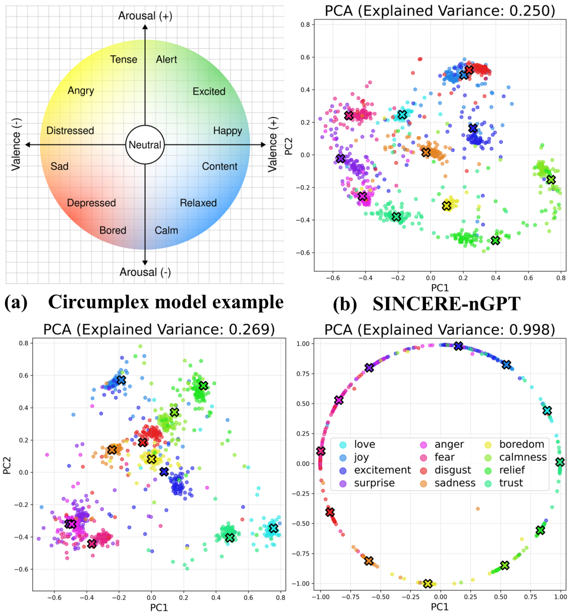
Mapping the Circumplex of Affect: Geometric Analysis of Emotion Representations via Hyperspherical Contrastive Learning

Yusuke Yamauchi[1] **Akiko Aizawa**[2,1]
[1]The University of Tokyo, [2]National Institute of Informatics
{y_yamauchi@is.s.u-tokyo.ac.jp, aizawa@nii.ac.jp}

Abstract

Psychological research has long utilized circumplex models to structure emotions, placing similar emotions adjacently and opposing ones diagonally. Although frequently used to interpret deep learning representations, these models are rarely directly incorporated into the representation learning of language models, leaving their geometric validity unexplored. This paper proposes a method to induce circular emotion representations within language model embeddings via contrastive learning on a hypersphere. We show that while this circular alignment offers superior interpretability and robustness against dimensionality reduction, it underperforms compared to conventional designs in high-dimensional settings and fine-grained classification. Our findings elucidate the trade-offs involved in applying psychological circumplex models to deep learning architectures. Our code is available at https://github.com/yama11235/EmpiricalCircumplexModel



(a) **Circumplex model example** (b) **SINCERE-nGPT** (c) **SoftCSE-nGPT** (d) **CircularCSE-nGPT**

Figure 1: (a) An example of the psychological circumplex model of emotion[1]. (b)(c)(d) PCA plots of the embeddings from the models trained in this study.

1 Introduction

In recent years, the mechanistic interpretability of large language models (LLMs) has emerged as a pivotal field for ensuring AI safety and controllability (Zhao et al., 2024b; Bereska and Gavves, 2024; Opitz et al., 2025). Central to this field are the Linear Representation Hypothesis and Superposition Hypothesis, which posit that models represent human-interpretable concepts as linear directions within low-dimensional subspaces (Park et al., 2024; Elhage et al., 2022). This framework has been empirically validated through representation engineering, where intervening in specific directions allows for the direct manipulation of model behavior (Li et al., 2023; Arditi et al., 2024).

However, recent studies suggest that not all concepts are best captured by linear structures. Periodic concepts, such as days of the week or months of the year, have been found to form circular embeddings, and models undergoing the "grokking" phenomenon in modular arithmetic ultimately arrange numerical representations in circular configurations (Engels et al., 2025; Park et al., 2025; Liu et al., 2022; Nanda et al., 2023). These findings raise a fundamental question: Are there other critical human concepts that are characterized by non-linear, specifically circular, manifold structures?

In psychology, the Circumplex Model of Affect has long posited that emotions are arranged in a circle defined by two axes: valence and arousal (Figure 1(a)). Psychological emotion models are frequently employed in the analysis of machine learning models, with research investigating whether language models actually reflect these underlying structures (Wang and Zong, 2025; Zhao et al., 2025; Reichman et al., 2025). However, since these anal-

[1]Based on the circumplex model by Russell (1980). Retrieved from Wikimedia Commons (https://en.wikipedia.org/wiki/File:Circumplex_model_of_emotion.svg).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16981–17004
July 2-7, 2026 ©2026 Association for Computational Linguistics

yses are often limited to post-hoc observations, such as centroid distances or label co-occurrence probabilities, they remain confined to analyzing general trends and do not extend to an analysis at the concept manifold level. Whether a language model expresses emotions in the same way humans do, or whether it should do so, cannot be rigorously discussed unless the manifold structure of the model's embedding representations is explicitly reconstructed.

This paper investigates the validity of the circular emotion structure by explicitly inducing it within the embedding space. To this end, we first collect and synthesize a text dataset annotated with emotions assumed to follow a circular structure. We then train emotion representations using nGPT (Loshchilov et al., 2025), an architecture designed for hyperspherical representation learning, in combination with three contrastive loss functions: SINCERE (Feeney and Hughes, 2023), SoftCSE (Zhuang et al., 2024), and our proposed CircularCSE. These loss functions differ in the degree to which they incorporate a circular manifold structure as a constraint. Finally, we evaluate the resulting representations across three backbone categories (BERT-like models, LLM-based encoders, and decoder-only LLMs) using a comprehensive set of metrics, including V-Measure for discriminative power and Pearson correlation with Circumplex Distance (CD-r) for psychological alignment.

Our results reveal a stark structural dilemma. While the circular structure (CircularCSE) offers superior interpretability and remains robust in low-dimensional spaces or with few labels (Figure 1(d)), its performance degrades significantly in high-dimensional or fine-grained label settings compared to conventional designs. We provide a theoretical explanation for this: SINCERE thrives in high dimensions by arranging labels as an orthogonal simplex (90° margins), whereas CircularCSE's 2D ring geometry forces much tighter boundary margins, inherently limiting discriminability as the number of labels increases.

This conflict suggests that aligning models with human interpretation, which implicitly assumes low-dimensional manifold structures, comes at a structural cost to discriminative power. However, aligning model representations with human psychology offers diverse benefits, ranging from intuitive mechanistic interpretation to reduced computational costs. Our research re-examines the validity of incorporating human interpretability into model design from the perspective of deep learning.

2 Related Work

Psychological models of emotion are generally classified based on whether they represent emotion in a continuous or discrete space. A representative example of a continuous model is Russell's Circumplex Model of Affect, which posits that emotions are represented along two axes: valence and arousal (Russell, 1980). Along with the PAD model, which adds a dominance axis, it has been widely adopted in recent years (Mehrabian and Russell, 1974). In contrast, discrete emotion models construct a space from a set of basic emotions, where diverse emotional states are expressed through the compounding or subdivision of these basics. Plutchik proposed the "Wheel of Emotions," consisting of eight basic emotions (Plutchik, 1980). Regardless of the continuous or discrete space, most foundational psychological models represent emotions as a circular structure (Shaver et al., 1987; Ekman, 1992). This design reflects the clear similarities (e.g., joy and excitement) and polarities (e.g., positive and negative) inherent in emotions.

Geometry of LLM Representations. While the analysis of emotion representations in LLMs is common, many studies focus on examining the macroscopic relationships between emotions in existing models based on co-occurrence probabilities or the Euclidean distances between centroids (Guo and Choi, 2021; Wang and Zong, 2025; Zhao et al., 2025; Reichman et al., 2025). Outside the domain of emotion, it has been discovered that periodic concepts and numerical values in modular addition tasks exhibit circular structures (Engels et al., 2025; Park et al., 2025; Liu et al., 2022; Nanda et al., 2023). However, these findings are limited to specific conditions, such as those involving sparse autoencoders (Bricken et al., 2023) or grokking. To the best of our knowledge, no study has successfully constructed a circular structure for emotion representations within a standard language model.

Design of Contrastive Learning. Although many studies employ improved or custom variants of contrastive loss functions, they are primarily designed to enhance discriminative performance, with few focusing on the explicit induction of manifold structures (Choi et al., 2020; Yang et al., 2021; Deng et al., 2022). In the domain of image emotion classification, a prior study has incorporated angular

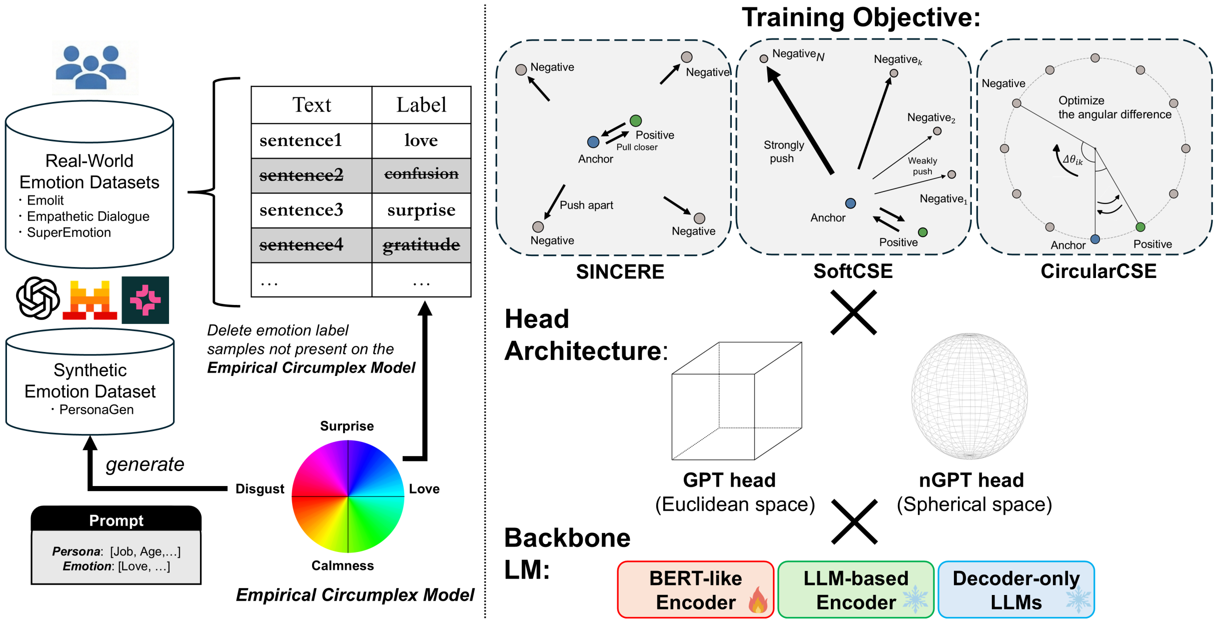


Figure 2: **Overview of our experimental framework.** (Left) Dataset construction procedure. Corresponding emotion labels are extracted or synthesized to reproduce the circumplex emotion structure. (Right) Training of GPT or nGPT heads across three backbone architectures using three distinct loss functions.

differences based on a circular assumption into the loss function (Yang et al., 2021). However, because this method operates within Euclidean space and combines this term with other loss functions, it is highly probable that the resulting manifold structure does not converge to a true circle.

In this study, we design a loss function that induces a circular structure on a hypersphere, thereby explicitly reproducing this geometry within the language model's embedding space. This allows us to rigorously discuss the utility and validity of applying psychological circumplex models of emotion to deep learning.

3 Methodology

3.1 Overview

Figure 2 presents an overview of our experimental framework. We collect datasets where emotions are assumed to be equally spaced on a circle and train the model to encode their representations. To replicate the circular structure, we employ the following two learning strategies:

Geometry of the embedding space. By learning emotion representations in a spherical space, we encourage the model to reproduce the circular structure, ensuring that differences are represented solely by angle distance (Section 3.3).

Contrastive learning design. We align the relationships between emotion labels by calibrating pairwise distances or gradient weights between anchor and negative samples according to their positions on the Circumplex Model of Affect (Section 3.4).

3.2 Preliminary

Figure 3 illustrates the basic circumplex emotion arrangement employed in this study, which consists of 12 emotion categories (hereafter referred to as the Empirical Circumplex Model, or ECM). While our model is primarily based on Russell's Circumplex Model of Affect (Russell, 1980; Yik et al., 2011), certain emotions have been substituted to align with the labels available in real-world datasets. For instance, although terms like "sleepy" or "quiet" would ideally represent the state of deactivation at $3\pi/2$, we adopt "calmness" due to the absence of datasets annotated with these specific labels. Formally, we define each emotion label as $y \in \mathcal{E}$, where y denotes the class of emotion and $\mathcal{E}$ denotes the complete set of emotion labels of size $E := |\mathcal{E}|$. In our experiments, we utilize a set of N text-label pairs for a given emotion classification dataset, denoted as $D = \{(x_i, y_i) \mid y_i \in \mathcal{E}\}_{i=1}^{N}$.

