G-10

ペルソナを活用した大規模言語モデルによる感情テキスト生成
Persona-Enhanced Emotion Text Generation with Large Language Models

井下 敬翔†‡
Keito Inoshita

1. はじめに

現代社会は,あらゆる活動がデータとして記録・蓄積されるデータ社会へと移行しつつある.人間の行動,発言,位置情報,購買履歴など,様々な記録が日々生成され,それらを活用した人工知能(AI)の応用は加速度的に進展している.特に自然言語処理(NLP)の分野では,大量のテキストデータを用いて,翻訳,要約,質問応答といった多様なタスクにおいて顕著な成果を挙げてきた [1].近年では,大規模言語モデル(LLM)が台頭し,これらのタスクにおいてさらなる精度向上を実現している [2].

しかし,人間の内面に関わる感情認識のようなタスクにおいては,依然として性能の向上が困難な状況にある [3].その大きな要因として,翻訳や要約などのタスクに比べて感情に関する学習データが著しく不足している点が挙げられる.これは,感情に関するデータ収集が倫理的・法的制約を強く受けることに起因する.たとえば,職場や医療現場での感情データ取得は,被収集者の心理状態に深刻な影響を与える可能性があり,実務上も倫理審査のハードルが高い.また,児童や高齢者といった層からのデータ取得は,現行の倫理的枠組みでは難しく,感情認識技術を多様な人々に適用するためのデータは決定的に欠如している.また,感情という対象はテキスト上で明示的に表現されるとは限らず,その解釈には高度な文脈理解が求められる.さらに,感情表現は本質的に主観的で多様性に富み,社会的・文化的背景や個人の性格によっても大きく左右されるため,汎用的かつ高品質なデータセットの構築が極めて困難である.

このような課題に対して近年注目されているのが,LLM を活用したデータ生成および拡張である.LLM は膨大なテキストコーパスを用いて事前学習されており,文脈に応じた自然な文章を生成できる能力を備えている [4].また,特定の状況や話者属性を条件として指定することで,それに応じた多様かつ一貫性のある言語表現を再現することが可能である.このような LLM の能力を感情データの拡張に応用すれば,現実世界では倫理的・法的・実務的制約により取得が困難な感情テキストを,人工的にかつ柔軟に生成できる可能性がある.さらに,このように獲得された生成データは,低リソース領域における学習データの補完に留まらず,再現性のある感情表現に基づくシステム評価や,新たなデータ拡張手法の設計などにも応用できる.したがって,LLM を活用した感情データ生成は,今後の感情理解技術の発展において重要な鍵を握るアプローチの一つであるといえる.

そこで本研究では,感情データの属人性や倫理的制約といった収集上の困難を回避しつつ,多様かつ文脈的な感情テキストを LLM によって生成する新たなフレームワークを提案する.感情カテゴリごとに架空のペルソナを合成し,年齢・性別・職業・文化的背景といった属性を与えることで,現実に近い多様な視点からの感情表現を実現する.こうしたペルソナを LLM のプロンプトに組み込み,感情に対応した自然なテキストを生成する.生成されたテキストはベクトルに変換された後,類似度計算を通して事前に設定された閾値を下回る場合のみをデータベースに保存される.これにより,現実世界での収集が困難な「多様な個人に根差した感情テキスト」を代替的に構築可能となる.本研究ではこのフレームワークの提案に伴い,以下の 3 つの研究課題(RQ)に取り組む.

RQ1. LLM が感情カテゴリを指示された際にどの程度正確かつ多様性に富んだ感情を表現できるか.
RQ2. 生成時のパラメータ設定やデータ保存時の閾値の違いは,感情テキストの正確性・多様性にどのような影響を与えるか.
RQ3. LLM が生成した感情テキストは,人間が実際に記述した感情表現とどの程度類似しているか.

これらの RQ に対して,可視化や意味的類似度による感情テキストの多様性評価に加え,LLM-as-a-Judge [5]により人間が記述した感情テキストとの類似性評価を行い,提案フレームワークの有効性を示す.本研究は,感情認識の性能向上に寄与するだけでなく,感情データの再現性や拡張性を高め,倫理的に安全なデータ生成基盤の構築に資することが期待される.

本研究の主な貢献は以下の 3 点である：

i) ペルソナ属性を活用し,文脈の多様性を組み込んだ感情テキストを LLM によって自動生成する手法を構築し,属人性と倫理的制約を回避しつつ高品質なデータを得る仕組みを実現した.
ii) 可視化,分類モデルによる予測精度,意味的類似度や分散指標に基づく多様性評価などの多面的な評価によって,生成された感情テキストの妥当性を定量的に検証した.
iii) 人手による感情表現の Few-shot による LLM-as-a-Judge の評価により,LLM による感情データ生成が実際の人間の感情記述に近い性質を持つことを示し,データ拡張や感情理解応用への実用性を明らかにした.

本論文は以下のように構成される.第 2 節では関連研究について概観し,第 3 節では提案手法の詳細を述べる.第 4 節では実験設計と評価結果を報告し,第 5 節で結果に基づいた議論を行う.最後に第 6 節で本研究の結論と今後の展望を述べる.

2. 関連研究

2.1 ペルソナや文体制御による生成テキストの多様性向上

ペルソナを LLM に与えることで,生成される感情テキストの多様性と自然さを高めることができる.このアプローチは,近年注目されているペルソナ制御による生成挙動の多様化に関する一連の研究と強く関連している.まず,Gupta ら [6] は,LLM に社会的ペルソナを与えた場合,推論タスクの精度が低下しバイアスが顕在化することを示しており,ペルソナが生成挙動に与える影響の大きさとリスクの両面を指摘している.一方で,Hu と Collier [7] は,年齢や性格といった属

†関西大学, Kansai University
‡ 滋賀大学, Shiga University

性情報を LLM に与えることで,主観的応答タスクの再現精度が有意に向上することを示しており,属性条件付けが出力の制御手段として有効であることを裏付けている.さらに,Araujo と Roth [8] は,162 種にわたる多様なペルソナを設定した大規模実験を通じて,ペルソナが出力文の内容・スタイルに一貫した変化をもたらすことを示し,複数属性から構成されるペルソナがテキスト生成の多様性源となりうる可能性を示唆している.Fröhling ら [9] もまた,ペルソナを明示的に提示することで,データアノテーションの出力が多様かつ再現性のある形で変化することを報告しており,本研究のペルソナによる生成制御の意義と整合する.また,文化的背景が出力文に与える影響については,Kamruzzaman と Kim [10] が国籍別ペルソナを用いた調査で,LLM の生成文に地域バイアスが反映される傾向を示しており,本研究の言語スタイルや環境といった属性の設計方針と関連が深い.

