# 感情は環状?

山内悠輔 [1]　相澤彰子 [2,1]
[1] 東京大学 大学院情報理工学系研究科　[2] 国立情報学研究所
{yamauchi_y, aizawa}@nii.ac.jp

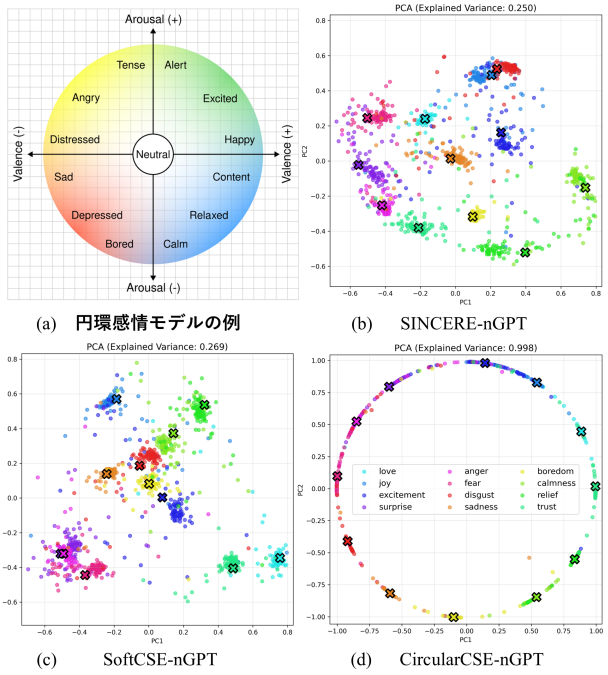
## 概要

心理学では感情を円形構造として定義する考えが一般的で，深層学習においてもこの考え方に沿ってモデルの感情表現の分析・解釈が行われてきた．しかし，言語モデルの埋め込み表現に直接取り入れた研究は少なく，多様体構造としての有効性は未知のままである．本研究では超球面上で対照学習を行うことで，言語モデルの埋め込み空間に明示的に感情円環を表現し，円形構造の是非を検討する．従来手法と比較することで解釈性と識別精度のトレードオフを実験・理論的に示し，人間の解釈をモデル設計に取り入れる際のジレンマを明らかにする．

## 1　はじめに

線形表現仮説 [1] と重ね合わせ仮説 [2] は言語モデルが概念をどのように表現するか解釈する上で広く用いられてきた．これらの仮説は言語モデルが様々な概念を，互いに干渉し合わないように直交する部分空間上で線形に表現するというものであり，実際に特定の方向のベクトルを推論時に加算することでモデルの振る舞いに介入できることが報告されている [3, 4]．一方で非線形な構造を持つ概念も発見されており，代表例として週や月などの周期性を持つ概念や mod 加算における数値表現は円形に表現されることがあげられる [5, 6, 7]．こうした非線形的な概念が他にも存在するかどうかについて，近年注目が集まっている [8]．

心理学では人間の感情構造をモデル化したものとして，円環感情モデルが広く用いられてきた (図 1(a)) [9, 10]．このモデルは機械学習モデルが人間の感情構造を反映しているかどうか検証するために，分析時の参照モデルとしてよく利用されている．しかし，既存研究は各感情表現の重心間距離や感情ラベルの共起予測確率をもとに感情間の関係性を分析するものが多く，事後的な分析かつ巨視的な考察にとどまっている [11, 12, 13]．言語モデルが人間の感情構造を反映しているか，また反映すべきかを議論するには，埋め込み空間上に多様体構造を明示的に構成する必要がある．




(a) 円環感情モデルの例　(b) SINCERE-nGPT　(c) SoftCSE-nGPT　(d) CircularCSE-nGPT

図 1: (a) 心理学で提唱される円環感情モデル例[1] (b)(c)(d) 本研究で訓練した言語モデルの埋め込み空間を PCA により次元削減したプロット図

そこで，本研究では言語モデルの埋め込み空間上に円環感情モデルを表現するように誘導し，その有効性を検討する．超球面上で学習を行うように設計された nGPT ヘッドと感情ラベル間のペアワイズ距離を取り入れた損失関数 CircularCSE を組み合わせて学習を行い，識別精度および人間の認識に沿うかどうかの異なる観点で，従来手法との比較を行う．結果として円環感情モデルを明示的に再現したモデルは高い解釈性や可視性を持つが，従来手法と比べて感情ラベルの識別精度が低下することを示す (図 1(d))．この現象を識別マージンの観点から説明

1) ラッセルの円環感情モデルに基づく．図は Wikimedia Commons から引用 (https://en.wikipedia.org/wiki/File:Circumplex_model_of_emotion.svg).

