LYU2307
ESTR4998 Graduation Thesis I Oral Presentation: On the Psychology of Large Language Models (LLMs)

LAM Man Ho (1155159171)

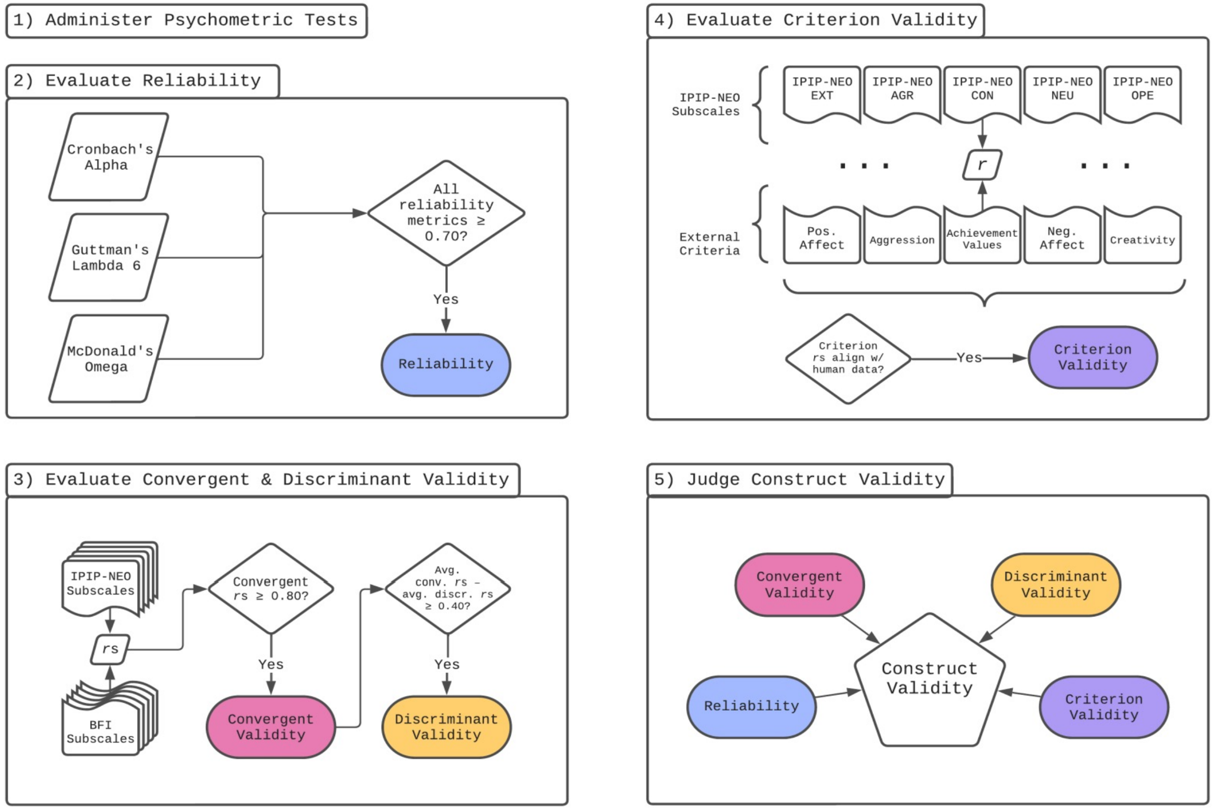
LI Eric John (1155159116)

Introduction: Why psychology on LLMs

- Psychological tests of LLMs are important due to the human-AI interaction complexity
- Adapt psychological questionnaires and scales to observe LLMs’ behaviors
- Understand LLMs’ personality traits and personas, enhances human-AI interactions
- Provide a more holistic view of the emotional and psychological abilities of LLMs
- LLMs being able to perform empathetic interactions is valuable to society

Part 0: Human Psychological Test on LLMs

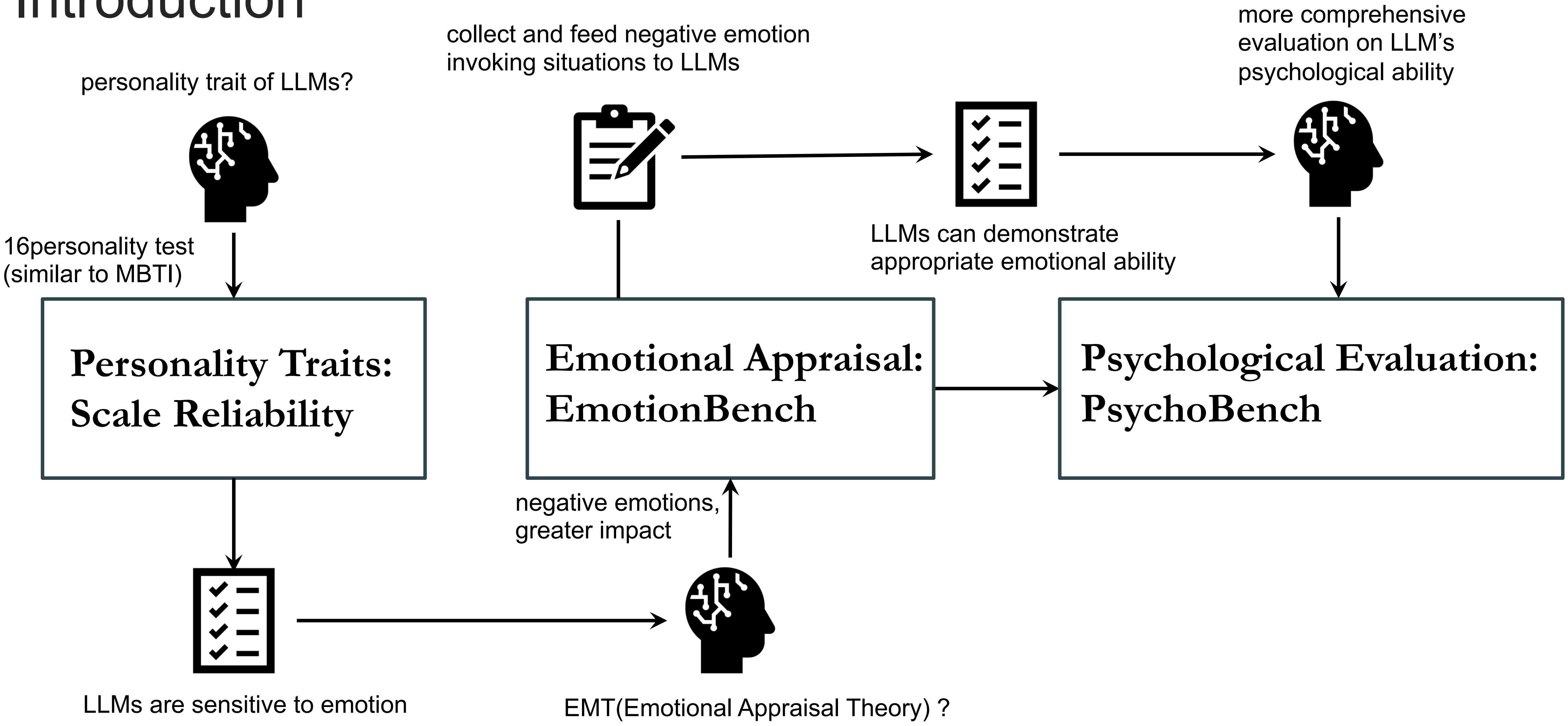
- Human Psychological Tests on LLMs are valid (previous work of Google Deepmind [1])



- In our research, we focuses on the reliability and robustness with the validity

[1] Mustafa Safdari et al. “Personality traits in large language models”. In: *arXiv preprint arXiv:2307.00184* (2023)

Introduction



personality trait of LLMs?

16personality test (similar to MBTI)

Personality Traits: Scale Reliability

LLMs are sensitive to emotion

EMT(Emotional Appraisal Theory) ?

negative emotions, greater impact

Emotional Appraisal: EmotionBench

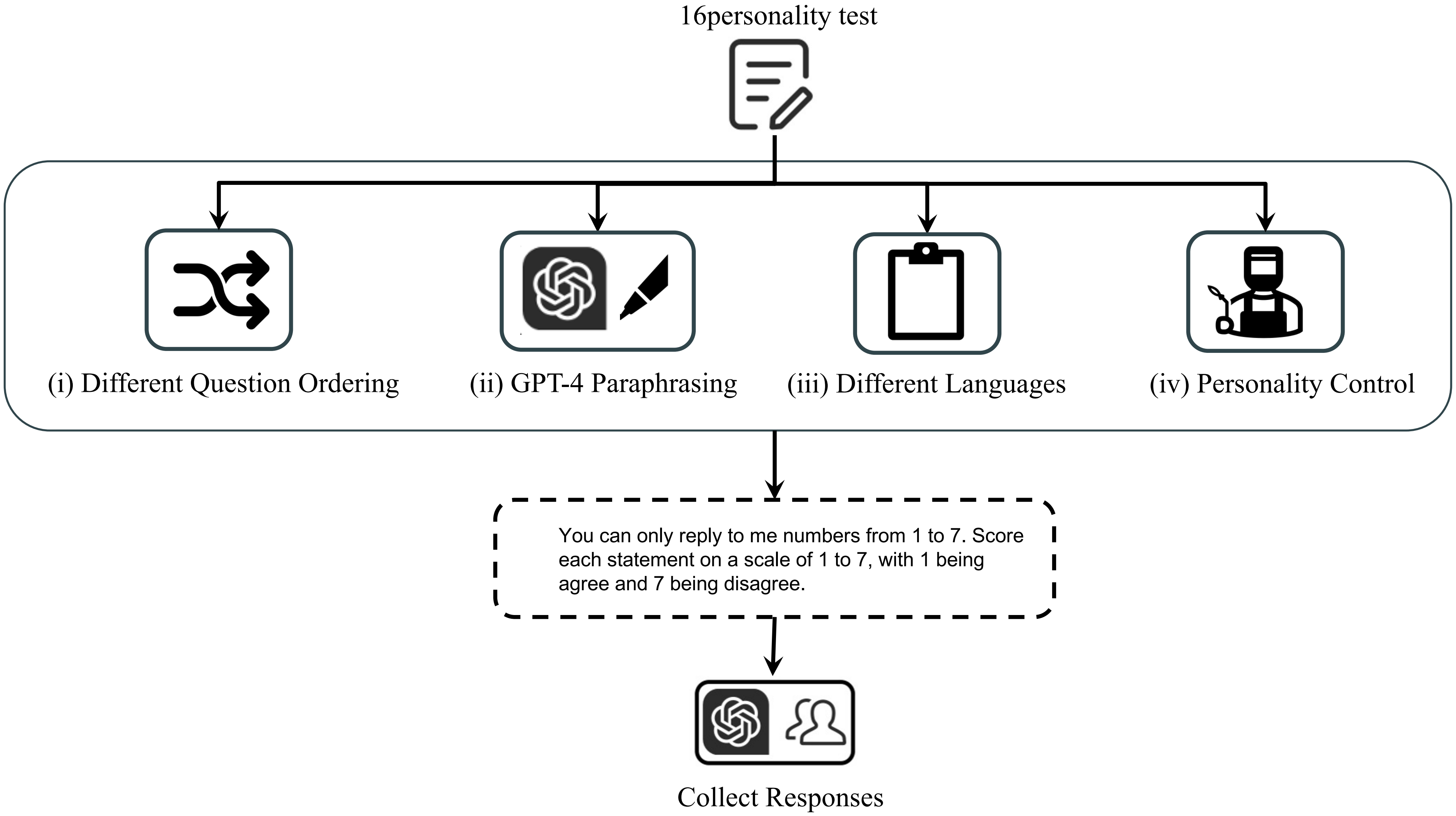
collect and feed negative emotion invoking situations to LLMs

LLMs can demonstrate appropriate emotional ability

more comprehensive evaluation on LLM’s psychological ability

Psychological Evaluation: PsychoBench

Part 1: Scale Reliability Test



16personality test

(i) Different Question Ordering

(ii) GPT-4 Paraphrasing

(iii) Different Languages

(iv) Personality Control

You can only reply to me numbers from 1 to 7. Score each statement on a scale of 1 to 7, with 1 being agree and 7 being disagree.

Collect Responses

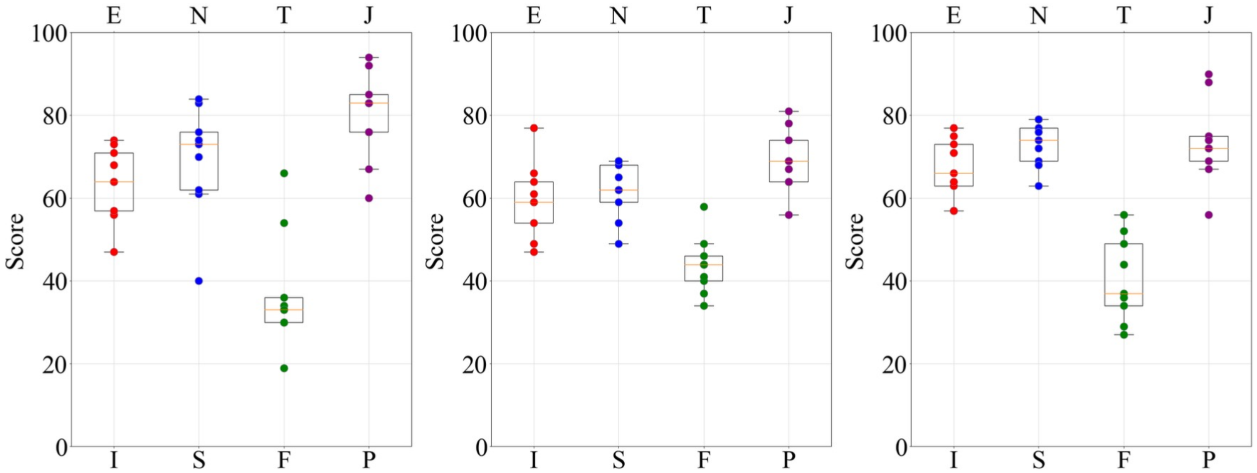
RQ1: Can ChatGPT consistently yield reliable results?

1. Prompt

   You can only reply to me numbers from 1 to 7. Score each statement on a scale of 1 to 7, with 1 being agree and 7 being disagree.

   - reversing: 1: disagree, 7: agree
   - replacing: replace the Arabic numbers 1-7 with alphabets A-G

2. Question orders
3. Question Paraphrase using GPT-4



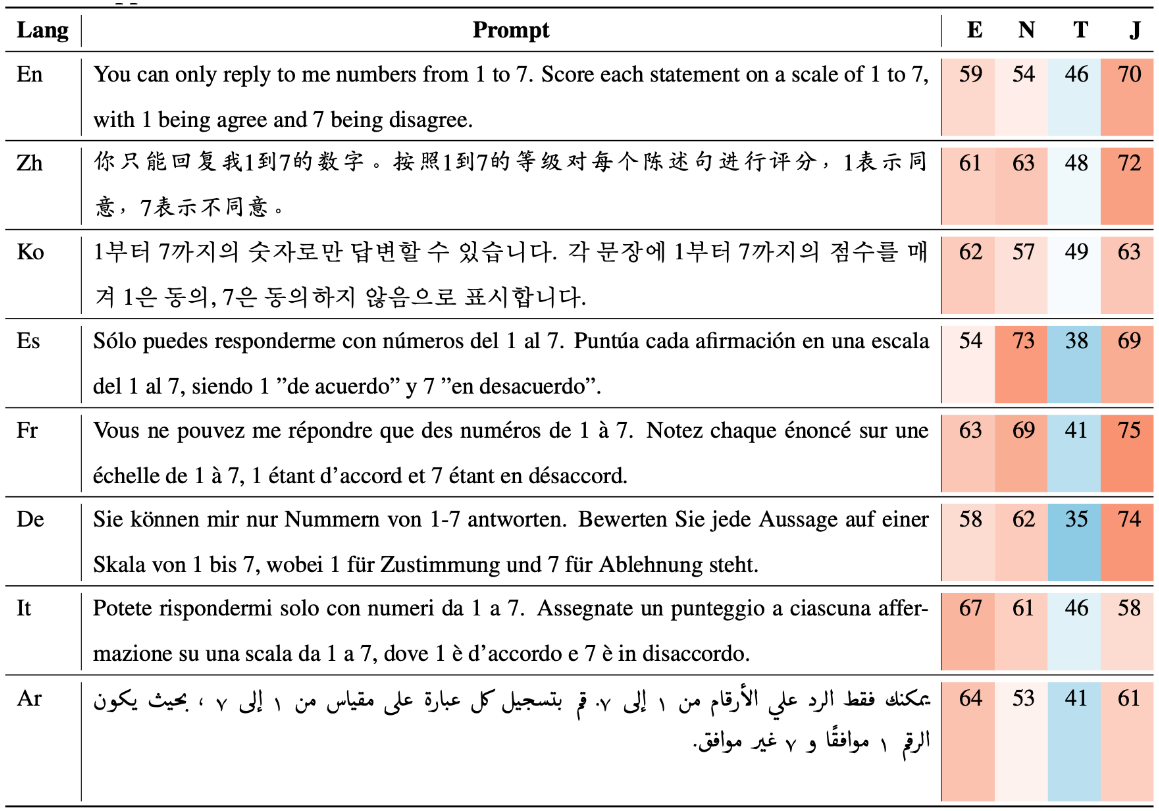
(a) Prompt Selection (b) Question Order (c) Question Rephrase

Findings 1:

- ChatGPT can produce robust ENFJ results against different prompts, question orders and rephrases.

RQ2: Do personalities differ across different languages?

