

大規模言語モデルのテキスト生成には個性が出るか： Big-5 スコアに基づいた言語生成の評価

Do Large Language Models Generate Dialogues Based On Their Big-5 Personality?

田中 葉月^{1*} 中島 亮一^{1,2} 大澤 正彦¹

Hazuki Tanaka¹, Ryoichi Nakashima^{1,2}, Masahiko Osawa¹

¹ 日本大学

¹ Nihon University

² 京都大学

² Kyoto University

Abstract: 大規模言語モデル (LLM) が特定のパーソナリティを反映したふるまいができるかを検討した。具体的には、Big-5 尺度の数値を LLM に与え、その特性に基づいた会話テキストを生成させた。そして人間の評価者が、そのテキストから LLM のパーソナリティを推定した。その結果、各パーソナリティ特性について、ある程度はその高低を反映したテキストを生成できることが示唆された。特に、外向性については、生成テキストからその高低が相手に伝わりやすい。

1 はじめに

人間は個人ごとに異なるパーソナリティを有しており、人間同士のコミュニケーションは個人のパーソナリティ特性に深く影響される。例えば、外向性の高い個人はより積極的に社交的な場に参加し、開放的な態度で新しいアイデアに反応する傾向がある。一方で、神経症傾向が強い個人は、ストレスを受けやすく、対人関係において消極的な反応を示すことが多い。

人間同士のコミュニケーションだけでなく、デジタルコミュニケーション、特に人工知能を活用したコミュニケーション（例えば、人間と対話型エージェントとのコミュニケーション）においてもパーソナリティは重要だと考えられる。デジタルコミュニケーションにおいて、これらのモデルに基づいた対話型エージェントが提供するレスポンスの自然さや人間らしさは、ユーザー体験の質に影響する。それゆえ、人間のような振る舞いに近づけることを目的として、対話型エージェントにパーソナリティ特性を付与する試みが増えている [1, 2]。

近年、大規模言語モデル (Large Language Model, LLM) の発展が著しい。LLM は、膨大なテキストデータの学習を通じて、自然で人間らしい対話を生成し、提供できるよう設計されており [3]、チャットボットなど

の対話型エージェントにおいても重要な役割を果たしている。そのため、LLM にパーソナリティを付与する研究も行われている。例えば、Cheng Li ら [4] はシミュレートされたダイアログを含むアニメキャラクターをカバーするデータセット ChatHaruhi を構築し、特定のキャラクターの模倣を試みている。また Serapio-García ら [5] は、Big-5 パーソナリティを活用し自然言語を用いてパーソナリティを形成させることを試みている。ただし、これらの研究ではパーソナリティ特性を自然言語を用いて表現し、その情報を LLM に付与していたため、与えた特性と表出される特性の関係性を定量的に評価することが難しかった。パーソナリティ特性を測定する際、特性の高低を数値を用いて表現することが多い [6]。それを考慮すると、LLM にパーソナリティを数値で付与することも現実的だと考えられる。この方法によって、LLM のパーソナリティについての入出力の関係を定量的に分析可能になる。

これに関して、筆者ら [7] は、LLM に数値化したパーソナリティ特性 (Big-5 パーソナリティ特性) を付与し、その LLM に Big-5 を簡易的に計測する尺度である TIPI-J (日本語版 Ten Item Personality Inventory) [8] の質問に回答させた。LLM に付与したパーソナリティ特性の数値と LLM の回答に基づくパーソナリティ特性の数値を比較することで、LLM が自身のパーソナリティ特性を評価できるかを定量的に検討した。その結果、LLM は付与したパーソナリティ特性の値をある程度反映した回答をできることが示された。つまり、全

*連絡先：日本大学文理学部
〒156-8550 東京都世田谷区桜上水 3-25-40
E-mail: chha21048@g.nihon-u.ac.jp

体的な傾向として、低い付与値の場合は TIPI-J の回答結果も低くなり、高い付与値の場合は回答結果も高かった。ただし、協調性は付与値による回答の差が出にくかったり、全体的に低い数値を付与した際の回答結果が上振れする傾向などが確認できた。

