From Syntax to Emotion: A Mechanistic Analysis of Emotion Inference in LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used in emotionally sensitive human–AI applications, yet little is known about how emotion recognition is internally represented. In this work, we investigate the internal mechanisms of emotion recognition in LLMs using sparse autoencoders (SAEs). By analyzing sparse feature activations across layers, we identify a consistent three-phase information flow, in which emotion-related features emerge only in the final phase. We further show that emotion representations comprise both shared features across emotions and emotion-specific features. Using phase-stratified causal tracing, we identify a small set of features that strongly influence emotion predictions, and show that both their number and causal impact vary across emotions; in particular, *Disgust* is more weakly and diffusely represented than other emotions. Finally, we propose an interpretable and data-efficient causal feature steering method that significantly improves emotion recognition performance across multiple models while largely preserving language modeling ability, and demonstrate that these improvements generalize across multiple emotion recognition datasets. Overall, our findings provide a systematic analysis of the internal mechanisms underlying emotion recognition in LLMs and introduce an efficient, interpretable, and controllable approach for improving model 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 variable performance across settings and tasks (Yongsatianchot et al., 2023; Liu et al., 2021; Shu et al., 2025). However, most existing studies assess EI primarily through model outputs, without examining the internal mechanisms by which emotional information is represented and processed (Tak and Gratch, 2023; Sabour et al., 2024). This lack of mechanistic understanding poses risks to safety and reliability, and limits our ability to systematically control model behavior or mitigate potential harms, such as incorrect emotion recognition or inappropriate emotional responses (Tak et al., 2025).

Mechanistic interpretability offers a foundational view of information processing and representation in LLMs, revealing their internal mechanisms and supporting targeted control of model reasoning (Rai et al., 2025; Feng et al., 2025). Only a small number of recent studies have begun to explore the mechanistic interpretability of emotion inference in LLMs. Prior work using linear probing has shown that information in LLMs evolves across layers, with early layers capturing surface linguistic patterns and later layers encoding more abstract semantic or behavioral signals (Skean et al., 2025; Ju et al., 2025, 2024). Focusing specifically on emotion inference, Tak et al. (2025) localize emotion-related signals to specific layers or regions of the model. In addition, Lee et al. (2025) and Wang et al. (2025) identify emotion-related neurons or attention heads and show that intervening on these units can influence model behavior. However, these approaches remain limited in several ways: they are either hypothesis-driven and non-observational, focus primarily on localizing emotion-related signals, or rely on polysemantic units that are difficult to interpret and scale for steering. Based on these gaps, we investigate the following research questions. **RQ1:** How is emotional information encoded, and how does it evolve across layers in LLMs during emotion inference? **RQ2:** Can we identify the internal representations that drive and causally influence the model's predictions? **RQ3:**

Can emotion inference in LLMs be systematically and reliably improved through targeted interventions?

Recent advances in mechanistic interpretability, particularly sparse autoencoders (SAEs), provide a feature-level view of model representations that enables more interpretable and fine-grained analysis of internal computations. Compared to prior probing- and neuron-level approaches, SAE-based analysis offers a more direct and scalable way to identify and intervene on behaviorally relevant features (Cunningham et al., 2023; Marks et al., 2025a).

To answer RQ1–RQ3, we systematically analyze the activation patterns and causal effects of sparse features during emotion inference using SAEs. Our contributions are as follows: (1) We identify a consistent three-phase information flow in LLMs, progressing from syntactic features to semantic concepts and finally to emotion-related features; (2) we show that emotions are represented through both shared and emotion-specific features, with uneven representation across emotions; (3) we provide mechanistic insights into emotion recognition failures, showing that *Disgust* is weakly and diffusely represented, while surprise-related features are frequently activated by other emotions; (4) we propose phase-conditioned circuit discovery to identify a small set of strongly causal features underlying emotion recognition; and (5) we introduce an interpretable and data-efficient causal feature steering method that improves emotion recognition performance while preserving language modeling ability.

2 Related Work

Mechanistic Interpretability. Mechanistic interpretability aims to understand how neural networks implement specific computations by analyzing their internal representations and causal pathways. A large body of work studies what information is encoded in model representations using probing classifiers and representation analysis (Belinkov, 2022; Alain and Bengio, 2018). Another line of research aims to identify circuits—sets of model components that jointly implement specific computations—via intervention techniques such as activation patching (Wang et al., 2023; Elhage et al., 2021; Geiger et al., 2025; Meng et al., 2023).

Recent work addresses the problem of polysemantic neurons by using SAEs to decompose model activations into large dictionaries of sparse, approximately monosemantic features (Bricken et al., 2023; Cunningham et al., 2023). These features correspond to directions in LLM latent spaces that align with human-interpretable concepts (Templeton et al., 2024). Building on this foundation, recent work proposes methods to construct interpretable sparse feature circuits for specific model behaviors (Marks et al., 2025b), and a growing body of work explores methods for controlling model behavior through targeted feature interventions based on feature activations (He et al., 2025; Soo et al., 2025; Bounhar et al., 2026). These approaches demonstrate that interpretable internal representations can be used not only to analyze models but also to reliably guide their behavior. Building on these lines of work, we leverage SAEs to study emotion recognition from a mechanistic perspective, enabling both analysis and control of emotion-related representations in LLMs.

Emotion-Related Mechanisms 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 recognition. Tigges et al. (2024) show that sentiment in language models 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. Moving beyond representation analysis, Lee et al. (2025) identify neurons that exhibit consistent activation patterns associated with specific emotions. Wang et al. (2025) further take a circuit-level perspective and construct emotion circuits that implement emotional computation through analytical decomposition and causal analysis.

While prior work provides valuable exploratory insights into the localization and controllability of emotion-related signals, it remains limited in three important ways. First, probing-based approaches primarily reveal whether emotion-related information can be predicted from the model's representations, but do not explain how this information is encoded or used in internal computations (Mueller et al., 2025). Second, these approaches are often hypothesis-driven rather than observational. Third, neuron- and attention-head–level analyses rely on highly polysemantic units whose semantic roles are difficult to interpret, and intervening on such units can have unintended or hard-to-predict ef-

fects; moreover, effective control often requires intervening on large sets of components, making these methods difficult to scale (Bricken et al., 2023; Arora et al., 2018; Elhage et al., 2022). In contrast, our work adopts an observational, approximately monosemantic approach to analyze the underlying mechanisms, enabling interpretable and scalable steering that improves emotion recognition beyond simple output-level control.

