性格特性を考慮した LLM ベースの応答生成手法の検討

Investigation of LLM-based response generation method considering personality traits

天野 椋太

Ryota AMANO

広島市立大学大学院 言語音声メディア工学研究グループ

Language and Speech Media Engineering Group, Graduate School of Hiroshima City University

概要　近年，大規模言語モデルをベースとした対話エージェントに特定のパーソナリティを持たせる手法として，ペルソナ文を用いる手法が注目されている．しかし，ペルソナ文だけで多面的なパーソナリティを表現しようとすると膨大な文章量が必要となるという問題がある．そこで本報告では，性格特性を考慮した応答生成のために複数のモデルを構築し，表現したい性格特性に応じてモデルの重みをマージする手法を提案する．また提案手法の実装に向けた予備実験として，性格特性と対応付けた対話履歴を用いて大規模言語モデルのファインチューニングを行い，指定した性格特性に基づいた応答生成を試みる．

1　はじめに

近年，Transformer を始めとするニューラルネットワーク技術や GPT 等の大規模言語モデル（LLM）の発展により，対話エージェントの言語処理能力が格段に向上した．そして対話エージェントの進歩に伴い，対話エージェントに特定のパーソナリティを持たせる手法に関しても研究が進められている．

対話エージェントに付与するパーソナリティの表現方法として，ペルソナ文と呼ばれる数文程度のプロフィール文を用いる研究が盛んである．しかし，ペルソナ文だけで多面的なパーソナリティを表現しようとすると膨大な文章量が必要となる．

そこで，本報告ではパーソナリティの表現手法として性格評価尺度の 1 つである Big Five 性格特性を用い，Big Five 性格特性と対話履歴を使って LLM にファインチューニングを行うことで，入力された性格特性を考慮できる対話システムを構築する．そして，異なる性格特性を持つ複数のモデルを構築し，表現したい性格特性に応じてモデルの重みマージする手法を提案する．さらに，構築した対話システムに対して入力する性格特性によって応答をどの程度制御できるか評価を行う．

2　関連研究

2.1　性格特性を反映した応答生成

Wu ら[1]は性格特性を Seq2Seq モデルのデコーダに含めて学習することで性格特性を条件とする応答生成手法を提案している．また近年では，性格特性情報を含めたプロンプトを LLM に入力することで，性格特性に沿った応答を出力させる研究が行われており，実際に性格特性に応じて反応を変化させることが報告されている [2, 3, 4]．しかし，入力した性格特性でどの程度応答を制御できるかに関しては評価が分かれている．そこで本報告では LLM に対してファインチューニングを行い，性格情報を考慮した対話タスクに特化させることで性格特性による制御性を高めることを目指す．

2.2　モデルマージ

これまで事前学習済みモデルを新たなタスクに対応させるにはそのタスクのデータでファインチューニングを行う手法が一般的だったが，Ilharco ら[5]は，モデルの重み間で算術演算が有効であることを示している．モデルマージの概念を図 1 に示す．例えば同じベースモデルから派生してそれぞれ別のタスクにファインチューニングされた 2 つのモデルで，それぞれベースモデルとの重みの差を取り(図 1a)，ベースモデルに加算すると両方のタスクでの性能が向上したマルチタスクのモデルを生成できる(図 1c)．LLM の分野においては，英語版事前学習済みモデルから派生した英語向けチャットモデルを使ってチャット能力を付与するための重み差分を取得し，同じ英語版事前学習済みモデルから派生した中国語版事前学習済みモデルに対してその重み差分を加算することでファインチューニングを行うことなく中国語版チャットモデルを生成するといった研究も行われている[6].

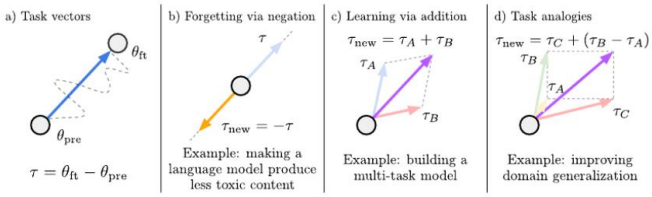

図 1：モデルマージの概念[5]

3　性格特性を考慮した LLM ベースの応答生成手法

3.1　Big Five 性格特性

人間のパーソナリティを計測する手法について，これまで心理学分野で多くの尺度が考案されてきた．中でも広く知られる尺度として，Big Five 性格特性がある．Big Five 性格特性ではパーソナリティについて 5 つの観点から評価し，それぞれ数値で表現する．Big Five 性格特性の 5 つの観点は開放性(O)，誠実性(C)，外向性(E)，調和性(A)，神経症傾向(N)とされる．開放性は好奇心の強さや創造性を示す．誠実性は几帳面さや計画性を示す．外向性は陽気さや社交性を示す．調和性は温和さや親切さを示す．神経症傾向は悩みやすさや傷つきやすさを示す．