3.3 Head Architecture

We append a single Transformer block (projection head) to a pre-trained backbone, utilizing the head's output embedding space to represent the geometric structure of emotions. Let t denote the index of the token in the input sequence $[1, \ldots, T]$ and d denote the dimension of the model. A standard

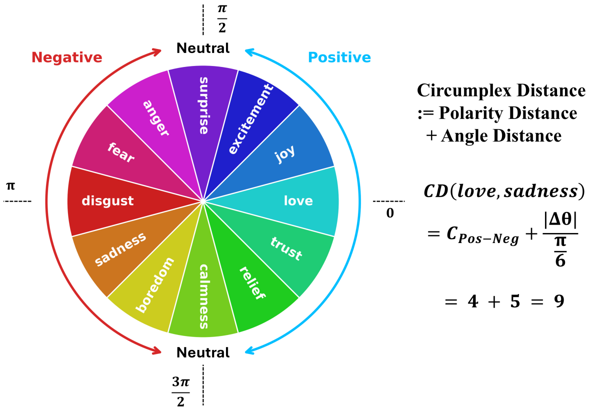


Figure 3: Our empirical circumplex model (ECM) and definition of Circumplex Distance (CD). CD is used only for evaluation (Section 4.2), not training.

Transformer block consists of an attention mechanism (ATTN), a multi-layer perceptron (MLP), and normalization modules (RMSNorm), formulated as follows:

$$\begin{aligned} h_t' &= h_t'' + \mathrm{ATTN}(\mathrm{RMSNorm}(h_t'')), \\ h_t &= h_t' + \mathrm{MLP}(\mathrm{RMSNorm}(h_t')), \end{aligned} \tag{1}$$

where $h_t, h_t', h_t'' \in \mathbb{R}^d$. Here, h_t'' denotes the backbone output, which serves as the head input, and h_t denotes the head output. This space corresponds to a d-dimensional Euclidean space where the arrangement of embeddings is unconstrained. Since the norm itself carries semantic meaning, it is inherently difficult to induce a circular geometry. Therefore, we adopt the normalized Transformer Block (nGPT), an architecture explicitly designed for spherical space (Loshchilov et al., 2025). The operations within the nGPT block are defined as follows:

$$\begin{aligned} h_t' &= \mathrm{Norm}((1-\alpha_A) \odot \mathrm{Norm}(h_t'') + \alpha_A \odot \mathrm{Norm}(\mathrm{ATTN}(h_t''))), \\ h_t &= \mathrm{Norm}((1-\alpha_M) \odot \mathrm{Norm}(h_t') + \alpha_M \odot \mathrm{Norm}(\mathrm{MLP}(h_t'))), \end{aligned} \tag{2}$$

where Norm() represents ℓ_2 normalization, $h_t, h_t' \in \mathbb{S}^{d-1}$, and $\alpha_A, \alpha_M \in \mathbb{R}^d$ are learnable parameters. nGPT removes normalization modules from standard transformer blocks and instead normalizes all hidden states and weights to unit norm along the feature dimension. Consequently, since the output of each module resides on a hypersphere and is updated along geodesics, the optimization process within the block can be viewed as traversing the spherical manifold. The rationale for adopting the nGPT architecture and further details regarding the intra-block processing are provided in Appendices A and B. In our experiments, we derive the final sentence embedding e by applying a pooling operation to the sequence of hidden states $h_{1:T}$, followed by normalization:

$$e = \mathrm{Norm}(\mathrm{Pooling}(h_{1:T})), \tag{3}$$

where $e \in \mathbb{S}^{d-1}, h_{1:T} \in \mathbb{R}^{T\times d}$. The specific pooling operation depends on the backbone model and is defined as follows:

$$\mathrm{Pooling}_{\mathrm{cls}} := h_1, \mathrm{Pooling}_{\mathrm{last}} := h_T, \mathrm{Pooling}_{\mathrm{mean}} := \frac{1}{T}\sum_{t=1}^{T} h_t. \tag{4}$$

For comparison, a conventional Transformer block (GPT head) is also trained as a baseline.

3.4 Training Objective

We perform contrastive learning by applying three distinct loss functions. Since our focus is on the quality of the manifold representation rather than emotion classification alone, we utilize the embeddings e directly during evaluation. Consequently, we employ neither additional linear classification layers nor cross-entropy loss. This approach ensures that the distinct effects of each loss function are directly reflected in the characteristics of the embedding space. We employ the Supervised InfoNCE REvisited (SINCERE) loss (Feeney and Hughes, 2023) as our baseline to achieve maximum discriminability while avoiding the problematic intra-class repulsion inherent in the original Supervised Contrastive Loss (Khosla et al., 2020). It computes the mean loss across all positive pairs within the batch $\mathcal{B}$ of size $B := |\mathcal{B}|$ for a given anchor as follows:

$$\mathcal{L}_{\mathrm{SINCERE}} = \frac{1}{B}\sum_{i\in\mathcal{B}}\left(\frac{-1}{|\mathcal{P}|}\sum_{j\in\mathcal{P}}\log\frac{\exp(e_i^T e_j/\tau)}{Z_i}\right) \tag{5}$$

where τ is a temperature, $\mathcal{P}$ is the in-batch positive set of the anchor sentence x_i, and Z_i represents the term corresponding to the positive sample as well as the in-batch negative samples:

$$Z_i = \exp(e_i^T e_j/\tau) + \sum_{k\in\mathcal{N}} \exp(e_i^T e_k/\tau) \tag{6}$$

where $\mathcal{N}$ is the in-batch negative set. Since this loss function computes the loss for all pairs within the batch, it induces a strong separation between positive and negative samples. However, because it applies equal weight to all negative samples, it fails to account for the specific degree of separation required between the anchor and each individual negative sample. Given that the ECM positions similar emotions nearby and opposing emotions at antipodes, the loss function should ideally dictate specific pairwise distances. Accordingly, we utilize SoftCSE for soft constraints and CircularCSE for hard constraints. We refine the 'weight individualization' method proposed in previous SoftCSE research, as it offers a more intuitive formulation and greater numerical stability (Zhuang et al., 2024). SoftCSE assigns individual weights to the negative sample terms in the SINCERE loss:

$$Z_i = \exp(e_i^T e_j/\tau) + \sum_{k\in\mathcal{N}} w_{ik}\exp(e_i^T e_k/\tau), \tag{7}$$

$$w_{ik} = \frac{1-\cos(\Delta\theta_{ik})}{\frac{1}{|\mathcal{N}|}\sum_{k\in\mathcal{N}}(1-\cos(\Delta\theta_{ik}))} \tag{8}$$

where $\Delta\theta_{ik}$ is the angular difference on the ECM. The numerator of w_{ik} is inversely proportional to the similarity of the emotion labels i and k i.e., $w_{ik} \propto -\cos(\Delta\theta_{ik})$, meaning that the closer the labels are on the circle, the smaller the weight and the weaker the repelling force becomes. The denominator acts as a normalization factor within the batch, adjusting the scale of the negative terms to match Equation (6) i.e., $\sum_{k\in\mathcal{N}} 1 = \sum_{k\in\mathcal{N}} w_{ik} = |\mathcal{N}|$. While the original paper employed $e_i^T e_k$ computed via a frozen encoder model instead of $\cos(\Delta\theta_{ik})$, this approach incurs additional inference overhead during training and relies heavily on the performance of the encoder. Therefore, we pre-define pairwise distances based on the ECM. This design facilitates the straightforward assignment of individual weights to negative samples.

As an even stronger constraint, we propose CircularCSE, which directly learns distances on the circle:

$$\mathcal{L}_{\text{CircularCSE}} = \frac{1}{B(B-1)}\sum_{i,j\in\mathcal{B}:i\neq j}\ell_{ij},$$
$$\ell_{ij} = \begin{cases}[\max(0, |e_i^T e_j - \cos(\Delta\theta_{ij})| - m)]^2 & \text{if } y_i = y_j\\ (e_i^T e_j - \cos(\Delta\theta_{ij}))^2 & \text{otherwise}\end{cases} \tag{9}$$

Here, $m > 0$ is a margin hyperparameter that allows for tolerance within classes. Although a margin is desirable to account for variations in nuance and intensity among samples sharing the same label, it inevitably lowers the model's discriminability.

4 Experiments

4.1 Experimental Setup

Datasets. Our experiments utilize three real-world datasets: Emolit (Rei and Mladenić, 2023), Empathetic Dialogue (Rashkin et al., 2019), SuperEmotion (de Fortuny, 2025), and one synthetic dataset: PersonaGen (Inoshita and Harada, 2025). Featuring a wide range of emotion labels, these datasets allow us to evaluate the geometric fidelity of the embedding space by analyzing how emotional categories are structurally organized. We select samples annotated with labels that match or closely resemble those in the ECM shown in Figure 3. For each dataset, the training set consists of 500 instances sampled per emotion label (450 for SuperEmotion), and the test set comprises 100 instances per label. Detailed descriptions of each dataset and the construction pipeline for the synthetic dataset are provided in Appendix C.

Models. To account for variations in embedding space properties resulting from diverse architectures and pre-training objectives, we categorize the models into three groups, selecting two representatives from each: **BERT-like Encoders** (mE5 (Wang et al., 2024) and mxbai (Lee et al., 2024)), **LLM-based Encoders** (Qwen3-Embedding-4B (Zhang et al., 2025) and Llama-Embed-Nemotron-8B (Babakhin et al., 2025)), and **Decoder-only LLMs** (Llama-3.2-3B (Grattafiori et al., 2024) and OLMo-3-7B (OLMo Team et al., 2025)).

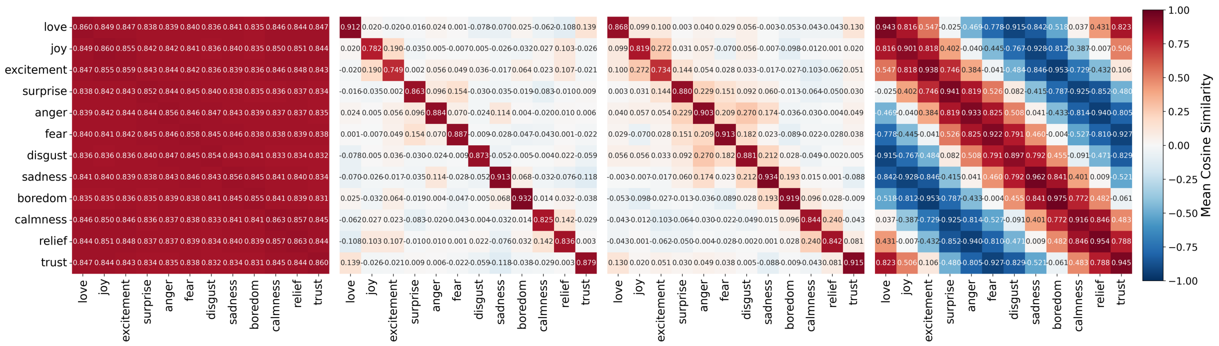
Implementation Details. We attach a single-layer Transformer Block (GPT head) or a normalized Transformer Block (nGPT head) to the final layer of each backbone model. The head shares the same attention mask, dimension size, and attention head count as the backbone. Final sentence embeddings are derived via specific pooling methods: the [CLS] token for BERT-like encoders, the last token for decoder-only LLMs and Qwen3-Embedding. For Llama-Embed-Nemotron, we utilize mean pooling, consistent with its original training methodology. During training, BERT-like encoders are fully fine-tuned, while the LLM backbones remain frozen with only the heads being trained. Further implementation details are provided in Appendix D.

4.2 Evaluation Metrics

We assess how well the trained embeddings capture emotional expressions via clustering analysis. Following the protocol of the MTEB clus-

Training Objective	Head Arch.	mE5		Qwen3-Embedding-4B		Llama-3.2-3B	
		V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r
Pretrained		0.342	0.574	0.495	0.522	0.094	0.217
SINCERE	- GPT	**0.760**	0.317	**0.756**	0.305	**0.725**	0.358
	- nGPT	0.744	0.221	0.739	0.545	0.577	0.425
SoftCSE	- GPT	0.755	0.477	0.751	0.552	0.710	0.548
	- nGPT	0.753	0.499	0.723	0.708	0.516	**0.728**
CircularCSE	- GPT	0.717	0.757	0.643	0.747	0.579	**0.728**
	- nGPT	0.720	**0.764**	0.659	**0.753**	0.382	0.708

Table 1: Summary of average performance across all datasets for selected models. Best results per model (excluding Pretrained) are bolded, and worst are underlined. V_{Measure} indicates clustering quality, and CD-r indicates correlation with the circumplex distance.



