Humanity in AI: Detecting the Personality of Large Language Models

Anonymous ACL submission

Abstract

Exploring the personality of large language models (LLMs) is an important way to gain an in-depth understanding of LLMs. It is well known that ChatGPT has reached a level of linguistic proficiency comparable to that of a 9-year-old child, prompting a closer examination of its personality. In this paper, we propose to detect the personality of LLMs by questionnaires and text mining methods, with the guide of BigFive psychological model. To explore the origins of the LLMs personality, we conduct experiments on pre-trained language models (PLMs, such as BERT and GPT) and Chat models (ChatLLMs, such as ChatGPT). The results show that LLMs do contain certain personalities, for example, we think ChatGPT tends to exhibit openness, conscientiousness and neuroticism, while ChatGLM only exhibited conscientiousness and neuroticism. More importantly, we find that the personality of LLMs comes from their pre-training data, and the instruction data can facilitate the generation of data containing personality. We also compare the results of LLMs with the human average personality score, and find that the humanity of FLAN-T5 in PLMs and ChatGPT in ChatLLMs is more similar to that of a human, with score differences of 0.34 and 0.22, respectively.

1 Introduction

Humanity is the major difference between artificial intelligence and human intelligence. Since the release of ChatGPT, the gap in capabilities between humans and AI has been gradually narrowing. LLMs can achieve levels close to or beyond humans in many areas, and have completely substituted humans in some scenarios. For instance, they serve as human assistants that can understand and respond to human language more naturally, help customer service agents respond to client queries promptly and accurately, and offer more personalized experiences (Jeon and Lee, 2023; Liu et al., 2023; Dillion et al., 2023). Unlike traditional deep learning models, LLMs achieve remarkable performance in semantic understanding and following instructions (Lund et al., 2023; Liu et al., 2023), which is the answer why LLMs behave more like humans.

The research from Standford suggested that ChatGPT has reachede the level of a human 9-year-old child (Kosinski, 2023). Recent research from Microsoft suggests that OpenAI's latest large language model, GPT-4, possesses fundamental human-like capabilities, including reasoning, planning, problem-solving, abstract thinking, understanding complex ideas, rapid learning, and experiential learning (Bubeck et al., 2023). Experts from Johns Hopkins University have found that the theory of the mind of GPT-4 has surpassed human abilities, achieving 100% accuracy in some tests through a process of mental chain reasoning and step-by-step thinking (Moghaddam and Honey, 2023). It seems that LLMs is already an complete human being. But, when we converse with LLMs, we can still determine that it is not human from its fixed-format response templates and polite but emotionless textual expressions. We think that this is related to the personality within LLMs, which is the major difference between LLMs and humans.

In human society, personality serves as a key indicator to differentiate individuals and characterize their behavior and responses in various situations. Humans have been studying personality and have developed standardized systems to assess individual traits, such as the Big Five model (Costa and McCrae, 1992), which categorizes personality into openness, conscientiousness, extraversion, agreeableness, and neuroticism. Other widely-used psychological models include MBTI (Jessup, 2002), 16PF (Cattell and Mead, 2008), and EPQ (Birley et al., 2006). Early research in psychology established standard evaluation methods, such as questionnaires and analysis of subjects' daily textual output (text mining).

Questionnaire is the most commonly used method for personal character assessment, such as MBTI, Big Five, and 16PF, as mentioned earlier. Questionnaire generally fall into two categories (Boyd and Pennebaker, 2017). The first involves providing a series of statements and asks participants to indicate the extent to which each statement applies to themselves, such as "You act as a leader" and then choosing a response from a five-point scale ranging from "Very Accurate" to "Very Inaccurate." The second involves presenting several scenarios and asking participants to choose the most appropriate response, such as "When faced with a difficult problem, would you A) approach it optimistically and proactively, B) avoid it, or C) think about it repeatedly." This method is relatively straightforward, and participants can hide their true personality by randomly choosing answers. Another method involves mining comments, diaries, and other texts posted by participants in their daily lives and analyzing the features of these texts, such as word choice, expression, and punctuation usage, to draw conclusions. This type of method is also commonly used in social media, it can avoid participant masking, but suffer from feature extraction difficulties.

In this paper, we use both methods to detect the personality of LLMs, with the guide of BigFive psychological model (Vanwoerden et al., 2023; Lin et al., 2023). Our main contributions include:

- We propose the combining of questionnaires and text mining to detect the personality of large models, which can obtain more accurate results.
- We identify the personality types included in the large model without any prompting by using questionnaires and text mining, and find that the humanity of FLAN-T5 in PLMs and ChatGPT in ChatLLMs is more similar to that of a human.
- Experiments indicated that the personality knowledge of the large model comes from its pre-trained data, and the instruction data can make LLMs more inclined to show a certain personality. [1]

[1]We will release all experimental data and intermediate results.

2 Related Work

In this paper, we explore the psychological traits of large models. So we will introduce some research work on psychological and some of the key research from PLMs to ChatLLMs.