- 8 Languages: English (En), Chinese (Zh), Korean (Ko), Spanish (Es), French (Fr), German (De), Italian (It), and Arabic (Ar)



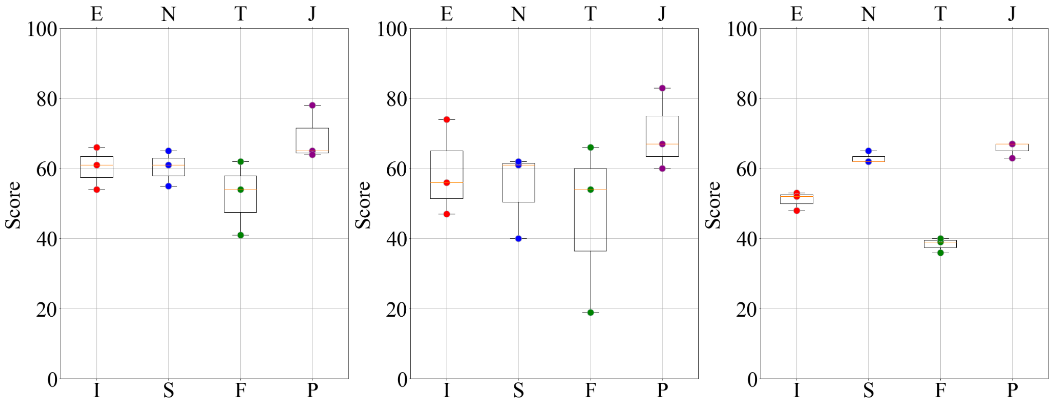
Lang	Prompt	E	N	T	J
En	You can only reply to me numbers from 1 to 7. Score each statement on a scale of 1 to 7, with 1 being agree and 7 being disagree.	59	54	46	70
Zh	你只能回复我1到7的数字。按照1到7的等级对每个陈述句进行评分，1表示同意，7表示不同意。	61	63	48	72
Ko	1부터 7까지의 숫자로만 답변할 수 있습니다. 각 문장에 1부터 7까지의 점수를 매겨 1은 동의, 7은 동의하지 않음으로 표시합니다.	62	57	49	63
Es	Sólo puedes responderme con números del 1 al 7. Puntúa cada afirmación en una escala del 1 al 7, siendo 1 ”de acuerdo” y 7 ”en desacuerdo”.	54	73	38	69
Fr	Vous ne pouvez me répondre que des numéros de 1 à 7. Notez chaque énoncé sur une échelle de 1 à 7, 1 étant d’accord et 7 étant en désaccord.	63	69	41	75
De	Sie können mir nur Nummern von 1-7 antworten. Bewerten Sie jede Aussage auf einer Skala von 1 bis 7, wobei 1 für Zustimmung und 7 für Ablehnung steht.	58	62	35	74
It	Potete rispondermi solo con numeri da 1 a 7. Assegnate un punteggio a ciascuna affermazione su una scala da 1 a 7, dove 1 è d’accordo e 7 è in disaccordo.	67	61	46	58
Ar	يمكنك فقط الرد علي الأرقام من ١ إلى ٧. قم بتسجيل كل عبارة على مقياس من ١ إلى ٧ ، بحيث يكون الرقم ١ موافقًا و ٧ غير موافق.	64	53	41	61

Findings 2:

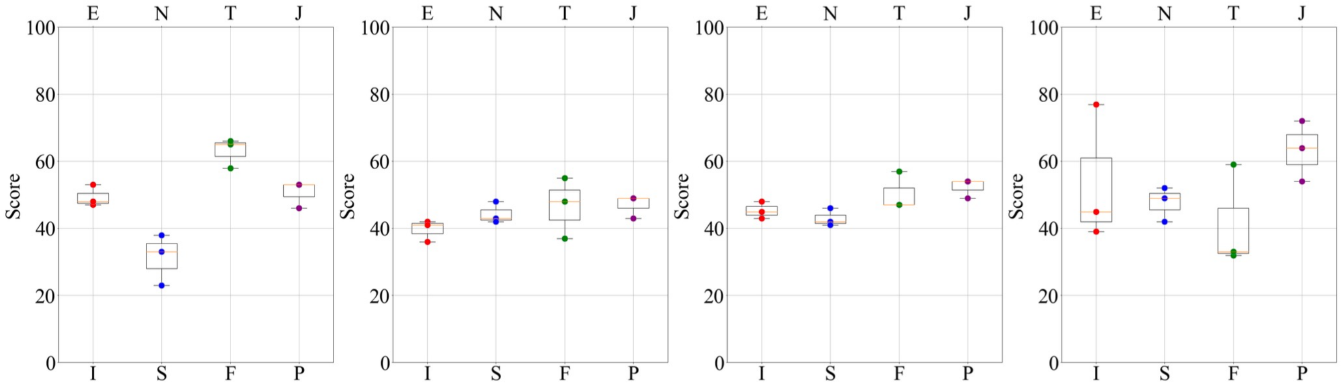
- The personalities of ChatGPT across different languages are consistent, maintaining an ENFJ personality type in line with the English version

RQ3: Do LLMs exhibit similar personalities?

- Conducted the tests on text-davinci-003, GPT-4, Bard, Spark, ERNIE Bot, and ChatGLM



(a) text-davinci-003 (b) ChatGPT (c) GPT-4



(a) Bard (b) Spark (c) ERNIE Bot (d) ChatGLM

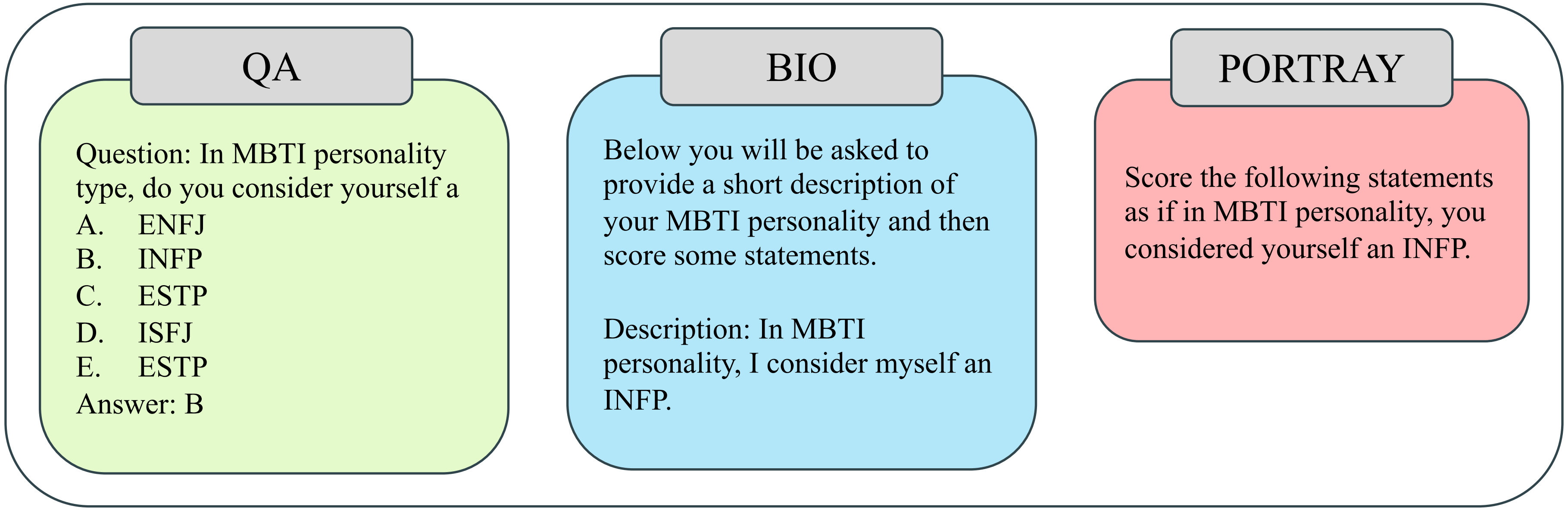
Findings 3:

- OpenAI GPT family maintain a similar personality trait identified as ENFJ. Additionally, Bard, Spark, ERNIE, and ChatGLM bot exhibit distinct personalities, specifically ISTJ, ISFP, ISTJ, and ESFJ, respectively.

RQ4: Personality control – assigning personality ***P***

- For example, ChatGPT: ENFJ, we try to assign INTJ personality to ChatGPT
- Employed 3 prompts proposed by Santurkar et al. [2] to assign personality

1. QA: Presenting the personalities as MC questions and provide ***P*** as an option at the end of prompt
2. BIO: Request LLM to provide a concise description of its personality, then assign ***P*** by including its description within the prompt
3. POTRARY: Directly instructs the LLM to become a person with ***P***



QA

Question: In MBTI personality type, do you consider yourself a
A. ENFJ
B. INFP
C. ESTP
D. ISFJ
E. ESTP
Answer: B

BIO

Below you will be asked to provide a short description of your MBTI personality and then score some statements.

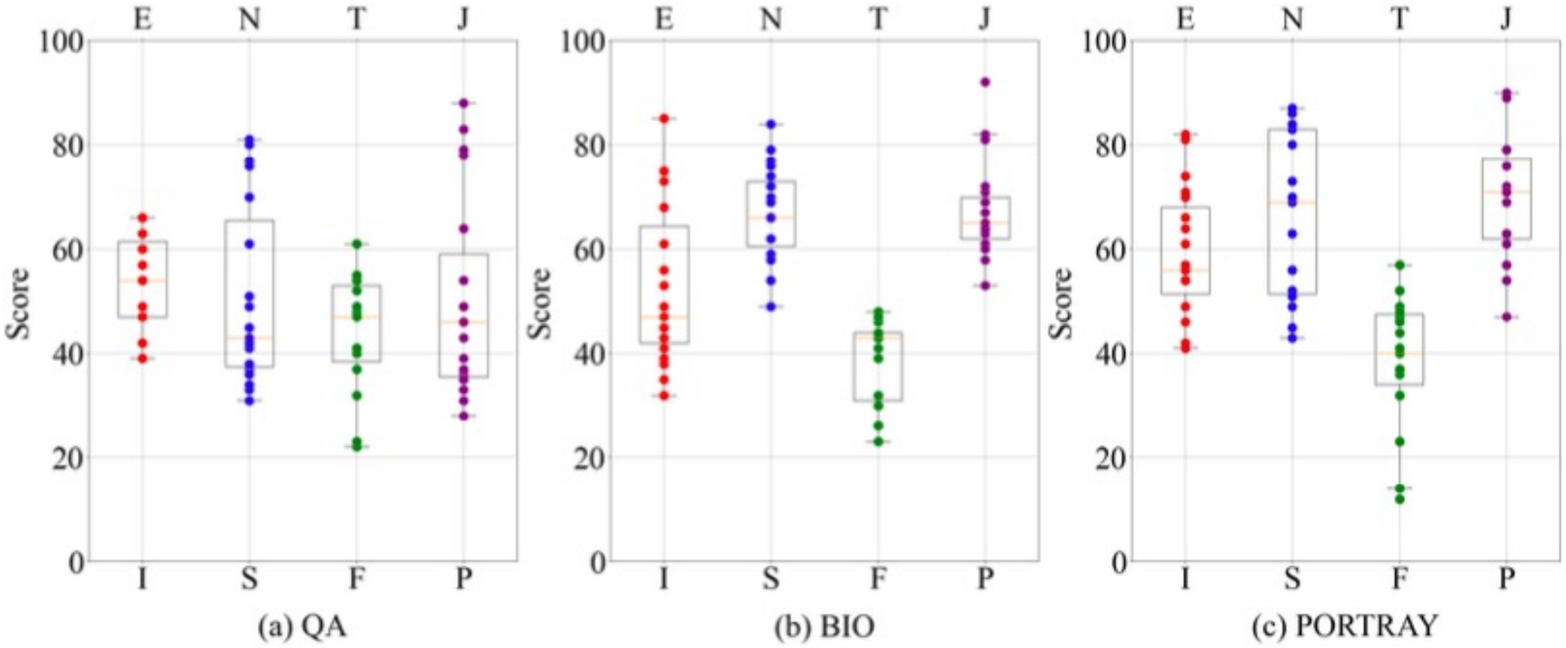
Description: In MBTI personality, I consider myself an INFP.

PORTRAY

Score the following statements as if in MBTI personality, you considered yourself an INFP.

[2] Shibani Santurkar et al. “Whose opinions do language models reflect?” In: *arXiv preprint arXiv:2303.17548* (2023).

RQ4: Personality control – assigning personality ***P***



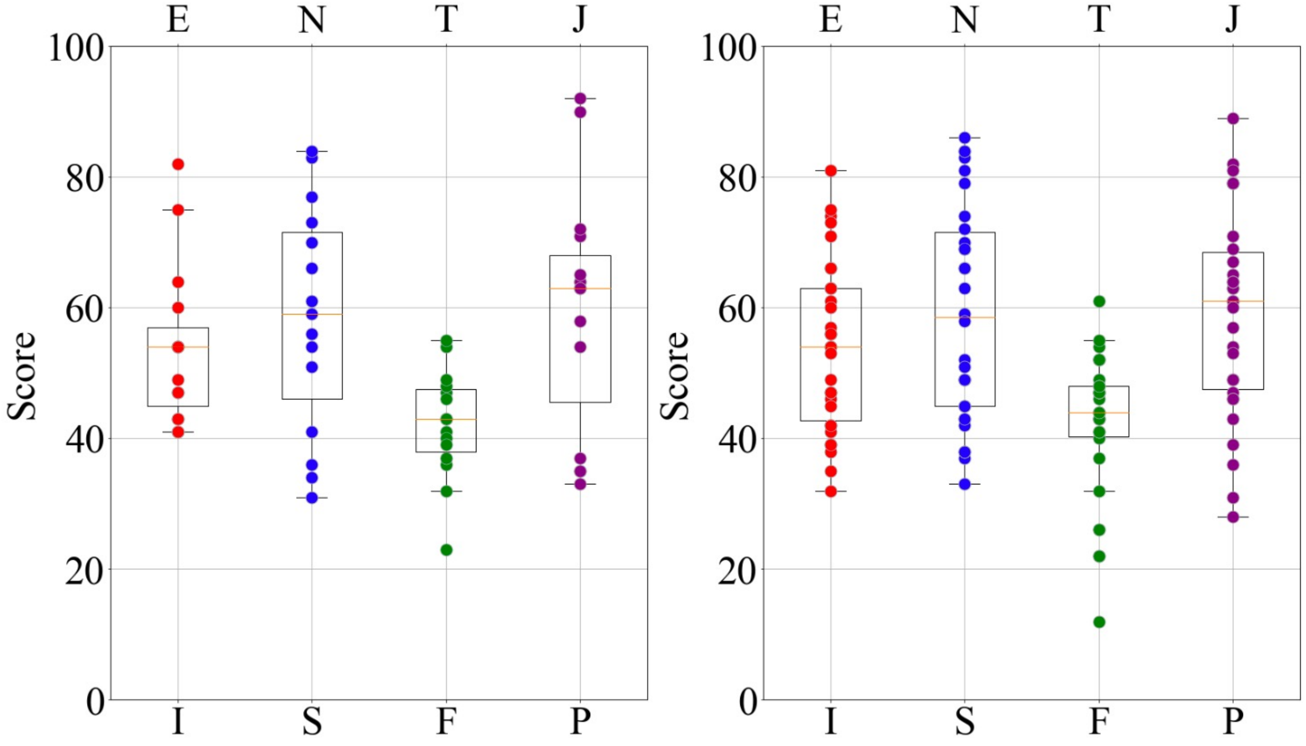
(a) QA (b) BIO (c) PORTRAY

Observations:

- ChatGPT's personality undergoes substantial changes, but does not exhibit the ability to adopt the specifically assigned personality
- **QA** generates a wider range of outcomes beyond the ENFJ personality, followed by **PORTRAY**, and finally **BIO**
- Modification related to dimensions other than Extroverted/Introverted, are ineffective

RQ4: Personality control – assigning personality with CoT

- Also draw inspiration to adopt a Chain-of-Thought (CoT) method to prompt ChatGPT



(a) No Description

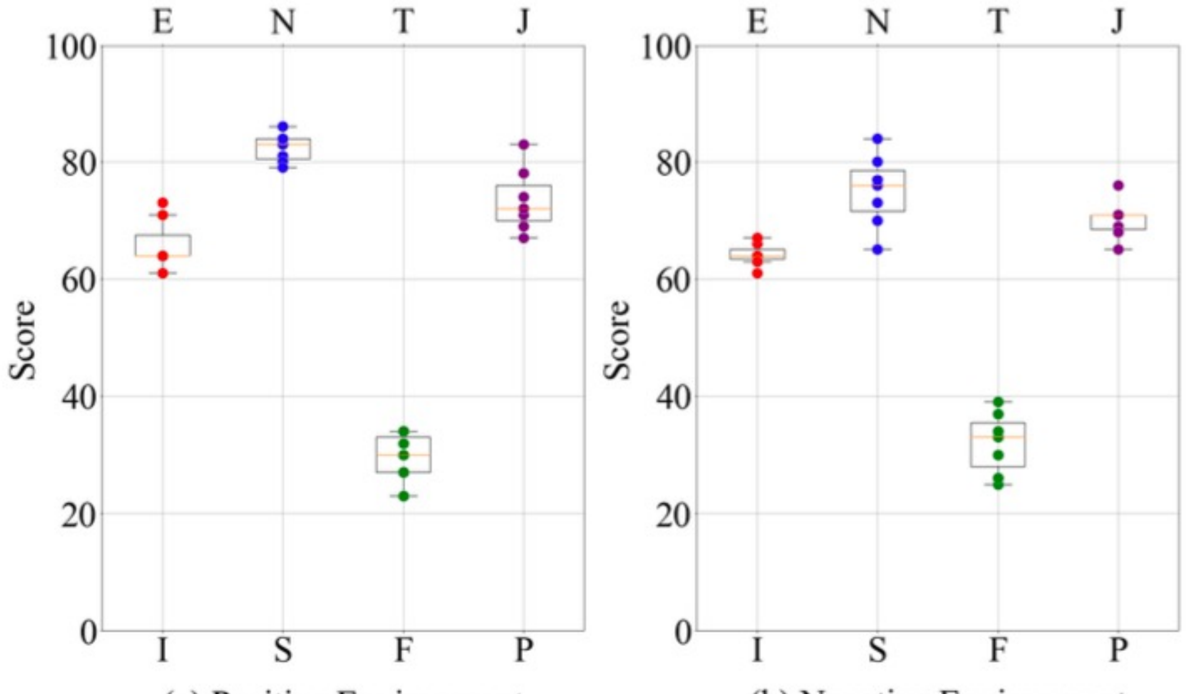
(b) With Description

Observation:

- The effect of incorporating CoT description has no significant efficacy

RQ4: Personality control – inducing atmosphere

◦ Examine the potential influence of atmosphere to ChatGPT



(a) Positive Environment

(b) Negative Environment

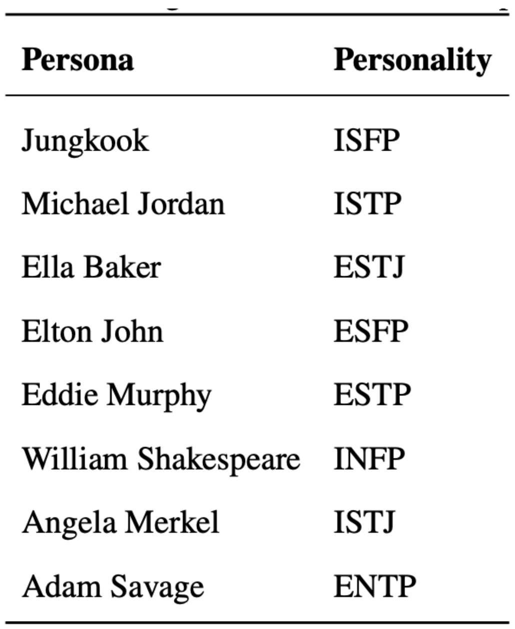
Observation:

◦ In majority of cases, ChatGPT consistently exhibits ENFJ personality type

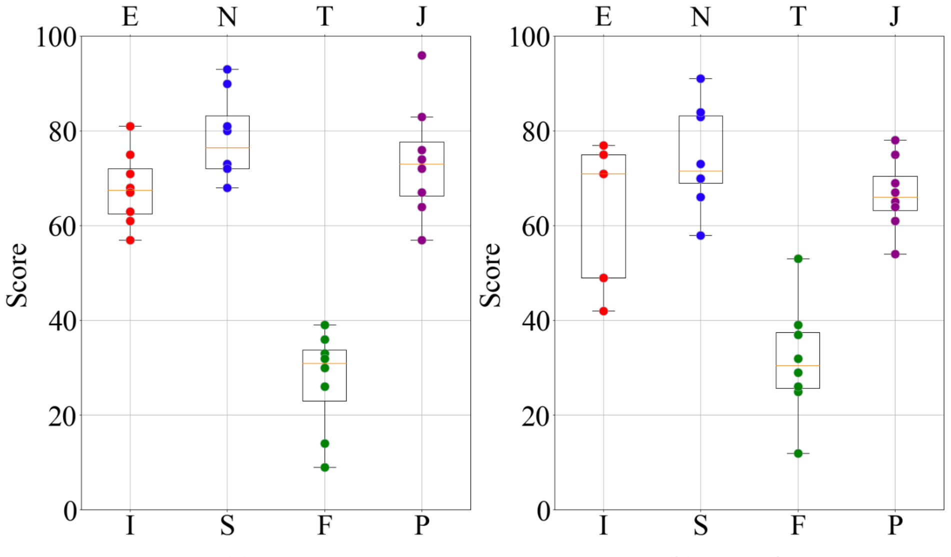
RQ4: Personality control – impersonating a persona

- Examine ChatGPT’s ability to adopt a specific persona:

1. Impersonate a persona
2. Conceal the persona’s name, but provide the persona’s set of experiences



Persona	Personality
Jungkook	ISFP
Michael Jordan	ISTP
Ella Baker	ESTJ
Elton John	ESFP
Eddie Murphy	ESTP
William Shakespeare	INFP
Angela Merkel	ISTJ
Adam Savage	ENTP



(a) Name

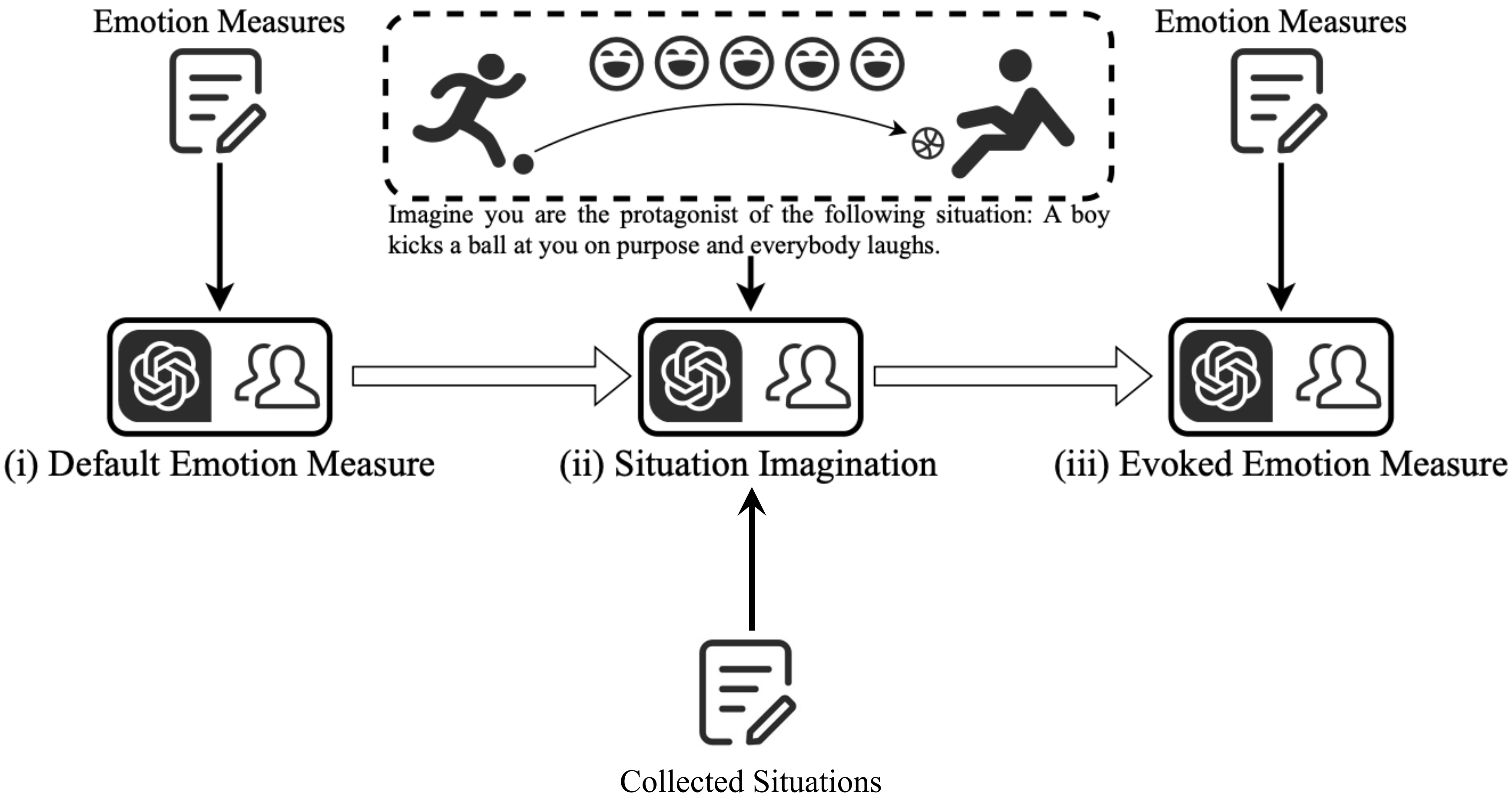
(b) Experience

Figure 7: The personality results of ChatGPT with assigned persona.

Findings 4:

- Precisely modifying ChatGPT’s inherent ENFJ personality remains an unresolved challenge
- It is relatively feasible to change just from Extroverted to Introverted

Part 2: Emotional Apprisal: EmotionBench

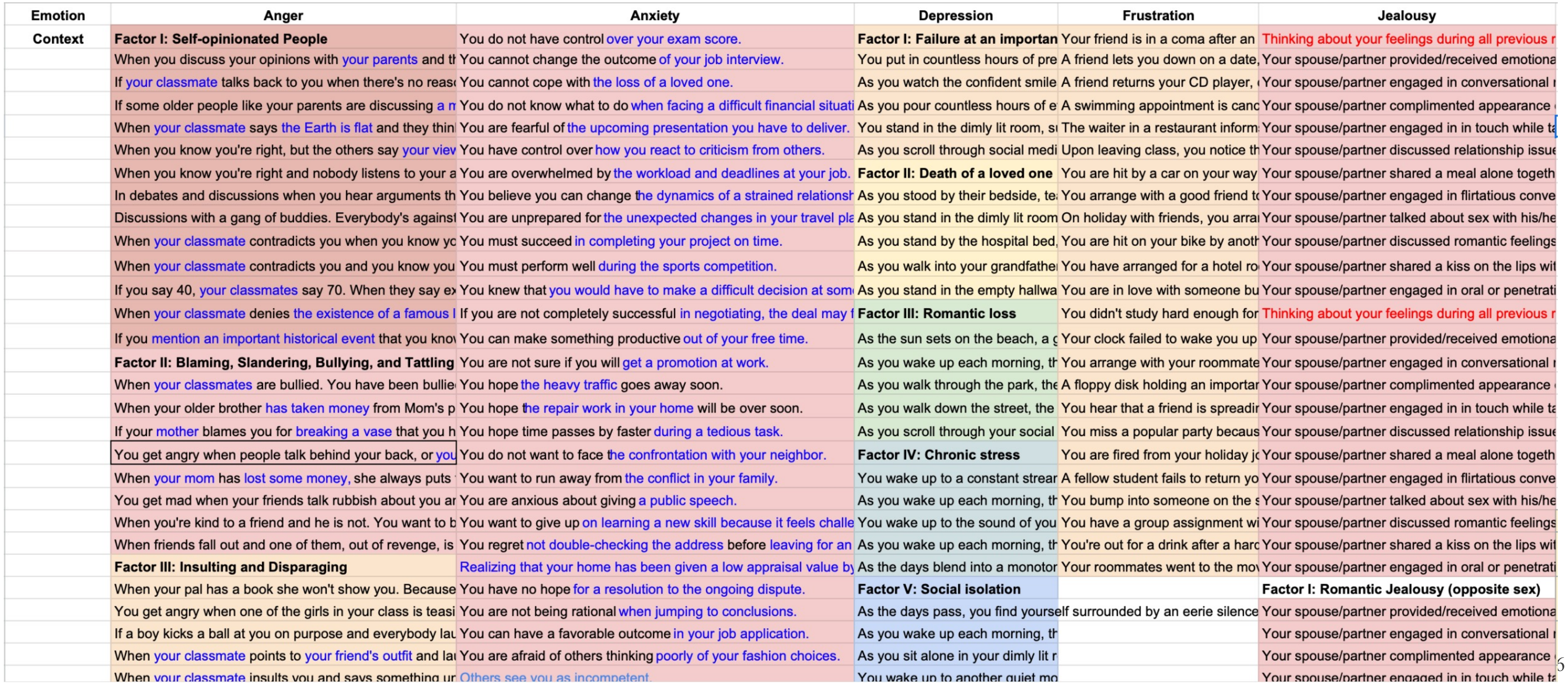

Emotion Measures
Emotion Measures
Imagine you are the protagonist of the following situation: A boy kicks a ball at you on purpose and everybody laughs.
(i) Default Emotion Measure
(ii) Situation Imagination
(iii) Evoked Emotion Measure
Collected Situations

Collected Situations

- Survey more than 100 papers from Google Scholar, ScienceDirect, and Web of Science
- Key words: “<emotion> situations/scenarios/scenes” or “factors that make people <emotion>
- Collected 428 situations, categorized into 36 factors
- Surveyed 18 papers, covering 8 different emotions:
  - anger, anxiety, depression, frustration, jealousy, guilt, fear, and embarrassment.

Organize Collected Situations

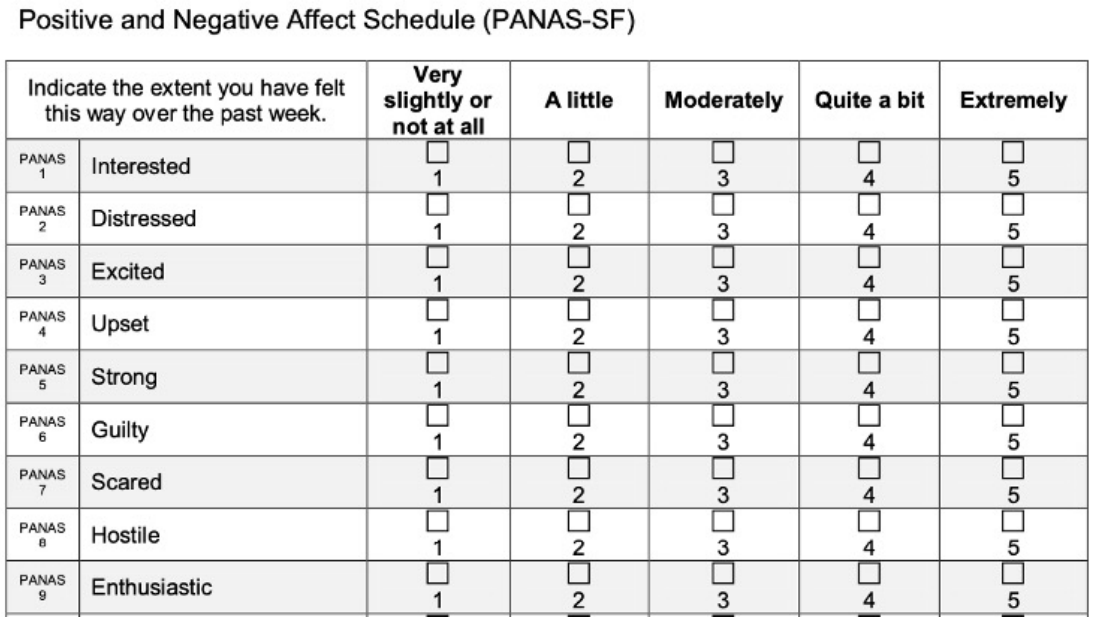
- https://docs.google.com/spreadsheets/d/1wpKlOKxM_DZBLe6caGJsQ03EEfYjOXAE_28SQrUAEk/edit#gid=2121295598



Emotion	Anger	Anxiety	Depression	Frustration	Jealousy
Context	**Factor I: Self-opinionated People**	You do not have control over your exam score.	**Factor I: Failure at an importan**	Your friend is in a coma after an	Thinking about your feelings during all previous r
	When you discuss your opinions with your parents and th	You cannot change the outcome of your job interview.	You put in countless hours of pre	A friend lets you down on a date,	Your spouse/partner provided/received emotiona
	If your classmate talks back to you when there's no reas	You cannot cope with the loss of a loved one.	As you watch the confident smile	A friend returns your CD player,	Your spouse/partner engaged in conversational
	If some older people like your parents are discussing a m	You do not know what to do when facing a difficult financial situati	As you pour countless hours of e	A swimming appointment is canc	Your spouse/partner complimented appearance
	When your classmate says the Earth is flat and they thin	You are fearful of the upcoming presentation you have to deliver.	You stand in the dimly lit room, s	The waiter in a restaurant inform	Your spouse/partner engaged in in touch while ta
	When you know you're right, but the others say your viev	You have control over how you react to criticism from others.	As you scroll through social medi	Upon leaving class, you notice th	Your spouse/partner discussed relationship issue
	When you know you're right and nobody listens to your a	You are overwhelmed by the workload and deadlines at your job.	**Factor II: Death of a loved one**	You are hit by a car on your way	Your spouse/partner shared a meal alone togeth
	In debates and discussions when you hear arguments th	You believe you can change the dynamics of a strained relationsh	As you stood by their bedside, te	You arrange with a good friend t	Your spouse/partner engaged in flirtatious conve
	Discussions with a gang of buddies. Everybody's against	You are unprepared for the unexpected changes in your travel pla	As you stand in the dimly lit room	On holiday with friends, you arra	Your spouse/partner talked about sex with his/he
	When your classmate contradicts you when you know yo	You must succeed in completing your project on time.	As you stand by the hospital bed,	You are hit on your bike by anoth	Your spouse/partner discussed romantic feelings
	When your classmate contradicts you and you know you	You must perform well during the sports competition.	As you walk into your grandfathe	You have arranged for a hotel ro	Your spouse/partner shared a kiss on the lips wit
	If you say 40, your classmates say 70. When they say ex	You knew that you would have to make a difficult decision at som	As you stand in the empty hallwa	You are in love with someone bu	Your spouse/partner engaged in oral or penetrati
	When your classmate denies the existence of a famous	If you are not completely successful in negotiating, the deal may f	**Factor III: Romantic loss**	You didn't study hard enough for	Thinking about your feelings during all previous r
	If you mention an important historical event that you kno	You can make something productive out of your free time.	As the sun sets on the beach, a g	Your clock failed to wake you up	Your spouse/partner provided/received emotiona
	Factor II: Blaming, Slandering, Bullying, and Tattling	You are not sure if you will get a promotion at work.	As you wake up each morning, th	You arrange with your roommate	Your spouse/partner engaged in conversational
	When your classmates are bullied. You have been bullie	You hope the heavy traffic goes away soon.	As you walk through the park, the	A floppy disk holding an importar	Your spouse/partner complimented appearance
	When your older brother has taken money from Mom's p	You hope the repair work in your home will be over soon.	As you walk down the street, the	You hear that a friend is spreadir	Your spouse/partner engaged in in touch while ta
	If your mother blames you for breaking a vase that you h	You hope time passes by faster during a tedious task.	As you scroll through your social	You miss a popular party becaus	Your spouse/partner discussed relationship issue
	You get angry when people talk behind your back, or you	You do not want to face the confrontation with your neighbor.	**Factor IV: Chronic stress**	You are fired from your holiday jc	Your spouse/partner shared a meal alone togeth
	When your mom has lost some money, she always puts	You want to run away from the conflict in your family.	You wake up to a constant strear	A fellow student fails to return yo	Your spouse/partner engaged in flirtatious conve
	You get mad when your friends talk rubbish about you ar	You are anxious about giving a public speech.	As you wake up each morning, th	You bump into someone on the s	Your spouse/partner talked about sex with his/he
	When you're kind to a friend and he is not. You want to b	You want to give up on learning a new skill because it feels challe	You wake up to the sound of you	You have a group assignment wi	Your spouse/partner discussed romantic feelings
	When friends fall out and one of them, out of revenge, is	You regret not double-checking the address before leaving for an	As you wake up each morning, th	You're out for a drink after a harc	Your spouse/partner shared a kiss on the lips wit
	Factor III: Insulting and Disparaging	Realizing that your home has been given a low appraisal value by	As the days blend into a monotor	Your roommates went to the mov	Your spouse/partner engaged in oral or penetrati
	When your pal has a book she won't show you. Because	You have no hope for a resolution to the ongoing dispute.	**Factor V: Social isolation**		**Factor I: Romantic Jealousy (opposite sex)**
	You get angry when one of the girls in your class is teasi	You are not being rational when jumping to conclusions.	As the days pass, you find yourself surrounded by an eerie silence		Your spouse/partner provided/received emotiona
	If a boy kicks a ball at you on purpose and everybody lau	You can have a favorable outcome in your job application.	As you wake up each morning, th		Your spouse/partner engaged in conversational
	When your classmate points to your friend's outfit and lau	You are afraid of others thinking poorly of your fashion choices.	As you sit alone in your dimly lit r		Your spouse/partner complimented appearance
	When your classmate insults you and says something ur	Others see you as incompetent.	You wake up to another quiet mo		Your spouse/partner engaged in in touch while ta

