider SAEs (e.g., 161k features) in order to maintain consistency across models. Wider SAEs may further reduce polysemanticity and capture more fine-grained concepts, which could lead to a more detailed characterization of emotion representations.

Ethical Considerations

Our study provides evidence of how large language models encode and process emotion-related information and identifies features that influence emotion predictions. However, these findings do not imply that LLMs possess human-like emotional reasoning abilities. Rather, the mechanisms identified in our analysis reflect statistical patterns in model representations. Emotional intelligence (EI) in LLMs remains an open research problem, and there is currently no widely agreed-upon definition of what constitutes safe or reliable emotional behavior in such systems.

While our method improves emotion recognition performance, improved performance on benchmark datasets does not guarantee safety or reliability in real-world applications. Emotion recognition in LLMs reflects learned patterns from data, which may be incomplete, biased, or culturally dependent. We observe uneven performance across emotions (e.g., weaker representation of *disgust*), which raises concerns about consistency and fairness. Misclassification of user emotions in real-world interactions, particularly in sensitive domains such as mental health support, may lead to inappropriate or harmful responses, potentially affecting vulnerable users (Ball and Breese, 2001). Moreover, interpretability methods may create a false sense of transparency and lead users to overestimate the model's understanding or reliability, as such features are approximations rather than complete causal explanations.

Our steering approach introduces additional considerations. Although it enables targeted modification of model behavior, it also presents dual-use risks, as similar techniques could be used to manipulate emotional interpretations or influence user behavior in undesirable ways. For these reasons, we emphasize that our work is intended to advance scientific understanding of model mechanisms rather than enable direct deployment in high-stakes or human-facing applications. Any real-world use should be approached with caution and accompanied by appropriate safeguards, evaluation, and human oversight.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Gene Ball and Jack Breese. 2001. *Emotion and personality in a conversational agent*, page 189–219. MIT Press, Cambridge, MA, USA.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, and 5 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024. EmotionQueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

Joel Davitz. 2013. *The Language of Emotion*. Academic Press.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. LLaMAs have feelings too: Unveiling sentiment and emotion representations in LLaMA models through probing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6124–6142, Vienna, Austria. Association for Computational Linguistics.

Paul Ekman. 1999. Basic emotions. In Tim Dalgleish and Mick Power, editors, *Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Joanne Frattaroli. 2006. Experimental disclosure and its moderators: A meta-analysis. *Psychological Bulletin*, 132(6):823–865.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Goodfire. 2025. Goodfire/Hackathon-gpt-oss-20b-SAE-l15 sae model. `https://huggingface.co/Goodfire/Hackathon-gpt-oss-20b-SAE-l15`. MIT License.

Aaron Grattafiori, Abhimanyu Dubey, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Lynn Greschner and Roman Klinger. 2025. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 628–646, Albuquerque, USA. Association for Computational Linguistics.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *Preprint*, arXiv:2410.20526.

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. 2025a. Sae-ssv: Supervised steering in sparse representation spaces for reliable control of language models. *Preprint*, arXiv:2505.16188.

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. 2025b. SAE-SSV: Supervised steering in sparse representation spaces for reliable control of language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2207–2236, Suzhou, China. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

Adam Karvonen, Can Rager, Samuel Marks, and Neel Nanda. 2024. Evaluating sparse autoencoders on targeted concept erasure tasks. *Preprint*, arXiv:2411.18895.

Anna Koufakou, Elijah Nieves, and John Peller. 2024. Towards a new benchmark for emotion detection in NLP: A unifying framework of recent corpora. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 196–206, Miami, Florida, USA. Association for Computational Linguistics.

Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. 2025. Do large language models have "emotion neurons"? investigating the existence and role. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15617–15639, Vienna, Austria. Association for Computational Linguistics.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *Preprint*, arXiv:2408.05147.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Aishwarya Maheswaran and Maunendra Sankar Desarkar. 2026. A unified view on emotion representation in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3589–3610, Rabat, Morocco. Association for Computational Linguistics.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025a. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025b. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *Preprint*, arXiv:2403.19647.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. 2025. Gemma scope 2: Technical report. Google Technical Report.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

James W. Pennebaker and Sandra K. Beall. 1986. Confronting a traumatic event: Toward an understanding of inhibition and disease. *Journal of Abnormal Psychology*, 95(3):274–281.

Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. 2024. Improving dictionary learning with gated sparse autoencoders. *Preprint*, arXiv:2404.16014.