3 Experimental Setup

Models, SAEs and Prompts. We use pretrained SAEs released by Gemmascope and Llamascope, which currently provide the most comprehensive publicly available SAE models (Lieberum et al., 2024; He et al., 2024). These releases include SAEs trained on the residual streams of all transformer layers for three LLMs: Gemma-2-2B, Gemma-2-9B, and Llama-3.1-8B (Team et al., 2024; Grattafiori et al., 2024). Together, these models span multiple architectures and sizes, enabling evaluation across both model families and scales. In particular, we use `GEMMASCOPE-RES-16K` for Gemma-2-2B and Gemma-2-9B, and `LLAMASCOPE-RES-32K` for Llama-3.1-8B. Throughout this work, we focus exclusively on residual-stream SAEs. The residual stream integrates information from both attention and multi-layer perceptron (MLP) sublayers, making it 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:`

We frame it as a classification problem following prior work (Zhu et al., 2021; Tak et al., 2025) and evaluate model outputs by examining logits restricted to six emotion labels (anger, joy, sadness, fear, surprise, 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. We restrict our selection to datasets whose emotion labels are self-reported by the original authors, as such labels 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.3.

Across all four datasets, we focus on six basic emotions—anger, sadness, joy, disgust, surprise, and fear—following Ekman's theory of basic emotions (Ekman, 1999). This framework is widely adopted in NLP and affective computing due to its cross-cultural validity, relative universality, and conceptual simplicity (Jurafsky and Martin, 2026). These emotions provide a well-established taxonomy for studying emotional representations, making them well suited for an initial mechanistic analysis of emotion inference in LLMs. The final combined dataset contains 9,870 instances.

4 RQ1: Encoding and Evolution of Emotion Representations

To understand the internal representations of emotions during emotion recognition in LLMs, we extract sparse feature activations across layers and analyze their corresponding feature explanations.

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

To select highly activated features and reduce noise, we filter features based on their activation strength. For each emotion e, we compute the mean activation of each feature across all instances labeled with e, and retain the features whose mean activation ranks within the top threshold $\tau_{\text{act}} = 0.05$ (i.e., the top 5%). We denote the resulting feature set for emotion e as $\mathcal{F}_e$, and the combined set of features across all emotions as $\mathcal{F}$.

Feature Explanation and Topic Modeling. After obtaining $\mathcal{F}_e$, we analyze the natural language feature explanations associated with these features, which are provided alongside the pretrained SAEs (Lieberum et al., 2024; He et al., 2024), to better understand what they represent. Prior work has interpreted SAE features by examining the input sequences that elicit the strongest acti-

vations (Templeton et al., 2024; Rajamanoharan et al., 2024; Bricken et al., 2023; Bills et al., 2023; Cunningham et al., 2023). A common line of work combines human evaluation and language-model–generated explanations to interpret large numbers of SAE features (Lieberum et al., 2024; He et al., 2024). In these approaches, raters assess whether a feature admits a coherent semantic interpretation based on its highest-activating examples and broader activation patterns, and provide a short description when applicable. To reduce the cost of human annotation, automated evaluation schemes have also been proposed, where a language model first generates an explanation for a feature and then predicts feature activations from that explanation (Lieberum et al., 2024).

The explanations of individual features vary, but many share overlapping semantics or express similar underlying concepts. To systematically analyze the emergence of concepts during emotion recognition, we perform topic modeling over the textual explanations of features. We apply BERTopic to cluster feature explanation into semantically coherent groups. Figure 1 shows the activation of a subset of feature topics across layers in Gemma-2-2B. Syntax-related features tend to activate in earlier layers, while emotion-related features emerge in later layers. Emotion-specific features show consistent activation patterns and occasionally co-activate across multiple emotions. For example, the topics *expression of excitement* and *celebration and community gathering* are frequently activated in both *Joy* and *Surprise* instances, while *expression of surprise* is also frequently activated in other emotions.

To better illustrate these activation patterns, we further group the 82 feature topics into three main categories: (1) *syntax (syn)*, which includes surface-level textual structures such as formatting, punctuation, grammar, code structure, programming constructs, or symbolic operations; (2) *concept (con)*, which captures semantic topics, domains, roles, events, or situational concepts that are neither syntactic nor explicit emotion expressions; and (3) *emotion (emo)*, which consists of direct expressions of emotional states, affective language, or explicit emotion-related terms. We use GPT-4o-mini to classify feature topics into these three categories and map $\mathcal{F}$ into $\mathcal{F}_{syn}$, $\mathcal{F}_{con}$, and $\mathcal{F}_{emo}$. Topic modeling details and prompts are provided in Appendix A.5, and examples of topics are shown in Table A.3.

Phase Structure of Emotion Inference. Across

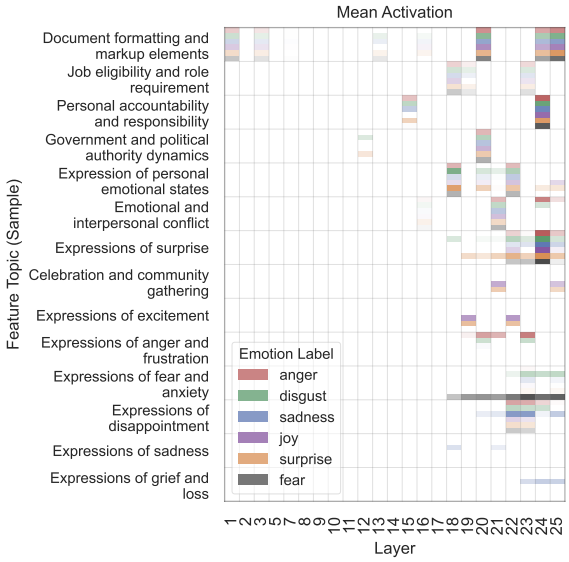

Figure 1: A subset of feature topics and their mean activation across layers for instances with different emotion labels for Gemma-2-2B. Topics are selected as representative examples from each category with clear activation patterns. Syntax-related features tend to activate in earlier layers, while emotion-related features emerge in later layers. Emotion-specific features show consistent activation patterns and occasionally co-activate across multiple emotions. Figure A.4 shows a more comprehensive activation map.

all emotions and models, we observe a consistent layer-wise pattern in the topics of $\mathcal{F}$. Specifically, we identify a consistent three-phase activation pattern: syntax-related features tend to appear earliest, followed by semantic concept features, and finally emotion-related features. This suggests that models first focus on syntactic structure, then extract semantic context, and subsequently form emotion-related representations.

