Emotions Where Art Thou: Understanding and Characterizing the Emotional Latent Space of Large Language Models

Benjamin Reichman
Georgia Institute of Technology
bzr@gatech.edu

Adar Avsian
Georgia Institute of Technology
aavsian3@gatech.edu

Larry Heck
Georgia Institute of Technology
larryheck@gatech.edu

Abstract

This work investigates how large language models (LLMs) internally represent emotion by analyzing the geometry of their hidden-state space. Using a synthetic dataset of emotionally rewritten sentences, we identify a low-dimensional emotional manifold via singular value decomposition and show that emotional representations are directionally encoded, distributed across layers, and aligned with interpretable dimensions. These structures are stable across depth and generalize to eight real-world emotion datasets spanning five languages. Cross-domain alignment yields low error and strong linear probe performance, indicating a universal emotional subspace. Within this space, internal emotion perception can be steered while preserving semantics using a learned intervention module, with especially strong control for basic emotions across languages. These findings reveal a consistent and manipulable affective geometry in LLMs and offer insight into how they internalize and process emotion.

1 Introduction

Large Language Models (LLMs) have become central tools for interacting with, analyzing, and generating human language. Their widespread deployment across domains has led to increasing interest in how they handle not just syntactic or semantic meaning, but also affective tone. Emotion is a fundamental part of language, shaping persuasion, social signaling, and narrative context. As such, understanding how LLMs process emotional content is essential for both interpretability and safe deployment.

A lot of literature on affect in NLP has focused on sentiment analysis, a task where models classify inputs into discrete emotional or affective categories (1; 2; 3; 4). While this demonstrates that LLMs can identify emotions, it offers little insight into how emotional meaning is represented internally. Classification accuracy is not equivalent to interpretability.

Other works have taken a behavioral view, exploring the emotional "intelligence" of LLMs. These include prompting models with hypothetical emotional scenarios and evaluating their responses (5), or probing how well they align with human judgments in affective tone (6; 7). Though these studies suggest some degree of affective sensitivity, they remain surface-level—they test outputs rather than investigating internal mechanisms.

Recent work has also examined emotion manipulation and decoding. For instance, models have been used to map text to dimensional emotion ratings like valence-arousal-dominance (VAD) (8; 9), or to generate emotionally inflected language on demand (10). LLMs have also been shown to be more likely to comply with emotionally framed requests (11). Yet even these studies largely treat emotion as a label or generation condition—not a latent internal representation.

While there has been some work examining how LLMs respond to or generate emotional language, the structure of emotional representations within their hidden states remains relatively underexplored. Most prior approaches focus on output behavior or classification accuracy, with comparatively few efforts aimed at interpreting the internal geometry of emotion encoding. This work addresses that gap by analyzing how emotions are represented within LLM hidden states across layers, datasets, and languages. We find that emotional encoding is directional, distributed, and remarkably consistent across varied textual modalities. We also investigate the model's internal "psychology": how emotions are separated, aligned, and—critically—how they can be steered via targeted interventions.

Our contributions are as follows: (1) We construct a low-dimensional emotional subspace via SVD and show that it captures interpretable, directionally encoded affective structure across LLM layers. (2) We demonstrate that this space generalizes across eight emotion datasets spanning five languages, with low alignment distortion and high cross-domain probe accuracy. (3) We introduce a learned steering module that manipulates internal emotional perception while preserving semantic content, with especially strong control over basic emotions.

2 Related Works

2.1 Models of Emotions

Psychological models of emotion are commonly categorized as either discrete or continuous. Discrete theories posit that emotions are fundamentally distinct categories—such as the six "basic" emotions proposed by Ekman (12): anger, surprise, disgust, enjoyment, fear, and sadness. Other taxonomies expand this set, including more nuanced affective states (13).

In contrast, continuous models view emotions as points in a low-dimensional latent space. A widely used formulation is the valence-arousal-dominance (VAD) model (14), where valence encodes hedonic tone, arousal measures intensity, and dominance reflects control or agency. Variants of this framework reduce or alter the axes (e.g., Russell's 2D circumplex (15)).

These representations offer an interpretive lens for analyzing learned emotion structure in LLMs: If models implicitly encode emotions in a geometric space, we may expect that certain latent directions align with these classic dimensions. Our work explores whether such structure emerges naturally in the hidden-state geometry of LLMs trained without explicit emotional supervision.

Neuroscientific models of emotion offer a parallel debate. Localist theories posit that discrete emotions correspond to specific, anatomically distinct brain regions, while constructionist theories argue that emotions emerge from distributed, domain-general processes (16; 17; 18). Our results, particularly from ML-AURA (Section 5), support a constructionist-style interpretation in LLMs: emotional content is not localized to a small subset of units but is instead widely distributed across neurons and layers, with high separability emerging from overlapping, multi-purpose components.

2.2 Emotions in Latent Space

Recent work has investigated how LLMs interact with emotional text, often focusing on behavior or output-level mappings. For example, ChatGPT has shown the ability to map emotions to Valence-Arousal-Dominance (VAD) values (9; 19), suggesting that emotion-relevant dimensions are accessible to the model. However, such studies do not analyze the internal structure or geometry of these latent representations.

Some prior work explicitly trains models to embed emotions into structured spaces, using classification objectives or external supervision. For instance, (20) and (21) train models to map between emotion spaces. Similarly, (22) learns an emotion space from labeled data, shows clustering by valence, and demonstrates transferability across datasets. However, in all of these works, the emotion space is imposed or supervised, not emergent.

A smaller body of work begins to analyze how emotions might be natively represented in pretrained models. For example, (23) finds that valence appears to be linearly embedded in contextual representations. (24) further show that while valence aligns well with linear probes, arousal and dominance are less separable—but their setup relies on encoder-only models and assumes fixed affective lexicons to ground the analysis. In contrast, we examine decoder-only LLMs and aim not to impose a psychological model onto the network, but to recover the emergent emotional structure from the geometry of its hidden states.

Other studies have shown that LLMs exhibit strong zero-shot emotion classification performance across languages (25), though subsequent work notes that language-specific tuning is sometimes necessary for culturally grounded affect (26). These findings suggest that emotion representations are at least partially transferable across linguistic domains—a hypothesis we test more directly through geometric alignment and projection-based analysis in Section 4.

3 Methods

To understand how emotions are represented in LLMs, a variety of tools were used. This section outlines those methods and their theoretical grounding. Empirical findings from these analyses are presented in Sections 4 and 5.

3.1 ML-AURA

ML-AURA quantifies how selectively a neuron responds to a specific concept by framing each neuron as a threshold-based detector (27). For a labeled dataset D, each neuron's output is summarized per example using the maximum activation across tokens. These scalar responses are then ranked and evaluated using the area under the precision-recall curve, comparing neuron output against the presence or absence of the target concept. Neurons with high AUC-PR are designated as "experts" for that concept.

In our adaptation, the concepts are emotion categories. We apply ML-AURA in a one-vs-all setup for each emotion, scoring each neuron by how well it distinguishes a target emotion from all others.

3.2 Centered-SVD

Prior work has shown that LLM hidden states lie on low-dimensional manifolds and that many semantic and syntactic properties are linearly recoverable in these subspaces (28; 29; 30). Leveraging this, we identify emotion-relevant subspaces using singular value decomposition (SVD).

We use the dataset introduced in (10), where each example consists of a human-authored neutral sentence paired with synthetic rewrites that express a specific target emotion $e \in \mathcal{E}$ while preserving semantic content. Each input x_i is passed through the model to extract hidden states at a given layer or sublayer. These activations are mean-pooled across tokens to produce sentence-level vectors $h_i \in \mathbb{R}^d$.

We aggregate these into a matrix $H \in \mathbb{R}^{N \times d}$, where N is the total number of emotionally labeled inputs. After centering H, we compute its singular value decomposition: $H = U\Sigma V^\top$. The rows of $V^\top$ define principal directions in the model's internal representation space. Given that emotional content is the primary structured variation across inputs, we hypothesize that the leading components capture dominant emotional axes. This hypothesis is evaluated through downstream alignment, probing, and causal manipulation.

