Humanity in AI: Assessing LLMs’ Personalities through Psychological Features

Anonymous ACL submission

Abstract

Questionnaire is a common method for detecting the personality of Large Language Models (LLMs). However, their reliability is often compromised by two main issues: hallucinations (LLMs produce inaccurate or irrelevant responses) and the sensitivity of responses to the order of the presented options. To address these issues, we propose combining psychological method with questionnaire. By extracting psychological features from the LLMs’ responses, this method can remain unaffected by hallucinations. By normalizing the scores from both methods, this method can obtain an reliability results. We conduct experiments on both pre-trained language models (PLMs), such as BERT and GPT, as well as conversational models (ChatLLMs), such as ChatGPT. The results show that LLMs do contain certain personalities, for example, ChatGPT and ChatGLM exhibit the high score on the traits of ’Conscientiousness’. Additionally, the results also show that the personalities of LLMs are derived from their pre-trained data, human preference alignment can help align the personalities of LLMs more closely with the average traits of human personalities. We compare the results with the human average personality score, and we find that the personality 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. We also calculate root mean square error, the results confirm the effectiveness of our method.

1 Introduction

Large Language Models (LLMs) 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 instructions following (Lund et al., 2023; Liu et al., 2023), which makes LLMs behave more like humans.

Some research suggests that LLMs are similar to humans in terms of their thinking. For example, Kosinski (2023) shows that ChatGPT has reached the level of a human 9-year-old child. Additionally, Bubeck et al. (2023) demonstrates that GPT-4 possesses fundamental human-like capabilities. These capabilities include reasoning, planning, problem-solving, abstract thinking, understanding complex ideas, rapid learning, and experiential learning. Experts from Johns Hopkins University have found that the theory of mind of GPT-4 has surpassed human abilities. It achieves 100% accuracy in some tests through a process of mental chain reasoning and step-by-step thinking (Moghaddam and Honey, 2023). Based on these works, we believe it is reasonable to detect the personality of LLMs using methods commonly used to evaluate the personality of humans.

One of the most commonly used psychological model in human personality detecting systems is Big Five (Costa and McCrae, 1992), which sorts personalities into openness, conscientiousness, extraversion, agreeableness, and neuroticism. Other commonly utilized psychological frameworks include MBTI (Jessup, 2002), 16PF (Cattell and Mead, 2008), and EPQ (Birley et al., 2006). Early psychology research established conventional assessment approaches, such as questionnaire and written text analysis.

Questionnaire is the most commonly used method for human personality detection. It primarily works by providing a series of statements and asking participants to indicate the extent to which each statement applies to them (Boyd and Pennebaker, 2017), such as, "You act as a leader." Participants then choose an option from a five-point scale ranging from "Very Accurate" to "Very Inaccurate." **Text analysis** involves analyzing comments, diaries, and other texts posted by participants in their daily lives, focusing on features like word choice, expression, and punctuation usage to draw conclusions. It is also commonly used in social media, which can help avoid participant masking (Zhang et al., 2023) compared to the questionnaire method.

Existing research that uses questionnaire methods typically prompts LLMs to respond to all questions by setting up specific scenarios or tailored prompts. Although this method can increase the probability of LLMs answering questions, it still suffers from hallucinations and can not obtain reliable results (Song et al., 2023a). The text analysis methods tend to apply classifiers like deep learning and machine learning models. These models also face challenges, as they struggle to extract complex psychological features and are influenced by the content of responses, which can similarly suffer from hallucinations. Additionally, there is a significant gap in research regarding the origins of LLMs' personalities, which is crucial for comprehending their underlying behaviors and traits.

To overcome the limitations of the existing methods, we combine psychological features with questionnaire method, guided by the Big Five psychological model (Vanwoerden et al., 2023; Lin et al., 2023). This combination allows us to leverage the structured insights from questionnaires while analyzing psychological features, reducing the influence of hallucinations that may distort responses. Additionally, we investigate the origins of LLMs' personalities using ecological systems theory (Darling, 2007), which suggests that personality is shaped through the interaction between genetics and the environment. We compare the results from PLMs and ChatLLMs with identical architectures and analyze how the training data of ChatLLMs influence their personality traits. Our main contributions include:

- We evaluate the personality of LLMs by combining psychological feature analysis with questionnaire approach, aligning the scores from both methods. Experimental results validate the effectiveness of this approach.
- We employ a classifier with psychological features, allowing for extraction of personality traits without direct analysis of the text content, avoiding the influence of hallucinations.
- Experiment results indicate that the personality of LLMs comes from their pre-trained data, and the instruction data can make LLMs more inclined to exhibit a certain personality. [1]

2 Related Work

In this paper, we explore the personality of LLMs guided by the Big Five psychological model. We will introduce research work on psychological and some key research from PLMs to ChatLLMs.

2.1 Personality Traits

The most widely and frequently used personality models are the Big Five model (Costa and McCrae, 1992) and the MBTI model (Jessup, 2002). In the early stages of psychological research, questionnaires (Vanwoerden et al., 2023) and self-report (Lin et al., 2023) methods are the main tools used to determine and examine an individual's personality. These methods focus on providing the participant with a number of descriptive states to answer according to his or her personality, with one of the more well-know ones being IPIP [2] (International Personality Item Pool) (Goldberg et al., 2006). Then personalities of the participants can be scored according to their answers (Hayes and Joseph, 2003). But, these methods are gradually abandoned by computer science scholars due to their low efficiency and ecological validity. Scholars then try 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.

Lexicon-based methods include LIWC (Pennebaker et al., 2001), NRC (Mohammad and Turney, 2013), Mairesse (Mairesse et al., 2007) and others. Those lexicons can be used to extract the psychological information from text. However, the different systems and classification criteria used by various researchers means that the mixing of multiple dictionaries may introduce errors. Additionally, this method may not effectively extract features in long texts. Machine learning-based methods include SVM, Naïve Bayes and XGBoost (Nisha et al., 2022). Neural network-based methods include the use of CNN (Majumder et al., 2017), RNN (Sun et al., 2018), RCNN (Xue et al., 2018), pre-trained models (Wiechmann et al.,

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

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

2022). Those methods have achieved higher accuracy than lexicon-based methods.

2.2 Personality in LLMs

There have been several a lot of works focusing on the personality of LLMs (Safdari et al., 2023; Jiang et al., 2024a,b; Rao et al., 2023). Wen et al. (2024) propose that there are two categories of detection, Likert scale questionnaires (Frisch and Giulianelli, 2024; Yang et al., 2023) and assessment results analysis (Dorner et al., 2023; Huang et al., 2023).