We then estimate the emergence layer of each topic category based on its normalized activation across layers. The emergence layer is defined as the earliest layer where activation exceeds $\tau = 0.05$ of the category's maximum activation for at least $k = 5$ consecutive layers. These emergence points are used to approximate the phase boundaries. Figure 2 visualizes the activation of the three categories across layers and the corresponding phase structure.

Phase One. In Phase One, syntax-related features $\mathcal{F}_{syn}$ are strongly activated and dominate the feature-topic activations.

Phase Two. In Phase Two, concept-related features $\mathcal{F}_{con}$ begin to emerge, and their activations continue to increase in subsequent layers. $\mathcal{F}_{syn}$

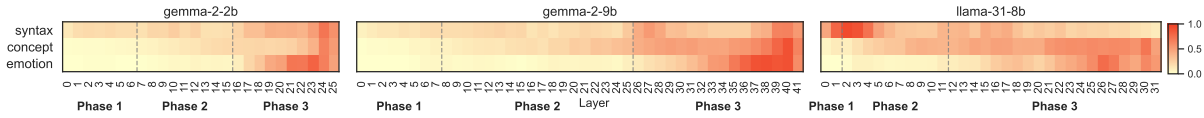

Figure 2: Average activation (smoothed) per category across layers, normalized within each category. The border of phases are marked.

remain active during this phase.

Phase Three. In Phase Three, emotion-related features $\mathcal{F}_{emo}$ begin to emerge and increase in activation. $\mathcal{F}_{syn}$ remain active, while $\mathcal{F}_{con}$ also reach their highest activation levels.

Overall, this three-phase progression reveals a consistent pattern of representational flow during emotion inference. This finding provides additional nuance beyond prior work (Tak et al., 2025), which suggests that emotion signals can be detected in mid-layers based on linear probe performance. Our results indicate that such mid-layer probe success likely reflects the emergence of emotion-correlated semantic concepts, whereas more direct emotion-related feature activations predominantly arise in later layers (e.g., Gemma-2-2B and 9B). This distinction has important implications for intervention: emotion-related features are more concentrated in later layers, making them more effective targets for control, while earlier layers contain more diffuse and indirect signals.

5 RQ2: Causal Feature Identification for Emotion Inference

We further conduct causal interventions to explore the causal pathways underlying emotion recognition. While prior work discovers circuits via gradient-based indirect effect approximation across the full computation graph (Marks et al., 2025a; Conmy et al., 2023), we propose *phase-stratified causal tracing*, which exploits the three-phase structure identified in Section 4 to perform phase-conditioned circuit discovery. We trace causal influences backwards across phases using a greedy top-τ_{causal} selection procedure, conditioning each phase's discovery on the features identified in the phase above it, recovering sparse causal pathways grounded in the model's representational hierarchy.

Identifying Causal Emotion Features. Since emotion-related features primarily emerge in Phase 3, we estimate the causal contribution of each feature in $\mathcal{F}_e$ to the model's emotion prediction within this phase. For each feature $f_i \in \mathcal{F}_e$, we perform feature ablation by zeroing its activation in Phase 3 and measuring the resulting change in the output logit corresponding to emotion e. The magnitude of this logit change quantifies the causal influence of f_i on the final emotion prediction (Marks et al., 2025a). We then collect the subset of features whose mean causal effect ranks within the top ratio $\tau_{\text{causal}} = 0.05$ (i.e., the top 5%), denoted as $\mathcal{F}^{(3)}_{e,\text{causal}}$.

Backtracking Feature Dependencies. To determine the structure of the causal pathways, we backtrack the Phase 2 features that causally influence $\mathcal{F}^{(3)}_{e,\text{causal}}$. For each feature f_j in Phase 2, we ablate its activation and record the resulting change in the activation of each downstream feature $f_i \in \mathcal{F}^{(3)}_{e,\text{causal}}$. The magnitude of this activation change defines a directed causal edge from f_j to f_i.

Repeating this procedure across all feature pairs yields the set of causal edges from Phase 2 to Phase 3 features. We then select Phase 2 features whose mean causal effect ranks within the top ratio τ_{causal}, denoted as $\mathcal{F}^{(2)}_{e,\text{causal}}$. We repeat the same procedure from Phase 1 to Phase 2 to identify the corresponding set $\mathcal{F}^{(1)}_{e,\text{causal}}$.

Using these feature sets, we construct emotion-specific sparse feature causal pathways for each of the six basic emotions. These pathways represent the most causally influential features contributing to the final emotion prediction and enable systematic analysis of how intermediate concepts are formed, interact, and ultimately influence the model's output.

Structure of Emotion Pathways. After identifying causally influential features across all phases for all emotions, denoted as $\mathcal{F}_{e,\text{causal}}$, we combine them into a unified feature set $\mathcal{F}_{\text{causal}}$. The resulting feature set contains 105 unique features for Gemma-2-2B, 256 for Gemma-2-9B, and 92 for Llama-3.1-8B, as summarized in Table 1. For Gemma-2-2B, 28 out of the 105 causal features originate from Phase 1 and Phase 2. Among these, 22 features are shared by more than four emotions, and most correspond to syntax-related features. This suggests that early-phase features capture gen-

