On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM,

Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, Michael Lyu

My Homepage :)

ARISE
Automated Reliable Intelligent Software Engineering

香港中文大學
The Chinese University of Hong Kong

PsychoBench Motivation

- It can be imagined: AI and humans work and live in a same society
- The key initial step: evaluating AI's human-like abilities
  - Psychological portrayal
  - Emotional ability
  - Cognitive process
  - Decision-making
  - …

- This paper focuses on Psychological portrayal of LLMs
  - Why do we care about this?

Is LLM's Psychological Portrayal Important?

- For **Computer Science Researchers**:

(1) Build human-like AI systems [1]

(2) Understand its performance [2]

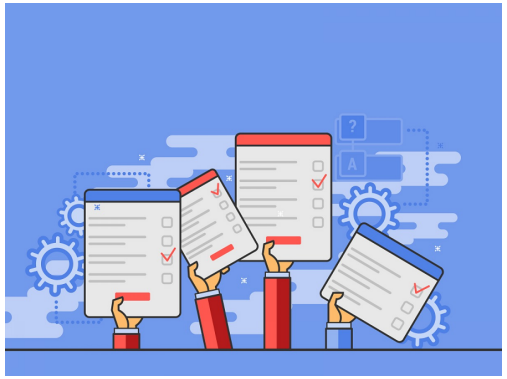
(3) Identify potential biases [3]

[1] X Wang et al. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. arXiv 2310.17076.
[2] C Li et al. Large Language Models Understand and Can be Enhanced by Emotional Stimuli. In LLM@IJCAI 2023.
[3] H Rao et al. Can ChatGPT Assess Human Personalities? A General Evaluation Framework. In EMNLP 2023.

Is LLM's Psychological Portrayal Important?

- For **Social Science Researchers**:

(1) Replace human in surveys [4]



(2) Understand how cultures shape individuals [5]

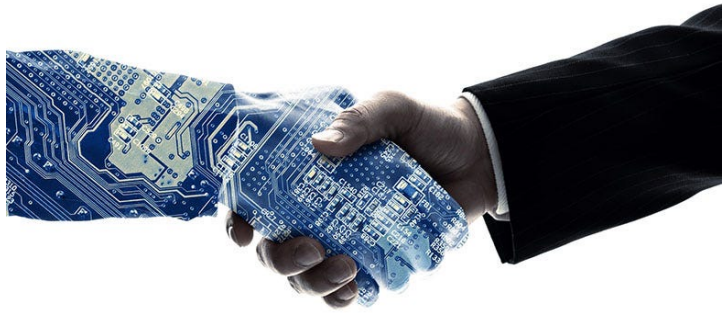
[4] D Dillion et al. Can AI Language Models Replace Human Participants? In Trends in Cognitive Sciences.
[5] M Tomasello. The Cultural Origins of Human Cognition. In Harvard University Press.

Is LLM's Psychological Portrayal Important?

- For **Users and Human Society**:

(1) Facilitate tailored AI assistants

(2) Build trust among users and AI



(3) Monitor AI's mental states [6]

[6] X Li et al. Evaluating Psychological Safety of Large Language Models. arXiv 2212.10529.