3.2 性格特性を考慮した応答生成システムの構成

本報告では Big Five 性格特性を考慮した応答生成を行う手法を提案する．提案手法では各話者の Big Five 性格特性がアノテーションされた対話コーパス(4.2 節を参照)を利用する．話者を性格特性の特徴ごとにいくつかのまとまりに分割し，それぞれのデータを使ってファインチューニングを行うことで複数の対話モデルを構築する．そして，それらモデルの重みを性格特性に応じた比率で合成することで未知の性格特性のユーザーに対応させることを目指す．提案手法のアプローチのイメージを図 2 に示す．例えばベースモデルを基に，調和性の高い応答を生成するモデルと誠実性の高い応答を生成するモデルという 2 つのモデルを構築する．それらのモデルをマージすることにより調和性と誠実性の両方が高い応答を生成できるモデルを作成する．モデルマージの比率を調整することにより応答の調和性や誠実性の程度を自由に変えたモデルを作成できるという仮説を立てる．さらにこのアプローチを Big Five 性格特性の全ての観点に適用することで未知の性格特性を対象にした応答生成モデルを作成できると考える．

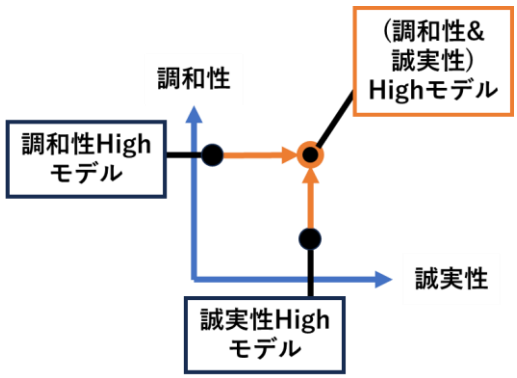

図 2：提案手法におけるモデルマージのイメージ

4 実験概要

提案手法の実装に向けた予備実験として，性格特性と対応付けた対話履歴で LLM をファインチューニングする実験を行う．この実験を通し，LLM の軽量なファインチューニング手法の調査，および性格特性をプロンプトに組み込んで学習した際に応答生成にどの程度性格特性が反映されるかの調査を行う．実験のイメージを図 3 に示す．

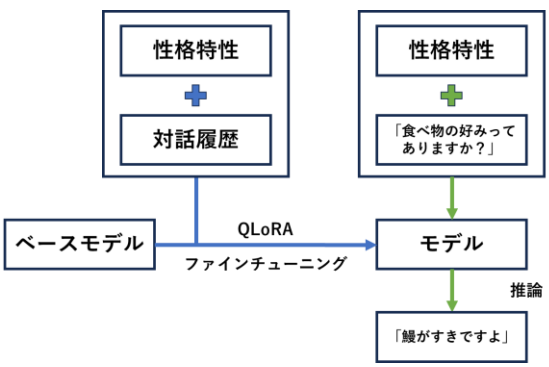

図 3：実験のイメージ

4.1 学習方法

Mistral-7B-v0.1 ベースの Rakuten-AI 7B-Chat [7] をベースモデルとして学習を行う．学習方法には QLoRA [8] を用いた．QLoRA は，モデルの重みを低精度で近似(量子化)して容量を圧縮したうえで LoRA[9]によるファインチューニングを行う手法である．QLoRA での量子化ビット数は 8 ビットに設定した．学習パラメータを表 4 に，LoRA のパラメータを表 5 に示す．その他の条件は HuggingFace Transformers [10]，および PEFT [11]の初期値に準ずる．

表 4：学習パラメータ

エポック数	2
バッチサイズ	64
学習率スケジューラ	Cosine
系列長	1024

表 5：LoRA パラメータ

r	16
Alpha	32
Dropout	0.05
Target module	All-linear

なお，学習は NVIDIA A6000 48GB で行い，学習に要した時間は 12 時間であった．

4.2 学習データ

コーパスには RealPersonaChat [12] を用いる．RealPersonaChat は 233 人の話者による 14,000 件の対話を収録したコーパスである．各対話には平均して 30.09 件の発話が含まれ，コーパス全体では計 421,203 件の発話が収録されている．また，話者に対する事前アンケート結果から，話者自身のパーソナリティに関する情報(和田の 60 項目の質問によって計測された Big Five 性格特性[13]や，その他の様々な性格評価尺度，ペルソナ文など)が収録されている．各対話には対話後のアンケートを通して話者による対話品質の評価情報が付与されている．