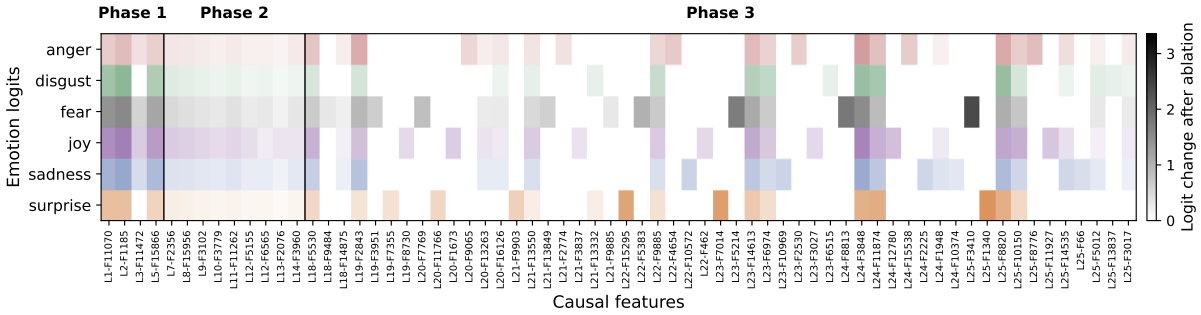

Figure 3: Influence of causal sparse features on emotion logits in Gemma-2-2B. Columns correspond to causal SAE features identified in Section 5 (filtered to those with logit influence greater than 0.4 on at least one emotion), and rows correspond to emotion logits. Color intensity indicates the magnitude of logit change after ablating each feature. Features in Phase 1 and Phase 2 are largely syntax-related and tend to influence multiple emotion logits with relatively uniform effects. In Phase 3, several features influence multiple emotions, while others exhibit strong emotion-specific effects. Emotions such as *anger* and *joy* have more causal features, whereas *Surprise* and *Disgust* have fewer. In particular, *Disgust* has only two emotion-specific features above the threshold, suggesting weaker representation of this emotion in the model. In contrast, although *Surprise* has fewer emotion-specific features, several of these features produce large logit changes, indicating the presence of strongly specialized features.

eral linguistic structures that broadly contribute to emotion prediction across instances rather than encoding emotion-specific semantics. In contrast, Phase 3 contains the majority of emotion-related features and exhibits substantially greater specialization across all three models.

As shown in Figure 3, the sizes of $\mathcal{F}^{(3)}_{e,\text{causal}}$ vary across emotions. *Anger*, *Joy*, and *Fear* are influenced by a larger number of emotion-specific causal features, suggesting richer causal representations. In contrast, *Disgust* and *Surprise* are influenced by fewer such features, and the causal effects of *Disgust*-related features are generally weaker. Furthermore, based on the feature explanations, the features in $\mathcal{F}^{(3)}_{disgust,\text{causal}}$ are not explicitly associated with *Disgust*; instead, they correspond to concepts such as negative sentiment and emotional turmoil. This suggests that the model predicts *Disgust* by combining related concepts, as no disgust-specific features are observed. The full list of $\mathcal{F}^{(3)}_{e,\text{causal}}$ is provided in Appendix A.3.

Beyond emotion-specific features, we also observe a set of general emotion-expression features shared across multiple emotions, suggesting that the model relies on both shared and emotion-specific features during emotion recognition.

6 RQ3: Improving Emotion Inference via Feature-Level Interventions

After obtaining the unified feature set $\mathcal{F}_{\text{causal}}$. We then apply a parameter-efficient optimization procedure to learn an interpretable steering vector that

			Phase 3					
Model	Total	Phase1–2	Anger	Joy	Sadness	Fear	Disgust	Surprise
Gemma-2-2B	105	28	31	36	28	30	19	17
Gemma-2-9B	256	64	72	87	74	75	56	43
Llama-3.1-8B	92	12	30	39	32	32	22	18

Table 1: Structure of causal feature pathways across models. We combine features from Phases 1 and 2 because they are largely $\mathcal{F}_{syn}$ and shared across emotions, whereas Phase 3 features exhibit stronger emotion specificity. Columns under Phase 3 report the number of features associated with each emotion based on their causal effects on emotion logits.

adjusts the activations of these features and improves the emotion recognition ability of the model, and test the generalization of the vector on multiple emotion recognition datasets.

Feature Steering Method and Setup. We learn a steering vector $S \in \mathcal{R}^d$ that scales the activations of features in $\mathcal{F}_{\text{causal}}$. Given the original feature activations F_{causal}, the steered activations are defined as

$$F'_{\text{causal}} = S \odot F_{\text{causal}},$$

where $\odot$ denotes element-wise scaling.

To train the steering vector, we optimize a composite objective function that balances emotion prediction accuracy, language modeling ability, and sparsity of the steering intervention:

$$L = L_{\text{CE}}(y, \hat{y}) + \alpha \cdot D_{\text{KL}}(p_o \| p_s) + \lambda \sum_{i=1}^{d} |s_i - 1|.$$

Here, L_{CE} denotes the cross-entropy loss for emotion recognition, D_{KL} measures the difference

between the original and steered token probability distributions p_o and p_s, and α controls the strength of the language-preservation regularization. This use of KL divergence is inspired by prior work that leverages it to preserve output coherence when optimizing against a secondary objective, most notably as a reward penalty in RLHF (Ziegler et al., 2020) and as a primary training objective in KTS, where model weights are adapted to accommodate fixed steering vectors (Stickland et al., 2024). In contrast to these settings, we apply KL regularization to constrain feature scaling parameters, ensuring that targeted interventions do not degrade language modeling ability. Finally, the L_1 penalty encourages sparse steering by limiting the number of features whose activations are significantly modified from a baseline of 1.

We conduct steering experiments on Gemma-2-2B, Gemma-2-9B, and Llama-3.1-8B. The training and validation sets are drawn from the combined dataset described in the experimental setup. To evaluate the generalization of the enhanced emotion recognition capability of the steered models, we further test them on three additional emotion recognition datasets: EmoEvent, Emotion Cause, and XED (Plaza del Arco et al., 2020; Ghazi et al., 2015; Öhman et al., 2020). More training details are provided in the Appendix A.6.

Results. Table 2 reports the per-emotion F1 scores before and after steering, together with generalization performance on three additional datasets and language modeling perplexity. The baseline models perform particularly poorly on *Disgust* and *Surprise*. This observation is consistent with our earlier analysis: *Disgust* lacks explicit internal representations, while surprise-related features are frequently co-activated with other emotions, especially *Joy*. After applying steering, the F1 scores for these two emotions improve substantially across all models, while the other emotions also show moderate gains. The improvements also generalize well to external datasets with different forms, especially to EmoEvent and XED. Meanwhile, model perplexity changes only marginally (within 0–8%), indicating that steering preserves language modeling ability while improving emotion recognition.

