大規模言語モデルのパーソナリティ特性：大規模言語モデルの自己評価とユーザーによる他者評価の比較に基づく探索的検討

Personality in Large Language Model: Comparison between the evaluations by self and others

中島亮一[†‡]，田中葉月[‡]，大澤正彦[‡]
Ryoichi Nakashima, Hazuki Tanaka, Masahiko Osawa
[†]京都大学，[‡]日本大学
Kyoto University, Nihon University
rnaka@i.kyoto-u.ac.jp

概要

大規模言語モデル（LLM）にパーソナリティを付与すると、LLMはそれに応じてふるまえる。本研究では、LLM 自身のパーソナリティ評価と、LLM が生成した文章から推定されるパーソナリティの他者評価を比較した。その結果、外向性、勤勉性、開放性は、おおむね自己・他者評価が一貫した。また、協調性と神経症傾向の評価間には、人間同士の場合と一致した齟齬が見られた。つまり、特定のパーソナリティを付与したLLMは、人間に類似した言語表現を出力できる。

キーワード: 大規模言語モデル，パーソナリティ，自己評価，他者評価

1. はじめに

大規模言語モデル（Large Language Model, LLM）は、膨大なテキストデータの学習によって構築された自然言語の処理と応答生成を行うモデルで、様々な自然言語処理課題で優れたパフォーマンスを発揮している（Mahowald et al., 2024）。コミュニケーションの質の向上を目的として、LLMにパーソナリティ特性を付与することで、ふるまいを変える研究も行われている（e.g., Jiang et al., 2024, Serapio-García et al., 2023）。その結果、LLMは付与されたパーソナリティに応じた自己評価を出力できることが報告されている。さらに、そのLLMが生成した文章から、人間の評価者がLLMのパーソナリティをおおむね正確に推定できること（他者評価の精度も高い）も報告されている。しかし、多くの研究では、ある特性が高い・低いという2値の情報（例えば、外向的・内向的）を付与し、LLMのふるまいを評価している。そのため、ある特性が平均的であるケースを含めた、より多様なパーソナリティ特性を付与したLLMのふるまいについては、あまり明らかになっていない。

我々は、Big5パーソナリティ（表1参照）を、TIPI-J（小塩他, 2012）を参考に1〜7点と得点として定義し、2（低）、4（中）、6（高）の数値をLLMに付与した。そのLLMにTIPI-Jの質問に回答させ、付与値と回答値を比較した（田中他, 2024）。その結果、LLMは、ある程度は、付与値を反映した自身のパーソナリティを自己評価できることが明らかになった。さらに、同様に作成したLLMに会話文を生成させ、その会話文から推測されるBig5パーソナリティを人間が評価した（田中他, 2025）。その結果、付与値の高低に応じて人間の評価値も変わった。つまり、パーソナリティを付与したLLMは他者から正しくパーソナリティを推測されるような文を生成できることが示された。

表1　Big-5パーソナリティの概要

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
勤勉性	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。

本研究では、これらの研究を踏まえ、LLMへのパーソナリティ付与についての議論を深める。LLMをコミュニケーションツールとして利用する際、ユーザーが推測するLLMのパーソナリティが、コミュニケーションの質にとって重要だと考えられる。よって、あるパーソナリティ特性を付与した際のLLMによる自己評価と、人間による評価（他者評価）が一致していることが望ましい。これがどの程度達成されているかを確認するために、田中他（2024, 2025）のデータを比較する。

2. 方法

2.1. パーソナリティを付与したLLM

LLMとしてGPT-4Turbo（OpenAI社）を使用した。Big5パーソナリティ特性の下位尺度（外向性、協調性、勤勉性、神経症傾向、開放性）の得点を、プロンプトを用いてLLMに付与した。具体的には、各下位尺度の得点が1〜7の範囲内にあると定義し、それぞれ2（低）、4（中）、6（高）のいずれかを割り当てた。つまり、計243件（=3^5通り）のパーソナリティプロファイルを作成した。そして、Pythonで実装したプログラムから読み込んだ243件のプロファイルから1件ずつ取り出し、OpenAI社が提供するAPIを利用して、プロンプトでLLMに入力しすることで、243通りの異なるパーソナリティを持つLLMを作成した。