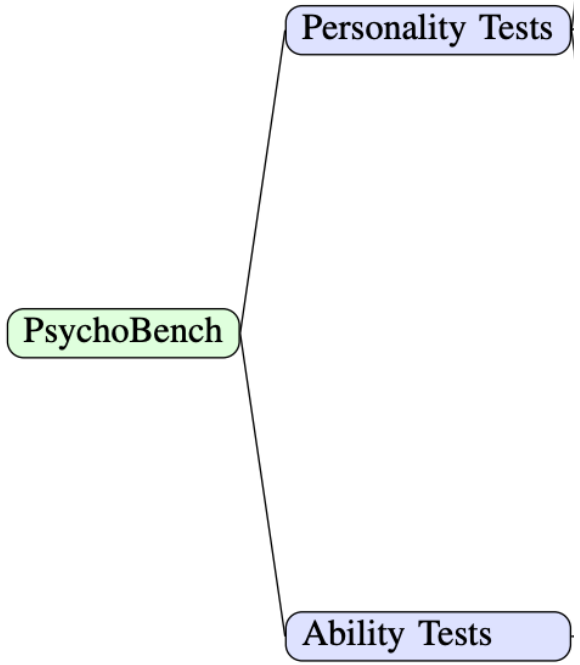
Introducing PsychoBench

PsychoBench

- Psychometrics
  - The field of assessing psychological attributes

Introducing PsychoBench

- Psychometrics

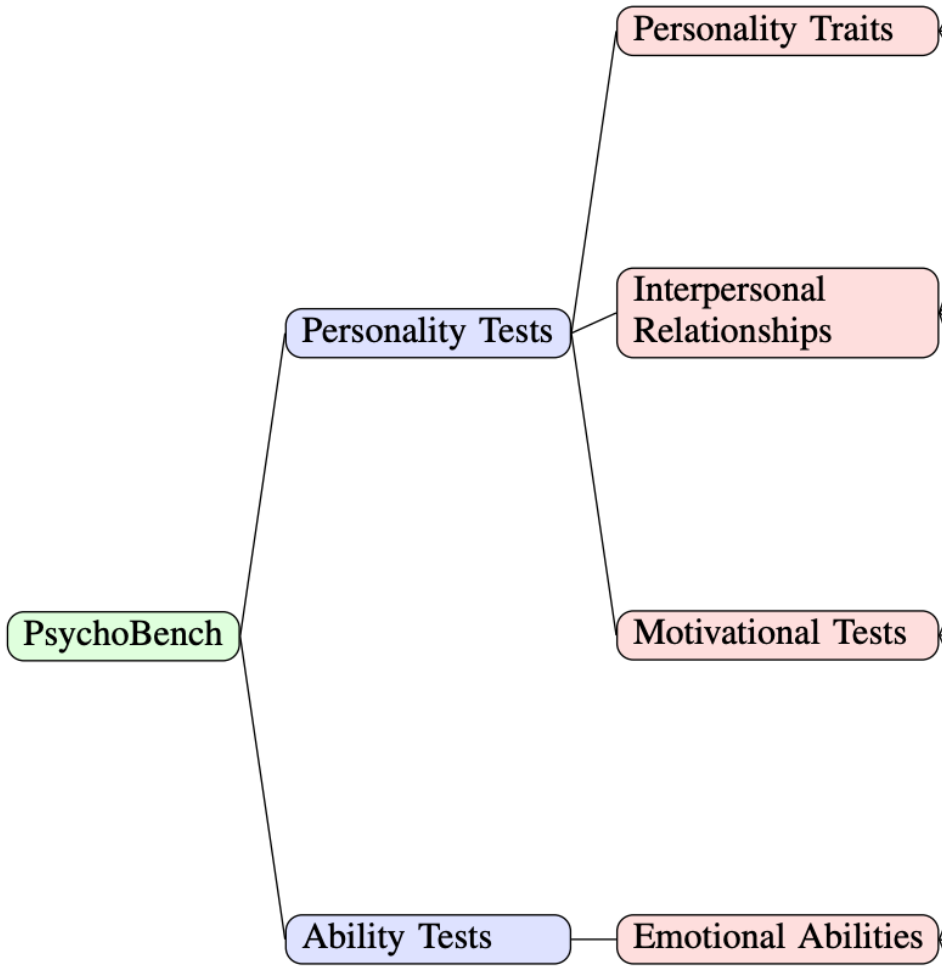


- Personality tests
  - Individual's attitudes, beliefs, values
  - Without absolute right/wrong answers

- Ability tests
  - Individual's proficiencies in specific domains
  - With objectively correct answers

J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.