Douglas Reynolds. 2009. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Saeedi, Alireza Kheirandish, Sirwe Saeedi, Hossein Sahour, Aliakbar Panahi, and Iman Naeeni. 2025. GT-NLP at SemEval-2025 task 11: EmoRationale, evidence-based emotion detection via retrieval-augmented generation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 640–650, Vienna, Austria. Association for Computational Linguistics.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Bangzhao Shu, Isha Joshi, Melissa Karnaze, Anh C. Pham, Ishita Kakkar, Sindhu Kothe, Arpine Hovasapian, and Mai ElSherief. 2025. Fluent but unfeeling: The emotional blind spots of language models. *Preprint*, arXiv:2509.09593.

Samuel Soo, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Ming YAN. 2025. Interpretable steering of large language models with feature guided activation additions. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120, Vienna, Austria. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Aidan Templeton, Tim Conerly, Jacob Marcus, John Lindsey, and 1 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`. Transformer Circuits Thread, Accessed: 2025-02-01.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.

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

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. 2025. Do llms "feel"? emotion circuits discovery and control. *Preprint*, arXiv:2510.11328.

Nutchanon Yongsatianchot, Parisa Ghanad Torshizi, and Stacy Marsella. 2023. Investigating large language models' perception of emotion using appraisal theory. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8.

Hengyuan Zhang, Zhihao Zhang, Mingyang Wang, and 1 others. 2026. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. *Preprint*, arXiv:2601.14004.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.

A Appendix

A.1 Sparse Autoencoders

Neural networks represent information in high-dimensional activation spaces, but the number of concepts a model must encode far exceeds the number of available neurons. This leads to *superposition*, where multiple concepts are simultaneously encoded as overlapping linear combinations of neuron activations (Elhage et al., 2022). As a result, individual neurons are typically *polysemantic*, activating for multiple unrelated concepts, making the direct interpretation of neuron activations unreliable (Elhage et al., 2022; Arora et al., 2018).

Sparse autoencoders (SAEs) are trained to decompose these polysemantic activations into a larger set of approximately monosemantic features (Cunningham et al., 2023; Bricken et al., 2023). Formally, an SAE consists of an encoder f_{enc} and a decoder f_{dec}. Given a hidden state $h^{(\ell)} \in \mathcal{R}^{d_{\text{model}}}$ from layer ℓ, the encoder maps it to a sparse latent representation:

$$z^{(\ell)} = \text{ReLU}(W_{\text{enc}}\, h^{(\ell)} + b_{\text{enc}}) \tag{1}$$

where $W_{\text{enc}} \in \mathcal{R}^{d_{\text{SAE}} \times d_{\text{model}}}$, $d_{\text{SAE}} \gg d_{\text{model}}$, and the ReLU activation enforces sparsity by zeroing out weakly activated features. The decoder reconstructs the original activation as:

$$\hat{h}^{(\ell)} = W_{\text{dec}}\, z^{(\ell)} + b_{\text{dec}} \tag{2}$$

The SAE is trained to minimize a combination of reconstruction error and an ℓ_1 sparsity penalty:

$$\mathcal{L}_{\text{SAE}} = \|h^{(\ell)} - \hat{h}^{(\ell)}\|_2^2 + \beta \|z^{(\ell)}\|_1 \tag{3}$$

where β controls the sparsity–reconstruction tradeoff. The ℓ_1 penalty encourages each input to activate only a small number of latent dimensions, enabling improved interpretability and concept localization (Bricken et al., 2023). Each dimension of $z^{(\ell)}$ corresponds to a learned feature direction in the residual stream, and the columns of W_{dec} define the corresponding feature vectors. Since $d_{\text{SAE}} \gg d_{\text{model}}$, the SAE can represent far more features than there are neurons, providing a richer and more disentangled foundation for interpreting a model's internal representations (Cunningham et al., 2023; Templeton et al., 2024).

Some SAE variants replace the standard ReLU activation with a learned threshold, known as a JumpReLU, which activates a feature only when its pre-activation exceeds a learnable per-feature threshold θ:

$$z^{(\ell)} = \text{JumpReLU}_{\theta}(W_{\text{enc}}\, h^{(\ell)} + b_{\text{enc}}) \tag{4}$$

This formulation allows the model to more precisely control the sparsity level of each feature independently, yielding more faithful reconstructions under the same sparsity constraints (Rajamanoharan et al., 2024). The SAEs provided by GemmaScope and LlamaScope, which we use in this work, employ this JumpReLU architecture (Lieberum et al., 2024; He et al., 2024).