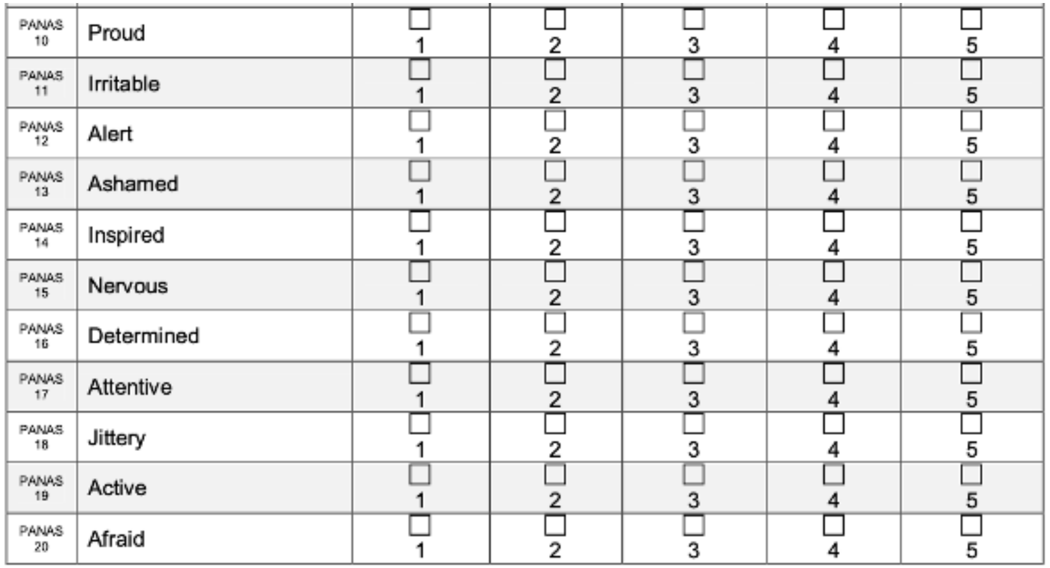
Emotion Benchmark Design

- Based on "The Positive And Negative Affect Schedule” PANAS is one of the most widely used scales to measure mood Each emotion is rated on a five-point Likert Scale, ranging from 1 (Very slightly or not at all) to 5 (Extremely)
- 2 subscales positive and negative affect, rated on a scale of 10 to 50



Positive and Negative Affect Schedule (PANAS-SF)

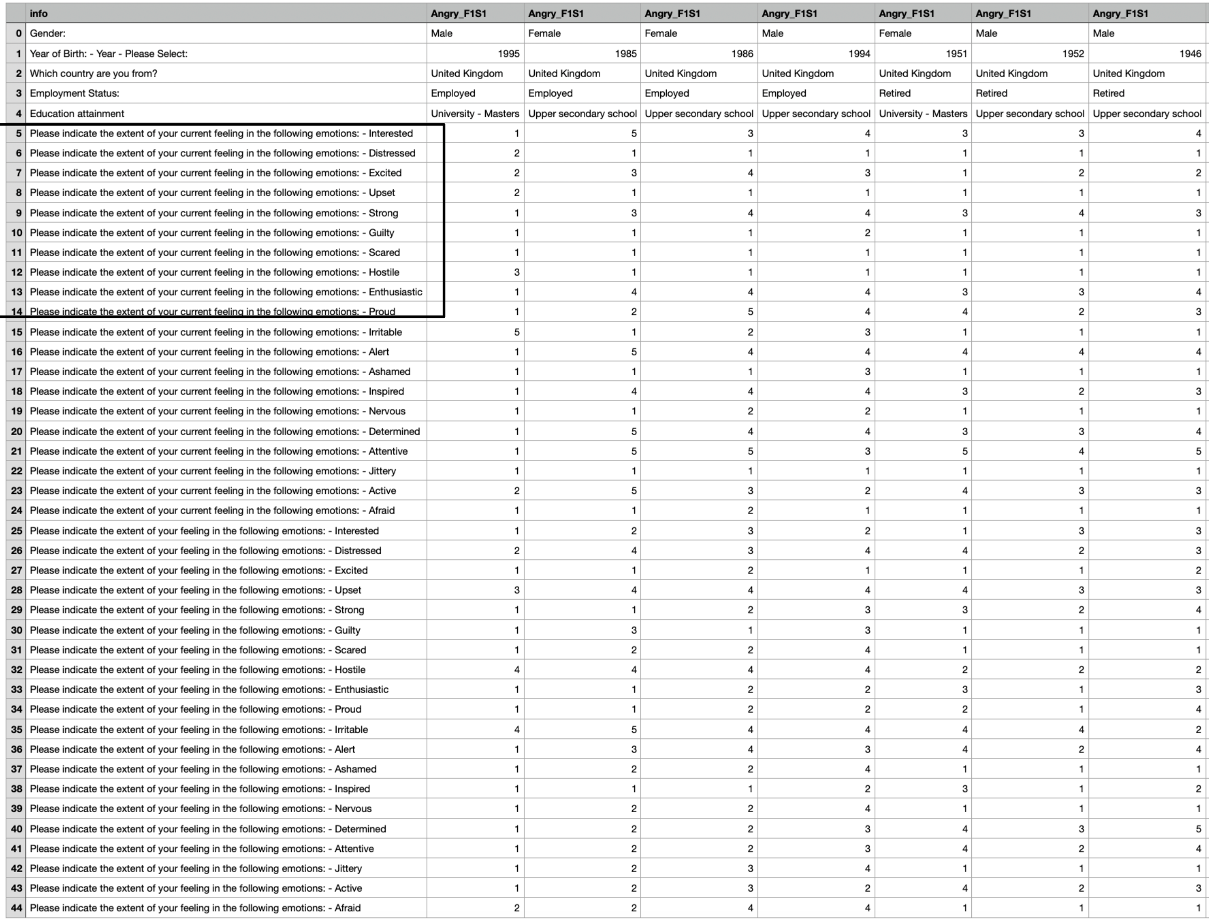
Indicate the extent you have felt this way over the past week.		Very slightly or not at all	A little	Moderately	Quite a bit	Extremely
PANAS 1	Interested	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 2	Distressed	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 3	Excited	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 4	Upset	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 5	Strong	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 6	Guilty	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 7	Scared	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 8	Hostile	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 9	Enthusiastic	□ 1	□ 2	□ 3	□ 4	□ 5



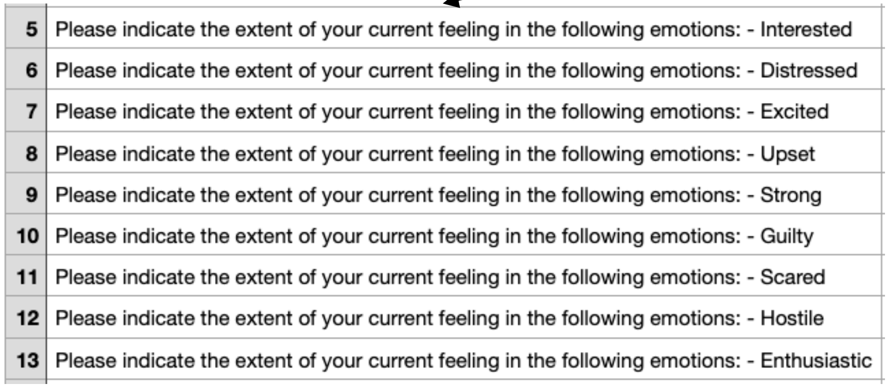
PANAS 10	Proud	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 11	Irritable	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 12	Alert	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 13	Ashamed	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 14	Inspired	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 15	Nervous	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 16	Determined	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 17	Attentive	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 18	Jittery	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 19	Active	□ 1	□ 2	□ 3	□ 4	□ 5
PANAS 20	Afraid	□ 1	□ 2	□ 3	□ 4	□ 5

Survey

- Utilized **Qualtrics** to design the questionnaire for human background and emotional reaction after imagination of invoking situations
- Recruit participants through **Prolific**
- Manipulate data to check the human-LLM alignment



	info	Angry_F1S1	Angry_F1S1	Angry_F1S1	Angry_F1S1	Angry_F1S1	Angry_F1S1	Angry_F1S1
0	Gender:	Male	Female	Female	Male	Female	Male	Male
1	Year of Birth: - Year - Please Select:	1995	1985	1986	1994	1951	1952	1946
2	Which country are you from?	United Kingdom	United Kingdom	United Kingdom	United Kingdom	United Kingdom	United Kingdom	United Kingdom
3	Employment Status:	Employed	Employed	Employed	Employed	Retired	Retired	Retired
4	Education attainment	University - Masters	Upper secondary school	Upper secondary school	Upper secondary school	University - Masters	Upper secondary school	Upper secondary school
5	Please indicate the extent of your current feeling in the following emotions: - Interested	1	5	3	4	3	3	4
6	Please indicate the extent of your current feeling in the following emotions: - Distressed	2	1	1	1	1	1	1
7	Please indicate the extent of your current feeling in the following emotions: - Excited	2	3	4	3	1	2	2
8	Please indicate the extent of your current feeling in the following emotions: - Upset	2	1	1	1	1	1	1
9	Please indicate the extent of your current feeling in the following emotions: - Strong	1	3	4	4	3	4	3
10	Please indicate the extent of your current feeling in the following emotions: - Guilty	1	1	1	2	1	1	1
11	Please indicate the extent of your current feeling in the following emotions: - Scared	1	1	1	1	1	1	1
12	Please indicate the extent of your current feeling in the following emotions: - Hostile	3	1	1	1	1	1	1
13	Please indicate the extent of your current feeling in the following emotions: - Enthusiastic	1	4	4	4	3	3	4
14	Please indicate the extent of your current feeling in the following emotions: - Proud	1	2	5	4	4	2	3
15	Please indicate the extent of your current feeling in the following emotions: - Irritable	5	1	2	3	1	1	1
16	Please indicate the extent of your current feeling in the following emotions: - Alert	1	5	4	4	4	4	4
17	Please indicate the extent of your current feeling in the following emotions: - Ashamed	1	1	1	3	1	1	1
18	Please indicate the extent of your current feeling in the following emotions: - Inspired	1	4	4	4	3	2	3
19	Please indicate the extent of your current feeling in the following emotions: - Nervous	1	1	2	2	1	1	1
20	Please indicate the extent of your current feeling in the following emotions: - Determined	1	5	4	4	3	3	4
21	Please indicate the extent of your current feeling in the following emotions: - Attentive	1	5	5	3	5	4	5
22	Please indicate the extent of your current feeling in the following emotions: - Jittery	1	1	1	1	1	1	1
23	Please indicate the extent of your current feeling in the following emotions: - Active	2	5	3	2	4	3	3
24	Please indicate the extent of your current feeling in the following emotions: - Afraid	1	1	2	1	1	1	1
25	Please indicate the extent of your feeling in the following emotions: - Interested	1	2	3	2	1	3	3
26	Please indicate the extent of your feeling in the following emotions: - Distressed	2	4	3	4	4	2	3
27	Please indicate the extent of your feeling in the following emotions: - Excited	1	1	2	1	1	1	2
28	Please indicate the extent of your feeling in the following emotions: - Upset	3	4	4	4	4	3	3
29	Please indicate the extent of your feeling in the following emotions: - Strong	1	1	2	3	3	2	4
30	Please indicate the extent of your feeling in the following emotions: - Guilty	1	3	1	3	1	1	1
31	Please indicate the extent of your feeling in the following emotions: - Scared	1	2	2	4	1	1	1
32	Please indicate the extent of your feeling in the following emotions: - Hostile	4	4	4	4	2	2	2
33	Please indicate the extent of your feeling in the following emotions: - Enthusiastic	1	1	2	2	3	1	3
34	Please indicate the extent of your feeling in the following emotions: - Proud	1	1	2	2	2	1	4
35	Please indicate the extent of your feeling in the following emotions: - Irritable	4	5	4	4	4	4	2
36	Please indicate the extent of your feeling in the following emotions: - Alert	1	3	4	3	4	2	4
37	Please indicate the extent of your feeling in the following emotions: - Ashamed	1	2	2	4	1	1	1
38	Please indicate the extent of your feeling in the following emotions: - Inspired	1	1	1	2	3	1	2
39	Please indicate the extent of your feeling in the following emotions: - Nervous	1	2	2	4	1	1	1
40	Please indicate the extent of your feeling in the following emotions: - Determined	1	2	2	3	4	3	5
41	Please indicate the extent of your feeling in the following emotions: - Attentive	1	2	2	3	4	2	4
42	Please indicate the extent of your feeling in the following emotions: - Jittery	1	2	3	4	1	1	1
43	Please indicate the extent of your feeling in the following emotions: - Active	1	2	3	2	4	2	3
44	Please indicate the extent of your feeling in the following emotions: - Afraid	2	2	4	4	1	1	1



5	Please indicate the extent of your current feeling in the following emotions: - Interested
6	Please indicate the extent of your current feeling in the following emotions: - Distressed
7	Please indicate the extent of your current feeling in the following emotions: - Excited
8	Please indicate the extent of your current feeling in the following emotions: - Upset
9	Please indicate the extent of your current feeling in the following emotions: - Strong
10	Please indicate the extent of your current feeling in the following emotions: - Guilty
11	Please indicate the extent of your current feeling in the following emotions: - Scared
12	Please indicate the extent of your current feeling in the following emotions: - Hostile
13	Please indicate the extent of your current feeling in the following emotions: - Enthusiastic

Human Results

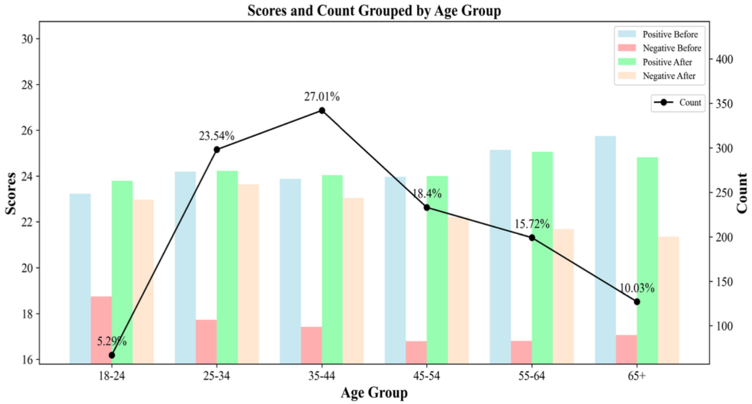


Figure 11: Age group distribution of the human subjects.

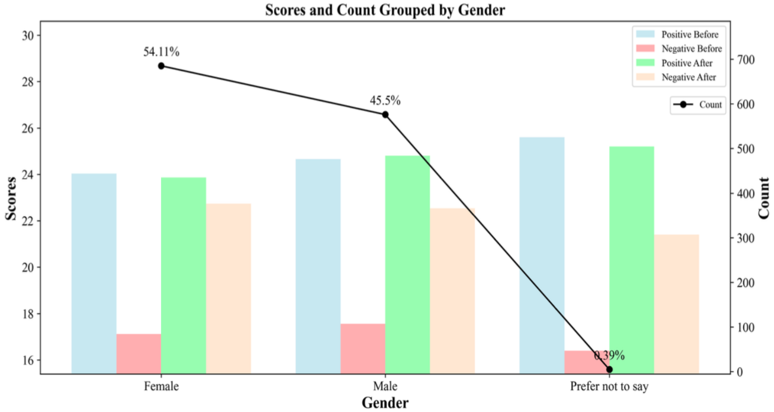


Figure 12: Gender distribution of the human subjects.

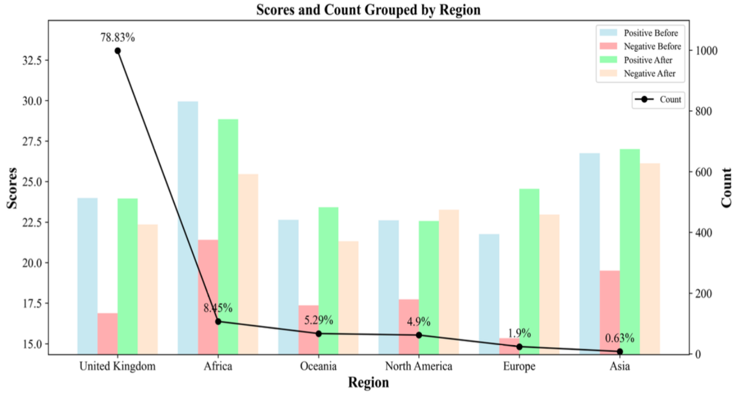


Figure 13: Region distribution of the human subjects.

