心理測定テストに関する LLM のメタ知識の検証

山本有起 [1]　Arjav Singh[2]　Yin Jou Huang[1]　Chenhui Chu[1]　村脇有吾 [1]
[1] 京都大学　[2] インド工科大学マドラス校
{yamamoto,huang,chu,murawaki}@nlp.ist.i.kyoto-u.ac.jp
mm20b007@smail.iitm.ac.in

概要

大規模言語モデル (LLM) に対する心理測定テストにおいて、LLM に内在するメタ知識が測定結果に与える影響を検証するため、ビッグファイブ理論に基づく IPIP-NEO に関するケーススタディを行う。本論文では、LLM に対する IPIP-NEO テストの実施、LLM の訓練時における IPIP-NEO データセットの混入の検証、LLM の IPIP-NEO に関するメタ知識の理解度の検証を行う。検証した 3 種のモデルに関して、性格の指示に対する敏感さが明らかになり、データセットの混入も確認された。また、3 種のモデルのメタ知識への理解度の傾向が確認された。

1　はじめに

大規模言語モデル (Large Language Model; LLM) はさまざまな文脈で人間に似た振る舞いを示す。特に、特定の振る舞いや性格を模倣するように LLM にプロンプトで指示を与えることで、ユーザーが意図した特性を LLM に与えることができる。このような特性を利用して、LLM をエージェントとして扱うことで、被験者実験における人間の参加者を LLM で代替できる可能性に注目が集まっている [1, 2]。

LLM がユーザーが指示した性格特性を反映しているかを確認する簡便な手段として、LLM に人間向けの心理測定テストを受けさせるという手法が採用されてきた [3, 4, 5][1)]。LLM は、事前学習時に様々な性格特性を示す人々が書いたテキストを読むことで、多様な性格特性を反映させる能力を獲得することが期待される。しかし、事前学習に使用されるテキストには、心理測定テスト自体に関する議論が含まれている可能性が高く、そうしたテキストを通じて LLM が心理測定テストに関するメタ知識を獲得する可能性も高い。このメタ知識が LLM の心理測定テストの結果にどのように影響するかは不明である。

本論文では、LLM に内在する心理測定テストのメタ知識がテストの結果に与える影響を検証するため、IPIP-NEO [7, 8] に関するケーススタディを行う。IPIP-NEO とは、ビッグファイブ理論という人間の性格を 5 つの主要な特性で説明する心理学の理論に基づいた心理測定テストである。まず準備として、3 種類のモデルに対して IPIP-NEO テストを行い、指示に基づく性格特性の模倣の様子を観察する。次に、2 種類の検証を行う。第一に訓練時における IPIP-NEO データセットの混入の有無を 3 種類のモデルで検証する。第二に、IPIP-NEO とビッグファイブ理論の対応関係を LLM に答えさせ、両者に対する LLM の理解度を測る。

実験の結果、検証した 3 モデルすべてが性格に関する指示に敏感に反応すること、そして IPIP-NEO データセットを訓練時に読んだことが確認された。また、各モデルのビッグファイブ理論および IPIP-NEO に対する理解度の現状を把握することができた。今後は、IPIP-NEO とビッグファイブ理論に対する LLM の理解度を下げた状態を実現し、その状態でのテストの結果をもとの結果と比較することで、メタ知識の影響をより直接的に検証したい。

2　IPIP-NEO・ビッグファイブ理論

本論文では、心理測定テストのうちの 1 つである IPIP-NEO を扱う。IPIP-NEO の背景にあるビッグファイブ理論は、個人の性格を理解・比較するための心理学の標準的な枠組みであり、人間の性格を 5 つの主要な特性で説明する。具体的には Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness である。

IPIP-NEO テストの概要を図 1 に示す。IPIP-NEO テストには 300 個の質問項目が用意されており、被験者自身がその項目の内容にどの程度

1) LLM に対する心理測定テストの有効性を疑問視する報告もある [6]。

	Very Inaccurate	Moderately Inaccurate	Neither Accurate Nor Inaccurate	Moderately Accurate	Very Accurate
1. Worry about things.	1	2	3	4	5
2. Make friends easily.	1	2	3	4	5
3. Have a vivid imagination.	1	2	3	4	5
4. Trust others.	1	2	3	4	5
5. Complete tasks successfully.	1	2	3	4	5

特性	スコア (0~ 100)
Extraversion	52.0
Agreeableness	30.0
Conscientiousness	20.0
Neuroticism	25.0
openness	9.0

