1X-06

線形表現仮説に基づく大規模言語モデルの感情概念空間のファインチューニング耐性

立谷 拓海 †　　伏見 卓恭 ‡
東京工科大学 †　　東京工科大学 ‡

1 はじめに

大規模言語モデル (Large Language Models: LLM) は，感情予測や感情認識などのタスクにおいて人間と同等あるいは一部タスクでは人間を上回るという結果が報告されている [1]. Park らによる線形表現仮説 [2] では，LLM は概念を表現空間上の線形な方向 (概念ベクトル) として保持しており，因果的に独立した概念同士は直交すると述べられている．LLM を特定タスクに特化させる方法として，ファインチューニング（Fine Tuning: FT）があるが，LLM が事前学習により獲得した感情概念が FT 後にどのように変化するのかについては十分に検証されていない．感情分類予測タスクで FT させた LLM は，精神衛生支援や SNS の動向分析などの領域における活用が期待されるため，感情概念空間の FT 耐性の検証が必要である．そこで，感情分類予測タスクで FT した LLM を用いて，線形表現仮説に基づいた感情概念空間の FT 耐性を検証する．

2 線形表現とステアリング

2.1 線形表現仮説とステアリング (Steering)

本研究では，概念を特定の方向ベクトルとして抽出する．ある層のブロック出力を $\mathbf{h}_{base}$，特定の概念を指すステアリングベクトルを $\mathbf{h}_{steering}$ とすると，ステアリング後の $\mathbf{h}_{new}$ は式 (1) で定義される．なお，α はステアリング強度を調整するためのスカラー値である．

$$\mathbf{h}_{new} = \mathbf{h}_{base} + \alpha \mathbf{h}_{steering} \tag{1}$$

この演算によって，モデルの最終的な出力 (ロジット) が意図した方向へ変化するかを観測することで，概念空間が破壊されるか否かを検証することができる．

2.2 感情概念ベクトルの抽出

今回の実験では，感情ラベル e を持つ文章群の最終ブロック出力の平均 $Mean(\mathbf{h}_e)$ から，感情を含まない「中立」文章群の平均 $Mean(\mathbf{h}_{neutral})$ との差をとったものを感情の概念ベクトルとし，式 (2) として与えられる．

$$\mathbf{h}_{steering} = Mean(\mathbf{h}_e) - Mean(\mathbf{h}_{neutral}) \tag{2}$$

出力ロジットへの影響が直接的である最終ブロックの出力を対象とする．

2.3 実験設定

■モデルとデータ: sarashina2.2-1b-instruct-v0.1 を用い，全 24 層の最終ブロックを対象とする．データセットは Troiano らによる crowd-enVENT[3] の感情を喚起したイベントについて述べた文章とその感情ラベルを日本語に機械翻訳したもので，3 エポック（990 ステップ）の感情分類 FT を実施した．

■プロンプトとラベル: FT，フィルタリング，解析の全工程で感情リストを提示して選択させる同一の指示プロンプトを用いた．フィルタリングにより「恥」が消失したため，ステアリングでは 12 感情ラベルを対象とした．また，感

Robustness against Fine-Tuning in the Emotional Concept Space of Large Language Models Based on the Linear Representation Hypothesis
† Takumi Tachiya, Tokyo University of Technology
‡ Takayasu Fushimi, Tokyo University of Technology

"次の感情リストを考えてください：怒り、退屈、嫌悪、恐怖、罪悪感、喜び、誇り、安堵、悲しみ、恥、驚き、信頼、中立 次の文脈から推測される感情は何ですか？ 文脈：'サンプル' 答え："

図 1　使用したプロンプト

情ラベルの先頭のトークンのロジットを観測対象とした．このとき，使用した指示プロンプトは図 1 のようになっている．' サンプル' にはデータセットのサンプルが入る．

3　実験結果と考察

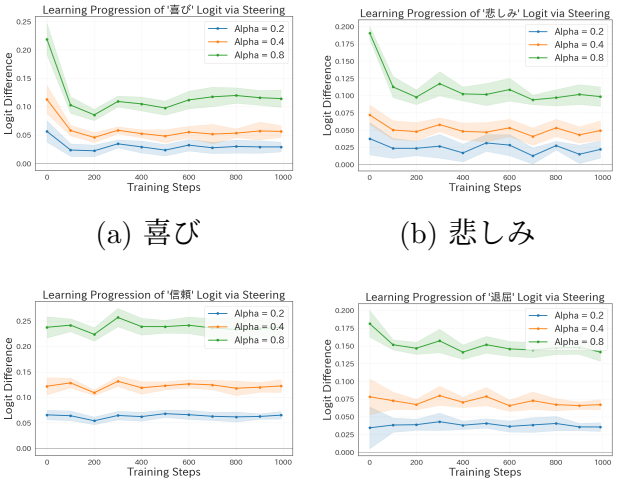
事前実験として実施した「同一概念内の類似性」や「概念間の直交性」,「線形プローブ機能」の検証では，FT 前後で幾何学的な構造に有意な変化は見られなかった．しかし，ステアリング実験では特筆すべき結果が確認された．

3.1　学習ステップごとのステアリング感度

図 2 に，各ステップのモデルにおけるステアリング介入によるロジット変化量を示す．なお，このグラフは事前にサンプリングした 10 件の「中立」ラベルの文章を用いて感情概念ベクトルを求めて抽出したロジットの平均を取った結果である．「喜び」「悲しみ」や図示はしていないが多くの感情で，FT が進むにつれてステアリングによるロジット上昇幅が減少する傾向が見られた．これは，特定の感情分類タスクに最適化される過程で，事前学習由来の汎用的な感情概念軸と出力層（Unembedding）との整合性が失われた，あるいは特定のタスクに合わせて再構築されたことを示している．

3.2　特定感情における頑強性

一方で，「信頼」「退屈」については，FT 後もステアリング感度が維持，あるいは一部のステップ（300 ステップ付近）では FT 前を上回る上昇を見せた．これは，この 2 つの感情概念が本タスクの学習過程において破壊されにくい耐性を持っていることを示している．この要因として，データセット内での概念の明確さや，FT によりモデルが指示文の意図をこの二つの感情に強く結びつけた可能性が考えられる．また，検証誤差が最小となった地点（300 ステップ付近）でロジットの変化量が増加する現象が確認された．これは，学習初期において感情概念の線形性が一時的に洗練されるものの，それ以降のステップでは過学習に伴い，ステアリングの有効性が減少していく傾向を示している．



(a) 喜び　(b) 悲しみ

(c) 信頼　(d) 退屈

図 2　感情ごとのプロット結果

4　おわりに

本研究では，FT による感情概念空間の変容を，ステアリング感度という動的指標を用いて検証した．実験の結果，幾何構造は維持される一方で，出力への影響力 (感度) は感情の種類によって異なる耐性を示すことが明らかになった．今後の課題として，感情情報の局在性が指摘されている中期層での解析 [4] や，異なるモデルサイズでの比較検証が挙げられる．

参考文献

[1] Z. Elyoseph et al. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, Vol. 14, , 2023.

[2] K. Park et al. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[3] E. Troiano et al. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics, Vol.49, No.1, pp.1 – 71*, 2023.

[4] Ala N. Tak et al. Mechanistic interpretability of emotion inference in large language models. *arXiv preprint arXiv:2502.05489v2*, 2025.