Table 3 shows a subset of the learned steering vector for Gemma-2-2B, including features whose steering multipliers change by more than 10%. The results indicate that the steering vector is highly sparse: only 12 features are significantly modified, yet these changes lead to substantial performance improvements. Several semantically meaningful features are amplified. For example, feature L21–F13332, associated with expressions of sympathy, L25–F13837, related to emotion *Disgust*, and feature L25–F1340, related to expressions of surprise, are upweighted. In contrast, some features are suppressed, including L23–F2530 (expressions of anger) and L22–F4654 (instances of intense emotional states). These results suggest a trend in which underrepresented emotions, such as *Disgust* and *Surprise*, are amplified, while relatively overrepresented emotions, such as *anger*, are suppressed. These patterns suggest that the steering procedure selectively adjusts a small set of semantically interpretable features, providing a transparent and targeted mechanism for controlling model behavior while maintaining overall model stability. Since the steering only learns a small number of parameters, we conduct additional ablation studies to examine the contribution of features from different phases. The results are provided in the Appendix A.7.

Data Efficiency of Feature Steering. To evaluate the data efficiency of our feature steering method, we conduct experiments using the Gemma-2-2B model. We vary the number of labeled training examples per emotion while keeping all other components fixed. We sample k examples per emotion class from the training set, where $k \in 1, 2, 5, 10, 50, 100, 200, 400$. For $k \leq 100$, we randomly sample the training set five times and report the average F1 scores. Performance is evaluated on a held-out validation set using macro-F1. Table 4 shows that feature steering is highly data-efficient, and that five examples per emotion are sufficient to achieve near-maximal performance across all emotions. This result suggests that the discovered emotion features can be adapted to other emotion datasets using only a small number of examples to train the steering vector.

7 Conclusion

Large language models are increasingly used in emotionally sensitive human–AI interactions. While many studies evaluate the emotion recognition capabilities of LLMs through benchmark-based assessments of model outputs, relatively few examine the internal mechanisms underlying these behaviors. Understanding such mechanisms is crucial for diagnosing model failures and enabling safer, more interpretable, and more controllable

Model	Method	Main Dataset (F1)							EmoEvent	EmotionCause	XED	PPL
		Anger	Disgust	Fear	Joy	Sadness	Surprise	Macro				
Gemma-2-2B	Baseline	0.54	0.28	0.72	0.79	0.69	0.28	0.55	0.34	0.58	0.39	17.58
	Steered	0.60	0.64	0.75	0.80	0.74	0.34	0.65	0.48	0.62	0.44	17.53
Gemma-2-9B	Baseline	0.57	0.60	0.74	0.84	0.72	0.24	0.62	0.44	0.68	0.45	13.25
	Steered	0.60	0.70	0.81	0.85	0.76	0.43	0.69	0.50	0.68	0.51	13.38
Llama-3.1-8B	Baseline	0.65	0.49	0.75	0.83	0.77	0.29	0.63	0.39	0.64	0.46	11.49
	Steered	0.68	0.68	0.80	0.85	0.79	0.38	0.70	0.49	0.68	0.51	12.41

Table 2: Emotion recognition performance before and after steering. The main dataset reports F1 scores for six emotions and macro-F1. EmoEvent, EmotionCause, and XED evaluate generalization to external datasets. Perplexity (PPL) measures language modeling performance to assess the preservation of general language ability. Cells highlighted in green indicate the largest improvements in single-emotion F1. Macro-F1 improves by 0.10, 0.07, and 0.07 for the three models, respectively. The largest improvement is observed for *Disgust* on Gemma-2-2B (+0.36), while the smallest improvement is for *Joy* on Gemma-2-2B and Gemma-2-9B (+0.01).

Feature	Description	Multiplier
L23–F6515	References to power dynamics and social interactions	1.611
L25–F1340	Phrases and expressions related to surprise or unexpected events	1.303
L24–F1380	Negative sentiments and internal conflicts regarding personal beliefs and experiences	1.265
L25–F13837	References to personal identity crises and the emotional turmoil associated with societal expectations	1.268
L23–F6974	Expressions of interest or emotional engagement related to personal relationships and social dynamics	1.259
L21–F13332	Expressions of condolence and sympathy	1.167
L22–F9885	Emotional responses to diagnoses and health-related situations	1.142
L24–F11661	Words and phrases related to spiritual experiences and feelings of connection	1.103
L23–F2530	Expressions of frustration and anger	0.877
L15–F9079	Special characters and formatting elements within the text	0.853
L22–F4654	Instances of intense emotional states	0.701
L19–F2843	Mentions of specific significant events or figures	0.559

Table 3: Causal features and their learned steering multipliers (multipliers with changes exceeding 10%).

# Samples / emotion	Macro F1	Anger	Disgust	Fear	Joy	Sadness	Surprise
1	0.60	0.55	0.58	0.72	0.70	0.68	0.35
2	0.62	0.59	0.56	0.74	0.79	0.72	0.32
5	0.64	0.60	0.60	0.74	0.79	0.74	0.36
10	0.64	0.61	0.57	0.74	0.80	0.74	0.36
50	0.64	0.60	0.58	0.74	0.80	0.73	0.35
100	0.64	0.59	0.61	0.75	0.80	0.75	0.35
200	0.63	0.59	0.59	0.74	0.80	0.72	0.36
400	0.63	0.60	0.59	0.75	0.80	0.74	0.33

Table 4: Effect of training data size on steering performance.

improvements.

In this work, we investigate the internal representations of LLMs during emotion inference using SAEs. We identify a consistent three-phase information flow, in which representations progress from syntactic features to semantic concepts and finally to emotion-related features. We further show that emotions are encoded through a combination of shared and emotion-specific features, with substantial imbalance across emotions. Our analysis also provides mechanistic insights into common failure modes: *Disgust* is weakly and diffusely represented, while features associated with *Surprise* are frequently activated across multiple emotions, leading to misclassification.