より構造的な人物設定に基づく研究として,Li ら [11] は,1000 人以上の詳細な人物設定に基づいて,連続的な行動をシミュレーションするベンチマークを提案し,ペルソナが文脈依存の挙動再現に与える限界を実証的に評価し,Giorgi ら [12] は,信念や生活背景などの人間的要素を明示・暗示的に埋め込んだペルソナを用いて,信念生成や感情検出の出力傾向を評価しており,主観的コンテンツ生成への応用性と課題の両面を明らかにしている.最後に,Inoshita [13] は,LLM に多様な年齢の文体を模倣させる実験を行い,人間の年齢特有の文体特徴が部分的に再現可能であることを定量的に確認した.これは年齢・教育歴といった属性の導入を強く支持する知見である.これらの結果はペルソナ設計の際に人間の性格やより深い心情を考慮することで出力により多様性が生まれることを示している.

2.2 LLM を用いたデータ拡張技術

LLM を活用してテキストを自動生成し,コーパスの質的・量的拡張が可能である.このような LLM ベースのデータ拡張は,近年さまざまな NLP タスクに応用されており,汎用性と精度向上の両面で注目されている.まず,Dauber ら [14] は,BERT による分類タスクにおいて,LLM によるデータ生成を活用し,学習に有益であることを示している.低リソース環境への応用として,Evuru ら [15] は,制約付きプロンプトを用いた CoDa フレームワークを提案し,従来法を上回る性能を達成している.多様性の観点からは,Wang ら [16] の DoAug フレームワークが,パラフレーズの多様性最大化を通じて平均 10%以上の性能向上を示しており,生成データの分布的な広がりの重要性を裏付けている.これは本研究で行う類似度による選別とも対応し,高品質かつ多様性あるデータの重要性を支持する.また,Jayawardena と Yapa [17] は,ParaFusion という LLM 生成による大規模パラフレーズデータセットを開発し,語彙・構文多様性の 25%以上の向上を報告しており,LLM は文体や構文の幅を広げられることを示した.

次に,精度とロバスト性を目的としたデータ補完の事例として,Gopali ら [18] は GPT-3.5 で MBTI データのマイノリティクラスを補完し,クラスバランスの改善による F1 スコア向上を達成している.Zhou ら [19] は,関係抽出タスクにおいて,元データからのパラフレーズと新規文生成の 2 種を組み合わせてモデル性能を向上させており,多角的な生成データ設計が有効であることを示している.さらに,Lippmann ら [20] は,モデルの誤分類領域を LLM で補完するという「教師付きデータ拡張」手法を提案しており,特定条件下での出力制御や補完的生成の重要性を指摘している.最後に,Ding ら [21] は LLM を用いたデータ拡張の技術的俯瞰を行い,制御性・多様性・効率性といった評価軸が今後の研究で鍵となることを示しており,本研究の評価指標設計とも強く整合している.

2.3 感情コーパスの生成・拡張

感情認識モデルの開発には,大規模かつバランスの取れた感情コーパスが不可欠であるが,感情という主観的で多次元的な概念の性質上,その構築は困難である.そこで近年は,感情コーパスの構築に関する研究が増えてきている.まず,Peng と Zhang ら [22] は,Chain-of-Thought [23] を用いたプロンプト制御によって,感情極性や視点を自由に操作しながらデータ拡張を行う手法を提案している.また,Jayawardena と Yapa らは,GPT を用いて語彙・構文的に多様なパラフレーズ感情文データセットを構築し,多様性と自然性を同時に向上させている.Inoshita [24] は,ルールベースで GPT に感情に関係する表現を拡張したコーパスを拡張し,多様性を実現している.これらの研究は LLM を活用して感情テキストを生成する点で,本研究と目的が共通している.

また,Gopali らは,NLP タスクにおけるクラス不均衡に注目し,その差を補完するために少数派クラスを GPT で合成し,分類性能を向上させた.感情分析タスクでもこの有効性を示している.さらに,Lee と Lee ら [25] は,SNS 由来のカジュアル文体を ChatGPT で正規化し,感情認識精度を高める効果を示した.これは,LLM が感情を一定程度理解し,適切に文体を変更できることを示し,感情テキスト生成に対する応用可能性を示唆する.

これらの研究は,LLM に対するペルソナを活用したプロンプト制御により,多様性担保,文体操作,クラス均衡が実現できることを示し,本研究の提案するフレームワークが感情データの生成・拡張が可能であることを強く補強する.また,より多様性のあるペルソナ設計や詳細なパラメータ調整を行うことで,さらに品質の高い感情データの生成ができる可能性を示している.

3. ペルソナを活用した多様な感情テキストの生成

3.1 フレームワークの全体像

本研究が提案するペルソナを活用した LLM による感情テキスト生成フレームワークの全体像を図 1 に示す.本フレームワークは,多様なペルソナ合成,ペルソナを活用した LLM による感情テキスト生成,意味的類似度に基づく冗長排除及びデータベースへの保存という 4 段階の処理で構成される.具体的にはそれぞれ,次の処理を行う.まず,ペルソナ合成では,年齢,性別,職業,性格特性,教育歴,言語スタイル,コミュニケーション環境の 7 カテゴリからそれぞれランダムに 1 属性ずつを抽出し,多次元的なペルソナを合成する.次に,このペルソナ情報を,任意に設定された感情カテゴリと共にプロンプトを構築し,システムプロンプトと併せて LLM に入力する.LLM により生成された感情テキストは Sentence-BERT により埋め込み, ベクトルに変換され,既にデータベースに格納されているテキストとの最大コサイン類似度を計算する.ここで最大コサイン類似度が,任意に設定された閾値θ未満であれば新規性が認められると判断しデータベースに感情ラベルと共に格納され,θ以上であれば棄却する.最終的に全てのデータは CSV 形式で格納され, 感情カテゴリごとに均衡なコーパスが構築される.

