THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Self-assessment, Exhibition, and Recognition: a Review of LLMs‘ Understanding of Personality

Target: EMNLP’24 (DDL: 16, June)

Zhiyuan Wen

Email: zyuanwen@polyu.edu.hk
http://preke.github.io

Department of Computing
電子計算學系

Abstract

With large language models (LLMs) appearing to behave increasingly human-like in interactions, there is a growing interest in investigating LLMs‘ understanding of personality. However, due to the multiplicity in psychology personality research and the rapid development of LLMs, numerous existing studies exhibit substantial diversity in research focus, methods, and LLMs being investigated. Consequently, it is difficult for researchers to not only have a holistic understanding of this field but also further apply research findings to real-world applications. In this paper, we present a comprehensive review of LLMs’ understanding of personality, dividing existing studies into three research problems: Personality Self-assessment of LLMs, Personality Exhibition in LLMs, and Personality Recognition in LLMs. For each research problem, we further categorize and summarize the investigations and solutions in existing work, identifying the key findings and open challenges. Finally, we outline the future research trends to inspire researchers towards this emerging field. Our paper is the first comprehensive survey on personality understanding in LLMs. By providing a clear taxonomy of existing studies and identifying promising future directions, we aspire to facilitate researchers for the understanding of this interdisciplinary field.

Outline

- Research problem
- Importance
- Solution
- Results
- Impact

The Research Problem

- What problem I want to solve?
  - A comprehensive review of existing literature in LLMs' understanding of personality with clear organization
- What questions I want to answer?
  - What are the research problems in LLMs' understanding of personality?
  - How existing studies investigate and solve these problems?
  - Are there any challenging issues remain unsolved in existing studies?
  - What are the promising future research trends?

The Importance

- Why people care about LLMs' understanding of personality?
  - Widely application and human-like interaction experience of LLMs
  - Better understanding of LLMs
    - Investigating LLMs' capabilities in personality-related tasks
    - Uncovering potential biases and safety issues in LLMs
  - Better utilization of LLMs
    - Building conversational agents based on LLMs
    - Conducting personality-related tasks with LLMs
- Why we need a survey paper?
  - Extensive studies have been published in recent two years
    - Variety in research focus
    - Variety in personality models
    - Variety in LLMs
  - No existing survey paper for this emerging area

The Solution

- We conduct a statistical analysis of recent publications on research focus, personality models, and LLMs
- We propose a hierarchical taxonomy of existing literature in problem-level and method-level
  - Personality Self-assessment of LLMs
    - Likert-scale Questionnaires
    - Text analysis on Responses
  - Personality Exhibition in LLMs
    - Editing LLM's Personality
    - Inducing LLM's Personality
  - Personality Recognition in LLMs
    - Personality Recognition by LLM
    - LLM-enhanced Personality Recognition
- We synthesize and consolidate the findings from existing studies, and further analyze the underlying causes of these results

The Results

- A clear taxonomy to organize existing literature
- Finding summarization in existing literature:
  - Most existing studies are prompting-engineering
  - No consensus on general personality traits of LLMs, but LLMs are observed to exhibit darker traits than human average
  - Compared with fine-tuning, inducing specific personality in LLM by prompting renders more effective results and has less resource constraints
  - The zero-shot prompting ability of LLMs easily outperforms traditional NN models and pre-trained language models in personality recognition, but it still underperforms supervised (even small) models
- Research gap identification:
  - Reliability and validity in personality self-assessment
  - Consistency and robustness in personality exhibition
  - Bias elimination in personality recognition
- Future research trends:
  - Benchmarking for LLMs' personality understanding
  - Psychometrics tailored to LLMs
  - Conversational agents based on LLMs exhibiting specific personality traits

