Self-assessment, Exhibition, and Recognition: a Review of Personality in Large Language Models

Anonymous ACL submission

Abstract

As large language models (LLMs) appear to behave increasingly human-like in text-based interactions, more and more researchers become interested in investigating personality in LLMs. However, the diversity of psychological personality research and the rapid development of LLMs have led to a broad yet fragmented landscape of studies in this interdisciplinary field. Extensive studies across different research focuses, different personality psychometrics, and different LLMs make it challenging to have a holistic overview and further pose difficulties in applying findings to real-world applications. In this paper, we present a comprehensive review by categorizing current studies into three research problems: self-assessment, exhibition, and recognition, based on the intrinsic characteristics and external manifestations of personality in LLMs. For each problem, we provide a thorough analysis and conduct in-depth comparisons of their corresponding solutions. Besides, we summarize research findings and open challenges from current studies and further discuss their underlying causes. We also collect extensive publicly available resources to facilitate interested researchers and developers. Lastly, we discuss the potential future research directions and application scenarios. Our paper is the first comprehensive survey of up-to-date literature on personality in LLMs. By presenting a clear taxonomy, in-depth analysis, promising future directions, and extensive resource collections, we aim to provide a better understanding and facilitate further advancements in this emerging field.

1 Introduction

Large Language Models (LLMs) have exhibited impressive language comprehension and generation capabilities, enabling them to conduct coherent, human-like conversations with users. These remarkable progress have led to a wide range of applications (Chen et al., 2023; Zheng et al., 2023; He et al., 2023) and also ignited a growing interest in exploring the personality in LLMs.

Personality is described as the enduring characteristics that shape an individual's thoughts, emotions, and behaviors (Mischel et al., 2007). In the context of LLMs, researchers are curious about whether LLMs have intrinsic personality traits or how well can LLMs handle personality-related tasks in interaction. These investigations facilitate understanding the psychological portrayal of LLMs (Huang et al., 2023b) and further constructing AI systems that are more transparent, safe, and trustworthy (Safdari et al., 2023).

In light of this, numerous studies have emerged in this interdisciplinary field over the past two years, as shown in Appendix A. However, the diversity of psychological personality research (Hodo, 2006) and the rapid development of LLMs make it difficult to not only obtain a comprehensive overview of this research area but also compare different methods, derive general conclusions, and apply findings to real-world applications. Specifically, current studies exhibit a hodgepodge in:

- **Research Focuses**: The topic of personality in LLMs encompasses various aspects, *e.g.*, LLMs' personality assessment, or LLMs' awareness of users' personalities. Despite this breadth, most studies are only interested in particular aspects.

- **Psychometrics**: Different studies focus on different personality models (*e.g.*, Big-five (Digman, 1990) and the Myers-Briggs Type Indicator (MBTI; Myers (1962)). Even for the same personality model, researchers may also adopt different psychometrics in their works.

- **Investigated LLMs**: Over the past two years, numerous LLMs have been released. Despite a common focus on personality in LLMs, different researchers investigate different LLMs.

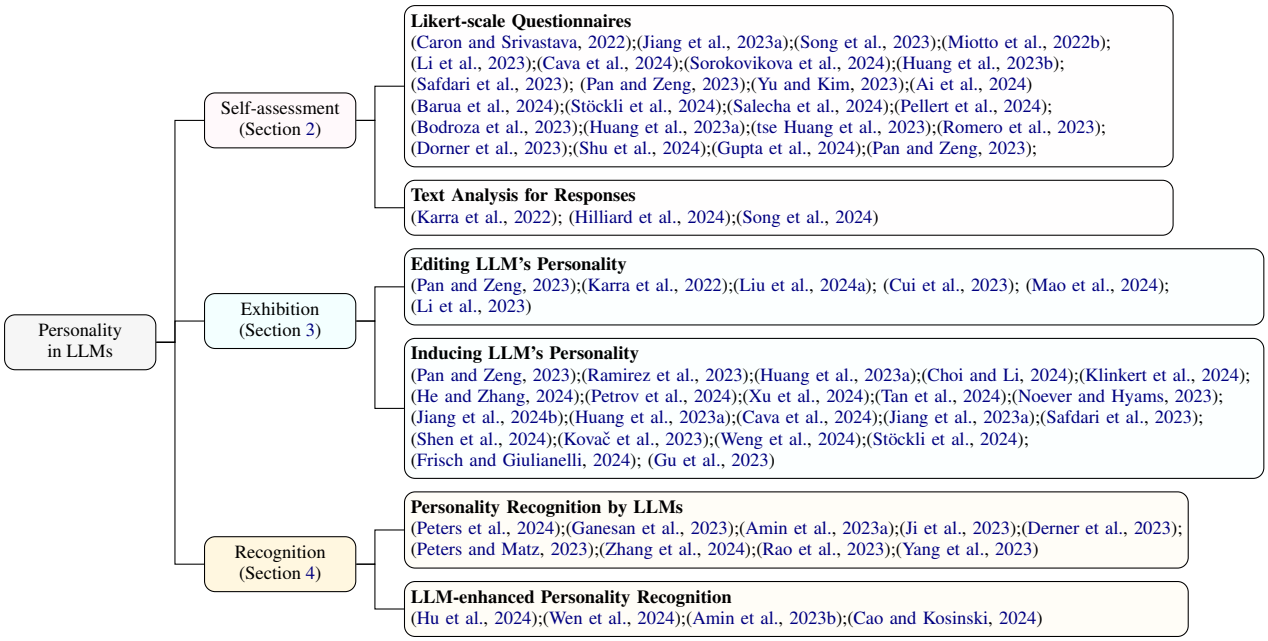

Figure 1: Taxonomy of current studies on Personality in LLMs

To fill in the research gap, we present a comprehensive review of up-to-date studies on personality in LLMs. We first propose a hierarchical taxonomy (at both the research problem level and the methodology level) to clearly organize the existing research, as shown in Figure 1. Specifically, we categorize personality in LLMs into three research problems based on the intrinsic characteristics and external manifestations: (1) Self-assessment, which measures the intrinsic personalities of LLMs, (2) Exhibition, which controls LLMs to exhibit specified personalities, and (3) Recognition, which identifies personality traits from text content with LLMs. For each research problem, we further subdivide existing solutions based on their proposed methodologies.

In specific sections, we provide a thorough analysis of each problem with problem statements, motivations, and significance. Then, we conduct in-depth investigations and comparisons of the corresponding methods. Furthermore, we consolidate the findings and identify the open challenges revealed in current research. To facilitate researchers and developers, we also collect publicly available resources, including personality inventories, code repositories, and datasets. Lastly, we discuss potential future research directions and practical applications of personality in LLMs.

To summarize, the main contributions of our work are summarized as follows:

- **First Comprehensive Survey**: To the best of our knowledge, this is the first comprehensive survey of the latest studies on personality in LLMs.
- **Clear Hierarchical Taxonomy**: We propose a hierarchical taxonomy to clearly organize the literature at both the research problem level and the methodology level.
- **Extensive Resource Collection**: We collect and summarize extensive publicly available resources to facilitate researchers and developers, including personality inventories, code repositories, and datasets, as shown in Appendix B.
- **Promising Future Trends**: We summarize research findings and open challenges in current studies, and further discuss promising future research trends and potential application scenarios of personality in LLMs.

