2023/12/12 M1 中原 慧

Hang J, Xiajie Z, Xubo C, Cynthia B, Jad K(2023)

PersonaLLM: Investigating the Ability of Large Language Models to Express Big Five Personality Traits

1. イントロダクション

・LLM は人間のような会話をすることが出来ると期待されているため，様々な文脈で人間と相互作用し，人間をサポートする擬人化された AI エージェントの構築に関心が集まっている.

・Character AI や Replica のような新興企業は，それぞれのプラットフォーム上で，バーチャル・キャラクターを通じてユーザーの関心を集めることに成功している．一方で，学術的な研究(Park et al., 2023;Wang et al., 2023b)でも，ジェネレイティブ・エージェントは人間らしい行動を示し，社会科学研究のシミュレーションに利用できる可能性が示唆されている.

・ビッグファイブ性格モデル(Goldberg, 2013)の豊富な研究に基づき，LLM がビッグファイブ性格特性(外向性・協調性・誠実性・神経症傾向・開放性)を人間に近い形で採用する能力を検証することを目的とする.

・本論文では，LLM ベースのエージェントの初期プロンプト設定で性格特性が定義されているエージェントを「LLM ペルソナ」と定義する．LLM ペルソナを作成し，性格テストや個人的なストーリーを描くように促す．そして，LLM ペルソナの言語行動を分析し，以下の研究課題(RQ)を探求する.

RQ1：LLM は、Big Five Personality Inventory（BFI）評価を完了する際に、割り当てられたペルソナの行動を反映することができますか？

RQ2：LLM のペルソナが生み出すストーリーには，どのような言語的パターンが見られるか？

RQ3：人間と LLM は，LLM ペルソナが生成したストーリーをどのように評価するか？

RQ4：LLM のペルソナから生成された統計から，人間と LLM はビッグファイブ性格特性を正確に認識することができるか？

・性格と言語使用との間には強い相関関係があることが，研究によって一貫して示されている(Pennebaker and King, 1999; Pennebaker and Graybeal, 2001; Lee et al.)．Pennebaker ら(2001)は，人間の文章から特徴を要約する LIWC(Linguistic Inquiry and Word Count)というディクショナリーを導入し，ビッグファイブ性格特性との相関を実証した．ほとんどの先行研究では人間の言語使用に焦点を当てているが，本研究ではこの研究を LLM に拡張した.

2. 実験デザイン

異なる性格特性を持つ LLM ペルソナを作成する.

↓

彼らに性格評価を行う.

↓

これらの LLM ペルソナに物語を書かせる.

↓

広く採用されている LIWC フレームワークで彼らの作品を分析する.

↓

人間と LLM の両方を募り，これらの物語を 6 つの次元で評価する.

↓

最終的に人間と LLM の両方を用いて，物語の作者の性格特性を推測する.

・モデル設定

ChatGPT(GPT-3.5-turbo-0613)と GPT-4(GPT-4-0613)は，最先端のチャットベースの LLM であり，多ターンの対話に適しているため，この実験に使用した．ペルソナの行動に多様性を持たせるため，temperature を 0.7 に設定し，人間の個人間の自然な違いをエミュレートしている．その他のパラメータはすべての OpenAI のデフォルト設定のままである.

・LLM ペルソナ作成

ChatGPT と GPT-4 では，ビッグファイブの性格タイプの組み合わせごとに 10 人の LLM ペルソナを作成し，合計 320(2^5*10)人のペルソナを作成した．これらをそれぞれ，ChatGPT ペルソナ，GPT-4 ペルソナと呼ぶ.

システムプロンプトで LLM ペルソナを作成する方法は，“You are a character who is [TRAIT 1, ..., TRAIT 5].”というプロンプトを入力した．TRAIT はビッグファイブの各性格特性の
(1) extroverted / introverted,
(2) agreeable/antagonistic,
(3) conscientious / unconscientious
(4) neurotic / emotionally stable
(5) open / closed to experience
の組み合わせである．

・BFI 性格テスト

性格タイプを指定した後，LLM ペルソナに 44 項目のビッグファイブ目録(BFI)を行ってもらう．BFI を完了した後各スコアを受け取り，事前に設定された性格プロファイルと比較する．ここで，BFI を使用する理由は，(1)LIWC を対象とした多くの研究を含め，パーソナリティ関連の研究に広く利用されているため，先行研究との比較が可能であること．(2)16 タイプしかない MBTI や 16PF とは対照的に，32 タイプという，より詳細で包括的な評価が可能であること．

・ストーリー執筆

320 人の LLM ペルソナに，次のようなプロンプトで個人的なストーリーを書いてもらう．“Please share a personal story in 800 words. Do not explicitly mention your personality traits in the story.”