The Impact

- Our survey is the first comprehensive survey on personality understanding in LLMs
- For young researchers in this area
  - We provide a clear and comprehensive grasp the current research status
  - We identify the open issues and promising trends for further research
- For developers in LLM-based conversational agents with specific personalities
  - We summarize the most extensively studied LLMs and personality models, along with their research findings. This will expedite their investigations process for developers
- For companies/institutions developing LLMs
  - We identify the potential biases and safety issues uncovered in existing studies, providing insights for LLM development or further refinement


thank
you

Outline

- Introduction
- Personality Self-assessment of LLMs
- Personality Exhibition in LLMs
- Personality Recognition in LLMs
- Future Research Trends
- Abstract of the Manuscript

Introduction

- What is the topic about?
  - Personality is a set of relatively stable individual traits and characteristics that define a person's distinctive pattern of thinking, feeling, and acting[1]
    - Big five personality traits, MBTI types, Dark Traid
  - Large Language Models (LLMs)' understanding of Personality
    - **Personality Self-assessment of LLMs:** measure LLMs' intrinsic personality traits (if have)
    - **LLMs' Personality Exhibition:** reflect the specified personality in generated text content by LLMs
    - **Personality Recognition with LLMs:** classify the personality traits from the input text content facilitated by LLMs

[1] Gordon Willard Allport. 1937. Personality: A psychological interpretation

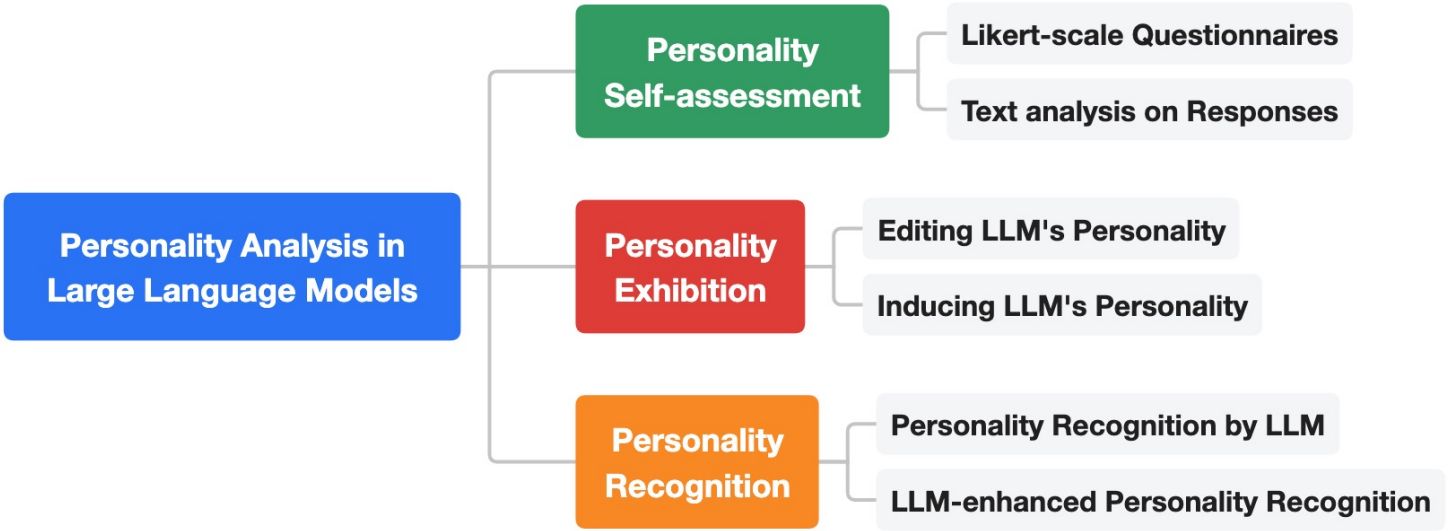
Introduction

- Why people care about this topic?
  - Better Understanding of LLMs
  - Better Utilization of LLMs

- Why we need a survey paper?
  - Extensive studies have been published in recent two years
    - Variety in research focus
    - Variety in personality assessment approaches
    - Variety in LLMs
  - No existing survey paper for this emerging area

