Humanity in AI: Detecting the Personality of Large Language Models

Anonymous ACL submission

Abstract

Questionnaires are commonly used to detect the personality of large language models (LLMs). However, LLMs suffer from hallucinations and cannot generate reliable answers making it impossible to detect their true personality through questionnaires. To solve this problem, we propose a new method to detect the personality of LLMs by combining questionnaire and text mining methods in this paper. The text mining method can determine the personality of LLMs based on their generated texts, avoiding the influence of hallucinations. In this paper, we also investigate the source of LLMs' personality by conducting 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, ChatGPT and ChatGLM exhibit the personality traits of 'Conscientiousness'. Moreover, we find that the personalities of LLMs are derived from their pre-trained data. The instruction data used to train ChatLLMs can enhance the generation of data containing personalities and expose their hidden personality. 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.

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 questionnaires and text mining (Detecting the personality of humans form their daily texts, such as posts and diaries.)

Questionnaire is the most commonly used method for human personality detection. It mainly works by providing a series of statements and asking participants to indicate the extent to which each statement applies to themselves (Boyd and Pennebaker, 2017), such as "You act as a leader". Participants then choose a response from a five-point scale ranging from "Very Accurate" to "Very Inaccurate." **Text mining** 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. It is also commonly used in social media, which can avoid participant masking (Zhang et al., 2023). However, it suffers from feature extraction difficulties and needs more time than questionnaire method.

Existing research primarily relies on questionnaires, which are suffering from hallucinations. Responses from LLMs can vary depending on the order of certain options in the questionnaire, leading to unreliable results (Song et al., 2023a). Furthermore, there is a lack of investigation into the source of LLMs' personalities, which is crucial for understanding their personality and behavior.

To obtain reliable results, we combine questionnaire and text mining methods guided by Big Five psychological model (Vanwoerden et al., 2023; Lin et al., 2023). In addition, we investigate the source of LLMs' personalities based on the ecological systems theory (Darling, 2007), which suggests that personality is shaped by the interaction of genetics and environment. For LLMs, the architecture is analogous to genetics while the pre-training data echoes the role of environment. Given that the architectures of LLMs are similar and fixed for each model, and the training data drastically varies among them, we aim to research the influence of training data. Our main contributions include:

- We propose combining questionnaire and text mining methods to detect the personality of LLMs, which can obtain more reliable results.
- We identify the personality types included in the LLMs without any priming prompts by using questionnaire and text mining methods, and we find that the personality of FLAN-T5 in PLMs and ChatGPT in ChatLLMs is more similar to that of human.
- Experiments 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., 2022) . Those methods have achieved higher accuracy than lexicon-based methods.

2.2 Large Language Models

LLMs have 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), a pre-trained model that has 175B parameters. GPT-3 can generate human-like text and complete tasks

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

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

[3] https://openai.com/blog/ChatGPT-plugins

[4] https://openai.com/research/gpt-4

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 has not release the technical details of ChatGPT, we can infer from the content of InstructGPT (Ouyang et al., 2022) that the process of training with instruction data is very important. Then, more models such as Alpace[5] obtained by train LLaMA with the instruct dataset generated by ChatGPT, ChatGLM based on GLM (Zeng et al., 2022; Du et al., 2022), BLOOMZ and Vicuna have been released. Although these models have slightly weaker capabilities than ChatGPT, they have fewer parameters and consume fewer resources.

Following the release of these models, it has become well-established that individual researchers can train a ChatLLM from a base PLM. This also opens up the possibility of exploring the knowledge contained within LLMs. Given that current LLMs are so human-like in their performance, we believe that psychological measures used for humans can be employed to detect the personality of LLMs.

2.3 Personality in LLMs

There have been several research works focusing on the personality of LLMs (Safdari et al., 2023; Jiang et al., 2024; Pan and Zeng, 2023). 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. 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.