・評価方法

ChatGPT と GPT-4 のペルソナによって作成されたストーリーの LIWC 分析を行う．(1)ChatGPT ペルソナによって作成されたストーリーの大部分は，提供されたプロンプトに忠実に従っておらず，しばしば性格的特徴へのあからさまな言及が含まれていたためテキストベースの性格評価に適していない．
(2)ChatGPT ペルソナによって生成されたストーリーの質は，GPT-4 ペルソナによって生成されたものより明らかに劣っていた．

ストーリーの評価と性格予測の段階で，クラウドワーカーは，ストーリーが AI によって書かれたことを意識させられるか，意識させられないかの２つの条件のどちらかにランダムに振り分けられる．この方法論的デザインは，「AI が作者であることの認識」が物語の評価や性格予測の精度にどのような影響を与えるかを調査することを目的としている．

・LIWC 分析

LIWC-22 を使用して，ストーリーから心理学的特徴を抽出し，これらの特徴とパーソナリティ特性の相関関係を調べる．

・ストーリー評価

GPT-4 ペルソナによって生成されたストーリーを人間と LLM の両方に評価してもらう．予算上の制約から，各人格タイプから１０個づつあるストーリーのうち，性格特性の明示的な言及がないものに焦点を当てて，１個をサンプリングすることで，評価のための３２のストーリーを評価対象とする．ストーリーを評価する６つの側面は以下の通りである．

(1)読みやすさ：ストーリーが読みやすく，構成がしっかりしており，自然な流れになっているか.
(2)個人性：ストーリーが個人的なもので，書き手の考えや感情，生活が表れているか.
(3)冗長性：ストーリーが簡潔で，不要な内容がないか.
(4)まとまり：ストーリーの文章が上手くまとまっているか.
(5)好感度：読んでいて楽しいか，面白いか.
(6)信憑性：ストーリーに説得力があり，現実の状況に即したリアルなものであるか.
・性格予測
　３２のストーリーのコレクションについて，人間のアノテーターと LLM の評価者はそれぞれ，ストーリーから作者のビッグファイブ性格特性を１～５のスケールで予測するよう求められる．この目的は，人間と LLM の両方が，ストーリーだけから割り当てられた性格特性を正確に推測できるかどうかを判定することである.

3. 結果

・RQ１：BFI 評価における行動
　ChatGPT と GPT-4 で作成された３２０人の LLM ペルソナを用いて，彼らのパーソナリティ BFI スコアを計算し，５つのパーソナリティ尺度ごとにその分布を分析する．具体的には，パーソナリティ・スコアの平均値間の差異を評価する為に一元配置分散分析を適用する．その結果，**５つの性格特性すべてにおいて統計的に有意な差があることが明らかになった.** Cohen の d は以下の通り.
ChatGPT(EXT: 7.81; AGR: 5.93; CON: 1.56;NEU: 1.83; OPN: 2.90)
GPT-4(EXT: 5.47; AGR: 4.22; CON: 4.39; NEU: 5.17;OPN: 6.30)

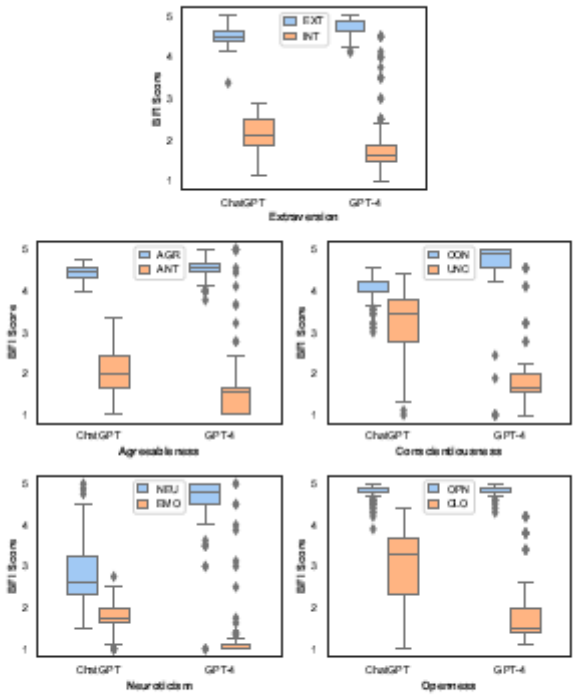

Figure 2: BFI assessment in five personality dimensions by ChatGPT and GPT-4 personas. Significant statistical differences are found across all dimensions.

