Psychological Investigation of Personality Knowledge in a Large Language Model

Zhao Zicheng[1,2] Iwai Ritsuko [2] Asai Nichika[1] Kumada Takatsune[1,2]
[1]Graduate School of Informatics, Kyoto University [2]GRP, R-IH, RIKEN
{zhao.zicheng.55d,asai.nichika.37h}@st.kyoto-u.ac.jp
ritsuko.iwai@riken.jp, kumada.takatsune.7w@kyoto-u.ac.jp

Abstract

The purpose of this study is to investigate the knowledge of personality in a LLM using a psychometric methodology. In Experiment 1, a standard psychological questionnaire is used to measure the personality profile of the LLM and showed that the model has some knowledge about personality. Experiment 2 examined the scores of Big Five personality questions in the LLM when a wide range of personality descriptions were submitted as prompts to the model. The results showed that the LLM has similar personality knowledge as humans. Implications for LLM research and psychological research are discussed.

1 Introduction

In human psychological research, personality refers to the relatively stable and enduring set of emotional, cognitive, and behavioral patterns exhibited by individuals[1]. These patterns not only shape how people perceive and respond to various situations but also manifest in interpersonal interactions, decision-making, and emotional regulation. Consequently, understanding personality is crucial for predicting and explaining human behavior, making it a central focus in psychology and related fields.

Among personality theories, the Big Five Personality Theory is widely recognized and applied to depict human personality structure. This framework describes personality into five traits: Openness to experience (OPE), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU). As a universal framework for understanding human personality, the Big Five is frequently measured using standardized inventories (e.g., the BFI-2：Big Five Inventory-2[2] and IPIP-120：International Personality Item Pool[3]), enabling a comprehensive description and comparison of individual personality profiles.

In recent years, LLMs have demonstrated outstanding capabilities in natural language generation and human-agent conversational interactions. Furthermore, LLMs are tuned to a specific personality by a prompt for giving perception of personality for users in a conversation system. However, in order to give users consistent perception of personality, LLMs should have similar knowledge as humans. The lexical approach to human personality is based on the notion that personality words in natural language and used to describe individual differences in personalities in daily social context. LLMs learn from a large collection of usage of natural language in the context including personality words and the contexts where such personalities are described. Thus, it is plausible to consider that LLMs have knowledge of personality similar to that of humans. However, to our best knowledge, there is no study examining the similarity of personality knowledge of LLMs to human by a convincing methodology.

The purpose of this study is to investigate the knowledge about personality in a LLM using a psychometric methodology. More specifically, we examine to what extent the LLM has knowledge about human personality in the same way as humans. For this purpose, we submit prompts to a LLM model for asking to be a person having predisposition related to a Big Five trait and to answer each item in a big five questionnaire. After giving a set of prompts that cover all the Big Five traits, all responses are analyzed using an exploratory factor analysis, which is a standard method for examining the latent structure of the responses. If the LLM learns knowledge of human personality, five dimensions corresponding to the big five will be identified.

2 Relate work

Lexical approach to personality The lexical approach works for defining personality because it assumes that the most important and widely recognized traits are naturally reflected in language. Allport and Odbert[4] collected thousands of descriptive words from dictionaries to systematically map personality traits through language. Later, researchers refined these words into measurable dimensions, leading to the Big Five Personality traits[5]. Representative adjectives to describe Big Five Personality traits are as follows[6]: OPE: imaginative, curious, artistic; CON: organized, responsible, hardworking; EXT: energetic, talkative; AGR: kind, cooperative, and trusting; NEU: anxious, self-conscious, and vulnerable. In this study, we used the Big Five personality framework to understand the personality of LLMs.

Personality in LLMs Studies on LLMs have shown their ability to mimic human behaviors across domains, including cognitive tests and social simulations[7][8] [9][10]. Recent work on personality in LLMs has introduced methods for evaluating the personality traits[11][12]. However, these studies primarily focus on evaluating the trait scores of the models. In contrast, little research has delved into whether LLMs possess personality knowledge structure similar to that of humans. In addition, there is no systematic study employing a reliable psychometric methodology to verify whether the personality knowledge of LLMs aligns with that of humans.

3 Experiment

3.1 Model

Mistral 7B Mistral 7B is an open-source LLM released by the French startup Mistral AI in September 2023[13], with 7.3 billion parameters. In this study, its primary role is to serve as the "subject."All experiments were conducted on an NVIDIA RTX 4090 GPU.

3.2 Experiment 1

The purpose of Experiment 1 is to evaluate the personality traits of the Mistral model using the BFI-2 as a baseline, without giving prompts to modify personality of the model.