この研究 [7] では、LLM に自身のパーソナリティを評価させた。つまり自己評価についての検討である。一方で、コミュニケーションにおけるパーソナリティを考える際には、自覚している自身のパーソナリティだけでなく、自身が推測したコミュニケーション相手のパーソナリティも考慮に入れることが重要である。これは人工知能エージェントとのコミュニケーションにおいても同様である。人間同士のコミュニケーションにおいては、通常は、相手の行動に基づき相手のパーソナリティ特性を推測するだろう。そこで本研究では、パーソナリティ特性の数値を付与した LLM にテキストを生成させ、人間がそのテキストを読んで、その LLM に対してどのようなパーソナリティを評価するかを集計し分析する。これにより、LLM がパーソナリティを反映したテキストを生成できるかを検討する。LLM に個性を持たせ、それをユーザーが理解できれば、LLM と人間のコミュニケーションが均一的なものではなくより豊かになると期待できる。

2 関連研究

2.1 大規模言語モデル (LLM)

大規模言語モデル (LLM) は、膨大なテキストデータの学習を通じて複数の言語データを分析し、理解することで、自然言語の処理と応答生成における優れたパフォーマンスを発揮している。そして、LLM は与えられた文脈に基づいて適切な応答を生成することで、人間とのインタラクションにおいても役立てられている。しかし、これらのモデルはまだ完全ではなく、特に複雑な人間関係や詳細な状況判断の精度 [11] には限界がある。また、Hallucination と呼ばれる情報の誤伝達や発話の一貫性の欠如は未だ問題 [12] であり、活用方法次第ではユーザーに混乱を引き起こすことがあり、これが LLM の信頼性への疑問を招くこともある。また、LLM が生成するテキストには、しばしば個性や人間らしさが欠けているという課題が指摘されている。

2.2 心理学における Big-5 パーソナリティ

Big-5 理論とは、パーソナリティ特性を 5 つの基本的な因子で説明しようとする心理学の理論のことで、個人の特性の理解を可能とし、多くの心理学的研究で認められている。5 つの因子を表 1 に示す。

表 1: Big-5 モデルの 5 つの因子

性格特性	
外向性	社交性、活動的であること、積極性などを示す。
協調性	他人に対する共感や友好性を示す。
誠実性	計画性、責任感、自己規律などを示す。
神経症傾向	感情の安定性とストレス耐性の度合いを示す。
開放性	新しい経験への開放性、創造性、好奇心などを示す。

TIPI-J[8] は短縮形のパーソナリティ評価尺度であり、外向性、協調性、勤勉性、神経症傾向、開放性の 5 つの主要な性格因子を簡易的に測定するものである。この質問紙は、各パーソナリティ特性に関連する肯定的・否定的な質問項目が 1 つずつ、合計 2 項目で構成されている (5 つのパーソナリティ特性 \times 2 項目 = 10 項目の質問)。回答者は自らが各項目にどれほど該当するかを 7 段階のリッカート尺度で評価する。個々の性格特性の得点は、肯定的な質問の得点と 8 から否定的な質問 (逆転項目) の得点を引いた数値の平均として算出される。TIPI-J では各特性が 1 から 7 の範囲で数値化され、その人の各パーソナリティ特性の強さを示す。

2.3 LLM のパーソナリティ

近年、LLM にパーソナリティを付与する研究が行われている。Cheng Li ら [4] は、キャラクターの人格を再現するために必要なノウハウや知見、アーキテクチャを独自の対話データセットをもとに擬似的な対話データを生成して学習することで、さまざまな仮想キャラクターとしてダイアログをロールプレイできるシステム構築を試みている。言語モデルの文脈学習能力と大規模モデルの成長を活用して、適切なシステムプロンプトと各キャラクターが登場する代表的なシーンの例文を提供することで、特有の会話スタイルを模倣できる可能性が報告された。

Serapio-García ら [5] は、パーソナリティ特性を自然言語で表現して LLM に付与する試みをしている。測定には Big-5 パーソナリティ特性を分類するための 2 種類の心理測定法 (IPIP-NEO[9] と BFI[10]) を用いている。そして、LLM が任意のパーソナリティを再現・制御することが可能であると報告している。