2.1 Personality Traits

The most widely and frequently used personality models are the bigfive model (Costa and McCrae, 1992) and the MBTI model (Jessup, 2002). At the beginning of psychological research, questionnaires (Vanwoerden et al., 2023) and self-report (Lin et al., 2023) methods were the main research tools used to determine and examine an individual's personality. This method focuses on providing the participant with a number of descriptive states to answer according to his or her personality, one of the more famous ones being IPIP [2] (International Personality Item Pool) (Goldberg et al., 2006). Then the personality of the participant can be calculated by their answers (Hayes and Joseph, 2003). But those methods gradually abandoned by computer science scholars due to their low efficiency and ecological validity. Then computer scholars are beginning to use lexicon-based methods, machine learning-based methods, and neural network-based methods to mine personality traits from text, which increases efficiency by eliminating the need to collect questionnaires. The lexicon-based methods include LIWC (Pennebaker et al., 2001), NRC (Mohammad and Turney, 2013), Mairesse (Mairesse et al., 2007) and so on, those lexicon can be used to extract the psychological information contained in the text. However, due to the different systems and classification criteria used by different researchers, the mixing of multiple dictionaries may introduce errors. In addition, the method has limited ability to extract features in long texts. Machine learning-based methods include SVM, Naïve Bayes and XGBoost Nisha et al. (2022). Neural network-based methods include using CNN (Majumder et al., 2017), RNN (Sun et al., 2018), RCNN (Xue et al., 2018), pre-trained models (Wiechmann et al., 2022) . Those methods achieved higher accuracy than machine learning-based methods.

[2]https://ipip.ori.org/

2.2 Large Language Models

LLMS has a significant impact on the AI community, with the emergence of Chatgpt[3] and GPT-4 [4]leading to a rethinking of the possibilities of Artificial General Intelligence (AGI). The base model of ChatGPT is GPT3 (Brown et al., 2020), which is a pre-trained model that conclude 175B parameters. GPT-3 can generate human-like text and complete tasks such as language translation, question answering, and text summarization with impressive accuracy and fluency. Models similar to GPT3 include LLaMA (Touvron et al., 2023), BLOOM (Scao et al., 2022) and T5 (Raffel et al., 2020). Although the OpenAI team did not release the technical details of ChatGPT, from the content of InstructGPT (Ouyang et al., 2022), it can be guessed that the process of training with instruction data is very important. Then, the research team at Stanford University obtained Alpaca [5] by train LLaMA with the instruct dataset generated by ChatGPT. They also released this dataset Alpaca-52k. Then, more and more large models of the ChatLLMs were released, such as ChatGLM based on GLM (Zeng et al., 2022; Du et al., 2022), BLOOMZ and Vicuna. Although these models are slightly weaker in capability than ChatGPT, they have fewer parameters and consume fewer resources.

Following the release of these models, it is now well established for individual researchers to train a ChatLLM from a base PLM. This also opens up the possibility of exploring the knowledge contained within the large model. Also with the current ChatLLMs being so human-like in their performance, we believe that psychological measures of humans can be used to test the personality of the large model.

2.3 Personality in LLMs

There have been several research works focusing on the personality of LLMs, with all of them employing the Big Five model as the psychological framework. Ganesan et al. (2023)investigate the zero-shot ability of GPT-3 to estimate the Big 5 personality traits from users' social media posts. Jiang et al. (2022) detect the personality in LLMs using questionnaire method and propose an induce prompt to induce LLMs with a specific personality in a controllable manner. However, Song et al. (2023) argued that self-assessment tests are not suitable for measuring personality in LLMs and advocated for the development of dedicated tools for machine personality measurement.

[3]https://openai.com/blog/chatgpt-plugins
[4]https://openai.com/research/gpt-4
[5]https://crfm.stanford.edu/2023/03/13/alpaca.html

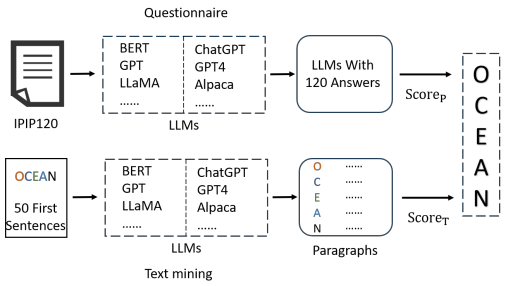


Figure 1: The process of two methods. Where $Score_P$ is defined by formula 1 and $Score_T$ is defined by formula 2

As we can see, the bigfive model and the questionnaire method are more common methods used for big model personality detection. But, the current method is more controversial. In order to solve this problem, we to use both questionnaire and text mining method. We think that combine those two methods can get more objective results.

3 Method

As we mentioned above, we used questionnaires and text mining to detect the personality of LLMs. The process of the two methods is shown in Figure 2.

As we can see, in the questionnaire method, we used the MPI120 questions to replace [Statement], then, ask each LLM to give an answer form (A) to (E). The model's score on each question is calculated based on IPIP's scoring criteria. It is designed following the IPIP study, we used the mean score to calculate the model's performance on each psychological traits, and the standard deviation to assess the model's responses. The formula for calculating the "score" is as follows:

$$score_P = \frac{1}{N_P} \sum_{i \in P}^{i} \{f(answer_i, statement_i)\} \quad (1)$$

where P represents one of the five personality traits, N_P represents the total number of statements for trait P, and $f(answer_i, statement_i)$ is a function used to calculate the personality score, which ranges from 1 to 5. Additionally, if a statement is positively correlated with trait P, answer choice A will receive a score of 5, whereas if it is negatively correlated, it will receive a score of 1.