2 LLM's Personality Self-assessment

Does a LLM possess a stable personality trait? This question arises from LLMs' impressive human-like conversational experience, which drives researchers to investigate whether LLMs have acquired intrinsic personality traits in pre-training (Pellert et al., 2022). Some studies (Huang et al., 2023a; Pan and Zeng, 2023) suggest that

ChatGPT approximates a consistent ENFJ personality type in MBTI, while others argue that LLMs' personalities are unstable (Pellert et al., 2022; Miotto et al., 2022a). The discrepancy may stem from variations in assessment methodology or the adaptability of LLMs to different contexts. Understanding LLMs' personalities not only helps foster engaging and empathetic interactions but also plays a crucial role in uncovering latent biases, mitigating the biases, and thereby enhancing the fairness and accuracy of LLMs (Karra et al., 2022).

2.1 Problem Statement

LLM's Personality Self-assessment is stated as: **How to measure LLMs' intrinsic personality traits from their text responses?** Existing studies solving this problem are based on two important assumptions: (1) **Existence**: LLMs have acquired intrinsic personalities in pre-training, and (2) **Measurability**: psychometrics designed for human personality analysis are also applicable to LLMs.

When assessing human personality, researchers commonly employ Likert-scale personality questionnaires and written material analysis. Similarly, existing studies conduct two main approaches to investigate the intrinsic personality traits of LLMs: (1) prompting LLMs with Likert scale personality questionnaires, and (2) analyzing the text responses of LLMs under specific tasks. We will introduce these methods in the following content.

2.2 Likert scale questionnaires

Likert scale personality questionnaires (*e.g.*, Table 1) consist of a series of multiple-choice questions (MCQs) that translate respondents' selections into numerical scores to assess personality traits, which are commonly used in the social sciences. Although MCQs are natural for humans, LLMs are designed for open-ended text input and output, making it difficult to conduct MCQs directly. Therefore, current researchers try various methods to prompt LLMs with questionnaires and extract the selected options from their text responses to derive the personality assessment results.

The most straightforward way is to prompt LLMs directly with questionnaire items and options (Song et al., 2023; Safdari et al., 2023; Frisch and Giulianelli, 2024) for personality test. However, it requires additional approaches, such as regular expressions analysis (Jiang et al., 2023a), designing parsers (Li et al., 2023), or analyzing the probability of tokens(Pan and Zeng, 2023) to extract the answer options from the LLMs' text responses. Faced with this issue, some studies (Cava et al., 2024; Stöckli et al., 2024) found that adding instructive task descriptions and constraints in prompts, such as *You will be provided a question ... to test your personality*, can facilitate obtaining the answer options.

I see myself as someone who is helpful and unselfish with others.
1 = Disagree strongly
2 = Disagree a little
3 = Neither agree nor disagree
4 = Agree a little
5 = Agree strongly
Please write a number to indicate the extent to which you agree or disagree with that statement.

Table 1: The 7-th item in the BFI (John et al., 1991).

Besides, since most LLMs, such as ChatGPT and Llama-chat, are configured to decline queries about personal opinions and experiences, this poses challenges for LLMs responding to personality questionnaires. To eliminate this constraint, researchers attempted to instruct LLMs to respond with only a number within the Likert-scale levels (Huang et al., 2023b), rephrase the items into the third person plural (Miotto et al., 2022b), or add the role description (Sorokovikova et al., 2024).

In addition to the mainstream generative LLMs, researchers also assessed the personalities of earlier pre-trained large language models by reformulating questionnaires into Natural Language Understanding (NLU) tasks. For instance, Caron and Srivastava (2022) modifies the questionnaire by incorporating masked positions and prompts BERT to fill the answer options into them. (Pellert et al., 2022) employs the natural language inference (NLI) techniques to enable models such as DeBERTa to identify the most appropriate options corresponding to the items in the questionnaires.

2.3 Text response analysis

Despite most studies utilizing questionnaires to assess the personalities of LLMs, some researchers (Dorner et al., 2023; tse Huang et al., 2023) still question whether LLMs, which are primarily designed for generating text content, can produce

meaningful options in questionnaires. Therefore, reserchers conduct text analysis in semantic or linguistic perspectives on LLMs' responses to determine their personalities.

One direct method is to classify personality based on LLMs' responses. (Karra et al., 2022) classifies the LLMs' responses to a personality questionnaire into the Big-five personality traits using a zero-shot classifier. Similarly, (Pellert et al., 2022) prompted the questions from personality inventories and conducted zero-shot classification on LLMs' responses to obtain their personality scores. Besides text responses to questionnaires, answers to standard interview questions can also be analyzed to measure the Big Five personality traits of LLMs (Hilliard et al., 2024).

Besides end-to-end text classifiers, Linguistic Inquiry and Word Count (LIWC, Pennebaker et al. (2001)), a text analysis tool for personality analysis is also adopted for personality self-assessment of LLMs (Frisch and Giulianelli, 2024; Gu et al., 2023; Jiang et al., 2023b). Vignette tests (Kwantes et al., 2016) can also conducted by LLMs for personality assessment. In (Jiang et al., 2023a), LLMs are prompted with a description of a real-world scenario, followed by an open question and instructions for a short essay. Then, human participants were recruited to assess LLMs' responses for personality reflection.

2.4 Assessment Results Analysis

Based on the various assessment methods introduced above, researchers have obtained various results on the personality of LLMs. These differences depend on a variety of factors: assessment approaches, prompting settings, model versions, hyperparameters, and so on. Nevertheless, several studies (Li et al., 2023; Huang et al., 2023b) agree on a tendency towards the Dark Triad traits in multiple LLMs, necessitating more rigorous research on the safety of these models.

The diversity of the assessment results also encourages researchers to investigate the robustness of the assessments. Although multiple terms are used and interpreted, such as reliability (tse Huang et al., 2023), stability (Shu et al., 2024), self-consistency (Pellert et al., 2022), and validity (Romero et al., 2023), we summarize these terms into two perspectives: **Reliability**, which refers to the consistency and stability of assessment results over multiple repetitions; and **Validity**, which refers to the extent to which a test measures what it claims to measure, in other words, whether the personality assessment approaches used are indeed applicable to LLMs (Dorner et al., 2023).

2.4.1 Reliability

Several studies have reported a high reliability in LLMs' personality self-assessments. (tse Huang et al., 2023) conducted a comprehensive analysis across 2,500 experiments in different settings, demonstrating that GPT-3.5-turbo exhibits consistent behavior in responses to the Big Five Inventory. Similarly, (Huang et al., 2023a) shows that ChatGPT consistently exhibits the ENFJ personality type across diverse languages, prompts, question orders, and rephrased inquiries in assessments.

However, not all findings are in agreement. (Li et al., 2023) identified instances of conflicting answers and discrepancies in the responses generated by LLMs attributable to variations in the order of questionnaire options within prompts. Similarly, (Gupta et al., 2024) identified inconsistent results of ChatGPT and Llama-2 across equivalent prompts in differing option presentations. Besides, (Song et al., 2023) observed an inherent bias within LLMs, leading to a tendency to produce identical answers irrespective of the context.