図 1　IPIP-NEO テストの概要。被験者の回答に基づいて、各性格特性のスコアを算出する。このスコアが高ければその特性が高いと評価されていることになる。

あてはまるかを 5 段階 (Very Inaccurate, Moderately Inaccurate, Neither Accurate Nor Inaccurate, Moderately Accurate, Very Accurate) で評価する。その自己評価に基づいて 5 つの性格特性のスコア (0–100) が算出される。各質問項目は 5 つの性格特性のうちの 1 つに関連を持ち、他の性格特性の評価には影響を及ぼさない。また、IPIP-NEO においては、ビッグファイブ理論の 5 つの主要特性をそれぞれ 6 つに細分化した計 30 個の下位特性のスコアも同様に算出される。例えば、主要特性の 1 つである Extraversion は Friendliness, Gregariousness, Assertiveness, Activity Level, Excitement-Seeking, Cheerfulness の 6 つの下位特性に細分化され、それぞれ評価される。

3　LLM に対する IPIP-NEO テスト

準備として、llama-3-8b, llama-3-70b [9], llm-jp-3-13b [10] の 3 つのモデルに対して性格に関する指示を入力し、IPIP-NEO による心理測定テストを行う。

IPIP-NEO の採点には年齢と性別が必要であり、各モデルに対して 40 歳の男性を模倣するよう指示する。性格を指示するプロンプトは、文献 [3] を参考に、各主要特性に関連する 12 個の形容詞を用いて 5 段階のレベルで性格を調整する。レベル 3 に比べ、レベル 4 は少し高い、レベル 5 はかなり高い主要特性を示す性格になるようプロンプトを調整する。また、レベル 2 は少し低い、レベル 1 はかなり低い主要特性を示す性格になるようプロンプトを調整する。主要特性と形容詞の関係やプロンプトの詳細は付録 A、 B に記す。

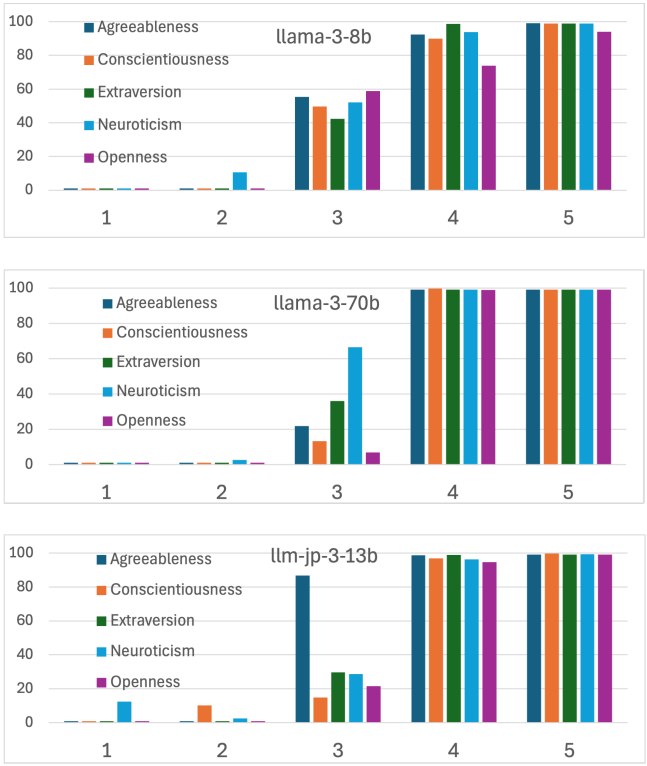

図 2　各レベルに対して各 LLM が示す主要特性のスコア。横軸がレベル、縦軸が各主要特性のスコアを表す。

各モデルに対する IPIP-NEO テストの実行結果を図 2 に示す。全体の傾向として、レベルが大きくなるほど IPIP-NEO テストによって算出された各性格特性のスコアが大きくなった。つまり、プロンプトに基づく性格調整が、少なくとも心理測定テストの結果に対してはある程度指示通りに反映されたと言える。

詳細を見ると、レベル 3 とレベル 2、4 での各特性のスコアには大幅な差が発生しているが、レベル 1、2、レベル 4、5 間での各特性のスコアの差はほとんどない。レベル 2 やレベル 4 は、各特性に関連する形容詞に “a bit” という副詞をつけることで少し偏った性格特性を示すよう意図したものであるが、算出されたスコアは大きな偏りを示した。つまり、各モデルはプロンプトによって与えられた性格に関する指示に大いに敏感に反応したと考えられる。また、他の 2 つのモデルに比べて、llama-3-8b はレベル 3 において各特性のスコアが 50 付近に固まっている。レベル 3 は各特性が平均的な高さである性格になるよう指定したものであり、平均的に振る舞うという指示に忠実にしたがえたのは llama-3-8b だけであるという結果になった。

