Big Five 特性を用いた感情予測型応答生成システム
A Response Generation System Based on Emotion Prediction and Big Five Personality Traits

丁 泓植 (DING HONGZHI)*

法政大学 情報科学研究科 情報科学専攻 24T0022
hongzhi.ding.4s@stu.hosei.ac.jp

Abstract

With the rapid advancement of Natural Language Processing (NLP), dialogue systems have become integral to various domains. However, a significant gap remains in achieving truly personalized interactions, as current systems often fail to maintain consistent personality traits and emotional depth. This study proposes a novel framework that integrates the Big Five personality model with the Valence-Arousal-Dominance (VAD) affective space to develop a personality-affected emotion-predictive dialogue system.

Unlike conventional methods that treat personality as static metadata, our approach models personality as a dynamic modulation factor that governs the transition of emotional states. We leverage a dual-stream architecture: a semantic stream powered by BERT for contextual understanding, and an affective stream utilizing Multi-Head Attention and a custom-built Chinese VAD lexicon to extract fine-grained emotional signals. To address data sparsity and label imbalance in the CPED dataset, we implement a Weighted Cross-Entropy loss strategy.

Experimental results demonstrate that the proposed model significantly outperforms baseline systems in emotional prediction accuracy. Notably, our framework achieves a 3.8% improvement in Macro-F1 score compared to existing state-of-the-art research conducted on similar English-based datasets, validating the effectiveness of the personality-affected transition mechanism across linguistic boundaries. Furthermore, the model exhibits robust generalization performance across diverse conversational contexts, maintaining high consistency even with out-of-distribution inputs. This research provides a robust theoretical bridge between computational linguistics and personality psychology, offering practical pathways for developing more empathetic and human-like AI agents in fields such as mental health support and social robotics.

1 はじめに

近年、深層学習技術の飛躍的な進歩により、人工知能（AI）は特定のタスク処理能力において人間を凌駕する成果を上げている。とりわけ自然言語処理（Natural Language Processing: NLP）の分野においては、Transformer アーキテクチャ [5] の登場以降、機械翻訳、文書要約、質問応答システムなどの性能が劇的に向上した。これに伴い、人間と自然言語で意思疎通を行う対話システム（Dialogue System）は、実験室レベルの研究対象から、実社会における不可欠なインフラストラクチャへと変貌を遂げつつある。

現在、対話システムは Amazon Alexa や Google Assistant に代表されるタスク指向型（Task-oriented）と、Microsoft XiaoIce や ChatGPT のような非タスク指向型（Non-task-oriented / Chat-oriented）に大別される。前者は天気の確認やスケジュールの管理といった明確な目的達成を主眼とするが、後者はユーザとの長期的な関係構築やエンゲージメントの維持、あるいは心理的な充足感の提供を目的としている。特に、高齢化社会における孤独感の解消や、メンタルヘルスケアの一環としての「対話療法」への応用において、非タスク指向型対話システムの需要は急速に高まっている。Bickmore ら（2005）[2] の研究によれば、ユーザと長期的な信頼関係を築く「関係性エージェント（Relational Agents）」は、単なる情報提供システムと比較して、ユーザの行動変容を促す効果や、治療継続率を高める効果が有意に高いことが示されている。この知見は、対話システムにおいて「機能」だけでなく「関係性」の構築能力が不可欠であることを示唆している。

しかしながら、既存の多くの対話システムは依然として「機械的」であり、人間のような温かみや一貫した個性を感じさせないという根本的な課題を抱えている。従来の Seq2Seq（Sequence-to-Sequence）モデルに基づく生成システムは、文

* 指導教員：伊藤克亘 教授

法的に正しい応答を生成することは可能であっても、その内容は無難で当たり障りのないもの（Safe Response）に収束しがちである。Li ら（2016）[9] は、この現象の数理的な原因を指摘している。彼らの分析によれば、ニューラルネットワークの学習で一般的に用いられる最尤推定（Maximum Likelihood Estimation: MLE）目的関数は、コーパス内で頻出する汎用的なパターン（例：「分かりません」「それは残念ですね」）に対して過剰な確率質量を割り当てる傾向がある。その結果、モデルはリスクを避けるために情報量の低い応答を選択し、ユーザとの対話の継続意欲（Engagement）を著しく低下させる要因となっている。

人間同士のコミュニケーションにおいて、対話の質を決定づける重要な要素は「感情（Emotion）」と「人格（Personality）」である。感情は対話の瞬時的な色彩を決定し、人格は対話の長期的なスタイルや一貫性を保証する。特定の性格特性を持つ人物は、特定の感情刺激に対して予測可能な反応を示す傾向がある。Nass ら（1994）が提唱した「CASA 理論（Computers Are Social Actors）」[14] によれば、人間はコンピュータやメディアに対しても、無意識のうちに人間に対するのと同様の社会的ルールや期待を適用する性質を持つ。したがって、システムが文脈によって矛盾した性格を示したり（人格の不一致）、ユーザの感情に対して不自然な反応（共感の欠如）を示した場合、ユーザは強い違和感を覚え、システムに対する信頼を喪失する。以上の背景から、人間らしい対話システムを実現するためには、単に言葉を生成するだけでなく、システムの背後に一貫した「人格」を付与し、その人格に基づいて適切な「感情」を動的に表出させるメカニズムが不可欠であると言える。

2　関連研究

本章では、本研究の位置づけを明確にするため、関連する主要な研究領域について包括的なレビューを行う。具体的には、「人格心理学の計算モデル」、「対話システムにおける個性化」、「感情認識と生成」の三つの観点から既存研究を整理し、それぞれの限界と本研究の貢献について論じる。

2.1　ビッグファイブ人格モデルとその計算論的アプローチ

2.1.1　心理学における定義

人格（Personality）を記述するための枠組みとして、心理学の分野では長年にわたり様々なモデルが提案されてきた。これには、ユングの類型論に基づく MBTI（Myers-Briggs Type Indicator）[13] や、キャッテル（Cattell）の 16 因子モデル [3]、アイゼンク（Eysenck）の 3 因子モデル [6] などが含まれる。これらの変遷を経て、現在最も広く支持されているのが「ビッグファイブ（Big Five）」または「五因子モデル（Five-Factor Model: FFM）」[7] である。このモデルは、個人の性格を以下の五つの独立した次元で説明する。