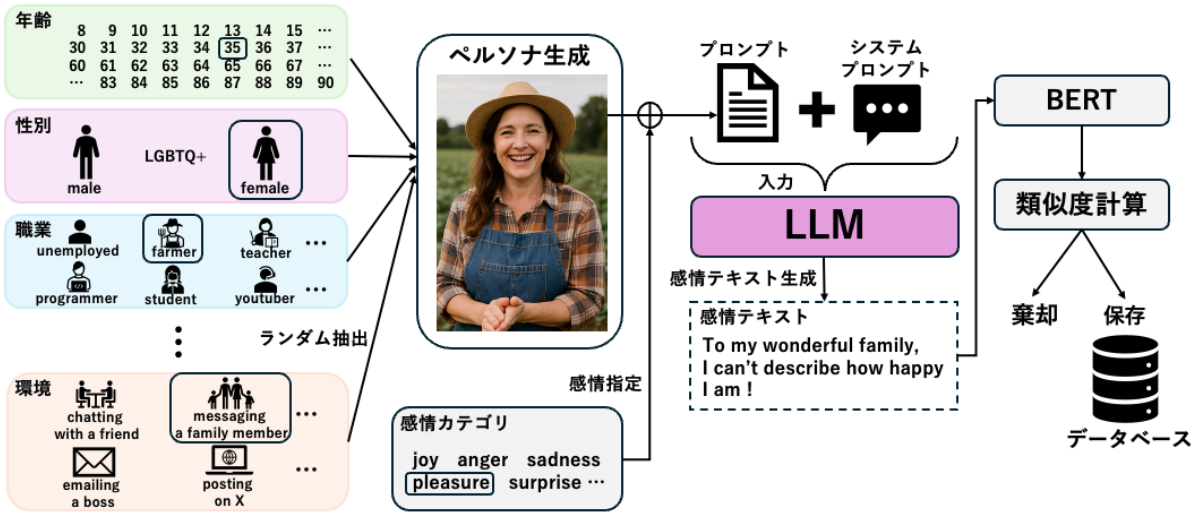

図 1 感情テキスト生成フレームワーク.

このように,本フレームワークは,ペルソナによる文脈的多様性と類似度フィルタによる意味的多様性の双方を両立させることで，制約なしに高品質な感情データセットを生成することを目的とする.

3.2 多様性に基づくペルソナ合成

感情テキスト生成における多様性の確保は,単に異なる感情カテゴリを生成するだけでは不十分である.特に,LLM を活用する際には,指示された感情を一様な文体や立場から表現する傾向が生じることがある.これに対処するために,本研究では,あらかじめ合成した多次元的なペルソナ情報を LLM への入力プロンプトに組み込むことで,感情表現にバリエーションを導入する手法を採用する.本研究におけるペルソナは,表 I に示す属性カテゴリで構成される.ペルソナ合成の際には,これらの属性カテゴリの中からランダムに 1 つずつの属性が選択され,それらを組み合わせることで,感情テキスト生成の際に多様な人間像をシミュレートする.

表 I ペルソナ生成に使用する属性カテゴリ

属性カテゴリ	属性数	例
年齢	82	8~90
性別	3	male, female, LGBTQ+
職業	20	teacher, programmer, farmer など
性格特性	18	high extraversion, cautious, impulsive など
教育レベル	6	junior high school〜PhD
言語スタイル	6	casual, poetic, internet slang など
環境	12	chatting with a friend, emailing a boss など

本研究で定義した各カテゴリは,感情表現の多様性に直結する側面を意識して設計されており,以下のような効果をもたらす.まず,年齢や教育レベルは,言語能力や語彙の選択に影響を与える重要な要素である.たとえば,小学生の語り口と高齢の博士号取得者の語り口では,同じ感情でも文体や表現が大きく異なる.性別やLGBTQ+属性もまた,文化的背景や社会的役割の違いにより,感情の表出傾向や選択される語彙に差異をもたらす.職業は,その人物が置かれている社会的文脈や日常的な言語使用のスタイルを形成する.営業職や介護士などの対人職では感情表現が豊かである一方で,研究者やプログラマーのような論理的言語使用が多い職業では,抑制的かつ構造的な表現が観察される.性格特性については,心理学的に確立された Big Five の次元をベースに設計しており,たとえば高い神経症傾向をもつ人物は否定的感情を強く表出しやすいという傾向がある.環境と言語スタイルは,生成される文がどういったコミュニケーション場面に属するかを規定する.家族との LINE や匿名掲示板のようなカジュアルな場面では私的で感情的な表現が多く,上司へのメールやカスタマーサポートへの連絡といった形式的な文脈では,抑制的かつ丁寧な表現が生まれやすい.これにより,同じ感情カテゴリにおいても異なる文脈的表現が実現される.

このように合成されたペルソナカテゴリの構成では,全組み合わせで約 3,800 万通りのバリエーションを持ち,極めて高い文脈的多様性を保証する.各ペルソナ情報は,感情テキスト生成時に LLM へとプロンプトを通して提示され,LLM はその人物像・背景・状況に即した自然な感情テキストを生成するよう指示される.これにより,従来の定型的なテンプレート文とは異なり,多様な話者の視点を内包した感情表現が生成され,感情の多様性・文体の多様性・社会的背景の多様性を備えた高品質なデータセットの構築が可能となる.