全ての性格次元において，ポジティブな特性からネガティブな特性への BFI スコアの一貫した低下を観察した．全体として，この図は RQ1 に答えており，LLM ペルソナが BFI 評価を完了する際に，指定されたペルソナの行動を確かに模倣できることを確認している．

・RQ2：ライティングにおける言語的パターン

LLM ペルソナが執筆した個人的なストーリーから LIWC 特徴を抽出し，これらの心理学的特徴とパーソナリティ特性の間のスピアマン相関テストを実施する．パーソナリティ特性は 2 値としてエンコードし，これは人間が執筆したストーリーの Public Essays データセット(Pennebaker and King, 1999)で使用されているエンコーディングと同じである．LLM と人間との文章の比較分析は，LIWC メトリクスの中の心理学的なカテゴリーと語彙に焦点をあてて行う．

表は，性格特性と統計的に相関のある代表的な性格特性をまとめたものである．**各人格特性は，LLM ペルソナの異なる代表的な言語行動と関連していることがわかる**．例えば，開放性の特性が ChatGPT/GPT-4 ペルソナと人間の好奇心の語彙集の使用と正の相関があることを発見した．神経症傾向は，ChatGPT/GPT-4 ペルソナと人間の不安，ネガティブなトーン，およびメンタルヘルスに関連する語彙集と正の相関がある．外向性は，すべての著者におけるポジティブなトーンと所属感の語彙集と正の相関がある．

一方で，**ChatGPT/GPT-4 ペルソナは言語的特性において人間と異なることがある**．今回使用したエッセイ・データセットの人間の文章は，LLM の表現力を理解するための参考資料として使用されているこ

とに注意することが重要である．したがって，人間が書いたストーリーと LLM が生成したストーリーは同じプロンプトで生成されたものではないため，決定的な基準として扱うべきではない．

Trait	Selected LIWC Features	Lexicons	ChatGPT	GPT-4	Humans	ChatGPT#	GPT-4#
EXT	Positive Tone	good, well, new, love	+	+	+	**16/18**	10/18
	Affiliation	we, our, us, help	+	+	+		
	Certitude	really, actually, real	-	-			
	Social Behavior	said, love, care	+	+			
	Friends	friend	+		+		
AGR	Moralization	wrong, honor, judge	-	-	-	**16/23**	13/23
	Interpersonal Conflict	fight, attack	-	-	-		
	Affiliation	we, our, us, help	+	+	+		
	Negative Tone	bad, wrong, hate	-	-	-		
	Prosocial Behavior	care, help, thank	+	+			
CON	Drives	we, our, work, us		+	+	1/31	**11/31**
	Achievement	work, better, best	+	+			
	Lifestyle (Work, Money)	work, price, market		+	+		
	Moralization	wrong, honor, judge	-		-		
	Interpersonal Conflict	fight, attack			-		
	Time	when, now, then			+		
NEU	Anxiety	worry, fear, afraid	+	+	+	7/27	**15/27**
	Negative Tone	bad, wrong, hate	+	+	+		
	Mental Health	trauma, depressed	+	+	+		
	Sadness	sad, disappoint, cry	-	+	+		
	Anger	hate, mad, angry		+	+		
	Perception (Feeling)	feel, hard, cool		+	+		
OPN	Curiosity	research, wonder	+	+	+	2/36	**17/36**
	Insight	know, how, think		+	+		
	Affiliation	we, our, us, help		-	-		
	Perception (Visual)	see, look, eye		+	+		
	Future Focus	will, going to		-	-		

Table 1: Correlated metrics between LIWC features and binary personality traits with Point-biserial Correlation. The analysis is done on personal stories generated by ChatGPT and GPT-4 and the human Essays corpus (Pennebaker and King, 1999). This analysis focuses on the psychological and extended vocabulary metrics (81 in total). We report the representative personality LIWC features (+ means positive correlation, − means negative correlation) and the # of overlapped significant LIWC features for ChatGPT and GPT-4 with human writings.