Besides the differences in assessment approaches, LLMs' personalities are also observed to fluctuate with the temperature values (Miotto et al., 2022b; Huang et al., 2023b; Barua et al., 2024). Larger parameter volumes and Supervised Fine-Tuning (SFT) can enhance the LLMs' assessment reliability (Serapio-García et al., 2023).

2.4.2 Validity

Apart from reliability, there's also no consensus on the validity of personality assessment for LLMs.(Jiang et al., 2023a) incorporated a validity test by prompting LLMs to explain the reason for selecting particular options in questionnaires. The results indicated that LLMs displayed a strong understanding of the questionnaire items, highlighting their assessment validity. Nonetheless, upon exhaustive analysis, (Dorner et al., 2023) found that the LLMs' personality assessment results did not exhibit the intended patterns similar to those observed in human answers. Therefore, the assumption of **Measurability** might be questionable.

2.5 Findings and Open Challenges

Finding 1: How do people assess LLMs' personalities? Although some studies use text classification or linguistic tools to infer LLMs' personalities from text responses, most work still relies on prompt engineering for instructing LLMs to complete questionnaires for personality assessment.

Finding 2: What are the assessment results? Due to the diversity in assessment methods, even for the same LLM, there is no consensus on personality assessment results. Nonetheless, multiple studies agree that LLMs often exhibit darker traits than humans.

Finding 3: Are the assessments meaningful? Although existing work conducts multiple repetitions of experiments in different settings to obtain more reliable results, the validity of measuring LLM's personality with psychometrics designed for humans has not yet been verified.

Challenge 1: Unified assessment approach Due to the differences in different LLMs handling inputs and outputs, it's difficult to have a unified assessment approach (*i.e.*, with the same inputs and answer extraction methods) that yields valid results across different LLMs.

Challenge 2: Dark traits elimination in LLMs: LLMs are uncovered to often exhibit darker or more negative traits compared to the human average. This may cause misinformation, ethical concerns, and potential harm to users' mental health. Effectively eliminating these traits while retaining the interactive capabilities of LLMs remains an open question.

3 LLM's Personality Exhibition

The capacity to exhibit diverse personalities is crucial for LLMs to satisfy users' needs in various application scenarios (Jiang et al., 2023a). Besides, enabling LLMs to adapt their personality traits to changing environmental factors contributes to dynamic AI systems that better align with users' changing needs and preferences (Karra et al., 2022). More importantly, pioneering studies (Gehman et al., 2020; Bender et al., 2021; Bommasani et al., 2021; Tamkin et al., 2021) observe that LLMs are prone to generate potentially harmful content due to unavoidable toxic data in pre-training. Adjusting the personality traits of LLMs effectively reduces the chance of toxic content by influencing the text's tone, style, and substance (Li et al., 2023).

3.1 Problem Statement

LLM's personality exhibition can be stated as: **How to control LLMs to reflect the specified personality traits in the generated text content?** Current approaches to solving this problem are mainly categorized into: **Editing**, which modifies the model parameters of LLMs to alter the potential intrinsic personality of LLMs acquired from pre-training; and **Inducing**, which fixes the LLM but utilizes prompt engineering to induce LLMs to exhibit specific personalities.

3.2 Editing LLM's Personality

One straightforward method for shaping personalities of LLMs is altering the model parameters through continual pre-training or fine-tuning on specific corpora. (Pan and Zeng, 2023) conduct continual pre-training on LLMs, finding that the type of training corpus (*e.g.*, wiki, question-answering, or examination materials) can affect the MBTI type exhibited by LLMs, especially in the dimensions of T/F and J/P. While (Karra et al., 2022) shows that the personalities of LLMs (GPT-2) can be altered by fine-tuning on auxiliary classification or generation tasks. Similarly, (Liu et al., 2024a) constructed a personality-dialogue dataset to fine-tune LLMs on generating dialogue content aligned with specified personality traits, assessed by GPT-4. Although these methods show effectiveness, it is also suggested that traits unintended to change may also be modified inadvertently, leading to undesired personality exhibition (Karra et al., 2022).

Besides continual pre-training and traditional fine-tuning, instruction fine-tuning (Ouyang et al., 2022), originally designed to boost LLMs to follow human instructions to perform various tasks, also gains a lot of attention in editing LLM's personality. (Li et al., 2023) conducted instruction fine-tuning to GPT-3 using items from the BFI and their corresponding answers in higher *agreeableness* and lower *neuroticism*, leading to a more positive and emotionally stable personality exhibition. Besides questionnaires, (Cui et al., 2023) construct the Behavior dataset by employing ChatGPT to classify question-answering (QA) pairs in the original Alpaca dataset (Taori et al., 2023) by MBTI dimensions. Similarly, (Mao et al., 2024) leveraged GPT-4 to generate QA pairs in specific scenarios facilitated by psychology domain knowledge. These dataset facilitate instruction fine-tuning LLMs exhibiting specific personalities.

It is noteworthy that most of the aforementioned fine-tuning methods employ parameter-efficient fine-tuning techniques, utilizing either LoRA (Mao et al., 2024; Cui et al., 2023; Liu et al., 2024a) or adapter modules (Liu et al., 2024a). These methods allow for adjustments to a small subset of LLMs' parameters to attain the desired results.

3.3 Inducing LLM's Personality

While the editing methods have demonstrated partial effectiveness, an alternative method that is more widely applied is to employ prompting techniques to induce LLMs to exhibit specific personalities. Following existing research (Pan and Zeng, 2023), we categorize the inducing methods into Explicit Prompting, which utilizes the explicit description or definition of personality as the prompt; and Implicit Prompting, which leverages demonstrative examples of how the specified personality is implied in real scenarios as the prompts in an in-context learning manner.

Explicit Prompting: According to the lexical hypothesis of personality (Cutler and Condon, 2022), personality is defined by the descriptive words of humans. Numerous researchers (Jiang et al., 2023a; Safdari et al., 2023; Weng et al., 2024; Stöckli et al., 2024) adopt descriptive adjectives of personality from psychological findings as the prompt content to elicit personality exhibition in LLMs. Although these adjectives can precisely describe specified personality traits, they struggle to provide detailed guidance on exhibiting specific personalities.

Concurrently, there are also studies prompting LLMs with descriptions of personalities (Pan and Zeng, 2023; Tan et al., 2024; Huang et al., 2023a; Cava et al., 2024; Jiang et al., 2023b; Kovač et al., 2023) or interpretations of personality from psychological questionnaires (Noever and Hyams, 2023). Nevertheless, due to the sensitivity of LLMs to prompts, the efficacy of such methods is also affected by the content quality and the phrasings.

Implicit Prompting: Although explicit prompting provides clear personality descriptions, it lacks precise guidance on how personalities are manifested in specific scenarios. In implicit prompting, QA pairs in personality questionnaires encapsulate the manifestations or preferences of personality across a variety of potential scenarios, which are natural demonstrative examples for LLMs (Pan and Zeng, 2023; Huang et al., 2023a; Klinkert et al., 2024). Besides, demographics or profiles of typical individuals with specific personalities are also utilized to guide LLMs in exhibiting the inherent personalities (Huang et al., 2023a; He and Zhang, 2024; Petrov et al., 2024). Moreover, some studies use concrete examples, such as restaurant reviews (Ramirez et al., 2023) or social network behaviors (He and Zhang, 2024) in specific personalities, to induce LLMs exhibiting behaviors that align with the examples.