(a) Pretrained (b) SINCERE (c) SoftCSE (d) CircularCSE

Figure 4: Average cosine similarity between emotion label pairs of mE5

tering tasks (Muennighoff et al., 2023), we partition the test set into clusters using k-means (MacQueen, 1967), setting the number of clusters k equal to the number of ground truth labels, and evaluate performance using V-Measure (Rosenberg and Hirschberg, 2007). Given that our models are optimized for cosine similarity, we utilize Spherical k-means (Dhillon and Modha, 2001), which uses cosine distance instead of the Euclidean ℓ_2 metric. The algorithm is run 10 times with varied initializations, and the result with the minimum inertia is selected. We evaluate not only the discriminative power of the embedding space but also the extent to which its structure aligns with human perception. Building upon the approaches of (Zhao et al., 2016, 2024a), we define the Circumplex Distance (CD) on the ECM as follows: $\text{CD}(y_i, y_j) := C + \text{AngleDistance}(y_i, y_j)$ where C represents the constant inter-polarity distance, and $\text{AngleDistance}(y_i, y_j)$ denotes the number of steps between labels on the ECM (see Figure 3). We define the distance between Neutral and Positive/Negative polarities as 2, the distance between Positive and Negative polarities as 4, and the distance between identical polarities as 0. This configuration ensures that the distance between opposing polarities consistently exceeds the distance between identical polarities. By incorporating the premise that polarity differences are cognitively more significant than mere steps on the circle, CD reflects human psychological emotion structure more faithfully. To measure the alignment with this metric, we propose the Pearson correlation with CD (CD-r):

$$\text{CD-r} := \text{Pearson}(\text{CD}(y_i, y_j), 1 - \text{AvgCosSim}(y_i, y_j)) \tag{10}$$

$$\text{AvgCosSim}(y_i, y_j) := \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} e_k^T e_l \tag{11}$$

with N_i and N_j representing the total number of samples for each emotion label in the test set.

4.3 Results

Table 1 presents the average V-Measure and CD-r across datasets for mE5, Qwen3-Embedding-4B, and Llama-3.2-3B (comprehensive results are provided in Table 4 of Appendix E).

Regarding the training objectives, SINCERE and SoftCSE yield higher V-Measure scores, whereas CircularCSE underperforms. Conversely, for CD-r, SINCERE scores are low, while CircularCSE achieves higher results. This phenomenon can be intuitively explained by referencing Figure 4, which illustrates the average cosine similarity between emotion label pairs for each head. Pre-trained models typically form an anisotropic embedding space, resulting in high similarity across all emotion label pairs (Ethayarajh, 2019). In contrast, SINCERE attempts to position negative samples orthogonally, driving similarities toward zero. SoftCSE relaxes this constraint, and CircularCSE aligns the similarity of each emotion label pair with the corresponding cosine similarity on the circular manifold. Since clustering requires high discriminative power between labels, SINCERE is advantageous; however, it fails to sufficiently capture the relational structure between labels, leading to a lower CD-r. While CircularCSE successfully captures the ordinal relationships between labels, it makes distinguishing between adjacent labels more difficult, resulting in a lower V-Measure. These results highlight a fundamental conflict between the objectives of deep learning, which prioritizes discriminative accuracy, and psychology, which emphasizes alignment with human perception, creating a trade-off between accuracy and interpretability.

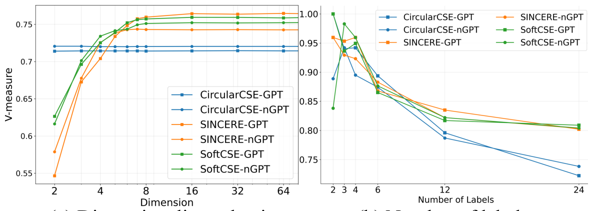
Revisiting Table 1, we observe distinct trends across models; notably, for Llama-3.2-3B, the nGPT model exhibits a lower V-Measure. This suggests that Decoder-only models rely more heavily on vector norms to encode contextual information compared to Encoder models, implying that the normalization process leads to information loss. Furthermore, the significantly low accuracy of the pretrained models indicates that while emotion information is latent within the embeddings, it is not explicitly separable; instead, the embedding space prioritizes contextual information or features required for next-token prediction. These findings are consistent across multilingual datasets (Appendix F) and across various parameter scales (Appendix G).

5 Analysis

5.1 Conflict between Psychological and Deep Learning Models

To better elucidate the differences between the individual methods, we conduct further analysis. Figures 1(b), (c), and (d) visualize the results of applying PCA to the Emolit test set embeddings for each mE5 head. For SINCERE-nGPT and SoftCSE-nGPT, although the emotions separate into distinct clusters, the structures lack clear regularity, and the explained variance ratio of the principal components is low. In contrast, CircularCSE-nGPT exhibits a clear circular arrangement of emotions, with a significantly higher explained variance ratio. Hypothesizing that this difference in arrangement would manifest clearly under distinct scenarios, we conducted the following two experiments:

1. Robustness to dimensionality reduction.
2. Robustness to the number of emotion labels.



(a) Dimensionality reduction (b) Number of labels

Figure 5: Clustering performance of mE5 heads under different conditions. (a) Impact of PCA dimensionality reduction. (b) Impact of the number of emotion labels.

The experimental results are presented in Figure 5. Figure 5(a) shows the average V-Measure across datasets when clustering with Spherical k-means after reducing the dimensions of each mE5 head via PCA. Figure 5(b) illustrates the change in V-Measure on the Emolit dataset as the number of emotion labels varies (results for other models and details on label configurations are provided in Appendix H). CircularCSE maintains stable accuracy even in low dimensions and performs comparably to SINCERE and SoftCSE when the number of emotion labels is small. However, its performance degrades in high-dimensional settings or with a large number of labels. These results can be explained by the optimal solutions and optimal margins of the respective loss functions. The SINCERE loss function, given by Equations (5) and (6), attains its theoretical lower bound when the cosine similarity between all positive-negative pairs equals $\frac{-1}{E-1}$ (Proof in Appendix I). In practice, however, due to the curse of dimensionality, the positive-negative similarity often settles at a local optimum of 0, resulting in an average boundary margin of 90° (orthogonality). In contrast, since CircularCSE arranges labels on a ring, the geometry of its optimal solution is consistently a 2-dimensional circle, and the boundary margin be-

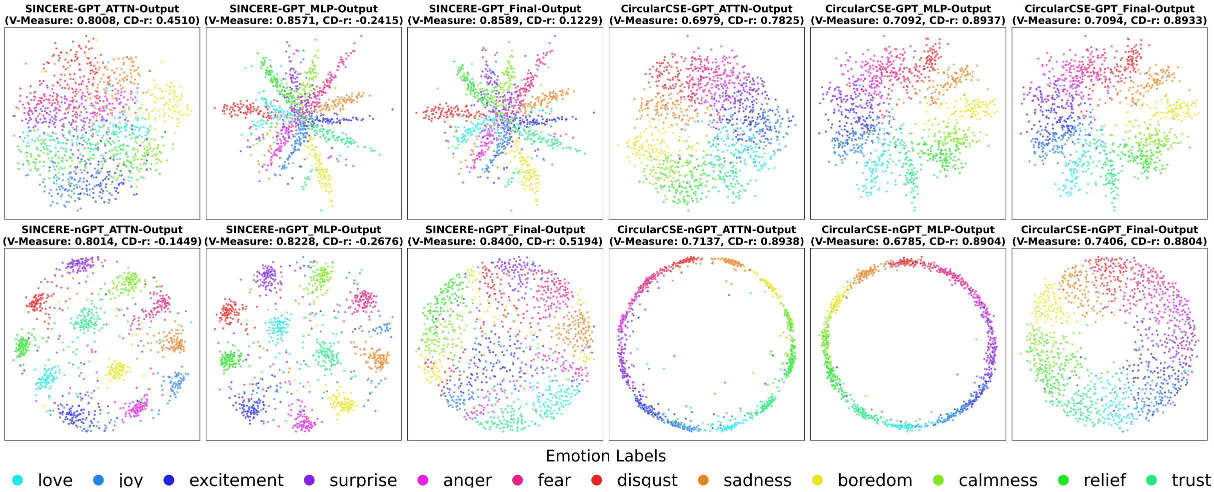


Figure 6: Visualization of embedding representations from each module of the GPT and nGPT heads using Multidimensional Scaling (MDS). Additional results for other models are presented in Appendix J.

tween any positive-negative pair is at most $\frac{\pi}{E}$ (for a 12-class classification problem, the maximum margin is 30°). The optimal boundary margins of SINCERE and CircularCSE coincide only in 2 dimensions; otherwise, the gap in discriminative power widens as the model dimensionality or the number of labels increases. This indicates that arranging emotions on a circular manifold, i.e., attempting to make representations visualization-friendly or semantically interpretable, implicitly imposes a low-dimensional manifold structure, which conflicts with the high-dimensional representations typical of deep learning. This limitation becomes more pronounced as model capacity increases (higher dimensionality) or as task demands rise (distinguishing among a diverse set of emotions).

5.2 Effects of Spherical Constraints

We perform a qualitative evaluation of the head architectural differences using dimensionality reduction. Specifically, we extract intermediate representations from within the transformer block of Qwen3-Embedding-4B and employ Multidimensional Scaling (MDS) for dimensionality reduction (de Leeuw, 2005), followed by clustering and visualization. Since MDS arranges points to reconstruct the pairwise distance matrix of the samples, it is particularly effective at preserving global geometry in high-dimensional spaces.

The results are presented in Figure 6. The impact of the spherical constraint is prominently reflected in the structural differences of the representations. In the GPT head, the output of the MLP layer dominates the final result; the MLP amplifies differences in magnitude (norm), effectively overwriting prior intermediate representations with vectors of larger norms. Additionally, clusters exhibit linear elongation, where concepts appear as directional basis vectors. Although this expansion increases intra-cluster variance—allowing for a wider range of captured representations—it hinders Euclidean-based classification, as points may inadvertently lie closer to external clusters than to their own.

In contrast, in the nGPT head, the removal of the norm component results in the Attention and MLP layers exhibiting similar output geometries, with clusters adopting complex, non-linear structures rather than straight lines. We attribute this to the fact that on the hypersphere, differences between concepts are expressed solely via angular components rather than norms, causing directions to encode composite rather than singular concepts.

These findings suggest that even when quantitative evaluation metrics appear identical, the underlying representation structures can differ vastly. Manifold-aware research remains scarce, and determining the optimal manifold structure for specific tasks is an open question for future work.