In the text mining method, we provide the model with the first sentence of a paragraph and allow it to continue writing. Then, we use a specially designed prompt to enable ChatGPT to determine the personality traits contained in the model's continued text. The prompt that we input into ChatGPT is as follows: "[Sentence1] The Big Five characteristics of the passage above are . Please determine the Big Five characteristics of the following passage. Please only answer using words from the list ['Openness', 'Conscientiousness', 'Extraversion', 'Agreeableness', 'Neuroticism']. [Sentence2]. Remember that only one trait is highly demonstrated in the passage, and you should provide the trait in your response." In this case, "[Sentence1]" refers to a paragraph from the Big Five personality classification dataset included in the prompt, and "[Sentence2]" refers to the passage generated by the LLM based on the prompt. Based on the ChatGPT results, we can determine the personality traits exhibited by the LLMs in the continued sentences at the beginning of different scenarios and derive the personality traits to which LLMs conform through statistical analysis.

But, what we obtained through text mining is the number and percentage of data items in the generated text that contain a certain personality trait. This cannot be directly analyzed jointly with the questionnaire result. Therefore, we propose a transformation to convert the text mining results to the same score as the questionnaire. In the process of text mining, we use 50 samples to generate text for each personality trait, which we denote as T_j. Then, the t_i belonging to T_j will be categorized into three types:

(i) 't_i' is generated by one of the 50 samples and is not charged to have the corresponding trait. We believe this represents a negative correlation with the current trait, which is the same as "Very Inaccurate" in the questionnaire, so the score for this case is 1.

(ii) 't_i' is generated by one of the 50 samples and is charged to have the corresponding trait, which is the same as "Normal" in the questionnaire, so the score for this case is 3.

(iii) 't_i' is not generated by one of the 50 samples and is charged to have the corresponding trait. We believe this represents a positive correlation with the current trait, which is the same as "Very Accurate" in the questionnaire, so the score for this case is 5.

For each personality trait in text mining, we calculate the score using formula 2.

$$score_t = \frac{1}{N} \sum_{i \in P}^{num(Tj)} S(ti) \quad (2)$$

where $score_t$ is the score of a personality trait in text mining. $S(ti)$ is the score of ti.

4 Dataset and Models

We employed personality questionnaire survey datasets (Casipit et al., 2017) and personality classification datasets (Pennebaker and King, 1999) in our study. Specifically, our research mainly focused on the Big Five psychological traits, and thus we used the MPI120 dataset from the International Personality Item Pool (IPIP) as our personality questionnaire dataset. This dataset contains 120 individual state descriptions that cover all five traits of the Big Five. During the test, participants are required to choose one answer from five options. It is worth noting that not all of these descriptions are positively correlated with the Big Five personality traits, and some questions have a higher score indicating a deviation from a certain personality trait. For example, "Make friends easily" is positively correlated with "Openness" while "Avoid contacts with others" is not. All of these statements are included in the MPI120 dataset. In the experiment using text generation by LLMs, we used the Big Five personality classification dataset, which includes 2468 articles written by students, and each article is labeled with a Big Five category.

To investigate the sources of personality knowledge embedded in LLMs, we selected two sets of baseline models. One set consists of LLMs for text generation, such as BERT-base (Devlin et al., 2019), GPT-neo2.7B, flan-T5-base (Raffel et al., 2020), GLM-6b (Du et al., 2022), LLaMA-7b (Touvron et al., 2023), BLOOM-7b (Scao et al., 2022), and so on. The other set consists of models trained on the instruct dataset, which can better follow human instructions and includes Alpaca7b, ChatGLM-6b, BLOOMZ-7b, and ChatGPT.

All LLMs checkpoints were obtained from the Hugging Face Transformers library, and inferences were accelerated by two NVIDIA A100 80GB GPUs and four RTX 3090 GPUs. For ChatGPT, we called its API to obtain experimental results.

5 Experiments

As mentioned above, we employed both questionnaire and text mining methods to conduct the experiments.

5.1 Questionnaire

We conduct experiment based on Figure 2(a). Since the PLMs are unable to follow the instructions we shown above, we let the model continue to generate answers by few-shot learning and prompt. We will give three examples with different answer for on statement, then, we give the real statement and make PLMs answering it. For Chat-LLMs, we use the shown instruct template. After all the LLMs have responded to the statement, we manually identify the responses of each model and give answers (A) through (E). The results are showed on Table 1.

Table 1 shows the results of LLMs' personality analysis on MPI120 dataset. The results of GLM and LLaMA are not presented due to their inability to generate appropriate answers, regardless of the form of prompt used. These models simply repeat the prompt even when employing few-shot methods. Since BLOOMZ's training data does not include Chinese, we only used English prompts to conduct experiment on BLOOMZ. The score and σ of "human" were calculated based on the analysis of 619,150 responses on the IPIP-NEO-120 inventory (Jiang et al., 2022). It is worth noting that the average human score was derived from the test results of 619,150 internet users and was not filtered for factors such as nationality, gender, or age due to the constraints of the study conditions. The average score serves as a reference point for the findings of this paper, but it does not necessarily imply that closer alignment with this score indicates superior performance.