3.3 Space Alignment

Prior work has shown that latent spaces arising from related tasks often exhibit similar internal geometry, with relationships between them approximately rigid or linear up to rescaling and rotation (31). While some approaches lift these spaces into anchor-relative

Dataset	Stress-1 ↓	Stress-2 ↓	Sammon ↓	Avg Dist ↓	ℓ_2 ↓	σ ↓	Probe Acc. ↑
Go-Emotions*	0.33 ± 0.33	0.13 ± 0.13	*	*	*	*	0.52 ± 0.52
CARER (Twitter)	0.33 ± 0.33	0.13 ± 0.13	0.17 ± 0.17	1.16 ± 1.16	1.23 ± 1.23	0.12 ± 0.12	0.60 ± 0.60
SemEval	0.29 ± 0.29	0.11 ± 0.11	0.18 ± 0.18	1.09 ± 1.09	1.16 ± 1.16	0.09 ± 0.09	0.65 ± 0.65
EmoEvent (EN)	0.31 ± 0.31	0.11 ± 0.11	0.13 ± 0.13	1.02 ± 1.02	1.09 ± 1.09	0.12 ± 0.12	0.71 ± 0.71
EmoEvent (ES)	0.32 ± 0.32	0.12 ± 0.12	0.14 ± 0.14	0.97 ± 0.97	1.05 ± 1.05	0.13 ± 0.13	0.72 ± 0.72
Bhaav (Hindi)	0.32 ± 0.32	0.12 ± 0.12	0.14 ± 0.14	0.96 ± 0.96	1.03 ± 1.03	0.13 ± 0.13	0.53 ± 0.53
German Drama*	0.40 ± 0.40	0.29 ± 0.29	*	*	*	*	0.57 ± 0.57
MultiEmotions-It	0.39 ± 0.39	0.17 ± 0.17	0.22 ± 0.22	1.24 ± 1.24	1.33 ± 1.33	0.15 ± 0.15	0.62 ± 0.62
EmoTextToKids (FR)	0.33 ± 0.33	0.12 ± 0.12	0.15 ± 0.15	1.01 ± 1.01	1.09 ± 1.09	0.14 ± 0.14	0.68 ± 0.68
Average (Full-Space)	0.34 ± 0.17	0.14 ± 0.56	0.14 ± 0.44	1.10 ± 0.20	1.14 ± 0.38	0.13 ± 0.27	0.62 ± 0.09
Average (50D-Space)	0.45 ± 0.16	0.23 ± 0.55	0.24 ± 0.46	0.82 ± 0.29	0.91 ± 0.45	0.22 ± 0.29	0.54 ± 0.11

Table 1: Per-dataset distortion metrics and probe accuracy in the synthetic emotional subspace. Lower distortion values indicate greater geometric consistency. Probe accuracy reflects how well emotion labels can be decoded via a linear probe trained on the synthetic manifold. Datasets marked with * were identified as outliers; asterisks in cells indicate anomalously high values omitted for readability.

representations to handle isometric variance, recent work demonstrates that direct alignment via linear or rigid transformations is often sufficient and easier to apply in practice (32). Following this approach, we use linear regression to align the emotional subspace derived from synthetic data with that derived from human-authored emotion classification datasets. This alignment allows us to test whether the structure found in the synthetic manifold reflects transferable emotional encodings or artifacts specific to the synthetic generation process.

4 Emotion Universality

Using the tools presented in Section 3, we provide evidence that emotional representations in LLMs are structurally universal. We show that emotions are encoded in similar geometric subspaces across datasets, languages, and writing styles. This and all subsequent sections focus on LLaMA 3.1; the appendices provide analogous results for Qwen 2.5 and Mistral 7B.

4.1 Datasets

To evaluate the universality of emotional representations in LLMs, we selected a diverse set of emotion classification datasets spanning multiple languages, modalities, and writing styles. Only datasets with explicit categorical emotion labels were included; datasets with only polarity (e.g., positive/negative) or star ratings were excluded due to insufficient granularity. In total, we use eight datasets: (1) Go-Emotions (33): English Reddit comments; (2) CARER (34): English tweets; (3) SemEval-2007 Task 14 (35): English news headlines; (4) EmoEvent (36): English and Spanish tweets; (5) Emotions in Drama (37): German plays from the 18th–19th century; (6) Bhaav (38): Hindi short stories; (7) MultiEmotions-It (39): Italian YouTube and Facebook comments; (8) EmoTextToKids (40): French journalistic and encyclopedic text aimed at children. The chosen languages are those for which high-quality emotion datasets exist and which are officially supported by LLaMA 3.1, as specified in its technical report.

4.2 Universality Analysis

The first step of the universality analysis was to collect the mean-pooled hidden-states of the model when text x_i from dataset D is input into it. Then the mean-pooled hidden state is either projected onto the space described in Section 3.2 or statistics are directly derived from comparing the mean-pooled hidden states.

The first analysis compares the cosine similarity of emotion-specific centroids across datasets. For each emotion shared between a dataset and the synthetic corpus (e.g., happy or joy), we compute the mean-pooled hidden-state centroid and compare it to its synthetic counterpart. Averaged across all shared emotions, layers, and sublayer types, the centroid cosine

similarity is 0.838 ± 0.12. Figure 1 shows the centroid cosine similarity across each of the datasets. This high similarity suggests that LLMs encode emotional categories in consistent directions across diverse domains and languages.

Next, following the method described in Section 3.3, we used least-squares regression to assess how well the latent spaces of real-emotion datasets align with the hidden-state space of the synthetic dataset. This was performed both on the raw hidden states and on their centered SVD representations. When aligning the 50-dimensional SVD-projected spaces, we observed a mean squared error of 1.79 ± 1.97, indicating strong alignment between the subspaces. To further characterize the learned transformation, we computed its spectral flatness and Frobenius norm. The spectral flatness averaged 2.09 ± 0.39, suggesting that the projection distributes energy across multiple dimensions rather than collapsing onto a low-rank subspace. The Frobenius norm of 7.70 ± 2.67 reflects a moderate overall transformation strength. Together, these results indicate that the alignment transformation is neither degenerate nor axis-dominated, consistent with a view of emotional geometry as distributed and multi-axial. The low regression error, combined with the spectral richness of the mapping, supports the conclusion that emotional subspaces from real-world datasets can be affinely aligned with the synthetic manifold with minimal distortion, providing further evidence for a shared latent emotional geometry across domains.

Having established the alignment between emotional spaces, we next examine whether the internal geometry and topology of emotional representations are consistently preserved across datasets. To assess the geometric consistency of emotional representations, we evaluate a range of distortion metrics that compare the pairwise distances between embeddings of shared emotion labels. These metrics are computed over the full sample distribution of representations.

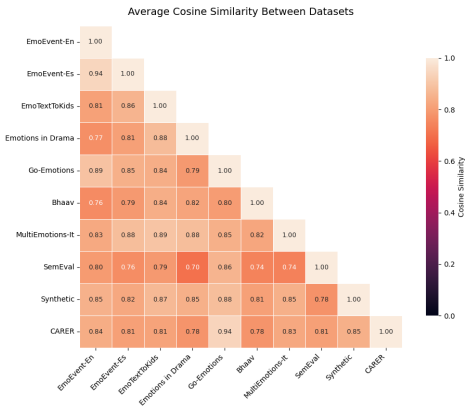

Figure 1: Cosine similarity of emotional centroids between datasets.

We report three classical distortion scores from the multidimensional scaling literature—Kruskal's Stress-1, Stress-2, and Sammon Stress—alongside three additional embedding distortion measures derived from recent work in machine learning geometry. These include average distortion, ℓ_2 distortion, and σ-distortion (a variance-based metric from Chennuru et al. (41)). Results are summarized in Table 1.

