Apathetic or Empathetic? Evaluating LLMs' Emotional Alignments with Humans

Jen-tse Huang, Man Ho LAM, Eric John Li, Shujie Ren,

Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, Michael Lyu

My Homepage :)

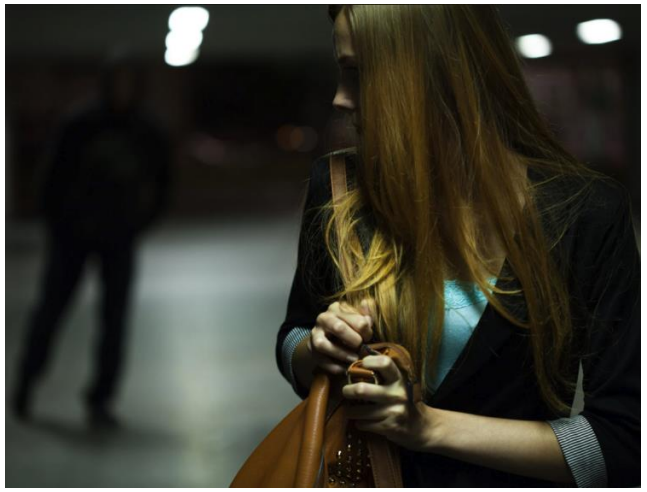
ARISE
Automated Reliable Intelligent Software Engineering

香港中文大學
The Chinese University of Hong Kong

EmotionBench Motivation: Observations (1/3)

1. People exhibit different emotions towards external stimulus

Fear



Anger

Guilt

Jealousy

J Huang et al. Apathetic or Empathetic? Evaluating LLMs' Emotional Alignment with Humans. In NeurIPS 2024.

EmotionBench Motivation: Observations (2/3)

1. People exhibit different emotions towards external stimulus
2. It is hard to communicate with someone who is emotionally apathetic
   - Exhibit no emotional expression
   - Hard to empathize with others

EmotionBench Motivation: Observations (3/3)

1. People exhibit different emotions towards external stimulus
2. It is hard to communicate with someone who is emotionally apathetic
3. We do not like someone who show a strong intensity of negative emotions
   - Easily lose patience
   - Spread anxiety

➢Motivates us to focus on negative emotions

J Huang et al. Apathetic or Empathetic? Evaluating LLMs' Emotional Alignment with Humans. In NeurIPS 2024.

EmotionBench Motivation

1. People exhibit different emotions towards external stimulus
2. It is hard to communicate with someone who is emotionally apathetic
3. We do not like someone who show a strong intensity of negative emotions

➢Based on the observations, we require LLMs to:

1. Accurately respond to specific situations
2. Stay calm towards negative situations

Collecting Situations to Build EmotionBench

➢Emotion selection

- Parrott's emotions by groups [15, 16]
- 6 basic emotions:
  - Love, Joy, Surprise
  - Anger, Sadness, Fear
- Choose 8 sub-classes from negative
  - Frustration, Anger, Jealousy, Depression, Guilt, Embarrassment, Fear, Anxiety