6 Discussion and Conclusion

We compared conventional models with those reproducing psychological circular structures by performing contrastive learning in Euclidean and spherical spaces. Through this comparison, we discovered an unavoidable structural dilemma: a trade-off between discriminative performance and human interpretability. This highlights the importance of selecting the appropriate method based on the intended application. Although this study focused on clustering, our framework is transferable

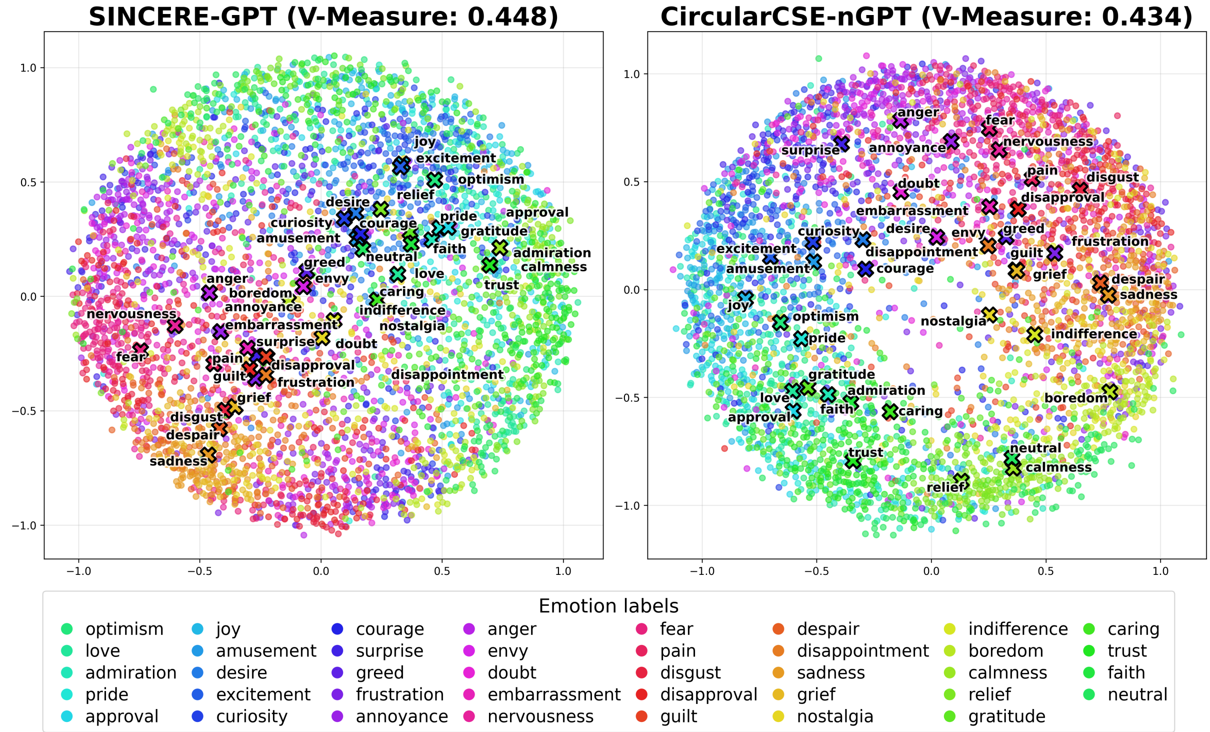


Figure 7: MDS visualization of 39 emotion labels from the Emolit dataset for Qwen3-Embedding-4B trained with a 12-class ECM. The plotted label names represent the centroids for each emotion class.

to representation learning for other tasks or concepts. For instance, interpretable approaches can potentially be applied to tasks such as identifying underrepresented labels and verifying the validity of emotion placement.

Figure 7 illustrates the MDS visualization of the full Emolit dataset, encompassing all 39 emotion labels including those not encountered during training. While SINCERE-nGPT tends to collapse into a bipolar positive-negative structure for these unseen labels, CircularCSE-nGPT maintains its circular geometry, situating novel labels in semantically plausible locations relative to known categories. Interestingly, the gap in V-Measure scores between these methods, which was approximately 0.1 on the training set, diminishes to nearly 0.01 in this context. This suggests that grounding models in human-centric psychological priors may offer a more robust rationale for representing complex emotional landscapes beyond the training distribution. Notably, while Russell's original circumplex model places "Neutral" at the center of the circle (Figure 1(a)), our induced representations position it near low-arousal states such as "calmness" and "relief". From a linguistic perspective, this indicates that neutral affect is closely related to deactivation. Such findings demonstrate the potential for this framework to serve as a mathematical tool for verifying or refining theories regarding the structural arrangement of human emotions. Furthermore, these visualizations serve as a practical tool for examining the bias and comprehensiveness of emotion labels within a dataset. For example, the region surrounding "boredom" is sparsely populated, revealing that texts annotated with such deactivated emotions are rare in current corpora.

While this study focused on emotion, our approach is versatile and can be extended to various domains. For example, the theory of Interpersonal Styles (Wiggins, 1979) and Basic Human Values (Schwartz, 2012) are also posited to exhibit a circular structure. Not only psychological models, but also subjective concepts with inherent polarities (e.g., political ideologies) could be modeled intuitively through this framework. This allows for the representation of fluid concepts that traditional classifiers fail to capture.

Finally, we envision the ultimate objective of interpretable models as enabling humans to steer model behavior through direct interaction within visualized representation spaces. If a text generation mechanism can be developed based on the approach proposed in this study, it would facilitate intuitive and flexible model manipulation. Therefore, extending this framework to generation tasks represents a key direction for future research.

Limitations

Task Simplification. We simplified the task formulation for our experiments. Real-world emotions are highly complex and often involve simultaneous conflicting states (e.g., bittersweet, calm anger). However, in this study, we operate under the assumption that each text corresponds to a single emotion label. Furthermore, regarding the "Neutral" state, while we treated it as a distinct category belonging to neither Positive nor Negative extremes, it is intrinsically located at the center (origin) of Russell's circumplex model. Integrating these complex cases into our current framework would require specialized mechanisms; developing a more natural representation for such states remains a challenge for future work.

Modality Constraints. Emotion analysis spans not only text but also modalities such as images and audio. Certain emotional states (e.g., boredom or drowsiness) are more readily manifested through gestures or non-verbal cues rather than text. Capturing these nuances necessitates a multimodal architectural design.

Connection to Psychological Models. Although the circumplex model used in this study references Russell's model, some emotion placements differ from the original configuration to align with the specific labels available in our dataset. Additionally, numerous circumplex models exist beyond Russell's proposal, and it is possible that our specific arrangement does not perfectly reflect the true relationships between emotions (discussed in Appendix C.4). However, since our conclusions are largely driven by the dimensionality of the embedding space, we anticipate obtaining similar results with other 2- or 3-dimensional visualizable psychological models. Reproducing higher-dimensional and more complex emotional models is an objective for future research.

Evaluation Metrics. In this study, we introduced and evaluated CD-r as a new metric to assess model interpretability. As an intrinsic measure, CD-r does not necessarily correlate directly with downstream model performance, as demonstrated by the trade-offs observed in our experiments. Since interpretability and explainability remain ambiguous concepts, there are currently no universally established evaluation metrics in this field. While metrics such as circularity or curvature could potentially characterize the geometry of embedding spaces, further investigation is required to determine which specific aspects of model behavior they quantify. Developing valid and general-purpose metrics for model interpretability remains a subject for future research.

Datasets. Our research utilized supervised emotion classification datasets annotated with fine-grained emotion labels. Given that datasets with such detailed annotations are scarce across diverse domains , devising a framework that can be applied to unsupervised or self-supervised settings is a key direction for future work.

Ethical Considerations

This study primarily utilizes publicly available and properly licensed datasets and models, all of which permit research use. The sole exception is the generation of synthetic data, where we employed both open-source and proprietary models to create a new dataset. While this dataset may potentially contain harmful content, it was used exclusively for training classification tasks; consequently, the trained models themselves pose no safety risks. All datasets and models were used in accordance with their intended research purposes.

We used AI tools for proofreading and mathematical typesetting in the preparation of this paper. All content has been thoroughly reviewed and verified by the human authors.

Acknowledgments

This work was supported by Research and Development Center for Large Language Model/National Institute of Informatics, and JSPS KAKENHI Grant Number 24K03231.

References

Allen Institute for AI. 2025. Olmo-3-7b-instruct model card. https://huggingface.co/allenai/Olmo-3-7B-Instruct. Accessed: 2025-12-27.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *ArXiv*, abs/2406.11717.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *Preprint*, arXiv:2511.07025.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions* *on Machine Learning Research*. Survey Certification, Expert Certification.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Hongjun Choi, Anirudh Som, and Pavan Turaga. 2020. Amc-loss: Angular margin contrastive loss for improved explainability in image classification. *Preprint*, arXiv:2004.09805.

Enric Junqué de Fortuny. 2025. The super emotion dataset. *arXiv preprint arXiv:2505.15348*.

Jan de Leeuw. 2005. Modern multidimensional scaling: Theory and applications (second edition). *Journal of Statistical Software*, 14.

Google DeepMind. 2025. gemma3-27b-it model card. `https://huggingface.co/google/gemma-3-27b-it`. Accessed: 2025-12-17.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. 2022. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979.

Inderjit S Dhillon and Dharmendra S Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1):143–175.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

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.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Patrick Feeney and Michael C Hughes. 2023. Sincere: Supervised information noise-contrastive estimation revisited. *arXiv preprint arXiv:2309.14277*.

Michael Freenor and Lauren Alvarez. 2025. Steering embedding models with geometric rotation: Mapping semantic relationships across languages and models. *Preprint*, arXiv:2510.09790.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. 2021. Dissecting supervised contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3821–3830. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuting Guo and Jinho D Choi. 2021. Enhancing cognitive models of emotions with representation learning. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 141–148.

Keito Inoshita and Rushia Harada. 2025. Persona-based synthetic data generation using multi-stage conditioning with large language models for emotion recognition. *arXiv preprint arXiv:2507.13380*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embedding model.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. 2022. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. 2025. nGPT: Normalized transformer with representation learning on the hypersphere. In *The Thirteenth International Conference on Learning Representations*.

J MacQueen. 1967. Multivariate observations. In *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, volume 1, pages 281–297.

Albert Mehrabian and James A. Russell. 1974. *An Approach to Environmental Psychology*. The MIT Press, Cambridge, MA.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. 2019. Spherical text embedding. *Advances in neural information processing systems*, 32.

Meta AI. 2024. Llama-3.3-70b-instruct model card. `https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/`. Accessed: 2025-12-27.

Microsoft. 2025. Phi-4-mini-instruct model card. `https://huggingface.co/microsoft/Phi-4-mini-instruct`. Accessed: 2025-12-27.

Mistral AI. 2025. Ministral-3-14b model card. `https://huggingface.co/mistralai/Ministral-3-14B-Base-2512`. Accessed: 2025-12-27.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

OLMo Team, Allyson Ettinger, David Heineman, Amanda Bertsch, Bailey Kuehl, Dirk Groeneveld, Hamish Ivison, and 1 others. 2025. Olmo 3 technical report. Technical report, Allen Institute for AI. Accessed: 2025-12-17.

OpenAI. 2025. Gpt-5 nano model overview. `https://openai.com/ja-JP/index/gpt-5-system-card/`. Accessed: 2025-12-27.

Juri Opitz, Lucas Moeller, Andrianos Michail, Sebastian Padó, and Simon Clematide. 2025. Interpretable text embeddings and text similarity explanation: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22314–22330, Suzhou, China. Association for Computational Linguistics.

Vardan Papyan, X. Y. Han, and David L. Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. 2025. ICLR: In-context learning of representations. In *The Thirteenth International Conference on Learning Representations*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Robert Plutchik. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, New York, USA.

Qwen. 2025. Qwen3-30b-a3b-instruct-2507 model card. `https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507`. Accessed: 2025-12-27.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.

Luis Rei and Dunja Mladenić. 2023. Detecting fine-grained emotions in literature. *Applied Sciences*, 13(13).

Benjamin Reichman, Adar Avsian, and Larry Heck. 2025. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models. *arXiv preprint arXiv:2510.22042*.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

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

Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061.

Chris Van Pelt and Alex Sorokin. 2012. Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 765–766.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Xiangyu Wang and Chengqing Zong. 2025. Learning emotion category representation to detect emotion relations across languages. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4752–4767.

Jerry S Wiggins. 1979. A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, 37(3):395–412.

Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, and Xinbo Gao. 2021. A circular-structured representation for visual emotion distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4237–4246.

Michelle Yik, James A Russell, and James H Steiger. 2011. A 12-point circumplex structure of core affect. *Emotion*, 11(4):705.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Bo Zhao, Maya Okawa, Eric J. Bigelow, Rose Yu, Tomer Ullman, Ekdeep Singh Lubana, and Hidenori Tanaka. 2025. Emergence of hierarchical emotion organization in large language models. *Preprint*, arXiv:2507.10599.

Chenxi Zhao, Jinglei Shi, Liqiang Nie, and Jufeng Yang. 2024a. To err like human: Affective bias-inspired measures for visual emotion recognition evaluation. *Advances in Neural Information Processing Systems*, 37:134747–134769.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024b. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. 2016. Predicting personalized emotion perceptions of social images. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1385–1394.

Haojie Zhuang, Wei Emma Zhang, Jian Yang, Weitong Chen, and Quan Z Sheng. 2024. Not all negatives are equally negative: Soft contrastive learning for unsupervised sentence representations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3591–3601.

A Justification to use nGPT

In this section, we provide a detailed rationale for adopting the nGPT architecture to reproduce the circular structure within the model's embedding space. Simply, normalizing the model's embeddings at every step would satisfy the condition; this ensures all outputs lie on the unit hypersphere, meaning differences are expressed solely as angular differences. However, this design introduces significant issues for gradient descent during training.