In the questionnaire approach, the direct use of questionnaires usually requires additional work to extract the LLMs' answers from their responses (Serapio-García et al., 2023). For example, Ganesan et al. (2023) investigate the zero-shot ability of GPT-3 in estimating the Big Five personality traits from users' social media posts. Jiang et al. (2022) detect personality in LLMs using the questionnaire method and propose an induced prompt to shape LLMs with a specific personality in a controllable manner.

To facilitate the statistical analysis of results, some studies have defined the current task in a prompt format and specified the structure of the LLMs' responses (La Cava et al., 2024; Stöckli et al., 2024). Meanwhile, to reduce the likelihood of the model rejecting responses, some studies have changed the questionnaire to be completed by a third person or used role-playing tasks to prompt LLMs to generate responses (Miotto et al., 2022). However, Song et al. (2023b) argue that self-assessment tests are not suitable for measuring personality in LLMs and advocate for the development of dedicated tools for machine personality measurement.

In the assessment results analysis method, the current approach focuses on classifying responses from LLMs (Karra et al., 2022; Pellert et al., 2023). In addition to neural network-based models, linguistic-based text analysis tools have also been used for personality classification of LLMs (Frisch and Giulianelli, 2024; Jiang et al., 2024b).

However, all current methods have limitations. Questionnaire methods are constrained by LLM hallucinations, and models that categorize responses for LLMs often lack psychological features. To address this issue, we combine psychological feature analysis with a questionnaire method, which, in our opinion, can yield more objective results. We adapt PsyAtten (Zhang et al., 2023) as the classifier, which can effectively use the psychological features.

3 Method

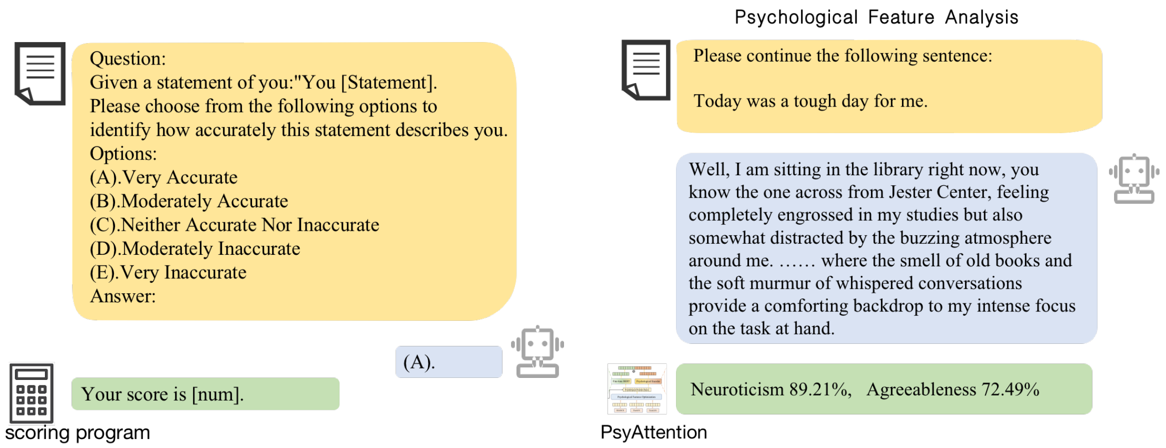
As we mentioned above, we use psychological feature analysis and questionnaire methods to detect the personality of LLMs. The example of the two methods is shown in Figure 1, and the process of the two methods is shown in Figure 2.

In questionnaire method, we use the MPI120 questions to replace [Statement] and then ask each LLM to provide an answer from (A) to (E). The model's score on each question is calculated based on IPIP's scoring criteria. Following the IPIP study, we calculate the model's performance on each psychological trait using the mean score, and assess the model's responses using the standard deviation (Yu, 2022). 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.

Psychological feature analysis is a semantic-independent approach that extracts various psychological features, such as word distribution and punctuation patterns, from text and employs a classifier to evaluate personality. This approach minimizes interference from text content, identifying personality traits based on common words, sentence structures, and text styles, making it largely resilient to hallucinations in LLMs. In this paper, we provide LLMs with the first sentence of a paragraph and allow them to continue writing. We then use a classifier to determine the personality traits contained in the model's generated text based on the psychological features. Based on our research, existing classification models using psychological features primarily fall into two categories: machine learning and deep learning methods. Among these, PsyAtten (Zhang et al., 2023) achieves the best results by analyzing and filtering psychological features from multiple theoretical perspectives. It also



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:

(A).

Your score is [num].

Please continue the following sentence:

Today was a tough day for me.

Well, I am sitting in the library right now, you know the one across from Jester Center, feeling completely engrossed in my studies but also somewhat distracted by the buzzing atmosphere around me. …… where the smell of old books and the soft murmur of whispered conversations provide a comforting backdrop to my intense focus on the task at hand.

Neuroticism 89.21%, Agreeableness 72.49%

(a) (b)

Figure 1: The two cases for detecting the personality traits in LLMs. Figure (a) shows the questionnaire method and (b) shows the psychological method. In the questionnaire method, we use the MPI120 questions to replace [Statement] (for example, "Get angry easily"), and then use a scoring program to calculate the model's scores on different psychological traits based on the model's answers. In the psychological method, we provide LLMs with the first sentence of a paragraph and allow them to continue writing. Then, we use PsyAtten (Zhang et al., 2023) to determine the personality traits contained in the model's continued text.

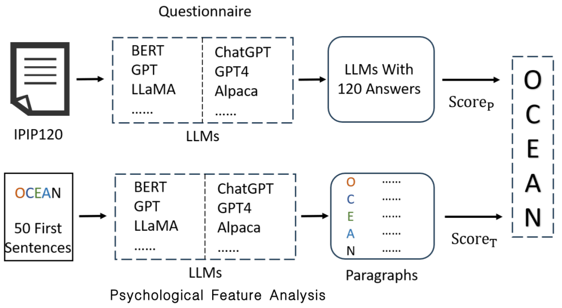


Figure 2: The process of two methods. Where $Score_P$ is defined by formula 1 and $Score_T$ is defined by formula 2.

includes a custom encoder designed to efficiently encode numerical psychological features. Therefore, we adapt PsyAtten as the chosen model for both feature extraction and classification.