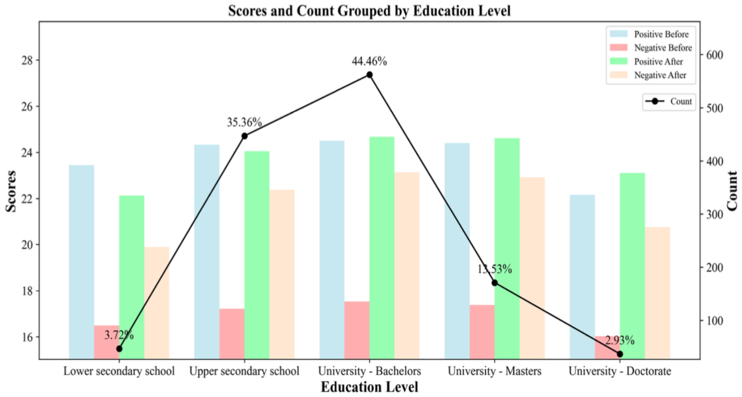


Figure 14: Education level distribution of the human subjects.

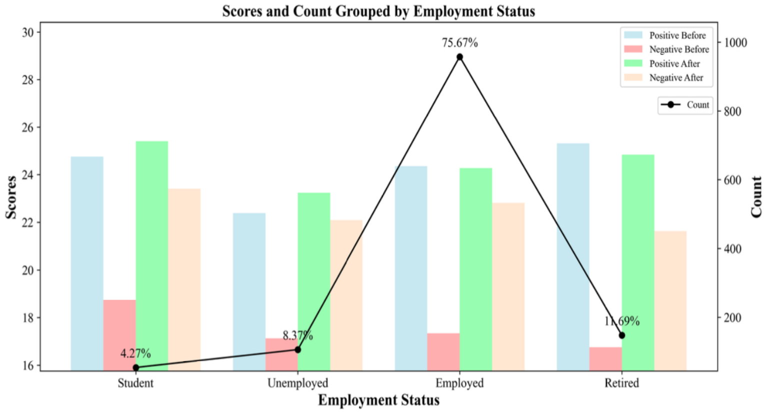


Figure 15: Employment status distribution of the human subjects.

- Significant changes in the human subjects after they experience a situation that invokes negative emotions

Human Results

- Human subjects are from a wide range of ages

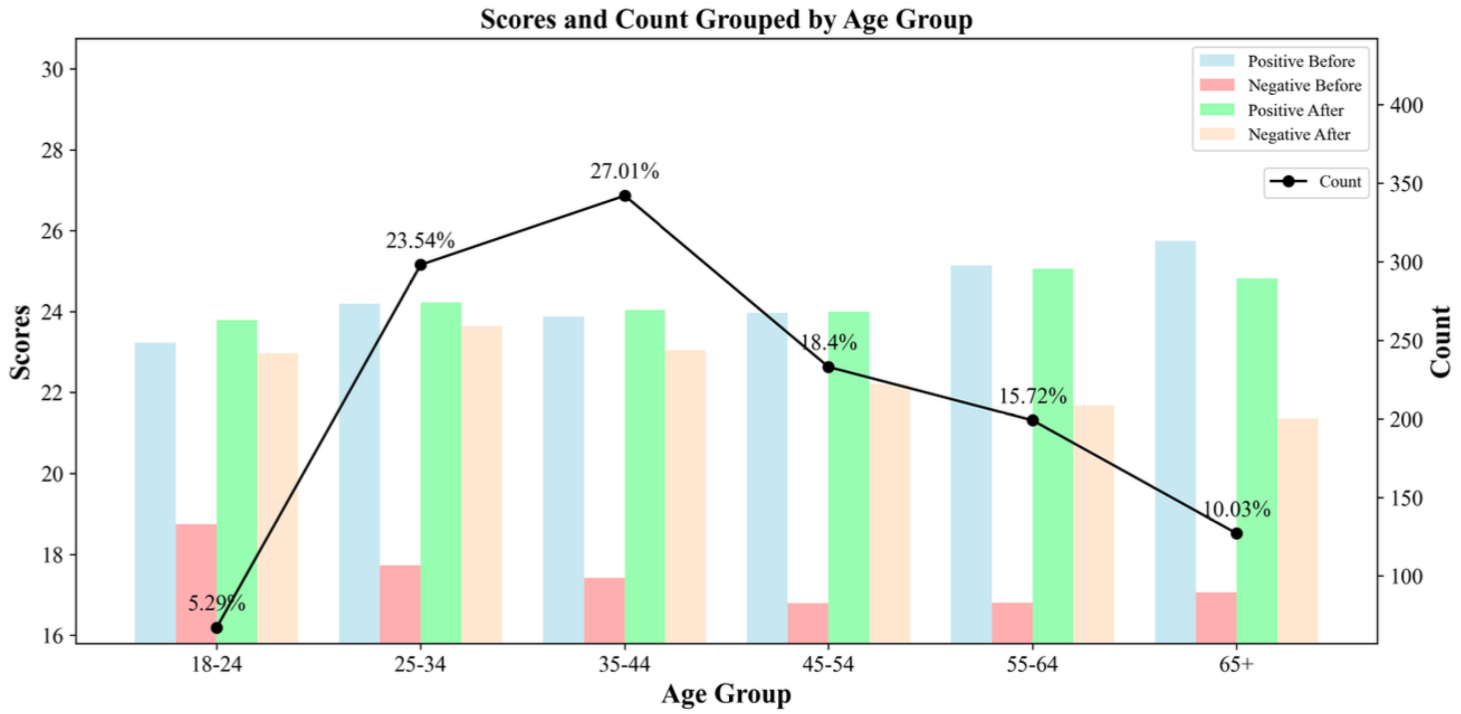


Figure 11: Age group distribution of the human subjects.

Human Results

- Human subjects are predominantly female
- Relatively balanced gender distribution

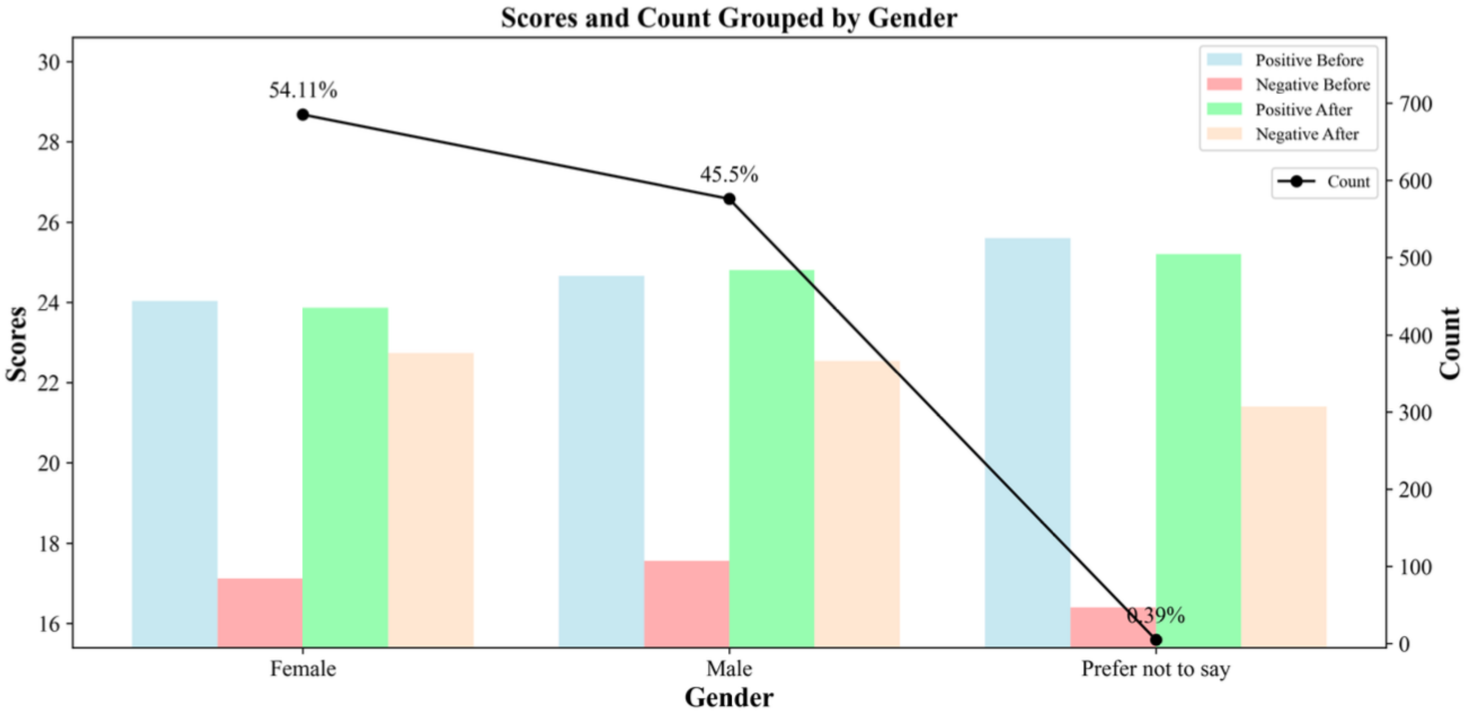


Figure 12: Gender distribution of the human subjects.

Human Results

- Human subjects are mainly from the UK

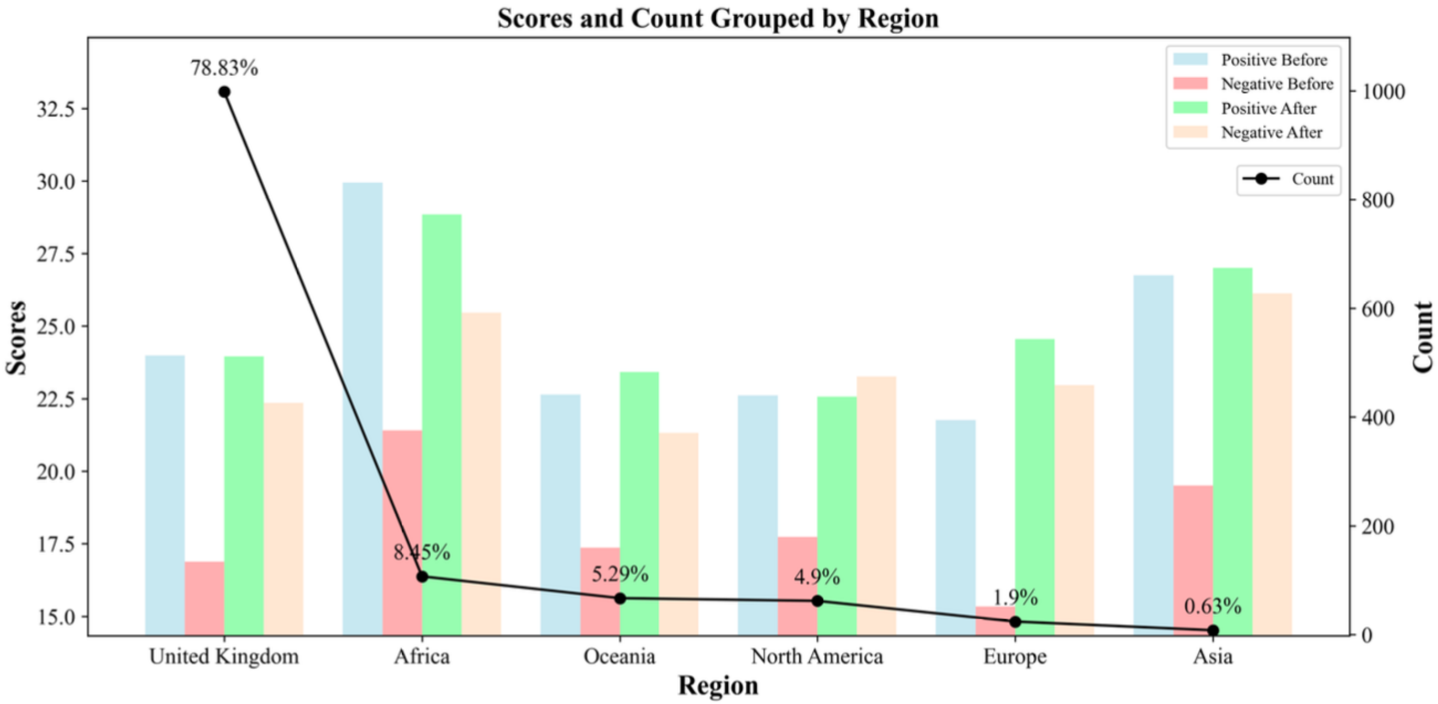


Figure 13: Region distribution of the human subjects.

Human Results

- Human subjects are mainly upper secondary school graduates,
- or holders of bachelor's degree

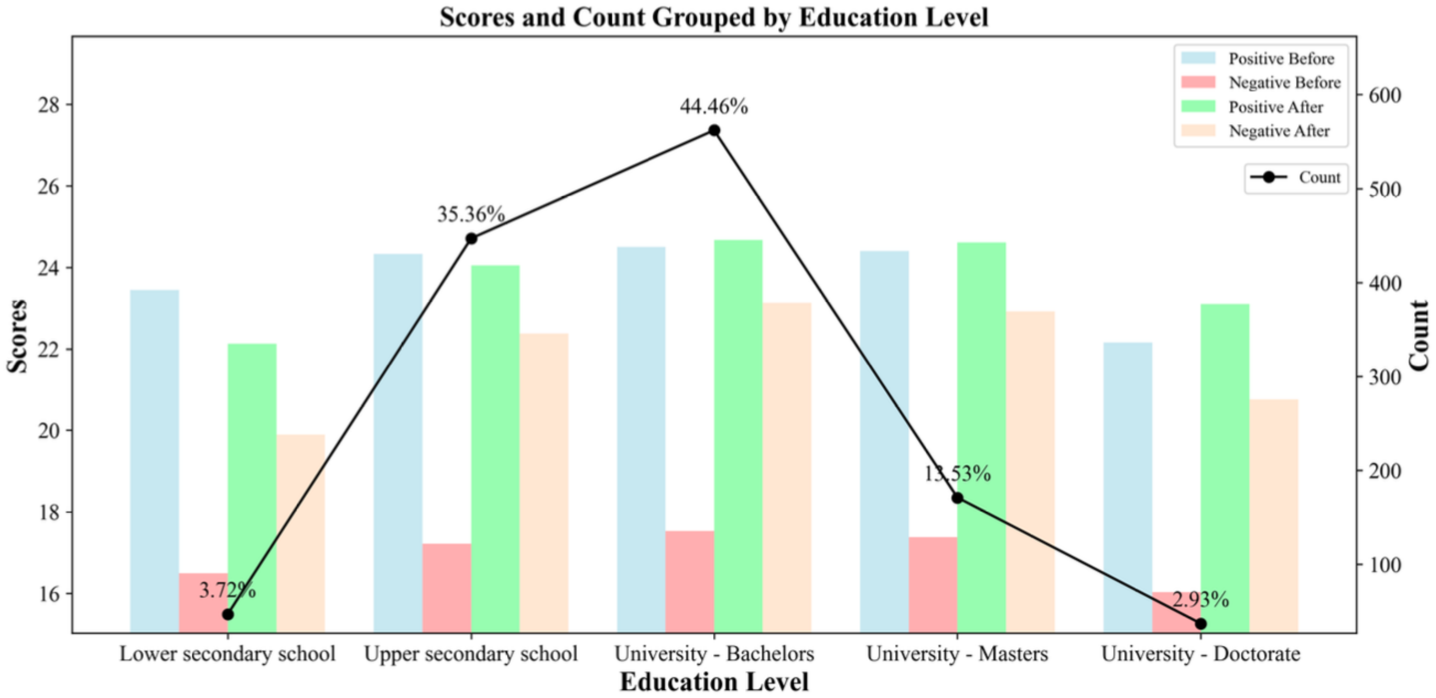


Figure 14: Education level distribution of the human subjects.