We evaluate three classical distortion scores from the multidimensional scaling literature—Kruskal's Stress-1, Stress-2, and Sammon Stress—to assess how well global and local relational geometry is preserved when aligning real-world datasets with the synthetic emotional manifold. While traditional usage defines Stress-1 below 0.2 as acceptable and below 0.1 as strong (42), these thresholds were developed for low-dimensional embeddings (e.g., 2D). In our setting—mapping between high-dimensional hidden states—no canonical thresholds exist, but consistently low scores (e.g., Stress-1 under 0.4 across most datasets) still indicate robust relational preservation. Projection into the 50D synthetic subspace increases stress slightly (e.g., Stress-1 rises from 0.34 ± 0.17 to 0.45 ± 0.16), consistent with expected compression effects, but overall scores remain low enough to support the presence of a coherent emotional geometry.

To complement these, we report three high-dimensional embedding distortion metrics: average distortion, ℓ_2-distortion, and σ-distortion. Values near 1 for the first two suggest that pairwise distances are preserved up to global scaling, while low σ-distortion indicates stable proportionality in relative distances. Most datasets fall within these expected ranges. Two datasets—Go-Emotions and German Drama—are notable exceptions. Go-Emotions involved collapsing a large number of fine-grained emotional categories into broader groups for cross-dataset comparability, which likely introduced label-level ambiguity and elevated distortion. German Drama presents a different challenge: the texts are plays written in early

modern German, with substantial lexical, syntactic, and stylistic divergence from contemporary usage. Language from works like Faust exemplifies this drift—rich in archaisms, poetic structure, and theological allusion—posing a mismatch with the modern German data the model was trained on. We flag these datasets as outliers and report both raw and outlier-excluded averages. In all, these results suggest that the emotional manifold's coarse topology is stable, with distortion emerging mainly in dataset-specific fine structure and post-projection local detail.

Despite modest projection-related distortion, the subspace remains functionally expressive. Linear probes trained on the synthetic 50D space achieve an average accuracy of 0.54 ± 0.11 across datasets that, on average, contain more than four emotion classes. When trained in the full hidden-state space, accuracy improves to 0.62 ± 0.09, indicating that emotional structure is not only geometrically consistent but also linearly decodable across diverse domains.

5 Model Psychology

Having established the external consistency of emotional geometry across datasets, we now turn inward, asking how these emotions are internally structured within the model, and what this reveals about the model's implicit psychological architecture.

The first perspective we examine is neural encoding patterns. Using the ML-AURA method described in Section 3.1, we assess the degree to which individual neurons respond selectively to emotional inputs. In this framework, each neuron is treated as a potential 1-vs-all classifier, and its classification performance is quantified via AUROC. We report results in terms of the percentage of neurons per layer achieving an AUROC above 0.9, interpreted as strong evidence of emotion-selective activation. For the six Ekman emotions, we find that on average, 80% of neurons per layer exceed this threshold, indicating that emotion information is widely and reliably encoded. Among these, sadness (98%) and surprise (96%) show the most widespread selectivity, while fear is more sparsely encoded, with 53% of neurons exceeding the threshold. This does not suggest weak separability, but rather that fewer neurons are strongly specialized for fear relative to other emotions. The non-Ekman emotions—envy, neutral, and excitement—also exhibit high separability, with an average of 88% of neurons surpassing the 0.9 AUROC threshold.

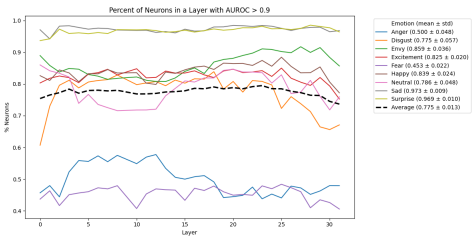

Figure 2: Results of ML-AURA by layer and emotion. Results are in terms of percent of neurons with an AUROC score above 0.9.

When analyzing by architectural component, MLP layers show slightly higher selectivity than attention layers (79% vs. 76.5%). Differentiability fluctuates across depth, with no clear monotonic trend: while the first layer starts at 76% and the final layer ends at 76.3%, several peaks and troughs occur in between, with the highest selectivity observed at layer 26 (79%). These patterns support the conclusion that emotional information is not confined to late layers or specialized regions, but is distributed broadly and redundantly throughout the network. These patterns are visualized by emotion in a layer-by-layer fashion in Figure 2.

To understand how emotions are geometrically represented in the network, we examine their structure within the derived SVD subspace. This subspace provides a low-dimensional lens into the model's internal affective organization. Our first goal is to assess how consistently emotions are arranged along the principal axes across layers and layer types. To this end, we analyze the rank ordering of emotion centroids along each principal component, controlling for possible polarity flips.

We find that the emotional structure is remarkably stable across the model, particularly for the top three components. Across layers, the average Spearman correlation in emotion

rankings is 0.87, 0.83, and 0.80 for PC1, PC2, and PC3, respectively; the corresponding Kendall's Tau values are 0.82, 0.77, and 0.74. These results indicate that, while the magnitude and orientation of the components may shift, their semantic content remains intact.

Even when using a more fine-grained labeling scheme, as in the Go-Emotions dataset, which contains nearly three times as many emotion categories, we observe similar consistency. Rank-order correlations for Go-Emotions along the top three PCs remain high: Spearman values of 0.92, 0.74, and 0.73, and Kendall's Tau of 0.86, 0.68, and 0.68. These findings reinforce the conclusion that the model's emotional manifold is structurally stable, with interpretable axes.

Having established the stability of the SVD subspace across layers and datasets, we now examine the semantic content of the leading principal components. By analyzing the relative positions of emotion centroids along each axis, we infer the underlying affective dimensions implicitly encoded by the model. Figure 3 visualizes the first three emotion axes that we describe below.

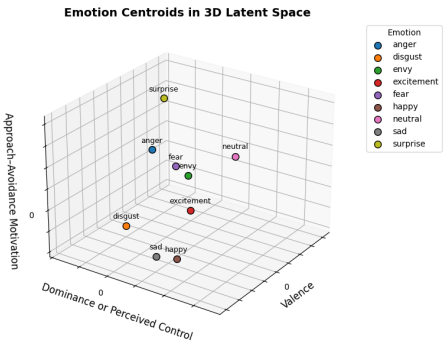

Figure 3: Emotion centroids plotted on the emotional axis found

- PC1 strongly resembles a valence dimension. Emotions such as happy, surprise, and excitement lie at the positive end, while anger and fear occupy the negative end—suggesting a pleasure–displeasure continuum common to many psychological models.
- PC2 appears to reflect dominance or perceived control. Emotions high on this axis (e.g., fear, sadness) are often associated with low control or submission, whereas those at the opposite end (e.g., happy, surprise) may reflect more autonomous or socially detached states.
- PC3 maps onto approach–avoidance motivation. Emotions like excitement, happy, and envy—typically associated with goal-seeking behavior—score high, while anger and fear, linked to avoidance or defensive responses, score low.
- PC4 may correspond to arousal or urgency. Surprise and fear rank highly, consistent with high physiological activation, while happy and neutral lie at the calmer end.
- PC5 appears to encode emotional volatility or temporal intensity. Emotions such as surprise and excitement dominate the upper end, while sadness anchors the low end—indicating a spectrum from sudden, reactive states to more sustained affect.
- PC6, though flatter, may touch on aspects of self-conscious affect. Emotions like envy and anger cluster at one end, and disgust at the other, possibly hinting at a distinction between self-evaluative and other-directed moral emotions. However, its interpretation remains tentative.

These dimensions are not explicitly supervised, but show surface-level resemblance to constructs proposed in affective science, such as valence, arousal, dominance, and approach–avoidance tendencies (cf. (15; 14; 43)). While these alignments are not exact, and many components blend multiple emotional signals, the emergence of such patterns suggests that LLMs may implicitly encode affective distinctions that overlap with long-standing psychological taxonomies. This correspondence invites further investigation into the extent to which models trained solely on text internalize latent emotion structures, and whether these can serve as proxies or tools for understanding affective semantics in language.