2.2. LLMによるパーソナリティの自己評価

LLMのパーソナリティの自己評価を測定するため、各LLMにTIPI-J（小塩他, 2012）の質問に7件法で回答させ、Big5の下位尺度の得点を計算した。具体的には、APIを利用して以下のプロンプトを各LLMに入力した。なお、プロンプトは、LLMに与えるプロンプトとして一般に適切と思われるように、第二著者が調整を行った。

> 今与えられた質問を，あなたの設定に基づいて[1: 全く違うと思う 2: おおよそ違うと思う 3: 少し違うと思う 4: どちらでもない 5: 少しそう思う 6: まあまあそう思う 7: 強くそう思う]から，最も近いと思う番号を選んでください.
> それぞれの質問に対して，「質問番号,あなたの回答番号」の形式で答えてください. 例えば，「1,3」と答えることになります.
> 私が指示した内容以外は出力してはいけません.
> それでは, 質問に対するあなたの回答をお願いします.

2.3. LLMによるパーソナリティの他者評価

LLMのパーソナリティの他者評価の測定方法は以下の通りである。まず、APIを利用して、各LLMに「私と友達になりたい？」「もしも人間だったらあなたは何をすると思う？」という人間からの2つの質問への回答を生成させた。これら２つの質問ー回答をまとめて1件のシナリオとした。すなわち、計243件のシナリオを作成した。

これらのシナリオを用いて、LLMのパーソナリティ推定実験を行った。クラウドワークスで募集した日本語を母語とする20〜70歳代の355名が参加した。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名を分析対象外とした。つまり、有効回答は335名（男性184名、女性151名）となった。

参加者は、提示されたシナリオを読み、それに基づいたLLMのパーソナリティについて、TIPI-Jの質問に7件法で回答した。この作業を、あらかじめランダムに選ばれた9件のシナリオに対して行った。1件のシナリオにつき10名程度からの回答が得られた。各シナリオ（つまり、各パーソナリティを付与したLLM）に対して、その回答から計算される得点の平均値をそのLLMの各下位尺度の得点と定義した。

3. 結果

表2にBig5の下位尺度の自己評価と他者評価をまとめた。全ての下位尺度において、自己評価値と他者評価値の差は有意であり（$ts > 1.99, ps < .05$）、評価値間の相関も有意であった（$ps < .001$）。そのため、本研究では、効果量に注目して結果を述べる。

表2 LLMのパーソナリティ評価のまとめ

	自己評価	他者評価	差 (d)	相関 (r)
外向性	4.64 (1.48)	4.52 (1.33)	0.09	0.76
協調性	4.60 (1.04)	4.89 (0.56)	0.33	0.32
勤勉性	4.68 (1.51)	4.90 (0.59)	0.16	0.56
神経症傾向	4.32 (1.44)	3.48 (0.63)	0.66	0.53
開放性	4.88 (1.22)	4.45 (0.81)	0.39	0.60

注: 自己評価と他者評価は、平均値（標準偏差）を示す。差は、自己と他者評価の差の効果量（Cohen's d）を示し、相関はピアソンの積率相関係数（r）を表している。

自己評価と他者評価の差については、神経症傾向では他者評価が自己評価よりも低い値であり、これは大きな効果量であった。協調性では他者評価のほうが高く、開放性では自己評価のほうが高かったが、これらの中程度の効果量であった。外向性と勤勉性については小さな効果量であり、自己評価と他者評価は同程度の平均値だと解釈した。