Procedure We used the BFI-2[2] as a personality test. The BFI-2 consists of 60 items, with 12 items per each factor. In BFI-2, each item presents a descriptive statement (e.g., Q1 is “I am outgoing and sociable”) accompanied by five response options ranging from “Very Accurate” to “Very Inaccurate”. The model is asked to select the option it thinks the most appropriate based on its own understanding. Each chosen response is converted into a numerical score: for positive items (e.g., Is outgoing, sociable), “Very Accurate” corresponds to 5 points and “Very Inaccurate” to 1 point; for negative items (e.g., "Tends to be disorganized"), the scoring is reversed (e.g., “Very Accurate” corresponds to 1 point and “Very Inaccurate” to 5 points). After collecting all responses, we summed the scores of all items that belong to the same trait and then calculated the mean to obtain the score for that trait.

Table 1 Mistral 7B's numerical values of personalities

	OPE	CON	EXT	AGR	NEU
Neutral	3.79	4.21	2.70	4.67	2.46

Results and Discussion The scores of BFI-2 are shown in Table 1.The model shows slightly lower score below average scores for EXT and NEU and relatively high scores in OPE, AGR and CON. Because the model can choose one value in the range of 1-5, the average is 3. This shows that the model chooses a value for each question based on the knowledge about the meaning of questions, suggesting that the model has some knowledge about personality.

3.3 Experiment 2

The purpose of Experiment 2 is to examine the knowledge structure of personality in LLMs using a psychometric methodology.

Table 2 An Example of a Prompt

Template
You have the personality with:
e.g., "Making friends easily" from IPIP-120's EXT
Please evaluate this statement: I am
e.g., "Is compassionate, has a soft heart." from BFI-2's AGR
Please rate how accurately this describes you on a scale from 1 to 5.
Options:
(5). Very Accurate
(4). Moderately Accurate
(3). Neither Accurate Nor Inaccurate
(2). Moderately Inaccurate
(1). Very Inaccurate
I would rate this statement as:

Procedure In this experiment, IPIP-120, a Big Five personality traits with 24 items per trait for a total of 120 items[3], is used as a set of "personality prompts" to guide the LLMs in adopting specific personality traits. Because IPIP-120 covers more ground and has richer items used to create diverse prompts, aiming to create the framework in experiment 2. We use a new approach that involves creating a prompt (Table 2): We select one item from IPIP-120 as a personality prompt (e.g., “Making friends easily” from the EXT group). We fixed this as a personality prompt, and then test all 60 questions from BFI-2 (e.g., Q2: “Is compassionate, has a soft heart [AGR].” , Q3: “Tends to be disorganized [CON].”). This process creates a 1 × 60 matrix. Then, we use the next item from IPIP-120 as a fixed personality prompt and repeat the same steps. After doing this, we end up with a 120 × 60 matrix.

Results and Discussion We conducted an exploratory factor analysis using the principal factor solution with a promax rotation on this 120 × 60 matrix to investigate whether the responses of the LLM reflect the five-factor structure(Table 3). Items with high factor loadings were mostly aligned with the corresponding personality traits. For instance, 8 of 12 CON items in BFI-2 questions loaded on F1. In Table 4, we list some examples of items that are loaded to an expected factor and unexpected factor, respectively. Expected items are most strongly related to the factor, with the highest loadings, and they usually match the theoretical meaning of the factor. For example, in F4 OPE, items like Q20 OPE ("Fascinated by art, music, or literature"), Q35 OPE ("Values art and beauty"), and 60 OPE ("Original, comes up with new ideas") clearly reflect traits of OPE ("imaginative, curious, artistic") .This indicate that the high-loading expected items largely align with the corresponding theoretical personality traits, indicating that the model's internal knowledge structure supports its understanding of the Big Five framework.

These results also suggest that when the LLM processes ambiguous statements, it may associate the given trait with other personality traits in the similar way as humans do. For example, when processing “Shows a lot of enthusiasm” (EXT), the model might infer that the person also has traits like being friendly or helpful (AGR). This associative reasoning could lead to the expected factor structure.

Unexpected items mean that items not belonging to an expected factor also show high loadings. For instance, in