A.2 Data Preprocessing

Among these datasets, EXPRESS differs from the others in several important ways. It contains instances with a wide range of text lengths and allows one or multiple emotion labels. In addition, the emotion label is masked from the original text, meaning that the annotated emotion may correspond only to a localized emotional expression rather than the overall emotional content. To ensure consistency across datasets, we apply the following filtering steps to EXPRESS: (1) excluding instances exceeding a fixed length threshold of 50 words, (2) retaining only instances annotated with a single emotion, and (3) using GPT-4o to filter out cases where the labeled emotion reflects only a local emotional span rather than the global emotional tone of the text.

A.3 Model Training Details

In the main steering experiment, we use an 80–20 train–test split on the main dataset and sample 400 training examples per emotion from the training set. We evaluate language modeling preservation using perplexity on a subset of the WikiText-103 test set (Merity et al., 2017). For Gemma-3 models, we use the small-sparsity SAE setting for steering.

B Statistical Significance Tests

We further conducted paired statistical significance tests across the seven evaluated models to assess whether sparse feature steering consistently improves emotion recognition performance compared to the zero-shot baseline. Since the comparisons involve paired measurements across the same set of models and the number of models is relatively small, we use one-sided Wilcoxon signed-rank tests. We additionally report paired effect sizes (d_z) for the mean paired difference.

Table A.2 summarizes the results. Compared to the original zero-shot baseline, sparse feature steering significantly improves macro-F1 on the main dataset ((0.607 to 0.691), (+13.9%), (p=0.0078)), as well as on the EmoEvent ((+12.8%), (p=0.0078)) and XED ((+13.2%), (p=0.0156)) transfer datasets. Improvements are particularly strong for the weakest emotions identified in our earlier analyses. On the main dataset, *disgust* F1 improves from 0.424 to 0.667 ((+57.5%), (p=0.0078)), while *surprise* improves from 0.330 to 0.457 ((+38.7%), (p=0.0078)).

Few-shot prompting also improves macro-F1 on the main dataset ((+8.0%), (p=0.0078)) and improves *disgust* recognition ((+29.8%), (p=0.0078)). However, its improvement on *surprise* is smaller and not statistically significant ((p=0.1484)), and its gains transfer less consistently to external datasets. In contrast, sparse feature steering consistently outperforms few-shot prompting across all evaluated datasets, including the main dataset ((+5.5%), (p=0.0078)), EmoEvent ((+13.8%), (p=0.0156)), and XED ((+13.8%), (p=0.0391)).

Model	Total	Number of Features						Average Logit Effect per Feature					
		Anger	Disgust	Fear	Joy	Sadness	Surprise	Anger	Disgust	Fear	Joy	Sadness	Surprise
Gemma-2-2B	121	13	38	10	28	33	18	0.564	0.155	1.354	0.346	0.236	0.873
Gemma-2-9B	264	35	86	26	57	67	35	0.287	0.109	0.704	0.220	0.154	0.718
Gemma-3-4B-PT-big	313	62	114	30	55	70	44	0.226	0.144	0.638	0.260	0.220	0.471
Gemma-3-4B-PT-small	134	25	36	14	30	38	19	0.432	0.256	1.274	0.551	0.330	1.154
Gemma-3-4B-IT-big	322	47	126	39	68	58	48	0.695	0.225	0.718	0.548	0.442	0.753
Gemma-3-4B-IT-small	186	30	64	33	33	55	22	1.099	0.430	1.002	1.166	0.442	1.875
Gemma-3-12B-PT-big	437	71	109	58	109	106	52	0.285	0.159	0.422	0.204	0.161	0.663
Gemma-3-12B-PT-small	170	29	37	22	40	31	19	0.499	0.284	1.097	0.518	0.450	1.684
Gemma-3-12B-IT-big	518	75	112	103	135	98	74	0.340	0.291	0.202	0.219	0.238	0.459
Gemma-3-12B-IT-small	161	26	34	20	34	26	23	1.104	0.664	2.073	0.901	1.026	2.001
Llama-3.1-8B	138	24	30	15	28	37	21	0.662	0.326	1.954	0.638	0.453	1.433
Average	251.3	39.7	71.5	33.6	56.1	56.3	34.1	0.563	0.277	1.040	0.507	0.378	1.098

Table A.1: Summary of selected causal sparse emotion features across models and SAE settings. For each model, we report the total number of unique selected features, the number of features associated with each emotion according to `affected_emotions`, and the average logit effect per feature for each affected emotion. The final row reports the average across model-SAE settings.