To better characterize the causal structure of emotion inference, we propose a phase-conditioned circuit discovery method that identifies a minimal set of strongly causal features. Building on these insights, we introduce an interpretable and data-efficient sparse feature steering approach that learns a steering vector over causally identified features. This method significantly improves emotion recognition performance while largely preserving language modeling capabilities.

Overall, our findings demonstrate the potential of mechanistic interpretability methods for advancing affective computing with large language models. Future work may further explore the causal pathways linking semantic and emotional representations, as well as how emotion-related features interact and compose to form complex affective states. The causal features identified in this work may also serve as reusable building blocks for other affective computing tasks under limited data settings. We will release our code and resources following the double-blind review process.

Limitations

Due to the limited availability of publicly released SAEs from GemmaScope and LlamaScope, we re-

strict our analysis to three models. While these models span different architectures and sizes and exhibit consistent patterns across model families, we do not evaluate larger models or additional families. Other available SAE releases are less suitable for our study: some are trained on models that are too small to reliably perform emotion inference (e.g., GPT-2, Pythia-70M) (Karvonen et al., 2024; Gao et al., 2025), others do not provide SAEs for all layers (e.g., GPT-OSS-20B SAE) (Goodfire, 2025), and some employ alternative architectures that are not directly comparable (e.g., Matryoshka SAEs) (Bussmann et al., 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

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

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.

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

Soren Bills, Nathan Cammarata, Dan Mossing, Henry Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeffrey Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html. Accessed: 2025-02-01.

Abdelaziz Bounhar, Rania Hossam Elmohamady Elbadry, Hadi Abdine, Preslav Nakov, Michalis Vazirgiannis, and Guokan Shang. 2026. Yapo: Learnable sparse activation steering vectors for domain adaptation. *Preprint*, arXiv:2601.08441.

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.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. 2025. Learning multi-level features with matryoshka sparse autoencoders. *Preprint*, arXiv:2503.17547.

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.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Preprint*, arXiv:2304.14997.

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.

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.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. 2025. Monitoring latent world states in language models with propositional probes. In *International Conference on Learning Representations*, volume 2025, pages 19337–19359.

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*.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Preprint*, arXiv:2301.04709.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *CICLing (2)*, pages 152–165.

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.

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. 2025. Sae-ssv: Supervised steering in sparse representation spaces for reliable control of language models. *Preprint*, arXiv:2505.16188.

Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. 2025. Probing then editing response personality of large language models. *Preprint*, arXiv:2504.10227.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8235–8246, Torino, Italia. ELRA and ICCL.

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.

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.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

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

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2025. The quest for the right mediator: Surveying mechanistic interpretability through the lens of causal mediation analysis. *Preprint*, arXiv:2408.01416.

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.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2025. A practical review of mechanistic interpretability for transformer-based language models. *Preprint*, arXiv:2407.02646.

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.

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.

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.

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.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*.

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*.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R. Bowman. 2024. Steering without side effects: Improving post-deployment control of language models. *Preprint*, arXiv:2406.15518.

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.

Ala N. Tak and Jonathan Gratch. 2023. Is gpt a computational model of emotion? In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

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.

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.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

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.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

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:

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

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 Sparse Feature Linear Probing

Method. To assess the presence and strength of emotion-related signals within SAE feature activations, and to identify features that are most discriminative across emotions, we train linear classifiers ("probes") to predict emotion labels from sparse feature vectors. We formalize the probing task as multinomial logistic regression over six emotion classes, where the probability of emotion class k given feature vector $\mathbf{f}$ is computed via a softmax over linear projections.

We employ logistic regression with an elastic net penalty, which combines L1 (lasso) and L2 (ridge) regularization to balance sparsity and stability. The elastic net objective consists of the negative log-likelihood augmented with a weighted combination of L1 and L2 penalties on the weight matrix, where λ controls the overall regularization strength (implemented as $C = 1/\lambda$) and α determines the L1/L2 mixing ratio.

We set $\alpha = 0.3$, prioritizing ridge regularization while retaining sufficient sparsity for feature selection. This setting encourages distributed reliance on multiple features while maintaining interpretability. The inverse regularization strength is set to $C = 0.002$ to account for the high dimensionality of the SAE feature space and the large dataset size. All probes are trained using the SAGA solver, which efficiently supports elastic net penalties via stochastic average gradient descent. We train each probe for up to 2,000 iterations with early stopping based on convergence tolerance.

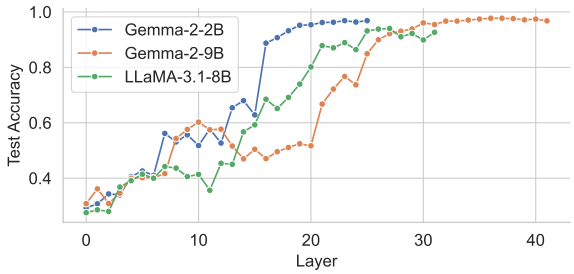

Figure A.1: Layer-wise linear probe accuracy for each model.

Probes are trained independently for each layer using a balanced training set and evaluated on a held-out balanced test set. Hyperparameters are empirically chosen to achieve strong probe performance in later layers while preserving a sufficient number of non-zero feature weights for downstream steering and circuit analysis.

Emotion Disentanglement. Figure A.1 shows linear probe accuracy across layers. Probe accuracy remains low in early layers, indicating that emotion-relevant representations are largely entangled. Accuracy then increases steadily in mid-layers, suggesting that the activated features associated with different emotions begin to diverge. In late layers, probe accuracy exceeds 90%, consistent with the emergence of explicit emotion-related features.

Given that feature interpretations indicate that emotion-related features emerge in Phase Three, our results suggest that while hidden states (or activated features) associated with different emotions become increasingly separable in the mid-layers, this separability *primarily reflects the differentiation of semantic concepts rather than the formation of explicit emotion representations*, although these semantic concepts may be highly correlated with emotion-related representations in later layers.

Importantly, emotion-related representations do not disentangle simultaneously across layers. As shown in Figure A.2, a linear probe initially defaults to predicting all emotions as anger (Layer 0), and first learns to distinguish positive from negative affect at early layers (Layer 5). It then begins to separate fear and sadness from anger around Layer 8, achieving reliable discrimination after Layer 13. In contrast, the probe can only distinguish disgust from anger at later layers (after Layer 17), and continues to struggle to separate surprise from joy even at the final layer.