3.3 ペルソナを活用した LLM による感情テキスト生成

ペルソナ情報および任意に設計された感情カテゴリに基づいて LLM を活用し,自然かつ文脈的妥当性の高い感情テキストを生成する.ここで,プロンプト設計や出力制御のための温度パラメータの設定は,生成結果を大きく左右するため,慎重な調整が必要である.本研究では,感情テキストの生成のために,3.2 節にて構築された多様なペルソナ情報を活用したプロンプト設計を行う.ペルソナの属性を自然文に埋め込むことで,モデル出力に人間らしい多様性と一貫性を誘導する.具体的には図 2 に示すプロンプトを活用する.このプロンプトでは,人物設定,使用環境,言語スタイルといった多面的属性を組み込み,さらに目標感情を強く喚起された状況を想定する.これにより,単なる感情表現文ではなく,ある文脈下における自然な発話としての感情文を生成することが

可能となる.モデルへの入力は,上述のプロンプトに加えて,図 3 に示すシステムプロンプトを設定することで出力の一貫性とペルソナの再現性を担保する.

```
You are a {年齢}-year-old {性別} working as a {職業}.
Your personality is described as {性格特性 }, and your highest level of education is {教育レベル }.
You are currently in a situation where you are {環境}, and you are feeling the emotion of \"{感情カテゴリ }\" very strongly.
Please use a {言語スタイル} tone.
Rather than describing the emotion, write one natural sentence that someone might actually say or write to express that emotion.
```

図 2 感情テキスト生成のためのペルソナを反映したプロンプト．それぞれの {} にペルソナ情報が入力される

```
You are to act as a single, real human being who matches the given persona and context. Express the specified emotion as naturally and authentically as possible in a single sentence.
```

図 3 感情テキスト生成のためのシステムプロンプト.

また,生成結果の多様性を制御するために,適切な温度パラメータτを設定する.τは確率分布からのサンプリングにおける乱数性を調整するものであり,値が高いほど創造性が増し,低いほど決定的な出力となる.表 II にパラメータの変化による出力への効果を示す.

表 II 温度パラメータと出力の関係性

τ	創造性	出力の特徴
0.7	低〜中	定型的・高一貫性
1.0	中	自然さと多様性の均衡
1.3	中〜高	語彙多彩・やや逸脱
1.6	高	比喩・自由度高いが散漫

以上のプロンプト及びシステムプロンプト設計に加え,τを適切に設定することで,感情テキストを生成する.

3.4 意味的類似度に基づく冗長排除

LLM を用いて感情テキストを生成する際,同一感情カテゴリ内で内容が類似した文が多数生成されてしまう可能性がある.これは,モデルが確率モデルであるが故に,指示された感情を反映する際に最も代表的な表現を繰り返し出力する傾向にあるためである.したがって,単にペルソナと感情カテゴリを指定するだけでは,意味的な多様性が十分に担保されたデータセットを構築することは困難である.

本研究では,この問題に対処するため,文の意味的な類似度を計算し,あらかじめ定めた閾値以上に類似した文を除外することで,生成データの多様性を担保する枠組みを導入する.具体的には,事前学習済みの Sentence-BERT 系列モデルである all-MiniLM-L6-v2 [26] を使用し,各テキストを 384 次元のベクトル空間に埋め込んだ上で,コサイン類似度に基づいた類似性評価を行っている.まず,生成された新規テキストについて,その埋め込みベクトルv_{new}を取得する.この際に,類似度計算の安定性と解釈生を高めるために,v_{new}に対して,以下に示す L2 正規化を適用し,$\hat{v}_{new}$を得る.

$$\hat{v}_{new} = \frac{v_{new}}{\sqrt{\sum_{i=1}^{d} v_{new,i}^2}}$$

ここで,dは次元数を意味する.

次に,$\hat{v}_{new}$とすでにデータベースに格納された文ベクトル集合$\{v_1, v_2, \ldots, v_n\}$との最大コサイン類似度を計算する.最大スコアが閾値θ未満である場合に限り,その文はデータベースに追加される.これにより,意味的に異なる感情表現のみが保持され,データセット全体としての語用的・文脈的多様性が向上する.なお,閾値θは生成データの多様性と収集効率の間のトレードオフを制御する重要なパラメータである.具体的には,表 III に示すように,閾値が高いほど,文の意味的な冗長性は抑制されるが,生成成功率は低下し,閾値が低いほど,多くの文が保持される反面,意味的な重複が増加する.

表 III 類似度閾値の変化による取得データの特徴

類似度閾値	特徴	データの特徴
0.75	緩い制限	冗長性やや高い
0.80	やや緩い	冗長性を抑え成功率向上
0.85	中庸	多様性と効率性のバランス
0.90	厳格な制限	意味的に広範な文のみ

このように,生成文に対して意味的類似度のフィルタリングを施すことで,LLM の出力傾向に依存しない,より広範で豊かな感情表現の収集が実現される.ペルソナによる多様性の導入とあわせて,本手法は高品質な感情テキストデータの構築における重要な要素として機能する.

4. 実験と分析

4.1 データセットと実験設計

提案フレームワークの性能を検証するために,基本となるコーパスを構築する.感情カテゴリには[joy, anger, sadness, pleasure, surprise, fear, neutral]の 7 種類を採用し,LLM には OpenAI 社の GPT-4o-mini [27] を使用し,十分な生成能力のもと,質の高い感情テキスト生成を実行した.コーパス作成の際には,τを LLM の初期値である 0.7, θを 0.80 に固定して生成を行った.各感情あたり 500 文を取得し,合計 3500 文のデータセット(Base-7k)を得た.これらの条件のもと,本研究では 3 つの RQ に以下の設計で取り組む.

RQ1. Base-7k を用いて 3 つの観点から生成品質を評価する.i) 各テキストの埋め込みベクトルに対して t-SNE で 2 次元へ射影し,感情クラスタが視覚的に分離しているかを確認する.ii) Mean Cosine Distance,Cluster Entropy,Centroid Distance を算出し,感情内の語用的広がりを定量化する.iii) 埋め込みベクトルを入力とし, LightGBM による 7 クラス分類器を学習し,生成文が正しく感情カテゴリに分類される割合を確認する.

RQ2. τ と θ が正確性,多様性,効率のバランスに与える影響を評価する.τを[0.7, 1.0, 1.3, 1.6],θを[0.75, 0.80, 0.85, 0.90]とし,16 通りの組合せを設定する.各条件で前述と同じ手順を適用し,感情ごとに 500 文を生成し,各コーパスについて,RQ1 と同様の性能比較を行う.

RQ3. RQ2 で検証したτとθ組み合わせのうち総合評価が最も高かった条件を使用し,LLM が生成した感情テキストはと人間が実際に記述した感情表現とどの程度類似しているか評価する.評価には生成に使用したモデルよりも大きい GPT-4o による LLM-as-a-Judge を用い,感情一致度と文体自然度を 5 段階で付与させる.