Comparison	Metric	Mean A	Mean B	Rel. Change	p	Holm-p	d_z
Steering vs. Baseline	Main Macro-F1	0.607	0.691	+13.9%	0.0078	0.0312	3.12
	EmoEvent Macro-F1	0.409	0.461	+12.8%	0.0078	0.0312	2.74
	XED Macro-F1	0.442	0.500	+13.2%	0.0156	0.0469	2.11
	Disgust F1	0.424	0.667	+57.5%	0.0078	0.0312	3.58
	Surprise F1	0.330	0.457	+38.7%	0.0078	0.0312	2.83
Few-shot vs. Baseline	Main Macro-F1	0.607	0.656	+8.0%	0.0078	0.0391	2.16
	EmoEvent Macro-F1	0.409	0.429	+4.9%	0.1094	0.2188	0.91
	XED Macro-F1	0.442	0.461	+4.3%	0.0781	0.1953	1.02
	Disgust F1	0.424	0.550	+29.8%	0.0078	0.0391	2.41
	Surprise F1	0.330	0.381	+15.5%	0.1484	0.2969	0.74
Steering vs. Few-shot	Main Macro-F1	0.656	0.691	+5.5%	0.0078	0.0391	1.97
	EmoEvent Macro-F1	0.429	0.488	+13.8%	0.0156	0.0469	2.09
	XED Macro-F1	0.461	0.525	+13.8%	0.0391	0.0781	1.36
	Disgust F1	0.550	0.667	+21.3%	0.0156	0.0469	2.01
	Surprise F1	0.381	0.457	+19.9%	0.0156	0.0469	1.88

Table A.2: Paired statistical significance tests across seven models comparing sparse feature steering, few-shot prompting, and the zero-shot baseline. We report one-sided Wilcoxon signed-rank tests, Holm-corrected p-values, and paired effect sizes (d_z). Sparse feature steering yields significant and transferable improvements, particularly for weakly represented emotions such as *disgust* and *surprise*.