Figure 4 visualizes how emotional states are organized in the full hidden-state space using a 2D t-SNE projection of mean-pooled hidden states labeled with their respective emotions. Despite the dimensionality reduction, emotion classes form distinguishable, partially overlapping clusters, with closely related emotions (e.g., happy and excitement) frequently co-localized and others (e.g., fear and joy) appearing more spatially distant. While not

Dataset	Sad	Happy	Fear	Anger	Neutral	Disgust	Envy	Excitement	Surprise
Semeval	14.8 → 99.3	22.6 → 90.5	8.2 → 96.6	23.0 → 59.2	0.0 → 98.9	0.0 → 97.4	0.0 → 100.0	0.0 → 95.8	18.9 → 86.7
CARER (Twitter)	46.3 → 98.8	15.7 → 88.4	7.0 → 89.2	10.3 → 42.7	0.0 → 98.8	0.0 → 94.8	0.0 → 100.0	0.0 → 84.6	7.7 → 77.6
EmoTextToKids (FR)	0.4 → 99.7	4.8 → 96.6	7.2 → 83.3	11.1 → 64.0	19.7 → 98.0	6.2 → 96.4	0.0 → 99.9	0.0 → 95.5	18.6 → 92.2
German Drama	10.4 → 100.0	3.6 → 97.8	8.6 → 50.9	8.6 → 71.9	0.0 → 97.5	0.0 → 97.6	0.0 → 94.6	0.0 → 93.5	9.6 → 72.6
EmoEvents (EN)	24.0 → 98.6	22.7 → 96.8	3.0 → 78.8	41.0 → 79.1	0.0 → 90.4	0.0 → 92.4	0.0 → 100.0	0.0 → 94.5	5.3 → 51.7
EmoEvents (ES)	19.0 → 99.1	21.7 → 92.2	8.7 → 91.6	30.0 → 79.7	0.0 → 89.2	2.0 → 95.1	0.0 → 100.0	0.0 → 95.4	3.3 → 60.2
MultiEmotions-It	9.2 → 98.7	33.2 → 99.7	0.0 → 81.9	21.5 → 51.1	5.7 → 91.8	6.3 → 99.2	0.0 → 100.0	5.1 → 96.0	4.4 → 72.6
Bhaav (Hindi)	8.5 → 100.0	0.0 → 51.0	2.4 → 32.7	0.0 → 59.1	51.5 → 98.3	0.0 → 82.7	0.0 → 99.3	0.0 → 57.9	0.0 → 19.5*
GoEmotions	3.4 → 99.2	15.5 → 89.5	8.2 → 8.2*	8.2 → 50.4	4.0 → 97.2	0.0 → 68.4	0.0 → 68.9	0.0 → 82.5	0.0 → 39.8*

Table 2: Top-1 prediction rates before and after learned steering for each target emotion across datasets. Each cell shows *baseline* → *post-steering* accuracy. *Indicates failure cases where target emotion remained under 10%.

all boundaries are sharp, the observed structure reinforces our earlier findings: emotional information is embedded in a distributed yet semantically coherent geometry.

Together, the distributed AUROC patterns, stable subspace directions, interpretable principal components, and emergent clustering structure suggest that LLMs encode emotion not as isolated tags, but as coherent, multidimensional structures—akin to a learned affective manifold.

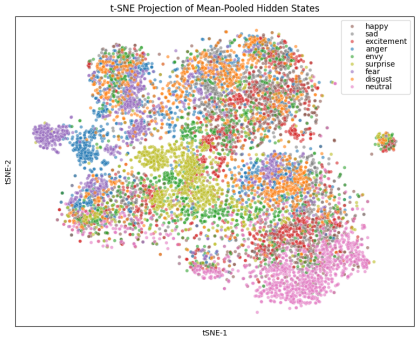

Figure 4: Plot of projected emotions in TSNE-space.

6 Steerability and the Limits of Control

Prior work on emotional steering in LLMs focuses primarily on shifting the emotional tone of generated text. (44) and (45) learn vectors to modify output valence or categorical emotion. More recently, (23) attempts to steer internal emotional representations but collapses emotion into a binary positive/negative axis, achieving valence shifts 53.5% of the time. In contrast, we aim for fine-grained control over the model's internal perception of emotion across a full categorical space, while preserving semantic content.

We train a learned module that operates within the previously constructed SVD-based emotional subspace. For each emotion, we select all layers and sublayers where adding the centroid direction to same-emotion hidden states improves 1-vs-all classification AUROC beyond a fixed threshold. These layers are used for steering and serve as a proxy for the more challenging task of controllable emotional manipulation. At each selected layer, the model's hidden state is projected into the subspace using the precomputed basis. The projected representation is passed through a one-layer MLP with GELU activation to compute a shift, which is then mapped back into hidden-state space and added residually. The MLP is trained to steer the model's internal representation to favor the target emotion token when prompted.

We define the overall training objective as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{sem}}$$

where $\mathcal{L}_{\text{sem}}$ preserves semantic meaning and $\mathcal{L}_{\text{token}}$ enforces perceptual control.

Semantic Preservation. The semantic consistency loss combines cosine and ℓ_2 distance between the original and shifted final-layer hidden states:

$$\mathcal{L}_{\text{sem}} = (1 - \cos(h_{\text{base}}, h_{\text{shifted}})) + \gamma \cdot \|h_{\text{base}} - h_{\text{shifted}}\|_2$$

Emotion Control. To ensure accurate emotion classification, we combine a standard cross-entropy loss with a token-level margin loss. The margin loss enforces that the logit for the

target emotion token e_i exceeds its synonyms s_i by a margin m_1 (0.5), and that both exceed all other emotions e_j and their synonyms by m_2 (10):

$$\mathcal{L}_{\text{margin}} = \max(0, m_1 - (\log p_{\text{e}_i} - \log p_{\text{s}_i})) + \max(0, m_2 - (\log p_{\text{s}_i} - \log p_{\text{e}_j}))$$

To prevent the model from optimizing by suppressing unrelated tokens, we weight the loss for emotion tokens more heavily in $\mathcal{L}_{\text{CE}}$:

$$\mathcal{L}_{\text{token}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{margin}}$$

We optimize the objective using AdamW with learning rate 1e-3 and weight decay 1e-2, using a cosine schedule with 50 warm-up steps. Steering uses the top 40 dimensions of the centered SVD-derived emotional subspace. The learned module is trained independently for each target emotion across all selected steering layers, using supervision from emotion-token prompts and hidden-state consistency targets. At evaluation, token sampling is performed with temperature 0 for determinism. Ablations on loss components and architecture are provided in the supplementary material.

Table 2 shows that our learned steering approach achieves consistent and accurate control over internal emotional representations across a diverse set of languages and datasets. For core emotions like sadness, anger, and excitement, post-steering accuracy typically exceeds 90%. Performance is robust even in multilingual settings, with particularly strong results in French, German, and Italian. Steerability remains high for many emotions in Hindi—a lower-resource language—but control over fear and surprise is less reliable, suggesting that lexical sparsity and data imbalance remain limiting factors for certain emotions in under-resourced settings.

7 Conclusion

Using a combination of probing, alignment, and causal intervention techniques, this work shows that emotional representations in LLMs are directionally consistent across layers, datasets, and languages. We find that emotions cluster in coherent, low-dimensional subspaces whose structure is stable across architectural depth and transferable across linguistic and cultural domains. The leading axes of this space correspond to psychologically interpretable dimensions, despite no explicit supervision. These emotional directions are not confined to isolated neurons or layers but are distributed and redundant, supporting high linear separability even under one-vs-all probing. Alignment experiments further reveal that the synthetic and real-world emotion spaces can be matched with minimal distortion, and linear probes trained in one domain generalize well to others. Together, these findings suggest that LLMs internalize a structured latent affective manifold during pretraining.

Crucially, this representational structure is not merely interpretable but also controllable. Our learned intervention module achieves accurate and emotion-specific steering across languages, reliably shifting the model's internal affective state toward the desired target. Steering is especially effective for basic emotions like sadness, anger, and fear, even in low-resource settings. However, control over more nuanced categories such as envy and excitement remains inconsistent, particularly in Hindi, highlighting the residual challenges of lexical sparsity and affective ambiguity.

These findings offer a structured account of how LLMs represent and modulate emotion. Future work should extend this analysis to multimodal models, investigating whether shared affective subspaces emerge across language, vision, and speech, and whether emotional representations in one modality can steer or constrain perception in another. Such models may yield a richer, more disentangled affective geometry, enabling both deeper interpretability and more naturalistic emotional reasoning.