Introduction

- Overview of Existing Studies
  - 48 published papers (up to 2024.05)
  - Most studied personality models: Big-five, MBTI, Dark Triad
  - Top three popular LLMs: GPT-series, Llama, Mixtral



LLM's Personality Self-assessment

- **Problem Statement:** how to systematically measure the personality traits exhibited in text responses generated from LLMs through psychometric methods?
  - Psychometric methods are statistical techniques and tools used in social sciences to analyze individual differences and psychological attributes[1]
- **Motivation:**
  - Uncovering latent biases and safety issues within LLMs
  - Providing insights for building conversational agents based on LLMs
- **Assumptions:**
  - LLMs have acquired intrinsic personality from pre-training data
  - Psychometric methods designed for human are also suitable for LLMs
- **Constraints:**
  - LLMs only have open ended text input and output

[1] Wijsen L D, Borsboom D, Alexandrova A. Values in psychometrics[J]. Perspectives on Psychological Science, 2022, 17(3): 788-804

LLM's Personality Self-assessment

How to assess the personalities of LLMs?

- Likert-scale questionnaires
  - Filling masked positions in items[1]
  - Adding options beside original items[2]
  - Adding instructive task descriptions[3] and response format constraints[4]
- Text analysis on responses
  - Personality classification[5]
  - Psycholinguistic statistical feature analysis[6]
  - Human evaluation[7]

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.

Questionnaire item example

[1] Caron G, Srivastava S. Identifying and manipulating the personality traits of language models[J]. arXiv preprint arXiv:2212.10276, 2022.
[2] Li X, Li Y, Joty S, et al. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective[J]. arXiv preprint arXiv:2212.10529, 2022.
[3] La Cava L, Costa D, Tagarelli A. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models[J]. arXiv preprint arXiv:2401.07115, 2024.
[4] Huang J, Wang W, Li E J, et al. Who is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench[J]. arXiv preprint arXiv:2310.01386, 2023.
[5] Karra S R, Nguyen S T, Tulabandhula T. Estimating the Personality of White-Box Language Models[J]. arXiv preprint arXiv:2204.12000, 2022.
[6] Jiang H, Zhang X, Cao X, et al. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences[J]. arXiv preprint arXiv:2305.02547, 2023.
[7] Jiang G, Xu M, Zhu S C, et al. Evaluating and inducing personality in pre-trained language models[J]. Advances in Neural Information Processing Systems, 2024, 36.

LLM's Personality Self-assessment

What are the assessment results?

- Even for the same LLM, different studies obtain different results
  - Different assessment approaches
  - Different inventories (questionnaires)
  - …
- LLMs generally exhibit more negative traits than human norms[1][2].

LLMs	Results	Source
ChatGPT	ENTJ	(Pan and Zeng, 2023)
ChatGPT	ENFJ	(Huang et al., 2023a)
GPT-4	INTJ	(Pan and Zeng, 2023)
GPT-4	ENFJ	(Huang et al., 2023a)
Llama-7b	ENFP	(Pan and Zeng, 2023)
Llama-2 (7b,13b)	ENFJ	(Pan and Zeng, 2023)

MBTI types of LLMs are like teachers, commanders, instructors..

Table 3: Results on personality traits.