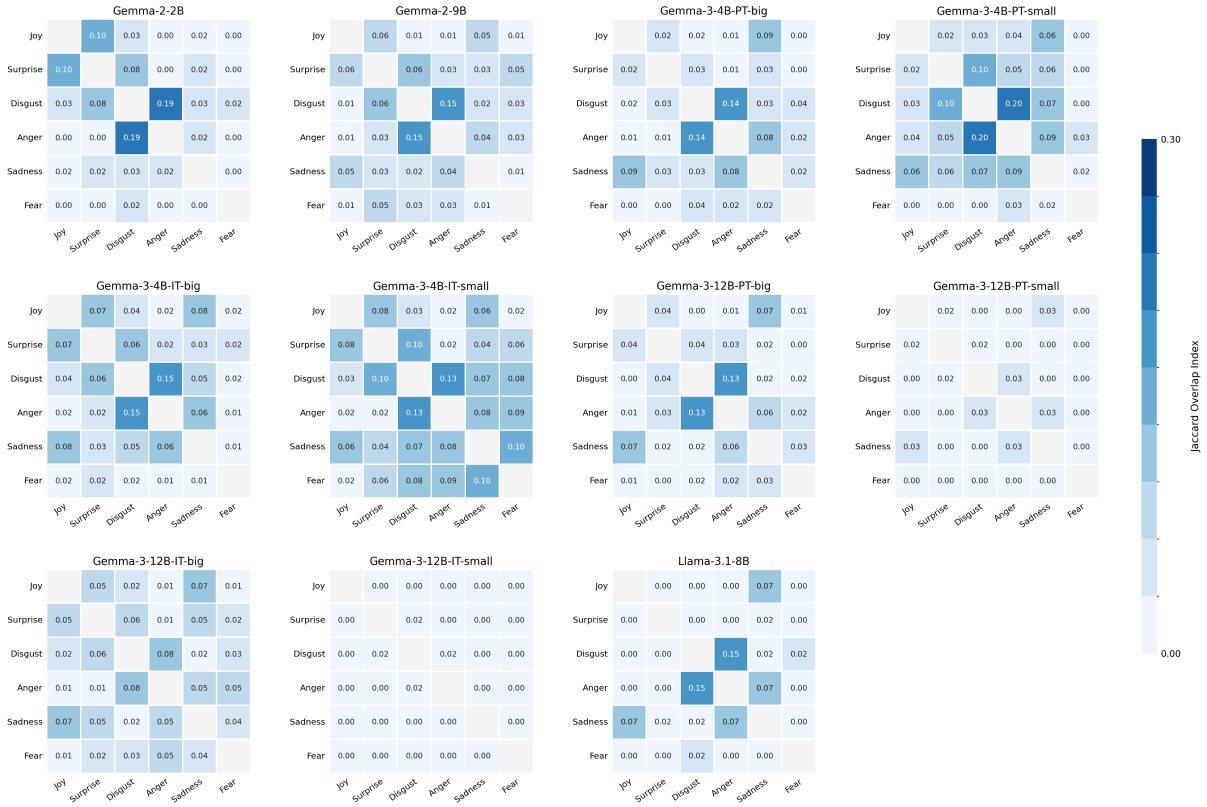

Figure A.1: Cross-emotion overlap of causal sparse emotion features across all models and SAE settings. Each subplot shows the Jaccard overlap between selected feature sets for each pair of emotions. Diagonal entries are masked, and darker cells indicate stronger feature sharing.

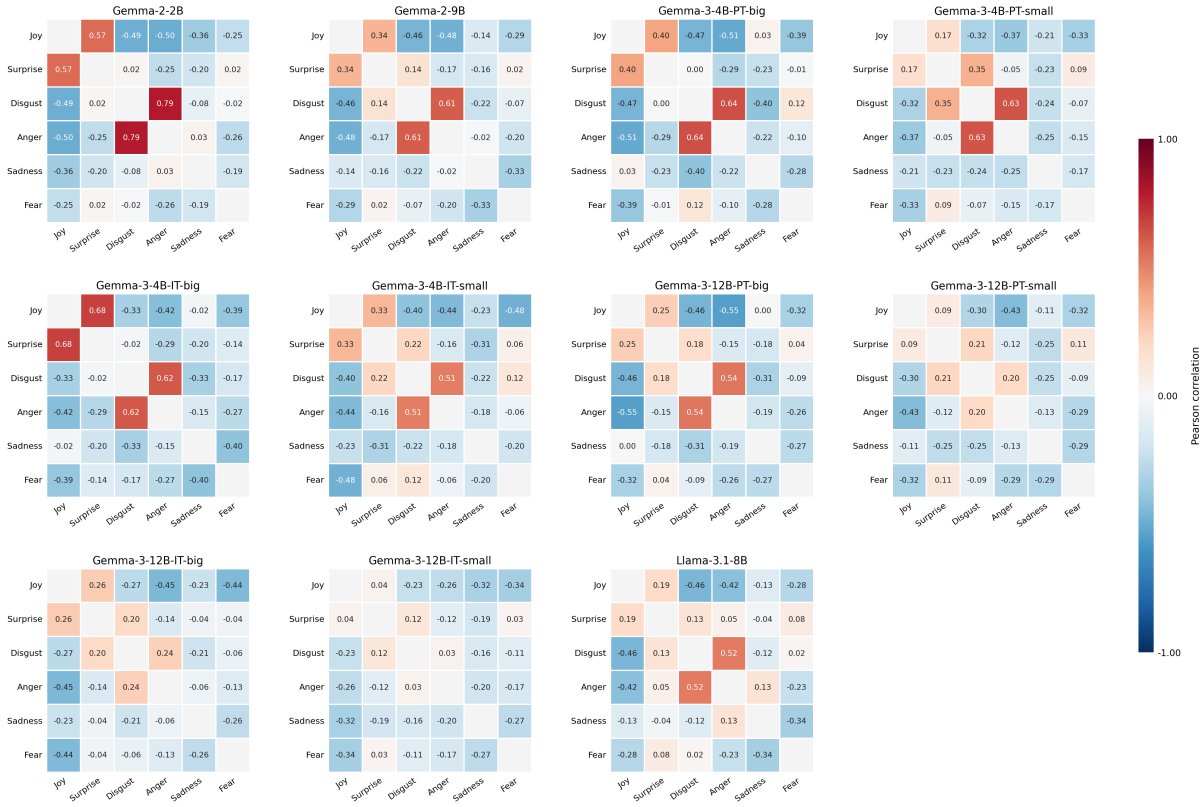

Figure A.2: Cross-emotion activation-strength correlations across all models and SAE settings. Each subplot shows Pearson correlations between the mean activation strengths of causal sparse feature sets for different emotions. Diagonal entries are masked. Redder cells indicate stronger positive correlations, while bluer cells indicate stronger negative correlations.