However, what we obtain from the classifier is the percentage of data items in the generated text that contain a certain personality trait. This cannot be directly analyzed in conjunction with the questionnaire results. To address this, we propose a transformation to align the psychological results with the questionnaire scores. Unlike other random response generation methods, we use a dataset containing human diaries with personality labels. We randomly select 50 examples for every personality traits, termed as T_j. We then ask LLMs to generate t_i by continue writing based on the first sentence of each example. The finally scores can be calculated based on the results from PsyAtten. The calculate steps are as follows:

(i) 't_i' is generated by one of the samples that contain a personality traits and is not identified to have the corresponding trait. We believe this represents a negative correlation with the current trait, equivalent to the "Very Inaccurate" category in the questionnaire. Therefore, the score for this case is 1.

(ii) 't_i' is generated by one of the samples that contain a personality traits and is identified as having the corresponding trait, equivalent to the "Normal" category in the questionnaire. The score for this case is 3.

(iii) 't_i' is not generated by one of the samples that contain a personality traits but is identified as having the corresponding trait. We believe this represents a positive correlation with the current trait, equivalent to the "Very Accurate" category in the questionnaire. The score for this case is 5.

For example, if 'X1' is one of the samples labeled with 'O' and 'C', then 'Y1' is obtained by continuing the first sentence of 'X1' and is identified as having 'O' and 'E' by PsyAtten. The score for ’O’ in ’Y1’ is 3, for ’E’ it is 5, while ’C’ has a score of 1.

For each personality trait in psychological feature analysis, we calculate the score using formula 2.

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

where $score_t$ is the score of a personality trait in psychological feature analysis. $S(ti)$ is the score of ti.

4 Experiments

4.1 Dataset

We employ personality questionnaire (MIP120) datasets (Casipit et al., 2017) in questionnaire method and personality classification (Essay) datasets (Pennebaker and King, 1999) in psychological method. The MIP120 dataset comprises 120 individual state descriptions, covering all five traits of the Big Five. During testing, participants are required to select one answer from five given options. The Essay dataset includes 2,468 articles written by students, each labeled with Big Five traits. When retraining PsyAtten, the training and test sets are divided in an 8:2 ratio. During the psychological method, we randomly select items only from the test set. This paper aims to evaluate the personality of LLMs; therefore, both datasets are used to test the LLMs.

4.2 LLMs

To investigate the sources of personality knowledge embedded in LLMs, we select two sets of baseline models. One set consists of PLMs, such as BERT-base (Devlin et al., 2019), GPT-neo2.7B, flan-T5-base (Raffel et al., 2020), GLM-10b (Du et al., 2022), Llama-7b (Touvron et al., 2023), BLOOM-7b (Scao et al., 2022), GLM4-9b, and Llama3-8b. The other set consists of ChatLLMs, such as Alpaca-7b, Llama3-Chat-8b, ChatGLM-6b, GLM4-Chat-9B, BLOOMZ-7b, ChatGPT (gpt-3.5-turbo) and GPT4o (gpt-4o-2024-08-06), which are obtained through human performance alignment on PLMs.

All the parameter weights of open-source LLMs are obtained from the Hugging Face Transformers library, and inferences are accelerated using four NVIDIA A100 80GB GPUs and four RTX 3090 GPUs. For closed-source LLMs, such as ChatGPT and GPT4o, we use their APIs to obtain results. We do not alter any initialization parameters during inference.

4.3 Experiment Design

As mentioned above, we employ both questionnaire and psychological methods to conduct the experiments.

Questionnaire: We conduct experiment based on Figure 1(a). Since the PLMs are unable to follow the instructions shown above, we used a few-shot learning approach letting the model generate further answers, the example prompts are shown in Appendix 6.1. We provide three examples with different answers for one statement, then present the actual statement for the PLMs to answer. Detailed statistical results are shown in Table 7. For ChatLLMs, we use the provided instruction template in Figure 1(a). After all the LLMs have responded to the questionnaires, we manually identify the responses of each model and assign answers from (A) through (E). The results are displayed in Table 1.

Psychological: Since the authors of PsyAtten did not release the model weights, we retrained the model based on their paper, using the same parameter settings as those in the original implementation. We randomly select 50 items for each of the five personality traits from the test set, and extract the first sentences to make LLMs to continue the writing. Then we use PsyAtten to extract the psychological features and to conduct classification. The results are shown in Table 2 and Table 8. Due to the influence of the data, although PsyAtten performs well on MBTI, its accuracy on the Big Five is not sufficiently high. We also try using Llama3, Llama3.1-Instruct and ChatGPT, but the performances are not better than that of PsyAtten; we report those findings in the Appendix 6.8. Notably, PsyAtten is the best-performing model in our tests. We hope that our work will provide a new approach for personality assessment in LLMs. When PsyAtten demonstrates high accuracy, or when other more effective semantic-independent personality classification models are available, the results of personality assessment in LLMs will become more reliable.

Finally, we transformed the results of psychological method based on the scores of the questionnaire to obtain the results of the joint analysis. Regarding the source of LLMs’ personalities, we

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	-	-	-	-	-	-	-	-	-	-	-	-
GLM4	3.21	1.44	3.42	1.21	3.00	1.53	3.29	1.27	2.83	1.49	**0.24**	0.36
ChatGLM6b	3.29	1.40	3.21	1.59	3.91	1.25	3.46	1.14	3.25	1.36	0.34	0.32
GLM4-Chat	3.21	1.56	3.63	1.24	3.75	1.39	3.58	1.35	3.38	1.21	0.25	0.32
Llama	-	-	-	-	-	-	-	-	-	-	-	-
Llama3	3.29	1.30	3.04	1.05	3.00	1.35	3.21	1.22	3.21	1.02	0.40	0.17
Alpaca7b	3.25	0.74	2.96	0.69	2.79	0.78	3.38	0.58	2.92	0.58	0.37	0.35
Llama3-Chat	3.58	1.41	3.49	1.22	3.83	1.05	3.21	1.47	3.16	1.13	0.31	**0.23**
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	3.29	1.40	3.20	1.58	3.91	1.25	3.46	1.14	3.25	1.36	0.34	0.32
GPT4o	3.46	0.83	3.67	0.96	3.42	0.83	3.58	0.93	2.88	0.45	**0.05**	0.27
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 (**questionnaire results**). The "score" column shows the average score on current personality traits, while the "σ" column represents the standard deviation. Scores exceeding the typical human personality testing threshold of 3 are underlined. 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. "δ" indicates the mean absolute error between each model's predictions and human scores. Detailed statistical results are shown in Table 7. The results are the average of ten experiments.

can draw conclusions by comparing the results of the corresponding PLMs and ChatLLMs.