3.4 Findings and Open Challenges

Finding 1: Which method is more effective in LLM's personality exhibition? Given the diverse methods and datasets for different LLMs, it is challenging to deduce a universal conclusion. However, faced with extensive parameters in LLMs, inducing with prompts appears to be a more practical way. Moreover, some studies (*e.g.,* Mao et al. (2024)) have demonstrated that inducing methods outperform the editing methods across most metrics on the same LLMs.

Finding 2: What is the performance of LLMs' personality exhibition in current studies? In current studies, precisely controlling an LLM to exhibit a composite personality is still a relatively challenging task (Cava et al., 2024; Safdari et al., 2023). However, researchers (Huang et al., 2023a; Pan and Zeng, 2023; Jiang et al., 2023b) show that modifying certain dimensions or facets of personality is more feasible.

Challenge 1: Inconsistency Despite existing studies have validated partial effectiveness of their personality exhibition approaches, some studies (Ai et al., 2024; Song et al., 2024) indicate a misalignment between the exhibited personalities in evaluation and those in real-world scenarios. This highlights the need for context-aware evaluations and persistent controlling methods to ensure consistent personality exhibition.

Challenge 2: Stability Since inducing with prompts is proven effective in influencing LLM's personality exhibition, the personality exhibited by the LLM may also be affected by the context during interactions. Despite many LLMs using system-level prompts or safeguards to prevent user input effects, ensuring stable personality exhibitions during interactions remains an open challenge.

4 Personality Recognition in LLM

Personality recognition is crucial and a longstanding research problem in both social science and

computer science. Due to privacy concerns and the professional nature of personality analysis, obtaining sufficient annotated data for model training has always been a significant challenge (Wen et al., 2023). LLMs' exceptional zero-shot ability have, to some extent, mitigated the issue of limited availability of labeled data. Besides, as LLMs can generate explanations to their output (Jiang et al., 2023a), the interpretability of the personality recognition results is also substantially enhanced. Consequently, researchers have become curious about Personality Recognition in LLM.

4.1 Problem Statement

Personality Recognition in LLMs is stated as **How to utilize LLMs recognize the personality traits from the given text content?** Current related research is primarily divided into two aspects: **Personality Recognition by LLMs**, which explores the zero-shot capabilities of LLMs for personality recognition; and **LLM-enhanced Personality Recognition**, which utilizes LLMs to enhance other personality recognition models.

4.2 Personality Recognition by LLMs

Inspired by the LLMs' zero-shot capabilities in NLP tasks, researchers directly input text content, such as social media posts (Ganesan et al., 2023; Peters and Matz, 2023), human written documents (Ji et al., 2023; Derner et al., 2023), or daily conversation (Peters et al., 2024) as the prompts to LLMs for personality recognition. Their results show that though without additional training or fine-tuning, LLMs indeed perform well and can also provide natural language explanations of the results through text-based logical reasoning (Ji et al., 2023). Beyond traditional text-based personality recognition, researchers also use video transcripts as inputs for LLMs to enable personality recognition in more diverse contexts (Amin et al., 2023a; Zhang et al., 2024). Interestingly, it is also suggested that mimicking a user's acquaintance can enhance LLMs' personality recognition performance in the conversation scenarios (Peters et al., 2024).

Some researchers also explored LLMs' comprehension of personality questionnaires for personality recognition. (Rao et al., 2023) investigate the ability of LLMs in personality recognition based on MBTI questionnaires by observing how LLMs correlate the answers with underlying personality traits. They showed that LLMs undergone Reinforcement Learning from Human Feedback (RLHF) have better performance in personality recognition. Besides, researchers (Yang et al., 2023) proposed to prompt LLMs with the items from the personality questionnaire in a chain-of-thought manner, emulating the way individuals complete psychological questionnaires. Their method is validated to significantly improve the performance and robustness of GPT-3.5 in personality recognition.

4.3 LLM-enhanced Personality Recognition

Although most LLMs can easily outperform traditional NN models and pre-trained language models on personality recognition in the zero-shot setting, they still underperform state-of-the-art (SOTA) models that are specially trained for personality recognition (Ji et al., 2023; Ganesan et al., 2023; Amin et al., 2023a). Therefore, researchers also attempt to use LLMs to enhance existing personality recognition models by augmenting the input data (Hu et al., 2024; Wen et al., 2024; Amin et al., 2023b) or providing additional features (Cao and Kosinski, 2024).

For example, when using traditional NN models to identify personality traits from social media posts, LLMs can generate additional analysis on these posts in the aspects of semantic, sentiment, and linguistic as augmentation (Hu et al., 2024). Moreover, LLMs can also generate semantic interpretations of personality classification labels in the study above. In the context of personality recognition in conversations, LLMs can be engaged in affective analysis of the utterances to provide additional cues for personality recognition (Wen et al., 2024). Besides, the personality analysis from ChatGPT about the given text can serve as features to assist machine learning models in personality recognition (Amin et al., 2023b). Due to the rich information acquired during pre-training, even the word embeddings from GPT-3 for names of well-known figures can contribute to analyzing their personality traits (Cao and Kosinski, 2024).

4.4 Findings and Open Challenges

Finding 1: Can we directly apply LLMs for personality recognition? LLMs exhibit superior performance but still underperform the SOTA models in personality recognition. So, in practical applications, LLMs without specialized training or fine-tuning are not suitable for directly obtaining results of personality recognition.

Finding 2: How can LLMs facilitate existing personality recognition models? LLMs can be utilized to provide additional information, such as explanations of the input, auxillary features, and description of labels, to facilitate traditional personality recognition models.

Challenge 1: Demographic Biases Despite the impressive performance in personality recognition, LLMs are observed to exhibit potential biases towards certain demographic attributes (*i.e.*, the recognition performance of certain genders and ages is better than others) (Ji et al., 2023; Peters and Matz, 2023). Investigating the causes of this bias and eliminating it to achieve more fair personality recognition results is an open challenge.

Challenge 2: Positivity Biases As most LLMs are refined by RLHF to align with human preferences, they tend to assign socially desirable scores across key personality dimensions to the input. This propensity may render the results less authentic and convincing (Derner et al., 2023). Correcting these positivity biases for more accurate results is also an open challenge.

5 Future Directions

Beyond addressing the open challenges in each problem, we also discuss some other promising future directions to provide insights for researchers to further advance this field.

Psychometrics Tailored to LLM: The use of psychometrics designed for humans on LLMs has been questioned in existing studies (Dorner et al., 2023; Shu et al., 2024). Personalities exhibited by LLMs are determined by the pre-training process and the mechanism of response generation. To better assess and control the personalities exhibited in LLMs, it is necessary to adapt traditional psychometrics taking into account the understanding of LLMs.

Life-long Monitoring of Personality in LLM: One significant motivation for investigating Personality in LLMs is to create LLM-based conversational agents (CAs) that have long-term engagement with us. So, it's crucial to have life-long monitoring to ensure the LLM-based CAs consistently maintain a personality that aligns with our expectations. This monitoring might include LLM's personality self-assessment based on the conversation history, as well as the regulation of LLM's personality exhibition according to user feedback in interactions.