Since the hypersphere is a non-Euclidean space where distance is geodesic, the standard linear vector addition inherent in residual connections updates vectors in the direction of steepest descent without accounting for curvature. Consequently, the updated representations deviate from (or 'fall off') the hypersphere. Therefore, we must employ Riemannian optimization instead of standard gradient descent to account for the manifold's curvature during training (Meng et al., 2019; Freenor and Alvarez, 2025), necessitating the design of a more sophisticated architecture. The nGPT architecture is particularly compatible with representation learning for the following three reasons:

Elimination of Weight Decay: Research on grokking suggests that weight decay is critical for the emergence of generalization circuits (Liu et al., 2022; Nanda et al., 2023). Models often memorize solutions within the magnitude (norm) of the weights; thus, weight decay is typically required to suppress these norm components and encourage structural generalization. In nGPT, however, weights are normalized by design, compelling the network to learn generalized structures directly without relying on weight decay.

Feasibility of Head Training: To the best of our knowledge, pre-trained weights for full-scale nGPT models are not currently available, and training from scratch is computationally expensive. However, in embedding tasks, it is standard practice to freeze the backbone and train only a projection head. Adopting an nGPT head aligns with this convention while bypassing the cost of full pre-training.

Angle-based Optimization: In the original paper, nGPT was evaluated primarily as a causal language model, focusing on convergence speed rather than the implications of learning on a spherical manifold. In embedding tasks, however, cosine similarity (i.e., distance on the unit hypersphere) is the standard metric. Therefore, nGPT enables consistent angle-based optimization of representations throughout both training and inference.

These advantages make this architecture highly effective, not only for the current experiment but also as a general design principle for training embedding models.

B Transformer Block

This section describes the detailed computational flow of the GPT head and the nGPT head adopted in this paper. Let t denote the index of the token in the input sequence $[1, \dots, T]$ and d denote the dimension of the model. A standard Transformer block consists of an attention mechanism (ATTN), a multi-layer perceptron (MLP), and normalization modules (RMSNorm), formulated as follows:

$$\begin{aligned} h_t^{'} &= h_t^{''} + \text{ATTN}(\text{RMSNorm}(h_t^{''})), \\ h_t &= h_t^{'} + \text{MLP}(\text{RMSNorm}(h_t^{'})), \end{aligned} \tag{12}$$

where $h_t, h_t^{'}, h_t^{''} \in \mathbb{R}^d$. Here, $h_t^{''}$ denotes the input to the block, and h_t denotes the output. In our experiments, we derive the final sentence embedding e by applying a pooling operation to the sequence of hidden states $h_{1:T}$, followed by normalization:

$$e = \text{Norm}(\text{Pooling}(h_{1:T})), \tag{13}$$

where $e \in \mathbb{S}^{d-1}, h_{1:T} \in \mathbb{R}^{T \times d}$. The specific pooling operation depends on the backbone model and is defined as follows:

$$\begin{aligned} &\text{Pooling}_{\text{cls}} := h_1, \text{Pooling}_{\text{last}} := h_T, \\ &\text{Pooling}_{\text{mean}} := \frac{1}{T} \sum_{t=1}^{T} h_t, \end{aligned} \tag{14}$$

The ATTN performs the following computations:

$$\text{ATTN}(h_t^{''};\ h_{1:t}^{''}) = \text{Concat}(\text{A}_1, ..., \text{A}_{n_{heads}}) \boldsymbol{W}_O, \tag{15}$$

$$\text{A}_n = \text{softmax}\left(\frac{(h_t^{''} \boldsymbol{W}_Q^n)(h_{1:t}^{''} \boldsymbol{W}_K^n)^T}{\sqrt{d_{head}}} \right) (h_{1:t}^{''} \boldsymbol{W}_V^n), \tag{16}$$

where $\boldsymbol{W}_Q^n, \boldsymbol{W}_K^n, \boldsymbol{W}_V^n \in \mathbb{R}^{d \times d_{head}}, \boldsymbol{W}_O \in \mathbb{R}^{d \times d}, d_{head} = d / n_{heads}$. For unidirectional models, the attention mask references tokens from 1 to t, whereas for bidirectional models, it references the entire sequence from 1 to T. The MLP operation is defined as follows

$$\text{MLP}(h_t^{'}) = (\text{SiLU}(h_t^{'} \boldsymbol{W}_u) \odot h_t^{'} \boldsymbol{W}_v) \boldsymbol{W}_{o\text{MLP}}, \tag{17}$$

where $\boldsymbol{W}_u, \boldsymbol{W}_v \in \mathbb{R}^{d \times d_{\text{MLP}}}, \boldsymbol{W}_{\text{oMLP}} \in \mathbb{R}^{d_{\text{MLP}} \times d}$. Next, we describe the architecture of the Normalized Transformer Block (nGPT). The first primary modification is that residual connections are computed along geodesics:

$$\begin{aligned} h_t^{'} &= \text{Norm}((1 - \alpha_A) \odot \text{Norm}(h_t^{''}) + \alpha_A \odot \text{Norm}(\text{ATTN}(h_t^{''}))), \\ h_t &= \text{Norm}((1 - \alpha_M) \odot \text{Norm}(h_t^{'}) + \alpha_M \odot \text{Norm}(\text{MLP}(h_t^{'}))), \end{aligned} \tag{18}$$

where $h_t, h_t^{'}, h_t^{''} \in \mathbb{S}^{d-1}$, and $\alpha_A, \alpha_M \in \mathbb{R}^d$ are learnable parameters. By constraining the vector updates to the hypersphere, this process enables pseudo-Riemannian optimization. The ATTN and MLP modules incorporate the following modifications:

$$\begin{aligned} &\text{A}_n = \text{softmax}(\boldsymbol{q}\boldsymbol{k}^T \cdot \sqrt{d_{head}})(h_{1:t}^{''} \boldsymbol{W}_V^n), \\ &\text{where } \ \boldsymbol{q} = \text{Norm}(h_t^{''} \boldsymbol{W}_Q^n) \odot s_{qk}, \\ &\qquad\qquad \boldsymbol{k} = \text{Norm}(h_{1:t}^{''} \boldsymbol{W}_K^n) \odot s_{qk}, \end{aligned} \tag{19}$$

with $s_{qk} \in \mathbb{R}^{d_{head}}$, and

$$\begin{aligned} &\text{MLP}(h_t^{\prime}) = (\text{SiLU}(v) \odot u) \boldsymbol{W}_{o\text{MLP}}, \\ &\text{where} \quad u = (h_t^{\prime} \boldsymbol{W}_u) \odot s_u, \\ &\qquad\qquad v = (h_t^{\prime} \boldsymbol{W}_v) \odot s_v \sqrt{d_{\text{MLP}}}, \end{aligned} \tag{20}$$

with $s_u, s_v \in \mathbb{R}^{d_{\text{MLP}}}$. Here, s_{qk}, s_u, s_v are learnable scaling parameters introduced to compensate for the absence of the learnable affine parameters typically found in RMSNorm. Furthermore, all weight matrices $\boldsymbol{W}$ are normalized along their embedding dimension at every training step.

C Dataset Collection

This section outlines the selection criteria and construction procedures for the dataset.

C.1 Real-world Emotion Dataset

Our primary criterion for dataset selection was the availability of a diverse range of emotion labels. Reproducing a circular structure requires gathering emotions with distinct properties; in particular, low-arousal emotions (e.g., boredom, calmness) were critical yet often scarce in standard datasets.

Emolit (Rei and Mladenić, 2023) is a large-scale emotion text dataset sourced from Project Gutenberg, annotated with 38 distinct emotion labels.

Dataset	Train/Test	Labels
Emolit	6000/1200	love, joy, excitement, surprise, anger, fear, disgust, sadness, boredom, calmness, relief, trust
Empathetic Dialogue	4000/800	joyful, excited, surprised, angry, afraid, disgusted, sad, (trusting, faithful)
SuperEmotion	4050/900	love, joy, excitement, surprise, anger, fear, disgust, sadness, relief
PersonaGen	6000/1200	love, joy, excitement, surprise, anger, fear, disgust, sadness, boredom, calmness, relief, trust

Table 2: Details of the datasets used in Section 4.3. All datasets are balanced, containing an equal number of samples for each emotion label. For Empathetic Dialogue, we chose the labels that most closely match the categories in Figure 3.

Pretrained Model	intfloat/ multilingual-e5-large	mixedbread-ai/ mxbai-embed-large-v1	Qwen/ Qwen3-Embedding-4B	nvidia/ llama-embed-nemotron-8b	meta-llama/ Llama-3.2-3B	allenai/ Olmo-3-1025-7B
Backbone	unfreeze		freeze			
Torch dtype	bfloat16					
Epoch	15					
Learning Rate	5e-5					
LR scheduler type	constant					
Train Batch Size	128					
Random Seed	42					
Hidden size d	1024		2560	4096	3072	4096
Num attention heads	16		32		24	32
Pooling strategy	cls		last	mean	last	
(SINCERE, SoftCSE)-τ	0.05					
CircularCSE-margin m	0					

Table 3: Hyperparameters used to train models

Each text entry is accompanied by emotion prediction probabilities generated by a binary Natural Language Inference (NLI) model. The dataset is characterized by a well-balanced distribution of diverse emotion labels, achieved through rigorous preprocessing for noise removal and diversity enhancement. For our experiments, we assigned the emotion with the highest predicted probability as the ground truth label for each text.

Empathetic Dialogue (Rashkin et al., 2019) is a conversational emotion dataset. Each session consists of a Speaker and a Listener, where the Listener's responses are designed to empathize with the Speaker's emotional state. Although the dataset is annotated with 32 emotion labels, it is not grounded in a specific psychological emotion model; consequently, it contains numerous labels that are undefined in standard psychological frameworks. In our preprocessing, we used the speaker's first utterance, and mapped 'afraid' to 'fear' and included 'faithful' under the 'trust' category to compensate for the deficiency of samples for the latter.

SuperEmotion (de Fortuny, 2025) is a unified dataset merging existing sources into Shaver's six basic emotions and Neutral (seven classes). We employed subsets of GoEmotions (Demszky et al., 2020) and CrowdFlower (Van Pelt and Sorokin, 2012) from this collection. Although GoEmotions is highly popular, it is characterized by significant class imbalance, particularly regarding the scarcity of negative samples.

C.2 Synthetic Emotion Dataset

Due to the scarcity of datasets fully covering the circumplex emotion model, we employ an existing dataset synthesis framework to construct a new dataset specifically for our experiments. This approach mitigates the risk of representation space distortion caused by missing labels while ensuring a diverse range of emotional texts.

PersonaGen (Inoshita and Harada, 2025) is a framework for generating diverse, high-fidelity text. It assigns persona attributes using census-based statistical distributions and utilizes LLMs to validate the consistency of these attribute combinations. This process ensures the creation of realistic personas that are highly likely to exist in reality.

Figure 8 shows the PersonaGen prompt template. We employed seven distinct LLMs to guarantee diversity in the synthesized dataset:

- `gpt-5-nano` (OpenAI, 2025)

PersonaGen prompt template

```
###System prompt:### You are a roleplay AI.
###User prompt:### Roleplay as the persona
below.
Speak 1-2 natural English sentences expressing
the emotion.

[Persona] {'Age', 'Job', 'Education',
           'Location', 'Family'}
[Scene] {'Scene'}
[Style] {'Style'}
[Emotion] {'Emotion'}
Output:
```

Figure 8: Prompt template used in the dataset synthesis.

- `Ministral-3-14B-Instruct-2512` (Mistral AI, 2025)
- `Olmo-3-7B-Instruct` (Allen Institute for AI, 2025)
- `Qwen3-30B-A3B-Instruct-2507` (Qwen, 2025)
- `Phi-4-mini-instruct` (Microsoft, 2025)
- `gemma-3-27b-it` (DeepMind, 2025)
- `Llama-3.3-70B-Instruct` (Meta AI, 2024)

After randomly sampling from the model outputs, we removed duplicates and filtered for a length range of 3–50 tokens, ultimately collecting 600 texts for each emotion category.

C.3 Integration with ECM

The final dataset specifications are presented in Table 2. Emotion labels for each dataset were filtered based on the configuration shown in Figure 3. We manually selected emotions that either exhibited exact name matches or possessed equivalent semantic properties. Furthermore, we balanced the emotion label distribution in each dataset through under-sampling to avoid embedding space distortion arising from class imbalance. The findings in Section 4.3 are based on these datasets.

C.4 Differences between the ECM and Russell's Circumplex Model of Affect

Notable deviations from Russell's Circumplex Model of Affect include the placement of love at 0°, calmness at 270°, and trust at 330°.