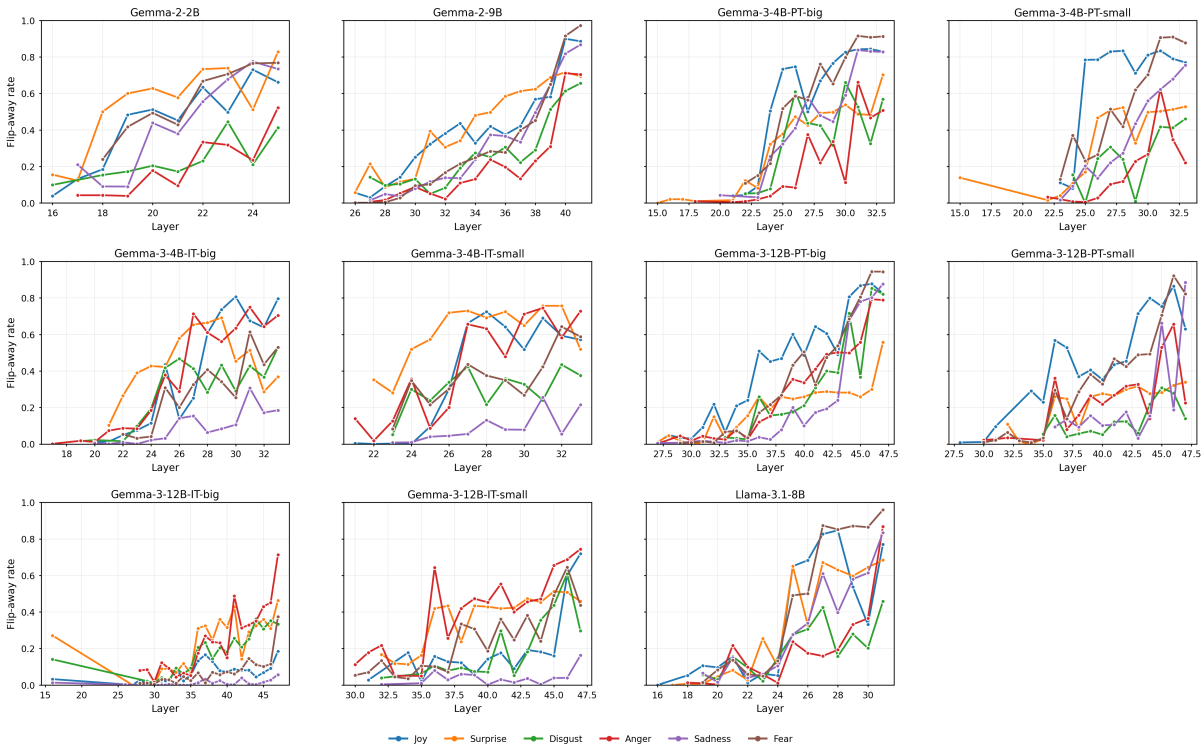

Figure A.3: Layer-wise ablation effects across all models and SAE settings. Each subplot shows the prediction flip-away rate after ablating selected causal sparse emotion features for each emotion across layers. Higher values indicate stronger behavioral necessity of the selected features at that layer.