Limitations While our interventions demonstrate strong control over internal emotional perception, they do not address downstream generation—leaving open the question of how internal shifts affect model outputs in practice. Additionally, although the steering directions yield causal effects, they are derived from statistically constructed subspaces

without formal guarantees of disentanglement. Finally, performance on certain emotions in low-resource settings (e.g., fear and surprise in Hindi) suggests limitations imposed by lexical sparsity and pretraining data imbalance, particularly for culturally specific or less frequently expressed affective states.

References

[1] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics*, vol. 3, no. 2, pp. 143–157, 2009.

[2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[3] R. Wadawadagi and V. Pagi, "Sentiment analysis with deep neural networks: comparative study and performance assessment," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 6155–6195, 2020.

[4] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[5] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu, "Emotional intelligence of large language models," *Journal of Pacific Rim Psychology*, vol. 17, p. 18344909231213958, 2023.

[6] J.-t. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, and M. R. Lyu, "Apathetic or empathetic? evaluating llms'emotional alignments with humans," in *Advances in Neural Information Processing Systems* (A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, eds.), vol. 37, pp. 97053–97087, Curran Associates, Inc., 2024.

[7] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is chatgpt equipped with emotional dialogue capabilities?," *arXiv preprint arXiv:2304.09582*, 2023.

[8] S. Shah, S. Reddy, and P. Bhattacharyya, "Retrofitting light-weight language models for emotions using supervised contrastive learning," *arXiv preprint arXiv:2310.18930*, 2023.

[9] J. Broekens, B. Hilpert, S. Verberne, K. Baraka, P. Gebhard, and A. Plaat, "Fine-grained affective processing capabilities emerging from large language models," in *2023 11th international conference on affective computing and intelligent interaction (ACII)*, pp. 1–8, IEEE, 2023.

[10] B. Reichman, A. Avsian, and L. Heck, "Reading with intent–neutralizing intent," *arXiv preprint arXiv:2501.03475*, 2025.

[11] R. Vinay, G. Spitale, N. Biller-Andorno, and F. Germani, "Emotional manipulation through prompt engineering amplifies disinformation generation in ai large language models," *arXiv preprint arXiv:2403.03550*, 2024.

[12] P. Ekman, T. Dalgleish, and M. Power, "Basic emotions," *San Francisco, USA*, 1999.

[13] R. Plutchik, *The emotions*. University Press of America, 1991.

[14] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, pp. 261–292, 1996.

[15] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[16] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett, "The brain basis of emotion: a meta-analytic review," *Behavioral and brain sciences*, vol. 35, no. 3, pp. 121–143, 2012.

[17] K. Vytal and S. Hamann, "Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis," *Journal of cognitive neuroscience*, vol. 22, no. 12, pp. 2864–2885, 2010.

[18] A. Celeghin, M. Diano, A. Bagnis, M. Viola, and M. Tamietto, "Basic emotions in human neuroscience: neuroimaging and beyond," *Frontiers in psychology*, vol. 8, p. 1432, 2017.

[19] N. Yongsatianchot, T. Thejll-Madsen, and S. Marsella, "What's next in affective modeling? large language models," in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–7, IEEE, 2023.

[20] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," *arXiv preprint arXiv:1912.02164*, 2019.

[21] S. Buechel, L. Modersohn, and U. Hahn, "Towards label-agnostic emotion embeddings," *arXiv preprint arXiv:2012.00190*, 2020.

[22] X. Wang and C. Zong, "Distributed representations of emotion categories in emotion space," 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)*, pp. 2364–2375, 2021.

[23] O. Hollinsworth, C. Tigges, A. Geiger, and N. Nanda, "Language models linearly represent sentiment," in *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 58–87, 2024.

[24] Y. Zhang, W. Chen, R. Zhang, and X. Zhang, "Representing affect information in word embeddings," *Experiments in Linguistic Meaning*, vol. 2, pp. 310–321, 2023.

[25] F. Bianchi, D. Nozza, and D. Hovy, "Xlm-emo: Multilingual emotion prediction in social media text," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pp. 195–203, 2022.

[26] L. De Bruyne, P. Singh, O. De Clercq, E. Lefever, and V. Hoste, "How language-dependent is emotion detection? evidence from multilingual bert," in *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pp. 76–85, 2022.

[27] X. Suau, P. Delobelle, K. Metcalf, A. Joulin, N. Apostoloff, L. Zappella, and P. Rodriguez, "Whispering experts: Neural interventions for toxicity mitigation in language models," in *Forty-first International Conference on Machine Learning*, 2024.

[28] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," *arXiv preprint arXiv:2012.13255*, 2020.

[29] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[30] T. Lizzo and L. Heck, "UNLEARN efficient removal of knowledge in large language models," in *Findings of the Association for Computational Linguistics: NAACL 2025*, (Albuquerque, New Mexico), pp. 7257–7268, Association for Computational Linguistics, Apr. 2025.

[31] L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodolà, "Relative representations enable zero-shot latent space communication (2023)," *arXiv preprint arXiv:2209.15430*, 2023.

[32] Z. Lähner and M. Moeller, "On the direct alignment of latent spaces," in *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pp. 158–169, PMLR, 2024.

[33] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[34] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds.), (Brussels, Belgium), pp. 3687–3697, Association for Computational Linguistics, Oct.-Nov. 2018.

[35] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14: Affective text," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (E. Agirre, L. Màrquez, and R. Wicentowski, eds.), (Prague, Czech Republic), pp. 70–74, Association for Computational Linguistics, June 2007.

[36] F. Plaza-del-Arco, C. Strapparava, L. A. Ureña-Lopez, and M. T. Martin-Valdivia, "EmoEvent: A Multilingual Emotion Corpus based on different Events," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 1492–1498, European Language Resources Association, May 2020.

[37] K. Dennerlein, T. Schmidt, and C. Wolff, "Computational emotion classification for genre corpora of german tragedies and comedies from 17th to early 19th century," *Digital Scholarship in the Humanities*, vol. 38, pp. 1466–1481, 07 2023.

[38] Y. Kumar, D. Mahata, S. Aggarwal, A. Chugh, R. Maheshwari, and R. R. Shah, "Bhaav-a text corpus for emotion analysis from hindi stories," *arXiv preprint arXiv:1910.04073*, 2019.

[39] R. Sprugnoli *et al.*, "Multiemotions-it: a new dataset for opinion polarity and emotion analysis for italian," in *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pp. 402–408, Accademia University Press, 2020.

[40] A. Étienne, D. Battistelli, and G. Lecorvé, "Emotion identification for french in written texts: Considering modes of emotion expression as a step towards text complexity analysis," in *Proceedings of the 14th ACL Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA)*, 2024.

[41] L. Chennuru Vankadara and U. von Luxburg, "Measures of distortion for machine learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[42] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[43] R. J. Davidson, "Cerebral asymmetry, emotion, and affective style.," 1995.

[44] N. Subramani, N. Suresh, and M. E. Peters, "Extracting latent steering vectors from pretrained language models," *arXiv preprint arXiv:2205.05124*, 2022.

[45] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, *et al.*, "Representation engineering: A top-down approach to ai transparency," *arXiv preprint arXiv:2310.01405*, 2023.

[46] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[47] M. AI, "Un ministral, des ministraux," October 2024. Accessed: 2025-05-20.

[48] W. Xu, A. Ritter, and R. Grishman, "Gathering and generating paraphrases from twitter with application to normalization," in *Proceedings of the sixth workshop on building and using comparable corpora*, pp. 121–128, 2013.

[49] W. Xu, A. Ritter, C. Callison-Burch, W. B. Dolan, and Y. Ji, "Extracting lexically divergent paraphrases from twitter," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 435–448, 2014.

A Reproduction of Emotional Universality on Other Models

We extend the analysis from Section 4 to the Qwen 3.0-8B (46) and Mistral-8B (47) models, using the same datasets and evaluation methods. Both models support similar language coverage; however, Mistral-8B lacks support for Hindi.