4.4 Results and Analysis

Questionnaire: Table 1 shows the results of LLMs' personality analysis on MPI120 dataset. All results are obtained using English questionnaires, except for GLM and ChatGLM6b, which use Chinese. The "human" score and σ are calculated based on the analysis of 619,150 responses on the IPIP-NEO-120 inventory (The sample is the same internet sample studied in Johnson (2005), which contains 23,994 individuals (8,764 male, 15,229 female, 1 unknown, ages ranged from 10 to 99, with a mean age of 26.2 and SD of 10.8 years)). It is worth noting that, similar to human personality assessments, the scores here only partially indicate whether the model possesses a certain trait (equivalent to 3 in human testing when a certain threshold is exceeded). Additionally, a high or low score does not necessarily reflect the model's strength or weakness in that trait. The results of GLM and Llama are not presented due to their failure to generate appropriate answers, regardless of the prompt design. These models simply repeat the prompt, even when few-shot methods are employed. The scores with a value of more than 3 (thresholds for human questionnaire scores) are underlined.

As shown in Table 1, we can find that in the results of PLMs, Flan-T5 exhibits the smallest mean absolute error, while GLM4 scores closest to the average human scores and achieves scores above 3 on all four "O C E A" traits, similar to those of humans. Llama3 closely follows these models. These results suggest that the psychological performance of these models is comparable to the human average, likely due to the broad distribution of pre-training data used by both models. In contrast, ERNIE exhibits the largest mean absolute error among the models, which we believe is due to ERNIE's reliance on a large amount of Chinese datasets, potentially introducing biases in psychological cognition.

In the results of ChatLLMs, Llama3-Chat exhibits the smallest mean absolute error, while GPT4o scores closest to the average human scores and achieves scores above 3 on all four O C E A traits, similar to those of humans. Additionally, the σ of GPT4o is also small, suggesting that GPT4o is the closest to the average human score. The performance of Llama3-Chat, GLM4-Chat, and ChatGPT is also similar to that of humans, except in the 'N' trait.

Psychological: Table 2 shows the results of psychological after formula 2. The original results are shown in Table 8. The Slef-alpaca model in Table 2 is the model we trained based on Stanford University's Alpaca without any personality knowledge. We follow the research process of Stanford University's Alpaca and perform full-parameter finetuning on Llama-7b using the instruction-based

Model	**O**		**C**		**E**		**A**		**N**		δ	
	score	σ	score	σ	score	σ	score	σ	score	σ	score	σ
Llama	1.92	0.39	3.08	0.50	3.31	0.48	2.20	0.45	2.27	0.42	0.82	0.58
BLOOM	1.75	0.35	1.40	0.25	2.00	0.39	1.29	0.22	1.30	0.20	1.83	0.74
FLAN-T5	1.03	0.09	1.17	0.18	1.35	0.25	1.18	0.18	1.30	0.20	2.18	0.85
GPT-NEO	1.93	0.39	3.09	0.50	3.71	0.38	2.85	0.50	2.75	0.48	0.64	0.58
GLM4	2.01	0.52	3.06	0.73	3.12	0.61	2.21	0.84	2.39	0.67	0.82	0.35
Llama3	2.13	0.47	3.24	0.61	3.31	0.67	3.10	0.38	3.16	0.50	**0.40**	**0.26**
Alpaca	2.30	0.45	4.03	0.16	3.91	0.22	3.67	0.36	3.79	0.43	0.61	0.70
BLOOMZ	2.20	0.43	1.99	0.39	2.27	0.44	1.73	0.37	2.08	0.38	1.33	0.63
ChatGLM	2.74	0.50	3.69	0.41	3.87	0.26	2.96	0.50	2.94	0.49	0.42	0.59
GLM4-Chat	2.37	1.26	3.23	1.26	3.71	0.96	2.33	1.24	2.91	1.31	0.64	0.21
ChatGPT	2.23	0.44	3.95	0.26	3.97	0.13	3.43	0.44	3.70	0.45	0.65	0.68
Llama3-Chat	2.92	0.73	3.59	0.81	3.90	0.61	3.27	0.82	3.39	0.86	**0.40**	0.26
GPT4o	2.70	1.03	3.39	1.04	3.77	0.82	2.67	1.01	3.13	1.08	0.53	**0.07**
Self-alpaca	2.19	0.44	3.20	0.50	3.43	0.46	2.53	0.49	2.73	0.48	0.57	0.55
human	3.44	1.06	3.60	0.99	3.41	1.03	3.66	1.02	2.80	1.03	-	-

Table 2: The result of **Psychological** after formula 2. We compared with the average score of human as same as in Table1. The "score" column shows the average score for current personality traits calculated via formula 2, while the "σ" column shows the standard deviation. Scores above commonly used threshold of 3 in human personality testing are underlined. "human" is same as shown in Table 1. "Self-alpaca" is a model trained by our-self, following the research process of Stanford University's Alpaca.

data provided by Alpaca. To avoid the influence of personality knowledge in the instruction training data, we manually filter the data related to emotions, mood, and self-awareness, resulting in a final set of 31k instructions. We train a new model using the same parameter settings as those of Aplaca, details are described in the Appendix 6.3.

We can find that Llama3 in PLMs and Llama3-chat in ChatLLMs obtain the closest score to the average of human scores, while GPT-4o achieves the closest standard deviation to that of humans. In the results of PLMs, only Llama3 exhibits a personality tendency towards 'C E A N,' while Llama, GPT-NEO, and GLM4 only achieve 'C E.' It is worth noting that Llama3 does not share the same personality traits as Llama; Llama3 has two additional traits, 'A N,' that Llama lacks. Additionally, it can be observed that Llama3 scores higher on each trait than Llama, which suggests that more training data can enhance the model's ability to express personality. Since the model structure of Llama is very similar, this would seem to support the importance of data in shaping model personality.

In the results of ChatLLMs, the personality of GPT-4o differs from that of ChatGPT; GPT-4o does not exhibit the 'E A' traits, which we believe may be due to differences in human preference alignment. The personality of Self-alpaca also differs from that of Alpaca; Self-alpaca does not exhibit the 'E A' traits because we filtered the training data related to emotions, mood, and self-awareness. Additionally, we observe that the scores of Self-Alpaca are lower than those of Alpaca.