ここで,Mean Cosine Distance,Cluster Entropy,Centroid Distance は,生成文に付与された感情カテゴリごとの語用的まとまり

を評価するための指標である.まず,Mean Cosine Distance は,各感情カテゴリ内の全てのテキストペアにおけるコサイン距離の平均を算出したものである.この指標は,同一感情カテゴリに属するテキスト同士の意味的な多様性を示す.値は 0〜2 を取り,値が大きいほど同カテゴリ内で多様な表現が含まれていることを示し,多様性の観点では望ましい傾向とされる.次に,Cluster Entropy は,各感情カテゴリ内のテキストをさらに小さなクラスタに分割した際のエントロピーを算出し,その感情内で表現がどれだけ均等に広がっているかを評価する指標である.理論的には,エントロピーの下限は 0 であり,上限はクラスタ数に依存するが,ここではおおよそ 2 を超えない範囲に収まる.値が大きいほど語用的なバリエーションが豊富であり,より自然な感情表現が生成されていることを示唆する.最後に,Centroid Distance は,各感情カテゴリの埋め込みベクトルの重心間の距離を算出し,感情カテゴリ間の分離性を評価する.値の範囲はデータの分布に依存するが,値が大きいほどカテゴリ間の重なりが少なく,感情の判別が容易であることを意味する.特に隣接する感情カテゴリ間でこの距離が大きい場合,生成文が明確に分類可能な語用的特徴を持っていると判断できる.これら 3 つの指標を総合的に評価することで,感情ごとの生成文の多様性と識別性の両面を定量的に把握することが可能である.

以上の設計により，RQ1〜RQ3 を通じて，提案フレームワークが生成する感情テキストの正確性，多様性，および人間らしさを多面的に評価する.

4.2 Base-7k を用いた 3 観点による生成品質の評価

τを 0.7，θを 0.80 に固定して生成を行なった結果得られた 3500 件の感情テキストで構成される Base-7k の品質を,感情の分離性・内部多様性・分類可能性という 3 つの観点から評価した.

まず,各生成文に対して all-MiniLM-L6-v2 を用いて得られた埋め込みベクトルを 2 次元に射影し,t-SNE を適用した結果を図 4 に示す.結果として,anger,sadness,neutral,fear のクラスタは比較的明確に分離しており,テキストに含まれる感情が埋め込み空間上でも識別可能である傾向が見られた.一方で,joy と pleasure のクラスタは重なりが大きくその境界は曖昧で,surprise と混同する場合も見られる.このことは,τを 0.7,θを 0.80 として作成された Base-7k において,怒りや悲しみ,中立,恐怖といった感情は比較的特徴的な語彙や文構造を持つため,その感情を明確に示すテキストを生成することができているが,喜びや快楽といった感情は語用的・意味的に似たようなテキストを生成しており,感情カテゴリ間の重なりが生じやすいことが示唆される.

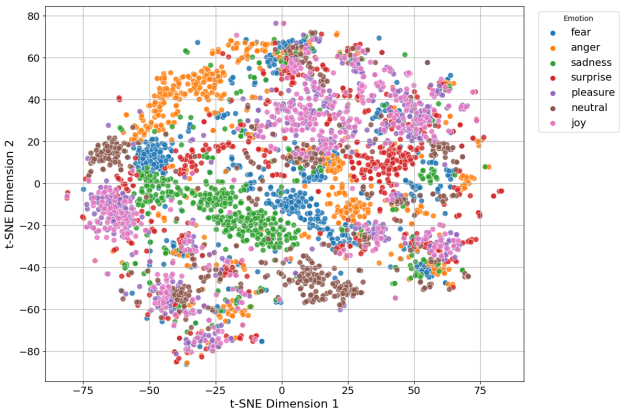

図 4 生成文の埋め込みベクトルに対する t-SNE の結果.

次に,感情ごとの語用的多様性を,Mean Cosine Distance,Cluster Entropy,Centroid Distance の 3 指標で定量化した.その結果を表 IV に示す.Mean Cosine Distance においては,anger や surprise,fear が高い値を示し,同一感情内においても多様な表現が含まれていることを示している.反対に,sadness や joy は比較的低く,一定のパターンに収束しやすい生成傾向があることが示唆される.これは,高覚醒感情においては多様かつ識別的な表現を生成できているが,低覚醒感情に関しては多様性が限定的であることを示している.Cluster Entropy は全体的に 1.53〜1.61 の間で分布しており,すべての感情カテゴリにおいて中程度の分散性を有していた.特筆すべきは pleasure が最も高く,喜びよりも快楽に関する表現の方が語用的に多様である可能性がある点である.Centroid Distance は全体として 0.38 と比較的小さく,各感情クラスタの中心が互いに近接している傾向があった.これは,u 埋め込み空間において感情の分離が完全ではなく,特に joy・pleasure といったカテゴリ間での距離が近いことが要因と考えられる.

表 IV Base-7k に対する 3 観点による生成品質評価の結果

評価指標	joy	anger	sadness	pleasure	surprise	fear	neutral
Mean Cosine Distance	0.70	0.79	0.69	0.70	0.78	0.75	0.76
Cluster Entropy	1.58	1.53	1.57	1.61	1.57	1.59	1.57
Centroid Distance				0.38			

最後に,埋め込みベクトルを用い,LightGBM による 7 クラス分類モデルを構築し,感情がどの程度正確に分類されるかを評価した.その結果を表 V に示す.分類性能は全体として高く,特に anger,sadness,surprise,fear,neutral といった感情は高い F1 スコアを示した.一方で,joy および pleasure は他の感情と比較して識別精度が著しく低く,Recall が特に低下している.これは,これらの感情カテゴリに対する埋め込み特徴が他カテゴリと区別しにくく,モデルにとって判別が難しいことを示している.実際に図 5 に示す混同行列を見ると,joy と pleasure の混同が非常に多いことがわかる.この結果は,前述の t-SNE および定量評価と整合的であり,joy や pleasure は語用的・意味的に類似性が高く,Base-7k で使用したパラメータ設定では独立性をもった生成が難しいことを意味している.