We begin by examining centroid cosine similarities across all shared emotions, layers, and sublayer types. Qwen-3.0 yields a mean similarity of 0.86 ± 0.04, while Mistral-8B achieves 0.92 ± 0.04. Figures 5 and 6 show per-dataset similarities. These values, higher than those observed for LLaMA, suggest that emotional categories are encoded in consistent directions across domains and languages in all three models.

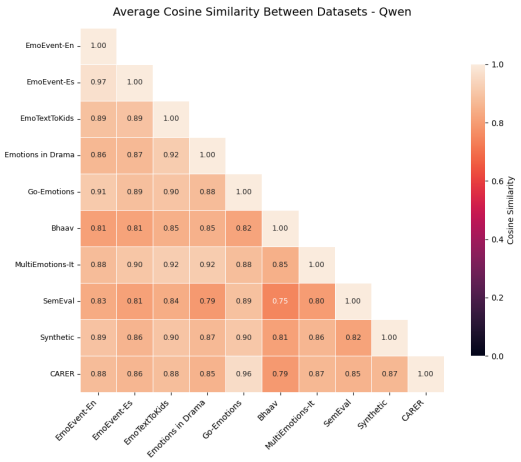

Figure 5: Cosine similarity of emotional centroids between datasets for Qwen.

Next, we used least-squares regression to assess how well the latent spaces of real-emotion datasets align with the synthetic emotional manifold. For each model, we aligned 50-dimensional SVD-projected spaces between different datasets within the same model. Qwen-3.0 yields an average MSE of 6.23 ± 9.55, while Mistral-8B yields 1.44 ± 1.47. The higher error in Qwen reflects greater variability in cross-dataset alignment, though both models maintain sufficiently low distortion to support an underlying shared geometry. Spectral flatness is 2.11 ± 0.38 for Qwen and 2.22 ± 0.44 for Mistral, indicating that the alignment transformations retain multi-dimensional structure rather than collapsing into low-rank subspaces. Frobenius norms are similar—7.64 ± 1.45 (Qwen) and 7.60 ± 0.91 (Mistral)—suggesting comparable transformation magnitudes. These results support the conclusion that, across datasets, emotional spaces within each model are structurally coherent and affinely alignable.

Tables 3 and 4 summarize stress and distortion metrics for Qwen 3.0 and Mistral-8B. Each model has a distinct set of outlier datasets. For Qwen 3.0, Go-Emotions and CARER are outliers—Go-Emotions likely for the same aggregation-related reasons as in LLaMA. For Mistral, outliers include EmoEvent (EN and ES), German Drama, Bhaav, and CARER. German Drama again likely reflects the difficulty of the archaic and out-of-distribution text in the dataset. Mistral's lack of Hindi support plausibly explains the Bhaav results. EmoEvent and CARER both derive from Twitter, which is known to exhibit platform-specific linguistic variation (48; 49). As Mistral's pretraining corpus is not publicly disclosed, it remains unclear whether these divergences reflect distributional mismatch or deeper structural disalignment.

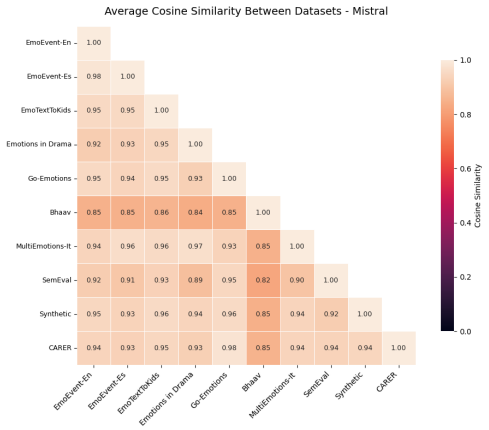

Figure 6: Cosine similarity of emotional centroids between datasets for Mistral.

Qwen 3.0. Qwen shows higher stress metrics than LLaMA across both full and 50D spaces (e.g., Stress-1: 0.43 ± 0.62 vs. 0.34 ± 0.17), indicating less faithful preservation of relational structure. Distortion scores (*average*, ℓ_2, and σ) are modestly worse but still broadly

Dataset	Stress-1 ↓	Stress-2 ↓	Sammon ↓	Avg Dist ↓	ℓ_2 ↓	σ ↓	Probe Acc. ↑
Go-Emotions*	0.41 ± 0.18	0.20 ± 0.22	*	*	*	*	0.31 ± 0.10
CARER (Twitter)*	0.41 ± 0.15	0.19 ± 0.17	*	*	*	*	0.31 ± 0.08
SemEval	0.48 ± 0.40	0.39 ± 1.58	0.65 ± 4.25	1.41 ± 0.63	1.49 ± 0.98	0.09 ± 0.14	0.42 ± 0.15
EmoEvent (EN)	0.33 ± 0.13	0.13 ± 0.14	0.16 ± 0.29	1.09 ± 0.17	1.16 ± 0.29	0.12 ± 0.24	0.48 ± 0.14
EmoEvent (ES)	0.33 ± 0.13	0.12 ± 0.15	0.15 ± 0.36	1.05 ± 0.15	1.12 ± 0.30	0.14 ± 0.39	0.46 ± 0.16
Bhaav (Hindi)	0.37 ± 0.11	0.15 ± 0.12	0.16 ± 0.23	0.90 ± 0.16	0.98 ± 0.30	0.16 ± 0.28	0.37 ± 0.08
German Drama*	0.52 ± 0.79	0.89 ± 9.44	*	*	*	*	0.27 ± 0.08
MultiEmotions-It	0.60 ± 1.33	2.14 ± 27.05	4.76 ± 65.12	1.34 ± 0.33	1.66 ± 3.48	0.74 ± 8.67	0.42 ± 0.11
EmoTextToKids (FR)*	0.36 ± 0.13	0.14 ± 0.14	*	*	*	*	0.40 ± 0.12
Average (Full-Space)	0.43 ± 0.62	0.57 ± 10.87	1.17 ± 26.73	1.17 ± 0.37	1.31 ± 1.67	0.24 ± 3.40	0.40 ± 0.09
Average (50D-Space)	0.46 ± 0.54	0.50 ± 9.65	1.27 ± 31.91	0.91 ± 0.39	1.07 ± 1.90	0.36 ± 4.72	0.40 ± 0.09

Table 3: Per-dataset distortion metrics and probe accuracy in the synthetic emotional subspace for Qwen 3.0. Lower distortion values indicate greater geometric consistency. Probe accuracy reflects how well emotion labels can be decoded via a linear probe trained on the synthetic manifold. Datasets marked with * were identified as outliers; asterisks in cells indicate anomalously high values omitted for readability.

comparable, suggesting reasonable preservation of pairwise distances and topological features. Elevated variance in datasets like German Drama and MultiEmotions-It inflates Qwen's averages, though most non-outlier datasets remain stable. However, Qwen lags behind LLaMA in probe accuracy, suggesting that the emotional manifold it induces encodes affective distinctions in a less linearly decodable—or more nonlinearly entangled—form. Qwen also exhibits a greater number of outlier datasets than LLaMA, underscoring reduced geometric consistency across both linguistic and domain boundaries.

Mistral-8B. Mistral likewise shows increased stress (Stress-1: $0.43 \pm 0.16 \rightarrow 0.51 \pm 0.14$), with a consistent rise in compression-related error compared to LLaMA. Nonetheless, distortion metrics remain stable and close to 1, especially in the ℓ_2 and average distortion columns, indicating that the underlying emotional manifold is preserved up to global scaling. Note that several high-distortion outliers were excluded, so these metrics likely reflect a cleaner subset. Overall, despite the increase in stress, Mistral maintains a coherent and decodable emotional geometry.

Inter-Model. Having shown that the synthetic dataset hidden-state space serves as a "canonical" emotion space within each model, we next mapped these spaces across models. While some layers proved incompatible, most aligned successfully. Table 5 reports distortion and stress scores, with the number of excluded outlier layers (out of 224) in parentheses. Only for Qwen's Sammon Stress metric were more than 50% of layers excluded. Most incompatibilities occur in the key and value projection layers, suggesting model-specific differences in how attention mechanisms process emotional content, whereas semantic representations remain more aligned.