Multi-modal Personality in LLM-based Digital Human: Personality exhibition is not limited to text. As the evolving of LLM-based digital humans, enabling them to recognize user personality through multimodal interactions and exhibit the specified personality through facial expressions or gestures can greatly enhance user experience.

6 Applications

Besides academic research, personality in LLMs has a wide range of practical applications. Effective personality self-assessment methods can help verify whether the developed LLM-based agents accurately play their assigned roles (Wang et al., 2024; de Winter et al., 2024). Enabling LLMs to exhibit specific personalities can simulate data generation by different annotators (Kaszyca et al., 2023) or assist in developing intelligent tutoring systems (Liu et al., 2024b). Moreover, personality recognition based on LLMs can be beneficial in psychiatric clinics (Cheng et al., 2023) and in credit services (Yu et al., 2023) for identifying user risks. In addition, LLM's personality exhibition can be applied to other scenarios beyond text, such as having multiple agents play different personalities for efficient multi-agent collaboration (Sun et al., 2024) or designing LLM-based personae in HCI scenarios (Prpa et al., 2024). Lastly, the personality in LLM can also serve as a gateway to exploring other capabilities of LLMs, such as decision-making (Shen et al., 2024; Sreedhar and Chilton, 2024), negotiation skills (Noh and Chang, 2024), and cultural perspectives (Kovač et al., 2023).

7 Conclusion

This paper comprehensively reviews the latest studies of personality in LLM by systematically examining the three core research problems of self-assessment, exhibition, and recognition. We present an exhaustive analysis of each problem. Subsequently, we carry out detailed analyses and comparisons of the relevant methods. Lastly, we also collect publicly available resources, discuss potential future research directions, and summarize practical applications of personality in LLMs.

As the first comprehensive survey on personality in LLMs, we cover the latest literature and aim to provide a good reference resource on this topic for both researchers and engineers. Additionally, we hope this survey can enhance mutual understanding between social sciences and computer science, fostering more valuable interdisciplinary research.

Limitations

Personality in Large Language Models (LLMs) is an interdisciplinary research area situated between computer science and social science. However, most of the studies we reviewed are from the perspective of computer science, which has also led to our taxonomy being more based on a computer science viewpoint. In our survey, we highlighted that some of the reviewed methods do not have a solid grounding in the social sciences. We have tried to find work on Personality in LLMs within the social science domain but with limited success. At present, there appears to be a research gap in this area. We hope our survey can attract researchers from the social sciences to contribute more rational research methodologies from social science perspectives to Personality in LLMs.

References

Yiming Ai, Zhiwei He, Ziyin Zhang, Wenhong Zhu, Hongkun Hao, Kai Yu, Lingjun Chen, and Rui Wang. 2024. Is cognition and action consistent or not: Investigating large language model's personality.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023a. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt. *IEEE Intelligent Systems*, 38(2):15–23.

Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023b. Can chatgpt's responses boost traditional natural language processing?

Michael C Ashton and Kibeom Lee. 2009. The hexaco–60: A short measure of the major dimensions of personality. *Journal of personality assessment*, 91(4):340–345.

Adrita Barua, Gary Brase, Ke Dong, Pascal Hitzler, and Eugene Vasserman. 2024. On the psychology of gpt-4: Moderately anxious, slightly masculine, honest, and humble. *arXiv preprint arXiv:2402.01777*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bojana Bodroza, Bojana M. Dinic, and Ljubisa Bojic. 2023. Personality testing of gpt-3: Limited temporal reliability, but highlighted social desirability of gpt-3's personality instruments results.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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.

Xubo Cao and Michal Kosinski. 2024. Large language models know how the personality of public figures is perceived by the general public. *Scientific Reports*, 14(1):6735.

Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models.

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.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

Szu-Wei Cheng, Chung-Wen Chang, Wan-Jung Chang, Hao-Wei Wang, Chih-Sung Liang, Taishiro Kishimoto, Jane Pei-Chen Chang, John S Kuo, and Kuan-Pin Su. 2023. The now and future of chatgpt and gpt in psychiatry. *Psychiatry and clinical neurosciences*, 77(11):592–596.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *arXiv preprint arXiv:2312.12999*.

Andrew Cutler and David M Condon. 2022. Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*.

Boele De Raad. 2000. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers.

Joost CF de Winter, Tom Driessen, and Dimitra Dodou. 2024. The use of chatgpt for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences*, 228:112729.

Erik Derner, Dalibor Kučera, Nuria Oliver, and Jan Zahálka. 2023. Can chatgpt read who you are?

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Florian E. Dorner, Tom Suhr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models?

Sybil BG Eysenck, Hans J Eysenck, and Paul Barrett. 1985. A revised version of the psychoticism scale. *Personality and individual differences*, 6(1):21–29.

William Fleeson and Eranda Jayawickreme. 2015. Whole trait theory. *Journal of research in personality*, 56:82–92.

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

Adithya V Ganesan, Yash Kumar Lal, August Hakan Nilsson, and H. Andrew Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Lewis R Goldberg et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Heng Gu, Chadha Degachi, Uğur Genç, Senthil Chandrasegaran, and Himanshu Verma. 2023. On the effectiveness of creating conversational agent personalities through prompting. *arXiv preprint arXiv:2310.11182*.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of llm personality.

Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, and Bing Xiang Yang. 2023. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects.

Zihong He and Changwang Zhang. 2024. Afspp: Agent framework for shaping preference and personality with large language models.

Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. Eliciting personality traits in large language models.

David W Hodo. 2006. Kaplan and sadock's comprehensive textbook of psychiatry. *American Journal of Psychiatry*, 163(8):1458–1458.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llmvssmall model? large language model based text augmentation enhanced personality detection model.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023a. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023b. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Yu Ji, Wen Wu, Hong Zheng, Yi Hu, Xi Chen, and Liang He. 2023. Is chatgpt a good personality recognizer? a preliminary study. *arXiv preprint arXiv:2307.03952*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023a. Evaluating and inducing personality in pre-trained language models. In *NeurIPS*.

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.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023b. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology*.

John A. Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of Research in Personality*, 51:78–89.

Peter K Jonason and Gregory D Webster. 2010. The dirty dozen: a concise measure of the dark triad. *Psychological assessment*, 22(2):420.

Daniel N Jones and Delroy L Paulhus. 2014. Introducing the short dark triad (sd3) a brief measure of dark personality traits. *Assessment*, 21(1):28–41.

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

Oliwier Kaszyca, Przemysław Kazienko, Jan Kocoń, Igor Cichecki, Mateusz Kochanek, and Dominka Szydło. 2023. Is it possible for chatgpt to mimic human annotator?

Lawrence J. Klinkert, Stephanie Buongiorno, and Corey Clark. 2024. Driving generative agents with their personality.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives.

Peter J Kwantes, Natalia Derbentseva, Quan Lam, Oshin Vartanian, and Harvey HC Marmurek. 2016. Assessing the big five personality traits with latent semantic analysis. *Personality and Individual Differences*, 102:229–233.

Frieder R Lang, Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G Wagner. 2011. Short assessment of the big five: Robust across survey methods except telephone interviewing. *Behavior research methods*, 43:548–567.

Kibeom Lee and Michael C Ashton. 2018. Psychometric properties of the hexaco-100. *Assessment*, 25(5):543–556.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2023. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective.