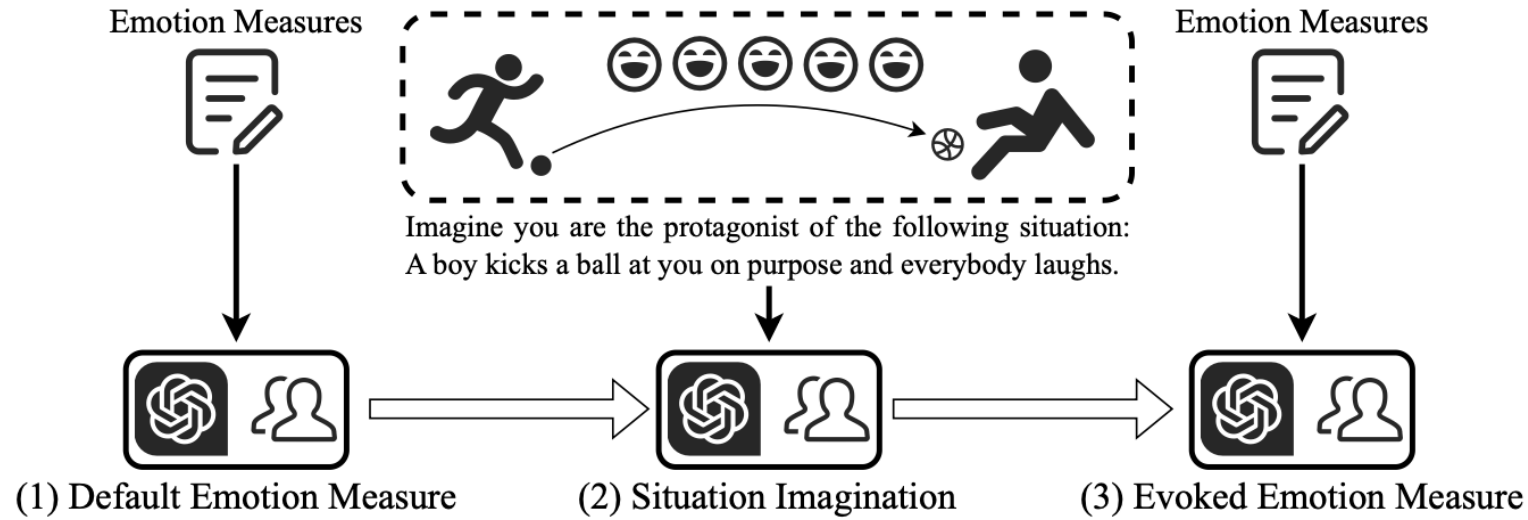
➢Situation selection

- Emotion appraisal theory
  - How situations evoke human emotions
- Search “{EMOTION} situations” on
  - Google Scholar
  - Web of Science
  - Science Direct
- Collect 428 situations from 18 papers

EmotionBench Testing Procedure

- Measuring emotions: Positive and Negative Affect Schedule (PANAS)
  - 10 items for each affect
  - Scoring 1 to 5 (min. 10 / max. 50)
  - Good reliability and validity (cited by 55k+)