Overall, the remaining layers exhibit strong cross-model similarity, with Mistral showing greater compatibility with LLaMA than Qwen. Even without excluding layers, inter-model mappings achieve low MSE (0.03 ± 0.05), indicating that the spaces are, in principle, linearly alignable. However, the associated spectral flatness and Frobenius norms are high, implying that these transformations are complex and energetically distributed. This combination—low distortion, low MSE, high transformation complexity—suggests that while the models encode emotion with consistent structure, they do so using different internal bases.

The emotional structure is preserved across all models, with Mistral showing the highest consistency and Qwen slightly lower than LLaMA. Across layers, average Spearman correlations in emotion rankings along PC1, PC2, and PC3 are 0.94, 0.83, and 0.79 for Mistral, and 0.75, 0.81, and 0.76 for Qwen. Corresponding Kendall's Tau values are 0.91, 0.78, 0.74 (Mistral) and 0.70, 0.76, 0.71 (Qwen).

Using the fine-grained Go-Emotions labels yields consistent results. For Mistral, Spearman values are 0.93, 0.74, and 0.71; for Qwen, 0.85, 0.69, and 0.70. Kendall's Tau values are 0.90, 0.68, 0.66 (Mistral) and 0.79, 0.64, 0.65 (Qwen).

Dataset	Stress-1 ↓	Stress-2 ↓	Sammon ↓	Avg Dist ↓	ℓ_2 ↓	σ ↓	Probe Acc. ↑
Go-Emotions	0.44 ± 0.17	0.22 ± 0.36	0.30 ± 0.75	1.17 ± 0.23	1.30 ± 0.40	0.20 ± 0.16	0.32 ± 0.05
CARER (Twitter)*	0.46 ± 0.11	0.22 ± 0.11	*	*	*	*	0.30 ± 0.10
SemEval	0.38 ± 0.21	0.19 ± 0.53	0.24 ± 0.91	1.13 ± 0.19	1.22 ± 0.39	0.16 ± 0.52	0.37 ± 0.11
EmoEvent (EN)*	0.40 ± 0.11	0.17 ± 0.11	*	*	*	*	0.49 ± 0.11
EmoEvent (ES)*	0.44 ± 0.12	0.20 ± 0.12	*	*	*	*	0.47 ± 0.13
Bhaav (Hindi)*	0.42 ± 0.11	0.19 ± 0.10	*	*	*	*	0.32 ± 0.08
German Drama*	0.50 ± 0.25	0.32 ± 0.98	*	*	*	*	0.26 ± 0.09
MultiEmotions-It	0.48 ± 0.13	0.25 ± 0.14	0.30 ± 0.22	1.16 ± 0.11	1.30 ± 0.19	0.24 ± 0.27	0.50 ± 0.11
EmoTextToKids (FR)	0.42 ± 0.11	0.19 ± 0.12	0.23 ± 0.23	1.13 ± 0.12	1.24 ± 0.23	0.19 ± 0.16	0.37 ± 0.09
Average (Full-Space)	0.44 ± 0.16	0.22 ± 0.40	0.27 ± 0.61	1.15 ± 0.17	1.26 ± 0.32	0.20 ± 0.31	0.37 ± 0.14
Average (50D-Space)	0.52 ± 0.14	0.29 ± 0.42	0.34 ± 0.63	1.03 ± 0.21	1.19 ± 0.37	0.31 ± 0.34	0.41 ± 0.11

Table 4: Per-dataset distortion metrics and probe accuracy in the synthetic emotional subspace for Mistral. Lower distortion values indicate greater geometric consistency. Probe accuracy reflects how well emotion labels can be decoded via a linear probe trained on the synthetic manifold. Datasets marked with * were identified as outliers; asterisks in cells indicate anomalously high values omitted for readability.

Model	Stress-1 ↓	Stress-2 ↓	Sammon ↓	Avg Dist ↓	ℓ_2 ↓	σ ↓
Mistral → LLaMA	0.49 ± 0.31 (0)	0.26 ± 0.38 (2)	0.34 ± 0.48 (59)	0.95 ± 0.67 (58)	0.89 ± 0.40 (67)	0.21 ± 0.26 (66)
Qwen → LLaMA	1.33 ± 1.16 (39)	0.88 ± 0.95 (73)	1.02 ± 0.97 (115)	1.99 ± 1.27 (98)	1.96 ± 1.26 (109)	0.16 ± 0.24 (66)

Table 5: Distortion metrics for mapping each model's SVD-projected emotional manifold back to LLaMA. Values shown are outlier-excluded means ± standard deviations, with the number of excluded layers indicated in parentheses. Lower values indicate better preservation of geometry.

This structure is not only stable within models, but also consistent across them in terms of relative emotion positioning. Spearman correlations between Qwen and LLaMA are 0.75, 0.83, and 0.77 (PC1–PC3), with corresponding Kendall's Tau of 0.69, 0.77, and 0.71. Mistral–LLaMA shows similar values: 0.76, 0.84, and 0.79 (Spearman), and 0.69, 0.77, and 0.71 (Kendall). These high rank-order correlations suggest that the emotional geometry described in Section 5 reflects a shared conceptual structure across models. However, results from the inter-model alignment analysis indicate that these shared structures are embedded in distinct internal coordinate systems, requiring high-complexity transformations to align. Thus, while the emotional manifolds are topologically consistent, their parameterizations remain model-specific—likely shaped by architectural and pretraining differences.

Figures 7 and 8 apply the ML-AURA method from Section 3.1 to assess neuron-level selectivity for emotional inputs in a 1-vs-all classification setting. These reproduce the ML-AURA results presented in Section 5 using the Qwen and Mistral models. Consistent with LLaMA, both models show that sadness and surprise elicit the most widespread neuron selectivity, while fear and anger are more sparsely represented—fewer neurons exceed the AUROC threshold of 0.9 for these emotions. This suggests that, across architectures, a greater number of neurons specialize in distinguishing sadness and surprise from other emotions. Nonetheless, neurons in all three models exhibit reliable separation across emotional categories, indicating distributed but consistent encoding.

B Ablations for Emotional Steering

In Section 6, we introduced a method for steering how LLMs internally represent and perceive emotion. This appendix presents ablation studies identifying which components are essential for successful steering. We evaluate the impact of: (1) the number of steering dimensions in the SVD subspace, (2) the presence of the GELU nonlinearity, (3) the use of synonyms in the loss function, (4) the weight of the target-token term in the cross-entropy loss, (5) individual components of the semantic similarity loss, (6) the structure of the margin loss, and (7) the choice of target layers for intervention.

To reduce evaluation cost while capturing variance in performance, we selected three emotion-dataset pairs representing high, moderate, and poor performance in the main re-

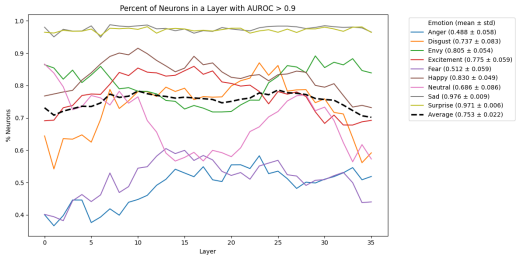

Figure 7: Results of ML-AURA by layer and emotion for Qwen3-8B. Results are in terms of percent of neurons with an AUROC score above 0.9.

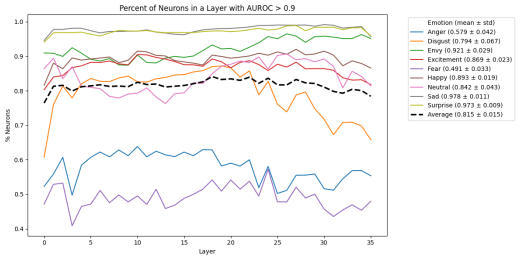

Figure 8: Results of ML-AURA by layer and emotion for Mistral. Results are in terms of percent of neurons with an AUROC score above 0.9.

sults: sad (EmoTextToKids), anger (CARER), and fear (Bhaav). All ablations were conducted using these fixed emotion-dataset combinations.