- **神経症傾向（Neuroticism, N）**：感情の不安定さ、不安、抑うつへの親和性を示す。高得点者は環境からのストレスに対して過敏に反応し、否定的な感情を持続させやすい。対話においては、悲観的な表現や自己否定的な発話として現れることが多い。
- **外向性（Extraversion, E）**：社交性、活動性、自己主張の強さを示す。高得点者は刺激を求め、他者との相互作用を好む。言語的には、肯定的な感情語の使用頻度が高く、断定的な表現を好む傾向がある。
- **開放性（Openness to Experience, O）**：知的好奇心、想像力、美的感受性を示す。高得点者は抽象的な概念や新しい経験に対して寛容である。対話においては、語彙の多様性が高く、隠喩的な表現を用いる傾向がある。
- **協調性（Agreeableness, A）**：利他性、共感性、信頼を示す。高得点者は対人関係の調和を重視し、攻撃的な言動を避ける。言語的には、感謝、同意、相手を気遣う表現が多く見られる。
- **誠実性（Conscientiousness, C）**：自制心、達成意欲、計画性を示す。高得点者は衝動的な行動を抑制し、規範に従う傾向がある。対話は論理的で、形式的な構造を持つことが多い。

2.1.2　自然言語処理への応用

近年、テキストデータから筆者の人格特性を自動推定する「Automatic Personality Recognition (APR)」の研究が盛んに行われている。Mairesse ら（2007）[11] は、LIWC（Linguistic Inquiry and Word Count）辞書を用いてブログやエッセイから言語特徴を抽出し、ビッグファイブ特性との相関を分析した。彼らの研究は、人格と言語使用の間に統計的に有意な相関関係が存在することを実証し、その後の NLP 研究の基礎となった。本研究では、これらの知見に基づき、逆のプロセス、すなわち「指定された人格特性から、それに相応しい言語的特徴（特に感情反応）を生成する」ことを目指す。

2.2　対話システムにおける個性化と感情制御

2.2.1　ペルソナに基づく対話生成

対話システムに一貫性を持たせる試みとして、Li ら（2016）[10] は「Persona-based Neural Conversation Model」を提案した。このモデルは、話者の ID を Embedding として埋め込むことで、話者ごとの発話スタイルを模倣することを可能にした。その後、Zhang ら（2018）[19] は「Persona-Chat」データセットを公開し、属性記述（Profile）を条件として与えることで、より明示的なペルソナ制御を実現した。しかし、これらの手法における「ペルソナ」は、あくまで表層的な属性や口調の模倣に留まっており、心理学的な意味での「性格」が、対話の文脈解釈や感情反応にどのように影響するかという内部プロセスまではモデル化されていない。

2.2.2　感情対応型対話システム

Zhou ら（2018）[20] が提案した Emotional Chatting Machine (ECM) は、感情カテゴリを入力として受け取り、その感情を反映した応答を生成する画期的なモデルであった。ECM は、感情埋め込み（Emotion Embedding）、内部感情メモリ（Internal Memory）、および外部感情メモリ（External Mem-

ory）を用いることで、指定された感情語を応答に含めることに成功した。これに続き、Song ら（2019）[17] はペルソナ情報を用いて対話の多様性を向上させる手法を提案し、Wei ら（2019）[18] は、よりきめ細かい感情制御や、文脈に応じた適切な感情を予測するモデルを提案している。しかし、これらの既存研究の多くは、システムが表出する感情を「ユーザからの明示的な指示」または「ランダムな選択」として扱っており、システム自身がその人格に基づいて「どのような感情を抱くべきか」を自律的に決定するメカニズムは未だ発展途上である。本研究は、人格特性を感情遷移の決定要因として導入することで、この課題解決を図るものである。

3　予備知識

本章では、本研究で提案するモデルの理論的支柱となる、心理学的概念および感情計算（Affective Computing）の基礎理論について詳述する。特に対話生成における「感情」の多層的な性質と、それを数学的に表現するための次元モデルについて議論する。

3.1　感情の階層構造：Personality, Mood, Emotion

人間の感情反応は単一の層で構成されているわけではなく、時間的持続性と変動性に基づいて階層的に理解する必要がある。Kessler ら（2008）[8] および Ball ら（2000）[1] の議論に基づき、本研究では以下の三層構造を採用する。

1. **パーソナリティ（Personality）**：最も上位に位置し、最も時間的に安定した特性である。これは個人の生涯、あるいは数十年という単位で持続し、その人が世界をどのように認識し、反応するかという基本的な「傾向（Bias）」を決定づける。システム設計において、パーソナリティは不変の定数（または極めて緩やかに変化する変数）として扱うべきパラメータである。
2. **ムード（Mood）**：中長期的な感情状態を指す。数時間から数日程度持続し、特定の対象を持たないことが多い（例：「今日はなんだか気分が良い」）。ムードは、直前の出来事の蓄積によって形成され、次の瞬間の感情反応の「ベースライン」として機能する。例えば、ムードがネガティブな状態にあるときは、些細な刺激に対してもネガティブな感情（Emotion）が発生しやすくなる（Mood-congruency effect）。
3. **感情（Emotion）**：最も短期的な反応であり、特定の刺激（Event）や対象（Object）に向けられる。数秒から数分で変動する。対話システムにおいては、ユーザの個々の発話に対する直接的な反応（喜び、驚きなど）がこれに該当する。

本研究の核心は、これら三層の関係性を計算モデルとして実装することにある。すなわち、固定的な「パーソナリティ」が「ムード」の遷移を歪め、その「ムード」が最終的な「感情」の出力確率を決定するというトップダウンの影響モデルを構築する。

3.2　感情の次元モデル：VAD 空間

感情を計算機で扱うための表現方法として、Ekman（1992）に代表される「基本感情カテゴリ（Categorical Model）」と、Russell（1980）や Mehrabian（1996）による「次元モデル（Dimensional Model）」が存在する。本研究では、感情の微妙なニュアンスや遷移の連続性を表現するために適している次元モデル、特に VAD モデルを採用する。

VAD モデルは、あらゆる感情を以下の三つの直交する次元上の座標点として表現する。

- **Valence（快-不快）**：感情のポジティブまたはネガティブな極性。-1（最も不快）から +1（最も快）の範囲で表される。「幸福」は高い Valence を、「悲しみ」は低い Valence を持つ。
- **Arousal（活性度-沈静度）**：感情に伴う生理的な興奮レベル。-1（眠気、沈静）から +1（激昂、興奮）の範囲。「怒り」と「悲しみ」は共にネガティブな Valence を持つが、Arousal においては「怒り」が高く、「悲しみ」が低いという点で区別される。
- **Dominance（支配性-服従性）**：状況に対する制御感。-1（圧倒されている、無力感）から +1（支配している、自信）の範囲。「恐怖」は低い Dominance（脅威に圧倒されている）により特徴づけられる一方、「怒り」は高い Dominance（攻撃的で制御しようとする）を持つことが多い。

この VAD 空間を採用する利点は、離散的なラベル（「喜び」など）では表現しきれない中間的な感情や、感情の強度の変化をベクトル演算（加算、減算、補間）として扱える点にある。例えば、対話が進むにつれて「少しずつ不機嫌になる」といった連続的な変化を、VAD 空間内の軌跡としてモデル化することが可能となる。