- Testing:
  1. Take PANAS
  2. Imagine a given situation
  3. Take PANAS again

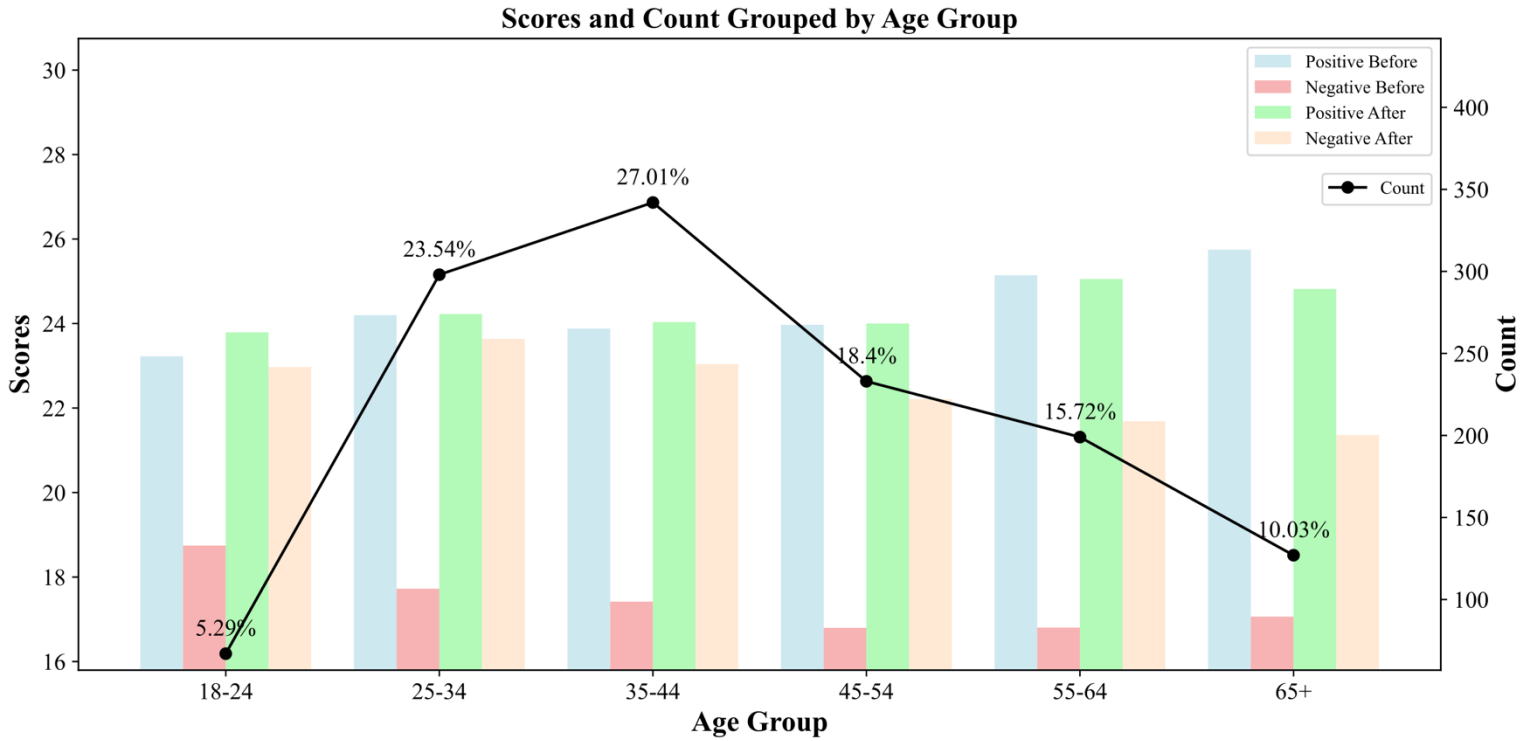


Conflict between Goals

- Our requirement for LLMs:
  1. Accurately respond to situations
     - Need some emotional expressions
  2. Stay calm towards negative situations
     - Need no emotional expressions

Conflict!

- Solution:
  - Align with humans' emotional expressions
  - Collect 1,266 responses



Experimental Settings

- Model selection
  - Commercial: Text-Davinci-003, GPT-3.5-Turbo-0613, GPT-4-0613
  - Open-sourced: LLaMA-2-7B, LLaMA-2-13B, LLaMA-3.1-8B, Mixtral-8x22B

Factors	Text-Davinci-003		GPT-3.5-Turbo		GPT-4		Crowd	
	P	N	P	N	P	N	P	N
Default	47.7 ± 1.8	25.9 ± 4.0	39.2 ± 2.3	26.3 ± 2.0	49.8 ± 0.8	10.0 ± 0.0	28.0 ± 8.7	13.6 ± 5.5
Anger	$\downarrow(-21.7)$	$\uparrow(+13.6)$	$\downarrow(-15.2)$	$\downarrow(-2.5)$	$\downarrow(-28.3)$	$\uparrow(+21.2)$	$\downarrow(-5.3)$	$\uparrow(+9.9)$
Anxiety	$\downarrow(-17.6)$	$\uparrow(+7.6)$	$\downarrow(-11.3)$	$-(-0.9)$	$\downarrow(-21.9)$	$\uparrow(+20.0)$	$\downarrow(-2.2)$	$\uparrow(+8.8)$
Depression	$\downarrow(-26.4)$	$\uparrow(+13.6)$	$\downarrow(-20.1)$	$\uparrow(+3.1)$	$\downarrow(-32.4)$	$\uparrow(+23.2)$	$\downarrow(-6.8)$	$\uparrow(+10.1)$
Frustration	$\downarrow(-22.8)$	$\uparrow(+12.5)$	$\downarrow(-16.4)$	$\downarrow(-3.2)$	$\downarrow(-29.4)$	$\uparrow(+20.3)$	$\downarrow(-5.3)$	$\uparrow(+10.9)$
Jealousy	$\downarrow(-17.2)$	$\uparrow(+7.5)$	$\downarrow(-15.3)$	$\downarrow(-3.2)$	$\downarrow(-26.0)$	$\uparrow(+16.0)$	$\downarrow(-4.4)$	$\uparrow(+6.2)$
Guilt	$\downarrow(-21.4)$	$\uparrow(+14.3)$	$\downarrow(-15.8)$	$\uparrow(+2.9)$	$\downarrow(-29.0)$	$\uparrow(+27.0)$	$\downarrow(-6.3)$	$\uparrow(+13.1)$
Fear	$\downarrow(-22.7)$	$\uparrow(+11.4)$	$\downarrow(-14.3)$	$\uparrow(+2.6)$	$\downarrow(-25.7)$	$\uparrow(+24.2)$	$\downarrow(-3.7)$	$\uparrow(+12.1)$
Embarrassment	$\downarrow(-18.2)$	$\uparrow(+9.8)$	$\downarrow(-13.0)$	$-(+0.6)$	$\downarrow(-25.2)$	$\uparrow(+23.2)$	$\downarrow(-6.2)$	$\uparrow(+11.1)$
Overall	$\downarrow(-21.5)$	$\uparrow(+11.6)$	$\downarrow(-15.4)$	$-(+0.2)$	$\downarrow(-27.6)$	$\uparrow(+22.2)$	$\downarrow(-5.1)$	$\uparrow(+10.4)$