表 1 Sharded Rank Comparison Test における片側 t 検定の p 値。3 種類とも p 値が十分低く、IPIP-NEO データセットの混入が裏付けられる。

モデル	p 値
llama-3-8b	0.00145
llama-3-70b	0.00113
llm-jp-3-13b	0.00565

4 訓練時における IPIP-NEO の混入の検証

次に、llama-3-8b, llama-3-70b, llm-jp-3-13b の 3 つの事前訓練モデルの訓練データに IPIP-NEO データセットが含まれるか検証する。混入の検証手法として Sharded Rank Comparison Test [11] を採用する。このテストは訓練データを直接参照しないブラックボックス手法[2)]であり、テキスト断片の順序に着目する。正規の順序のデータセットとランダムにシャッフルしたデータセットを用意し、LLM による生成確率を比較したとき、LLM が当該データセットを事前訓練時に読んだ、つまりデータ混入があった場合は、正規の順序を選好すると期待できる。このアイデアを具体化し、両者の平均生成確率が同じであるという帰無仮説が棄却できるかを片側 t 検定によって検定する。

IPIP-NEO は質問項目が特定の順序で並ぶことは確かだが、具体的にどのようフォーマットのテキストデータとして LLM に提示されたかは明らかではない。Sharded Rank Comparison Test は順序に着目することで、フォーマットの揺らぎに頑健に混入の検証を実現すると期待できる。なお、この手法では検証するデータセットをシャードという単位に分割するが、IPIP-NEO はサイズが小さいため、シャード数を 2 とした。

実験結果を表 1 に示す。検証した全てのモデルに関して算出された p 値は 0.05 を大きく下回る値であり、有意水準が 5%で帰無仮説は棄却される。つまり、検証したすべてのモデルが正規の順序のデータセットをシャッフルしたデータセットよりも有意に選好しており、訓練データに IPIP-NEO データセットが含まれていると結論づけられる。

2) llm-jp-3-13b については訓練データも公開されていることから、別途検証を進めたい。

5 メタ知識の理解度の検証

本節では、4 節で訓練時に IPIP-NEO データセットを読んだと結論付けた LLM の instruction tuning されたバージョンに対して、IPIP-NEO およびビッグファイブ理論への理解度を検証する。

5.1 検証方法

2 節で述べたように、IPIP-NEO の質問項目と 5 つの性格特性には 1 対 1 の対応がある。例えば、“Worry about things.” という質問項目は Neuroticism に対応する。そこで第 1 の検証方法では、全 300 個の質問項目に対応する性格特性を LLM に直接答えさせ、正答できた割合で IPIP-NEO・ビッグファイブ理論に対する理解度を測る (以下、性格特性検証とする)。このテストでは、LLM にビッグファイブ理論の 5 つの性格特性を選択肢として与えるプロンプトと与えないプロンプトの 2 種類を用意し、この 2 つの条件で LLM の理解度を測る。

次に、IPIP-NEO の質問項目に対応する性格特性には方向性の概念が存在する。例えば、“Make friends easily.” という質問項目に当てはまる場合は高い Extraversion を持つ傾向があり、この質問項目は+Extraversion に対応する。一方、“Keep others at a distance.” に当てはまる場合は Extraversion が低い傾向があり、この質問項目は-Extraversion に対応する。そこで第 2 の検証方法では、各質問項目に対応する性格特性とその方向性の両方を LLM に問い合わせ、両方正答できた割合で理解度を測る (以下、方向性検証とする)。この検証方法においても、5 つの性格特性を選択肢として与えるプロンプトと与えないプロンプトの両方を用いる。

最後に、IPIP-NEO の 5 つの主要な性格特性はそれぞれ 6 つの下位特性に細分化され、5 つの主要な性格特性と同様に、各質問項目は 1 つの下位特性と 1 対 1 対応する。そこで第 3 の検証方法では、各質問項目に対応する下位特性を LLM に問い合わせ、正答できた割合で理解度を測る (以下、下位特性検証とする)。前述した 2 つの検証方法とは異なり、下位特性を選択肢として与えない条件では検証を行わず、選択肢として与えた条件でのみ検証する。また、プロンプトにはその質問項目がどの主要特性に該当するかを提示し、その下位特性に当たる 6 つの特性を選択肢として与えることにする。