表 V Base-7k に対する LightGBM の性能評価

感情カテゴリ	Precision	Recall	F1
joy	0.51	0.58	0.54
anger	0.81	0.88	0.85
sadness	0.87	0.83	0.85
pleasure	0.51	0.43	0.46
surprise	0.81	0.87	0.84
fear	0.85	0.79	0.82
neutral	0.84	0.83	0.83

このように,τを 0.7, θを 0.80 に固定した生成により作成された Base-7k は,高識別性かつ多様な感情生成できていることを示した一方で,特に快楽系感情において表現の曖昧性や分離性の低さが課題として浮き彫りとなった.

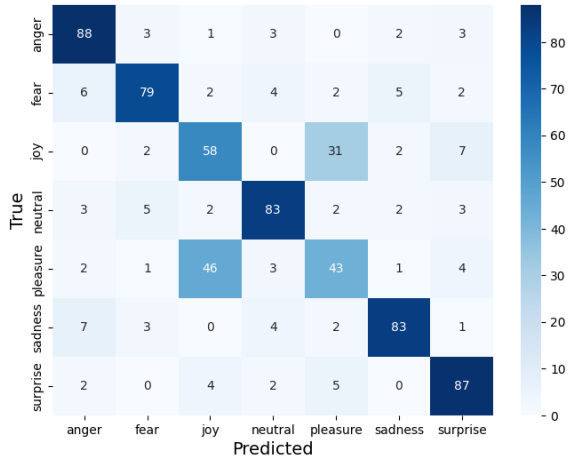

図 5 LightGBM の予測に対する混同行列の可視化.

4.3 温度・類似度閾値の感情生成品質に対する影響

出力の品質に影響を与えるとされるτとθが正確性,多様性,効率のバランスに与える影響を評価する.それぞれτ = [0.7, 1.0, 1.3, 1.6],θ = [0.75, 0.80, 0.85, 0.90],からなる 16 通りの組み合わせを設定した.各条件に対して感情ごとに 500 文を生成し,RQ1 と同様の観点から評価を行った.まず,Mean Cosine Distance の結果を表 VI に示す.全体として,τ = 1.6 に設定した場合に最も高い値を示し,特に(τ = 1.6, θ = 0.75)や(τ = 1.6, θ = 0.80)では全感情カテゴリにおいて距離が相対的に大きく,多様性の観点で優れていた.一方,τが低い場合や θ が高い場合では多様性が減少する傾向が見られた.これは,出力がより保守的になり,表現の幅が狭まったためと考えられる.次に,Cluster Entropy の結果を表 VII に示す.全体的に 1.50〜1.60 の範囲で安定しており,特に(τ = 1.3, θ = 0.75),(τ = 1.0, θ = 0.90)などにおいて多数の感情カテゴリでエントロピーが高く,多様な表現が分散して現れていることが確認された.一方で,(τ = 0.7, θ = 0.85)や(τ = 1.6, θ = 0.85)のように一部でエントロピーが低下する条件も見られ,出力が定型的になる可能性を示唆している.最後に表 VII に示す Centroid Distance の結果を見ると(τ = 0.7, θ = 0.85)が最も高いスコアを示し,クラスタの識別性が高いことが示唆された.一方で,τが高くθが低い条件では,文の多様性が高まる一方でクラスタ間距離は縮まり,感情分類が困難になる可能性も示唆される.

3 つの指標の結果を総合すると,以下の 2 つの条件が,正確性・多様性・識別性のバランスが取れており,特に有望なパラメータ設定であると判断できる.

- (τ = 1.6, θ = 0.75)最大の多様性と高い Cluster Entropy を実現.
- (τ = 1.3, θ = 0.75)中程度の多様性を保ちつつ,最も高い Centroid Distance と比較的高いエントロピーを実現.

これらの 2 条件において,LLM が生成した感情テキストは人間が実際に記述した感情表現とどの程度類似しているか 4.4 節で評価する.

表 VI パラメータ組み合わせごとの Mean Cosine Distance の結果

(τ, θ)	joy	anger	sadness	pleasure	surprise	fear	neutral
(0.7, 0.75)	0.709	0.791	0.703	0.716	0.788	0.760	0.757
(0.7, 0.80)	0.698	0.788	0.687	0.701	0.783	0.750	0.764
(0.7, 0.85)	0.694	0.780	0.685	0.695	0.780	0.741	0.758
(0.7, 0.90)	0.691	0.783	0.670	0.702	0.790	0.741	0.760
(1.0, 0.75)	0.711	0.784	0.698	0.716	0.786	0.761	0.772
(1.0, 0.80)	0.694	0.791	0.697	0.710	0.798	0.751	0.761
(1.0, 0.85)	0.694	0.787	0.687	0.701	0.784	0.743	0.761
(1.0, 0.90)	0.689	0.791	0.684	0.706	0.785	0.750	0.768
(1.3, 0.75)	0.703	0.783	0.706	0.716	0.799	0.760	0.765
(1.3, 0.80)	0.697	0.790	0.691	0.709	0.793	0.752	0.767
(1.3, 0.85)	0.690	0.791	0.690	0.701	0.792	0.759	0.767
(1.3, 0.90)	0.699	0.787	0.694	0.699	0.785	0.744	0.761
(1.6, 0.75)	**0.719**	0.797	0.710	0.713	**0.810**	**0.762**	**0.780**
(1.6, 0.80)	0.710	**0.802**	0.709	**0.718**	0.791	0.758	0.777
(1.6, 0.85)	0.707	0.796	**0.715**	0.718	0.797	0.757	0.773
(1.6, 0.90)	0.701	0.795	0.714	0.711	0.793	0.758	0.779

4.4 生成感情文の感情一致度と人間らしさに関する評価

4.3 節の結果に基づき,最終評価として,LLM が生成した感情テキストが人間によって書かれた表現とどの程度類似しているかを検証した.評価は,LLM-as-a-judge で実行し,生成時に用いたモデルの GPT-4o-mini よりも高性能な LLM である GPT-4o を評価モデルとして使用した.感情の再現性と人間らしさを各テキストに対して以下の 4 つの観点から 5 段階評価(1：非常に低い〜5：非常に高い)を行った.