- 0° (love): While this angle typically represents pleasant emotions such as "satisfied" or "pleased," we adopted "love" because it is a predominant emotion in real-world datasets and is considered the polar opposite of "disgust."
- 270° (calmness): This angle normally corresponds to "tired" or "sleepy." However, to the best of our knowledge, no text datasets annotated with these specific labels exist. Consequently, we substituted them with "calmness," which serves as the closest proxy for a deactivated emotional state.
- 330° (trust): While "serene" is the standard representative for this angle, it is rare within corpora. We therefore adopted "trust" as it approximates the polar opposite of "fear."

As discussed in the *Limitations* section, since our experimental results are primarily driven by dimensional constraints, the specific arrangement or selection of emotion types does not significantly alter our conclusions. However, effectively capturing rare emotions within training corpora remains a challenge for future work. We posit that CircularCSE, with its ability to model inter-emotional relationships, is particularly advantageous for learning low-frequency emotion labels.

D Implementation Details

Table 3 lists the hyperparameters used for training. These values were determined through preliminary experiments involving various combinations, selected to ensure training stability. For any parameters not explicitly listed, we employ the default settings. In our experiments, we used a consistent margin m across all backbone models. Sensitivity analysis revealed that varying m had a negligible impact on results. This is due to the gradient signal distribution: for a batch size of 128 (approx. 120 for simplicity), the expected number of positive pairs is $\binom{10}{2} \times 12 = 540$, while negative pairs account for $\binom{120}{2} - 540 = 6600$. Since positive pairs represent only $\approx 7.5\%$ of the total, the gradient signal is dominated by negative pairs, diminishing the sensitivity of the margin m.

E Overall Results

Table 4 presents the results obtained by training six distinct backbone models across each dataset, applying six combinations of head architectures and loss functions. Observing the trends across datasets, we find that all methods consistently demonstrate significantly high accuracy on PersonaGen, whereas performance remains low on Empathetic Dialogue and SuperEmotion.

This suggests that compared to real-world data, synthetic data lacks diversity and ambiguity, making it an inherently easier dataset for emotion prediction. Conversely, Empathetic Dialogue and SuperEmotion originate from conversations and social media, respectively; these results imply that accurate emotion prediction on such datasets is difficult without clear context.

F Validation across Multilingual Datasets

Table 5 presents the evaluation results on the multilingual subsets (French: Fr, Italian: It, and Dutch: Nl) of the Emolit dataset using the multilingual models mE5, Qwen3-Embedding-4B, and Llama-3.2-3B.. Although the discriminative performance of all models is lower compared to the English version, the overall trend remains consistent, with a negative correlation observed between V-Measure and CD-r. These results suggest that the tradeoff between discriminative accuracy and interpretability is a structural issue and is language-independent.

G Analysis of Parameter Scaling for Decoder-only Models

Recent decoder-only models benefit from large parameter sizes, with open-source models in the tens-of-billions scale being widely utilized. In light of this, we conducted additional experiments with Qwen3-14B to evaluate the impact of parameter scale within decoder-only architectures. Table 6 presents the evaluation results for three decoder-only models of varying sizes. Our findings indicate that while information loss due to normalization tends to be mitigated as model size increases , the structural dilemma remains persistent. We offer the following theoretical perspectives:

Structural Dilemma. This trade-off is inherently independent of parameter size. Maximizing discriminative accuracy requires widening the margin between adjacent labels, which fundamentally treats similar concepts as unrelated from a human perspective. Since any model must adopt a specific margin, these conflicting objectives remain a challenge regardless of capacity.

Information Loss. In Transformers with residual connections, the hidden representation norm is expected to grow at $O(\sqrt{L})$, where L is the number of layers. While larger models with more layers could potentially suffer greater information loss during normalization, many recent architectures employ normalization schemes that maintain constant norms. Therefore, the extent of loss depends more on specific architectural tuning than on scale alone.

Given that decoder-only models are more diverse and possess distinct properties compared to encoder models, further investigation is required.

H Robustness Experiments

We illustrate the dimensionality robustness of mE5, Qwen3-Embedding-4B, and Llama-3.2-3B in Figure 9, and the robustness to label count variations in Figure 11. Experiments regarding variations in the number of labels were conducted using the Emolit dataset. Figure 10 illustrates the arrangement and types of emotions within the circular model. We manually determined these label placements based on Russell's Circumplex Model (Russell, 1980; Yik et al., 2011).

I Lower Bound of SINCERE

In this section, we provide a simplified proof regarding the lower bound of SINCERE. To simplify the derivation, we posit the following assumption.

Assumption 1 (Class Prototype).
The representation of each class $y_i \in \mathcal{E} = \{y_1, \ldots, y_E\}$ *is concentrated at a single point, denoted by the unit vector* $e_i \in \mathbb{S}^{d-1}$, $\|e_i\| = 1$, *and the positive example corresponding to an anchor is always* e_i.

Assumption 2 (Class Uniformity).
The dataset is class-balanced, i.e., $p(y = i) = \frac{1}{E}$. *We analyze the population (expected) SINCERE objective, under which an anchor from class* y_i *contrasts against the representations of all other classes in expectation. Consequently, the negative set consists of the* $E - 1$ *class prototypes* $\{e_j\}_{j \neq i}$.
Under these assumptions, given the formulation:

$$\mathcal{L}_{\text{SINCERE}} = \mathbb{E}_i \mathbb{E}_{j\sim\mathcal{P}} \left[-\log \frac{\exp(e_i^T e_j/\tau)}{\exp(e_i^T e_j/\tau) + \sum_{k\in\mathcal{N}} \exp(e_i^T e_k/\tau)} \right] \tag{21}$$

the following holds:

Theorem.
When there are E *emotion classes, the lower bound of* $\mathcal{L}_{SINCERE}$ *is achieved when the class representations form a regular simplex, resulting in a positive-negative inner product (similarity) of* $-\frac{1}{E-1}$.
Here, a regular simplex is a configuration where the class representations satisfy the following conditions:

- Zero Sum: $\sum_{i=1}^{E} e_i = 0$
- Equal Norms
- Equal Pairwise Inner Products

It is well-established that Cross Entropy Loss and Supervised Contrastive Loss also converge to a regular simplex under certain conditions (Papyan et al., 2020; Graf et al., 2021). For a more rigorous proof, please refer to (Papyan et al., 2020; Graf et al., 2021).

Proof.

$$
\begin{aligned}
&\mathcal{L}_{\text{SINCERE}} = \\
&\mathbb{E}_i\mathbb{E}_{j\sim\mathcal{P}}\left[-\log\frac{\exp(e_i^T e_j/\tau)}{\exp(e_i^T e_j/\tau)+\sum_{k\in\mathcal{N}}\exp(e_i^T e_k/\tau)}\right] \\
&= \mathbb{E}_i\mathbb{E}_{j\sim\mathcal{P}}\left[-\log\frac{\exp(1/\tau)}{\exp(1/\tau)+\sum_{k\in\mathcal{N}}\exp(e_i^T e_k/\tau)}\right] \\
&= \mathbb{E}_i\mathbb{E}_{j\sim\mathcal{P}}\left[\log\left(1+\sum_{k\in\mathcal{N}}\exp\left(\frac{e_i^T e_k-1}{\tau}\right)\right)\right]
\end{aligned} \tag{22}
$$

Given that $\tau > 0$, $\mathcal{L}_{\text{SINCERE}}$ is monotonically increasing with respect to $\exp(e_i^T e_k/\tau)$. Consequently, minimizing $\mathcal{L}_{\text{SINCERE}}$ is equivalent to minimizing the negative sample term $\sum_{k\in\mathcal{N}}\exp(e_i^T e_k/\tau)$ for each sample. For the set of class vectors $\{e_1,\ldots,e_E\}\subset\mathbb{S}^{d-1}$, the squared L^2 norm of their sum is given by:

$$
\left\|\sum_{i=1}^{E} e_i\right\|^2 = \sum_{i=1}^{E}\|e_i\|^2+\sum_{i\neq k}e_i^T e_k = E+\sum_{i\neq k}e_i^T e_k \geq 0 \tag{23}
$$

Therefore, it follows that: $\sum_{i\neq k}e_i^T e_k \geq -E$. Let $\overline{e_i^T e_k}$ denote the mean of the inner products for all pairs. Using the relationship $\sum_{i\neq k}e_i^T e_k = E(E-1)\overline{e_i^T e_k}$, we derive the inequality

$$
\overline{e_i^T e_k} \geq -\frac{1}{E-1} \tag{24}
$$

This implies that the global average of the inner products between a class vector and vectors of other classes is lower-bounded by $-\frac{1}{E-1}$. We consider the lower bound of the negative term. Since the exponential function is convex, applying Jensen's inequality:

$$
\sum_{k\in\mathcal{N}}\frac{1}{E-1}\exp\left(\frac{e_i^T e_k}{\tau}\right) \geq \exp\left(\sum_{k\in\mathcal{N}}\frac{1}{E-1}\frac{e_i^T e_k}{\tau}\right) \tag{25}
$$

Based on Assumption 2, we derive:

$$
\sum_{k\in\mathcal{N}}\exp\left(\frac{e_i^T e_k}{\tau}\right) \geq (E-1)\exp\left(\frac{\overline{e_i^T e_k}}{\tau}\right) \tag{26}
$$

Applying Jensen's inequality again,

$$
\begin{aligned}
&\sum_{i=1}^{E}\sum_{k\in\mathcal{N}}\frac{1}{E}\exp\left(\frac{e_i^T e_k}{\tau}\right) \\
&\geq (E-1)\sum_{i=1}^{E}\frac{1}{E}\exp\left(\frac{\overline{e_i^T e_k}}{\tau}\right) \\
&\geq (E-1)\exp\left(\sum_{i=1}^{E}\frac{1}{E}\frac{\overline{e_i^T e_k}}{\tau}\right)
\end{aligned} \tag{27}
$$

Based on Equation (24),

$$
\sum_{i=1}^{E}\sum_{k\in\mathcal{N}}\exp\left(\frac{e_i^T e_k}{\tau}\right) \geq E(E-1)\exp\left(\frac{-1}{(E-1)\tau}\right) \tag{28}
$$

This is the lower bound of SINCERE Loss based on the Assumption. The condition for equality in Jensen's inequality dictates that all pairwise inner products between class vectors must be equal, and their mean must be $-\frac{1}{E-1}$. Consequently, the class similarity between positive and negative samples required to achieve the lower bound is given by

$$
e_i^T e_k = -\frac{1}{E-1}\ \forall\, i\neq k \tag{29}
$$

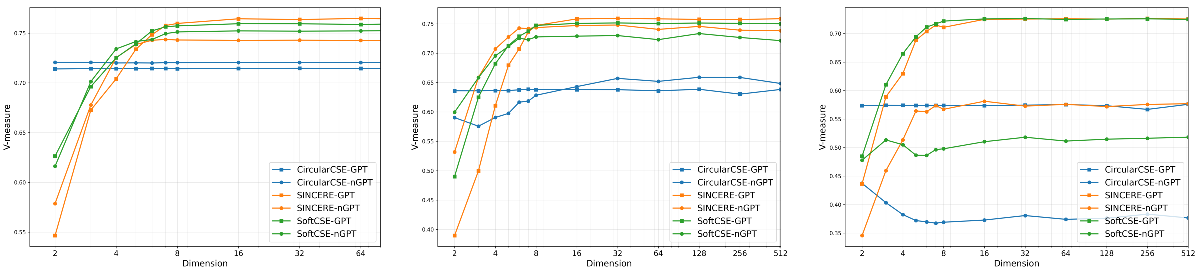
Furthermore, the condition for equality in Equation (24) is given by:

$$
\left\|\sum_{i=1}^{E}e_i\right\|^2 = 0 \iff \sum_{i=1}^{E}e_i = 0 \tag{30}
$$

Summarizing these conditions, it is shown that the geometry required to achieve the minimum loss satisfies the properties of a regular simplex.

J MDS Visualization

Figures 12, 13, and 14 display the visualization results of the embedding representations from each module of mE5, Qwen3-Embedding-4B, Llama-3.2-3B on the Emolit dataset, projected using MDS. As an overall trend, the embedding representations of mE5 form dense clusters, whereas those of Qwen3-Embedding-4B and Llama-3.2-3B are more dispersed. This difference likely reflects the impact of the training strategy, as BERT-like encoders (mE5) were fully fine-tuned while the LLM backbones (Qwen3 and Llama) remained frozen. For Llama-3.2-3B, the outputs from the standalone Attention or MLP layers exhibit higher discriminative accuracy than the representations after final residual connections. This suggests that the low inherent discriminability of the original embeddings significantly influences the final performance.



(a) mE5 (b) Qwen3-Embedding-4B (c) Llama-3.2-3B

Figure 9: Robustness to dimensionality reduction.