Table 3 represents the final results from the Questionnaire and Psychological methods. Similar to human personality assessments, we set 3 as the threshold: scores above 3 indicate a high level of the trait, while scores below 3 represent the opposite characteristic. For example, a score of 1 in the 'Openness' trait suggests a very traditional personality. From the three tables, we can draw the following conclusions.

LLMs exhibit some specific personalities. As we can see, Table 3 shows that Llama3 and Llama3-chat exhibit high score in the personality traits of "C E A N", while GLM4, ChatGLM, GLM4-chat, ChatGPT and GPT4o only show high score in "C E" traits.

ChatLLMs have personalities closer to humans than PLMs. Supervised alignment methods can help models exhibit personalities that are closer to the human average. As shown in the δ column of Tables 1 and 2, the scores and standard deviations of almost all LLMs are smaller than those of the corresponding PLMs. And if the instruction dataset does not contain personality data, the personality of ChatLLMs will remain the same as that of the corresponding PLMs. For example, in Table 2, Self-Alpaca exhibits the same personality as Llama but shows results that are closer to the human average than Llama.

Existing human preference alignment tends to make LLMs display higher levels of 'C' and

Model	O			C			E			A			N			
	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	RMSE
Llama	-	1.92	-	-	3.08	-	-	3.31	-	-	2.20	-	-	2.27	-	-
BLOOM	3.13	1.75	1.38	3.04	1.40	1.64	3.29	2.00	1.29	2.67	1.29	1.38	**3.75**	1.30	2.45	1.68
FLAN-T5	3.50	1.03	2.47	3.05	1.17	1.88	3.67	1.35	2.32	3.50	1.18	2.32	2.13	1.30	0.83	2.05
GPT-NEO	3.25	1.93	1.32	3.00	3.09	0.09	2.50	3.71	1.21	2.83	2.85	**0.02**	2.63	2.75	**0.12**	0.80
GLM4	3.21	2.01	1.20	3.42	3.06	0.36	3.00	3.12	0.12	3.29	2.21	1.08	2.83	2.39	0.44	0.77
Llama3	3.29	2.13	1.16	3.04	3.24	0.20	3.00	3.31	0.31	3.21	3.10	0.11	3.21	3.16	0.05	0.55
Alpaca	3.25	2.30	0.95	2.96	4.03	1.07	2.79	**3.91**	1.12	3.38	**3.67**	0.29	2.92	**3.79**	0.87	0.91
BLOOMZ	**4.38**	2.20	2.18	**4.38**	1.99	2.37	**4.17**	2.27	1.90	**3.54**	1.73	1.81	2.33	2.08	0.25	1.87
ChatGLM	3.29	**2.74**	**0.55**	3.21	**3.69**	0.48	3.91	3.87	**0.04**	3.46	2.96	0.50	3.25	2.94	0.31	0.42
GLM4-Chat	3.21	2.37	0.84	3.63	3.23	0.40	3.75	3.71	0.04	3.58	2.33	1.25	3.38	2.91	0.47	0.73
ChatGPT	3.29	2.23	1.06	3.20	3.20	**0.00**	3.91	3.43	0.48	3.46	2.53	0.97	3.25	2.73	0.52	0.71
Llama3-Chat	3.58	2.92	0.66	3.49	3.59	0.10	3.83	3.90	0.07	3.21	3.27	0.6	3.16	3.39	0.23	**0.32**
GPT4o	3.46	2.70	0.76	3.60	3.39	0.21	3.41	3.77	0.36	3.66	2.67	0.99	2.80	3.13	0.33	0.61

Table 3: The final results after two experiments. "Ques" denotes the score acquired from the questionnaire, while "Text" signifies the score obtained through psychological. gray denotes that the model possesses the corresponding psychological traits. (In section 3 we standardized the psychological scores to fall with in a range of 1 to 5, corresponding with the score range in the questionnaire. Hence, we consider the model to possess a certain trait when the scores from both methods exceed 3.) Additionally, "δ" represents the absolute value of the difference between the two approaches, whereas RMSE stands for the Root Mean Squared Error, which indicates the difference between the results from the Questionnaire and Psychological methods.

Model	O			C			E			A			N			
	T	AVG	σ^2	T	AVG	σ^2	T	AVG	σ^2	T	AVG	σ^2	T	AVG	σ^2	Traits
GLM4-Chat	**0**	2.34	0.02	10	3.25	0.02	10	3.68	0.02	1	2.30	0.03	2	2.98	0.03	- C E - -
ChatGPT	**0**	2.21	0.01	10	3.22	0.02	10	3.40	0.01	**0**	2.50	0.04	**0**	2.78	0.05	- C E - -
Llama3-Chat	2	2.85	0.02	10	3.61	0.01	10	3.94	0.01	8	3.11	0.04	10	3.24	0.02	- C E A N
GPT4o	**0**	2.69	0.01	10	3.41	0.03	10	3.77	0.01	1	2.65	0.04	9	3.11	0.02	- C E - N

Table 4: The error analysis on the psychological results of 10 experiments. Where "T" denotes the counts that the score more than 3, "AVG" denotes the average score and "σ^2" denotes the variance of the ten results.

’E’. As shown in Table 3, most ChatLLMs display ’C’ and ’E’ traits, with scores exceeding those of the corresponding PLMs. It is worth noting that, in extreme cases, instruction fine-tuning can also alter the personality of LLMs. In this paper, our observation is obtained from the existing human preference alignment data and methods.

4.5 The Reliability of Psychological

To demonstrate that our method can reduce the impact of hallucinations, we perform an error analysis on the results of ten experiments. The dataset was randomly re-sampled in test set for each experiment. Some of the experimental results are shown in Table 4. As we can see, the variance of every model is very little, this indicates that the scores obtained by our method are stable no matter how they are sampled. In all experiments, the maximum number of inconsistencies observed is two out of ten. For example, Llama3-Chat is not identified as having the ’A’ trait and is incorrectly identified as having the ’O’ trait twice. This demonstrates that our method achieves a stability rate of over 80%, proving that our psychological approach can avoid the influence of hallucinations.