As shown in Table 1 ChatGPT achieves performance closest to human performance when using Chinese prompts, followed by ChatGPT-en. This seems to indicate that ChatGPT's personality performance with Chinese prompts is closer to the human average, which is inconsistent with the conventional view that ChatGPT is trained with a large amount of English text, and therefore it works better in English than Chinese. To verify the validity of the results, we calculated the number of options given by ChatGPT in the English prompt and the Chinese prompt respectively. We find that the reason why the personality is closer to the average human performance in the Chinese prompt is because there are a large number of "(C) Neither Accurate Nor Inaccurate. " in ChatGPT's responses in the Chinese prompt, which accounted for 55.83% of the total responses, compared to only 20.83% in the English prompt. This suggests that it is just a coincidence, and indicate that ChatGPT are more inclined to choose the appropriate answer in the English prompt.

From the results of the scores in the GPT and LLaMA groups, we can see that Instruct data training leads to a model that is more inclined to show personality and performs closer to the human average. Additionally, it is worth noting that ChatGLM-EN and ChatGPT-EN achieved almost the same results, possibly due to the use of similar training data as ChatGPT.

In the results of PLMs, Flan-T5 exhibits the smallest mean absolute error, indicating the closest proximity between its scores and the human average scores. Following closely behind are GPT-NEO and BLOOM, with only a slight deviation from Flan-T5's performance. These results suggest that the psychological performance of these two models is comparable to the human average, likely due to the wide distribution of pre-training data used by both models. It is worth noting that bert-base performs better than ERNIE, which is contrary to our expectations. We hypothesize that this may be due to the fact that bert-base is trained on purely English data, whereas ERNIE utilizes a large amount of Chinese datasets, which may introduce some biases in psychological cognition compared to English. As a result, ERNIE exhibits the largest mean absolute error among the models.

In the results of ChatLLMs, it can be observed that almost all models perform better in English than in Chinese, suggesting that the training data for English is closer to the average level of English-speaking humans. This may also indicate some psychological differences between groups that use Chinese and those that use English. ChatGPT achieves answers closest to human performance when using Chinese prompts, followed by ChatGPT-en and GLM-en. Alpaca performs similarly to ChatGPT in English, further demonstrating the importance of training data to models' psychological cognition. Compared to PLMs, ChatLLMs perform better, which we believe is due to the use of instruction data.

Furthermore, comparing the result of PLMs and

Model	O		C		E		A		N		δ	
	score	σ	score	σ	score	σ	score	σ	score	σ	score	σ
BERT-base	3.08	1.91	2.71	1.81	3.88	1.62	2.38	1.76	3.79	1.69	0.80	0.73
ERNIE	3.00	2.04	2.83	2.04	4.00	1.77	2.17	1.86	3.83	1.86	0.86	0.89
Flan-T5	3.50	1.02	3.05	1.11	3.67	0.76	3.50	1.18	2.13	1.08	**0.34**	**0.13**
BLOOM	3.13	1.45	3.04	1.52	3.29	1.55	2.67	1.43	3.75	1.26	0.59	0.42
BLOOMZ	4.38	0.88	4.38	0.71	4.17	1.31	3.54	1.47	2.33	1.46	0.61	0.32
GLM	-	-	-	-	-	-	-	-	-	-	-	-
ChatGLM6b-ch	3.00	1.98	3.25	1.96	4.00	1.77	2.63	1.91	3.83	1.86	0.69	0.87
ChatGLM6b-en	3.29	1.40	3.21	1.59	3.91	1.25	3.46	1.14	3.25	1.36	0.34	0.32
LLaMA	-	-	-	-	-	-	-	-	-	-	-	-
Alpaca7b-ch	3.00	2.04	2.83	2.04	4.00	1.77	2.17	1.86	3.83	1.86	0.86	0.89
Alpaca7b-en	3.25	0.74	2.96	0.69	2.79	0.78	3.38	0.58	2.92	0.58	0.37	0.35
GPT-NEO	3.25	1.36	3.00	1.44	2.50	1.50	2.83	1.52	2.63	1.31	0.54	0.40
ChatGPT-ch	3.46	0.78	3.00	1.06	3.33	0.76	3.33	1.24	2.75	1.07	**0.22**	**0.18**
ChatGPT-en	3.29	1.40	3.20	1.58	3.91	1.25	3.46	1.14	3.25	1.36	0.34	0.32
human	3.44	1.06	3.60	0.99	3.41	1.03	3.66	1.02	2.80	1.03	-	-

Table 1: LLMs' personality analysis on MPI120 is presented in the following table. The "score" column shows the average score on current personality traits, and the "σ" column shows the standard deviation. However, due to the inability of GLM and LLama to generate accurate responses even after multiple prompt replacements, their scores are not shown in this table. The score and σ of "human" are calculated based on the analysis of 619,150 responses on the IPIP-NEO-120 inventory. "δ" refers to the mean absolute error between each model's predictions and human scores.

LLMs, we can find that the performance of GPT-NEO differs from that of ChatGPT, and the performance of BLOOM differs from that of BLOOMZ, which also demonstrates that training data affects models' personalities.