- 感情一致度：指示された感情が文中に適切に表現されているか
- 文法的自然さ：文構造や語順,句読点などの自然さ
- 語彙と表現の多様性：語彙のバリエーションや感情への適合度
- 文構造と論理の流れ：意味や構文のつながりの自然さ

対象としたデータは,4.3 節で最もバランスが良いと判断された 2 条件(τ = 1.6, θ = 0.75)および (τ = 1.3, θ = 0.75)で生成されたテキストそれぞれ 3500 文である.

図 6 に各項目のスコア分布を箱ひげ図として示す.全体的に平均スコアはいずれも 4.0 以上を維持しており,いずれの観点においても LLM が高品質な感情表現を生成できてい

表 VII パラメータ組み合わせごとの Cluster Entropy 及び Centroid Distance(CD)の結果

(τ, θ)	joy	anger	sadness	pleasure	surprise	fear	neutral	CD
(0.7, 0.75)	1.585	1.573	1.591	1.605	1.460	1.565	1.594	0.366
(0.7, 0.80)	1.577	1.534	1.572	**1.606**	1.569	1.593	1.573	0.376
(0.7, 0.85)	1.481	1.586	1.589	1.529	1.489	1.574	1.587	**0.389**
(0.7, 0.90)	1.540	1.558	1.559	1.547	1.558	1.588	1.582	0.387
(1.0, 0.75)	1.574	1.587	1.564	1.593	1.514	1.588	**1.597**	0.362
(1.0, 0.80)	1.605	1.577	1.581	1.518	1.564	1.581	1.548	0.369
(1.0, 0.85)	1.554	1.585	1.574	1.545	1.530	1.580	1.577	0.381
(1.0, 0.90)	1.602	**1.588**	1.591	1.524	1.581	1.597	1.549	0.372
(1.3, 0.75)	1.582	1.583	1.590	1.525	**1.607**	**1.607**	1.555	0.359
(1.3, 0.80)	1.576	1.552	1.601	1.556	1.597	1.572	1.585	0.369
(1.3, 0.85)	1.572	1.525	1.594	1.549	1.579	1.594	1.555	0.374
(1.3, 0.90)	**1.606**	1.576	**1.603**	1.562	1.500	1.606	1.506	0.375
(1.6, 0.75)	1.593	1.572	1.600	1.545	1.572	1.593	1.573	0.351
(1.6, 0.80)	1.588	1.543	1.581	1.588	1.522	1.581	1.549	0.344
(1.6, 0.85)	1.451	1.534	1.566	1.542	1.562	1.595	1.571	0.351
(1.6, 0.90)	1.586	1.554	1.543	1.517	1.557	1.579	1.573	0.356

ることが確認された.特に文法的自然さは,τ =1.3 において平均 4.96(SD=0.20)と非常に安定したスコアを示しており,LLM が定型文法を確実に再現できていることを裏付けている.また,感情一致度においても,両条件とも平均が 4.2 以上を記録しており,感情カテゴリが明示されたプロンプトに対して,LLM が文脈に即した感情を表現する能力を有していることが示された.一方で,すべての評価項目においてスコア 1 または 2 が散見され,特にτ =1.6 においては全ての指標で複数の外れ値が確認された.これは,τが高い場合に,創造的な表現が文の一貫性を損なうリスクがあることを示唆している.しかし,τ =1.3 のケースと比較して多くの指標で上回っていることから,許容できる範囲と言える.

これらの結果から,LLM-as-a-Judge は,感情テキストの自然さや感情との整合性を多面的に評価する手段として一定の有効性を持つことが確認できた.また,本研究で検証した条件の中では,(τ =1.6, θ =0.75)が最も安定して高い評価を得ており,多様性と一貫性の両面で優れた出力を生成する傾向が確認された.

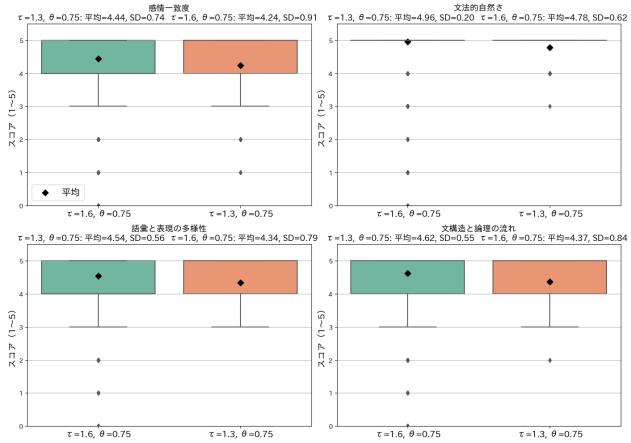

図 6 LLM-as-a-Judge を用いた 4 指標による生成感情テキストの比較評価.

5. 議論

本研究では,ペルソナ情報と意味的類似度フィルタを組み合わせた感情テキスト生成フレームワークを構築し,その生成品質を多角的に評価した.実験の結果,提案フレームワークが感情の識別性,語用的多様性,文体の自然さにおいて一定の水準を満たす出力を実現できることが確認された.特に温度パラメータτは出力の多様性に大きな影響を与え,τ =1.6 において Mean Cosine Distance が最大となり,創造性の高い出力が得られた.一方,類似度閾値θの調整は多様性への影響が限定的であり,冗長性の抑制には一定の効果を持つが,性能向上への貢献は限定的であった.また,最終的に LLM-as-a-Judge による主観評価を行った結果,(τ =1.6, θ=0.75)が全体として最も高い評価を獲得し,多様性と文体整合性のバランスが取れた出力を実現できていた.

ただし,本研究にはいくつかの制約が存在する.第 1 に,ペルソナ生成において現実的でない属性の組み合わせ(例：8 歳の PhD 取得者など)が発生する可能性があり,今後は属性間の整合性制約を導入する必要がある.第 2 に,生成件数が各感情カテゴリあたり 500 件に限定されており,大規模データによる安定性の検証には至っていない.第 3 に,人手評価の属人性を排除するため LLM-as-a-Judge を用いたが,この手法自体も評価モデルの能力に依存しており,より信頼性の高い評価スキームの構築が求められる.

これらの点を踏まえつつも,本研究の成果は,感情データの多様かつ自然な自己生成を実現する一つの有効な枠組みを提示するものであり,今後の感情理解タスクや生成データ拡張の基盤技術としての応用が期待される.