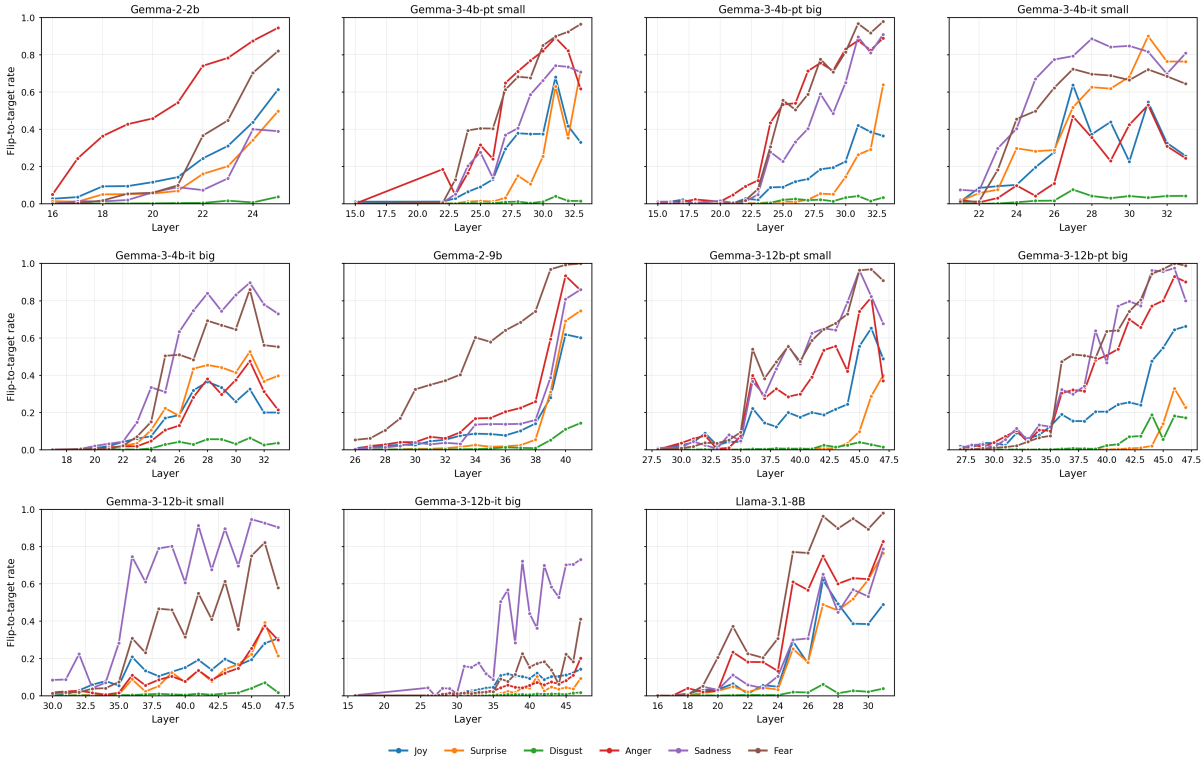

Figure A.4: Layer-wise activation patching effects across all models and SAE settings. Each subplot shows the flip-to-target rate after patching selected causal sparse emotion features for each target emotion across layers. Higher values indicate stronger sufficiency of the selected features for inducing the target emotion prediction.

Model	#	Cluster 0	Cluster 1
Gemma-2-2B (joy)	1	Winning a race after many weeks of training.	Finally getting a first 5-star reward before the set ends.
	2	Realizing one's desires and aims.	Obtaining money needed for a planned purchase.
	3	Going to the aquarium with a boyfriend and enjoying the animals together.	Being accepted as supervisor for a student teacher.
	4	Finally feeling well again after a long battle with depression.	Opening the TE score envelope and getting into Physiotherapy.
	5	Singing with a band in front of a large school audience.	Achieving a first 6 at university after lower previous scores.
	6	Finishing a long hike with a sense of accomplishment.	Receiving a positive examination result for a Psychology degree.
	7	Winning in a video game.	Being accepted into a first-choice university.
	8	Favorite national team winning an international sports event.	Being accepted to Villanova University.
Llama-3.1-8B (joy)	1	Recovering from depression and finally feeling well again.	Receiving a university acceptance letter after waiting nearly a year for exam results.
	2	Reuniting with a long-distance partner after COVID border restrictions ended.	Winning a close tennis match after coming from behind.
	3	Jumping into a pond during a summer camp night swim with a friend.	Passing school leaving exams and being selected for college.
	4	Beating Virtual Clash for the first time after two years.	Watching one's team win the cup.
	5	Singing and laughing during a sunset drive in South Africa with loved ones.	Achieving a first 6 at university after a long series of lower scores.
	6	Winning an auction for a long-desired item.	Being accepted into a first-choice university.
	7	Recovering from bulimia and depression and finally feeling like dancing again.	Watching a favorite football team win the cup final.
	8	Traveling to visit one's grandmother and enjoying the seaside with family.	Watching one's team win.
Gemma-2-2B (disgust)	1	Seeing someone pick their nose and eat it.	Detesting the cruel behavior of a fellow soldier.
	2	Watching people eat with bad manners such as smacking or belching.	Feeling revulsion toward someone behaving like a scoundrel.
	3	Accidentally eating a spoiled pistachio nut.	Being approached and insulted by a drunk man in the underground station.
	4	Eating unpleasant food after a cold and exhausting mountain climb.	Reading a book with obscene and tasteless content.
	5	Seeing someone eat live worms.	Feeling disgust toward brutal violence and physical aggression.
	6	Watching someone noisily gulp food in a cafeteria.	Speaking to a man avoiding responsibility for his family's suffering.
	7	Eating a spoiled brownie from the fridge.	Seeing someone spit on the floor in the street.
	8	Seeing someone use meat after dropping it on the floor.	Seeing someone spit inside a shopping centre.