・RQ3：ストーリーの評価

人間と LLM の評価者は，GPT-4 ペルソナによって生成されたストーリーを評価する．評価の主観的な性質のため，Lee(2023)と同様に，３人のアノテーター間の IAA(Inter-annotator Agreement)スコアが低い．したがって，高い一致を目指すのではなく，人間から多様な知覚を収集するために，ストーリーごとに 5 人の人間又は LLM 評価者を持つことにした，

GPT-4 ペルソナが作成したこれらのストーリーは，人間と LLM の両方の評価者から，読みやすさ・まとまり・信憑性の点で 4.0 に近いかそれ以上の高い評価を得ている．このことは，**ストーリーが言語的に流暢で，構造的にまとまっているだけでなく，説得力のあるものであることを示している**．その上，人間の評価者は個人性に高いスコアを割り当てており，これらの物語が実際に個人的な経験を描写していることを示している．

興味深いことに，これらの物語は人間の評価者から好感度の点で低いスコアを受けており，物語が信憑性があり個人的である一方で，楽しいものではないかもしれないことを示唆している．

当然のことながら，GPT-4 評価者はすべての次元で最も高い評価を与え，GPT-4 が作成したストーリ

ーを強く選好していることを示している．これは，LLM が，LLM の作成したコンテンツを好むという以前の知見を裏付けるものである(Liu et al. 2023).

人間の評価者の，読みやすさ・冗長性・まとまり・好感度・信頼性に対する評価は，コンテンツが AI によって生成されたものであることを認識しているかどうかに関わらず一貫しているようである．また，書き手が AI であることを知らされている場合，コンテンツの個人的な印象は著しく低下する．最後に，GPT-4 評価者が提供する評価は一貫して高く，書き手が AI であるかどうかを知らされている条件と知らされていない条件との間の変動は最小であり，GPT-4 コンテンツへの強く一貫した偏りを示している．

Evaluator	Readability	Redundancy	Cohesiveness	Likability	Believability	Personalness
Uninformed Condition – *Evaluation Scores (Mean$_{STD}$)*						
Human	$4.28_{0.85}$	$3.70_{1.17}$	$4.23_{0.88}$	$3.74_{1.00}$	$3.96_{1.02}$	$4.32_{0.85}$
ChatGPT	$4.75_{0.43}$	$3.04_{0.40}$	$4.97_{0.17}$	$4.22_{0.48}$	$3.93_{0.25}$	$3.55_{0.61}$
GPT-4	$4.94_{0.24}$	$4.96_{0.22}$	$5.00_{0.00}$	$4.84_{0.36}$	$4.93_{0.25}$	$5.00_{0.00}$
Informed Condition – *Evaluation Scores (Mean$_{STD}$)*						
Human	$4.38_{0.70}$	$3.62_{1.16}$	$4.12_{0.82}$	$3.80_{0.98}$	$3.97_{0.80}$	$3.99_{0.90}$
ChatGPT	$4.97_{0.17}$	$2.99_{0.35}$	$5.00_{0.00}$	$4.22_{0.41}$	$3.97_{0.17}$	$3.31_{0.77}$
GPT-4	$5.00_{0.00}$	$4.92_{0.33}$	$5.00_{0.00}$	$4.84_{0.36}$	$4.91_{0.28}$	$5.00_{0.00}$

Table 2: LLM and human evaluation results of GPT-4 generated personal stories **across six dimensions**. **Uninformed** and **informed** conditions indicate whether human or LLM evaluators are informed that the stories are generated by AI. For each evaluated attribute, we report its mean Likert scale and the standard deviation. Temperature is set to 0 for both GPT-3.5 and GPT-4.