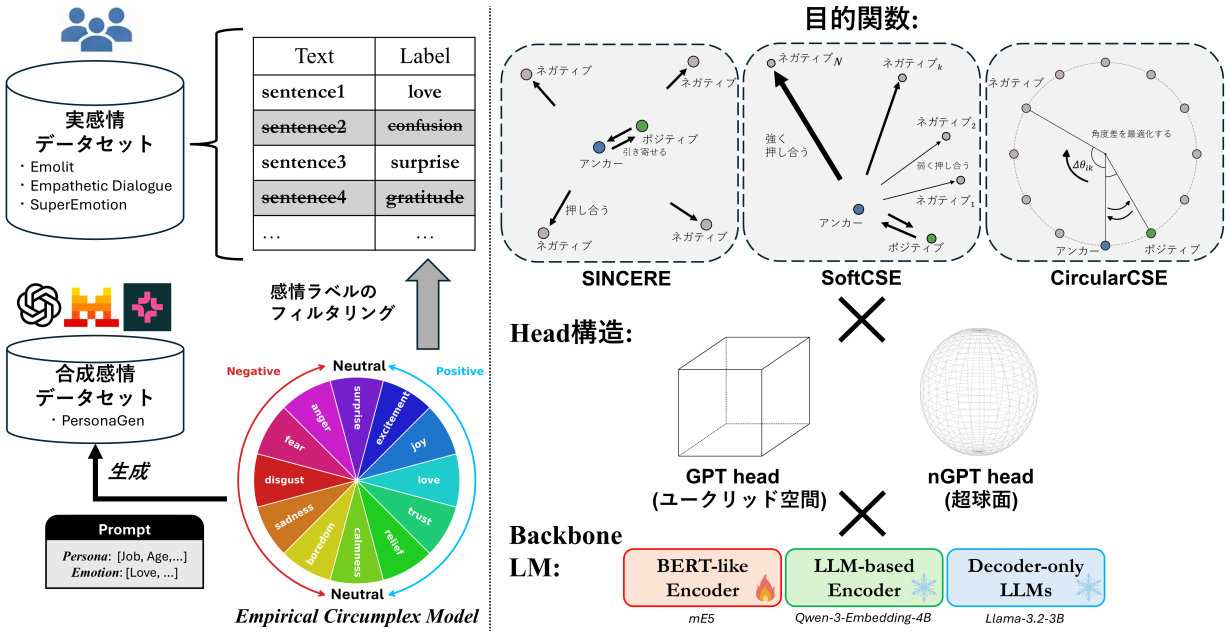



図 2: **本実験フレームワークの概要図.** (左) 円環感情モデルに対応する感情ラベルを持つデータセットを構築. (右) 異なる Backbone モデル，Head 構造，目的関数の組み合わせを適用して訓練を行う

し，人間が求める解釈性と深層学習が求める識別精度の間にはトレードオフが存在することを明らかにする.

## 2 感情表現の学習

図 2 に本研究の実験フレームワークの概要図を示す. 実験では円環上に等間隔に配置されると想定される感情テキストデータを収集し (図 2(左)), そのデータセットを用いて感情表現の学習を行う (図 2(右)). 本実験ではラッセルの円環感情モデルに倣い [9, 10]，12 種類の感情ラベルから構成される円環感情モデルを定義する (以降 Empirical Circumplex Model，または ECM と呼称する). このモデルに基づいて，実感情データセットからサンプルを抽出し，さらに合成感情データセットを構築する. 以下，各感情データセット（$D$）について $N$ 個のテキスト（$x_i$）-ラベル（$y_i$）ペアを使用するものとし，これを $D = \{(x_i, y_i) \mid y_i \in \mathcal{E}\}_{i=1}^{N}$ と表記する. ここで $\mathcal{E}$ はラベルの種類が $E := |\mathcal{E}|$ の感情ラベルの全集合と定義する.