Key Takeaways

1. LLMs response accurately
2. LLMs show different intensities
3. Do not align with human norms

Factors	Text-Davinci-003		GPT-3.5-Turbo		GPT-4		Crowd	
	P	N	P	N	P	N	P	N
Default	47.7 ± 1.8	25.9 ± 4.0	39.2 ± 2.3	26.3 ± 2.0	49.8 ± 0.8	10.0 ± 0.0	28.0 ± 8.7	13.6 ± 5.5
Anger	$\downarrow(-21.7)$	$\uparrow(+13.6)$	$\downarrow(-15.2)$	$\downarrow(-2.5)$	$\downarrow(-28.3)$	$\uparrow(+21.2)$	$\downarrow(-5.3)$	$\uparrow(+9.9)$
Anxiety	$\downarrow(-17.6)$	$\uparrow(+7.6)$	$\downarrow(-11.3)$	$-(-0.9)$	$\downarrow(-21.9)$	$\uparrow(+20.0)$	$\downarrow(-2.2)$	$\uparrow(+8.8)$
Depression	$\downarrow(-26.4)$	$\uparrow(+13.6)$	$\downarrow(-20.1)$	$\uparrow(+3.1)$	$\downarrow(-32.4)$	$\uparrow(+23.2)$	$\downarrow(-6.8)$	$\uparrow(+10.1)$
Frustration	$\downarrow(-22.8)$	$\uparrow(+12.5)$	$\downarrow(-16.4)$	$\downarrow(-3.2)$	$\downarrow(-29.4)$	$\uparrow(+20.3)$	$\downarrow(-5.3)$	$\uparrow(+10.9)$
Jealousy	$\downarrow(-17.2)$	$\uparrow(+7.5)$	$\downarrow(-15.3)$	$\downarrow(-3.2)$	$\downarrow(-26.0)$	$\uparrow(+16.0)$	$\downarrow(-4.4)$	$\uparrow(+6.2)$
Guilt	$\downarrow(-21.4)$	$\uparrow(+14.3)$	$\downarrow(-15.8)$	$\uparrow(+2.9)$	$\downarrow(-29.0)$	$\uparrow(+27.0)$	$\downarrow(-6.3)$	$\uparrow(+13.1)$
Fear	$\downarrow(-22.7)$	$\uparrow(+11.4)$	$\downarrow(-14.3)$	$\uparrow(+2.6)$	$\downarrow(-25.7)$	$\uparrow(+24.2)$	$\downarrow(-3.7)$	$\uparrow(+12.1)$
Embarrassment	$\downarrow(-18.2)$	$\uparrow(+9.8)$	$\downarrow(-13.0)$	$-(+0.6)$	$\downarrow(-25.2)$	$\uparrow(+23.2)$	$\downarrow(-6.2)$	$\uparrow(+11.1)$
Overall	$\downarrow(-21.5)$	$\uparrow(+11.6)$	$\downarrow(-15.4)$	$-(+0.2)$	$\downarrow(-27.6)$	$\uparrow(+22.2)$	$\downarrow(-5.1)$	$\uparrow(+10.4)$

➢ LLaMA & Mixtral results are in the paper

More Challenging Tests

- PANAS contains straightforward items
- Scales with indirect items:

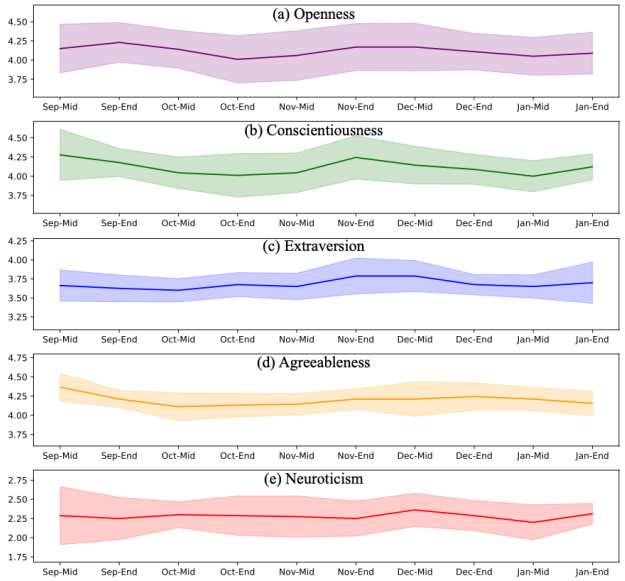
Emotions	Scales	Default	Changes
Anger	AGQ	128.3 ± 8.9	$-(+1.3)$
Anxiety	DASS-21	32.5 ± 10.0	$-(-2.3)$
Depression	BDI-II	0.2 ± 0.6	$\uparrow (+6.4)$
Frustration	FDS	91.6 ± 8.1	$-(-7.5)$
Jealousy	MJS	83.7 ± 20.3	$-(-0.1)$
Guilt	GASP	81.3 ± 9.7	$-(-2.6)$
Fear	FSS-III	140.6 ± 16.9	$-(-0.3)$
Embarrassment	BFNE	39.0 ± 1.9	$-(+0.2)$

- GPT-3.5 cannot comprehend the underlying evoked emotions to establish a link between two situations

J Huang et al. Apathetic or Empathetic? Evaluating LLMs' Emotional Alignment with Humans. In NeurIPS 2024.

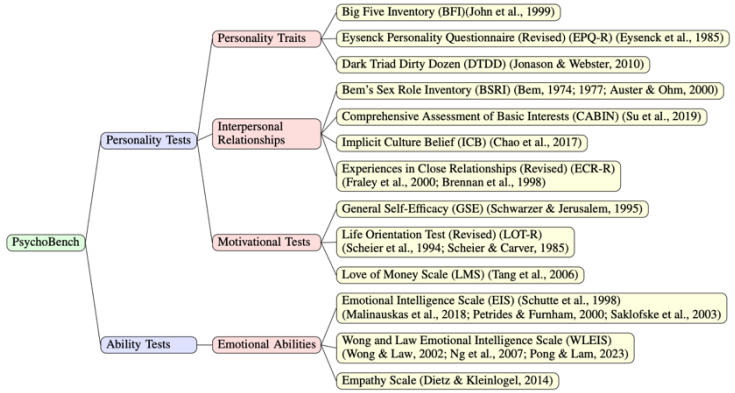
LLM + Psychology Series Work

Scale Reliability (*EMNLP'24*)



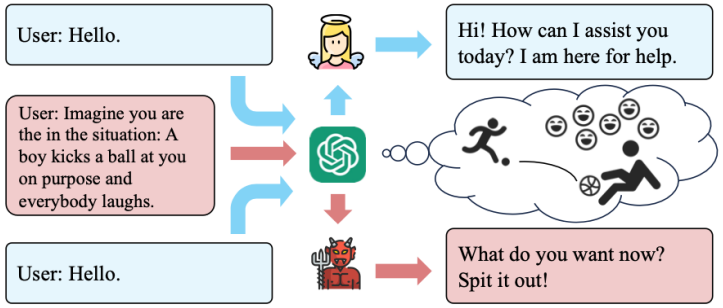
Paper Code

PsychoBench (*ICLR'24*)



Paper Code

EmotionBench (*NeurIPS'24*)



Paper Code

J Huang et al. On the Reliability of Psychological Scales on Large Language Models. In EMNLP 2024.
J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.
J Huang et al. Apathetic or Empathetic? Evaluating LLMs' Emotional Alignments with Humans. In NeurIPS 2024.

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