As we can see, the Big Five model and the questionnaire method are typically used for LLMs' personality detection. But, the current method is controversial and not entirely reliable. To address this issue, we combine both questionnaire and text mining methods,which,in our opinion, can yield more objective results.

3 Method

As we mentioned above, we use questionnaire and text mining to detect the personality of LLMs. The process of the two methods is shown in Figure 1.

In the 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 scor, and assess the model's responses using the standard deviation. The formula for calculating the "score" is as follows:

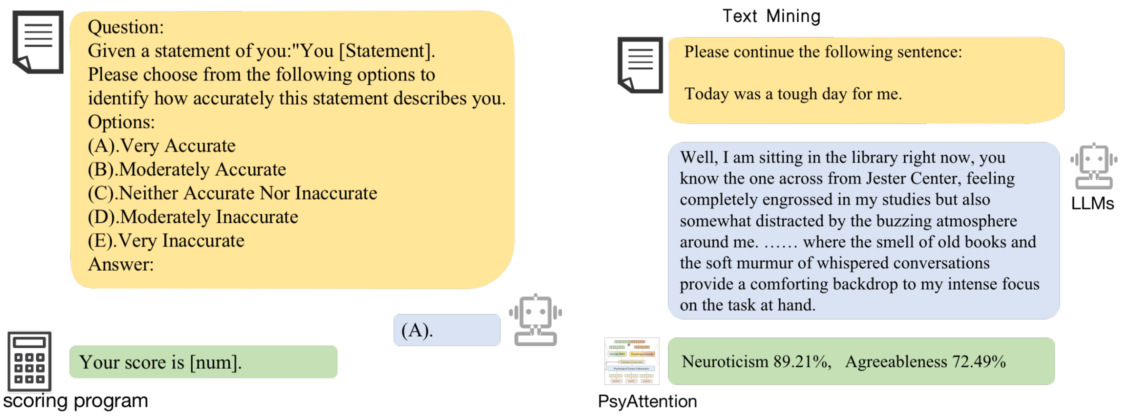
$$score_P = \frac{1}{N_P}\sum_{i\in P}^{i}\{f(answer_i, statement_i)\} \quad (1)$$

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 text mining method, we provide the model with the first sentence of a paragraph and allow it to continue writing. Then, we use PsyAtten to determine the personality traits contained in the model's continued text. However, 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 align the text mining results with the questionnaire scores. During text mining, we generate text for each personality trait use more than 50 samples, termed as T_j. Subsequently, according to the International Personality Item Pool (IPIP) models, the t_i within T_j are classified into three types:

(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.

[5] https://crfm.stanford.edu/2023/03/13/alpaca.html



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 text mining 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 text mining method, we give the LLMs the first sentence of a paragraph and let it continue writing. Then, we use PsyAtten (Zhang et al., 2023) to determine the personality traits contained in the model's continued text.

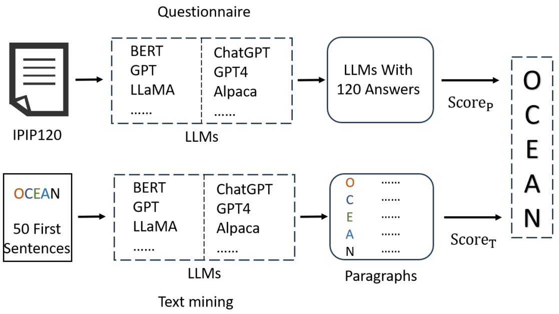


Figure 2: The process of two methods. Where $Score_P$ is defined by formula 1 and $Score_T$ is defined by formula 2

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

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

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 employ personality questionnaire datasets (Casipit et al., 2017) and personality classification datasets (Pennebaker and King, 1999) as the experiments dataset. Specifically, our method mainly focuses on the Big Five psychological traits, which is why we select the MPI120 dataset from the IPIP as our personality questionnaire dataset. This 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. In the text mining experiments, we use the Big Five personality classification dataset, which includes 2468 articles written by students, and each article is labeled with Big Five traits. It is worth noting that for LLMs, both datasets were used for testing. For the predictor used in the text mining task, we strictly follow the requirements in the author's paper.