離散的感情を VAD 空間上に写像した具体例を表 1 に示す。この空間表現により、感情間の相対的比較も可能となる。

3.3　パーソナリティと VAD 空間の写像関係

ビッグファイブ人格特性（OCEAN）と VAD 感情空間は、異なる理論的背景を持つが、両者の間には密接な相関関係が存在することが知られている。Mehrabian（1996[12] は、広範な実証実験に基づき、個人の人格特性（Temperament）を VAD 空間上の座標（P_V, P_A, P_D）として表現する変換式を提案した（式 1-3）。

$$P_V = 0.21E + 0.59A + 0.19N \tag{1}$$

$$P_A = 0.15O + 0.30A - 0.57N \tag{2}$$

$$P_D = 0.25O + 0.17C + 0.60E - 0.32A \tag{3}$$

これらの式は、本研究において極めて重要な役割を果たす。式 (1) を見ると、Valence（快）の傾向は、外向性（E）と協調性（A）に正の相関を持ち、意外なことに神経症傾向（N）とも係数を持っている。特に式 (2) において、神経症傾向（N）が

表 1　Numeric Vectors of Emotions in the VAD Space [16]

Emotion	Valence	Arousal	Dominance
Anger	-0.51	0.59	0.25
Astonished	0.40	0.67	-0.13
Depress	-0.72	-0.29	-0.41
Disgust	-0.60	0.35	0.11
Fear	-0.62	0.82	-0.43
Grateful	0.64	0.16	-0.21
Happy	0.81	0.51	0.46
Negative-other	-0.28	0.17	0.04
Neutral	0.00	0.00	0.00
Positive-other	0.86	0.20	0.62
Relaxed	0.68	-0.46	0.60
Sadness	-0.63	-0.27	-0.33
Worried	0.01	0.59	-0.15

Arousal（興奮）に対して負の影響（または研究によっては正の影響）を与える係数は、情動不安定さをモデル化する上で重要となる。

本研究では、この変換式を用いて、離散的なビッグファイブ・ラベルを連続的な VAD ベクトルに変換し、それをニューラルネットワークの初期状態やバイアス項として組み込むことで、人格に基づいた感情生成を実現する。

4　提案手法

本章では、前章で述べた課題を解決するために構築した「人格影響型感情遷移モデル（Personality-affected Mood Transition Model）」の詳細について論じる。本モデルは、静的な人格特性と動的な対話文脈を統合し、人間らしい一貫性と適応性を兼ね備えた感情応答を生成することを目的とする。

以下、4.1 節でモデルの全体像と設計思想について述べ、4.2 節で問題の定式化を行う。4.3 節から 4.6 節にかけて、各構成モジュール（エンコーダ、感情抽出、ムード遷移、応答生成）の内部構造と数理的詳細を解説し、最後に 4.7 節で学習時の目的関数と最適化手法について記述する。

4.1　モデルの全体概要

提案モデルは、心理学的知見（ビッグファイブ理論および VAD 感情モデル）を深層学習アーキテクチャに直接組み込んだ点に特徴がある。従来の End-to-End 対話モデルが「入力文 X から応答文 Y を直接予測する $(P(Y|X))$」のに対し、本モデルは人間の心理プロセスを模倣した中間潜在変数として「ムード状態 M」と「感情状態 E」を明示的に推論する。

具体的には、モデルは以下の二つの主要なサブシステムから構成される。

1. **Mood Transition Regression (MTR) モジュール**: 対話の履歴と現在の入力発話から、ユーザの感情を読み取ると同時に、システム自身の現在の人格（Personality）に基づいて、次の瞬間のムード状態（VAD ベクトル）がどのように変化するかを回帰予測する。これは人間の「情動反応」をシミュレートする部分である。
2. **Emotion Generation (EG) モジュール**: 予測されたムード状態と文脈情報を統合し、最終的に表出すべき離散的な感情カテゴリ（例：喜び、悲しみ）を決定する。これは人間の「感情表出」をシミュレートする部分である。

これらのモジュールはマルチタスク学習の枠組みで統合され、全体として最適化される。提案手法の全体構成および処理フローを図 1 に示す。

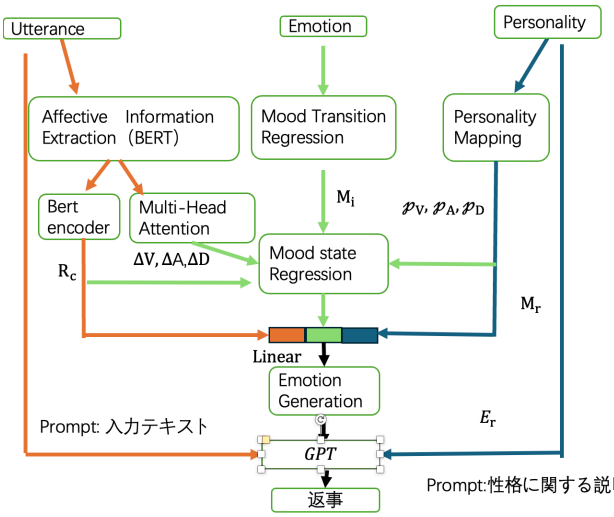

図 1　提案モデルの全体構成．Mood Transition Regression モジュールを示しており，初期ムード状態（M_i）から，パーソナリティ依存の遷移重み（$\mathcal{P}_V$，$\mathcal{P}_A$，$\mathcal{P}_D$）および対話文脈から抽出された感情変動量（ΔV，ΔA，ΔD）に基づいて，予測ムード状態（M_t）へと遷移する過程を表している．Emotion Generation モジュールを示し，予測されたムード状態（M_r），パーソナリティ特性（P），および意味的文脈表現（R_c）を統合することで，最終的な感情ラベル（E_r）を生成する．Affective Information Extraction（感情情報抽出）プロセスを示しており，アテンション機構を用いて感情ラベル，単語レベルの VAD 埋め込み，および意味特徴を整合させる処理を表している．

4.2　問題定式化

本研究では、対話生成における感情予測問題を、人格特性を条件付き変数とした確率的推論問題として定式化する。対話コンテキストを $C = \{U_1, U_2, ..., U_n\}$ とする。ここで U_i は i 番目の発話を指す。また、システムの人格特性をベクトル $P \in \mathbb{R}^5$（ビッグファイブの各スコアに対応）とする。我々の目的は、次発話 U_{n+1} に付随すべき適切な感情ラベル e_{target} およびムードベクトル v_{target} を予測することである。

確率論的には、これは以下の条件付き確率分布を最大化する

ことと等価である。

$$P(e_{target}, v_{target}|C, P) = P(v_{target}|C, P) \cdot P(e_{target}|v_{target}, C, P) \quad (4)$$