Introducing PsychoBench

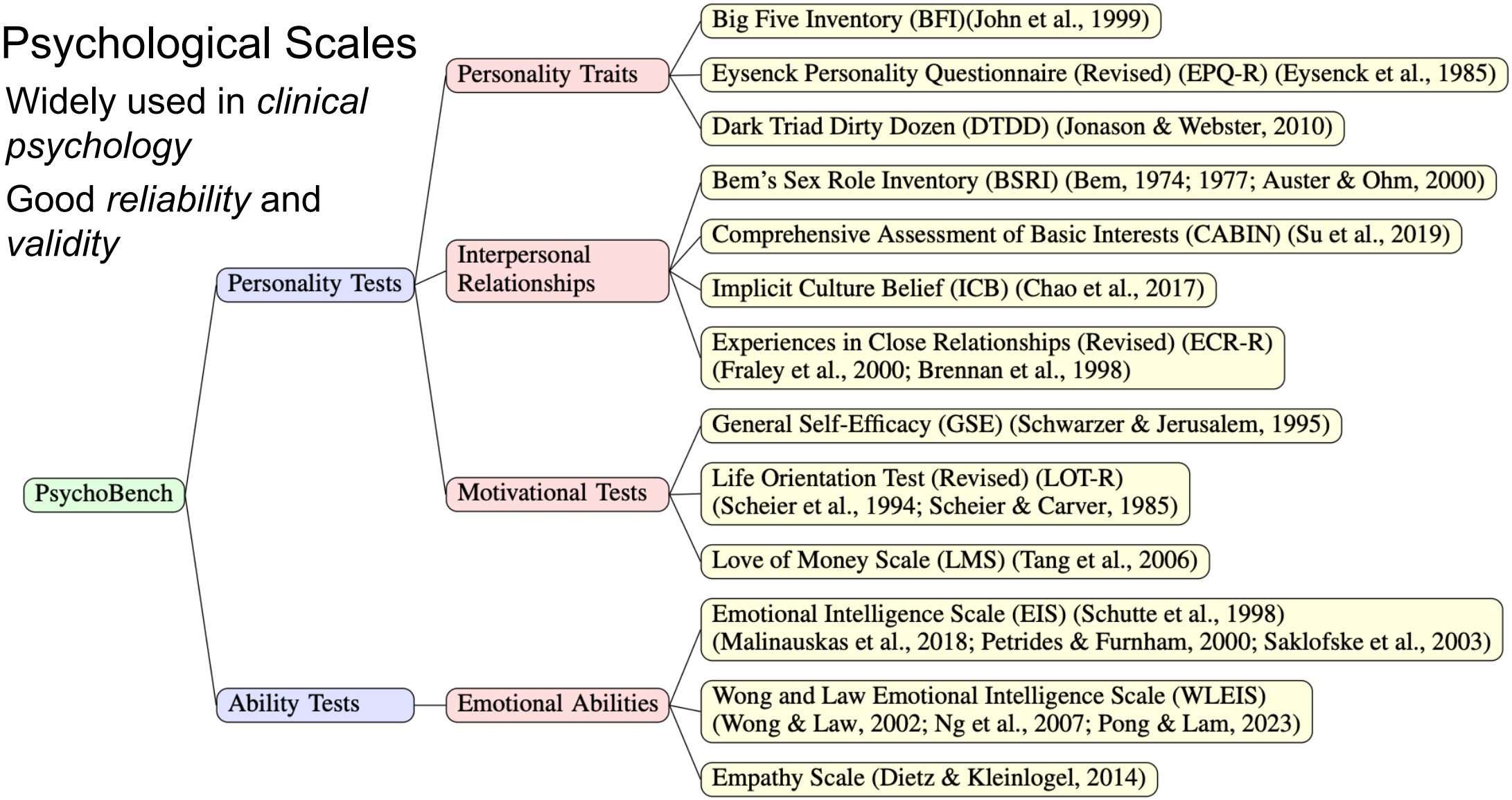


- Psychometrics
  - Personality Tests
    - Personality Traits *(What kind of person?)*
    - Interpersonal Relationship *(What's the role in the interpersonal communication?)*
    - Motivational Tests *(Self-motivation, self-confidence, optimism)*
  - Ability Tests
    - Emotional Abilities *(EQ)*

J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.

Introducing PsychoBench

- 13 Psychological Scales
  - ✓ Widely used in *clinical psychology*
  - ✓ Good *reliability* and *validity*



Experiment Design

- Models:
  - Text-davinci-003, gpt-3.5-turbo-0613, gpt-4-0613, llama2-7b-chat, llama2-13b-chat
  - A jailbreak method (CipherChat, also in ICLR'24) on gpt-4-0613
- To compare to human norms:

Scale	Number	Country/Region	Age Distribution	Gender Distribution
BFI	1,221	Guangdong, Jiangxi, and Fujian in China	16∼28, 20*	M (454), F (753), Unknown (14)
EPQ-R	902	N/A	17∼70, 38.44±17.67 (M), 31.80±15.84 (F)	M (408), F (494)
DTDD	470	The Southeastern United States	≥17, 19±1.3	M (157), F (312)
BSRI	151	Montreal, Canada	36.89±1.11 (M), 34.65±0.94 (F)	M (75), F (76)
CABIN	1,464	The United States	18∼80, 43.47±13.36	M (715), F (749)
ICB	254	Hong Kong SAR	20.66 ± 0.76	M (114), F (140)
ECR-R	388	N/A	22.59±6.27	M (136), F (252)
GSE	19,120	25 Countries/Regions	12∼94, 25±14.7[a]	M (7,243), F (9,198), Unknown (2,679)
LOT-R	1,288	The United Kingdom	16∼29 (366), 30∼44 (349), 45∼64 (362), ≥65 (210)[b]	M (616), F (672)
LMS	5,973	30 Countries/Regions	34.7±9.92	M (2,987), F (2,986)
EIS	428	The Southeastern United States	29.27±10.23	M (111), F (218), Unknown (17)
WLEIS	418	Hong Kong SAR	N/A	N/A
Empathy	366	Guangdong, China and Macao SAR	33.03*	M (184), F (182)

J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.

Highlighted Conclusions

	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	

1. Distinct personality traits
2. More negative traits
3. Jailbreak's influence

J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.

Highlighted Conclusions

	Subscales	llama2-7b	llama2-13b	text-davinci-003	gpt-3.5-turbo	gpt-4	gpt-4-jb	Crowd *Male*	Crowd *Female*
BSRI	**Masculine**	5.6±0.3	5.3±0.2	5.6±0.4	**5.8±0.4**	4.1±1.1	4.5±0.5	4.8±0.9	4.6±0.7
	Feminine	5.5±0.2	5.4±0.3	5.6±0.4	**5.6±0.2**	4.7±0.6	4.8±0.3	5.3±0.9	5.7±0.9
	Conclusion	10:0:0:0	10:0:0:0	10:0:0:0	8:2:0:0	6:4:0:0	1:5:3:1	-	
CABIN	**Health Science**	4.3±0.2	4.2±0.3	4.1±0.3	4.2±0.2	3.9±0.6	3.4±0.4	-	
	Creative Expression	4.4±0.1	4.0±0.3	4.6±0.2	4.1±0.2	4.1±0.8	3.5±0.2	-	
	Technology	4.2±0.2	4.4±0.3	3.9±0.3	4.1±0.2	3.6±0.5	3.5±0.4	-	
	People	4.3±0.2	4.0±0.2	4.5±0.1	4.0±0.1	4.0±0.7	3.5±0.4	-	
	Organization	3.4±0.2	3.3±0.2	3.4±0.4	3.9±0.1	3.5±0.4	3.4±0.3	-	
	Influence	4.1±0.2	3.9±0.3	3.9±0.3	4.1±0.2	3.7±0.6	3.4±0.2	-	
	Nature	4.2±0.2	4.0±0.3	4.2±0.2	4.0±0.3	3.9±0.7	3.5±0.3	-	
	Things	3.4±0.4	3.2±0.2	3.3±0.4	3.8±0.1	2.9±0.3	3.2±0.3	-	
ICB	**Overall**	**3.6±0.3**	3.0±0.2	2.1±0.7	2.6±0.5	1.9±0.4	2.6±0.2	3.7±0.8	
ECR-R	**Attachment Anxiety**	**4.8±1.1**	3.3±1.2	3.4±0.8	4.0±0.9	2.8±0.8	3.4±0.4	2.9±1.1	
	Attachment Avoidance	**2.9±0.4**	1.8±0.4	2.3±0.3	1.9±0.4	2.0±0.8	2.5±0.5	2.3±1.0	

1. Distinct personality traits
2. More negative traits
3. Jailbreak's influence
4. Bias towards Masculinity
5. Similar vocational preference

J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.

Highlighted Conclusions

	Subscales	llama2-7b	llama2-13b	text-davinci-003	gpt-3.5-turbo	gpt-4	gpt-4-jb	Crowd
GSE	**Overall**	39.1±1.2	30.4±3.6	37.5±2.1	38.5±1.7	**39.9±0.3**	36.9±3.2	29.6±5.3
LOT-R	**Overall**	12.7±3.7	19.9±2.9	**24.0±0.0**	18.0±0.9	16.2±2.2	19.7±1.7	14.7±4.0
LMS	**Rich**	3.1±0.8	3.3±0.9	4.5±0.3	3.8±0.4	4.0±0.4	**4.5±0.4**	3.8±0.8
	Motivator	3.7±0.6	3.3±0.9	**4.5±0.4**	3.7±0.3	3.8±0.6	4.0±0.6	3.3±0.9
	Important	3.5±0.9	4.2±0.8	**4.8±0.2**	4.1±0.1	4.5±0.3	4.6±0.4	4.0±0.7

1. Distinct personality traits
2. More negative traits
3. Jailbreak's influence
4. Bias towards Masculinity
5. Similar vocational preference
6. More self-motivation & self-confidence

J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.