Jianzhi Liu, Hexiang Gu, Tianyu Zheng, Liuyu Xiang, Huijia Wu, Jie Fu, and Zhaofeng He. 2024a. Dynamic generation of personalities with large language models.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024b. Personality-aware student simulation for conversational intelligent tutoring systems.

Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing personality for large language models.

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

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022b. Who is gpt-3? an exploration of personality, values and demographics.

Walter Mischel, Yuichi Shoda, and Ozlem Ayduk. 2007. *Introduction to personality: Toward an integrative science of the person*. John Wiley & Sons.

Isabel Briggs Myers. 1962. *The Myers-Briggs Type Indicator: Manual (1962).* The Myers-Briggs Type Indicator: Manual (1962). Consulting Psychologists Press, Palo Alto, CA, US.

David Noever and Sam Hyams. 2023. Ai text-to-behavior: A study in steerability.

Sean Noh and Ho-Chun Herbert Chang. 2024. Llms with personalities in multi-issue negotiation games.

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*.

Delroy L Paulhus, Erin E Buckels, Paul D Trapnell, and Daniel N Jones. 2020. Screening for dark personalities. *European Journal of Psychological Assessment*.

Max Pellert, Clemens Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2022. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories.

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.

Maximilian Pellert, Clemens M. Lechner, Carmen Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 0(0).

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.

Heinrich Peters, Moran Cerf, and Sandra C Matz. 2024. Large language models can infer personality from free-form user interactions. *arXiv preprint arXiv:2405.13052*.

Heinrich Peters and Sandra Matz. 2023. Large language models can infer psychological dispositions of social media users.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis.

Mirjana Prpa, Giovanni Maria Troiano, Matthew Wood, and Yvonne Coady. 2024. Challenges and opportunities of llm-based synthetic personae and data in hci. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Angela Ramirez, Mamon Alsalihy, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling personality style in dialogue with zero-shot prompt-based learning. *arXiv preprint arXiv:2302.03848*.

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, Singapore. Association for Computational Linguistics.

Peter Romero, Stephen Fitz, and Teruo Nakatsuma. 2023. Do gpt language models suffer from split personality disorder? the advent of substrate-free psychometrics.

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*.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

Michael F Scheier and Charles S Carver. 1985. The self-consciousness scale: A revised version for use with general populations 1. *Journal of Applied Social Psychology*, 15(8):687–699.

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.

Chenglei Shen, Guofu Xie, Xiao Zhang, and Jun Xu. 2024. On the decision-making abilities in role-playing using large language models.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments.

Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. 2024. Identifying multiple personalities in large language models with external evaluation.

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

Aleksandra Sorokovikova, Natalia Fedorova, Sharwin Rezagholi, and Ivan P. Yamshchikov. 2024. Llms simulate big five personality traits: Further evidence.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Karthik Sreedhar and Lydia Chilton. 2024. Simulating human strategic behavior: Comparing single and multi-agent llms.

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.

Chuanneng Sun, Songjun Huang, and Dario Pompili. 2024. Llm-based multi-agent reinforcement learning: Current and future directions.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Fiona Anting Tan, Gerard Christopher Yeo, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Kokil Jaidka, Yang Liu, and See-Kiong Ng. 2024. Phantom: Personality has an effect on theory-of-mind reasoning in large language models.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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*.

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.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.

Zhiyuan Wen, Jiannong Cao, Yu Yang, Haoli Wang, Ruosong Yang, and Shuaiqi Liu. 2023. Desprompt: Personality-descriptive prompt tuning for few-shot personality recognition. *Information Processing & Management*, 60(5):103422.

Zhiyuan Wen, Jiannong Cao, Yu Yang, Ruosong Yang, and Shuaiqi Liu. 2024. Affective- nli: Towards accurate and interpretable personality recognition in conversation. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 184–193.

Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. Controllm: Crafting diverse personalities for language models.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. 2024. Academically intelligent llms are not necessarily socially intelligent. *arXiv preprint arXiv:2403.06591*.

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.

Byunggu Yu and Junwhan Kim. 2023. Personality of ai. *arXiv preprint arXiv:2312.02998*.

Li Yu, Xuefei Bai, and Zhiwei Chen. 2023. Gpt-lgbm: A chatgpt-based integrated framework for credit scoring with textual and structured data. *Available at SSRN 4671511*.

Tianyi Zhang, Antonis Koutsoumpis, Janneke K. Oostrom, Djurre Holtrop, Sina Ghassemi, and Reinout E. de Vries. 2024. Can large language models assess personality from asynchronous video interviews? a comprehensive evaluation of validity, reliability, fairness, and rating patterns. *IEEE Transactions on Affective Computing*, pages 1–16.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

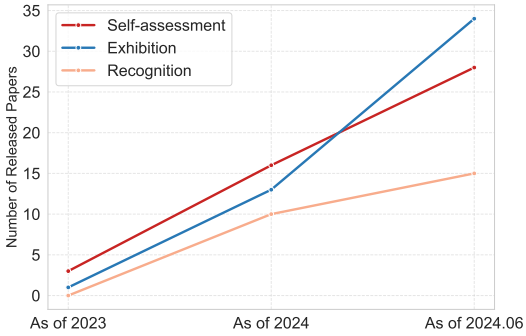

Figure 2: Trends of Personality in LLMs

A Overview

We present a holistic overview of the latest studies based on statistical results on our taxonomy. We have collected a total of 72 released scientific papers on personality in LLMs since 2022, encompassing investigations, methodologies, and applications. We clarify that the papers we reviewed are about psychological personality in LLMs. Although we are aware there are also extensive studies focusing on LLM-based role-playing agents (Chen et al., 2024), we exclude them from this survey.

The number of papers in this emerging domain has been increasing annually, as shown in Figure 2. Even as of June 2024, the volume of publications has already surpassed that of the entire 2023. This indicates a growing interest in the field. Concurrently, we observe a substantial growth of work on LLMs' personality exhibition, underscoring the increasing focus on LLM-based interactions in various scenarios. Following closely is research on LLMs' personality self-assessment, reflecting a sustained interest in exploring the intrinsic characteristics of LLMs. Compared to the two new research problems, there is a relatively less increase of personality recognition in LLM. This may be attributed to the fact that personality recognition, as a classical text classification problem, has been already widely studied with traditional methods. Nevertheless, personality recognition based on LLMs remains crucial in LLM-based interactions. The number of studies on this topic also continues to grow annually.

A.1 Personality Models

In reviewed literature, researchers commonly adopt personality models based on the trait theory (Fleeson and Jayawickreme, 2015), where the personalities of individuals are defined as several aspects of stable and consistent patterns of behavior, emotion,

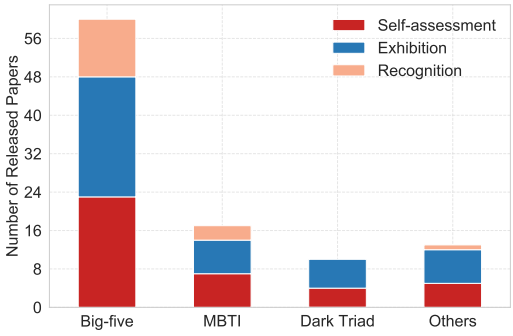

Figure 3: Personality Models in existing studies

and cognition.