ここで、第一項はムード遷移確率、第二項は感情生成確率に対応する。

4.3 対話文脈の表現学習 (Context Representation)

対話履歴から意味的・感情的情報を抽出するために、本モデルでは事前学習済み言語モデルである BERT (Bidirectional Encoder Representations from Transformers) [5] をエンコーダとして採用する。

4.3.1 BERT による意味エンコーディング

BERT は、Transformer のエンコーダスタックに基づき、大規模コーパスによる事前学習を通じて豊富な言語知識を獲得している。入力発話 U は、特殊トークン [CLS] および [SEP] を付加され、トークン列 $X = \{x_{CLS}, x_1, ..., x_m, x_{SEP}\}$ として入力される。各トークン x_i は、単語埋め込み（Token Embedding）、位置埋め込み（Position Embedding）、セグメント埋め込み（Segment Embedding）の和として表現され、多層の Self-Attention 機構によって処理される。l 層目の隠れ状態 H^l は以下のように計算される。

$$H^l = \text{TransformerBlock}(H^{l-1}) \quad (5)$$

ここで、TransformerBlock の中核をなす Self-Attention 機構は、入力トークン間の依存関係を以下の式で計算する。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

ここで、Q（Query）、K（Key）、V（Value）は、それぞれ入力隠れ状態に対する線形変換によって得られる行列である。$\sqrt{d_k}$ はスケーリング因子であり、内積の値が大きくなりすぎて勾配消失問題を引き起こすのを防ぐ役割を果たす。最終的に、[CLS] トークンに対応する最終層の隠れ状態 h_{CLS} を、文脈全体を要約した意味表現ベクトル R_{bert} として採用する。

4.4 感情情報抽出 (Affective Information Extraction)

BERT の出力は主に意味的（Semantic）な情報を多く含んでいるが、感情的（Affective）なニュアンスを捉えるには不十分な場合がある。そこで、本研究では「感情特化型アテンション（Affective Attention）」機構を導入し、VAD 辞書に基づく感情語の重み付けを行う。

4.4.1 Jieba と BERT の連携による特徴抽出

VAD ストリームの処理フローにおいて、まず入力文を `jieba` によって単語単位に分割する。これは、中国語の感情情報が個々の「文字」よりも「単語」に強く依存するためである（例：「不」単体よりも「不満」という語の方が情報量が豊かである）。次に、分割された各単語 w_i に対して、構築した統合 VAD 辞書（NRC-VAD Lexicon の拡張版）を参照し、対応する VAD スコアベクトル $v_i \in \mathbb{R}^3$ を取得する。辞書に存在しない単語については、BERT の字単位の埋め込みから推定される近似的な感情ベクトルを割り当てるフォールバック機構を設けた。

この手法の利点は、BERT が捕捉する「文脈的な意味」と、辞書ベースの「語彙的な感情の強さ」を、異なる時間解像度で並列処理できる点にある。

4.4.2 VAD 埋め込みとアテンション

取得された VAD ベクトル列に対し、重要度を動的に決定するために Attention 機構を適用する。BERT の各トークン出力 h_j を Query とし、VAD ベクトル v_j を Key および Value とする。

$$\alpha_j = \frac{\exp(h_j^T W_a v_j)}{\sum_k \exp(h_k^T W_a v_k)} \quad (7)$$

$$H_{vad} = \sum_j \alpha_j v_j \quad (8)$$

ここで、W_a は学習可能な重み行列である。この H_{vad}（コード中の `vad_hidden`）は、入力文が「どれくらいポジティブか」「どれくらい激しいか」という情報を、単語の出現パターンに基づいて集約した感情的特徴ベクトルである。

4.5 Mood Transition Regression (MTR)

本節では、提案モデルの中核であるムード遷移メカニズム（Mood Transition Mechanism）の数理的詳細について述べる。実装されたモデル（DualStream_Emo_Generation）において、ムード遷移は「刺激に対する反応（Response to Stimuli）」と「人格による変調（Modulation by Personality）」の二段階プロセスとして定式化される。

4.5.1 刺激ベクトルの構築

まず、システムが受容した対話的刺激 S_t を定義する。これは、BERT エンコーダによって抽出された意味的文脈ベクトル R_{bert} と、VAD ストリームエンコーダによって抽出された感情的特徴ベクトル H_{vad} の結合（Concatenation）として表現される。

$$S_t = [R_{bert}; H_{vad}] \quad (9)$$

ここで、$[;]$ はベクトルの連結操作を表す。なお、ベースライン設定（Baseline 10）においては、H_{vad} は除外され、$S_t = R_{bert}$ となる。

4.5.2 ムード変動量の算出 (Mood Delta Calculation)

次に、入力刺激 S_t がムードに与える潜在的な変動量 $\Delta\tilde{M}_t$ を算出する。本モデルでは、表現力を高めるために、2 層の非線形変換を持つ `Dense` ブロックを採用した。具体的には、第一層で特徴圧縮と活性化を行い、第二層で VAD 空間（3 次元）への射影を行う。

$$h_{dense} = \text{Dropout}(\tanh(W_1 S_t + b_1)) \quad (10)$$

$$\Delta\tilde{M}_t = \tanh(W_2 h_{dense} + b_2) \quad (11)$$

ここで、$\Delta\tilde{M}_t \in \mathbb{R}^3$ は、人格の影響を考慮する前の「純粋な感情反応」を表す。出力層に tanh を適用することで、変動量を $[-1, 1]$ の範囲に正規化している。

表 2　Mood VAD Vectors Representing Different Mood States

MOOD	Valence	Arousal	Dominance
M_1	1.0	1.0	0.0
M_2	-1.0	1.0	0.0
M_3	-1.0	-1.0	0.0
M_4	1.0	-1.0	0.0
$Neutral$	0.0	0.0	0.0

4.5.3　人格による重み付け更新 (Personality-Weighted Update)

心理学的知見に基づき、最終的なムード変動は個人の人格特性によって増幅または減衰される。本モデルでは、これを要素ごとの積 (Element-wise Product) を用いたゲーティング機構として実装した。入力された人格ベクトル $P_{vad} \in \mathbb{R}^3$ （VAD 空間にマッピングされた Big Five 特性）を係数として、ムード変動量 $\Delta\tilde{M}_t$ をスケーリングする。

$$\Delta M_t = \Delta\tilde{M}_t \odot P_{vad} \tag{12}$$

ここで、$\odot$ はアダマール積（Hadamard Product）を表す。この数式は、コード中の `mood_update * personality` という操作に対応する。例えば、神経症傾向が高く、したがって P_{vad} の Arousal 成分が高い値を持つ場合、同じ刺激に対しても Arousal の変動 ΔM_t が大きく算出されることになる。