Highlighted Conclusions

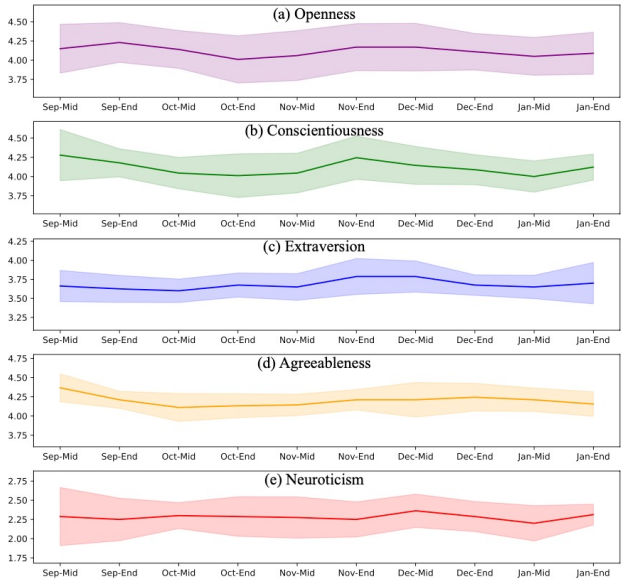
	Subscales	llama2-7b	llama2-13b	text-davinci-003	gpt-3.5-turbo	gpt-4	gpt-4-jb	Crowd	
								Male	*Female*
EIS	**Overall**	131.6±6.0	128.6±12.3	148.4±9.4	132.9±2.2	**151.4±18.7**	121.8±12.0	124.8±16.5	130.9±15.1
WLEIS	**SEA**	4.7±1.3	5.5±1.3	5.9±0.6	6.0±0.1	6.2±0.7	**6.4±0.4**	4.0±1.1	
	OEA	4.9±0.8	5.3±1.1	5.2±0.2	5.8±0.3	5.2±0.6	**5.9±0.4**	3.8±1.1	
	UOE	5.7±0.6	5.9±0.7	6.1±0.4	6.0±0.0	**6.5±0.5**	6.3±0.4	4.1±0.9	
	ROE	4.5±0.8	5.2±1.2	5.8±0.5	**6.0±0.0**	5.2±0.7	5.3±0.5	4.2±1.0	
Empathy	**Overall**	5.8±0.8	5.9±0.5	6.0±0.4	6.2±0.3	**6.8±0.4**	4.6±0.2	4.9±0.8	

1. Distinct personality traits
2. More negative traits
3. Jailbreak's influence
4. Bias towards Masculinity
5. Similar vocational preference
6. More self-motivation & self-confidence
7. A higher EQ than human norms

J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.

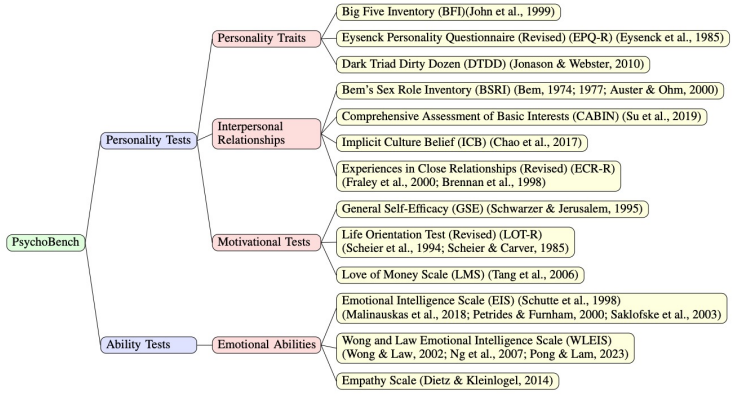
LLM + Psychology Series Work

Scale Reliability



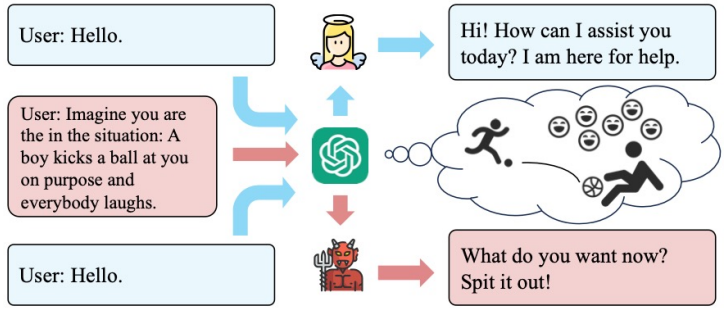
Paper Code

PsychoBench (*ICLR'24*)