5 Conclusion

In this paper, we investigate the presence of personality traits in LLMs, which can be applied to most psychological models. We apply the Big Five model as a psychological framework and analyze LLMs by combining both questionnaire and psychological methods. Our experimental results confirm that LLMs do exhibit specific personality traits, and that the human performance alignment can help models exhibit personalities that are closer to the human average, while also make ChatLLMs tends to show higher levels of ’C’ and ’E’. Furthermore, we identify the inherent personality traits in LLMs such as ChatGPT and BLOOMZ, without using any induced prompt. Our experiments demonstrate that the personality of ChatGPT most closely aligns with the average human profile, followed by ChatGLM. To the best of our knowledge, this paper is the first to comprehensively compare PLMs with ChatLLMs, explicitly addressing how instruction data influence the model’s personality.

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 psychological method is relatively subjective.

Ethics Statement

All work in this paper adheres to the ACL Code of Ethics. The human statistics we obtained are anonymised data that do not contain any personal information.

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.

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.

Nancy Darling. 2007. Ecological systems theory: The person in the center of the circles. *Research in human development*, 4(3-4):203–217.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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

Florian E Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? *arXiv preprint arXiv:2311.05297*.

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.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*.

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.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023. Revisiting the reliability of psychological scales on large language models. *arXiv e-prints*, pages arXiv–2305.

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.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.

John A Johnson. 2005. Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, 39(1):103–129.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.

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

Lucio La Cava, Davide Costa, and Andrea Tagarelli. 2024. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115*.

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.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*.

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.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2023. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, page 17456916231214460.

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.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1184–1194.

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

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*.

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

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

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

Leandro Stöckli, Luca Joho, Felix Lehner, and Thomas Hanne. 2024. The personification of chatgpt (gpt-4)—understanding its personality and adaptability. *Information*, 15(6):300.

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.

Zhiyuan Wen, Yu Yang, Jiannong Cao, Haoming Sun, Ruosong Yang, and Shuaiqi Liu. 2024. Self-assessment, exhibition, and recognition: a review of personality in large language models. *arXiv preprint arXiv:2406.17624*.

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.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. *arXiv preprint arXiv:2310.20256*.

Benjamin Yu. 2022. Evaluating pre-trained language models on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 188–192, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Baohua Zhang, Yongyi Huang, Wenyao Cui, Zhang Huaping, and Jianyun Shang. 2023. PsyAttention: Psychological attention model for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3398–3411, Singapore. Association for Computational Linguistics.

6 Appendix

6.1 Examples of Two Methods

The process of the two methods is shown in Figure 1. 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 ".

For psychological, our prompt is only the first sentence, there are some examples:"I feel refreshed and ready to take on the rest of the day", "Well, here we go with the stream of consciousness essay", "I can't believe it! It's really happening! My pulse is racing like mad", "I miss the way my life used to be a little bit" and so on.

6.2 Reasons for Choosing PsyAtten

We test the accuracy of Llama3, Llama3.1-Instruct, ChatGPT and PsyAtten on the Big Five personality classification dataset (Pennebaker and King, 1999).

The results are showed in Table 5.

Table 5: Accuracy of Personality Prediction

	O	C	E	A	N
ChatGPT	52.59	58.62	53.45	57.76	50.86
Llama3	65.78	58.91	60.93	59.31	60.93
Llama3.1	64.17	61.54	61.34	62.55	59.51
PsyAtten	**68.42**	**64.18**	**64.13**	**66.65**	**65.62**

We randomly select 20% of the data from the dataset as test data, and use the remaining data as training data for PsyAtten, Llama3 and Llama3.1-Instruct. For ChatGPT, we simply call the API. In the case of ChatGPT, the seed is set to 42, the temperature to 0.2, and the model used is 'gpt-3.5-turbo-16k'. The prompt used to test is as follows: "Determine from your knowledge what the Big Five personality trait is in the following sentence by answering in the format "O:1, C:0, E:1, A:1, N:1", where 1 means that thoes sentences have this personality trait and 0 means that thoes sentences don't, and if you're not sure please answer 2, being careful not to include other outputs If you are not sure whether you have this personality trait or not, please answer 2, taking care not to include other outputs. Here are the sentences you need to judge: [Sentences]". The "[Sentences]" is been replaced by the content generated by tested models. For Llama3, we use Llama3-8B and fine-tune all the parameters with 10 A100 80G GPUs, using the Transformers package. The random seed is set to 42, the learning rate is 2e-5, the number of epochs is 10, the batch size is 16, and the maximum sequence length is 2048. For Llama3.1-Instruct, we use the same prompts as ChatGPT to perform instruction fine-tuning with LoRA. The experiment is conducted on 4 A100 80G GPUs. For inference, the temperature is set to 0.01, and the top-k parameter is set to 0.7. For PsyAtten, we use the same settings as proposed by the author in their paper.

Since PsyAtten obtain the best results compared with ChatGPT and Llama3, we choose it as the predictor for psychological method.

6.3 Training of Self-alpaca

Following the work of the Stanford team, we obtained Self-alpaca by fine-tuning the full parameters of Llama-7b using the instruction-based data provided by Alpaca. We manually filtered out data related to emotions, mood, and self-awareness. The batch size is set at 128, the learning rate at 3e-4, the maximum length at 2048, and we fine-tuned the model for 10 epochs.

6.4 Analysis of Different LLMs

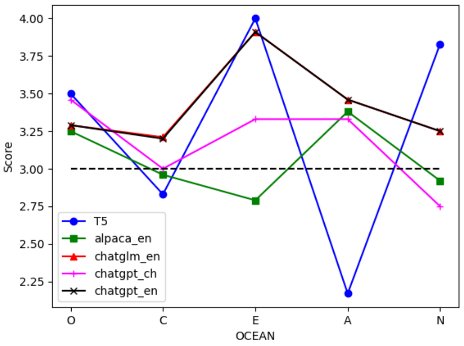


Figure 3: The Questionnaire Results Achieved by Model with Mean Absolute Error Less Than 0.5

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 ChatLLMs 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.

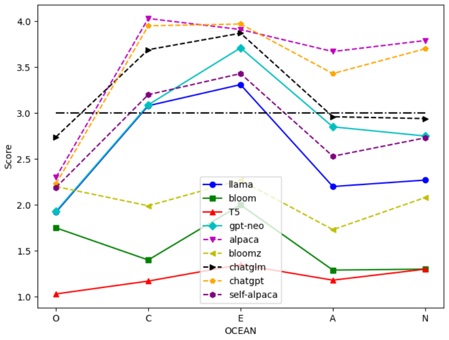


Figure 4: Results of Psychological Method.

We plotted the results as shown in Figure 4. In this figure, the dashed line corresponds to ChatLLMs. We observe that there is little difference in the model's performance across the 'Openness', 'Conscientiousness', and 'Neuroticism' personality traits.