	Subscales	llama2-7b	llama2-13b	text-davinci-003	gpt-3.5-turbo	gpt-4	gpt-4-jb	Crowd Male	Crowd Female
BFI	Openness	4.2±0.3	4.1±0.4	**4.8±0.2**	4.2±0.3	4.2±0.6	3.8±0.6	3.9±0.7	
	Conscientiousness	3.9±0.3	4.4±0.3	4.6±0.1	4.3±0.3	**4.7±0.4**	3.9±0.6	3.5±0.7	
	Extraversion	3.6±0.2	3.9±0.4	**4.0±0.4**	3.7±0.2	3.5±0.5	3.6±0.4	3.2±0.9	
	Agreeableness	3.8±0.4	4.7±0.3	**4.9±0.1**	4.4±0.2	4.8±0.4	3.9±0.7	3.6±0.7	
	Neuroticism	**2.7±0.4**	1.9±0.5	1.5±0.1	2.3±0.4	1.6±0.6	2.2±0.6	3.3±0.8	
EPQ-R	Extraversion	14.1±1.6	17.6±2.2	**20.4±1.7**	19.7±1.9	15.9±4.4	16.9±4.0	12.5±6.0	14.1±5.1
	Neuroticism	6.5±2.3	13.1±2.8	16.4±7.2	**21.8±1.9**	3.9±6.0	7.2±5.0	10.5±5.8	12.5±5.1
	Psychoticism	**9.6±2.4**	6.6±1.6	1.5±1.0	5.0±2.6	3.0±5.3	7.6±4.7	7.2±4.6	5.7±3.9
	Lying	13.7±1.4	14.0±2.5	17.8±1.7	9.6±2.0	**18.0±4.4**	17.5±4.2	7.1±4.3	6.9±4.0
DTDD	Narcissism	6.5±1.3	5.0±1.4	3.0±1.3	**6.6±0.6**	2.0±1.6	4.5±0.9	4.9±1.8	
	Machiavellianism	4.3±1.3	4.4±1.7	1.5±1.0	**5.4±0.9**	1.1±0.4	3.2±0.7	3.8±1.6	
	Psychopathy	4.1±1.4	3.8±1.6	1.5±1.2	4.0±1.0	1.2±0.4	**4.7±0.8**	2.5±1.4	

Dark Triad

- **Narcissism:** lack of empathy
- **Machiavellianism:** by any means necessary, fraud
- **Psychopathy:** anti-society

	Machiavellianism↓	Narcissism↓	Psychopathy↓
GPT-3	3.13 ± 0.54	3.02 ± 0.40	2.93 ± 0.41
GPT-3-I1	3.49 ± 0.39	3.51 ± 0.22	2.48 ± 0.34
GPT-3-I2	3.60 ± 0.40	3.43 ± 0.31	2.39 ± 0.35
FLAN-T5-XXL	3.93 ± 0.29	3.36 ± 0.21	3.10 ± 0.21
avg. human result	2.96 (0.65)	2.97 (0.61)	2.09 (0.63)

[1] Huang J, Wang W, Lam M H, et al. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models[J]. arXiv preprint arXiv:2305.19926, 2023.
[2] Li X, Li Y, Joty S, et al. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective[J]. arXiv preprint arXiv:2212.10529, 2022.

LLM's Personality Self-assessment

Are the assessments robust?

- **Reliability:** consistency and stability of test results over multiple repetitions
  - Although LLMs are sensitive to prompt design, thousands of repetitions (in different prompt settings) show satisfactory levels of reliability[1]
  - LLM's personality varies with temperatures[2], model sizes, and whether being instructive fine-tuned[3]

- **Validity:** measuring what it claims to measure
  - LLMs can explain the reason for selecting particular options within questionnaires[4]
  - Questionnaires applicable to humans may not be suitable for assessing personalities of LLMs[5]

[1] Huang J, Wang W, Lam M H, et al. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models[J]. arXiv preprint arXiv:2305.19926, 2023.
[2] Miotto M, Rossberg N, Kleinberg B. Who is GPT-3? An exploration of personality, values and demographics[J]. arXiv preprint arXiv:2209.14338, 2022.
[3] Safdari M, Serapio-García G, Crepy C, et al. Personality traits in large language models[J]. arXiv preprint arXiv:2307.00184, 2023.
[4] Jiang G, Xu M, Zhu S C, et al. Evaluating and inducing personality in pre-trained language models[J]. Advances in Neural Information Processing Systems, 2024, 36.
[5] Dorner F E, Sühr T, Samadi S, et al. Do personality tests generalize to Large Language Models?[J]. arXiv preprint arXiv:2311.05297, 2023.