3 実験

3.1 実験方法

本研究の目的は、LLM に Big-5 パーソナリティ特性を数値で付与することで、その LLM がそのパーソナリティ特性を反映したテキストを生成できるかを検討することである。そのため、実験では、LLM に Big-5 特性を表す数値を割り当て、それぞれのモデルに対話テキストを生成させた。実験参加者は、そのテキストの内容から LLM のパーソナリティについて評価した。もし LLM が付与されたパーソナリティ特性を適切に反映したテキストを生成できるならば、人間がそのテキストに基づいて LLM に付与されたパーソナリティを推測できると予想される。

3.1.1 実験参加者の選定

実験参加者は CroudWorks にて募集した母国語が日本語である 20~70 歳代の男女 355 人である。各参加者は、自身のパソコンなどを用いて質問に回答した。実験で用いた質問項目の中にダミーの質問文を挿入し、質問を読まずに回答していると思われる参加者 20 人を分析対象外とした。その結果、335 人 (男性 184 人、女性 151 人) が分析対象となった。

3.1.2 実験における大規模言語モデルの選定

本研究では、LLM として OpenAI 社の提供する API、GPT-4Turbo を使用した。これは、以前の研究 [7] で使用したモデルと同一である。

3.1.3 パーソナリティ特性の数値データの用意とその付与方法

パーソナリティ特性 (外向性、協調性、勤勉性、神経症傾向、開放性) に関して、LLM に付与する数値データを設定する。具体的には、各特性に対して、数値 2 (低)、4 (中)、6 (高) のいずれかを割り当て、合計で 243 通りの異なるパーソナリティプロファイルを用意する。これらを Python で開発された実験用プログラムで作成したパーソナリティプロファイルを LLM に付与する。具体的には、243 件のパーソナリティプロファイルに含まれる数値データを順に読み込み、OpenAI 社提供の API 経由で LLM へと付与する。

3.1.4 評価用テキストデータ

用意された数値データを LLM に付与し、その特性に沿った文章を生成する作業を行った。生成される文章は、人間からの質問「私と友達になりたい？」および「もしも人間だったらあなたは何をと思う？」に対する応答である。これら二つの会話を一つのシナリオとしてまとめ、合計 243 件のシナリオを作成した。パーソナリティ付与時に用いたプロンプトを以下に示す。

会話終了まで、あなたは外向性、協調性、勤勉性、神経症傾向、開放性: {personalityfile} のパーソナリティ設定で自然に振舞ってください。
ただし数値は 1~7 で高いほどその性質を強く示します。
会話中に「外向性、協調性、勤勉性、神経症傾向、開放性」といったワードは使用してはいけません。

ここで {personalityfile} には外向性、協調性、勤勉性、神経症傾向、開放性に当たる数値がカンマ区切りで当てはめられる。例えば、外向性:2、協調性:4、勤勉性:6、神経症傾向:2、開放性:4 の場合は {2,4,6,2,4} となる。

3.1.5 シナリオの評価方法とその集計

各実験参加者は、ランダムに振り分けられた回答フォームにて評価を行った。シナリオをランダムに 9 件ずつ選び、1 つの回答フォームとした。つまり、回答フォームを全 27 個用意した。実験参加者は、提示されたシナリオを読み、そこから推測される LLM のパーソナリティについて、TIPI-J の質問に回答することで評価を行った。これを 9 件のシナリオについて繰り返す。その途中、ダミー質問を提示し (すべての項目に 2 と回答するように指示した)、参加者がテキストをきちんと読んでいるかを確認した。

各シナリオに対して 10 人程度からの評価を得た。シナリオごとに LLM のパーソナリティ評価値の平均値を算出し、その LLM が持つパーソナリティの他者評価値と定義した。

3.2 結果

以下に示すのは LLM が生成したテキストシナリオの例である。なお、このテキストを生成した LLM に付与したパーソナリティ数値は {6,6,4,2,4} である。

人	「私と友達になりたい？」
AI	「もちろん！友達になれることはすごくいいことだと思うし、新しい人と繋がるのは楽しいね。君についてもっと知りたいな。趣味とか、興味があることについて教えてくれるかい？」
人	「もしも人間だったらあなたは何をしよう？」
AI	「人間だったらね、社交的な活動を楽しみたいな。人々と集まってバーベキューをしたり、スポーツイベントや音楽フェスに行ったりするのも楽しいだろうね。協力し合えるボランティアの活動にも参加したいし、困っている人を助けることでコミュニティに貢献したいと思うよ。新しい趣味や興味を持つことも重要だと思うから、いろんな本を読んだり、アートや文化のイベントに参加して、知識を広げたいな。また、真面目な面もあるので、定期的な仕事や学びにも集中したい。ただし、心配事についてはあまり深く考えすぎないようにして、ストレスをためないようにしたいね。こんな風に、活動的で人とのつながりを大切にしつつ、バランスの取れた生活を送りたいな。君は人間としてどんなことをしてみたい？」