本節の検証では、3、4 節で扱ったモデルの他に、

表 2 性格特性検証の実験結果。大文字小文字の違いを除き、2 節で紹介した各性格特性を表す英単語を正確に出力した割合を記載している。

モデル	正答率 (選択肢あり)	正答率 (選択肢なし)
gpt-3.5-turbo	69.33%	57.33%
gpt-4o	78.00%	70.00%
llama-3-8b-instruct	56.33%	42.33%
llama-3-70b-instruct	74.67%	52.00%
llm-jp-3-13b-instruct	57.00%	16.00%

表 3 方向性検証の実験結果。特性の名称とその方向性の両方を正答した割合を記載している。

モデル	正答率 (選択肢あり)	正答率 (選択肢なし)
gpt-3.5-turbo	68.67%	55.67%
gpt-4o	78.00%	70.00%
llama-3-8b-instruct	52.00%	36.67%
llama-3-70b-instruct	73.67%	51.67%
llm-jp-3-13b-instruct	55.00%	15.33%

GPT-3.5 および GPT-4 [12] も扱う。これら 2 つのモデルは、定性的にはビッグファイブ理論や IPIP-NEO の概要を説明できる程度の理解度を持っており、その他の 3 つのモデルの理解度を評価する上でのベースラインとして扱う。

5.2 実験結果

性格特性検証の結果を表 2 に示す。いずれの結果も偶然一致率の 20%を大幅に上回った。GPT シリーズは選択肢ありの場合で 70–80%の正答率であった。一般的に高性能とされ、ビッグファイブ理論や IPIP-NEO に対する説明を行えるほどの理解度を持った両モデルが少なくとも 2 割程度間違う簡単にはいかないタスクであることがわかる。また特性の選択肢がない場合は 1 割程度正答率が低下する。

llama-3-70b は GPT シリーズに匹敵する正答率を選択肢の有無にかかわらず示した。ただし、選択肢の有無による正答率の差は 2 割程度と GPT シリーズより大きい。対照的な結果となったのは llama-3-8b と llm-jp-3-13b である。この 2 つのモデルの選択肢ありでの正答率はほとんど同じだが、選択肢なしの条件下では後者の正答率が大きく低下した。

次に、方向性検証の結果を表 3 に示す。性格特性検証の結果と比較すると、llama-3-8b 以外のモデルの正答率はほとんど低下していない。つまり、各質問項目に対応する性格特性が分かる場合はその方向性も分かるということになる。また、llama-3-8b も 5.7%程度の低下にとどまっている。LLM のビッグファイブ理論や IPIP-NEO への理解度を測る際、性格特性の方向性に関してはあまり考慮する必要はないと結論づけられる。

表 4 下位特性検証の実験結果。下位特性として 6 つの選択肢が与えられたときの正答率を記載している。

モデル	正答率 (選択肢あり)
gpt-3.5-turbo	71.67%
gpt-4o	82.00%
llama-3-8b-instruct	55.67%
llama-3-70b-instruct	73.67%
llm-jp-3-13b-instruct	55.67%

次に、下位特性検証の結果を表 4 に示す。性格特性検証 (選択肢あり) と下位特性検証の結果を比較すると、全てのモデルにおいてその正答率が同程度の値になっていることがわかる。こちらに関しても GPT-4 で正答率 8 割程度のタスクとなり、llama-3-70b は十分に下位特性に対する理解があると考えられる。

本節の実験では、llama-3-70b は GPT シリーズに匹敵する高い理解度を持っていることが明らかとなり、よりパラメータ数が小さく理解度が低かった llama-3-8b と対比して、今後の検証 (理解度を下げたときに IPIP-NEO テストの結果がどのように影響を受けるか等) を進めると興味深いだろう。また、llm-jp-3-13b はそれらよりさらに理解度の低いモデルとして今後の検証で扱う余地があるかもしれない。

6 結論

各 LLM に対する IPIP-NEO テストの実施、訓練時の IPIP-NEO の混入の有無の検証、IPIP-NEO に関するメタ知識の理解度の検証を行った。テストを実施した 3 種のモデルはすべて性格に関する指示に敏感に反応しテストの結果が変化した。また、検証した 3 種のモデルすべてが IPIP-NEO データセットを訓練時に読んだことを確認した。また、IPIP-NEO に関するメタ知識の理解度について、理解度が概ねモデルのサイズに比例することを確認した。

今後は、ビッグファイブ理論や IPIP-NEO に関するメタ知識への各モデルの理解度を下げた状態を実現したい。モデルを一から再訓練することはコスト的に現実的ではないことから、machine unlearning 等の手法の応用が考えられる。こうして得られたモデルに対して再度 IPIP-NEO テストを実施し、その結果を今回の IPIP-NEO テストの結果と比較することで、メタ知識が心理測定テストの結果に与える影響をより直接的に検証したい。

謝辞

本研究は一部、文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けた。

参考文献

[1] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? **Trends in Cognitive Sciences**, Vol. 27, No. 7, pp. 597–600, 2023.