LLM's Personality Self-assessment

Findings:

- Most existing studies are prompting-engineering
- No consensus on general personality traits of LLMs, but LLMs are observed to exhibit darker traits than human average
- Seldom studies follow the standard psychometric assessment protocols

Open Challenges:

- Robust assessment approaches
- Standard benchmark for LLM's personality self-assessment
- Psychometrics tailored to LLMs

LLM's Personality Exhibition

- **Problem Statement:** how to reflect the specified personality trait in the generated text content by LLMs?

- **Motivation:**
  - **Adaptability:** Modulating LLM personality ensures compatibility with diverse application environments.
  - **Personalization:** Progressive personality tuning in LLMs accommodates the evolving preferences and requirements of users.
  - **Safety:** Integrating refined personality traits in LLMs diminishes the potential for generating objectionable content.

LLM's Personality Exhibition?

Method Categorization

- **Editing LLM's personality by modifying model parameters**
  - Continue-training: conduct pre-training tasks on selected datasets
  - Fine-tuning:
    - Traditional fine-tuning with personality-annotated text data on auxiliary tasks[1]
    - Instruction fine-tuning with question-answer pairs from personality questionnaires[2], LLM generated data[8][9]
- **Inducing LLM's personality by modifying input prompts**
  - Explicit prompting with personality or profile descriptions[3][4][7]
  - Implicit prompting with behavioral examples[5][6][7]

[1] Karra S R, Nguyen S T, Tulabandhula T. Estimating the Personality of White-Box Language Models[J]. arXiv preprint arXiv:2204.12000, 2022.
[2] Li X, Li Y, Joty S, et al. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective[J]. arXiv preprint arXiv:2212.10529, 2022.
[3] Huang J, Wang W, Lam M H, et al. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models[J]. arXiv preprint arXiv:2305.19926, 2023.
[4] Jiang G, Xu M, Zhu S C, et al. Evaluating and inducing personality in pre-trained language models[J]. Advances in Neural Information Processing Systems, 2024, 36.
[5] Safdari M, Serapio-García G, Crepy C, et al. Personality traits in large language models[J]. arXiv preprint arXiv:2307.00184, 2023.
[6] Jiang H, Zhang X, Cao X, et al. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences[J]. arXiv preprint arXiv:2305.02547, 2023.
[7] Pan K, Zeng Y. Do llms possess a personality? making the mbti test an amazing evaluation for large language models[J]. arXiv preprint arXiv:2307.16180, 2023.
[8] Cui J, Lv L, Wen J, et al. Machine Mindset: An MBTI Exploration of Large Language Models[J]. arXiv preprint arXiv:2312.12999, 2023.
[9] Mao S, Zhang N, Wang X, et al. Editing personality for llms[J]. arXiv preprint arXiv:2310.02168, 2023.

LLM's Personality Exhibition

Findings:

- Editing and Inducing are both efficient ways to control LLM exhibit specified personality traits
- Inducing by prompts rendering more effective results (in some studies) and has less resource constraints

Open Challenges:

- Consistency in personality exhibition
- Robustness of personality exhibition

Personality Recognition in LLMs

- **Problem Statement:** how to utilize LLMs to facilitate recognizing personality traits from the given text content?

- **Motivation of using LLMs:**
  - Excellent zero-shot capabilities of LLMs reduce annotation shortage
  - LLMs are able to generate explanations for identifying personality traits, the interpretability of the results is also substantially enhanced.