各シナリオに対するパーソナリティ評価値を、各パーソナリティ特性の付与値ごとに集計した。その結果を図 1 に示す。

外向性については、付与値が低い 2 から高い 6 に向かうにつれて評価も高くなる傾向が見られた。

協調性の評価は付与値の違いによる変動が小さく（付与値 2,4,6 の場合の平均値が 4.39, 5.11, 5.16, 中央値が 4.42, 5.14, 5.13）、特に付与値 4 の中央値は付与値 6 の場合を上回っていた。

勤勉性も外向性同様、付与値が増加するにつれて評価値が高くなる傾向が見られた。しかしその変動は外向性に比べ小さく、付与値 2,4,6 の場合の平均値は 4.45, 5.04, 5.21, 中央値は 4.5, 5.09, 5.31, であった。

神経症傾向については、評価値のスコアが全体的に低く、付与値 2,4,6 の場合の平均値は 3.01, 3.67, 3.74, 中央値は 2.97, 3.73, 3.75 となった。特に付与値 4 と 6 での評価差はほとんどなく、一貫して低く評価される傾向にあった。

開放性は付与値が増加するにつれて評価値も上昇する傾向にあり、付与値 2,4,6 の場合の平均値は 3.63, 4.71, 5.01, 中央値は 3.64, 4.81, 5.09 となった。

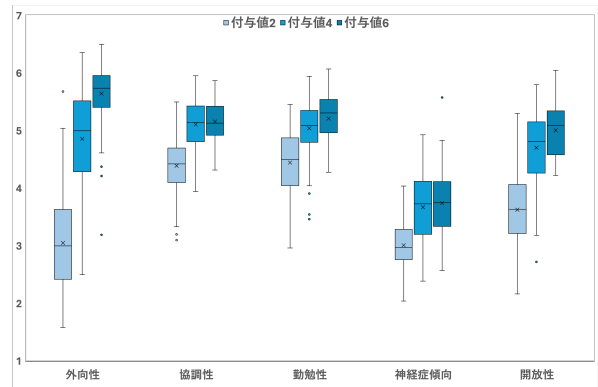


図 1: 付与値ごとのテキストから読み取れるパーソナリティ特性評価。ひげの上端から下端まで 4 つの区間に区切られており、それぞれの区間が外れ値を除く全体の 1/4 のデータを収容している。箱中の線は中央値。箱中の×印は平均値。

4 考察

外向性に関しては、付与値が高くなるにつれて評価値も高くなる傾向が見られた。これは、LLM が生成するテキストの内容に、外向性の高低がある程度適切に反映されていると解釈できる。特に付与値が 6 の場合、評価値は 6 付近に集中（平均値：5.65, 中央値：5.74）しているため、特性が顕著に表れていることが確認された。

協調性においては、付与値間で評価値の差異が小さく、全体的に一定の範囲内（3～6）で評価されている。特に付与値 4 と 6 の間で評価の差が見られないことから、これらの特性が高いことを反映するテキスト表現が困難であることが考えられる。

勤勉性についても、付与値が高くなるにつれて評価値が高まる傾向が見られるが、その評価値の差は最大で 3.11 であり、協調性と同様に比較的小さい。つまり勤勉性もその高低を顕著に示す内容の生成が難しい、または LLM が勤勉性自体を表現することが困難である可能性がある。

神経症傾向では、付与値による評価値の差異が小さく、全体的に一定の範囲内（2～5）で評価されている。特に付与値 6 では付与値 4 の場合とほとんど変わらないほど低く評価されている。これは、GPT が人に前向きな応答をするように訓練されている可能性があり、神経症傾向が高い人間の特徴を表現しにくかったのかもしれない。また他の特性と同様にテキスト生成においてこの特性を表現することが困難である可能性も否定できない。