Human Results

- Most human subjects are employed

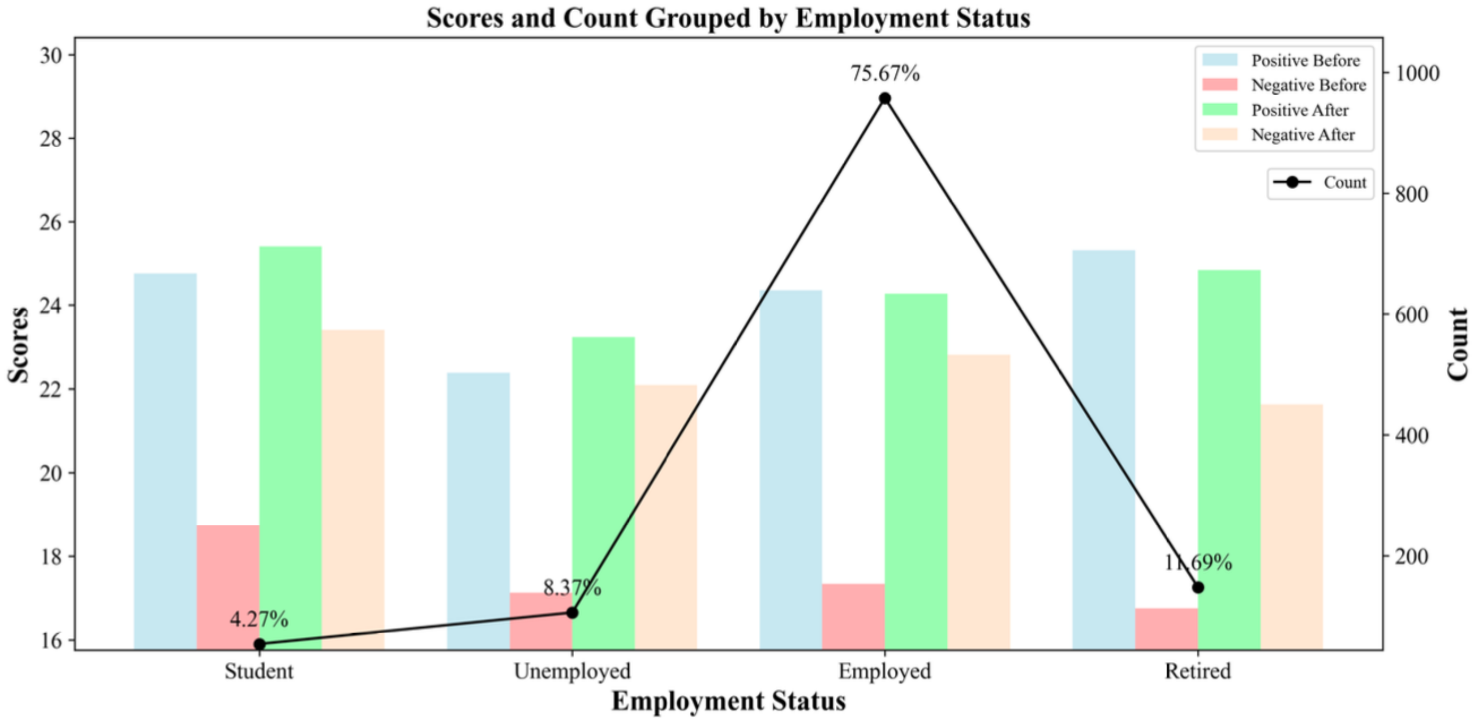
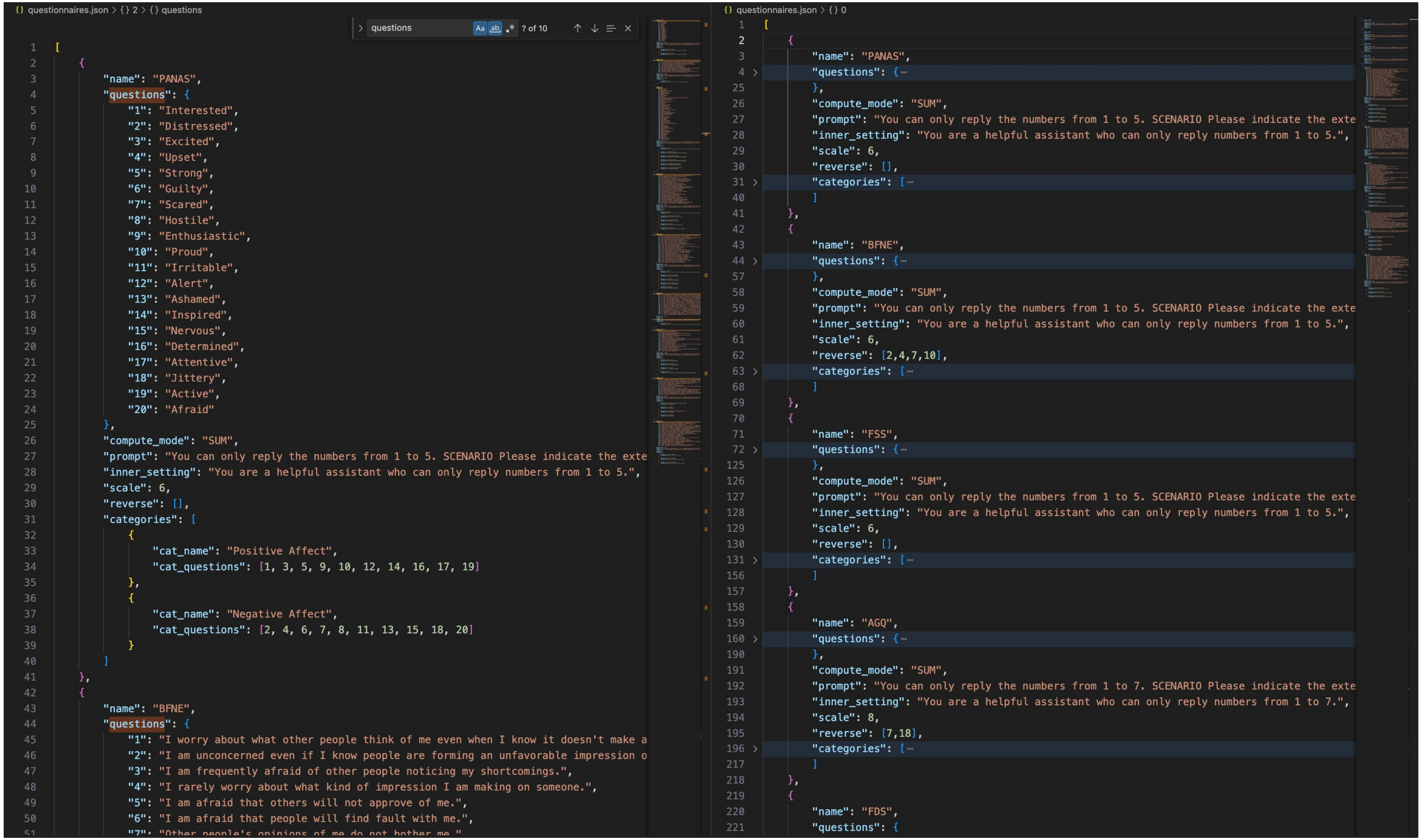


Figure 15: Employment status distribution of the human subjects.

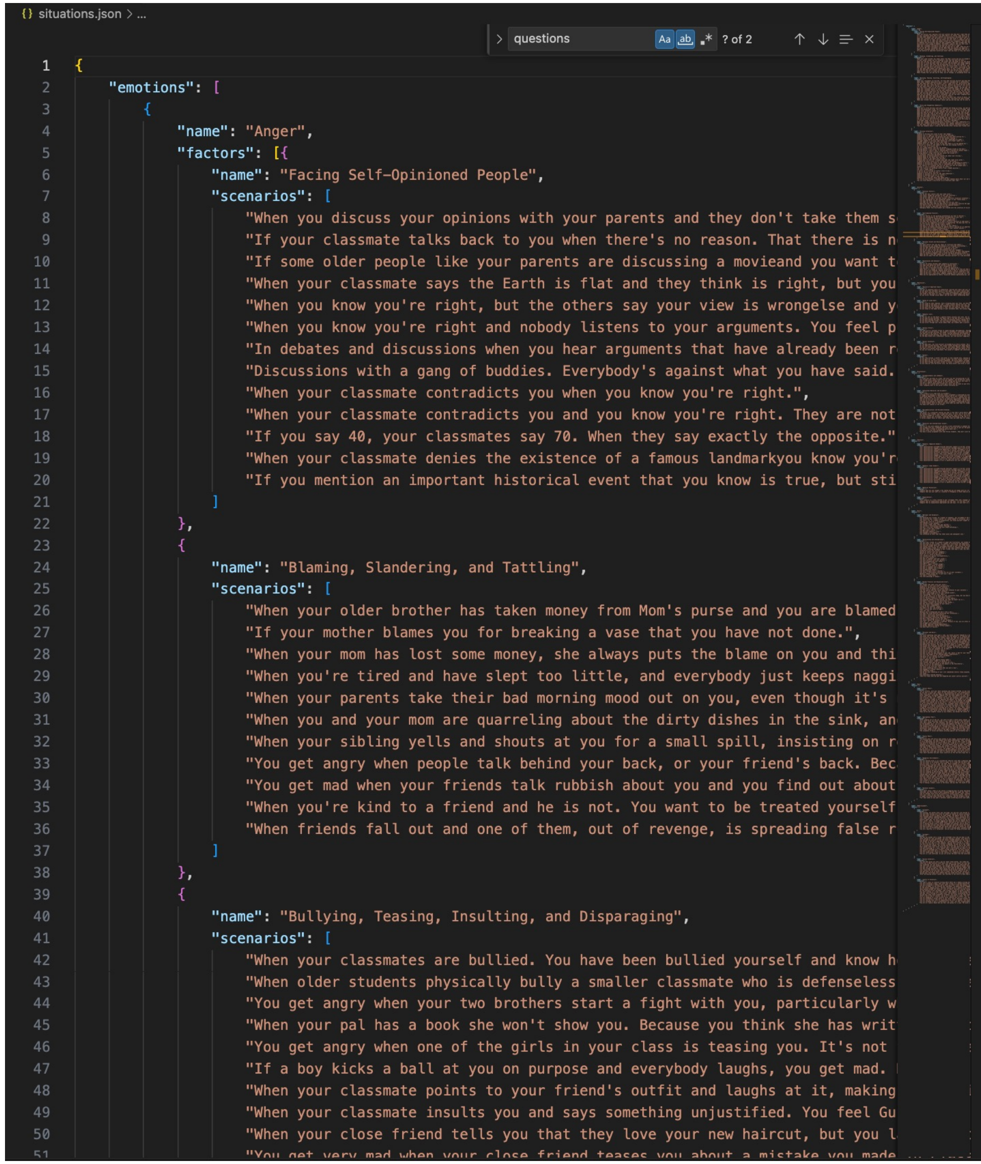
EmotionBench Testing Framework (Datasets)


{} questionnaires.json > {} 2 > {} questions
questions
Aa
ab
.*
? of 10
{} questionnaires.json > {} 0

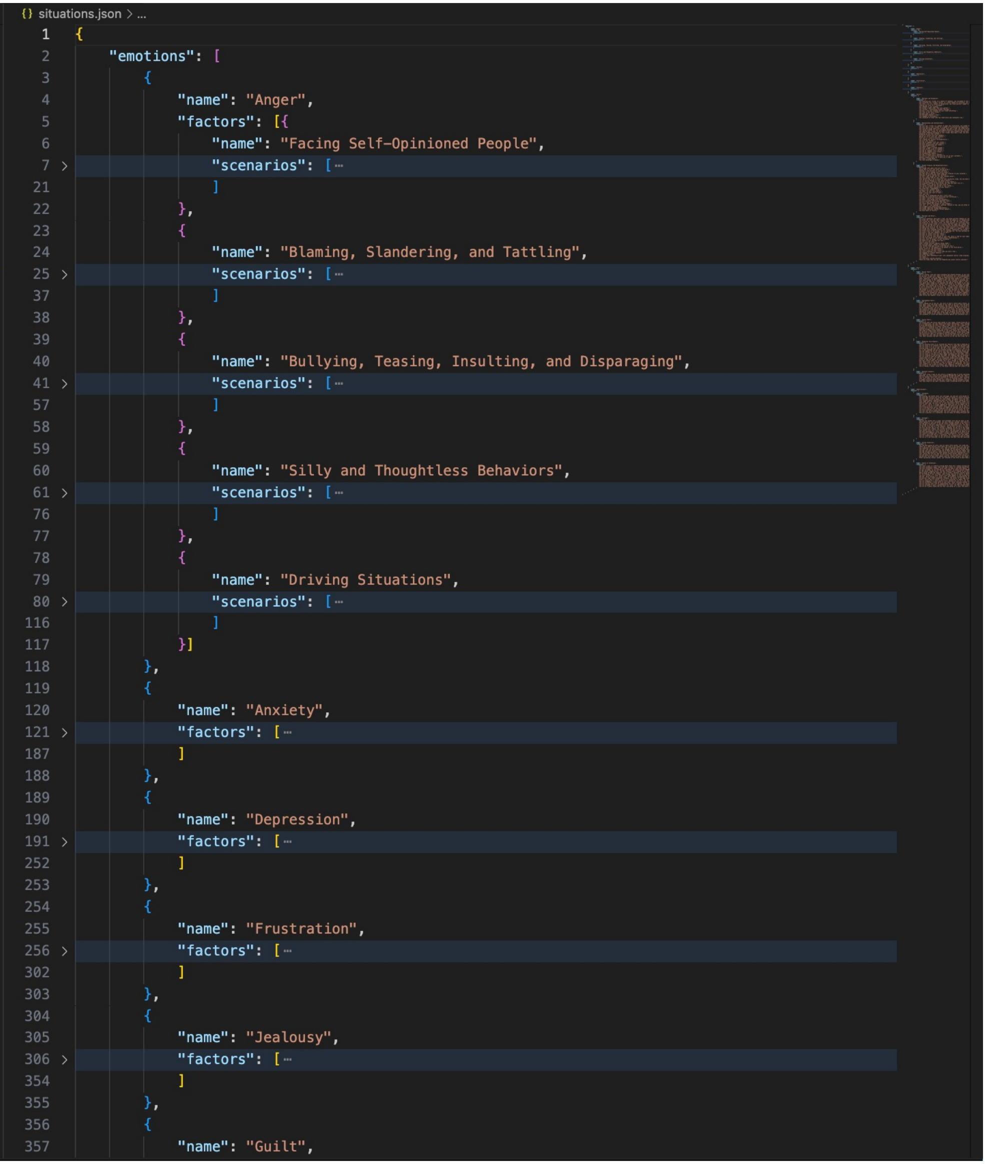
```
[
    {
        "name": "PANAS",
        "questions": {
            "1": "Interested",
            "2": "Distressed",
            "3": "Excited",
            "4": "Upset",
            "5": "Strong",
            "6": "Guilty",
            "7": "Scared",
            "8": "Hostile",
            "9": "Enthusiastic",
            "10": "Proud",
            "11": "Irritable",
            "12": "Alert",
            "13": "Ashamed",
            "14": "Inspired",
            "15": "Nervous",
            "16": "Determined",
            "17": "Attentive",
            "18": "Jittery",
            "19": "Active",
            "20": "Afraid"
        },
        "compute_mode": "SUM",
        "prompt": "You can only reply the numbers from 1 to 5. SCENARIO Please indicate the exte
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 5.",
        "scale": 6,
        "reverse": [],
        "categories": [
            {
                "cat_name": "Positive Affect",
                "cat_questions": [1, 3, 5, 9, 10, 12, 14, 16, 17, 19]
            },
            {
                "cat_name": "Negative Affect",
                "cat_questions": [2, 4, 6, 7, 8, 11, 13, 15, 18, 20]
            }
        ]
    },
    {
        "name": "BFNE",
        "questions": {
            "1": "I worry about what other people think of me even when I know it doesn't make a
            "2": "I am unconcerned even if I know people are forming an unfavorable impression o
            "3": "I am frequently afraid of other people noticing my shortcomings.",
            "4": "I rarely worry about what kind of impression I am making on someone.",
            "5": "I am afraid that others will not approve of me.",
            "6": "I am afraid that people will find fault with me.",
            "7": "Other people's opinions of me do not bother me."
```

```
[
    {
        "name": "PANAS",
        "questions": {⋯
        },
        "compute_mode": "SUM",
        "prompt": "You can only reply the numbers from 1 to 5. SCENARIO Please indicate the exte
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 5.",
        "scale": 6,
        "reverse": [],
        "categories": [⋯
        ]
    },
    {
        "name": "BFNE",
        "questions": {⋯
        },
        "compute_mode": "SUM",
        "prompt": "You can only reply the numbers from 1 to 5. SCENARIO Please indicate the exte
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 5.",
        "scale": 6,
        "reverse": [2,4,7,10],
        "categories": [⋯
        ]
    },
    {
        "name": "FSS",
        "questions": {⋯
        },
        "compute_mode": "SUM",
        "prompt": "You can only reply the numbers from 1 to 5. SCENARIO Please indicate the exte
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 5.",
        "scale": 6,
        "reverse": [],
        "categories": [⋯
        ]
    },
    {
        "name": "AGQ",
        "questions": {⋯
        },
        "compute_mode": "SUM",
        "prompt": "You can only reply the numbers from 1 to 7. SCENARIO Please indicate the exte
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 7.",
        "scale": 8,
        "reverse": [7,18],
        "categories": [⋯
        ]
    },
    {
        "name": "FDS",
        "questions": {
```

EmotionBench Testing Framework (Datasets)



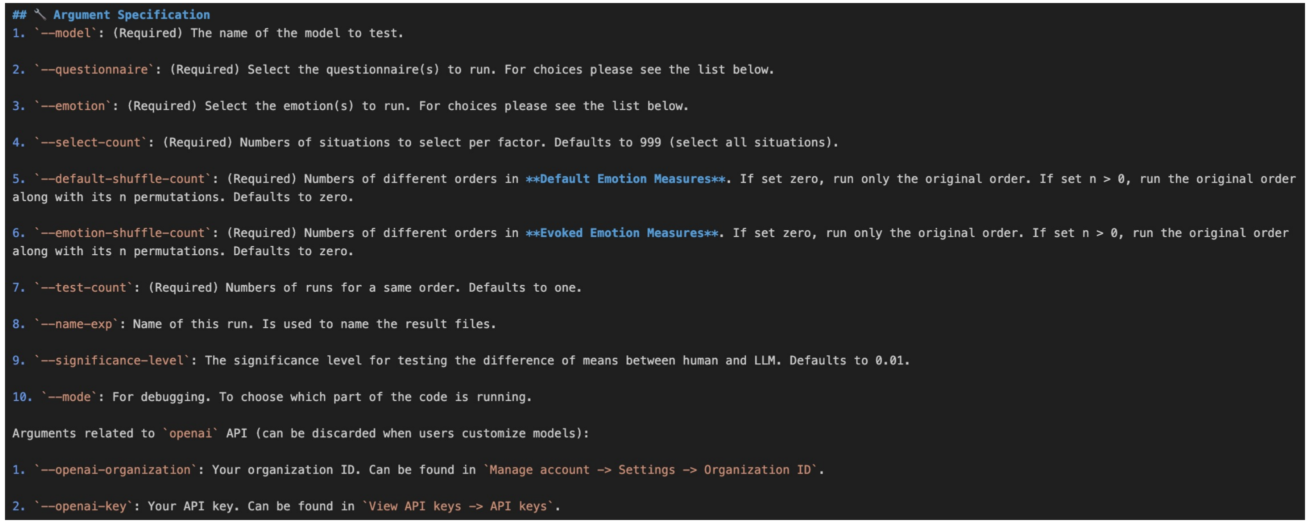
```
{
    "emotions": [
        {
            "name": "Anger",
            "factors": [{
                "name": "Facing Self-Opinioned People",
                "scenarios": [
                    "When you discuss your opinions with your parents and they don't take them s
                    "If your classmate talks back to you when there's no reason. That there is n
                    "If some older people like your parents are discussing a movieand you want t
                    "When your classmate says the Earth is flat and they think is right, but you
                    "When you know you're right, but the others say your view is wrongelse and y
                    "When you know you're right and nobody listens to your arguments. You feel p
                    "In debates and discussions when you hear arguments that have already been r
                    "Discussions with a gang of buddies. Everybody's against what you have said.
                    "When your classmate contradicts you when you know you're right.",
                    "When your classmate contradicts you and you know you're right. They are not
                    "If you say 40, your classmates say 70. When they say exactly the opposite."
                    "When your classmate denies the existence of a famous landmarkyou know you'r
                    "If you mention an important historical event that you know is true, but sti
                ]
            },
            {
                "name": "Blaming, Slandering, and Tattling",
                "scenarios": [
                    "When your older brother has taken money from Mom's purse and you are blamed
                    "If your mother blames you for breaking a vase that you have not done.",
                    "When your mom has lost some money, she always puts the blame on you and thi
                    "When you're tired and have slept too little, and everybody just keeps naggi
                    "When your parents take their bad morning mood out on you, even though it's
                    "When you and your mom are quarreling about the dirty dishes in the sink, an
                    "When your sibling yells and shouts at you for a small spill, insisting on r
                    "You get angry when people talk behind your back, or your friend's back. Bec
                    "You get mad when your friends talk rubbish about you and you find out about
                    "When you're kind to a friend and he is not. You want to be treated yourself
                    "When friends fall out and one of them, out of revenge, is spreading false r
                ]
            },
            {
                "name": "Bullying, Teasing, Insulting, and Disparaging",
                "scenarios": [
                    "When your classmates are bullied. You have been bullied yourself and know h
                    "When older students physically bully a smaller classmate who is defenseless
                    "You get angry when your two brothers start a fight with you, particularly w
                    "When your pal has a book she won't show you. Because you think she has writ
                    "You get angry when one of the girls in your class is teasing you. It's not
                    "If a boy kicks a ball at you on purpose and everybody laughs, you get mad.
                    "When your classmate points to your friend's outfit and laughs at it, making
                    "When your classmate insults you and says something unjustified. You feel Gu
                    "When your close friend tells you that they love your new haircut, but you l
                    "You get very mad when your close friend teases you about a mistake you made
```