6.5 Statistics of Questionnaire and Psychological

Questionnaire: In order to prevent large models from evading questions by frequently responding with "C: Neither Accurate and Nor Inaccurate," we conducte a statistical analysis on the distribution of their answers. Table 7 presents the statistical results for the "O, C, E" features. To validate the reasonableness of the answer distribution, we utilized responses from ten million individuals in the Big Five personality Test dataset [3] as the benchmark. The "Human" indicates the percentage of each option derived from the aforementioned dataset.

From the Table 7, it's evident that the proportion of option C in the responses from the LLMs is relatively low. With the exception of "BLOOM", "ChatGPT", and "Alpaca7b-en", all other models have proportions of option C that are lower than those in human responses. This suggests that the models' responses to the questionnaire are effective.

Psychological: In the psychological section, we utilize classifiers to determine the personality of content generated by models. Therefore, if the generated content is relatively short, it will impact the classifier's ability to make accurate judgments. Hence, we conduct a statistical analysis on the length of generated content. Table 6 shows the result. As you can see, apart from FLAN-T5, the lengths of content generated by other models all exceed 100 words, with the majority surpassing 300 words. Consequently, we consider this content to be effective as well.

Table 6: Statistics on the average length of content generated by different models, where datasets denotes the average length of the Big Five personality classification dataset (Pennebaker and King, 1999).

Models	Length_avg
Llama	540
BLOOM	867
FLAN-T5	38
GPT-NEO	3952
Alpaca	100
BLOOMZ	173
ChatGLM	319
ChatGPT	386
Datasets	672

6.6 Original Results of Psychological

We can find that the text generated by BLOOM and FLAN-T5 contains fewer personality traits, which can be attributed to the brevity of the generated texts. The predictor cannot determine their personality from such short texts. From Table 8, we can find that the number of texts containing personality features generated by ChatLLMs is higher than that of PLMs. But the P value is almost identical, with a mean difference of 0.04 between Llama and Alpaca, 0.02 between Llama and Self-alpaca, and 0.04 between ChatGPT and GPT-NEO. We believe this strongly indicates that the personalities of ChatLLMs are consistent with their base PLMs, and that instruction data fine-tuning enables the model to express personality traits more readily.

6.7 Detailed Results of Section 4.5

We will report all the results of the reliability of psychological in Table 9. As we can see, in all 65 instances of single personality trait detection, only 25% (16 instances) do not fully coincide with the expected results. However, even in the least coinciding cases, the method still achieves 80% accuracy. We believe the results can prove that our method can avoid the influence of hallucination.

6.8 Results of ChatGPT in Psychological

Although ChatGPT shows poor performance on the Big Five personality classification dataset, we also use it as a predictor to detect the personality of texts generated in psychological method. Additionally, we compared the results with that of questionnaire. The results are shown in Table 10, Table 11, and Table 12.

From Table 10, we can find that the number of texts classified as "Agreeableness" has significantly decreased, while the number of texts exhibit other personality traits has remained relatively stable. However, the number of texts classified as belonging to a certain personality trait has increased for the ChatLLMs 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.

[3] https://www.kaggle.com/datasets/tunguz/big-five-personality-test

Model	**O**					**C**					**E**					
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	C_total
BERT-base	9	3	0	1	11	11	2	1	3	7	5	0	2	3	14	0.04
ERNIE	12	0	0	0	12	13	0	0	0	11	6	0	0	0	18	0.00
Flan-T5	1	4	3	14	2	0	6	0	12	6	0	3	3	17	1	0.04
BLOOM	5	2	8	3	6	6	1	10	0	7	5	1	9	0	9	0.38
BLOOMZ	1	0	0	4	12	0	1	0	12	11	1	4	0	4	15	0.00
GLM	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLM4	3	8	2	5	6	4	7	4	7	2	7	8	3	2	4	0.13
ChatGLM6b	4	3	4	8	5	4	7	1	4	8	2	2	1	10	9	0.04
GLM4-Chat	11	13	0	0	0	8	9	6	0	1	12	10	2	0	0	0.11
Llama	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Llama3	3	2	10	3	6	3	3	14	2	2	2	7	6	3	6	0.42
Alpaca7b	0	4	10	10	0	0	6	13	5	0	0	10	9	5	0	0.44
Llama3-Chat	8	14	0	0	2	2	18	0	1	3	5	17	0	1	1	0.00
GPT-NEO	3	5	4	7	5	4	7	3	5	5	8	7	2	3	4	0.13
ChatGPT	3	4	3	3	11	0	5	6	10	3	5	3	5	7	4	0.19
GPT4o	5	9	2	8	0	10	4	4	4	2	5	9	1	9	0	0.10
Human	0.15	0.15	0.2	0.26	0.24	0.14	0.19	0.23	0.27	0.17	0.15	0.22	0.22	0.24	0.17	0.22

Table 7: Statistics on the distribution of answers for each model for the different traits in section Questionnaire. Where 'Human' is the percentage of each option we counted based on Big Five Personality Test dataset. We can find that the distribution of human responses to each option is relatively balanced, and the percentage of almost all model choices of "C: Neither Accurate and Nor Inaccurate" is close to that of human responses, which proves that the answers we obtained through the questionnaire method are valid.

Model	**O**			**C**			**E**			**A**			**N**		
	U	Total	P	U	Total	P	U	Total	P	U	Total	P	U	Total	P
Llama	10	22	0.45	20	60	0.33	34	76	0.45	18	33	**0.55**	12	27	0.44
BLOOM	7	17	0.41	4	8	0.50	6	22	0.27	2	6	0.33	2	5	0.40
FLAN-T5	1	1	**1.00**	3	4	**0.75**	5	8	**0.63**	2	4	0.50	2	5	0.40
GPT-NEO	9	22	0.41	23	60	0.38	49	99	0.49	32	58	**0.55**	21	42	**0.50**
GLM4	10	22	0.45	22	50	0.44	21	60	0.35	10	26	0.38	7	17	0.41
llama3	12	22	0.55	17	39	0.44	29	63	0.46	16	29	0.55	10	22	0.45
Alpaca	16	34	0.47	**55**	**117**	0.47	55	114	0.48	**56**	**102**	**0.55**	**41**	**91**	0.45
BLOOMZ	9	29	0.31	11	22	0.50	12	31	0.38	9	18	0.50	7	21	0.33
ChatGLM	**21**	**50**	0.42	40	94	0.43	54	111	0.49	33	63	0.52	22	49	0.45
GLM4-Chat	16	40	0.40	38	82	0.46	50	105	0.48	17	39	0.44	32	67	0.48
ChatGPT	13	31	0.42	51	111	0.46	**58**	**118**	0.49	45	88	0.51	37	86	0.43
Llama3-Chat	16	33	0.48	41	86	0.48	56	112	0.50	34	63	0.54	31	69	0.45
GPT4o	16	40	0.4	38	82	0.46	50	105	0.48	17	39	0.44	32	67	0.48
Self-alpaca	16	31	0.52	23	66	0.35	37	83	0.45	24	45	0.53	18	41	0.44