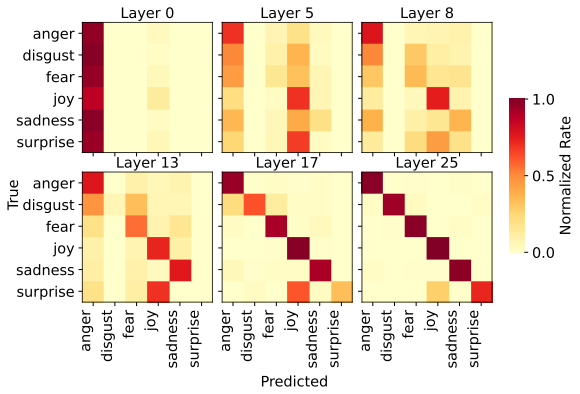

Figure A.2: Layer-wise confusion matrices of the linear probe for Gemma-2-2B.

A.3 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.4 Case Study

To better illustrate how the model makes emotion predictions and how information flows across the three phases, we present a concrete example using Gemma-2-2B. Figure A.3 visualizes a subset of the sparse feature causal graph for the prompt below. For visualization clarity, we retain only features with logit weights ≥ 0.15 across all layers.

"In this text: I felt … when an aeroplane I was on hit heavy turbulence and dropped a long way down suddenly, the emotion implied is:"

In the early layers, most causally important features are syntactic in nature, such as L1-F11070, L2-F1185, and L5-F15866. In later layers, more concept-level features emerge. For example, L13-F14818 is associated with phrases that establish analytical outcomes, which may be triggered by the task framing that asks the model to infer an emotional outcome from the described event. L16-11090 relates to stunts and action sequences, plausibly activated by the sudden airplane drop. This feature exhibits a strong causal influence on feature 7177 in the following layer, which is associated with evacuation and emergency responses. In turn, this activation causally impacts feature 9484 in the next layer, representing concepts related to fear and its psychological impact. Subsequently, emotion-relevant features begin to dominate: a sequence of fear-related features is activated across subsequent layers, culminating in feature 3410 in the final layer, which corresponds to intense emotional reactions and psychological responses associated with fear and horror.

Overall, this progression suggests a hierarchical inference process: the model first processes syntactic and structural information, then extracts salient actions and situational cues, and finally activates emotion-specific representations that drive the final prediction.

A.5 Topic Modeling

To preserve fine-grained emotional distinctions while mitigating spurious variation, we tune the dimensionality reduction and clustering stages of the BERTopic pipeline via a grid search. For UMAP, we vary the number of neighbors $n_{\text{neighbors}} \in \{5, 10, 15, 30\}$, the embedding dimensionality $n_{\text{components}} \in \{5, 10\}$, and the minimum distance parameter min_dist $\in \{0.0, 0.1\}$. For HDBSCAN, we tune the minimum cluster size min_cluster_size $\in \{5, 10\}$ and the minimum number of samples min_samples $\in \{\text{None}, 5\}$. The final configuration is selected by balancing emotional topic granularity, cluster coherence, and the proportion of points labeled as noise, resulting in 82 topics.

We then use GPT-4o-mini to classify the resulting feature topics into four categories: *syntax*, *concept*, *emotion*, and *other*. The prompt used for classification is shown in Table A.1.

A.6 Model Training Details

In our main steering experiment, we use an 80–20 train–test split on the main dataset and sample 450 training examples per emotion from the training set. For Gemma-2-2B, we set the KL-divergence regularization coefficient to $\alpha = 0.5$ and the L1 regularization coefficient to $\lambda = 0.01$. For Gemma-2-9B and Llama-3.1-8B, we use $\alpha = 0.05$ and $\lambda = 0.001$.

Figure A.3: Subset of the sparse feature causal graph for a representative case-study prompt. Nodes are filtered using a threshold of 0.15 on their contribution to the *fear* logit. Node color indicates the logit weight. We highlight the path leading to L25-F3410, which has the largest causal impact on the *fear* logit. Along this path, the model transitions from syntactic features (e.g., L0-F15092, "punctuation marks, colons, and parentheses") to semantic features, including L13-F14818 ("phrases that establish outcomes in an analytical context"), L16-F11090 ("references in stunts and action sequences in film and television"), and L17-F7177 ("terms related to evacuations and emergency responses"). L17-F7177 then causally influences the first fear-specific feature at L18-F9484 ("concepts and discussions related to fear and its psychological impact"). Subsequent nodes along the path are predominantly fear-related, culminating in L25-F3410 ("intense emotional reactions and psychological responses related to fear and horror").

Feature Topic Classification Prompt
Assign the following SAE feature topic name to exactly ONE category: 1) syntax–Surface-level textual structure, including formatting, punctuation, grammar, code structure, programming constructs, or symbolic and mathematical operations. 2) concept–Semantic topics, domains, roles, entities, activities, events, or situational knowledge that are neither syntactic nor explicit emotional expressions. 3) emotion–Direct expressions of emotional states, affective language, or explicit emotion terms. 4) other–Does not clearly belong to any of the above categories. Rules: –Do not invent new labels. –Base your decision only on the topic name. –Output only the label (one word). Topic: `{topic}`

Table A.1: Prompt used to classify SAE feature topics using GPT-4o-mini.

We evaluate language modeling performance using perplexity on a subset of the WikiText-103 test set (Merity et al., 2017).

A.7 Ablation Studies

To investigate the contribution of different phases to emotion recognition, we conduct an ablation study where the steering features $\mathcal{F}_{e,\text{causal}}$ are restricted to specific phases of the model. In each setting, we train the steering vectors using only the causal features extracted from the selected phase(s).

Table A.2 reports the resulting macro F1 and per-emotion F1 scores. When steering is applied using only Phase 1 or Phase 2 features, performance improves slightly compared to the baseline model. In contrast, using Phase 3 features leads to a substantially larger improvement, indicating that later layers contain features that more strongly distinguish emotions. Adding features from other phases on top of Phase 3 provides only minimal additional gains.