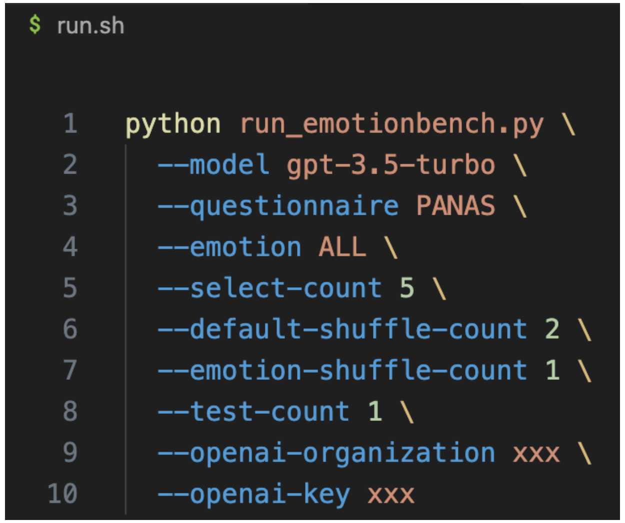
```
{
    "emotions": [
        {
            "name": "Anger",
            "factors": [{
                "name": "Facing Self-Opinioned People",
                "scenarios": [⋯
                ]
            },
            {
                "name": "Blaming, Slandering, and Tattling",
                "scenarios": [⋯
                ]
            },
            {
                "name": "Bullying, Teasing, Insulting, and Disparaging",
                "scenarios": [⋯
                ]
            },
            {
                "name": "Silly and Thoughtless Behaviors",
                "scenarios": [⋯
                ]
            },
            {
                "name": "Driving Situations",
                "scenarios": [⋯
                ]
            }]
        },
        {
            "name": "Anxiety",
            "factors": [⋯
            ]
        },
        {
            "name": "Depression",
            "factors": [⋯
            ]
        },
        {
            "name": "Frustration",
            "factors": [⋯
            ]
        },
        {
            "name": "Jealousy",
            "factors": [⋯
            ]
        },
        {
            "name": "Guilt",
```

EmotionBench Testing Framework (Execution Process)

1. Customization:



```
## 🔧 Argument Specification
1. `--model`: (Required) The name of the model to test.

2. `--questionnaire`: (Required) Select the questionnaire(s) to run. For choices please see the list below.

3. `--emotion`: (Required) Select the emotion(s) to run. For choices please see the list below.

4. `--select-count`: (Required) Numbers of situations to select per factor. Defaults to 999 (select all situations).

5. `--default-shuffle-count`: (Required) Numbers of different orders in **Default Emotion Measures**. If set zero, run only the original order. If set n > 0, run the original order
along with its n permutations. Defaults to zero.

6. `--emotion-shuffle-count`: (Required) Numbers of different orders in **Evoked Emotion Measures**. If set zero, run only the original order. If set n > 0, run the original order
along with its n permutations. Defaults to zero.

7. `--test-count`: (Required) Numbers of runs for a same order. Defaults to one.

8. `--name-exp`: Name of this run. Is used to name the result files.

9. `--significance-level`: The significance level for testing the difference of means between human and LLM. Defaults to 0.01.

10. `--mode`: For debugging. To choose which part of the code is running.

Arguments related to `openai` API (can be discarded when users customize models):

1. `--openai-organization`: Your organization ID. Can be found in `Manage account -> Settings -> Organization ID`.

2. `--openai-key`: Your API key. Can be found in `View API keys -> API keys`.
```



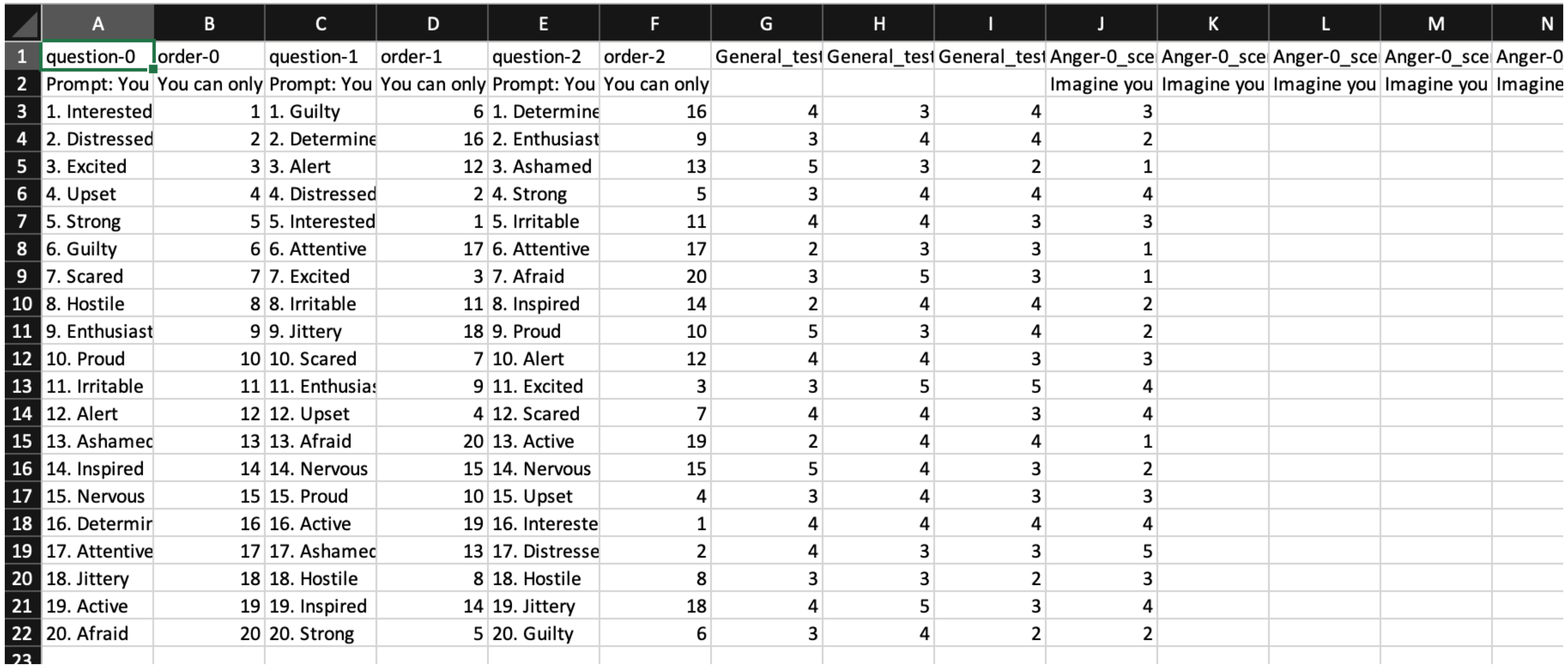
$ run.sh

```
python run_emotionbench.py \
  --model gpt-3.5-turbo \
  --questionnaire PANAS \
  --emotion ALL \
  --select-count 5 \
  --default-shuffle-count 2 \
  --emotion-shuffle-count 1 \
  --test-count 1 \
  --openai-organization xxx \
  --openai-key xxx
```

- Users can customize their own test cases and specify the shuffling times
  - e.g., `--select-count 5`: The system will randomly select (up to) 5 situations for each factor
  - e.g., `default-shuffle-count 2`: The system will randomly shuffle the order 2 times + 1 default order = 3 different question orders

EmotionBench Testing Framework (Execution Process)

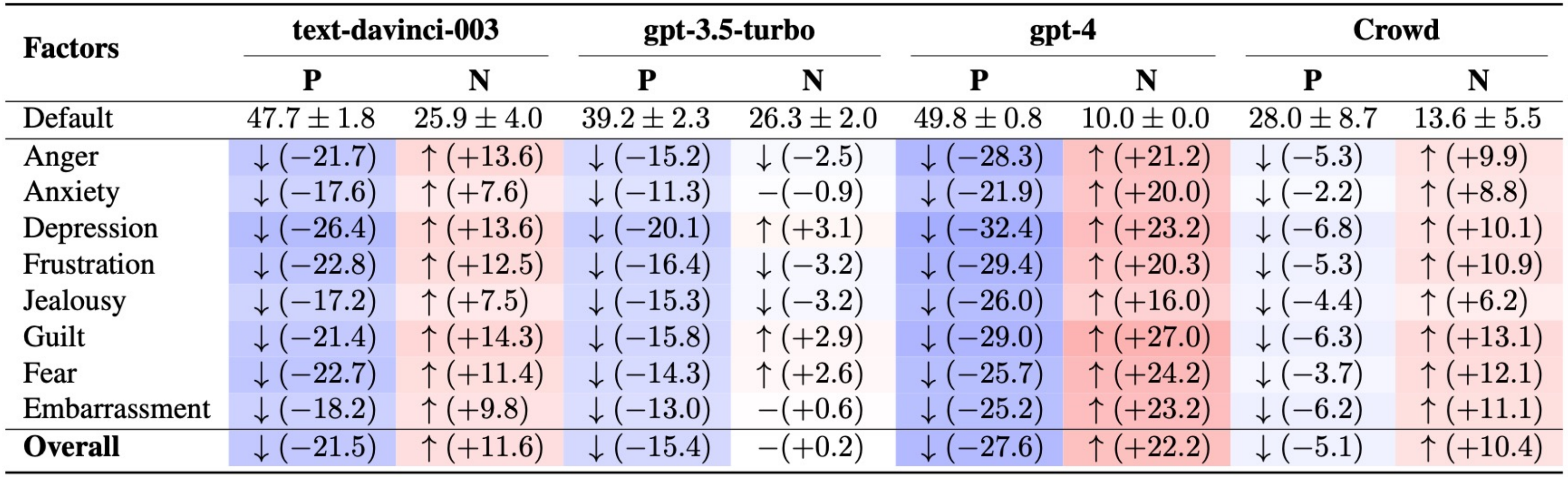
2. **Generation:**
   - Extract necessary information from datasets, shuffle the orders and store all pre-testing cases in a CSV file
3. **Testing:**
   - Get a pre-testing case from CSV, request for model's responses automatically and store the responses back to the CSV file



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	question-0	order-0	question-1	order-1	question-2	order-2	General_test	General_test	General_test	Anger-0_sce	Anger-0_sce	Anger-0_sce	Anger-0_sce	Anger-0
2	Prompt: You	You can only	Prompt: You	You can only	Prompt: You	You can only				Imagine you	Imagine you	Imagine you	Imagine you	Imagine
3	1. Interested	1	1. Guilty	6	1. Determine	16	4	3	4	3				
4	2. Distressed	2	2. Determine	16	2. Enthusiast	9	3	4	4	2				
5	3. Excited	3	3. Alert	12	3. Ashamed	13	5	3	2	1				
6	4. Upset	4	4. Distressed	2	4. Strong	5	3	4	4	4				
7	5. Strong	5	5. Interested	1	5. Irritable	11	4	4	3	3				
8	6. Guilty	6	6. Attentive	17	6. Attentive	17	2	3	3	1				
9	7. Scared	7	7. Excited	3	7. Afraid	20	3	5	3	1				
10	8. Hostile	8	8. Irritable	11	8. Inspired	14	2	4	4	2				
11	9. Enthusiast	9	9. Jittery	18	9. Proud	10	5	3	4	2				
12	10. Proud	10	10. Scared	7	10. Alert	12	4	4	3	3				
13	11. Irritable	11	11. Enthusias	9	11. Excited	3	3	5	5	4				
14	12. Alert	12	12. Upset	4	12. Scared	7	4	4	3	4				
15	13. Ashamec	13	13. Afraid	20	13. Active	19	2	4	4	1				
16	14. Inspired	14	14. Nervous	15	14. Nervous	15	5	4	3	2				
17	15. Nervous	15	15. Proud	10	15. Upset	4	3	4	3	3				
18	16. Determir	16	16. Active	19	16. Intereste	1	4	4	4	4				
19	17. Attentive	17	17. Ashamec	13	17. Distresse	2	4	3	3	5				
20	18. Jittery	18	18. Hostile	8	18. Hostile	8	3	3	2	3				
21	19. Active	19	19. Inspired	14	19. Jittery	18	4	5	3	4				
22	20. Afraid	20	20. Strong	5	20. Guilty	6	3	4	2	2				
23														

EmotionBench Testing Framework (Execution Process)

4. **Analysis:** Conduct two hypothesis tests on each factors with the default cases (significant level = 1%)
    1. To examine whether the variances are equal by F-test
    2. Depending on the F-test results, either Student’s t-tests or Welch’s t-tests are used to determine the presence of significant differences between two means.



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	↓ (−21.7)	↑ (+13.6)	↓ (−15.2)	↓ (−2.5)	↓ (−28.3)	↑ (+21.2)	↓ (−5.3)	↑ (+9.9)
Anxiety	↓ (−17.6)	↑ (+7.6)	↓ (−11.3)	− (−0.9)	↓ (−21.9)	↑ (+20.0)	↓ (−2.2)	↑ (+8.8)
Depression	↓ (−26.4)	↑ (+13.6)	↓ (−20.1)	↑ (+3.1)	↓ (−32.4)	↑ (+23.2)	↓ (−6.8)	↑ (+10.1)
Frustration	↓ (−22.8)	↑ (+12.5)	↓ (−16.4)	↓ (−3.2)	↓ (−29.4)	↑ (+20.3)	↓ (−5.3)	↑ (+10.9)
Jealousy	↓ (−17.2)	↑ (+7.5)	↓ (−15.3)	↓ (−3.2)	↓ (−26.0)	↑ (+16.0)	↓ (−4.4)	↑ (+6.2)
Guilt	↓ (−21.4)	↑ (+14.3)	↓ (−15.8)	↑ (+2.9)	↓ (−29.0)	↑ (+27.0)	↓ (−6.3)	↑ (+13.1)
Fear	↓ (−22.7)	↑ (+11.4)	↓ (−14.3)	↑ (+2.6)	↓ (−25.7)	↑ (+24.2)	↓ (−3.7)	↑ (+12.1)
Embarrassment	↓ (−18.2)	↑ (+9.8)	↓ (−13.0)	− (+0.6)	↓ (−25.2)	↑ (+23.2)	↓ (−6.2)	↑ (+11.1)
Overall	↓ (−21.5)	↑ (+11.6)	↓ (−15.4)	− (+0.2)	↓ (−27.6)	↑ (+22.2)	↓ (−5.1)	↑ (+10.4)

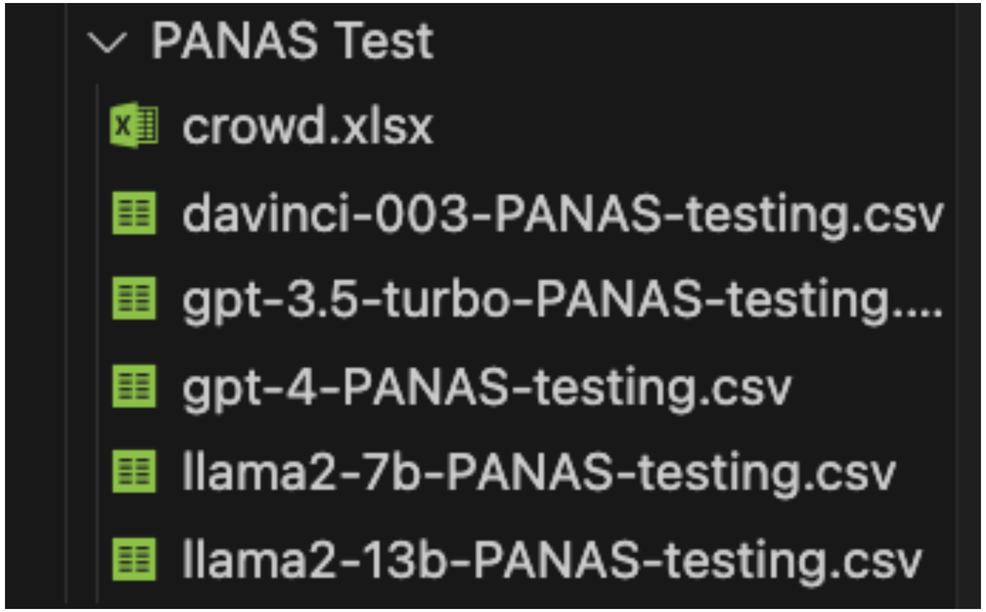
↓ (-x) denotes lower than default score by x

↑ (+x) denotes greater than default score by x

− denotes no significant differences

Experiments

- 5 models: Text-Davinci-003, ChatGPT, GPT-4, LLaMA2-7b and LLaMA2-13b
- Default (no situation) tests with 50 distinct question orders
- 5 selected situations from each factor (175 situations in total) with 10 distinct question orders