### 2.1 Head 構造

事前学習済みの backbone モデルの最終層に対して 1 層の Transformer block を追加し，その出力の埋め込み空間上で感情を表現する. $t$ を入力系列 $[1, \ldots, T]$ のインデックス，$d$ をモデルの埋め込み次元数とする. 従来の Transformer block は attention モジュール (ATTN)，多層パーセプトロン (MLP)，正規化層 (RMSNorm) から構成される:

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

ここで $h_t, h_t', h_t'' \in \mathbb{R}^d$ である. この空間は $d$ 次元の制約の無いユークリッド空間であり，埋め込み表現のノルムも意味を持つため円環構造を自然に形成するのは難しい. そこで，出力を超球面上に制約する normalized Transformer Block (nGPT)[14] を使用する:

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

Norm() は $\ell_2$ 正規化，$h_t, h_t' \in \mathbb{S}^{d-1}$，$\alpha_A, \alpha_M \in \mathbb{R}^d$ は学習可能なパラメータである. nGPT は正規化モジュールを取り除く代わりに，すべての隠れ表現と重みを各次元で正規化している. この結果，各モジュールの出力は超球面上に配置され，更新は測地線に沿って行われる. そのため，block 内の処理は疑似的にリーマン多様体上の最適化とみなせる. 本実験では出力系列の埋め込み表現 $h_{1:T}$ に対してプーリング処理を行い，最終的な埋め込み表現 $e$ を得る:

$$e = \mathrm{Norm}(\mathrm{Pooling}(h_{1:T})), \tag{2}$$

ここで $e \in \mathbb{S}^{d-1}, h_{1:T} \in \mathbb{R}^{T \times d}$ である. 比較のため，従来の Transformer block もベースラインとして学習する.

## 2.2 目的関数

本実験では 3 種類の損失関数を用いて対照学習を行う．我々は機械学習モデルの対照学習で一般的な Supervised Contrastive Loss [15] を拡張した Supervised InfoNCE REvisited (SINCERE) loss [16] をベースラインとする．この損失関数はバッチ $\mathcal{B}$(サイズが $B := |\mathcal{B}|$) 内の全てのポジティブペアの損失の平均を計算する:

$$\mathcal{L}_{\text{SINCERE}} = \frac{1}{B} \sum_{i \in \mathcal{B}} \left( \frac{-1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \log \frac{\exp(e_i^T e_j / \tau)}{Z_i} \right) \tag{3}$$

ここで $\tau$ は温度，$\mathcal{P}$ はアンカー文 $x_i$ の in-batch ポジティブ集合，$Z_i$ はポジティブサンプルと in-batch ネガティブサンプル項を表す:

$$Z_i = \exp(e_i^T e_j / \tau) + \sum_{k \in \mathcal{N}} \exp(e_i^T e_k / \tau) \tag{4}$$

($\mathcal{N}$ は in-batch ネガティブ集合)．SINCERE はバッチ内の全てのサンプルに対して損失を計算するため，ポジティブとネガティブサンプルの間で強力な分離力が働く．一方で，全てのネガティブサンプルに対して等しく押し合う力がかかるため，アンカーと個々のネガティブサンプル間の距離の違いを表現できない．ECM では近い感情が隣接し，反対の感情が対極に配置されることから，損失関数も感情ラベル間の距離を反映することが望ましい．したがって，円環構造を再現する損失関数では ECM 上のペアワイズ距離を制約として取り入れる．SoftCSE [5] はペアワイズ距離を弱い制約として導入し，SINCERE のネガティブ項に個別の重みを適用する:

$$\begin{aligned} Z_i &= \exp(e_i^T e_j / \tau) + \sum_{k \in \mathcal{N}} w_{ik} \exp(e_i^T e_k / \tau), \\ w_{ik} &= \frac{1 - \cos(\Delta\theta_{ik})}{\frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}} (1 - \cos(\Delta\theta_{ik}))} \end{aligned} \tag{5}$$

ここで $\Delta\theta_{ik}$ は ECM 上の角度差を表す．$w_{ik}$ は感情ラベル $y_i$ と $y_k$ のコサイン類似度に反比例し，円環上で近い感情は弱く押し合うようになる．

より強い制約として，円環上の距離を直接学習する CircularCSE を提案する:

$$\begin{aligned} \mathcal{L}_{\text{CircularCSE}} &= \frac{1}{B(B-1)} \sum_{i,j \in \mathcal{B}: i \neq j} \ell_{ij}, \\ \ell_{ij} &= \begin{cases} [\max(0, |e_i^T e_j - \cos(\Delta\theta_{ij})| - m)]^2 & \text{if } y_i = y_j \\ (e_i^T e_j - \cos(\Delta\theta_{ij}))^2 & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