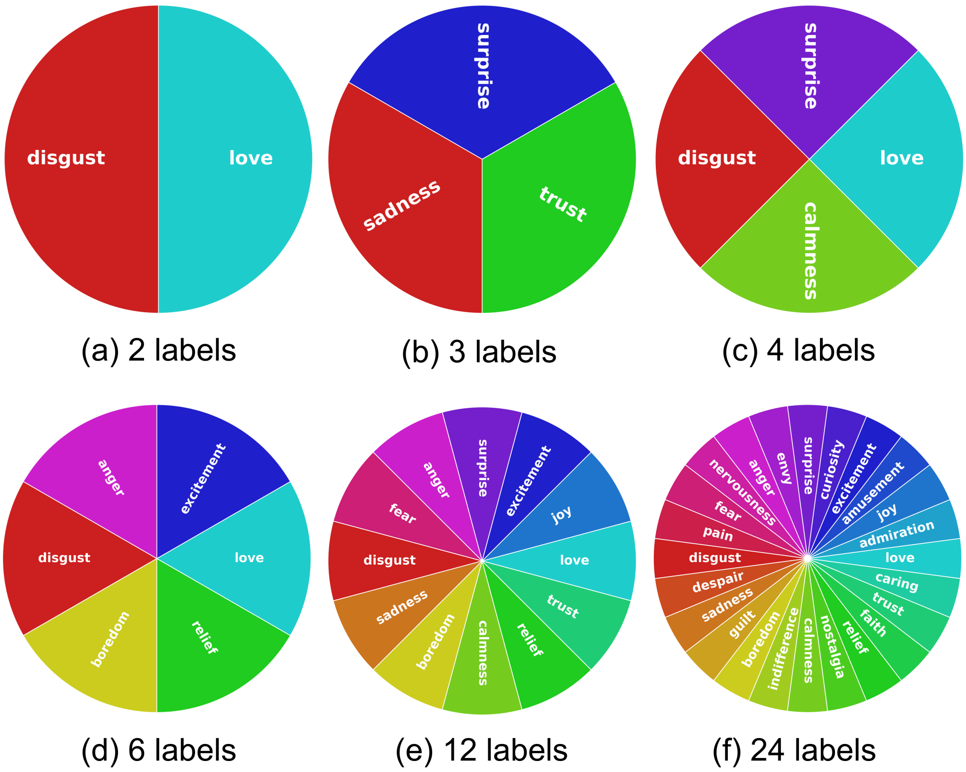
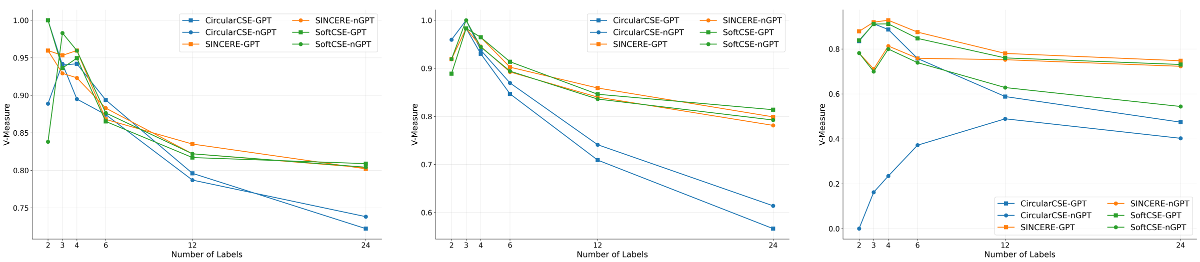


Figure 10: Empirical circumplex models used in label variation experiments



(a) mE5 (b) Qwen3-Embedding-4B (c) Llama-3.2-3B

Figure 11: Robustness to the number of emotion labels.