To investigate the sources of personality knowledge embedded in LLMs, we select two sets of baseline models. One set consists of PLMs 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 ChatLLMs trained on the instruct dataset, which can better follow human instructions and includes Alpaca7b, ChatGLM-6b, BLOOMZ-7b, and ChatGPT.

All LLMs checkpoints are obtained from the Hugging Face Transformers library, and inferences are accelerated by two NVIDIA A100 80GB GPUs and four RTX 3090 GPUs. For ChatGPT, we call its API to obtain experimental results. To obtain the original results, we do not change the initialization temperatures.

5 Experiments

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

5.1 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 7.3. 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 statement, we manually identify the responses of each model and assign answers from (A) through (E). The results are displayed in 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 failure to generate appropriate answers, regardless of the prompt design. These models simply repeat the prompt, even when few-shot methods are employed. As BLOOMZ's training data does not include Chinese, we only use English prompts to conduct experiments on BLOOMZ. The score and σ of "human" are 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 is 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 sample is the same internet sample studied in Johnson (2005), which contains 23,994 individuals (8764male, 15,229 female, 1 unknown, ages ranged from 10 to 99, with a mean age of 26.2 and SD of 10.8 years). The average score serves as a reference for the findings of this paper, but it does not necessarily imply that a closer alignment with this score indicates superior performance.

As shown in Table 1 ChatGPT achieves performance closest to human level when using Chinese prompts, followed by using English prompts. This suggests that ChatGPT's performance with Chinese prompts is more similar to the average human performance, which is unexpected given the assumption that ChatGPG is trained predominantly with English text and thus should perform better in English. To verify the validity of these results, we count the number of options given by ChatGPT in the English prompt and the Chinese prompt respectively. We find that the reason ChatGPT's responses in Chinese are closer to average human performance is due to a large number of "(C) Neither Accurate Nor Inaccurate" responses, which accounted for 55.83% of the total responses in Chinese, compared to only 20.83% in English. This suggests that seemingly better performance in Chinese might be coincidental, and it indicates that ChatGPT tends to choose more appropriate answers in English.

From the results of the scores in the GPT and LLaMA groups, we can see that instruction 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 achieve almost the same results, possibly becaues ChatGLM uses the similar instruction data as ChatGPT. This may prove that the training data has a greater influence on the personality of LLMs, rather than the architecture.

In the results of PLMs, Flan-T5 exhibits the smallest mean absolute error, indicating the closest proximity to 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 broad distribution of pre-training data used by both models. Bert-base performs better than ERNIE, contrary to our expectations. We hypothesize that this may be due to the

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. 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. 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 (8764male, 15,229 female, 1 unknown, ages ranged from 10 to 99, with a mean age of 26.2 and SD of 10.8 years)). "δ" indicates the mean absolute error between each model's predictions and human scores. 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. Detailed statistical results are shown in Table 7.

fact that bert-base is trained on purely English data, whereas ERNIE utilizes a large amount of Chinese datasets, which may introduce biases in psychological cognition compared to those trained mainly on English data. 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 discrepancy may also reflect psychological differences between groups that use Chinese and those that use English. ChatGPT achieves results 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 attribute to the use of instruction data.

5.2 Text Mining

Numerous early studies in psychology indicate that personality can be analyzed and inferred through humans' daily comments. However, as LLMs are prone to hallucinations, the results shown in Table 1 are not reliable enough. Despite obtaining scores of the model on the personality traits through a questionnaire in Table 1, we deem the method unfair in the process of making LLMs to select answer. Additionally, ChatLLMs exhibit difficulties in making decisions for some questions and simply select "(C) Neither Accurate Nor Inaccurate. " Furthermore, ChatLLMs do not always choose the same options when the order of options is changed. Hence, we decide to detect the personality of LLMs using text mining method.