ここで $m > 0$ はクラス内の分散を許容するマージンのハイパーパラメータである．

# 3 実験

## 3.1 実験設定

**データセット** 本実験では 3 つの実感情データセット（Emolit [17]，Empathetic Dialogue [18]，SuperEmotion [19]），と一つの合成感情データセット（PersonaGen [20]）を用いる．これらのデータセットは多様な感情ラベルを含むのが特徴で，各感情を空間上でどのように配置するかを評価することで人間の認識と比較を行う．我々は ECM 上の感情ラベルに適合するものを抽出・合成し，各感情ラベルごとに 500 件 (SuperEmotion のみ 450) 件を訓練セット，100 件をテストセットとした．

**モデル** アーキテクチャおよび事前学習時の目的関数により backbone モデルごとに埋め込み空間の特性が異なることを考慮し，以下の 3 種類の backbone モデル，mE5 [21]，Qwen3-Embedding-4B [22]，Llama-3.2-3B を用いる．モデルごとの詳細な訓練設定は付録 A に記載する．

## 3.2 評価指標

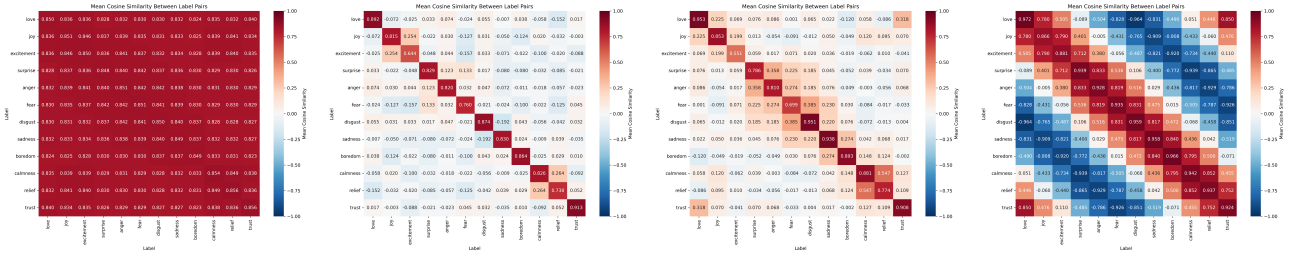
訓練された埋め込み表現がどの程度感情表現を捉えられているか評価するために，クラスタリングベースの評価を行う．我々はテストセットのサンプル全体を Spherical $k$-means [23] により感情ラベルの種類と同じクラスタ数に分け，当てはまりの良さを V-Measure [24] で評価する．また，識別精度に加えて人間の認識との一致率を評価する．先行研究 [25, 26] に基づいて，我々は ECM 上の距離 Circumplex Distance (CD) を次のように定義する: $\text{CD}(y_i, y_j) := C + \text{AngleDistance}(y_i, y_j)$．ここで $C$ は感情極性間の距離定数を表し，異なる極性間で追加の距離を加算する (例：$C_{Positive-Negative} := 4$)．$\text{AngleDistance}(y_i, y_j)$ は ECM 上での感情ラベル間のステップ数を表す．CD は同一極性の中で最も離れた感情間の距離が反対極性の最低距離よりも小さくなり，人間の正負を重要視する心理的な感情構造をより忠実に反映する．この CD を使用して，人間の認識との一致率をモデルの平均コサイン類似度とのピアソン相関 (CD-r) により計算する:

$$\text{CD-r} := \text{Pearson}(\text{CD}(y_i, y_j), 1 - \text{AvgCosSim}(y_i, y_j)) \tag{7}$$

CD-r が高いほど，モデルが感情ラベル間の関係性をより人間の認識に沿って捉えていることを表す．

表 1. データセット間の平均のパフォーマンス．モデルごとに最高の結果を太字，最低の結果を下線で示す．$V_{\text{Measure}}$ はクラスタリングの精度，CD-r は円環上の距離との相関を示す．

| 目的関数 | ヘッド構造 | mE5 | | Qwen3-Embedding-4B | | Llama-3.2-3B | |
|---|---|---|---|---|---|---|---|
| | | $V_{\text{Measure}}$ | CD-r | $V_{\text{Measure}}$ | CD-r | $V_{\text{Measure}}$ | 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 |



