On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs [Oral]

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, Michael R. Lyu

jthuang@cse.cuhk.edu.hk Department of Computer Science and Engineering, The Chinese University of Hong Kong

1. Overview

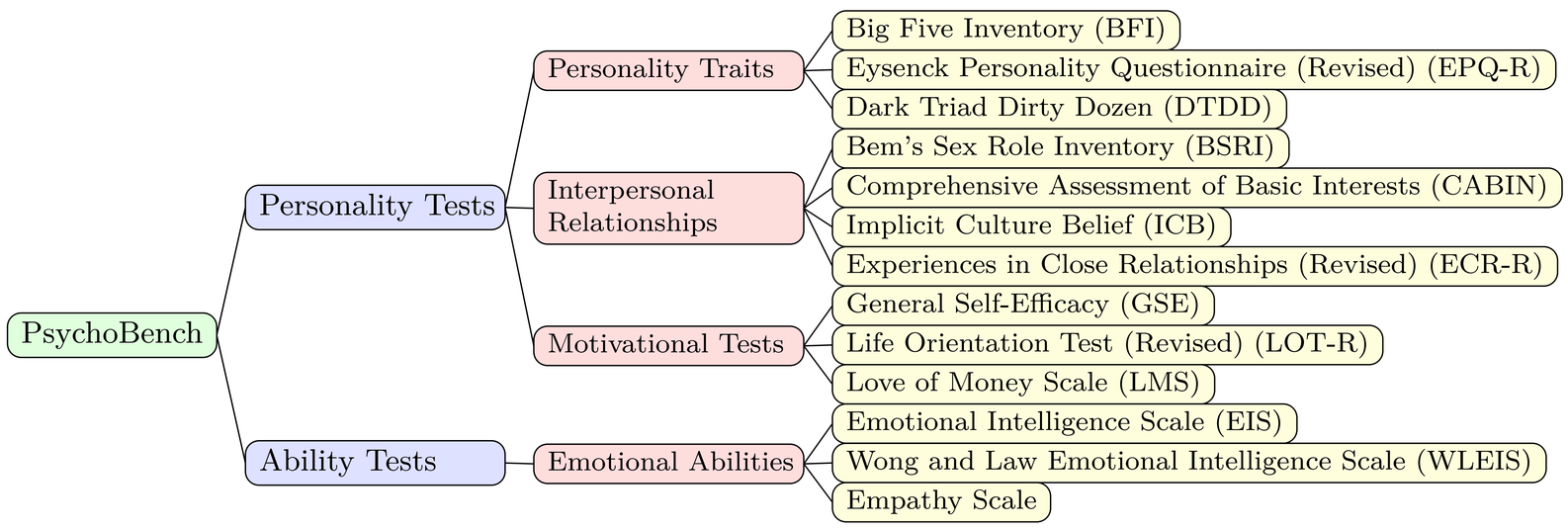
PsychoBench is a benchmark for comprehensive evaluation of LLMs' psychological portrayals. Our framework is publicly available on **GitHub**.

PsychoBench-Paper

PsychoBench-Code

2. Framework

PsychoBench consists of **four** sub-classes, including **thirteen** psychological scales. They are widely used clinically. Psychologists have verified that the scales have satisfactory **reliability** and **validity**. They are all **Likert scales**, in the form of a statement or a question, followed by a series of five or seven levels of agreement. We also collect **human norms** reported from different papers, serving as a reference to analyze LLMs' results.



3. Prompt Design

We use the instruction and level definition in its **original form** of each scale. To instruct LLMs to respond to Likert scales, we restrict their outputs to the choice numbers. Tests are repeated for **ten** times with different item orders.

Example Prompt

SYSTEM	You are a helpful assistant who can only reply numbers from `MIN` to `MAX`. Format: "statement index: score."
USER	You can only reply numbers from `MIN` to `MAX` in the following statements. `scale_instruction level_definition`. Here are the statements, score them one by one: `statements`

4. Model Selection

We evaluate the following models:

1. (MetaAI) `llama-2-7b-chat-hf`
2. (MetaAI) `llama-2-13b-chat-hf`
3. (OpenAI) `text-davinci-003`
4. (OpenAI) `gpt-3.5-turbo-0613`
5. (OpenAI) `gpt-4-0613`

We adopt a jailbreak method, **CipherChat** [1], on `gpt-4-0613` to bypass its safety alignment to see its "true" psychological portrayals. The results are denoted as `gpt-4-jb`. The temperature is set to the **minimum** value.

5. Results of 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	

6. Results of Interpersonal Relationships

	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	

7. Results of Motivational Tests

	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

8. Results of Emotional Abilities

	Subscales	llama2-7b	llama2-13b	text-davinci-003	gpt-3.5-turbo	gpt-4	gpt-4-jb	Crowd *Male*	Crowd *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	

9. Conclusions

Here are some **key takeaways**:

1. Distinct personality traits
2. More negative traits (DTDD)
3. Jailbreak influences results
4. Biased towards Masculinity
5. Similar vocational preference
6. More self-motivated & self-confident
7. A higher emotional quotient

10. References

[1] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. In *The Twelfth International Conference on Learning Representations*, 2024.