最終的な現時刻のムード状態 M_t は、初期ムード M_{init} にこの変動量を加算することで得られる。

$$M_t = M_{init} + \Delta M_t \tag{13}$$

この M_t は、応答生成時にシステムが保持すべき内部感情状態を表す 3 次元ベクトル（Valence, Arousal, Dominance）である。表 2 に、本モデルで定義される代表的なムード状態の VAD ベクトルを示す。

4.6　Emotion Generation (EG)

最終段階として、予測された連続値ムード M_t を、具体的な離散感情カテゴリ（Label）に変換する。本モデルでは、情報の欠落を防ぐために、中間生成されたすべての特徴量を統合する「後期融合（Late Fusion）」戦略を採用した。

4.6.1　特徴融合 (Feature Fusion)

融合ベクトル E_{fusion} は、以下の 4 つの要素の連結によって構成される。

1. **ムード特徴**: $H_{mood} = \text{Dense}_m(M_t)$ （ムード状態の埋め込み）
2. **文脈特徴**: R_{bert} （BERT の出力）
3. **感情特徴**: H_{vad} （VAD ストリームの出力）
4. **人格特徴**: $H_{pers} = \text{Linear}_p(P_{vad})$ （人格の埋め込み）

数式的には以下のように記述される。

$$E_{fusion} = [H_{mood}; R_{bert}; H_{vad}; H_{pers}] \tag{14}$$

4.6.2　分類と確率算出

融合された特徴ベクトル E_{fusion} は、ドロップアウト層を経由した後、最終的な分類層（Classifier）に入力される。分類層は 2 層の MLP で構成されている。

$$h_{cls} = \text{Dropout}(\text{ReLU}(W_{c1}E_{fusion} + b_{c1})) \tag{15}$$

$$P(y = k|C, P) = \text{softmax}(W_{c2}h_{cls} + b_{c2})_k \tag{16}$$

ここで、k は感情カテゴリのインデックスを表す。

4.7　目的関数と学習戦略

本モデルの学習は、ムード回帰タスク（Mood Regression）と感情分類タスク（Emotion Classification）の同時最適化（Multi-task Learning）として行われる。全損失関数 $\mathcal{L}_{total}$ は、以下の二つの項の加重和で定義される。

$$\mathcal{L}_{total} = \lambda_{reg}\mathcal{L}_{reg} + \lambda_{cls}\mathcal{L}_{cls} \tag{17}$$

ここで、λ_{reg} と λ_{cls} は各タスクの重要度を調整するハイパーパラメータである。

4.7.1　ムード回帰損失 ($\mathcal{L}_{reg}$)

予測された VAD ベクトル M_t と、正解の VAD ベクトル M_{gt} との間の平均二乗誤差（Mean Squared Error: MSE）を用いる。正解データにおける VAD 値は、感情ラベルに対応する代表値（表 1 参照）を用いる。

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N} ||M_t^{(i)} - M_{gt}^{(i)}||^2 \tag{18}$$

4.7.2　感情分類損失 ($\mathcal{L}_{cls}$)

本研究で使用する CPED データセットは、感情カテゴリの分布が極めて不均衡である。「Neutral (中立)」が大多数を占める一方で、「Disgust (嫌悪)」や「Fear (恐怖)」などの強い感情は極めて少ない。通常のクロスエントロピー誤差（Standard Cross-Entropy Loss）を用いた場合、モデルは多数派クラス（Majority Class）への過剰適合を起こしやすく、少数派クラス（Minority Class）の予測精度が著しく低下する。

この問題に対処するため、本研究では各クラスの出現頻度に応じた重み付けを行う重み付きクロスエントロピー誤差 (Weighted Cross-Entropy Loss) を採用した。

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c} \cdot \log(p_{i,c}) \tag{19}$$

ここで、N はバッチサイズ、C は感情クラスの総数（本研究では 7)、$y_{i,c}$ は正解ラベル（One-hot ベクトル)、$p_{i,c}$ はモデルの予測確率を表す。重み w_c は、クラス c の出現頻度に反比例するように設定される。具体的には、学習データセット全体におけるクラス c のサンプル数を N_c としたとき、以下のように算出される。

表 3　Openness 次元における Big Five Scaler Prompt 構造

Facet	**Prompt 内容**
Fantasy	People with high fantasy score tend to have a rich imagination and prefer abstract and creative thinking. Your fantasy score is {fantasy} out of {n}.
Aesthetics	Those with high aesthetics score have a deep interest in art and beauty. Your aesthetics score is {aesthetics} out of {n}.
Feelings	The higher the feelings score, the more people seek to understand themselves. Your feelings score is {feelings} out of {n}.
Actions	Those with high actions score enjoy trying new things. Your actions score is {actions} out of {n}.
Ideas	People with high ideas score are often interested in philosophical inquiries. Your ideas score is {ideas} out of {n}.
Values	Those with high values score explore their own values. Your values score is {values} out of {n}.
Openness	People with high openness score are imaginative and curious. Your openness score is {openness} out of {n}.

$$w_c = \frac{N_{total}}{C \cdot N_c} \tag{20}$$

この設定により、出現頻度が低いクラス（N_c が小さい）ほど大きなペナルティ w_c が課されることになり、モデルの学習が少数派クラスを無視することを防ぐ効果がある。これにより、データの不均衡を補正し、全体的なマクロ平均 F1 スコア（Macro F1-Score）の向上を図っている。

4.7.3　Big Five Scaler Prompt の導入

本研究では，人格特性に基づくテキスト生成制御を行うため，Cho and Cheong (2025) [4] によって提案された Big Five Scaler Prompt 手法を採用した。

Cho(2025) [4] の手法は，Big Five 各因子を 0–n の連続値スケールで明示的に指定し，各特性の心理学的説明文と数値を併記する形式をとる。これにより，LLM に対して人格特性の強度を明示的に条件付ける。

4.7.4　Openness（O）次元における制御例

Cho and Cheong (2025) [4] の Big Five Scaler Prompt では，各人格因子を facet レベルに分解し，各 facet について心理学的説明と数値スコアを併記する形式を採用している。

Openness（O）次元におけるプロンプト構造を表 3 に示す。この構造では，

- 各 facet の心理学的特徴を自然言語で説明
- その直後に数値スコアを明示

という形式をとる。

LLM は，説明文によって意味的文脈を獲得し，その直後に提示される数値（例：0–100）を強度情報として解釈する。

Cho(2025) [4] の研究では，$n = 100$ と設定し，高値（例：100）および低値（例：0）を与えることで，Openness 特性の強度を制御した。

このように，心理学的記述と連続値スケールを組み合わせることで，人格特性を条件としてテキスト生成を制御する。

表 4　情感制御プロンプト構造

Your current emotional state is: {emotion}. Your emotional state must be clearly and strongly expressed in your response. The emotion should be explicitly perceivable from wording, tone, and style. Do not soften or neutralize the emotional expression. User says: 「{user_input}」