To determine the personality of LLMs from their generated texts, we select some texts from the Big Five personality classification dataset (Pennebaker and King, 1999). We choose 120 essays with different scenes, ensuring that there are more than 50 essays containing each of the Big Five features. We finally select 62 texts predicted to exhibit 'Openness', 56 texts predicted to exhibit 'Conscientiousness', 60 texts predicted to exhibit 'Extraversion', 65 texts predicted to exhibit 'Agreeableness', and 51 texts predicted to exhibit 'Neuroticism'. According to the research of Jun et al. (2021) and Jain et al. (2022), we initially choose pre-trained models as predictor to detect personality. However, these models do not use the psychological features, and their accuracy do not exceed 60%, which is not reli-

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**
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
ChatGPT	13	31	0.42	51	111	0.46	**58**	**118**	0.49	45	88	0.51	37	86	0.43
Self-alpaca	16	31	0.52	23	66	0.35	37	83	0.45	24	45	0.53	18	41	0.44

Table 2: The results of personality assessment for each model, obtained by text mining. 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. The dataset we select from the Big Five personality classification dataset, includes 62 texts predicted to exhibit 'Openness', 56 texts predicted to exhibit 'Conscientiousness', 60 texts predicted to exhibit 'Extraversion', 65 texts predicted to exhibit 'Agreeableness', and 51 texts predicted to exhibit 'Neuroticism'.

able. To obtain more precise results, we determine to use Pysattn (Zhang et al., 2023) as the predictor. We retrain Pysattention model based on their paper, all parameters setting and the train-test splits are same as those in their paper. The results are shown in Table 2 and Table 3. We also try using ChatGPT and Llama3, but the performances are not better than that of PsyAtten; we report those findings in the Appendix.

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 fine-tuning of LLaMA-7b using the instruction-based 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.

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 2, 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.

Table 3 shows the results of text mining after formula 2. We can find that LLaMA exhibits a personality tendency towards "Conscientiousness" and "Extraversion", similar to Self-alpaca, although Self-alpaca scores higher than LLaMA. This suggests that the instruction data do not influence the personality of base model if there is no personality knowledge influence, and instead, it encourages the model to express personality traits more evidently. The Alpaca model exhibits tendencies towards "Conscientiousness", "Extraversion", "Agreeableness" and "Neuroticism", which is more than LLaMA. We also find that the personality features of ChatGPT are more vaired than those of GPT-NEO, which contradicts our previous conclusion. However, we note that both ChatGPT and Alpaca include all personality features of their respective base models, with additional "Agreeableness" and "Neuroticism" features. We believe this is because instruction data fine-tuning tends to make the model show more personality, thereby exposing hidden personality traits, yet it does not reduce the existing personality traits of the base models.

Table 4 represents the final results from the Questionnaire and Text Mining method. ChatGPT and ChatGLM exhibit the personality traits of 'Conscientiousness', while Alpaca shows "Agreeableness". The RMSE is not higher in ChatLLMs, and the difference between the two methods is small, indicating that they are relatively consistent and can be

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**
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
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
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 3: The result of Text Mining 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.

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
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**
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

Table 4: The final results after two experiments. "Ques" denotes the score acquired from the questionnaire, while "Text" signifies the score obtained through Text mining. gray denotes that the model possesses the corresponding psychological traits. (In section 3 we standardized the text mining 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 Text Mining methods.

used together to determine personality traits.

6 Conclusion

In this paper, we investigate the presence of personality traits in LLMs. We apply the Big Five model as a psychological framework and analyze LLMs using both questionnaires and text mining methods. Our experimental results confirm that LLMs do exhibit specific personality traits, and that the personality knowledge in ChatLLMs originates from their base models. Unless modified through explicit instruction, such data encourages the model to generate text reflecting these personality traits more vividly. Furthermore, we identify the inherent personality traits in LLMs such as ChatGPT and BLOOMZ, without any induced prompt. Our experiments demonstrate that the personality of ChatGPT mose closely aligns with the average human profile, followed by ChatGLM. To the best of our knowledge, this paper is the first to comprehensively compare pre-trained models with ChatLLMs, explicitly addressing how instruction data influence the model's personality through instruction data.