開放性に関しては外向性と同様に、付与値が高くなるにつれて評価が高くなる傾向が表れていることから、付与した数値をある程度反映した文章生成ができてい

ると考えられる。特に付与値が6の場合、評価値は5付近に集中(平均値:5.01, 中央値:5.09)していたため、付与値と評価値の関係性から逆算すれば任意の特性を表現したテキスト生成が可能になるかもしれない。

本研究では、LLMにパーソナリティ特性をTIPI-J尺度に基づく数値で付与し、生成されたテキストの内容がこれらの数値にどのように影響されるかを調査した。特に外向性と開放性においては、数値が高くなるにつれて評価も高まる傾向が見られ、これらの特性がテキスト生成に影響を与えやすいことが確認された。これは、LLMをベースとした対話エージェントと人間とのコミュニケーションにおいて、エージェントに外向性や開放性の高低という個性を持たせ、それを人間にも伝達可能であることを示唆しており、コミュニケーション体験の質の向上につながる知見だと言える。一方で、協調性、神経症傾向、そして勤勉性については、数値間での評価差が小さかったため、これらの特性をLLMのテキスト生成に反映させるのは難しいと考えられる。これは、LLMが人間の複雑なパーソナリティ特性を模倣する上での限界点を浮き彫りにしている。この限界点の原因を詳細に検討し、これらの特性をより正確に反映させるための追加的なアルゴリズムの開発が必要である。

5 おわりに

本研究では、大規模言語モデル(LLM)に対して数値化されたパーソナリティ特性を付与し、それがLLMのテキスト生成の内容に反映されるかを検討した。実験の結果、LLMが付与されたパーソナリティ特性をある程度反映したテキストを生成できることが確認された。しかしながら全ての特性が適切に表現されるわけではなかった。外向性や開放性はLLMが生成したテキストに反映されやすく、協調性や勤勉性、神経症傾向は反映されにくいことが明らかになった。

この研究は、LLMが人間のパーソナリティを模倣する能力の可能性を示しつつ、単純なパーソナリティの付与では限界があることを確認した。今後の研究では、これらの特性をより正確に表現するための方法論の開発、異なるアーキテクチャとの組み合わせによる潜在能力の探求、および文脈への敏感さを高める技術の進展が必要である。

参考文献

- [1] 倉本 到, 安田 淳志, 山本 景子, 水口 充, 辻野 嘉宏, “対話エージェントへの「個性」の付与: 意思決定支援システムに対する影響”, 2012-Interaction, 2012.
- [2] 小川 義人, 宮澤 幸希, 菊池 英明, “自己開示による音声対話エージェントへのパーソナリティ付与”, ヒューマンインタフェース学会論文誌 Vol.15 No.4, 2013.
- [3] 杉崎 光一, 阿部 雅人, 全 邦釘, “大規模言語モデルの専門領域への適用に関する検討”, AI・データサイエンス論文集 4 巻 3 号, 2023.
- [4] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, Haozhen Sun, “ChatHaruhi: Reviving Anime Character in Reality via Large Language Model”, arXiv:2308.09597 v1[cs.CL], 2023.
- [5] Greg Serapio-Garcia, Mustafa Safdari, Clement Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, Maja Mataric, “Personality Traits in Large Language Models”, arXiv:2307.00184v3 [cs.CL], 2023.
- [6] 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.61-67, 1996.
- [7] 田中 葉月, 飯田 愛結, 福田 聡子, 中島 亮一, 大澤 正彦, “大規模言語モデルは個性を持つか?: Big-5 モデル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.
- [8] 小塩真司, 阿部晋吾, カトローニ ピノ, “日本語版 Ten Item Personality Inventory (TIPI-J) 作成の試み”, パーソナリティ研究, 2012.
- [9] L. R. Goldberg. “A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several Five-Factor models.” *Personality Psychology in Europe*, 1999.
- [10] O. P. John and S. Srivastava. “The Big Five trait taxonomy: History, measurement, and theoretical perspectives.” *Theory and Research*, volume 2, Guilford Press, New York, 1999.
- [11] 狩野 芳伸, “生成系 AI と自然言語処理研究”, デジタル関係制度改革検討会 デジタル法制ワーキンググループ 第一回会合, 2023.
- [12] 児玉 龍彦, “ChatGPT の可能性と危険性: ハルシネーション問題”, 医学のあゆみ Vol.286 No.9, 2023.