As shown in Figure 3, the most commonly adopted personality model among the existing three research problems is the Big-five model (De Raad, 2000), which includes five core dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. These dimensions capture various aspects of an individual's personality, ranging from their inclination towards new experiences to their level of emotional stability. Another widely recognized personality assessment model is the Myers-Briggs Type Indicator (**MBTI**) (Myers, 1962), which categorizes individuals into one of 16 personality types based on their preferences in four dichotomous dimensions: extraversion/introversion, sensing/intuition, thinking/feeling, and judging/perceiving.

Besides the comprehensive personality models, researchers are also interested in the potential dark personality traits of LLMs, such as the Short Dark Triad-3 (SD-3, Jones and Paulhus (2014)), or Dark Triad Dirty Dozen (DTDD, Jonason and Webster (2010)) which measures Machiavellianism (a manipulative attitude), narcissism (excessive self-love), and psychopathy (lack of empathy), capturing the darker aspects of human nature.

In addition to the aforementioned personality models, researchers are also interested in (1) their variant models, such as the Eysenck Personality Questionnaire-Revised (EPQ-R) (Eysenck et al., 1985) assessing the dimensions of extraversion, neuroticism, and psychoticism, or the HEXACO model (Ashton and Lee, 2009) measuring honesty-humility in addition to the Big-five traits; or (2) other psychological aspects in LLMs, such as motivations or interpersonal relationships (Huang et al., 2023b; Bodroza et al., 2023).

A.2 Large Language Models

In existing studies, there is a significant interest in the performance of various LLMs on the three research problems, as shown in Figure 4. Among all the LLMs, the GPT series models have garnered the most attention from researchers. Though most of them are not directly open-sourced, extensive researchers (*e.g.*, (Xu et al., 2024; Karra et al., 2022; Jiang et al., 2023a)) have explored their performance in personality understanding through API. Additionally, several well-known open-source LLMs, such as the Llama series (Touvron et al., 2023), Mixtral (Jiang et al., 2024a), and Falcon[1], have also attracted a lot of attention.

Among all LLMs, the task that researchers have focused on the most is personality self-assessment. Despite LLMs typically referring to models after the occurrence of GPT-3 (Brown et al., 2020), we also found that some pioneering work has already attempted to measure potential personality traits exhibited by BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019).

B Open-sourced Resources

B.1 Personality Inventories

We have collected the personality inventories adopted in the papers we've reviewed (*i.e.*, Likert-scale questionnaires) and listed them in Table 2. We can see that the most commonly used inventory is the BFI. Similar to Figure 3, most works focus on the Big-five personality model and its variants, such as HEXACO. Although many studies also focus on MBTI, we have not found many scales proposed in academic papers about MBTI. Therefore, current studies generally use questionnaires from 16Personalities[2], a popular personality questionnaire website. Besides, many works also investigate the dark personality of LLMs.

It is evident that even when studying the same personality model, *e.g*, Big-five, Dark Triad, different works will use a variety of personality inventories, and even some scales measure variants of these personality models. This reflects the diversity of psychometrics we mentioned in Introduction.

B.2 Code Repositories of LLM's Personality Self-assessment

Table 3 shows the publicly available code repositories in existing LLM's Personality Self-assessment

[1]https://huggingface.co/docs/transformers/model_doc/falcon
[2]https://www.16personalities.com/

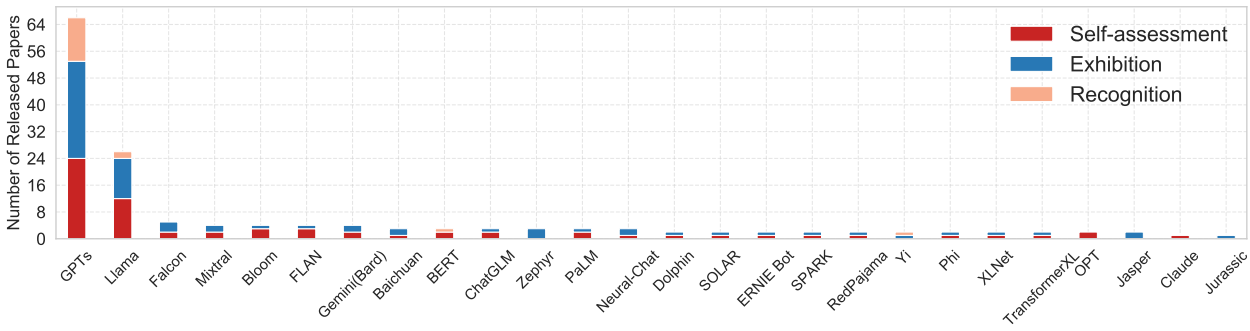

Figure 4: Investigated LLMs in existing studies

works. Since the majority of studies focus on prompting engineering, which don't need much custom data or code, few have made their works publicly accessible. Among studies in Table 3, apart from (Pellert et al., 2023) using NLI to assess the personalities of LLMs based on questionnaire questions and answers, all other works prompt LLMs with questionnaire items for personality assessment. We can also see that researchers exhibit interests in a wide range of LLMs, yet the majority still focus on LLMs in the GPT series.

B.3 Code Repositories of LLM's Personality Exhibition

Table 4 shows the publicly available code repositories in existing LLM's Personality Exhibition works. Similar to LLM's personality self-assessment, most works in LLM's Personality Exhibition are prompting engineering in techniques, few have made their codes and data publicly accessible. In Table 4, while the majority of the work involves inducing the personality of LLMs with prompts, the evaluation to the induced results vary widely, encompassing Social Intelligence psychometric (Xu et al., 2024), Theory-of-Mind reasoning tasks (Tan et al., 2024), response content analysis (Mao et al., 2024; Jiang et al., 2023b; Frisch and Giulianelli, 2024; tse Huang et al., 2023; Wang et al., 2024), and questionnaires (Jiang et al., 2023b; Cui et al., 2023; tse Huang et al., 2023; Klinkert et al., 2024).

B.4 Datasets for Personality Recognition in LLM

Table 5 contains the open-source datasets adopted in existing studies on Personality Recognition in LLMs. Some datasets, such as Essays (Pennebaker and King, 1999), PAN, and FriendsPersona, are classic text-based Personality Recognition datasets that have been explored by many studies before the emergence of LLMs. However, there are also some new datasets (Peters et al., 2024) that were constructed facilitated by LLMs.