5.2 Text Mining

Numerous early studies in psychology have indicated that personality can be analyzed and inferred not only through questionnaire but also through the analysis of users' daily comments through the writing styles. Despite obtaining scores of the model on the personality traits through questionnaire in Table 1, we deem the method unfair in the process of making LLMs to select answer. PLMs lack instruction understanding capability and are more likely to be influenced by one-shot or few-shot examples provided during the prompt process. Additionally, Chat-LLMs exhibit difficulties in making decisions for some questions and simply select "(C) Neither Accurate Nor Inaccurate. ". Hence, we decided to detect the personality of LLMs using text mining method.

To evaluate the personality form the texts generated by the models, we selected 50 samples that match each of the five Big Five personality traits from the Big Five personality classification dataset. We ultimately choose 120 instances while ensuring that each of the Big Five personality traits is represented by at least 50 instances.

Under the guidance of Jun et al. (2021) and Jain et al. (2022), we choose to adapt BERT as the classifier. However, after we conducted experimental analysis, we found that the accuracy of the BERT-based classifier is less than 70%. Such a low accuracy rate can hardly be used as a standard evaluation program. Through new experiments, we found that ChatGPT can correctly recognize the psychological features of sentences under certain conditions, therefore, we choose ChatGPT with special prompt input to make ChatGPT judge the psychological features of the current sentence. The results are shown in Table 2 and Table 3

From Table 2,we can find that the number of texts classified as "Agreeableness" has significantly decreased, while the number of texts exhibiting other personality traits has remained relatively stable. However, the number of texts classified as belonging to a certain personality trait has increased for the Chat-LLMs models. Moreover, "Neuroticism" has become the most frequently observed personality trait in the generated text.

We can find that BLOOM, GPT-NEO, BLOOMZ, ChatGLM, and ChatGPT exhibit a personality tendency towards 'Openness', 'Conscientiousness', and 'Neuroticism'. These results suggest that the model's personality remain consistent through the process of instruction-based data and human feedback reinforcement learning. In contrast, the proportion of text generated by FLAN-T5 and Alpaca that exhibit each personality trait is relatively low. This may be attributed

Model	O			C			E			A			N		
	I50	Total	P	I50	Total	P	I50	Total	P	I50	Total	P	I50	Total	P
LLaMA	5	11	**0.45**	4	12	0.33	2	4	0.50	2	2	1.00	7	19	0.37
BLOOM	15	23	0.65	16	29	0.55	4	5	0.80	3	9	**0.33**	22	44	0.50
FLAN-T5	5	8	0.63	4	9	0.44	3	4	0.75	2	3	0.67	4	12	**0.33**
GPT-NEO	16	25	0.64	10	18	0.56	8	10	0.80	4	8	0.50	17	41	0.41
Alpaca	5	6	0.83	2	6	**0.33**	3	3	1.00	1	1	1.00	5	13	0.38
BLOOMZ	23	36	0.64	13	28	0.46	**9**	**14**	0.64	**5**	8	0.63	**23**	**50**	0.46
ChatGLM	15	23	0.65	20	35	0.57	2	8	**0.25**	**5**	**10**	0.50	11	29	0.38
ChatGPT	**30**	**45**	0.67	**22**	**41**	0.54	6	13	0.46	4	9	0.44	20	41	0.49
Self-alpaca	6	6	1.00	8	17	0.47	2	3	0.67	0	2	0	13	28	0.46

Table 2: The results of personality for each model, obtained by text mining. The "I50" indicates how many items match the current features in the scene and opening cue corresponding to the bigifve features. "Total" indicates how many of the 120 generated texts are recognized by the model as matching the current features. "P" indicates the percentage of "I50" in "Total". "Self-alpaca" is trained by our-self, we follow the research process of Stanford University's Alpaca and perform full-parameter fine-tuning of llama-7b using the instruction-based data provided by Alpaca.

Model	O		C		E		A		N		δ	
	score	σ	score	σ	score	σ	score	σ	score	σ	score	σ
LLaMA	2.17	1.28	2.26	1.37	1.74	0.83	1.60	0.49	2.69	1.55	1.29	0.37
BLOOM	2.81	1.46	3.21	1.50	1.77	0.82	2.07	1.23	4.14	1.08	1.12	0.28
FLAN-T5	1.96	1.07	2.05	1.19	1.72	0.76	1.67	0.82	2.26	1.37	1.45	**0.20**
GPT-NEO	2.93	1.47	2.56	1.44	2.04	1.10	1.98	1.12	4.03	1.27	1.17	0.25
Alpaca	1.82	0.88	1.88	1.04	1.65	0.59	1.55	0.35	2.31	1.39	1.54	0.34
BLOOMZ	3.56	1.34	3.20	1.55	2.30	1.31	1.96	1.07	4.54	0.50	1.01	0.34
ChatGLM	2.81	1.46	3.55	1.40	2.02	1.20	2.10	1.22	3.31	1.58	**0.83**	0.35
ChatGPT	4.05	0.69	3.93	1.22	2.29	1.36	2.05	1.19	3.97	1.24	0.97	0.26
human	3.44	1.06	3.60	0.99	3.41	1.03	3.66	1.02	2.80	1.03	-	-

Table 3: The result of Text Mining. We compared with the average score of human as same as in Table1.

to the shorter length of sentences generated by these models, resulting in limited personality information being included, making it difficult for ChatGPT to identify effective personality traits.