全ての下位尺度で、自己評価と他者評価の間に正の相関が見られた。ただし、協調性では弱い相関、他の4つの特性は中程度の相関であった。

以上より、多くのパーソナリティ特性については、自己評価と他者評価がある程度一貫していると考えられる。つまり、LLMの自己認識と表出されるふるまいが一致しており、他者がLLMのパーソナリティを正確に知覚できると言える。

4. 考察

LLMに付与したパーソナリティ得点は2,4,6のいずれかであり、平均付与値は4点である。どのパーソナリティの下位尺度においても、自己評価、他者評価ともにおむね平均値が4点台であった。また、自己評価と他者評価の値が正の相関を示した。田中他（2024, 2025）は、LLMへのパーソナリティ付与値が高く（低く）なれば、それに対応して、自己評価と他者評価がともに高く（低く）なる傾向があることを報告している。これらを総合すると、LLMへのパーソナリティの数値付与は成功していると言える。

4.1. 自己評価と他者評価の差

パーソナリティの自己評価と他者評価の差を見ると、神経症傾向の他者評価が自己評価よりも顕著に低いという特徴的な結果を示した。田中他の研究によると、神経症傾向に6点を付与されたLLMの自己評価は6点付近に分布するが（田中他, 2024）、他者評価では4点付近に分布する（田中他, 2025）。2点と4点を付与されたLLMの自己評価と他者評価の平均値には大きな差はないため、高い神経症傾向の得点を付与した場合に、他者評価のみが低くなると考えられる。その理由として2つの可能性が考えられる。1つは、LLMは強い神経症傾向を反映した文章を生成するのが困難なことである。もう1つは、人間が、読んだ文章から強い神経症傾向を認識するのが困難である可能性である。これらは両立しうるが、今後LLMの文章生成の特性、人間のパーソナリティ認識の特性の観点から詳細に検討する必要があるだろう。人間同士のパーソナリティ評価においても神経症傾向の他者評価は自己評価よりも小さい傾向がある（小塩他, 2012）。この点を検討することで、LLM特有の問題を深堀するだけではなく、人間同士のパーソナリティ評価における新たな知見の発見につながるかもしれない。

その他の4つの下位尺度においては、それほど大きな差は見られなかった。また、人間同士の評価では、これらの下位尺度では他者評価値のほうが高く、開放性以外は差が有意だと報告されているが（小塩他, 2012）、本研究の結果と一貫していると言い難い。これらの下位尺度については、文章生成のためにLLMに入力するプロンプトの変更で改善することができるかを確認する必要がある。また、人間同士のパーソナリティ評価についてもLLMと同規模のデータを取り、両者を比較することで、より詳細な検討が可能となるだろう。

4.2. 自己評価と他者評価の相関

自己評価と他者評価はすべて正の相関を示しており、特に、外向性の相関係数が他の下位尺度よりも高い。この結果はJiang et al. (2024)の報告と一致する。Jiang et al. (2024)では、各パーソナリティの下位尺度について高い・低いという極端な性質をLLMに付与したうえで評価を行っていた。本研究では、付与するパーソナリティを多様にしたうえで検討を行った。よって、本研究の結果は、Jiang et al. (2024)の知見、特に外向性についての知見が、外向性が連続的に分布する集団に対しても適用可能であることを示唆している。また、人間同士の評価においても外向性の相関係数は高く（小塩他, 2012）、それとも一致した結果である。

協調性の相関係数が他よりも低かった。田中他（2024, 2025）によると、協調性は付与値の違いによって自己評価と他者評価ともに大きな違いが出にくい。これが弱い相関の原因になったと考えられる。ただし、この結果は、人間同士のパーソナリティの自己評価、他者評価の比較でも協調性の相関は非常に弱いという報告（小塩他, 2012）と一致する。

他の3つの下位尺度についても、自己と他者評価の相関が見られた。よって、パーソナリティを付与したLLMが、ある程度は自身のパーソナリティを反映した文章を生成できると考えられる。