[2] Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. Limited ability of LLMs to simulate human psychological behaviours: a psychometric analysis. arXiv:2405.07248, 2024.

[3] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. arXiv:2307.00184, 2023.

[4] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. arXiv:2305.02547, 2024.

[5] Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. On the reliability of psychological scales on large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 6152–6173, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[6] Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. Do personality tests generalize to large language models? In **Socially Responsible Language Modelling Research**, 2023.

[7] Goldbert Lewis R. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. **Personality Psychology in Europe**, Vol. 7, pp. 7–28, 1999.

[8] John A. Johnson. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. **Journal of Research in Personality**, Vol. 51, pp. 78–89, 2014.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and othres. The llama 3 herd of models. arXiv:2407.21783, 2024.

[10] LLM-jp. LLM-jp: A cross-organizational project for the research and development of fully open Japanese LLMs. arXiv:2407.03963, 2024.

[11] Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models. arXiv:2310.17623, 2023.

[12] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. GPT-4 technical report. arXiv:2303.08774, 2024.

A　IPIP-NEO における各特性の詳細

表 5　IPIP-NEO における主要特性と下位特性の対応およびプロンプトで使用した形容詞の一覧。各形容詞はそれぞれ 1 つの下位特性に関連するものであり、Low Marker と High Marker のどちらかに分類される。Low Marker と High Marker の使い分けに関しては図 3 に記載する。

主要特性	下位特性	Low Marker	High Marker
Extraversion	E1 - Friendliness	unfriendly	friendly
Extraversion	E2 - Gregariousness	introverted	extraverted
Extraversion	E3 - Assertiveness	timid	bold
Extraversion	E4 - Activity Level	inactive	active
Extraversion	E5 - Excitement-Seeking	unenergetic	energetic
Extraversion	E6 - Cheerfulness	gloomy	cheerful
Agreeableness	A1 - Trust	distrustful	trustful
Agreeableness	A2 - Morality	immoral	moral
Agreeableness	A3 - Altruism	unkind	kind
Agreeableness	A4 - Cooperation	uncooperative	cooperative
Agreeableness	A5 - Modesty	self-important	humble
Agreeableness	A6 - Sympathy	unsympathetic	sympathetic
Conscientiousness	C1 - Self-Efficacy	unsure	self-efficacious
Conscientiousness	C2 - Orderliness	messy	orderly
Conscientiousness	C3 - Dutifulness	irresponsible	responsible
Conscientiousness	C4 - Achievement-Striving	lazy	hardworking
Conscientiousness	C5 - Self-Discipline	undisciplined	self-disciplined
Conscientiousness	C6 - Cautiousness	impractical	practical
Neuroticism	N1 - Anxiety	relaxed	tense
Neuroticism	N2 - Anger	calm	angry
Neuroticism	N3 - Depression	happy	depressed
Neuroticism	N4 - Self-Consciousness	unselfconscious	self-conscious
Neuroticism	N5 - Immoderation	level-headed	impulsive
Neuroticism	N6 - Vulnerability	contented	discontented
Openness	O1 - Imagination	unimaginative	imaginative
Openness	O2 - Artistic Interests	uncreative	creative
Openness	O3 - Emotionality	unreflective	reflective
Openness	O4 - Adventurousness	uninquisitive	curious
Openness	O5 - Intellect	unintelligent	intelligent
Openness	O6 - Liberalism	socially conservative	socially progressive

B　IPIP-NEO テスト実施時のプロンプト

レベル	1	2	3	4	5
形容詞への修飾	Very {low marker}	a bit {low marker}	neither {lower marker} nor {high marker}	a bit {high marker}	very {high marker}

プロンプト例（Extraversion　レベル4の場合）
For the following task, respond in a way that matches this description: “You are a 40-year-old man. You are a bit friendly, a bit extraverted, a bit bold, a bit active, a bit energetic, a bit cheerful.”

図 3　IPIP-NEO テスト実施時の LLM に対するプロンプトの詳細と例。プロンプト例に示すように、各主要特性に関連する 6 つの形容詞を用いて LLM に対する性格の指示を構成する。特性の強さのレベルの調整は形容詞への修飾を図のように行うことで実現した。