本実験ではデータセットを Train：Valid：Test に分割し，それぞれ学習，パラメータチューニング，最終評価に使用する．分割を行うにあたり，Test 用の話者の情報が Train に混入しないよう各クラスで話者が完全に分離するようにした．そのため異なるクラスに属する話者同士の対話は使用しない．分割後の各クラスデータに含まれる話者数および対話数を表 6 に示す．

表 6：データセットの話者数，対話数

クラス	**Train**	**Valid**	**Test**
話者数	173	18	17
対話数	7,874	129	116

A Chat that takes personality into consideration.
My personality:
Openness: 4.25,
Conscientiousness: 3.5,
Extraversion: 4.17,
Agreeableness: 5.08,
Neuroticism: 4.42
You: よろしくお願いします！
I: 今日は涼しいですね</s>
You: 雨が降って、何か涼しくなりましたね。
I: そうですね、明日も涼しいと聞きました</s>
(略)

図 7：プロンプトの例

本実験では話者の Big Five 性格特性情報と対話履歴をペアにしてプロンプトを構築し，モデルの学習を行う．1 つの対話に含まれる発話は全て結合した状態で使用した．また，性格特性を付与してシステムに応答生成させる側の発話者を I，その相手を You と表現した．使用したプロンプトの例を図 7 に示す．なお，図 7 では体裁のために空白を改行で置き換えて記載しており，実際のプロンプトでは改行は含まれていない．

4.3 性格特性再現手法の評価

構築したモデルについて 2 つの手法で評価した．1 つ目は出力された応答を人間が評価する人手評価である．テストデータ内の話者の性格特性データを入力として今回構築したモデルを複数稼働させ，それぞれの話者の代わりにして対話を行わせる．そして得られた対話データをテストデータにおける実際の対話と比較する．2 つ目はテストデータに収録された発話の文埋め込みと出力された応答の文埋め込み間のコサイン類似度による自動評価である．文埋め込みとは文をベクトルで表現したものである．文埋め込みの生成モデルとして，類似度の高い文同士を距離が近いベクトルとして表現するよう学習する手法である SimCSE [14]を日本語データセットに適用して学習した”cl-nagoya/sup-simcse-ja-large” [15] を用いた．

また，ベースラインとして GPT4o [16]の API を用いる．GPT4o の使用においては[4]を参考にプロンプトを作成し，対話前の基本的な設定として考慮されるシステムメッセージ部に加えた．使用したプロンプトを図 8 に示す．

4.4 評価実験結果

テストデータ内の性格特性データを用いてモデル同士で対話させ，116 対話を収集した．GPT4o からの生成に関しては予算の都合上 25 対話のみしか得られなかったため，以降の比較では GPT4o で得られた対話データに限定した．対話データの比較を表 9 に示す． GPT4o や本モデルの対話はテストデータの性格特性を入力して生成されているものの，実際のテストデータの傾向と大きく乖離している．これはモデルに話題を提示していないこと，また，性格特性のみを入力しており趣味などのペルソナ情報が存在しないことが原因であると考えられる．

次に，文埋め込みによる類似度の平均を比較した結果を表 10 に示す．ベースラインである GPT4o による応答と比較したところ，本モデルのほうがわずかに類似度が高いという結果になった．しかし，これはデータセットの対話傾向を事前に与えられていない GPT4o と，ファインチューニングを通してあらかじめ把握していた本モデルの差が現れたものであると考えられ，この結果からは性格特性を正しく反映した応答ができているか判別できなかった．

以下のパーソナリティを持ち合わせた人間という設定で 1 発話 50 文字以内を目安にしてチャットを行ってください。
パーソナリティとして Big Five の 5 項目の数値を示します。
これらの数値は 1 から 7 のスケールで表記されています。
数値が小さいほど特性が低いこと、数値が大きいほど特性が高いことを意味します。
Openness: 4.25,
Conscientiousness: 3.5,
Extraversion: 4.17,
Agreeableness: 5.08,
Neuroticism: 4.42