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

図 3: mE5 における各感情ラベルペアの平均コサイン類似度

### 3.3 結果

表 1 に各訓練手法の平均 V-Measure と CD-r を示す．目的関数に関して，SINCERE と SoftCSE は高い V-Measure スコアを示すが，CircularCSE はそれを大きく下回る．対して CD-r では，SINCERE が低く CircularCSE がより高い値となる．この現象は図 3 に示す感情ラベルペアの平均コサイン類似度から，直感的に説明できる．Pretrained モデルは異方的な空間を形成するため [27]，どの感情ラベルペアに対しても高いコサイン類似度を示す．反対に，SINCERE は異なる感情表現を直交させようとするためコサイン類似度が 0 付近となる．SoftCSE ではその働きが緩和され，CircularCSE は各感情ラベルペアを円環上に配置する．そのため，識別能力が求められる V-Measure では SINCERE が優位な一方で，ラベル間の関係性を捉える必要がある CD-r では CircularCSE が優位となる．この結果はモデルの識別能力を最大化したい深層学習と人間の認識に沿った配置を求める心理学との間で摩擦があることに由来し，配置に意味を持たせるために直交する場所以外に配置しようとすると空間の自由度を下げるため，必然的にモデルの表現能力が低下するというジレンマがあることを示している．この乖離は特に高次元，多数ラベル設定で顕著で，CircularCSE は常に 2 次元の円環が最適解となるため次元数の恩恵を受けられない (付録 B に記載)．換言すれば，これは表現に必要な次元数を削減できることを示唆している．低次元空間では直交配置が不可能なため，特定の多様体構造を仮定することが有効なアプローチになる．(図 1(b)(c)(d)).

表 1 を改めて見ると，モデル間の違いについて Llama-3.2-3B では nGPT ヘッドが GPT ヘッドよりも特に低い V-Measure を示す事が分かる．これは Decoder-only モデルが Encoder モデルに比べて文脈情報をベクトルのノルムとして埋め込んでいることを示唆している．

## 4 おわりに

本研究では心理学の円環感情モデルを忠実に再現したモデルと，深層学習で用いられる従来手法を比較し，精度と解釈性の間に明確なトレードオフが存在することを発見した．用途に応じて精度と解釈性のどちらを優先するか選択する必要がある．

## 謝辞

合成データ生成用のコードを提供してくださった井下敬翔氏に感謝する．

## 参考文献

[1] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, **Proceedings of the 41st International Conference on Machine Learning**, Vol. 235 of **Proceedings of Machine Learning Research**, pp. 39643–39666. PMLR, 21–27 Jul 2024.

[2] 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. Toy models of superposition, 2022.

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

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

[5] Haojie Zhuang, Wei Emma Zhang, Jian Yang, Weitong Chen, and Quan Z Sheng. 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**, pp. 3591–3601, 2024.

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

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

[8] Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. **arXiv preprint arXiv:2505.18235**, 2025.

[9] James A. Russell. A circumplex model of affect. **Journal of Personality and Social Psychology**, Vol. 39, No. 6, pp. 1161–1178, 12 1980.

[10] Michelle Yik, James A Russell, and James H Steiger. A 12-point circumplex structure of core affect. **Emotion**, Vol. 11, No. 4, p. 705, 2011.

[11] Xiangyu Wang and Chengqing Zong. Learning emotion category representation to detect emotion relations across languages. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 47, No. 6, pp. 4752–4767, 2025.

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

[13] Bo Zhao, Maya Okawa, Eric J. Bigelow, Rose Yu, Tomer Ullman, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hierarchical emotion organization in large language models, 2025.

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

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

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

[17] Luis Rei and Dunja Mladenić. Detecting fine-grained emotions in literature. **Applied Sciences**, Vol. 13, No. 13, 2023.

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

[19] Enric Junqué de Fortuny. The super emotion dataset. **arXiv preprint arXiv:2505.15348**, 2025.

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

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

[22] 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. Qwen3 embedding: Advancing text embedding and reranking through foundation models. **arXiv preprint arXiv:2506.05176**, 2025.

[23] Inderjit S Dhillon and Dharmendra S Modha. Concept decompositions for large sparse text data using clustering. **Machine learning**, Vol. 42, No. 1, pp. 143–175, 2001.

[24] Andrew Rosenberg and Julia Hirschberg. 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)**, pp. 410–420, 2007.

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

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

[27] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics.

[28] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 3821–3830. PMLR, 18–24 Jul 2021.