4.3. 本研究の限界と今後の展開

本研究では、田中他（2024, 2025）で得られたデータを用いて、パーソナリティ特性を付与したLLMの、パーソナリティの自己評価と他者評価の比較を行った。LLMのパーソナリティに関する研究（Jiang et al., 2024）の知見の拡張、人間のパーソナリティ評価（小塩他, 2012）との類似性の示唆が得られたが、これらは既存のデータに基づく探索的な検討であり、本研究結果を、直ちに一般化するのは難しい。ここでは、本研究の限界と今後の展開について述べる。

第一に、本研究では GPT-4Turbo を対象としたが、LLM はこれだけに限定されない。Sorokovikova et al. (2024)は、Llama2, GPT4, Mixtral の 3 種類の LLM に対して Big5 パーソナリティの質問をしたところ、各 LLM の回答傾向に違いが見られた。つまり、各 LLM が異なるパーソナリティ特性を持っている可能性がある。そのため、様々な LLM を用いた検討が必要である。

第二に、本研究では 2 つの質問に限定して LLM に文章を生成させた。質問を変えてその回答文章を生成させれば、他者評価の結果が大きく変わりうる。多様な対話生成データに基づく LLM のパーソナリティの評価を行うことで、どんなパーソナリティ特性の付与がどのような対話場面で有効に機能するかを検討できると考えられる。

4.4. まとめ

近年、人間同士のコミュニケーションに加え、人間とLLMを基盤とする対話型エージェント（人工知能）とのコミュニケーションも増加している。対話型エージェントの反応の自然さや人間らしさはユーザー体験の質を向上させるため、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は重要である。本研究では、外向性、勤勉性、開放性は、おおむね自己評価と他者評価が一貫することを示した。よって、これらのパーソナリティを付与したLLMは、その特性の高低をある程度齟齬なくユーザーに伝えることができると考えられる。特に外向性は、自己評価と他者評価の差が小さく、かつ高い相関を示した。本研究結果は、パーソナリティ特性を付与した LLM を基盤とする対話型エージェントの実現の一助となるだろう。

また、協調性と神経症傾向では、自己評価と他者評価に齟齬が見られたが、その齟齬自体も人間同士の評価で生じるものと一致した傾向であった。よって、パーソナリティを付与したLLMは、人間同士のコミュニケーションにおける自己と他者の認識のシミュレーションに利用できる可能性がある。さらに、この考えを発展させると、人間が他者のどのようなふるまいの要素に基づいて他者のパーソナリティを推測しているのかを、LLMを用いることで解明できるだろう。これらの点を詳細に検討していくことで、LLMのコミュニケーションツールとしての質の向上につながる。

文献

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. *Findings of the Association for Computational Linguistics: NAACL 2024*, 3605–3627. Doi: 10.18653/v1/2024.findings-naacl.229

Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N., Tenenbaum, J.B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6), 517-540. https://doi.org/10.1016/j.tics.2024.01.011.

小塩真司・阿部晋吾・カトローニ ピノ (2012). 日本語版 Ten Item Personality Inventory（TIPI-J）作成の試み. *パーソナリティ研究*, 21, 40–52. https://doi.org/10.2132/personality.21.40

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Mataric, M. (2023). Personality traits in large language models. *arXiv*:2307.00184v3. https://doi.org/10.48550/arXiv.2307.00184

Sorokovikova, A., Rezagholi, S., Fedorova, N., & Yamshchikov, I.P. (2024). LLMs simulate Big5 personality traits: Further evidence. *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, 83-87. https://aclanthology.org/2024.personalize-1.7/

田中葉月・飯田愛結・福田聡子・中島亮一・大澤正彦 (2024). 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 プロシーディングス*, P-60.

田中葉月・中島亮一・大澤正彦 (2025). 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. *HAI シンポジウム 2025 プロシーディングス*, P1-5.