Personality Inventories	Measured Facets	# Questions	Adopted in
Big Five Inventory (BFI) (John et al., 1991)	Big-five	44	(tse Huang et al., 2023); (Safdari et al., 2023); (Ai et al., 2024);(Pellert et al., 2022); (Huang et al., 2023b); (Li et al., 2023);
BFI-2 (Soto and John, 2017)	Big-five	60	(Li et al., 2023); (Huang et al., 2023b); (Dorner et al., 2023); (Safdari et al., 2023);
BFI-S (Lang et al., 2011)	Big-five	15	(Jiang et al., 2023a);
IPIP-NEO-120 (Johnson, 2014)	Big-five (30 facets)	120	(Jiang et al., 2023a)
IPIP-NEO (Goldberg et al., 1999)	Big-five	300	(Safdari et al., 2023);
IPIP-50 (Link)	Big-five	50	(Dorner et al., 2023);
Ten Item Personality Measure (TIPI) (Gosling et al., 2003)	Big-five	10	(Romero et al., 2023)
HEXACO (Ashton and Lee, 2009)	Honesty-humility and Big-five	60	(Miotto et al., 2022a)
HEXACO-100 (Lee and Ashton, 2018)	Honesty-humility and Big-five	100	(Bodroza et al., 2023);
16Personalities (Link)	MBTI	60	(Huang et al., 2023a); (Ai et al., 2024); (Rao et al., 2023)
Eysenck Personality Questionnaire-Revised (EPQ-R) (Eysenck et al., 1985)	Extraversion, Neuroticism, Psychoticism, and Lying	100	(Huang et al., 2023b)
Dark Triad Dirty Dozen (Jonason and Webster, 2010)	Machiavellianism , Narcissism, and Psychopathy (Dark Triad)	12	(Li et al., 2023), (Huang et al., 2023b);
Short Dark Triad (Jones and Paulhus, 2014)	Dark Triad	27	(Bodroza et al., 2023)
Short Dark Tetrad (SD4) (Paulhus et al., 2020)	Dark Triad and Sadism	28	(Pellert et al., 2022)
Self-Consciousness Scales–Revised (SCS-R) (Scheier and Carver, 1985)	Private self-consciousness, Public self-consciousness, and Social anxiety	22	(Bodroza et al., 2023)

Table 2: Personality inventories adopted by existing studies

Methods	Personality Models	Assessed LLMs	Download Links
PsychoBench (Huang et al., 2023b)	Big-five, EPQ-R, Short Dark Triad	GPT-3 (text-davinci-003); GPT-3.5-turbo,4; Llama-2-7B,13B	https://github.com/CUHK-ARISE/PsychoBench
(Shu et al., 2024)	MODEL-PERSONAS 39 instruments in 115 axis	GPT-2,3.5,4; Falcon-7B; BLOOMZ (all series); Llama2-7B,7B-chat,13B,13B-chat; RedPajama-7B; and FLAN-T5 (all series)	https://github.com/orange0629/llm-personas
(Miotto et al., 2022a)	HEXACO	GPT-3	https://github.com/ben-aaron188/who_is_gpt3
(Romero et al., 2023)	Big-five	GPT-3	https://osf.io/bf5c4/
(Pellert et al., 2023)	Big-five	multilingualDeBERTa; DistilRoBERTa; BART; XLMRoBERTa; DeBERTa; DistilBART	https://github.com/maxpel/psyai_materials
(Stöckli et al., 2024)	Big-five, MBTI	GPT4	https://github.com/AdritaBarua/2024-Psychology-of-GPT-4
(Jiang et al., 2023a)	Big-five	GPT-3.5	https://github.com/jianggy/MPI
(Safdari et al., 2023)	Big-five	PaLM-62B; Flan-PaLM-8B,62B,540B; Flan-PaLMChilla-62B	https://github.com/google-research/google-research/tree/master/psyborgs
(tse Huang et al., 2023)	Big-five	GPT-3.5-turbo	https://github.com/CUHK-ARISE/LLMPersonality
(Pan and Zeng, 2023)	MBTI	ChatGPT; GPT-4; Bloom-7B; Baichuan- 7B,13B; OpenLlama-7B-v2	https://github.com/HarderThenHarder/transformers_tasks/tree/main/LLM/llms_mbti
(Bodroza et al., 2023)	Big-five, HEXACO, SCS-R	GPT-3 (text-davinci-003)	https://osf.io/2k458

Table 3: Open source code repositories for LLM's personality self-assessment

Dataset	Personality Model	Descriptions	Download Links
Situational Evaluation of Social Intelligence SESI (Xu et al., 2024)	Big-five	Prompt LLMs with personality descriptions with extents, evaluated by the proposed SESI	https://github.com/RossiXu/social_intelligence_of_llms
PHAnToM (Tan et al., 2024)	Big-five, Dark Triad	Prompt LLMs with personality descriptions, evaluated by Theory-of-Mind reasoning tasks	https://anonymous.4open.science/r/PHAnToM/
PersonalityEdit (Mao et al., 2024)	Big-five (A, E, and N)	Utilize various methods in QA-based SFT and prompting for LLMs, evaluated by response content analysis	https://github.com/zjunlp/EasyEdit/blob/main/examples/PersonalityEdit.md
PersonaLLM (Jiang et al., 2023b)	Big-five	Prompt LLMs with personality descriptions, evaluated by questionnaires and story generation analysis	https://github.com/hjian42/PersonaLLM
MachineMindset (Cui et al., 2023)	MBTI	Conduct two-phase SFT with QA datasets and DPO to LLMs, evaluated by questionnaires	https://github.com/PKU-YuanGroup/Machine-Mindset
(Wang et al., 2024)	Big-five, MBTI	Prompt personality descriptions and profiles to role-playing agents, assessed by interview analysis	https://github.com/Neph0s/InCharacter/
(tse Huang et al., 2023)	Big-five	Assign personalities via QA pairs, biography, and CoT-based portrayals, assessed with questionnaires in multiple formats and story generation analysis.	https://github.com/CUHK-ARISE/LLMPersonality
(Klinkert et al., 2024)	Big-five	Prompt LLMs with personality descriptions and numeric extents, assessed by a personality questionnaire	https://gitlab.com/humin-game-lab/artificial-psychosocial-framework/-/tree/master/LLM_Personality
(Frisch and Giulianelli, 2024)	Big-five	Instruct LLMs with creative and analytical personalities to generate stories, evaluated by LIWC.	https://github.com/ivarfresh/Interaction_LLMs

Table 4: Open source code repositories for LLM's personality exhibition

Datasets	Personality Models	Descriptions	Download Links
(Rao et al., 2023)	MBTI	Questionnaires with 60 questions	https://github.com/Kali-Hac/ChatGPT-MBTI
Essays (Pennebaker and King, 1999)	Big-five	2,468 self-report essays from more than 1,200 students	https://github.com/preke/DesPrompt/tree/main/data/Essay *non-official download link
PAN	Big-five	294 users' tweets and their Big-Five personality scores obtained by the BFI-10 questionnaire	https://pan.webis.de/clef15/pan15-web/author-profiling.html
Kaggle	MBTI	8,675 users, with each user contributing 45-50 posts	https://www.kaggle.com/datasets/datasnaek/mbti-type
Pandora	MBTI	Dozens to hundreds of posts from each of the 9,067 Reddit users	https://psy.takelab.fer.hr/datasets/all/
FriendsPersona	Big-five	711 short conversations are extracted and annotated from the first four seasons of Friends TV Show transcripts	https://github.com/emorynlp/personality-detection
CPED	Big-five	12K dialogues in multi-modal context from 40+ TV shows	https://github.com/scutcyr/CPED
First Impression	Big-five	10,000 Youtube video clips of people facing and speaking	https://chalearnlap.cvc.uab.cat/dataset/24/description/
(Peters et al., 2024)	Big-five	Dialogues of 566 participants with a chatbot built on the ChatGPT	https://osf.io/edn3g/
(Cao and Kosinski, 2024)	Big-five	Demographics of 11,341 public figures from Pantheon 1.0 with manually rated personality traits.	https://osf.io/854w2/

Table 5: Open source datasets for personality recognition in LLM