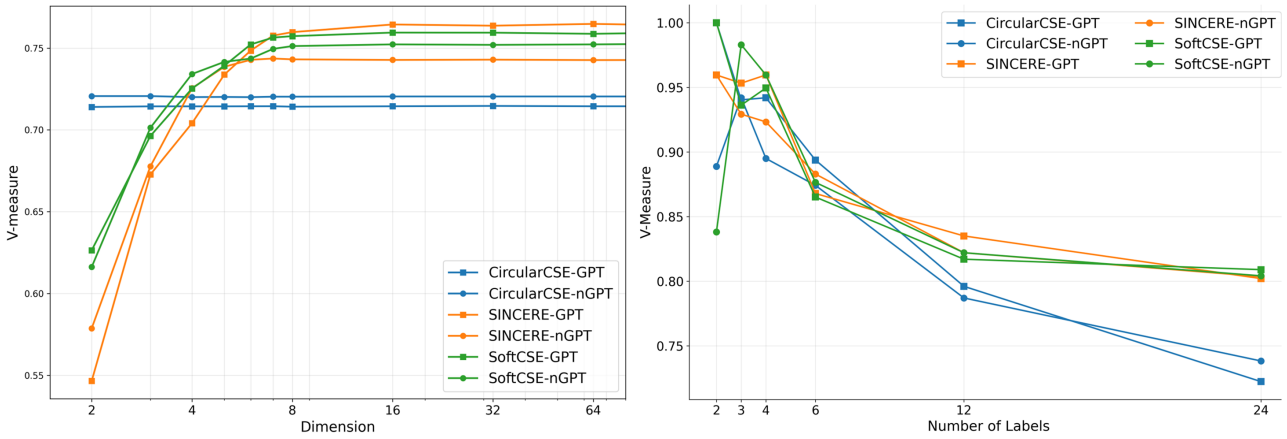
[29] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. **Proceedings of the National Academy of Sciences**, Vol. 117, No. 40, pp. 24652–24663, 2020.

## A　ハイパーパラメータ

表 2. 訓練時に使用したハイパーパラメータ．記載のないものはデフォルト設定を使用

| Pretrained Model | intfloat/<br>multilingual-e5-large | Qwen/<br>Qwen3-Embedding-4B | meta-llama/<br>Llama-3.2-3B |
|---|---|---|---|
| Backbone の凍結 | unfreeze | freeze | |
| Torch dtype<br>訓練エポック数<br>学習率<br>学習率のスケジューラー<br>訓練バッチサイズ<br>シード値 | bfloat16<br>15<br>5e-5<br>constant<br>128<br>42 | | |
| 隠れ次元数 $d$ | 1024 | 2560 | 3072 |
| attention heads の数 | 16 | 32 | 24 |
| プーリング処理 | cls | last | |
| (SINCERE, SoftCSE)-温度 $\tau$ | 0.05 | | |
| CircularCSE-マージン $m$ | 0 | | |

## B　各損失関数の性質




(a) 次元削減に対するロバスト性　　(b) ラベル数の増減に対するロバスト性

図 4: mE5 の異なる条件下での各手法のクラスタリング精度

図 4 に mE5 の異なる条件下 (a. PCA による次元削減 b. 感情ラベル数を増減して訓練) でのクラスタリング精度の変化を示す．CircularCSE は SINCERE や SoftCSE と比べて低次元，少数ラベル設定では同等以上の精度を示すが，高次元，多数ラベル設定では大きく下回るようになる．これは損失関数の最適解の形状とその時の異なるラベル間の最大マージンにより説明できる．SINCERE は損失関数の下界が E-1 次元の正単体で達成することが知られており [28, 29]，異なるラベル間のコサイン類似度が $\frac{1}{E}$ となる (実際は次元の呪いにより局所解の 0 に収束することが多い)．このとき，ラベル間の角度マージンは 90°(直交) 以上となるが，CircularCSE は常に 2 次元の円環が最適解となるため最大でも $\frac{\pi}{E}$ となる．したがって，埋め込み表現の次元数，ならびに感情ラベル数が増加するほど SINCERE と CircularCSE の間にはラベル間の最大マージンに差が生まれ，識別能力に差が生じるようになる．これらの結果は，モデルの表現に対して可視化したときに意味や解釈性を持たせようとすることが，暗黙的に低次元の多様体構造を仮定することを示唆している．