Phase	Macro F1	Anger	Disgust	Fear	Joy	Sadness	Surprise
Baseline	0.55	0.54	0.28	0.72	0.79	0.69	0.28
Phase 1	0.56	0.55	0.29	0.72	0.80	0.70	0.29
Phase 2	0.57	0.56	0.34	0.72	0.80	0.71	0.29
Phase 3	0.63	0.60	0.58	0.75	0.80	0.73	0.34
Phase 1 + 2	0.57	0.56	0.35	0.72	0.80	0.71	0.29
Phase 2 + 3	0.64	0.60	0.59	0.74	0.80	0.74	0.34
Phase 1 + 2 + 3	0.63	0.60	0.58	0.74	0.80	0.73	0.33

Table A.2: Phase ablation results showing macro F1 and per-emotion F1 scores.

Feature ID	Emotion	Explanation	Feature Topic	Category
L16-F5280	joy	expressions of emotional conflict and interpersonal relationships	Emotional and Interpersonal Conflict Dynamics	concept
L17-F2792	anger	expressions of anger and frustration	Emotional Expressions of Anger and Frustration	emotion
L17-F7026	anger	punctuation and formatting in the text	punctuation and formatting in text	syntax
L18-F5530	anger, disgust, fear, joy, sadness, surprise	emotions and their expressions in various contexts	Expressions of Personal Emotional States	emotion
L18-F9484	fear	concepts and discussions related to fear and its psychological impact	Expressions and Terms of Fear and Anxiety	emotion
L18-F11186	sadness	emotional expressions related to loss and sadness	Emotional Expressions of Sadness and Loss	emotion
L18-F14875	anger, fear, joy, sadness	punctuation marks and indicators of inclusion or lists	Colons and Punctuation in Textual Lists	syntax
L19-F7355	surprise	situations involving surprise or unexpected outcomes	Expressions and Reactions to Surprise and Unexpected Events	emotion
L19-F8730	joy	expressions and sentiments related to personal experiences and excitement	Expressions of Excitement and Positive Sentiments	emotion
L20-F1673	joy	expressions of celebration and emotional responses	Expressions of Personal Emotional States	emotion
L20-F16126	anger, disgust, fear, joy, sadness	colons and their associated structures	Colons and Punctuation in Textual Lists	syntax
L21-F3837	joy	expressions of excitement and congratulations in relation to personal achievements or milestones	Celebrations and Community Gatherings	concept
L21-F9885	fear	statements that convey emotional conflict or resentment	Emotional and Interpersonal Conflict Dynamics	concept
L21-F13550	anger, disgust, fear, joy, sadness, surprise	concepts and discussions centered around values, especially those related to love and family	concept Considerations of Compassion and Moral Values in Community and Governance	concept
L21-F13849	fear	expressions of fear and anxiety in various contexts	Expressions and Terms of Fear and Anxiety	emotion
L22-F462	joy	expressions of excitement or emotion related to personal achievements and shared experiences	Expressions of Excitement and Positive Sentiments	emotion
L22-F5383	fear	words and phrases related to fear and anxiety	Expressions and Terms of Fear and Anxiety	emotion
L22-F9885	anger, disgust, fear, joy, sadness, surprise	emotional responses to diagnoses and health-related situations	Expressions of Personal Emotional States	emotion
L22-F10572	sadness	expressions of disappointment and emotional reactions	Expressions of Disappointment and Emotional Responses	emotion
L22-F15295	surprise	unexpected reactions and emotions related to news or events	Expressions and Reactions to Surprise and Unexpected Events	emotion
L22-F15368	joy	elements related to quality, clarity, and efficiency in communication or presentation	Communication and Social Interaction Dynamics	concept
L23-F2670	joy	concepts related to symbolism and emotional connections in relationships	Celebrations and Community Gatherings	concept
L23-F5214	fear	concepts related to fear and anxiety in various contexts	Expressions and Terms of Fear and Anxiety	emotion
L23-F6515	disgust	references to power dynamics and social interactions	Interpersonal Dynamics in Personal and Social Relationships	concept
L23-F7014	surprise	expressions related to surprise and unexpected events	Expressions and Reactions to Surprise and Unexpected Events	emotion
L23-F10969	sadness	emotional expressions related to disappointment and regret	Expressions of Disappointment and Emotional Responses	emotion
L23-F14613	anger, disgust, fear, joy, sadness, surprise	terms and concepts related to programming and software development practices	Programming and Data Structures Concepts	concept
L24-F2225	sadness	references to emotional and physical well-being	Happiness and Fulfillment in Personal Well-Being	emotion
L24-F3848	anger, disgust, fear, joy, sadness, surprise	terms related to scientific and medical topics	Medical and Biological Terminology in Healthcare	concept
L24-F4406	fear	concerns related to legal and regulatory issues	Legal Proceedings and Terminology	concept
L24-F10374	sadness	emotional responses related to disappointment	Expressions of Disappointment and Emotional Responses	emotion
L24-F11661	joy	words and phrases related to spiritual experiences and feelings of connection	Spiritual Practices and Beliefs in Religion	concept
L24-F15538	anger	emotional reactions and interpersonal conflicts	Emotional and Interpersonal Conflict Dynamics	concept
L25-F3017	anger, disgust, fear, joy, sadness	financial report figures, including revenue, guidance, and detailed numerical metrics	Financial Metrics and Economic Data Analysis	concept
L25-F3410	fear	intense emotional reactions and psychological responses related to fear and horror	Expressions and Terms of Fear and Anxiety	emotion
L25-F5012	anger, disgust, fear, joy, sadness	themes related to government and qualifications for leadership	Leadership and Hosting in Political Contexts	concept
L25-F13749	joy	specific references to medical and clinical treatment settings or methodologies	Medical and Biological Terminology in Healthcare	concept
L25-F13837	disgust	references to personal identity crises and the emotional turmoil associated with societal expectations	Emotional and Psychological Struggles in Society	concept
L25-F15056	disgust	expressions indicating opinions, feelings, or subjective observations	Personal Opinions and Expressions	emotion

Table A.3: A representative subset of Phase-3 causal features $\mathcal{F}^{(3)}_{e,\text{causal}}$ identified for emotion prediction. Feature explanations are derived from SAE feature descriptions.

Figure A.4: A subset of feature topics and their mean activation across layers for instances with different emotion labels for Gemma-2-2B.