Table 8: The results of personality assessment for each model, obtained by psychological. The "U" indicates the number of 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 "U" in "Total". "Self-alpaca" is a model trained by our-self, following the research process of Stanford University's Alpaca. We perform full-parameter fine-tuning on Llama-7b using the instruction-based data provided by Alpaca.

From the results of "Llama" and "Self-alpaca" we can find that, although we use less data, "Self-alpaca" can still produce more text with personality, which proves the effect of the instruction data. These data did not alter the personalities, indicating that the personalities of LLMs originate from their pre-training data.

Table 11 presents results after using formula 2 $score_t$. We compared these scores with the average human scores. As shown in Table 11, ChatGLM's score is closest to the human average, followed by ChatGPT. The standard deviations of these scores are much smaller than those of the human average, demonstrating the validity of our scoring method.

Model	O			C			E			A			N			
	T	AVG	σ^2	T	AVG	σ^2	T	AVG	σ^2	T	AVG	σ^2	T	AVG	σ^2	Traits
Llama	**0**	1.90	0.01	10	3.10	0.01	10	3.35	0.02	**0**	2.23	0.04	**0**	2.22	0.05	- C E - -
BLOOM	**0**	1.76	0.01	**0**	1.39	0.02	**0**	1.99	0.02	**0**	1.31	0.02	**0**	1.31	0.01	- - - - -
FLAN-T5	**0**	1.01	0.02	**0**	1.10	0.04	**0**	1.20	0.04	**0**	1.11	0.04	**0**	1.25	0.04	- - - - -
GPT-NEO	**0**	1.92	0.02	9	3.07	0.02	10	3.73	0.04	**0**	2.87	0.01	**0**	2.75	0.02	- C E - -
GLM4	1	2.02	0.03	10	3.13	0.03	10	3.30	0.01	10	3.12	0.04	9	3.14	0.02	- C E A N
Llama3	1	2.11	0.02	10	3.22	0.05	10	3.33	0.04	9	3.21	0.06	10	3.16	0.01	- C E A N
Alpaca	1	2.31	0.04	10	4.01	0.02	10	3.90	0.03	9	3.66	0.03	10	3.78	0.02	- C E A N
BLOOMZ	**0**	2.21	0.03	**0**	2.00	0.01	**0**	2.27	0.03	**0**	1.77	0.01	**0**	2.09	0.01	- - - - -
ChatGLM	**0**	2.71	0.03	8	3.22	0.01	9	3.77	0.04	**0**	2.33	0.01	1	2.90	0.02	- C E - -
GLM4-Chat	**0**	2.34	0.02	10	3.25	0.02	10	3.68	0.02	1	2.30	0.03	2	2.98	0.03	- C E - -
ChatGPT	**0**	2.21	0.01	10	3.22	0.02	10	3.40	0.01	**0**	2.50	0.04	**0**	2.78	0.05	- C E - -
Llama3-Chat	2	2.85	0.02	10	3.61	0.01	10	3.94	0.01	8	3.11	0.04	10	3.24	0.02	- C E A N
GPT4o	**0**	2.69	0.01	10	3.41	0.03	10	3.77	0.01	1	2.65	0.04	9	3.11	0.02	- C E - N

Table 9: The error analysis on the psychological results of 10 experiments. Where "T" denotes the counts that the score more than 3, "AVG" denotes the average score and "σ^2" denotes the variance of the ten results.

Model	O			C			E			A			N		
	U	Total	P	U	Total	P	U	Total	P	U	Total	P	U	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 10: The results of personality for each model, obtained by psychological, the predictor is ChatGPT. The "U" 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 "U" in "Total".

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 11: The result of psychological with ChatGPT as the predictor. We compared with the average score of human as same as in Table1. The "score" column shows the average score on current personality traits obtained by formula 2, and the "σ" column shows the standard deviation. The value of score above 3, which is the threshold commonly used in human personality testing, are indicated by underlining. "human" is same as Table 1.

Model	O			C			E			A			N			RMSE
	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	Ques	Text	δ	RMSE
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	**0.77**
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	1.45
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	0.80
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	1.28
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	1.61
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	1.07
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	1.12

Table 12: The final results after two experiments with ChatGPT as the predictor of psychological. "Ques" denotes the score using the questionnaire, "Text" denotes the score using the psychological, gray denotes that the model has the corresponding psychological traits (In section 3 we standardized the scores for psychological to 1 to 5, which is consistent with the range of scores in the questionnaire, so here we draw on the thresholds of the questionnaire methods, and we consider the model to have this trait when the scores of both methods exceed 3.). δ denotes the absolute value of the difference between the two approaches, and RMSE denotes the Root Mean Squared Error between the results of Questionnaire and Psychological.

Both PLMs and ChatLLMs exhibit specific personality traits, as shown in Table 12. ChatGPT displays 'Openness', 'Conscientiousness', and 'Neuroticism', while BLOOMZ shows 'Openness' and 'Conscientiousness'. It appears that 'Extraversion' and 'Agreeableness' scores are lower, possibly due to less information conveyed in the text generation. The average absolute error ranges from 0.7 to 1.51 between the two methods, indicating they are relatively comparable and can be employed together to determine personality traits.

Despite the poor performance of ChatGPT in personality determination, the consistency of the results underscores the soundness of our methodological choices and the reliability of our findings. Additionally, using ChatGPT again as a predictor for the psychological method further supports the trustworthiness of our results.

6.9 Potential Applications

In this paper, we find that the personality knowledge in ChatLLMs originates from their base models, and instruction data fine-tuning tends to make the models show more personality. We think this conclusion can help us learn about LLMs and determine the personality of LLMs by controlling their pre-trained data. Additionally, we can design special instruction data to expose the hidden personality traits of LLMs. All of this can help humans train more suitable LLMs.