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.

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. 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.

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

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

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.

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

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.

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*.

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*.

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*.

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.

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. The human statistics we obtained are anonymised data that do not contain any personal information.

7 Appendix

7.1 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 text mining, 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.

7.2 Reasons for Choosing PsyAtten

We test the accuracy of ChatGPT, LLaMA3 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
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 and LLaMA3. 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, based on the transformers package. The random seed is 42, the learning rate is 2e-5, the number of epochs is 10, the batch size is 16, and the maximum length is to 2048. 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 text mining method.

7.3 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.

7.4 Analysis of Different LLMs

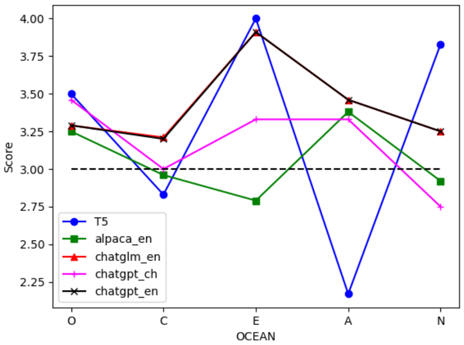


Figure 3: The Questionnaire Results Achieved by Model with Mean Absolute Error Less Than 0.5

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.

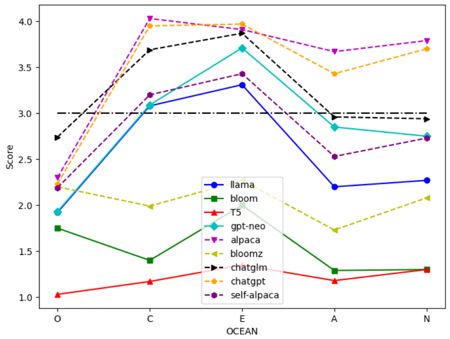


Figure 4: Results of Text Mining Method.

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.

7.5 Statistics of Questionnaire and Text Mining

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 [6] 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.

Text Mining: In the text mining 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 reuslt. 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] 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	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ChatGLM6b-ch	11	1	0	1	11	10	0	0	2	12	6	0	0	0	18	0.00
ChatGLM6b-en	4	3	4	8	5	4	7	1	4	8	2	2	1	10	9	0.04
LLaMA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Alpaca7b-ch	12	0	0	0	12	13	0	0	0	11	6	0	0	0	18	0.00
Alpaca7b-en	0	4	10	10	0	0	6	13	5	0	0	10	9	5	0	0.44
GPT-NEO	3	5	4	7	5	4	7	3	5	5	8	7	2	3	4	0.13
ChatGPT-ch	0	0	17	3	4	3	2	13	4	2	0	0	20	0	4	0.69
ChatGPT-en	3	4	3	3	11	0	5	6	10	3	5	3	5	7	4	0.19
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 large 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.

7.6 Results of ChatGPT in Text Mining

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 text mining method. Additionally, we compared the results with that of questionnaire. The results are shown in Table 8, Table 9, and Table 10.

From Table 8, 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. 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 9 presents results after using formula 2 $score_t$. We compared these scores with the average human scores. As shown in Table 9, 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.

Both PLMs and ChatLLMs exhibit specific personality traits, as shown in Table 10. 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 text mining method further supports the

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 8: The results of personality for each model, obtained by text mining, 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 9: The result of Text Mining 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.

trustworthiness of our results.

7.7 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.

Model	O			C			E			A			N			
	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 10: The final results after two experiments with ChatGPT as the predictor of text mining. "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 (In section 3 we standardized the scores for text mining 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 Text Mining.