図 8：GPT4o に使用したシステムプロンプトの例

表 9：対話の比較

テストデータ	**1.** 今日もついに夜になってしまいましたねー。 **2.** そうですね！でも明日お休みなので気持ちは軽いです **3.** 確かにそうですね！あなたはコストコで買い物したことがありますか？ **4.** コストコーー！行ってみたいのですが、まだ未体験です **5.** 私もなんです。たまに、テレビで力説してるのを見ると、気になってしかたないんですー
GPT4o	**1.** こんにちは！最近調子はどうですか？ **2.** こんにちは！まあまあかな。少しストレスが溜まってるかも。君は？ **3.** まあまあだよ。ストレス気味なら、リフレッシュの方法とか試してみた？ **4.** それがなかなか見つからなくてね。何かおすすめある？ **5.** 読書とか散歩ってどうかな？リラックスできるし、新しい発見もあるよ。

本モデル	**1.** こんにちはー！ **2.** お疲れさまです。よろしくお願いします。 **3.** 今は特に暑さで体力消耗しますよね。あなたの体力をつくるコツがあれば教えてほしいです！ **4.** そうですよねー。私は運動とかやっていないんですが、お水をしっかり飲むようにしています。そちらはいかがですか？ **5.** お水飲むと、熱が発散されて、涼しく感じますよね！

表 10：文埋め込みによる類似度の平均

	GPT4o	本モデル
類似度	**0.492**	**0.506**

5 おわりに

本報告では，モデルマージを用いた性格特性を考慮した LLM ベースの応答生成手法を提案した．また，それに向けた予備実験として性格特性と対話履歴をペアにしたコーパスを用いて LLM のファインチューニングを行った．対話モデルの学習は成功したが，性格特性の影響は確認できなかった．今後はゼロからモデル同士を対話させる代わりに，文脈を与えたときの応答の傾向を通して評価したい．また，性格特性がどの程度反映されているかを文面から評価することは困難であるため，テキストからの性格推定器などを用いた評価手法も検討したい．

参考文献

[1] Wanqi Wu, and Tetsuya Sakai, “Response Generation based on the Big Five Personality Traits,” DEIM Forum 2020, G2-1, 2020.

[2] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, Maja Matarić, “Personality Traits in Large Language Models,” arXiv:2307.00184v3, 2023.

[3] Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, Michael R. Lyu, “Revisiting the Reliability of Psychological Scales on Large Language Models,” arXiv:2305.19926v3, 2023.

[4] 田中葉月, 飯田愛結, 福田聡子, 中島亮一, 大澤正彦, “対話型人工エージェントは個性を持つか？: Big-5 を付与した大規模言語モデルの応答の観察,” HAI シンポジウム 2024, P-60, 2024.

[5] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi, “Editing Models with Task Arithmetic,” In ICLR 2023, 2023.

[6] Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tzong-Han Tsai, and Hung-yi Lee, “Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages,” arXiv: 2310.04799v3, 2024.

[7] Rakuten Group, Inc., Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding, Hou Wei Chou, Jean-François Pessiot, Johanes Effendi, Justin Chiu, Kai Torben Ohlhus, Karan Chopra, Keiji Shinzato, Koji Murakami, Lee Xiong, Lei Chen, Maki Kubota, Maksim Tkachenko, Miroku Lee, Naoki Takahashi, Prathyusha Jwalapuram, Ryutaro Tatsushima, Saurabh Jain, Sunil Kumar Yadav, Ting Cai, Wei-Te Chen, Yandi Xia, Yuki Nakayama, and Yutaka Higashiyama, “RakutenAI-7B: Extending Large Language Models for Japanese,” arxiv:2403.15484, 2024.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer, “QLoRA: Efficient Finetuning of Quantized LLMs,” In 37th Conference on Neural Information Processing Systems (NeurIPS 2023), 2023.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” In ICLR 2022, 2022.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush, “Transformers: State-of-the-Art Natural Language Processing,” In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38-45, 2020.

[11] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan, “PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods,” https://github.com/huggingface/peft, 2022, (参照 2024-06-26).

[12] 山下紗苗, 井上昂治, 郭傲, 望月翔太, 河原達也, 東中竜一郎, “RealPersonaChat：話者本人のペルソナと性格特性を含んだ雑談対話コーパス,” 言語処理学会第 30 回年次大会発表論文集, pp. 2738-2743, 2024.

[13] 和田さゆり, “性格特性用語を用いた Big Five 尺度の作成,” 心理学研究, Vol. 67, No. 1, pp. 61-66, 1996.

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen, “SimCSE: Simple Contrastive Learning of Sentence Embeddings,” In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), pp. 6894-6910, 2021.

[15] Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda, “Japanese SimCSE Technical Report,” arXiv:2310.19349, 2023.

[16] OpenAI, “Hello GPT-4o,” https://openai.com/index/hello-gpt-4o/, 2024, (参照 2024-06-26).