6. 結論

本研究では,ペルソナ情報と意味的類似度フィルタを組み合わせた感情テキスト生成フレームワークを提案し,その妥当性を多面的に評価した.実験の結果,特に温度パラメータτの調整が多様性に大きく寄与すること,LLM-as-a-Judge による自動評価が人間らしさの定量化に有効であることが確認された.また,提案手法により,感情表現の多様性・自然さ・識別性を兼ね備えた高品質なデータセットの構築が可能であることが示された.

今後は,ペルソナ生成の整合性向上,大規模コーパスの生成とその活用,評価 LLM のさらなる高度化に取り組みに加え,人間から収集した実データとの比較を目指す.また,文化

や言語の多様性を考慮したグローバル対応の感情データ拡張へと応用範囲を広げていくことが求められる.

謝辞

本研究は,JST 次世代研究者挑戦的研究プログラム JPMJSP2150 の支援を受けたものである.

参考文献

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713-3744, Jul. 2023.

[2] W. X. Zhao *et al.*, “A survey of large language models,” *arXiv [cs.CL]*, Mar. 2023.

[3] K. Inoshita, “Enhancing Sentiment Analysis Accuracy and Evaluating Task Affinity Using Large Language Models,” *IEEE ICoAILO*, Aug. 2025.

[4] A. T. Neumann, Y. Yin, S. Sowe, S. Decker and M. Jarke, “An LLM-Driven Chatbot in Higher Education for Databases and Information Systems,” *IEEE Transactions on Education*, vol. 68, issue. 1, pp. 103-116, Jun. 2024.

[5] J. Gu *et al.*, “A Survey on LLM-as-a-Judge,” *arXiv [cs.CL]*, Nov. 2024.

[6] S. Gupta *et al.*, “Bias runs deep: Implicit reasoning biases in persona-assigned LLMs,” *ICLR 2024*, pp. 1-26, Jan. 2024.

[7] T. Hu and N. Collier, “Quantifying the persona effect in LLM simulations,”in *Proceedings of the 62nd ACL*, pp. 10289-10307, Aug. 2024.

[8] P. H. L. de Araujo and B. Roth, “Helpful assistant or fruitful facilitator? Investigating how personas affect language model behavior,” *PLOS One*, Jul. 2024.

[9] L. Fröhling, G. Demartini, and D. Assenmacher, “Personas with attitudes: Controlling LLMs for diverse data annotation,” *arXiv [cs.CL]*, Oct. 2024.

[10] M. Kamruzzaman and G. L. Kim, “Exploring changes in nation perception with nationality-assigned personas in LLMs,” *arXiv [cs.CL]*, Jun. 2024.

[11] R. Li *et al.*, “How far are LLMs from being our digital twins? A benchmark for persona-based behavior chain simulation,” *arXiv [cs.CL]*, Feb. 2025.

[12] S. Giorgi *et al.*, “Modeling human subjectivity in LLMs using explicit and implicit human factors in personas,” in *Findings of the EMNLP 2024*, pp. 7174-7188, Nov. 2024.

[13] K. Inoshita, “Can a Large Language Model Generate Writing Styles across Ages,” *24th Forum on Information Technology*, Sep. 2025.

[14] S. D. Elliot Dauber, “Data Generation for NLP Classification Dataset Augmentation: Using Existing LLMs to Improve Dataset Quality,” *Stanford CS224N Custom Project*, Jun. 2024.

[15] C. K. R. Evuru, S. Ghosh, S. Kumar, R. S, U. Tyagi, and D. Manocha, “CoDa: Constrained Generation based Data Augmentation for low-resource NLP,”*in Findings of the NAACL 2024*, pp. 3754-3769, Jun. 2024.

[16] Z. Wang, J. Zhang, X. Zhang, K. Liu, P. Wang, and Y. Zhou, “Diversity-oriented data augmentation with large language models,” *in Proceedings of the 63rd ACL*, Jul. 2025.

[17] L. Jayawardena and P. Yapa, “ParaFusion: A large-scale LLM-driven English paraphrase dataset infused with high-quality lexical and syntactic diversity,” *arXiv [cs.CL]*, Apr. 2024.

[18] S. Gopali, F. Abri, A. Siami Namin, and K. S. Jones, “The applicability of LLMs in generating textual samples for analysis of imbalanced datasets,” *IEEE Access*, vol. 12, pp. 136451-136465, 2024.

[19] Y. Zhou, S. Shan, H. Wei, Z. Zhao, and W. Feng, “PGA-SciRE: Harnessing LLM on data augmentation for enhancing scientific relation Extraction,”in *Proceedings of the 23rd CCL*, pp. 352-369, Jul. 2024.

[20] P. Lippmann, M. T. J. Spaan, and J. Yang, “Illuminating blind spots of language models with targeted agent-in-the-loop synthetic data,” *arXiv [cs.CL]*, Mar. 2024.

[21] B. Ding *et al.*, “Data augmentation using large language models: Data perspectives, learning paradigms and challenges,” in *Findings of the 62th ACL*, pp. 1679-1705, Aug. 2024.

[22] L. Peng, Y. Zhang, and J. Shang, “Controllable data augmentation for few-shot text mining with chain-of-thought attribute manipulation,” in *Findings of the 64th ACL*, pp. 1-16, Aug. 2024,.

[23] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” in *Proceedings of the 36th NeulPS*, pp. 24824-24837, Jan. 2022.

[24] K. Inoshita, “Corpus development based on conflict structures in the security field and LLM bias verification,” in *Proceedings of the 4th NLP4DH*, pp. 504-512, Aug. 2024.

[25] S. J. Lee, H. Tony Lee, and K. Lee, “Enhancing emotion detection through ChatGPT-augmented text transformation in social media text,” in *Proceedings of the 33rd IEEE ROMAN*, Pasadena, pp. 872-879, Aug. 2024.

[26] Hugging Face, “sentence-transformers/all-MiniLM-L6-v2” Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. [Accessed: 08-Jun-2025].

[27] S. Shahriar *et al.*, “Putting GPT-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency,” *Appl. Sci. (Basel)*, vol. 14, no. 17, p. 7782, Sep. 2024.