Table 6 presents the effect of varying the number of steering dimensions R in the SVD subspace. We observe that extremely low ranks (e.g., $R = 1$) fail catastrophically, while small ranks like $R = 2$ surprisingly succeed on all three emotion-dataset pairs. However, this success is likely fragile—intermediate values such as $R = 15$ and $R = 10$ show inconsistent behavior, with performance collapses in some cases. As rank increases, steering generally improves, peaking around $R = 20$, which achieves near-perfect or perfect steering across all settings. Beyond this point, gains saturate or regress, particularly for fear, suggesting diminishing returns or overparameterization. We adopt $R = 20$ as the best-performing and most stable configuration.

Tables 7 and 8 examines the effect of varying the margin weights m_1 and m_2, which define separation constraints in the semantic loss. The margin m_1 enforces a minimum distance between the target emotion token and its synonyms, preventing collapse and encouraging meaningful local structure. We observe that performance remains relatively stable across m_1 values, though some instability appears for *fear*, suggesting mild sensitivity. In contrast, m_2 enforces separation between the target emotion token and all other emotion tokens (and their synonyms). Steering is highly sensitive to this margin: low m_2 values consistently fail, while performance improves monotonically as m_2 increases. At $m_2 = 20$, all emotion-dataset pairs steer successfully, indicating that strong inter-class separation is essential. We adopt $m_1 = 0.75$, $m_2 = 20$ as the best-performing configuration.

Table 9 shows the effect of varying the weight of the cross-entropy loss applied to the target emotion token and its synonyms. Lower weights lead to poor steering, particularly on *fear*, while higher values generally improve performance. The best overall results are observed at a weight of 25, suggesting that strongly emphasizing the generation of target emotion tokens is necessary for effective control.

Table 10 reports ablations over discrete architectural and training choices. Removing the GELU activation severely degrades performance across all tasks, indicating that nonlinearity is critical for steering. Omitting bias has a moderate effect, while removing synonyms from the loss function leads to failure on *fear*, suggesting their inclusion helps generalize the steering signal. Within the semantic similarity loss, the delta-norm and cosine components can be individually removed with limited degradation, but removing the full loss results in collapse—suggesting a synergistic effect where both components reinforce each other to guide the model's representation. The emotion margin loss is also crucial—its removal results in failure across all settings. Finally, applying steering across all layers performs worse than selectively targeting layers based on alignment with the emotion direction, underscoring the importance of precise and informed intervention over blanket modification.

Ablation Target	Sad (EmoTextToKids)	Anger (CARER)	Fear (Bhaav)
R=1	0.4 → 0	7.0 → 100	2.4 → 0
R=2	0.4 → 99.8	7.0 → 100	2.4 → 100
R=3	0.4 → 37.9*	7.0 → 100	2.4 → 100
R=5	0.4 → 100	7.0 → 2.4*	2.4 → 29.3*
R=10	0.4 → 64.3	7.0 → 99.6	2.4 → 44.2
R=15	0.4 → 30.8	7.0 → 17.6*	2.4 → 22.2*
R=20	0.4 → 93.2	7.0 → 99.1	2.4 → 81.3
R=25	0.4 → 84.8	7.0 → 96.0	2.4 → 6.3*
R=30	0.4 → 85.4	7.0 → 68.4	2.4 → 65.2
R=35	0.4 → 84.8	7.0 → 76.3	2.4 → 46.4
R=40	0.4 → 99.7	7.0 → 42.7	2.4 → 32.7*
R=45	0.4 → 95.4	7.0 → 51.0	2.4 → 61.1
R=50	0.4 → 99.2	7.0 → 99.3	2.4 → 27.2*
R=100	0.4 → 94.2	7.0 → 99.2	2.4 → 30.3*

Table 6: Ablation for number of steering directions. Top-1 prediction rates before and after steering under ablation conditions for selected emotion-dataset pairs. Each cell shows *baseline* → *post-ablation* accuracy. *Indicates failure cases where target emotion is not the most predicted Top-1 class.

Ablation Target	Sad (EmoTextToKids)	Anger (CARER)	Fear (Bhaav)
m1=0.1	0.4 → 99.2	7.0 → 66.7	2.4 → 37.3*
m1=0.25	0.4 → 97.8	7.0 → 99.0	2.4 → 27.1*
m1=0.5	0.4 → 99.7	7.0 → 42.7	2.4 → 32.7*
m1=0.75	0.4 → 96.1	7.0 → 99.8	2.4 → 22.2*
m1=1	0.4 → 93.3	7.0 → 65.4	2.4 → 37.25*

Table 7: Ablation for target synonym margin. Top-1 prediction rates before and after steering under ablation conditions for selected emotion-dataset pairs. Each cell shows *baseline* → *post-ablation* accuracy. *Indicates failure cases where target emotion is not the most predicted Top-1 class.

Ablation Target	Sad (EmoTextToKids)	Anger (CARER)	Fear (Bhaav)
m2=1	0.4 → 31.2	7.0 → 29.3	2.4 → 4.0*
m2=2	0.4 → 51.9	7.0 → 99.0	2.4 → 3.4*
m2=5	0.4 → 79.2	7.0 → 96.1	2.4 → 22.8*
m2=10	0.4 → 99.7	7.0 → 42.7	2.4 → 32.7*
m2=15	0.4 → 100	7.0 → 99.6	2.4 → 97.1
m2=20	0.4 → 99.6	7.0 → 100	2.4 → 100

Table 8: Ablation for margin between target and non-target classes. Top-1 prediction rates before and after steering under ablation conditions for selected emotion-dataset pairs. Each cell shows *baseline* → *post-ablation* accuracy. *Indicates failure cases where target emotion is not the most predicted Top-1 class.

Ablation Target	Sad (EmoTextToKids)	Anger (CARER)	Fear (Bhaav)
CE Loss Weight=1	0.4 → 96.3	7.0 → 95.1	2.4 → 1.4*
CE Loss Weight=2	0.4 → 92.8	7.0 → 54.2	2.4 → 6.0*
CE Loss Weight=5	0.4 → 94.9	7.0 → 98.7	2.4 → 12.7*
CE Loss Weight=10	0.4 → 80.0	7.0 → 65.7	2.4 → 56.2
CE Loss Weight=15	0.4 → 89.8	7.0 → 85.2	2.4 → 56.7
CE Loss Weight=20	0.4 → 99.7	7.0 → 42.7	2.4 → 32.7*
CE Loss Weight=25	0.4 → 98.0	7.0 → 99.8	2.4 → 93.2
CE Loss Weight=30	0.4 → 94.4	7.0 → 91.7	2.4 → 73.3

Table 9: Ablation for cross-entropy loss weight for emotion tokens. Top-1 prediction rates before and after steering under ablation conditions for selected emotion-dataset pairs. Each cell shows *baseline* → *post-ablation* accuracy. *Indicates failure cases where target emotion is not the most predicted Top-1 class.

Ablation Target	Sad (EmoTextToKids)	Anger (CARER)	Fear (Bhaav)
Baseline	0.4 → 99.7	7.0 → 42.7	2.4 → 32.7*
No GELU	0.4 → 25.9*	7.0 → 11.0*	2.4 → 1.3*
No Bias	0.4 → 88.2	7.0 → 91.7	2.4 → 26.9*
No Synonyms	0.4 → 98.9	7.0 → 99.3	2.4 → 15.9*
No Semantic Loss	0.4 → 30.2*	7.0 → 88.9	2.4 → 100
No Cosine Loss	0.4 → 74.3	7.0 → 100	2.4 → 76.3
No Delta-Norm Loss	0.4 → 100	7.0 → 97.7	2.4 → 100
No Emotion Margin Loss	0.4 → 23.9	7.0 → 13.3*	2.4 → 0.6*
Target Layers=All	0.4 → 66.1	7.0 → 64.9	2.4 → 12.9*

Table 10: Top-1 prediction rates before and after steering under various ablation conditions for selected emotion-dataset pairs. Each cell shows *baseline* → *post-ablation* accuracy. *Indicates failure cases where target emotion is not the most predicted Top-1 class.