Table 3 Factor Loading Scores

		Factor				
ID	**Trait**	**F1**	**F2**	**F3**	**F4**	**F5**
Q18	CON	0.82				
Q38	CON	0.79				
Q33	CON	0.78				
Q13	CON	0.75				
Q43	CON	0.72				
Q58	CON	0.64				
Q28	CON	0.63				
Q53	CON	0.54				
Q15	OPE	0.39				
Q30	OPE	-0.41				
Q5	OPE	-0.48			-0.41	
Q44	NEU	-0.61		0.48		
Q12	AGR		0.86			
Q47	AGR		0.77			
Q2	AGR		0.69			
Q32	AGR		0.67			
Q17	AGR		0.66			
Q37	AGR		0.60			
Q42	AGR		0.53			
Q7	AGR		0.52			
Q22	AGR		0.51			
Q48	CON	0.38	0.46		-0.45	
Q57	AGR		0.46			
Q52	AGR		0.45		0.38	
Q27	AGR		0.44			
Q29	NEU	-0.36		0.80		
Q19	NEU			0.74		
Q4	NEU			0.69		
Q59	NEU			0.58		
Q34	NEU	0.40		0.53	0.48	
Q14	NEU			0.50	0.40	
Q24	NEU	-0.46		0.50		
Q9	NEU			0.50		
Q49	NEU			0.38		
Q56	EXT			-0.37		
Q1	EXT			-0.37		
Q21	EXT			-0.44		
Q6	EXT			-0.48		
Q41	EXT			-0.51		
Q60	OPE				0.59	
Q31	EXT				0.52	
Q20	OPE				0.52	
Q10	OPE				0.50	
Q35	OPE				0.50	
Q45	OPE				-0.49	
Q50	OPE	-0.45			-0.53	
Q3	CON	0.45			-0.62	
Q51	EXT					0.68
Q11	EXT					0.52
Q16	EXT					0.47
Q8	CON				-0.39	0.45
Q39	NEU	0.36				-0.36
Q54	NEU					-0.51

the F1 CON factor, Q44 NEU and Q15 OPE showed high loading to the factor. A possible reason for this is **semantic ambiguity or overlap**. This means some sentences, while overall belonging to one trait, contain keywords that connect to other traits. For instance,Q56 "Shows a lot of enthusiasm expresses (EXT)", but the word "enthusiasm" on its own might also relate to AGR (friendly, cooperative, and trusting). Therefore, when the model processes the "Shows a lot of enthusiasm", it may focus more on the word "enthusiasm" rather than the meaning of the whole sentence. One reason for this may be that Mistral 7B, which has fewer parameters, cannot fully capture such subtle semantic nuances. As a result, it relies more on keywords instead of understanding the entire sentence's meaning.

Table 4 Examples of expected and unexpected personality traits for each factor

Factor	Expected	Unexpected
F1 CON	**Q13 CON** Is dependable,steady. **Q18 CON** Is systematic, likes to keep things in order. **Q33 CON** Keeps things neat and tidy.	**Q44 NEU** Keeps their emotions under control. **Q24 NEU** Feels secure, comfortable with self.
F2 AGR	**Q2 AGR** Is compassionate, has a soft heart. **Q7 AGR** Is respectful, treats others with respect. **Q12 AGR** Tends to find fault with others.	**Q48 CON** Leaves a mess, doesn't clean up. **Q56 EXT** Shows a lot of enthusiasm.
F3 NEU	**Q4 NEU** Is relaxed, handles stress well. **Q19 NEU** Can be tense. **Q29 NEU** Is emotionally stable, not easily upset.	**Q41 EXT** Is full of energy.
F4 OPE	**Q20 OPE** Is fascinated by art, music, or literature. **Q35 OPE** Values art and beauty. **Q60 OPE** Is original, comes up with new ideas.	**Q34 NEU** Worries a lot. **Q42 AGR** Is suspicious of others' intentions.
F5 EXT	**Q11 EXT** Rarely feels excited or eager. **Q51 EXT** Prefers to have others take charge. **Q16 EXT** Tends to be quiet.	**Q54 NEU** Tends to feel depressed, blue. **Q8 CON** Tends to be lazy.

3.4 General Discussion

In this study, we showed that the LLM (Mistral 7b) has the similar personality knowledge as humans. This suggests that the LLM has developed interrelated representation within a personality trait and across personality trait, through their training on large-scale textual data. However, knowledge of human personality could not be completely reproduced by the LLM. One reason for this may be that learning of the LLM was insufficient. Even if LLMs were able to learn all the data, it remains controversial whether it would be able to have exactly the same personality knowledge structure as humans.

On the other hand, the results of unexpected items suggest that they may not accurately measure human personality traits. It is almost impossible to assume that each factor is completely independent of or orthogonal to other factors in psychological constructs, because such constructs of our minds are complex and inter-related. In that sense, selecting items that load on mainly one single factor is very important to develop psychological questionnaires and measure psychological constructs including personality. Given the results that most of trait items loaded on the expected traits, it suggests that such items can be appropriately responded simply based on language knowledge. Responding the unexpected items, however, imply that it requires knowledge not simply through language, rather though experiences in the real world. Such subjective experiences may be different among individuals so that the items can be the causes of more cross loadings. In other words, LLMs have learned knowledge independent of individual experiences. This may give us an opportunity to refine the process of developing psychological questionnaires by comparing between humans' results and LLMs' ones in the future. This proposes a novel applicability of LLMs to psychology.