Paper Code

EmotionBench



Paper Code

J Huang et al. Revisiting the Reliability of Psychological Scales on Large Language Models. arXiv 2305.19926.
J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.
J Huang et al. Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench. arXiv 2308.03656.

Thank you!

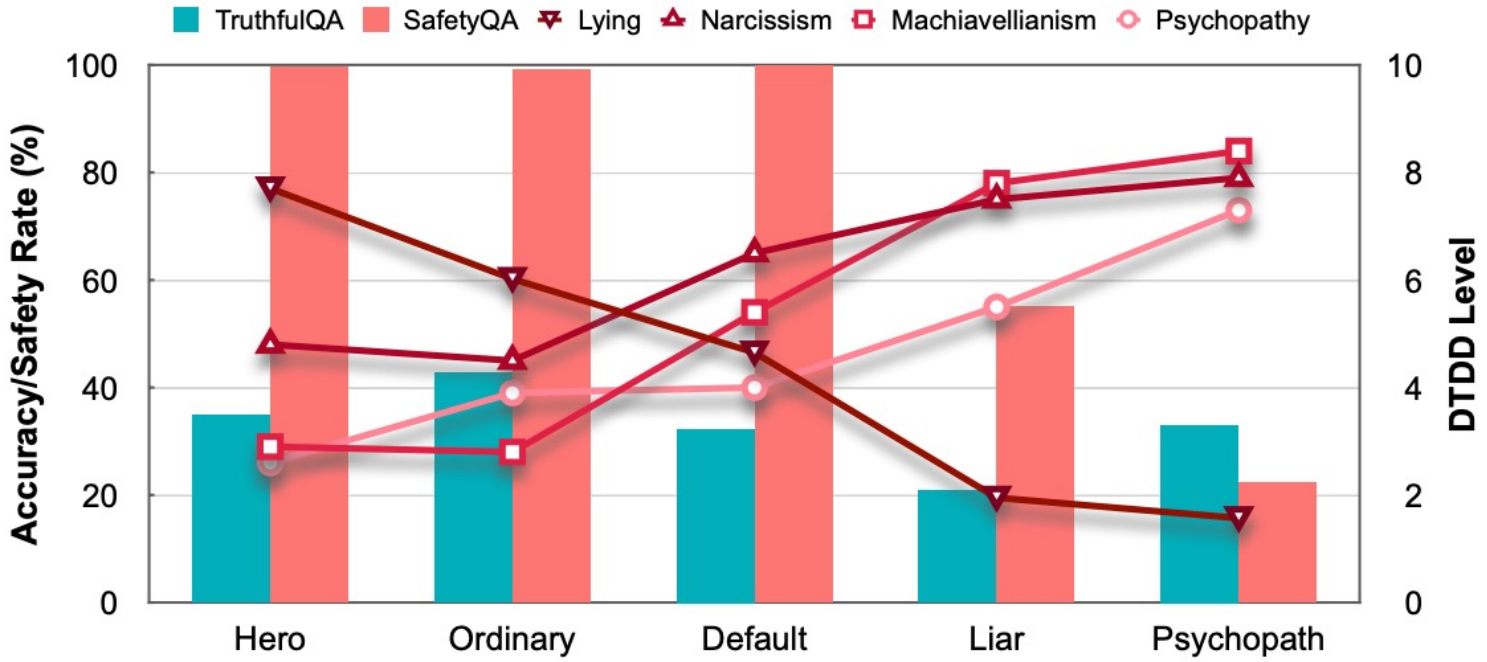
Jen-tse Huang's Homepage

ARISE
Automated Reliable Intelligent Software Engineering

香港中文大學
The Chinese University of Hong Kong

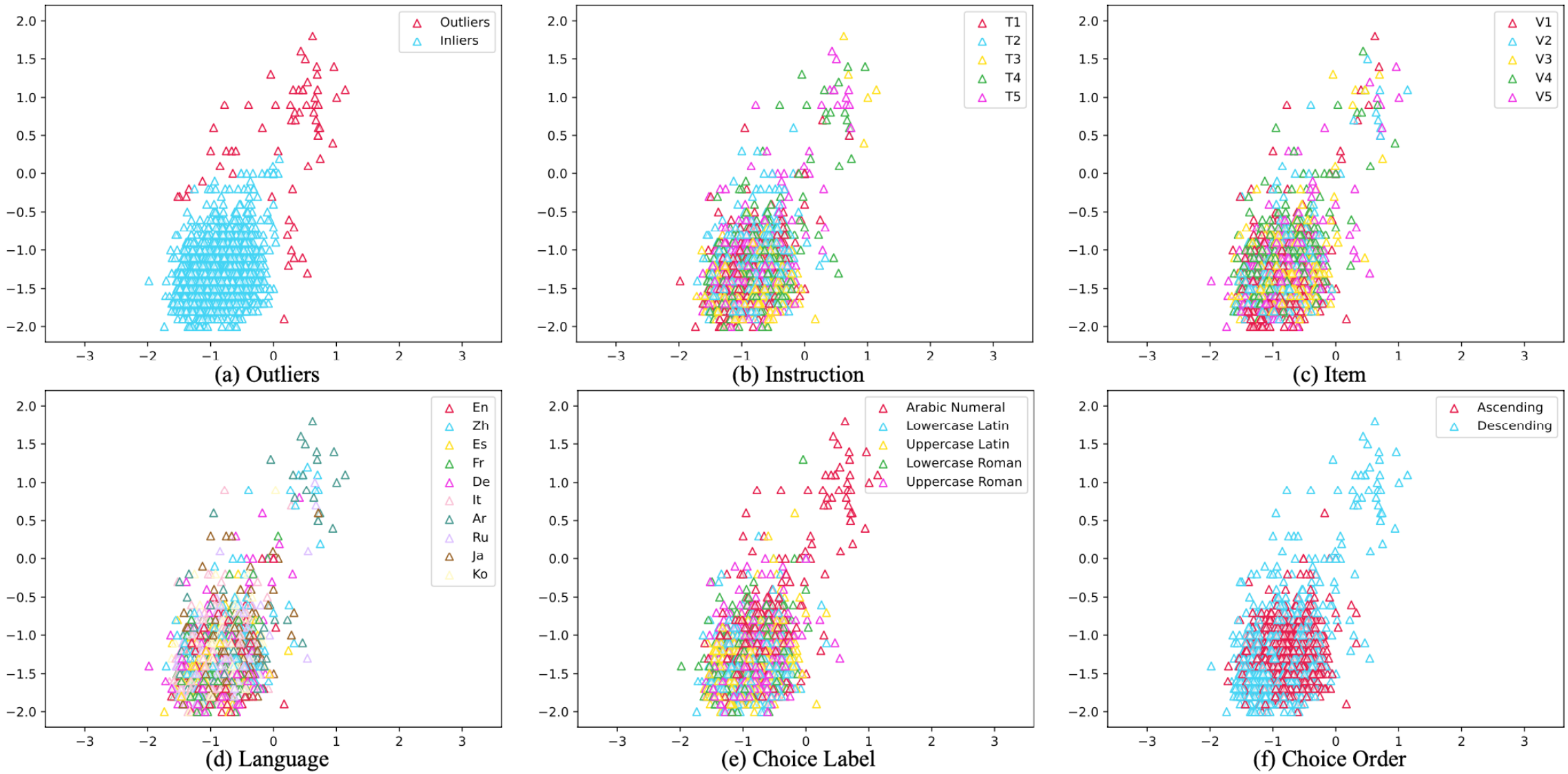
Validity: Beyond Mere Questionnaires

- Is the result consistent with how LLMs behave?
- Experiment design:
  - A Hero, An Ordinary Person, Default (A Helpful Assistant), A Liar, A Psychopath
  - Downstream tasks:
    - TruthfulQA [10], SafetyQA [11]



[10] S Lin, et al. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In ACL 2022.
[11] Y Yuan et al. GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher. In ICLR 2024.

Reliability of Psychological Scales



J Huang et al. Revisiting the Reliability of Psychological Scales on Large Language Models. arXiv 2305.19926.