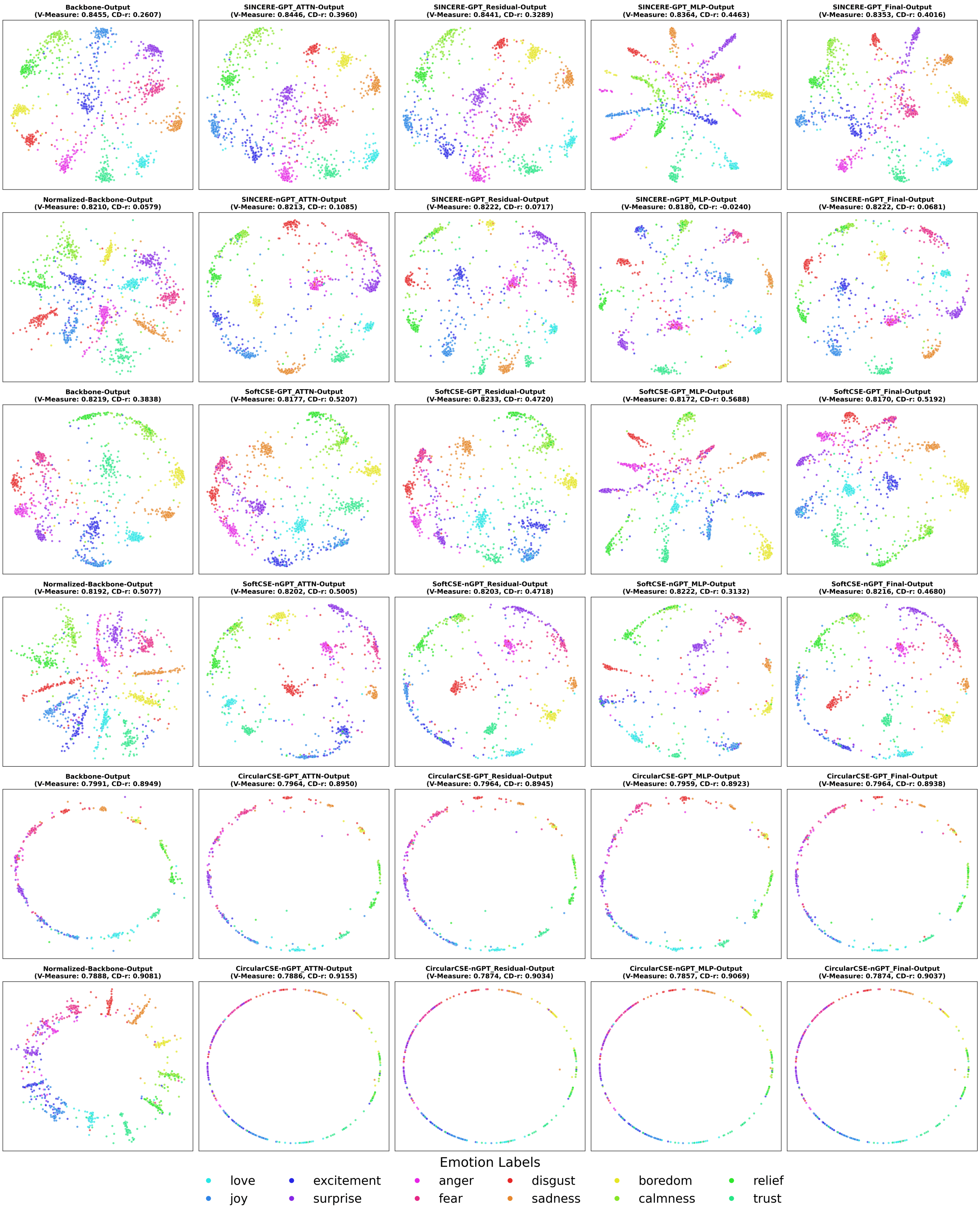


Figure 12: MDS visualization of mE5

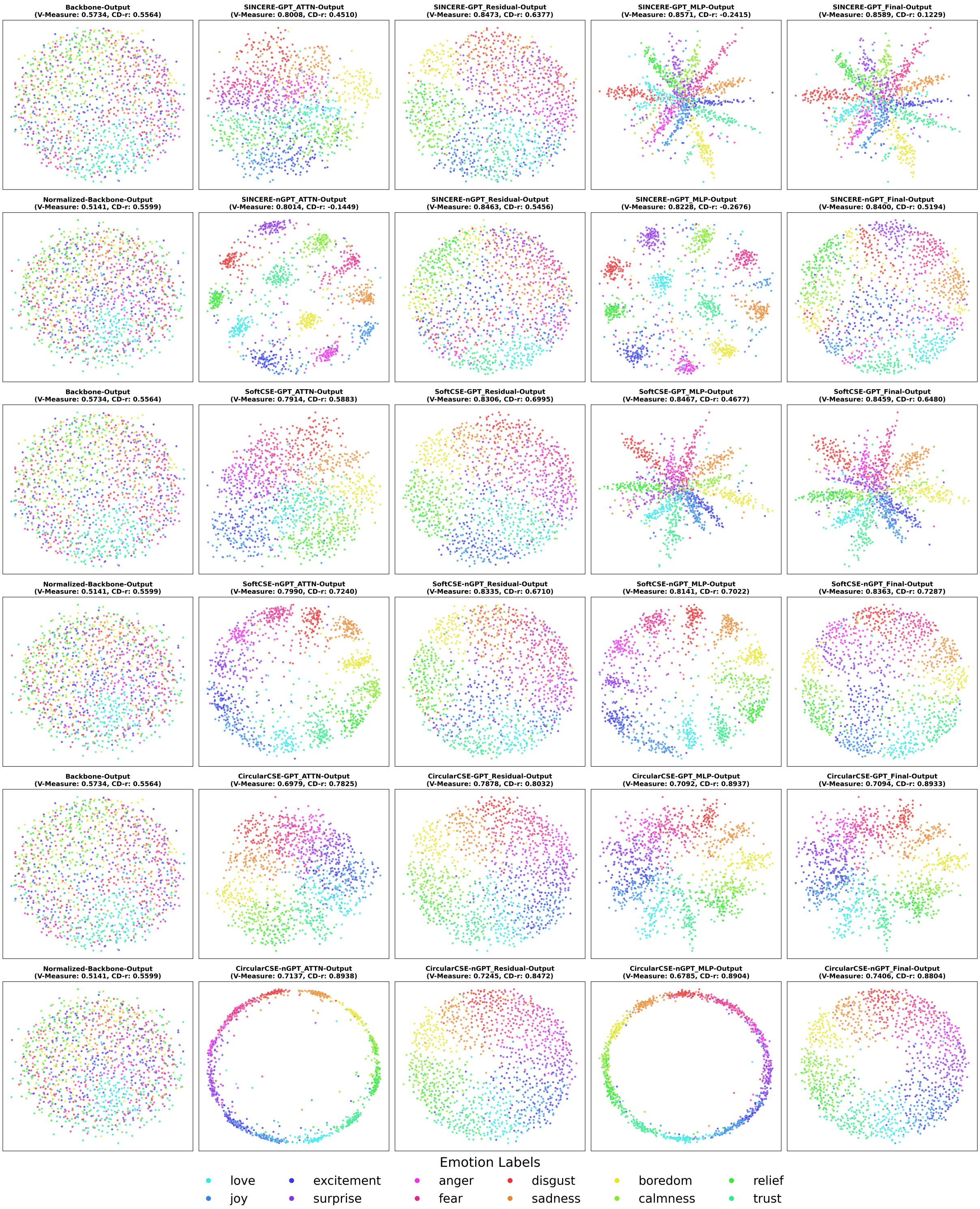


Figure 13: MDS visualization of Qwen3-Embedding-4B

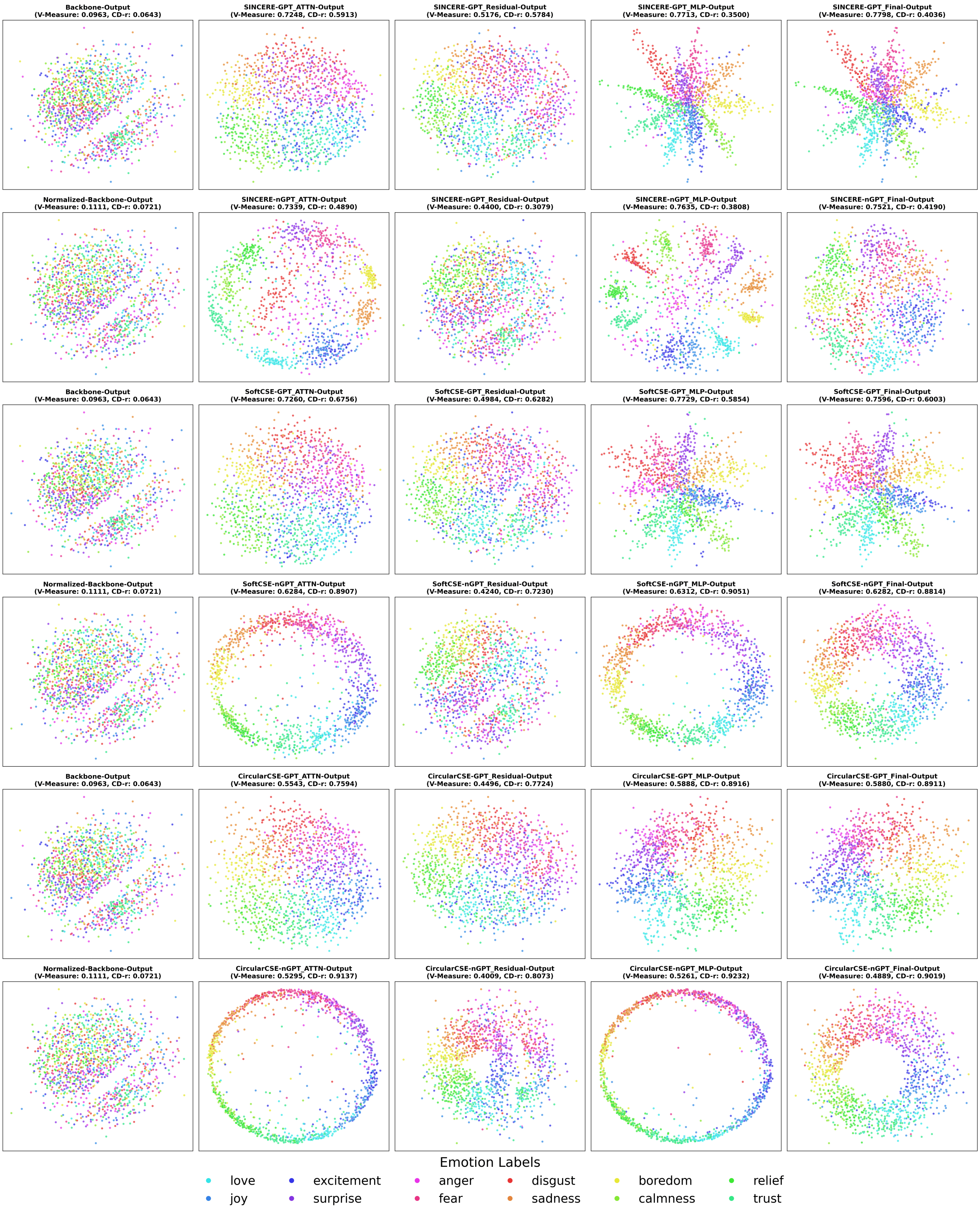


Figure 14: MDS visualization of Llama-3.2-3B

Base Model	Method	Emolit		Empathetic Dialogue		SuperEmotion		PersonaGen		Average	
		V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r
mE5	Pretrained	0.437	0.631	0.237	0.512	0.170	0.554	0.525	0.599	0.342	0.574
	SINCERE										
	- GPT	**0.835**	0.402	**0.667**	0.413	0.621	0.343	0.917	0.112	**0.760**	0.317
	- nGPT	0.822	0.068	0.643	0.504	0.595	0.186	0.916	0.125	0.744	0.221
	SoftCSE										
	- GPT	0.817	0.519	0.649	0.514	**0.636**	0.574	0.915	0.302	0.755	0.477
	- nGPT	0.822	0.468	0.660	**0.575**	0.621	0.559	0.909	0.395	0.753	0.499
	CircularCSE										
	- GPT	0.796	0.894	0.593	0.551	0.570	0.688	0.909	0.896	0.717	0.757
	- nGPT	0.787	**0.904**	0.610	0.537	0.565	**0.710**	**0.918**	**0.907**	0.720	**0.764**
mxbai	Pretrained	0.603	0.740	0.502	0.326	0.285	0.490	0.670	0.625	0.515	0.545
	SINCERE										
	- GPT	0.831	0.293	0.640	0.408	0.609	0.212	0.909	0.234	0.747	0.287
	- nGPT	0.825	0.159	0.648	0.492	0.595	0.482	0.912	0.226	0.745	0.340
	SoftCSE										
	- GPT	**0.841**	0.645	**0.657**	0.584	0.595	0.479	0.912	0.686	**0.751**	0.599
	- nGPT	0.826	0.577	0.628	**0.636**	**0.610**	0.506	**0.927**	0.619	0.748	0.585
	CircularCSE										
	- GPT	0.840	**0.893**	0.623	0.513	0.568	0.700	0.912	0.899	0.736	0.751
	- nGPT	0.793	0.886	0.604	0.520	0.572	**0.705**	0.907	**0.910**	0.719	**0.755**
Qwen3 -Embedding -4B	Pretrained	0.579	0.556	0.435	0.442	0.273	0.440	0.691	0.651	0.495	0.522
	SINCERE										
	- GPT	**0.859**	0.122	0.651	0.595	**0.588**	0.490	0.927	0.013	**0.756**	0.305
	- nGPT	0.840	0.519	**0.673**	0.471	0.554	0.579	0.895	0.613	0.741	0.545
	SoftCSE										
	- GPT	0.846	0.647	0.646	**0.673**	0.578	0.487	**0.932**	0.403	0.751	0.552
	- nGPT	0.836	0.729	0.643	0.587	0.523	**0.731**	0.890	0.786	0.723	0.708
	CircularCSE										
	- GPT	0.709	**0.893**	0.530	0.520	0.471	0.675	0.860	0.900	0.643	0.747
	- nGPT	0.741	0.880	0.603	0.560	0.443	0.660	0.850	**0.911**	0.659	**0.753**
Llama -Embed -Nemotron -8B	Pretrained	0.271	0.373	0.227	0.258	0.160	0.118	0.494	0.715	0.288	0.366
	SINCERE										
	- GPT	0.841	-0.041	**0.696**	0.591	0.601	0.461	0.940	-0.293	0.769	0.180
	- nGPT	0.834	0.386	0.693	0.345	**0.627**	0.393	0.938	0.617	**0.773**	0.435
	SoftCSE										
	- GPT	**0.847**	0.577	0.685	**0.691**	0.600	0.423	**0.941**	0.339	0.768	0.508
	- nGPT	**0.847**	0.735	0.692	0.591	0.599	0.657	0.938	0.756	0.769	0.685
	CircularCSE										
	- GPT	0.710	**0.902**	0.603	0.508	0.519	**0.675**	0.903	0.908	0.684	**0.748**
	- nGPT	0.685	0.901	0.605	0.483	0.447	0.671	0.880	**0.913**	0.654	0.742
Llama-3.2 -3B	Pretrained	0.100	0.064	0.069	0.067	0.050	0.101	0.159	0.636	0.094	0.217
	SINCERE										
	- GPT	**0.780**	0.404	**0.627**	0.548	**0.566**	0.363	**0.926**	0.116	**0.725**	0.358
	- nGPT	0.752	0.419	0.357	0.262	0.296	0.566	0.903	0.451	0.577	0.425
	SoftCSE										
	- GPT	0.760	0.600	0.616	**0.723**	0.549	0.476	0.913	0.391	0.710	0.548
	- nGPT	0.628	0.881	0.362	0.522	0.276	0.676	0.798	0.833	0.516	**0.728**
	CircularCSE										
	- GPT	0.588	0.891	0.484	0.457	0.425	0.657	0.817	**0.909**	0.579	**0.728**
	- nGPT	0.489	**0.902**	0.250	0.465	0.196	0.564	0.592	0.900	0.382	0.708
Olmo-3 -7B	Pretrained	0.047	-0.005	0.024	-0.036	0.048	-0.014	0.133	0.265	0.063	0.053
	SINCERE										
	- GPT	**0.805**	0.424	**0.615**	0.485	**0.581**	0.403	0.929	0.169	**0.732**	0.370
	- nGPT	0.786	0.444	0.409	0.297	0.376	0.467	0.913	0.243	0.621	0.363
	SoftCSE										
	- GPT	**0.805**	0.578	0.609	**0.579**	0.561	0.425	**0.941**	0.380	0.729	0.490
	- nGPT	0.742	0.800	0.351	0.546	0.335	0.636	0.898	0.701	0.582	0.671
	CircularCSE										
	- GPT	0.630	**0.903**	0.484	0.487	0.414	**0.682**	0.844	0.902	0.593	**0.744**
	- nGPT	0.494	0.876	0.248	0.418	0.254	0.533	0.605	**0.914**	0.400	0.685

Table 4: V-Measure and Circumplex Distance correlation (CD-r) across datasets and models. **Bold** indicates the maximum value and underlined indicates the minimum value across different configurations for each model.

Base Model	Method	En		Fr		It		Nl		Average	
		V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r
mE5	Pretrained	0.437	0.631	0.217	0.468	0.256	0.429	0.270	0.506	0.295	0.509
	SINCERE										
	- GPT	**0.835**	0.402	**0.767**	0.521	0.735	0.320	0.742	0.355	**0.770**	0.400
	- nGPT	0.822	0.068	0.745	0.342	0.750	0.314	**0.754**	0.249	0.768	0.243
	SoftCSE										
	- GPT	0.817	0.519	0.758	0.656	0.761	0.581	0.741	0.605	0.769	0.590
	- nGPT	0.822	0.468	0.742	0.695	**0.765**	0.582	0.748	0.549	0.769	0.574
	CircularCSE										
	- GPT	0.796	0.894	0.742	0.896	0.735	0.891	0.747	0.872	0.755	0.888
	- nGPT	0.787	**0.904**	0.704	**0.900**	0.734	**0.889**	0.709	**0.897**	0.734	**0.898**
Qwen3 -Embedding -4B	Pretrained	0.579	0.556	0.438	0.399	0.495	0.378	0.358	0.399	0.468	0.433
	SINCERE										
	- GPT	**0.859**	0.122	**0.782**	0.171	0.766	0.272	**0.773**	0.273	**0.795**	0.210
	- nGPT	0.840	0.519	0.767	0.459	0.748	0.473	0.758	0.492	0.778	0.486
	SoftCSE										
	- GPT	0.846	0.647	0.774	0.645	**0.773**	0.684	0.775	0.709	0.792	0.671
	- nGPT	0.836	0.729	0.775	0.688	0.743	0.721	0.732	0.702	0.772	0.710
	CircularCSE										
	- GPT	0.709	**0.893**	0.641	**0.894**	0.601	**0.899**	0.587	**0.902**	0.635	**0.897**
	- nGPT	0.741	0.880	0.683	0.858	0.647	0.886	0.638	0.879	0.677	0.876
Llama-3.2 -3B	Pretrained	0.100	0.064	0.083	0.018	0.082	0.017	0.083	0.011	0.087	0.028
	SINCERE										
	- GPT	**0.780**	0.404	0.677	0.423	**0.680**	0.365	**0.683**	0.553	**0.705**	0.436
	- nGPT	0.752	0.419	0.613	0.613	0.588	0.587	0.576	0.600	0.632	0.555
	SoftCSE										
	- GPT	0.760	0.600	**0.704**	0.573	0.652	0.614	0.680	0.649	0.699	0.609
	- nGPT	0.628	0.881	0.478	0.857	0.466	0.861	0.439	0.870	0.503	0.867
	CircularCSE										
	- GPT	0.588	0.891	0.477	**0.897**	0.465	**0.897**	0.437	**0.912**	0.492	**0.899**
	- nGPT	0.489	**0.902**	0.361	0.647	0.374	0.745	0.367	0.740	0.398	0.759

Table 5: Multilingual results for V-Measure and CD-r across the En, Fr, It, and Nl subsets of the Emolit dataset.

Training Objective	Head Arch.	Llama-3.2-3B		OLMo-3-7B		Qwen3-14B	
		V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r
Pretrained		0.094	0.217	0.063	0.053	0.053	0.157
SINCERE	- GPT	**0.725**	0.358	**0.732**	0.370	**0.694**	0.306
	- nGPT	0.577	0.425	0.621	0.363	0.621	0.402
SoftCSE	- GPT	0.710	0.548	0.729	0.490	0.682	0.518
	- nGPT	0.516	**0.728**	0.582	0.671	0.567	0.655
CircularCSE	- GPT	0.579	**0.728**	0.593	**0.744**	0.542	**0.757**
	- nGPT	0.382	0.708	0.400	0.685	0.388	0.663

Table 6: Comparison across decoder-only models with different parameter sizes.