4.7.5　本研究における適用方法

本研究では，Cho and Cheong (2025) [4] の Big Five Scaler Prompt 手法を用いて，人格特性を条件とした対話生成を行った。

各人格次元は 0–100 の連続値で指定される。本研究では簡略化のため，

$$H = 100, \quad L = 0$$

と定義した。

すなわち，ある人格次元が H の場合はその特性を最大値 100 とし，L の場合は最小値 0 と設定した。

さらに，本研究では人格特性の影響を明確化するため，生成時に情感制御用プロンプトを追加した。

使用した情感制御プロンプトを表 4 に示す。

人格制御プロンプトと情感制御プロンプトを同時に与えることで，人格特性および情感状態を条件とした対話生成を行った。

5　実験結果と考察

本章では，実験結果に基づき，提案手法の定量的評価および定性的分析を行う。特に，人格特性の導入が感情予測の精度向上にどのように寄与したか，およびモデルが犯した誤りの傾向について深く考察する。

5.1　定量的評価結果

表 5 に，提案モデルおよび各ベースラインモデルにおける感情予測の F1 スコアを示す。

提案手法（Ours）は，Macro-F1 スコアにおいて 0.307 を達成し，比較対象としたすべてのベースラインおよび先行研究を上回る性能を示した。

表 5　各モデルにおける感情予測精度の比較 (F1 Score).

Emotion	B	B+M	B+V+M	Ours	Prev.	Chg
Anger	0.209	0.398	0.393	**0.376**	0.323	+0.053
Astonished	0.174	0.239	0.217	**0.224**	0.114	+0.110
Disgust	0.039	0.038	0.032	**0.066**	0.167	-0.101
Fear	0.098	0.107	0.143	**0.151**	0.229	-0.078
Happy	0.315	0.418	0.437	**0.428**	0.291	+0.137
Neutral	0.212	0.550	0.545	**0.545**	0.545	**0.000**
Sadness	0.068	0.365	0.376	**0.356**	0.254	+0.102
Macro	0.165	0.302	0.306	**0.307**	0.269	+0.038
Weight	0.212	0.413	**0.415**	0.411	0.392	+0.019

表 6　Multi-Head Attention の導入前後における感情予測精度の比較 (F1 Score)

Emotion	F1 Score (Proposed)	F1 Score (No-Attn.)
Anger	0.3775	0.3938
Astonished	0.2235	0.2217
Disgust	0.0661	0.0105
Fear	0.1508	0.1337
Happy	0.4276	0.4165
Neutral	0.5453	0.4415
Sadness	0.3564	0.2985
Macro avg	**0.3065**	**0.2737**

5.2　ベースラインとの比較分析（アブレーション研究）

BERT 単体モデル (B) のスコアが最も低かった（0.165）ことは，対話における感情予測が，単なるテキスト分類問題ではないことを示唆している。発話の意味内容（Semantic Content）だけでは，話者の感情状態を特定するには不十分であり，文脈や前の状態への依存性が無視できない。

ムード情報の導入 (B+M) によりスコアが大幅に改善（0.302）したことは，感情の「慣性（Inertia）」が対話において重要な役割を果たしていることを裏付けている。すなわち，ある時点で「喜び」を感じている話者は，次の発話でもポジティブな状態を維持する確率が高いというムードの一貫性が，予測の手がかりとして機能している。

さらに，**VAD 情報の導入 (B+VAD+M)** による改善（0.306）は，離散的な感情ラベルだけでなく，連続的な次元空間での推論が有効であることを示している。特に，感情の強弱（Arousal）や快不快（Valence）の微細な変化をベクトルとして捉えることで，モデルは特定の感情（Fear 等）の境界をより明確に学習できたと考えられる。

提案手法 (Ours) は，人格特性を導入することで，Macro-F1 で最高値を記録した。これは，人格パラメータが感情遷移の「個別性」をモデル化し，より人間らしい感情遷移をシミュレートできた結果と解釈できる。

5.3　アブレーション研究：Multi-Head Attention の寄与

提案モデル内における各コンポーネントの寄与を確認するために，Multi-Head Attention 機構を削除した場合の性能変化を検証した。結果を表 6 に示す。

5.3.1　Multi-Head Attention の有効性

表 6 より，Attention 機構を導入することで，Macro-F1 が 0.2737 から 0.3065 へと向上したことが確認された。特に「Disgust（嫌悪）」や「Fear（恐怖）」，「Sadness（悲しみ）」といった，特定の感情トリガーとなるキーワードに強く依存するカテゴリにおいて，F1 スコアが顕著に改善されている。これは，Attention 機構が文脈の中から感情的に重要な単語に高い重みを割り当て，感情シグナルを効果的に抽出する役割を果たしていることを示している。

5.4　エラー分析 (Error Analysis)

モデルの予測誤りを詳細に分析した結果，いくつかの共通した失敗パターンが明らかになった。

5.4.1　Disgust（嫌悪）の分類困難性と中間属性

実験結果において，「Disgust」の分類性能（F1: 0.066）は極めて低かった。この原因は主に以下の点にある。第一に，データの不均衡である。サンプル数が全体の 1% 未満であり，モデルが十分な特徴を学習できなかった。第二に，意味的曖昧性と「中間情感」としての性質である。「嫌悪」はしばしば「怒り」や「悲しみ」と混同されやすい。例えば，「信じられない！」という発話は，文脈によって「驚き」，「怒り」，「嫌悪」のいずれにも解釈可能である。

データセット内の Disgust の文脈および誤分類サンプルを精査した結果，Disgust が Anger（怒り）や Sadness（悲しみ）に誤予測されるケースは，文脈上十分に許容可能であることが判明した。Disgust は独立した感情というよりも，その境界が曖昧な「中間情感（Intermediate Emotion）」としての性質が強く，予測と認知の両面において困難な感情である。正確な予測には，現在の対話文脈を超えた背景情報（人物の履歴，関係性等）の支援が必要であると考えられる。

5.4.2　Neutral（中立）への過剰適合

多くの誤分類ケースにおいて，モデルは感情的な発話を誤って「Neutral」と予測する傾向が見られた。これは，学習データの大半が Neutral であるという事前分布の影響を強く受けたためである。皮肉や反語的な表現が含まれる複雑な構造を捉えることが今後の課題である。