4 Conclusion

This study revealed that the knowledge structure of personality is similar to that of human. By combining diverse personality prompts with an exploratory factor analysis, we uncovered the latent knowledge structure of personality in a LLM. We provide a new approach for assessing the ability to implement personality in LLMs. We also propose a possibility that LLMs can contribute psychological studies of personality.

References

[1] Ian J. Deary Gerald Matthews and Martha C. Whiteman. **Personality traits**. Cambridge University Press, 2003.

[2] Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. **Journal of personality and social psychology**, Vol. 113, No. 1, p. 117, 2017.

[3] 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.

[4] Gordon W Allport and Henry S Odbert. Trait-names: A psycho-lexical study. **Psychological monographs**,

Vol. 47, No. 1, p. i, 1936.

[5] Oliver P John, Alois Angleitner, and Fritz Ostendorf. The lexical approach to personality: A historical review of trait taxonomic research. **European journal of Personality**, Vol. 2, No. 3, pp. 171–203, 1988.

[6] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. **Journal of personality**, Vol. 60, No. 2, pp. 175–215, 1992.

[7] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. **arXiv preprint arXiv:2207.07051**, Vol. 2, No. 3, 2022.

[8] Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. **Proceedings of the National Academy of Sciences**, Vol. 120, No. 6, p. e2218523120, 2023.

[9] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In **Proceedings of the 36th annual acm symposium on user interface software and technology**, pp. 1–22, 2023.

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

[11] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[12] Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. **arXiv preprint arXiv:2307.16180**, 2023.

[13] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. **arXiv preprint arXiv:2310.06825**, 2023.

A Appendix

The details of the BFI-2 used in this study is as follows

Table 5 BFI-2 Detailed Content

ID	Math	Trait	Statement
1	+	EXT	Is outgoing, sociable.
2	+	AGR	Is compassionate, has a soft heart.
3	-	CON	Tends to be disorganized.
4	-	NEU	Is relaxed, handles stress well.
5	-	OPE	Has few artistic interests.
6	+	EXT	Has an assertive personality.
7	+	AGR	Is respectful, treats others with respect.
8	-	CON	Tends to be lazy.
9	-	NEU	Stays optimistic after experiencing a setback.
10	+	OPE	Is curious about many different things.
11	-	EXT	Rarely feels excited or eager.
12	-	AGR	Tends to find fault with others.
13	+	CON	Is dependable, steady.
14	+	NEU	Is moody, has up and down mood swings.
15	+	OPE	Is inventive, finds clever ways to do things.
16	-	EXT	Tends to be quiet.
17	-	AGR	Feels little sympathy for others.
18	+	CON	Is systematic, likes to keep things in order.
19	+	NEU	Can be tense.
20	+	OPE	Is fascinated by art, music, or literature.
21	+	EXT	Is dominant, acts as a leader.
22	-	AGR	Starts arguments with others.
23	-	CON	Has difficulty getting started on tasks.
24	-	NEU	Feels secure, comfortable with self.
25	-	OPE	Avoids intellectual, philosophical discussions.
26	-	EXT	Is less active than other people.
27	+	AGR	Has a forgiving nature.
28	-	CON	Can be somewhat careless.
29	-	NEU	Is emotionally stable, not easily upset.
30	-	OPE	Has little creativity.
31	-	EXT	Is sometimes shy, introverted.
32	+	AGR	Is helpful and unselfish with others.
33	+	CON	Keeps things neat and tidy.
34	+	NEU	Worries a lot.
35	+	OPE	Values art and beauty.
36	-	EXT	Finds it hard to influence people.
37	-	AGR	Is sometimes rude to others.
38	+	CON	Is efficient, gets things done.
39	+	NEU	Often feels sad.
40	+	OPE	Is complex, a deep thinker.
41	+	EXT	Is full of energy.
42	-	AGR	Is suspicious of others’ intentions.
43	+	CON	Is reliable, can always be counted on.
44	-	NEU	Keeps their emotions under control.
45	-	OPE	Has difficulty imagining things.
46	+	EXT	Is talkative.
47	-	AGR	Can be cold and uncaring.
48	-	CON	Leaves a mess, doesn’ t clean up.
49	-	NEU	Rarely feels anxious or afraid.
50	-	OPE	Thinks poetry and plays are boring.
51	-	EXT	Prefers to have others take charge.
52	+	AGR	Is polite, courteous to others.
53	+	CON	Is persistent, works until the task is finished.
54	+	NEU	Tends to feel depressed, blue.
55	-	OPE	Has little interest in abstract ideas.
56	+	EXT	Shows a lot of enthusiasm.
57	+	AGR	Assumes the best about people.
58	-	CON	Sometimes behaves irresponsibly.
59	+	NEU	Is temperamental, gets emotional easily.
60	+	OPE	Is original, comes up with new ideas.