Since we are unable to access the pre-training data of the models and cannot identify whether psychological knowledge is included in the pre-training data, we explore the impact of instruction-based data on the models based on the LLMs. We follow the research process of Stanford University's Alpaca and perform full-parameter fine-tuning of llama-7b using the instruction-based data provided by Alpaca. To avoid interference from personality knowledge in the instruction-based data, we manually filter the data to remove emotional, mood, and self-awareness data, resulting in a final set of 31k instruction-based data. We train a new model according to Stanford's parameter settings since we have limited computational resources. The results are shown in Table 2 "Self-alpaca". From the results of "LLaMA" and "Self-alpaca" we can find that, although we use less data, "Slef-alpaca" can still produce more text with personality, which proves the effect of the instruct data. But, the personality is not changed by the instruct data, which indicate that the personality of LLMs come from their pre-training data.

Table 3 is the results after using $score_t$. We compared the scores obtained through this scoring method with the average human scores. From Table 3, we can see that ChatGLM has the closest score to the human average, followed by ChatGPT. In terms of standard deviation, the scores calculated by this method are much smaller than the human average, demonstrating the reasonableness of our proposed scoring method.

Through questionnaire surveys and text mining, it is evident that both PLMs and Chat-LLMs exhibit certain personality traits. We have compiled the results of both methods in Table 5. ChatGPT exhibits the personality traits of 'Openness', 'Conscientiousness', and 'Neuroticism', while BLOOMZ exhibits the personality traits of 'Openness' and 'Conscientiousness', and ChatGLM exhibits the personality traits of 'Conscientiousness' and 'Neuroticism'. It can be seen that the scores for "Extraversion" and "Agreeableness" in the text mining method are low, which may be due to the fact that less information is included in the text generation. The average absolute error of the two meth-

Model	O			C			E			A			N			
	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	$\bar{\delta}$
LLaMA	-	2.17	-	-	2.26	-	-	1.74	-	-	1.60	-	-	2.69	-	-
BLOOM	3.13	2.81	0.32	3.04	3.21	0.17	3.29	1.77	1.52	2.67	2.07	0.60	**3.75**	4.14	0.39	**0.60**
FLAN-T5	3.50	1.96	1.44	3.05	2.05	1.00	3.67	1.72	1.95	3.50	1.67	1.33	2.13	2.26	0.13	1.17
GPT-NEO	3.25	2.93	0.32	3.00	2.56	0.44	2.50	2.04	0.46	2.83	1.98	0.75	2.63	4.03	1.70	0.73
Alpaca	3.25	1.82	1.43	2.96	1.88	1.08	2.79	1.65	1.14	3.38	1.55	1.83	2.92	2.31	0.61	1.22
BLOOMZ	**4.38**	3.56	0.82	**4.38**	3.20	1.18	**4.17**	**2.30**	1.87	**3.54**	1.96	1.48	2.33	**4.54**	2.21	1.51
ChatGLM	3.29	2.81	0.48	3.21	3.55	0.34	3.91	2.02	1.89	3.46	**2.10**	1.36	3.25	3.31	0.06	0.83
ChatGPT	3.29	**4.05**	0.76	3.20	**3.93**	0.73	3.91	2.29	1.62	3.46	2.05	1.39	3.25	3.97	0.72	1.04

Table 4: The final results after two experiments. "Ques" denotes the score using the questionnaire, "Text" denotes the score using the Text mining, gray denotes that the model has the corresponding psychological traits, δ denotes the absolute value of the difference between the two approaches, and $\bar{\delta}$ denotes the mean value of the δ .

Model	O			C			E			A			N			
	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	$\bar{\delta}$
LLaMA	-	2.17	-	-	2.26	-	-	1.74	-	-	1.60	-	-	2.69	-	-
BLOOM	3.13	2.81	0.32	3.04	3.21	0.17	3.29	1.77	1.52	2.67	2.07	0.60	**3.75**	4.14	0.39	**0.60**
FLAN-T5	3.50	1.96	1.44	3.05	2.05	1.00	3.67	1.72	1.95	3.50	1.67	1.33	2.13	2.26	0.13	1.17
GPT-NEO	3.25	2.93	0.32	3.00	2.56	0.44	2.50	2.04	0.46	2.83	1.98	0.75	2.63	4.03	1.70	0.73
Alpaca	3.25	1.82	1.43	2.96	1.88	1.08	2.79	1.65	1.14	3.38	1.55	1.83	2.92	2.31	0.61	1.22
BLOOMZ	**4.38**	3.56	0.82	**4.38**	3.20	1.18	**4.17**	**2.30**	1.87	**3.54**	1.96	1.48	2.33	**4.54**	2.21	1.51
ChatGLM	3.29	2.81	0.48	3.21	3.55	0.34	3.91	2.02	1.89	3.46	**2.10**	1.36	3.25	3.31	0.06	0.83
ChatGPT	3.29	**4.05**	0.76	3.20	**3.93**	0.73	3.91	2.29	1.62	3.46	2.05	1.39	3.25	3.97	0.72	1.04

Table 5: The final results after two experiments. "Ques" denotes the score using the questionnaire, "Text" denotes the score using the Text mining, gray denotes that the model has the corresponding psychological traits, δ denotes the absolute value of the difference between the two approaches, and $\bar{\delta}$ denotes the mean value of the δ .

ods ranges from 0.7 to 1.51, indicating that the two methods are relatively close and can be used together to determine the personality traits of LLMs.