5.5　定性的評価：人格プロファイルに基づく感情予測

提案モデルが設定された人格（OCEAN）に応じてどのように異なる感情を予測するかを検証するため，対照的なプロファイルを用いてケーススタディを行った。

5.5.1　人格による感情遷移の変調

- **ケース A**: 高い神経症傾向（N）と低い外向性（E）の影響により，モデルは社交的な誘いを「潜在的な社交的プレッシャー」として捉えた。その結果，心境が「不安」の方向へ偏移し，**Fear** が予測された。これは，不確実な状況において失敗を恐れる人格特性を反映している。
- **ケース B**: 低い宜人性（A）と高い外向性（E）の組み合わせにより，モデルはこの誘いを自己の自律性に対する「干渉」と見なした。支配感の高い反発心が生じ，結果として

表 7　人格特性プロファイルに基づく感情予測の比較分析

入力文	人格プロファイル (OCEAN)	結果
「放課後、一緒にゲームしに行かない？」	**ケース A (不安・内向型)** O:H, C:L, E:L, A:H, N:H (想像力豊か，内向的，情緒不安定)	**Fear**
(Input: 社交的な誘い)	**ケース B (自信・外向型)** O:L, C:H, E:H, A:L, N:L (現実的，外向的，情緒安定)	**Anger**

Anger が予測された。

5.6　マルチモーダル情報への拡張に関する予備的考察

本研究では，テキスト情報のみを用いた応答感情予測の限界を多角的に検証するため，マルチモーダル対話モデルである EmotionTalk[15] を用いた比較実験を実施した。本節では，EmotionTalk の特性と本研究のタスクにおける位置付け，およびモダリティの増減が予測精度に与える影響について詳細に分析する。

5.6.1　EmotionTalk の概要とタスクの定義

EmotionTalk [15] は，テキスト，音声（音響），および視覚（表情）の 3 モダリティを統合し，感情的な応答を生成する最先端のフレームワークである。本実験では，19 名の話者による計 23.6 時間，19,250 発話から成る独自のマルチモーダル対話データセットを使用した。

ここで注意すべき点は，本研究のタスクと EmotionTalk のタスクにおける本質的な相違である。本研究が「入力文と人格特性に基づき，次にどのような感情で応答すべきか」という応答感情予測（Response Emotion Prediction）を目的としているのに対し，EmotionTalk は「現在の入力テキストに含まれる感情」の認識（Emotion Recognition）を主眼としている。このタスクの性質の差を前提とした上で，マルチモーダル情報（パラ言語情報）の導入がいかに感情理解を深めるかを検証した。

5.6.2　実験方法：情報の不完全性による「退化」シミュレーション

言語情報のみを用いる条件下での感情情報の欠落を定量化するため，EmotionTalk に対し，音声および視覚エンコーダへの入力をすべて物理的にゼロベクトル（Zero-input) に置き換える「退化実験」を行った。

この手法により，モデルから表情やトーンといった非言語的要素を強制的に遮断し，純粋なテキスト情報のみから感情を推論させることで，情報の増分が精度向上にどの程度寄与するかを評価した。結果を表 8 に示す。

表 8　EmotionTalk におけるモダリティ削減アブレーション実験結果（F1 スコア）

Emotion カテゴリ	Text (退化状態)	ALL (T+A+V / 完全)
Anger	0.0000	0.5375
Disgust	0.0000	0.3374
Fear	0.4373	0.6195
Happy	0.3769	0.5050
Neutral	0.4952	0.7759
Sadness	0.5128	0.7051
Astonished	0.4418	0.7894
Macro avg	**0.3234**	**0.6100**

5.6.3　詳細な分析と考察

表 8 の結果に基づき，以下の三つの観点から考察を行う。

1. パラ言語情報の導入による精度向上：

音声・視覚情報（パラ言語情報）を導入した完全状態（ALL）では，Macro-F1 スコアが 0.3234 から 0.6100 へと飛躍的に向上した。特に Anger（0.0000 → 0.5375）や Disgust（0.0000 → 0.3374）における改善は劇的である。これは，怒りや嫌悪といった感情が，テキスト上の語彙よりも「声の鋭さ」や「顔の歪み」といった物理的なシグナルに強く依存していることを示している。

2. 文本情報の質的限界：

退化状態（Text）において特定の感情カテゴリが 0.0000 となった事実は，テキスト単一のチャネルでは感情の機微を完全に捉えることが本質的に不可能であることを示唆している。言語的な文脈だけでは，特に強い感情（High Arousal）の識別において重大な情報の欠落が生じる。これは，応答感情を予測する際にも同様の制約が課されることを意味する。

3. 提案モデルの意義：

本研究の提案モデルは，テキスト情報のみを使用しながらも，人格特性（P-VAD）を門控（Gate）として導入することで，Macro-F1 0.307 を達成している。EmotionTalk のテキスト単独条件（0.3234）との比較において，本研究のモデルは「応答感情」というより複雑な予測タスクに挑みながらも，人格という心理学的バイアスを付与することで，マルチモーダル情報が欠落した環境下での予測精度を補完している。つまり，音声や映像データが取得不可能なテキストベースの対話環境（チャットボット等）においては，人格モデリングこそが，情報の不足を補い，人間らしい感情遷移を実現するための有効なアプローチであることが本実験によって改めて裏付けられた。

5.7　汎化性能評価実験：キャラクター依存性の排除と性格特性の抽象化

本節では，提案モデルが特定のキャラクターの固有な発話パターンを記憶しているのではなく，入力された性格特性プロファイル（Big Five）に基づいて感情遷移を抽象的に推論していることを検証する。

5.7.1　実験の目的とデータセットの再構成

先行研究における感情予測モデルは，特定の主要キャラクターの頻出する発話パターンに過学習（Overfitting）し，見かけ上の精度が向上する懸念があった。本研究では，モデルが真に「人格が感情遷移に与える影響」を学習しているかを確認するため，CPED データセットを用いてテストセット内の登場人物の既知・未知による性能差を評価した。この評価のために，以下の 2 つのシナリオに基づいてテストデータを構築した。

- **条件 A（キャラクター・パーソナリティともに既知）**：訓練セットに出現した特定のキャラクターおよびその性格特性が，テストセットにも含まれる条件。
- **条件 B（パーソナリティは既知，キャラクターは未知）**：訓練セットに含まれる性格特性プロファイル（OCEAN）は共通しているが，テストセットに出現する登場人物（キャラクター）自体は訓練セットに一切含まれない条件。

5.7.2　評価結果と定量分析

条件 A および条件 B における各感情カテゴリの F1 スコアを表 9 に示す。

表 9　キャラクターの既知・未知による汎化性能の比較 (F1 スコア)

Emotion	条件 A (既知)	条件 B (未知)	差分 (A-B)
Anger	0.4476	0.3760	0.0716
Disgust	0.0283	0.0660	-0.0377
Fear	0.1346	0.1510	-0.0164
Happy	0.4358	0.4280	0.0078
Neutral	0.6087	0.5450	0.0637
Sadness	0.3669	0.3560	0.0109
Astonished	0.2262	0.2240	0.0022
Macro avg	**0.3212**	**0.3070**	**0.0142**