・RQ4：パーソナリティ知覚

ストーリーから性格特性が予測可能かどうかを評価する為に，２つの分析を行った．第一に，各ペルソナの性格特性を二値分類問題として扱い，性格推論タスクにおける人間と LLM の精度を計算する．第二に，ペルソナの性格スコアを抽出し，人間の判断とペルソナの BFI スコアとの線形関係を分析する．

第一：性格予測

人間の評価者のパーソナリティ認識は，リッカート尺度を用いて収集され，その後，名目値に変換された．これらの値は，４と５をポジティブ・１と２をネガティブ・３をニュートラルに分類した．この二つの図から，**GPT-4 ペルソナが書いた個人的な物語に基づく性格特性の予測精度は，５つの次元で異なることがわかる**．人間の評価者が，AI が書いたかどうかを知らない場合，彼らは外向性で 0.68，協調性で 0.51 の精度を達成するが，他の Big Five Inventory（BFI）の次元ではランダム（0.50）よりも悪いパフォーマンスを示す．これは，このテキストベースの性格予測タスクが個々の人間の評価者にとって難しいことを示している．

各ストーリーの多数決に基づいて人間のアノテーターの投票を集計すると，外向性と協調性の精度は大幅に向上し，その他の性格特性も多数決によって向上する．これは，**性格特性がストーリーからグループレベルで人間の評価者に知覚可能(0.5 またはランダムな推測よりも良い)であることを示している**．

興味深いことに，**人間の評価者が，書き手が AI であることを知らされている場合，制度は様々な程度で低下することが分かった**．

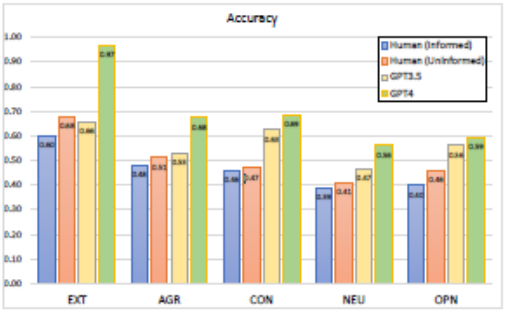

Figure 3: **Individual accuracy** of human and LLM evaluators in predicting personality.

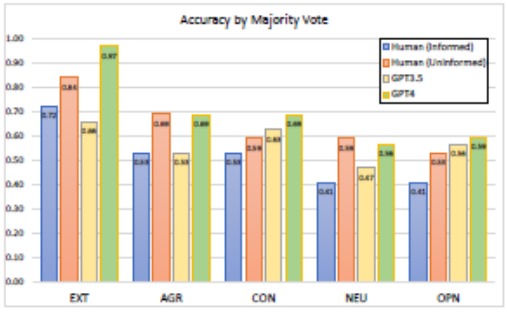

Figure 4: **Collective accuracy** of human and LLM evaluators in predicting personality with majority votes.

第二：BFI スコアとの相関

LLM ペルソナの BFI スコアと人間の認識との関係を掘り下げ，各特徴に関する人間の評価とペルソナの BFI スコアの間のスピアマンの r を計算した．その結果，LLM ペルソナの BFI スコアは，人間の認識と様々な程度で相関していることが明らかになった．

具体的には，人間の評価者が，書き手が AI であることを知らされていない場合，５つの全ての特性において有意な相関がみられた．(EXT: r = .64, p < .001; AGR: r = .33, p < .001;CON: r = .26, p < .001; NEU: r = .23, p < .005;OPN: r = .22, p < .005)逆に，書き手が AI であることを知らされていた場合は，４つの特性で相関が持続したが，開放性については有意ではなかった．(EXT: r = .42, p < .001;AGR: r = .32, p < .001; CON: r = .20, p < .05;NEU: r = .17, p < .05)人間の評価者が，書き手が AI であることを知らされた条件における BFI 相関の強さが減少したことは，以前我々が観察した，**書き手が AI であることの認識がパーソナリティの認識に影響を与えると**いうことを裏付けている．

4. 結論

この研究では，ChatGPT と GPT-4 が，十分に検証された性格尺度を用いて，一貫して性格特性を表現する能力を探った．

GPT-4 ペルソナは，ChatGPT ペルソナと比較して，BFI 評価スコアにおいてより大きな差異を示し，彼らのストーリーは，Essay コーパスを用いて，より顕著にパーソナリティを代表する特徴と一致していることが分かった．

書き手が AI であることが知らされると，人間の評価者はより個人的でないと感じたり，性格予測の精度に影響を与えたことは，人間の性格認識が，単なる言語的特徴を超えて，書き手の身元や背景など複雑な社会的推論や信念を含んでいることを示している．