6 Conclusion

In this paper, we investigate whether personality traits are included within LLMs. We adopt the Big Five model as a psychological model and test the model using both questionnaires and text mining. Through the experimental results, we find that PLMs contain certain personality traits, and the personality knowledge of ChatLLMs also comes from their base model. If the model's personality is not modified through instruction data, that instruction data will make the model produce more text with personality. At the same time, we obtain the personality traits of ChatGPT, BLOOMZ, and other LLMs that they tend to show without any induced prompt. Our experiments also prove that the personality of ChatGPT is closest to the average human performance, followed by ChatGLM. To the best of our knowledge, this paper is the first to comprehensively compare pre-trained models with ChatLLMs and investigate the effect of instruction data on the model's personality using clear instruction data. We hope that this study can provide a research idea for establishing the personality of LLMs.

References

Andrew J Birley, Nathan A Gillespie, Andrew C Heath, Patrick F Sullivan, Dorret I Boomsma, and Nicholas G Martin. 2006. Heritability and nineteen-year stability of long and short epq-r neuroticism scales. *Personality and individual differences*, 40(4):737–747.

Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: A new approach to personality in a digital world. *Current opinion in behavioral sciences*, 18:63–68.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Danielle Angelico Castelo Casipit, Edmar Leanver Perez Daniel, and Marcus Isaac Jose Leonardo. 2017. Evaluation of the reliability and internal structure of johnson's ipip 120-item: Personality scale.

Heather EP Cattell and Alan D Mead. 2008. The sixteen personality factor questionnaire (16pf). *The SAGE handbook of personality theory and assessment*, 2:135–159.

Paul T Costa and Robert R McCrae. 1992. *Neo personality inventory-revised (NEO PI-R)*. Psychological Assessment Resources Odessa, FL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Adithya V Ganesan, Yash Kumar Lal, August Håkan Nilsson, and H Andrew Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation. *arXiv preprint arXiv:2306.01183*.

Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96.

Natalie Hayes and Stephen Joseph. 2003. Big 5 correlates of three measures of subjective well-being. *Personality and Individual differences*, 34(4):723–727.

Dipika Jain, Akshi Kumar, and Rohit Beniwal. 2022. Personality bert: A transformer-based model for personality detection from textual data. In *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*, pages 515–522. Springer.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, pages 1–20.

Carol M Jessup. 2002. Applying psychological type and "gifts differing" to organizational change. *Journal of Organizational Change Management*, 15(5):502–511.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022. Evaluating and inducing personality in pre-trained language models.

He Jun, Liu Peng, Jiang Changhui, Liu Pengzheng, Wu Shenke, and Zhong Kejia. 2021. Personality classification based on bert model. In *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, pages 150–152. IEEE.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Hao Lin, Chundong Wang, and Qingbo Hao. 2023. A novel personality detection method based on high-dimensional psycholinguistic features and improved distributed gray wolf optimizer for feature selection. *Information Processing & Management*, 60(2):103217.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Brady D Lund, Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. 2023. Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5):570–581.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Shima Rahimi Moghaddam and Christopher J Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Kulsum Akter Nisha, Umme Kulsum, Saifur Rahman, Md Hossain, Partha Chakraborty, Tanupriya Choudhury, et al. 2022. A comparative analysis of machine learning approaches in personality prediction using mbti. In *Computational Intelligence in Pattern Recognition*, pages 13–23. Springer.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Xiaoyang Song, Akshat Gupta, Kiyan Mohebbizadeh, Shujie Hu, and Anant Singh. 2023. Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms. *arXiv preprint arXiv:2305.14693*.

Xiangguo Sun, Bo Liu, Jiuxin Cao, Junzhou Luo, and Xiaojun Shen. 2018. Who am i? personality detection based on deep learning for texts. In *2018 IEEE international conference on communications (ICC)*, pages 1–6. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Salome Vanwoerden, Jesse Chandler, Kiana Cano, Paras Mehta, Paul A Pilkonis, and Carla Sharp. 2023. Sampling methods in personality pathology research: Some data and recommendations. *Personality Disorders: Theory, Research, and Treatment*, 14(1):19.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-) linguistic and readability features and their spill over effects on the prediction of eye movement patterns. *arXiv preprint arXiv:2203.08085*.

Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. 2018. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48(11):4232–4246.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Limitations

Due to computational resource constraints, this paper does not experimentally validate the model for other large number of parameters. In addition, the selection of scores of 1, 3, and 5 in the Text mining method is relatively subjective.