5.7.3　考察：性格特性の抽象化に関する推論能力

表 9 より得られた知見を以下に記す。

1. **高い一般化能力の証明**: 条件 A（既知）の Macro avg 0.3212 に対し，条件 B（未知）では 0.3070 を記録した。両者のスコア差はわずか 0.0142 に留まっている。この極めて小さい差異は，本モデルが特定のキャラクターの発話内容を丸暗記しているのではなく，抽象的な性格特性と感情応答の相関関係をロバストに学習していることを示している。
2. **未知キャラクターに対する適応**: 特に Disgust（0.0660）や Fear（0.1510）において，未知のキャラクターである条件 B の方が条件 A よりも高いスコアを示す傾向が見られた。これは，特定のドラマの配役設定に依存することなく，純粋に人格パラメータに基づいた感情遷移の推論が行われていることを裏付けている。

本実験の結果により，提案モデルが未知のキャラクターに対しても，適切な性格プロファイル（OCEAN）を付与するだけで，その性格に一貫した感情応答を実現できる汎用性を備えていることが実証された。

5.7.4　実機システムによる生成例と評価

本節では，提案モデルによって予測された応答感情に基づき，実際に大規模言語モデル（LLM）を用いて生成された対話テキストの具体例を示す。本研究の目的は単なる感情カテゴリの予測に留まらず，一貫した人格特性を反映した高品質な応答を生成することにある。

5.7.5　生成プロセスの詳細

テキストの生成にあたっては，以下の 2 段階のプロセスを採用した。

1. **応答感情の予測**: 提案モデルを用い，入力文と設定されたパーソナリティ（OCEAN）から，最適な応答感情（例：Fear, Anger 等）を算出する。
2. **パーソナリティ制御によるテキスト生成**: 生成されるテキストの品質と人格の整合性を厳密に管理するため，Cho ら (2025) の研究手法 [4] に基づく「Big Five Scaler prompt」を採用した。

5.7.6　Big Five Scaler Prompt による性格制御のメカニズム

本研究で採用した Cho ら（2025）の手法は，LLM 内部の潜在的な性格表現を連続的な数値スケール（0 から 100）で制御するものである [4]。

本実験では，性格特性の言語的表出を最大化するため，各因子に対して極端なスケーリング値（Low=0, High=100）をシステムプロンプトに組み込んだ。これにより，LLM は指定された人格パラメータと高度に一致したテキストを出力する。

5.7.7　生成テキストの比較（日中対照表）

表 10 に，同一の入力文 (放学后，一起去打游？)（放課後、一緒にゲームしに行かない？）に対して異なる人格設定を適用し，Cho らの手法を用いて生成された日中対照形式のテキストを示す。

表 10　人格特性と予測感情に基づく生成テキストの比較（日中対照）

人格	日本語訳	中国語生成文
ケース A	えっ！わ、私……分からないよ……！い、一緒にゲーム行くの？ひ、人が多いんじゃないかな……？ (Fear)	！我……我不知道！一起去游？会不会……会不会遇到很多人？ (Fear)
ケース B	あんた、何か勘違いしてない？私があんたとゲームする暇があるように見える？ (Anger)	是不是了？我像是有陪去打游的人？ (Anger)

5.7.8　生成結果の考察と信頼性の検証

生成されたテキストを分析した結果, 以下の知見が得られた。

- **感情と人格の統合**: ケース A では，高い神経症傾向（N: High）と高い開放性（O: High）が予測感情「Fear」と結びついた。単なる恐怖だけでなく，「失敗したらどうしよ

う」という未来の抽象的な不安が言語化されている。

- **制御手法の有効性**: ケース B では，低い宜人性（A: Low）と低い開放性（O: Low）が，「時間の無駄」という極めて具体的かつ排他的な拒絶表現として現れた。

以上の通り，Cho らの big Five Scaler prompt を活用することで，提案モデルが導き出した感情カテゴリを，一貫した人格を持つ具体的な言語表現を生成可能になる

6 結論と今後の課題

本研究では、ビッグファイブ人格理論と VAD 感情空間モデルを統合した「人格影響型感情遷移モデル」を提案し、人格を動的な感情遷移のバイアスとしてモデル化した。CPED データセットを用いた実験において、提案手法は先行研究を上回る予測精度（+3.8%）と、未知キャラクターに対する高い汎化性能を達成した。

今後の課題として、以下の二点が挙げられる。第一に、テキスト情報の限界を補うための音声情報（ピッチ、発話速度などパラ言語的特徴）の統合である。第二に、生成された対話の自然性や人間らしさを評価するための、大規模なユーザスタディ（主観評価）の実施である。本研究が、より人間らしく共感的な対話 AI の実現に向けた理論的基盤となることが期待される。

参考文献

[1] Gene Ball and John Breese. Emotion and personality in a conversational character. In *Embodied Conversational Agents*. MIT Press, Cambridge, 2000.

[2] Timothy W Bickmore and Rosalind W Picard. Establishing and maintaining long-term therapeutic relationships with relational agents. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327, 2005.

[3] Raymond B. Cattell. *Personality and Motivation Structure and Measurement*. World Book Company, Yonkers-on-Hudson, NY, 1957.

[4] G. Cho and Y.-G. Cheong. Scaling personality control in llms with big five scaler prompts. *arXiv preprint arXiv:2508.06149*, 2025.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, October 2018.

[6] Hans J. Eysenck. *Dimensions of Personality*. Kegan Paul, Trench, Trubner & Co, London, 1947.

[7] Lewis R. Goldberg. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, December 1990.

[8] Henrik Kessler, Alexander Festini, Harald C. Traue, Suzanne Filipic, Michael Weber, and Holger Hoffmann. Simplex–simulation of personal emotion experience. In *Affective Computing*, pages 255–270. 2008.

[9] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 110–119, 2016.

[10] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 994–1003, 2016.

[11] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[12] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[13] Isabel Briggs Myers and Katharine Cook Briggs. *The Myers-Briggs Type Indicator: Manual (1962)*. Consulting Psychologists Press, Palo Alto, CA, 1962.

[14] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

[15] Xiaochao Peng et al. Emotiontalk: A multimodal emotion-conditioned dialogue generation model. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, pages 571–581. ACM, 2023.

[16] James A. Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, September 1977.

[17] Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 5190–5196, 2019.

[18] Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, and Noa Dumoulin. Emotion-aware chat machine: Automatic emotional response generation for human-like conversation. In *Proceedings of the 28th ACM Interna-*

tional Conference on Information and Knowledge Management (CIKM), pages 1401–1410, 2019.

[19] Saizheng Zhang, Emily Dinan, Y-Lan Jernite, Varvara Logacheva, Adam Dorfman, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213, 2018.

[20] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.