Table A.3: Top activating examples for joy and disgust sparse feature clusters across models. Joy clusters consistently separate experiential/social joy from achievement- and reward-based joy, while disgust clusters separate physical contamination from social or norm-based aversion.

Model	Emotion	X→M act. drop	X→Y loss	M→Y loss	Restore Δlogit	Recovery
google/gemma-2-2b	anger	21.08	0.62	0.81	0.55	0.85
	disgust	5.12	0.42	0.61	0.29	0.68
	fear	25.60	1.28	1.68	1.10	0.87
	joy	10.71	0.71	0.90	0.62	0.80
	sadness	10.06	0.70	0.91	0.60	0.79
	surprise	14.84	1.11	1.39	0.87	0.84
google/gemma-2-9b	anger	13.25	0.45	0.57	0.40	0.87
	disgust	6.12	0.57	0.66	0.45	0.72
	fear	26.10	0.88	1.09	0.82	0.89
	joy	11.12	0.75	0.92	0.64	0.84
	sadness	8.66	0.71	0.93	0.64	0.81
	surprise	11.61	1.04	1.28	1.02	0.87
google/gemma-3-12b-it	anger	2156.79	1.40	1.61	1.24	0.91
	disgust	1194.09	1.18	1.37	0.85	0.75
	fear	2922.09	1.74	1.95	1.50	0.95
	joy	1779.15	1.74	2.14	1.39	0.81
	sadness	2805.30	1.53	1.89	1.18	0.88
	surprise	2102.53	1.98	2.10	1.56	0.82
google/gemma-3-12b-pt	anger	1676.88	0.61	0.70	0.46	0.84
	disgust	833.75	0.47	0.52	0.30	0.72
	fear	2643.63	0.99	1.15	0.98	1.05
	joy	1734.65	1.07	1.19	0.90	0.91
	sadness	1747.14	0.79	1.08	0.61	0.78
	surprise	1958.14	1.24	1.36	1.06	0.83
google/gemma-3-4b-it	anger	590.45	2.06	2.38	1.62	0.68
	disgust	258.13	2.10	2.44	1.31	0.51
	fear	678.87	2.36	2.77	1.74	0.73
	joy	526.19	2.82	3.38	2.04	0.73
	sadness	439.46	1.93	2.30	1.31	0.59
	surprise	597.35	2.70	3.00	1.94	0.73
google/gemma-3-4b-pt	anger	330.26	0.69	0.81	0.53	0.73
	disgust	238.28	1.14	1.34	0.73	0.61
	fear	848.02	1.39	1.64	1.22	0.90
	joy	403.29	1.03	1.20	0.70	0.73
	sadness	390.71	1.04	1.27	0.80	0.73
	surprise	276.03	0.85	0.97	0.63	0.59
meta-llama/Llama-3.1-8B	anger	1.91	0.72	0.91	0.60	0.72
	disgust	1.05	0.52	0.63	0.32	0.54
	fear	3.35	1.50	1.82	1.47	0.99
	joy	1.74	0.93	1.09	0.74	0.79
	sadness	1.53	0.93	1.16	0.77	0.75
	surprise	1.06	0.74	0.90	0.68	0.73
ALL_MODELS_AVG	anger	684.37	0.93	1.11	0.77	0.80
	disgust	362.36	0.91	1.08	0.61	0.65
	fear	1021.09	1.45	1.73	1.26	0.91
	joy	638.12	1.29	1.54	1.01	0.80
	sadness	771.84	1.09	1.36	0.85	0.76
	surprise	708.79	1.38	1.57	1.11	0.77

Table A.4: Mean adjacent-layer group path-patching results per emotion across models.