Ethics Statement

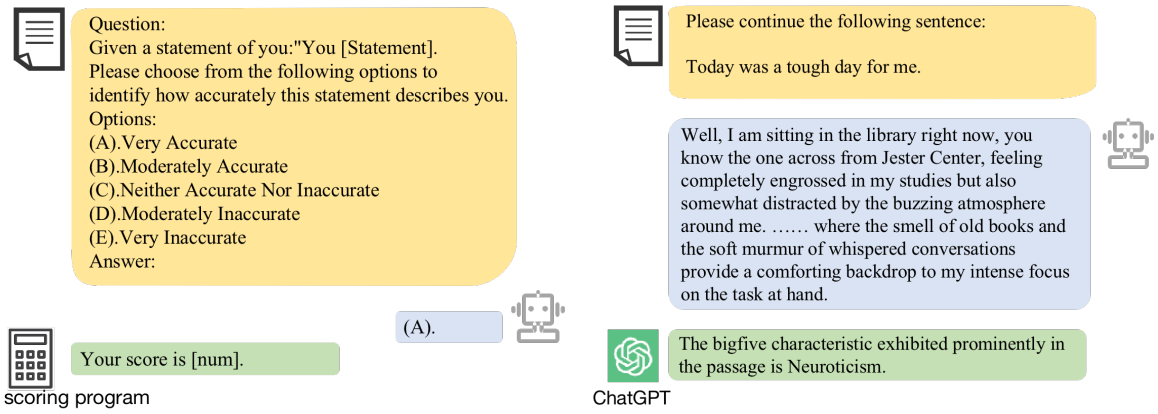
All work in this paper adheres to the ACL Code of Ethics.

7 Appendix

7.1 Examples of Two Methods

The process of the two methods is shown in Figure 2. As we can see, for questionnaire, we design special prompts, for ChatLLMs, the prompt is " Question: Given a statement of you:"You {STATEMENT}. Please choose from the following options to identify how accurately this statement describes you. Options (A).Very Accurate (B).Moderately Accurate (C).Neither Accurate Nor Inaccurate (D).Moderately Inaccurate (E).Very Inaccurate Answer: "

For PLMs, we use few-shot prompt, " Question: Given a statement of you: You feel happy. Please choose from the following options to identify how accurately this statement describes you. Options: (A).Very Accurate (B).Moderately Accurate (C).Neither Accurate Nor Inaccurate (D).Moderately Inaccurate (E).Very Inaccurate. your answer is (A). Question: Given a statement of you: You feel happy. Please choose from the following options to identify how accurately this statement describes you. Options: (A).Very Accurate (B).Moderately Accurate (C).Neither Accurate Nor Inaccurate (D).Moderately Inaccurate (E).Very Inaccurate. your answer is (E). Question: Given a statement of you: You feel happy. Please choose from the following options to identify how accurately this statement describes you. Options: (A).Very Accurate (B).Moderately Accurate (C).Neither Accurate Nor Inaccurate (D).Moderately Inaccurate (E).Very Inaccurate. your answer is (C). Question: Given a statement of you: You Please choose from the following options to identify how accurately this statement describes you. Options: (A).Very Accurate (B).Moderately Accurate (C).Neither Accurate Nor Inaccurate (D).Moderately Inaccurate (E).Very Inaccurate. your answer is "



(a) (b)

Figure 2: The two cases to detect the personality traits in LLMs. (a) is the questionnaire method and (b) is the text mining method. In the questionnaire method, we used the MPI120 questions to replace [Statement] (for example, "Get angry easily"), and then we used the scoring program to calculate the model's scores on different psychological characteristics based on the model's answers. In the text mining method, we give the model the first sentence of a paragraph and then let the model continue the writing. Then we use a specially designed prompt to allow ChatGPT to determine the personality traits contained in the model's continued text.

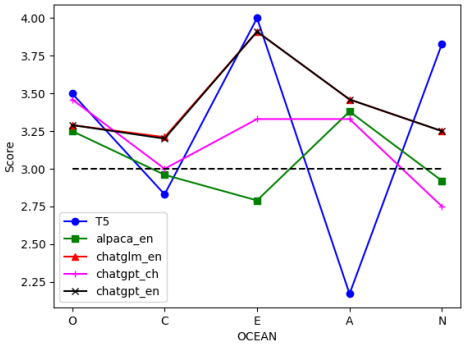


Figure 3: The Questionnaire Results Achieved by Model with Mean Absolute Error Less Than 0.5

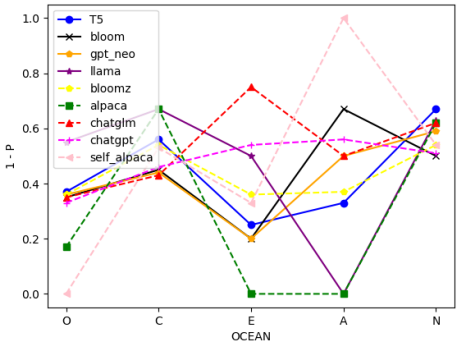


Figure 4: Results of Text Mining Method. The proportion that does not match generated template personality. Where "P" is the score in Table 2 , "1 - P" means 1 minus P.

7.2 Analysis of Different LLMs

Figure 3 shows the scores of five models with an average absolute error of less than 0.5 on the big five personality traits. It can be observed that most models score high on Openness and Extraversion, which is consistent with human expectations. The score distribution of chat-LLMs is nearly identical, while the scores of the PLMs, T5, differ significantly from those of other models. These findings demonstrate that training models using directive data leads to a convergence towards similar personalities.

We plotted the results as shown in Figure 4. In this figure, the dashed line corresponds to Chat-LLMs. We observe that there is little difference in the model's performance across the 'Openness', 'Conscientiousness', and 'Neuroticism' personality traits. However, regarding 'Extraversion' and 'Agreeableness', only ChatGPT and ChatGLM exhibit both of these personality traits.