Personality Recognition in LLMs

Method Categorization

- Personality Recognition by LLMs
  - Zero-shot prompting and in-context learning[1][2][3][4]
  - Parameter-efficient fine-tuning[8]
- LLM-enhanced Personality Recognition
  - Enriching the information (description, interpretation) of personality labels[7][8]
  - Chain-of-thought reasoning[9]

[1] 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.
[2] Ganesan A V, Lal Y K, Nilsson A H, et al. Systematic evaluation of gpt-3 for zero-shot personality estimation[J]. arXiv preprint arXiv:2306.01183, 2023.
[3] 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.
[4] T. Zhang, A. Koutsoumpis, J. K. Oostrom, D. Holtrop, S. Ghassemi and R. E. d. Vries, "Can Large Language Models Assess Personality from Asynchronous Video Interviews? A Comprehensive Evaluation of Validity, Reliability, Fairness, and Rating Patterns," in IEEE Transactions on Affective Computing
[5] Amin M M, Cambria E, Schuller B W. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt[J]. IEEE Intelligent Systems, 2023, 38(2): 15-23.
[6] Peters H, Matz S. Large language models can infer psychological dispositions of social media users[J]. arXiv preprint arXiv:2309.08631, 2023.
[7] Hu L, He H, Wang D, et al. LLM vs Small Model? Large Language Model Based Text Augmentation Enhanced Personality Detection Model[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(16): 18234-18242.
[8] Wen Z, Cao J, Yang Y, et al. Affective-NLI: Towards Accurate and Interpretable Personality Recognition in Conversation[C]//2024 IEEE International Conference on Pervasive Computing and Communications (PerCom). IEEE, 2024: 184-193.
[9] Yang T, Shi T, Wan F, et al. PsyCoT: Psychological Questionnaire as Powerful Chain-of-Thought for Personality Detection[J]. arXiv preprint arXiv:2310.20256, 2023.

Personality Recognition in LLMs

Findings:

- Zero-shot prompting ability of LLMs easily outperforms traditional NN models and pre-trained language models[1][4][6], but underperforms specialized personality recognition models[1][2][5]
- LLMs undergone RLHF has much better performance [3]
- LLMs can provide additional information to facilitate personality recognition models[7][8]

Open Challenge:

- Eliminating potential biases in LLMs towards certain demographic attributes[2][6]

[1] 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.
[2] Ganesan A V, Lal Y K, Nilsson A H, et al. Systematic evaluation of gpt-3 for zero-shot personality estimation[J]. arXiv preprint arXiv:2306.01183, 2023.
[3] 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.
[4] T. Zhang, A. Koutsoumpis, J. K. Oostrom, D. Holtrop, S. Ghassemi and R. E. d. Vries, "Can Large Language Models Assess Personality from Asynchronous Video Interviews? A Comprehensive Evaluation of Validity, Reliability, Fairness, and Rating Patterns," in IEEE Transactions on Affective Computing
[5] Amin M M, Cambria E, Schuller B W. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt[J]. IEEE Intelligent Systems, 2023, 38(2): 15-23.
[6] Peters H, Matz S. Large language models can infer psychological dispositions of social media users[J]. arXiv preprint arXiv:2309.08631, 2023.
[7] Hu L, He H, Wang D, et al. LLM vs Small Model? Large Language Model Based Text Augmentation Enhanced Personality Detection Model[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(16): 18234-18242.
[8] Wen Z, Cao J, Yang Y, et al. Affective-NLI: Towards Accurate and Interpretable Personality Recognition in Conversation[C]//2024 IEEE International Conference on Pervasive Computing and Communications (PerCom). IEEE, 2024: 184-193.
[9] Yang T, Shi T, Wan F, et al. PsyCoT: Psychological Questionnaire as Powerful Chain-of-Thought for Personality Detection[J]. arXiv preprint arXiv:2310.20256, 2023.

Future Research Trends

Future Trends

- Benchmarking for personality understanding in LLMs
- Psychometrics tailored to LLMs
- Life-long monitoring of psychometric properties in LLMs
- Conversational agents based on LLMs exhibiting specific personality traits
- Bias and harmful behavior elimination in LLMs


thank
you