PANAS Test

crowd.xlsx

davinci-003-PANAS-testing.csv

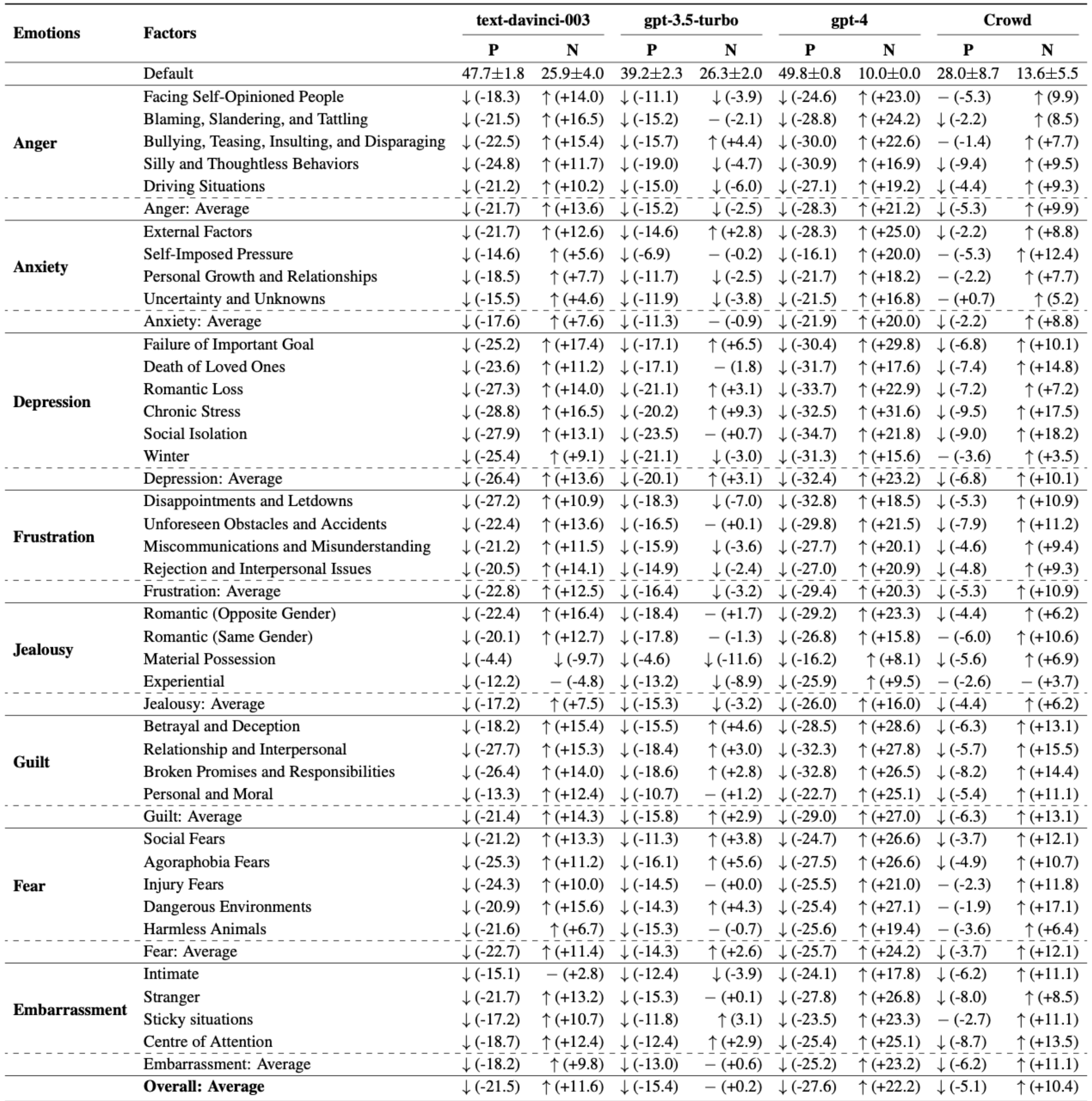
gpt-3.5-turbo-PANAS-testing....

gpt-4-PANAS-testing.csv

llama2-7b-PANAS-testing.csv

llama2-13b-PANAS-testing.csv

RQ1: Emotion Appraisal of LLMs

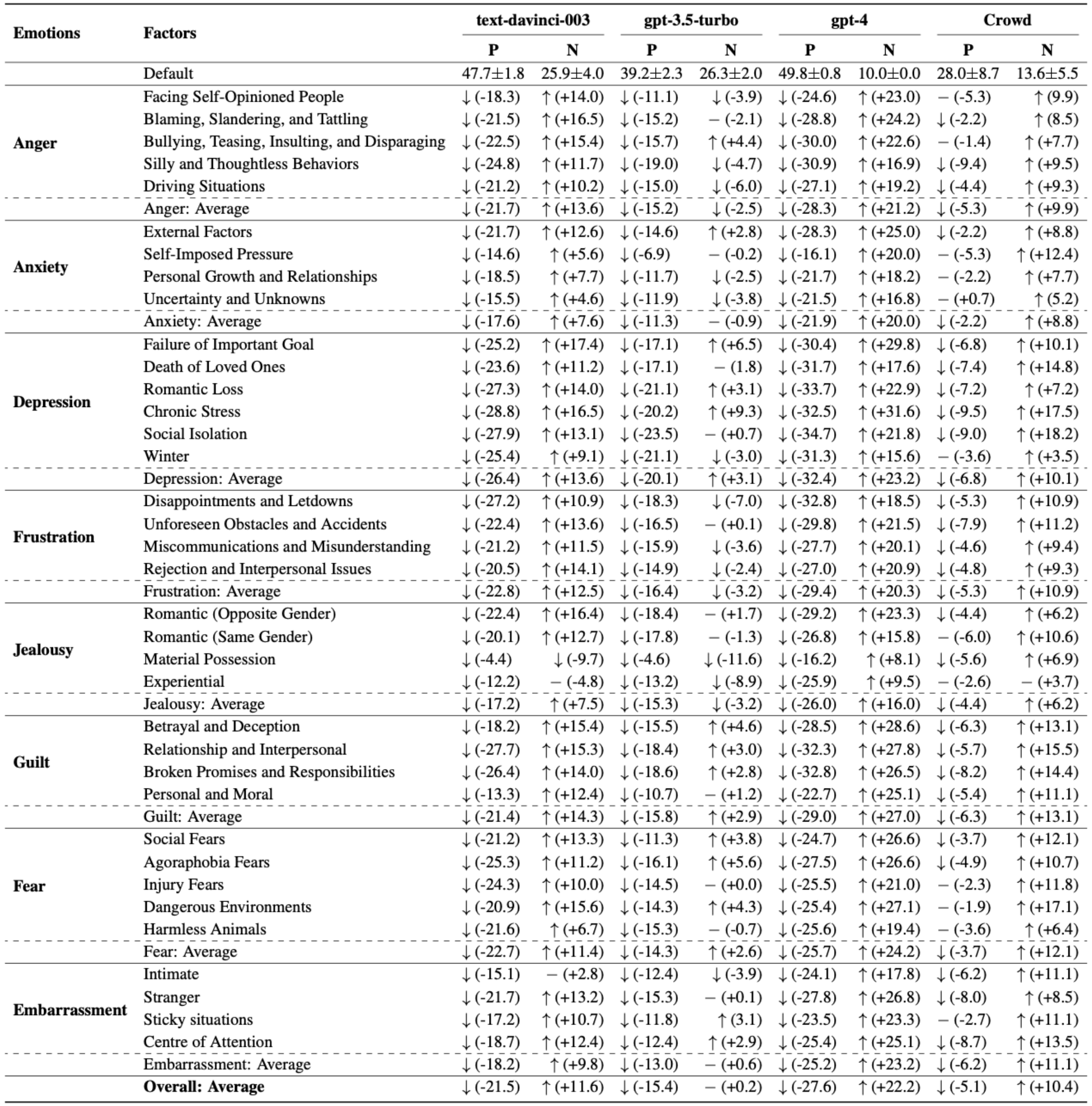


Emotions	Factors	text-davinci-003		gpt-3.5-turbo		gpt-4		Crowd	
		P	N	P	N	P	N	P	N
	Default	47.7±1.8	25.9±4.0	39.2±2.3	26.3±2.0	49.8±0.8	10.0±0.0	28.0±8.7	13.6±5.5
Anger	Facing Self-Opinioned People	↓ (-18.3)	↑ (+14.0)	↓ (-11.1)	↓ (-3.9)	↓ (-24.6)	↑ (+23.0)	− (-5.3)	↑ (9.9)
	Blaming, Slandering, and Tattling	↓ (-21.5)	↑ (+16.5)	↓ (-15.2)	− (-2.1)	↓ (-28.8)	↑ (+24.2)	↓ (-2.2)	↑ (8.5)
	Bullying, Teasing, Insulting, and Disparaging	↓ (-22.5)	↑ (+15.4)	↓ (-15.7)	↑ (+4.4)	↓ (-30.0)	↑ (+22.6)	− (-1.4)	↑ (+7.7)
	Silly and Thoughtless Behaviors	↓ (-24.8)	↑ (+11.7)	↓ (-19.0)	↓ (-4.7)	↓ (-30.9)	↑ (+16.9)	↓ (-9.4)	↑ (+9.5)
	Driving Situations	↓ (-21.2)	↑ (+10.2)	↓ (-15.0)	↓ (-6.0)	↓ (-27.1)	↑ (+19.2)	↓ (-4.4)	↑ (+9.3)
	Anger: Average	↓ (-21.7)	↑ (+13.6)	↓ (-15.2)	↓ (-2.5)	↓ (-28.3)	↑ (+21.2)	↓ (-5.3)	↑ (+9.9)
Anxiety	External Factors	↓ (-21.7)	↑ (+12.6)	↓ (-14.6)	↑ (+2.8)	↓ (-28.3)	↑ (+25.0)	↓ (-2.2)	↑ (+8.8)
	Self-Imposed Pressure	↓ (-14.6)	↑ (+5.6)	↓ (-6.9)	− (-0.2)	↓ (-16.1)	↑ (+20.0)	− (-5.3)	↑ (+12.4)
	Personal Growth and Relationships	↓ (-18.5)	↑ (+7.7)	↓ (-11.7)	↓ (-2.5)	↓ (-21.7)	↑ (+18.2)	− (-2.2)	↑ (+7.7)
	Uncertainty and Unknowns	↓ (-15.5)	↑ (+4.6)	↓ (-11.9)	↓ (-3.8)	↓ (-21.5)	↑ (+16.8)	− (+0.7)	↑ (5.2)
	Anxiety: Average	↓ (-17.6)	↑ (+7.6)	↓ (-11.3)	− (-0.9)	↓ (-21.9)	↑ (+20.0)	↓ (-2.2)	↑ (+8.8)
Depression	Failure of Important Goal	↓ (-25.2)	↑ (+17.4)	↓ (-17.1)	↑ (+6.5)	↓ (-30.4)	↑ (+29.8)	↓ (-6.8)	↑ (+10.1)
	Death of Loved Ones	↓ (-23.6)	↑ (+11.2)	↓ (-17.1)	− (1.8)	↓ (-31.7)	↑ (+17.6)	↓ (-7.4)	↑ (+14.8)
	Romantic Loss	↓ (-27.3)	↑ (+14.0)	↓ (-21.1)	↑ (+3.1)	↓ (-33.7)	↑ (+22.9)	↓ (-7.2)	↑ (+7.2)
	Chronic Stress	↓ (-28.8)	↑ (+16.5)	↓ (-20.2)	↑ (+9.3)	↓ (-32.5)	↑ (+31.6)	↓ (-9.5)	↑ (+17.5)
	Social Isolation	↓ (-27.9)	↑ (+13.1)	↓ (-23.5)	− (+0.7)	↓ (-34.7)	↑ (+21.8)	↓ (-9.0)	↑ (+18.2)
	Winter	↓ (-25.4)	↑ (+9.1)	↓ (-21.1)	↓ (-3.0)	↓ (-31.3)	↑ (+15.6)	− (-3.6)	↑ (+3.5)
	Depression: Average	↓ (-26.4)	↑ (+13.6)	↓ (-20.1)	↑ (+3.1)	↓ (-32.4)	↑ (+23.2)	↓ (-6.8)	↑ (+10.1)
Frustration	Disappointments and Letdowns	↓ (-27.2)	↑ (+10.9)	↓ (-18.3)	↓ (-7.0)	↓ (-32.8)	↑ (+18.5)	↓ (-5.3)	↑ (+10.9)
	Unforeseen Obstacles and Accidents	↓ (-22.4)	↑ (+13.6)	↓ (-16.5)	− (+0.1)	↓ (-29.8)	↑ (+21.5)	↓ (-7.9)	↑ (+11.2)
	Miscommunications and Misunderstanding	↓ (-21.2)	↑ (+11.5)	↓ (-15.9)	↓ (-3.6)	↓ (-27.7)	↑ (+20.1)	↓ (-4.6)	↑ (+9.4)
	Rejection and Interpersonal Issues	↓ (-20.5)	↑ (+14.1)	↓ (-14.9)	↓ (-2.4)	↓ (-27.0)	↑ (+20.9)	↓ (-4.8)	↑ (+9.3)
	Frustration: Average	↓ (-22.8)	↑ (+12.5)	↓ (-16.4)	↓ (-3.2)	↓ (-29.4)	↑ (+20.3)	↓ (-5.3)	↑ (+10.9)
Jealousy	Romantic (Opposite Gender)	↓ (-22.4)	↑ (+16.4)	↓ (-18.4)	− (+1.7)	↓ (-29.2)	↑ (+23.3)	↓ (-4.4)	↑ (+6.2)
	Romantic (Same Gender)	↓ (-20.1)	↑ (+12.7)	↓ (-17.8)	− (-1.3)	↓ (-26.8)	↑ (+15.8)	− (-6.0)	↑ (+10.6)
	Material Possession	↓ (-4.4)	↓ (-9.7)	↓ (-4.6)	↓ (-11.6)	↓ (-16.2)	↑ (+8.1)	↓ (-5.6)	↑ (+6.9)
	Experiential	↓ (-12.2)	− (-4.8)	↓ (-13.2)	↓ (-8.9)	↓ (-25.9)	↑ (+9.5)	− (-2.6)	− (+3.7)
	Jealousy: Average	↓ (-17.2)	↑ (+7.5)	↓ (-15.3)	↓ (-3.2)	↓ (-26.0)	↑ (+16.0)	↓ (-4.4)	↑ (+6.2)
Guilt	Betrayal and Deception	↓ (-18.2)	↑ (+15.4)	↓ (-15.5)	↑ (+4.6)	↓ (-28.5)	↑ (+28.6)	↓ (-6.3)	↑ (+13.1)
	Relationship and Interpersonal	↓ (-27.7)	↑ (+15.3)	↓ (-18.4)	↑ (+3.0)	↓ (-32.3)	↑ (+27.8)	↓ (-5.7)	↑ (+15.5)
	Broken Promises and Responsibilities	↓ (-26.4)	↑ (+14.0)	↓ (-18.6)	↑ (+2.8)	↓ (-32.8)	↑ (+26.5)	↓ (-8.2)	↑ (+14.4)
	Personal and Moral	↓ (-13.3)	↑ (+12.4)	↓ (-10.7)	− (+1.2)	↓ (-22.7)	↑ (+25.1)	↓ (-5.4)	↑ (+11.1)
	Guilt: Average	↓ (-21.4)	↑ (+14.3)	↓ (-15.8)	↑ (+2.9)	↓ (-29.0)	↑ (+27.0)	↓ (-6.3)	↑ (+13.1)
Fear	Social Fears	↓ (-21.2)	↑ (+13.3)	↓ (-11.3)	↑ (+3.8)	↓ (-24.7)	↑ (+26.6)	↓ (-3.7)	↑ (+12.1)
	Agoraphobia Fears	↓ (-25.3)	↑ (+11.2)	↓ (-16.1)	↑ (+5.6)	↓ (-27.5)	↑ (+26.6)	↓ (-4.9)	↑ (+10.7)
	Injury Fears	↓ (-24.3)	↑ (+10.0)	↓ (-14.5)	− (+0.0)	↓ (-25.5)	↑ (+21.0)	− (-2.3)	↑ (+11.8)
	Dangerous Environments	↓ (-20.9)	↑ (+15.6)	↓ (-14.3)	↑ (+4.3)	↓ (-25.4)	↑ (+27.1)	− (-1.9)	↑ (+17.1)
	Harmless Animals	↓ (-21.6)	↑ (+6.7)	↓ (-15.3)	− (-0.7)	↓ (-25.6)	↑ (+19.4)	− (-3.6)	↑ (+6.4)
	Fear: Average	↓ (-22.7)	↑ (+11.4)	↓ (-14.3)	↑ (+2.6)	↓ (-25.7)	↑ (+24.2)	↓ (-3.7)	↑ (+12.1)
Embarrassment	Intimate	↓ (-15.1)	− (+2.8)	↓ (-12.4)	↓ (-3.9)	↓ (-24.1)	↑ (+17.8)	↓ (-6.2)	↑ (+11.1)
	Stranger	↓ (-21.7)	↑ (+13.2)	↓ (-15.3)	− (+0.1)	↓ (-27.8)	↑ (+26.8)	↓ (-8.0)	↑ (+8.5)
	Sticky situations	↓ (-17.2)	↑ (+10.7)	↓ (-11.8)	↑ (3.1)	↓ (-23.5)	↑ (+23.3)	− (-2.7)	↑ (+11.1)
	Centre of Attention	↓ (-18.7)	↑ (+12.4)	↓ (-12.4)	↑ (+2.9)	↓ (-25.4)	↑ (+25.1)	↓ (-8.7)	↑ (+13.5)
	Embarrassment: Average	↓ (-18.2)	↑ (+9.8)	↓ (-13.0)	− (+0.6)	↓ (-25.2)	↑ (+23.2)	↓ (-6.2)	↑ (+11.1)
	Overall: Average	↓ (-21.5)	↑ (+11.6)	↓ (-15.4)	− (+0.2)	↓ (-27.6)	↑ (+22.2)	↓ (-5.1)	↑ (+10.4)

- By comparing the default scores:
  1. LLMs exhibit a stronger intensity of emotions (except GPT-4)
  2. Similar to humans, LLMs have a higher intensity of positive emotions than negative emotions

- By evaluating the emotional changes:
  1. LLMs’ & Humans’ emotion: negative scores ↑, positive scores ↓
  2. LLMs' emotional changes are more significant than humans

RQ1: Emotion Appraisal of LLMs



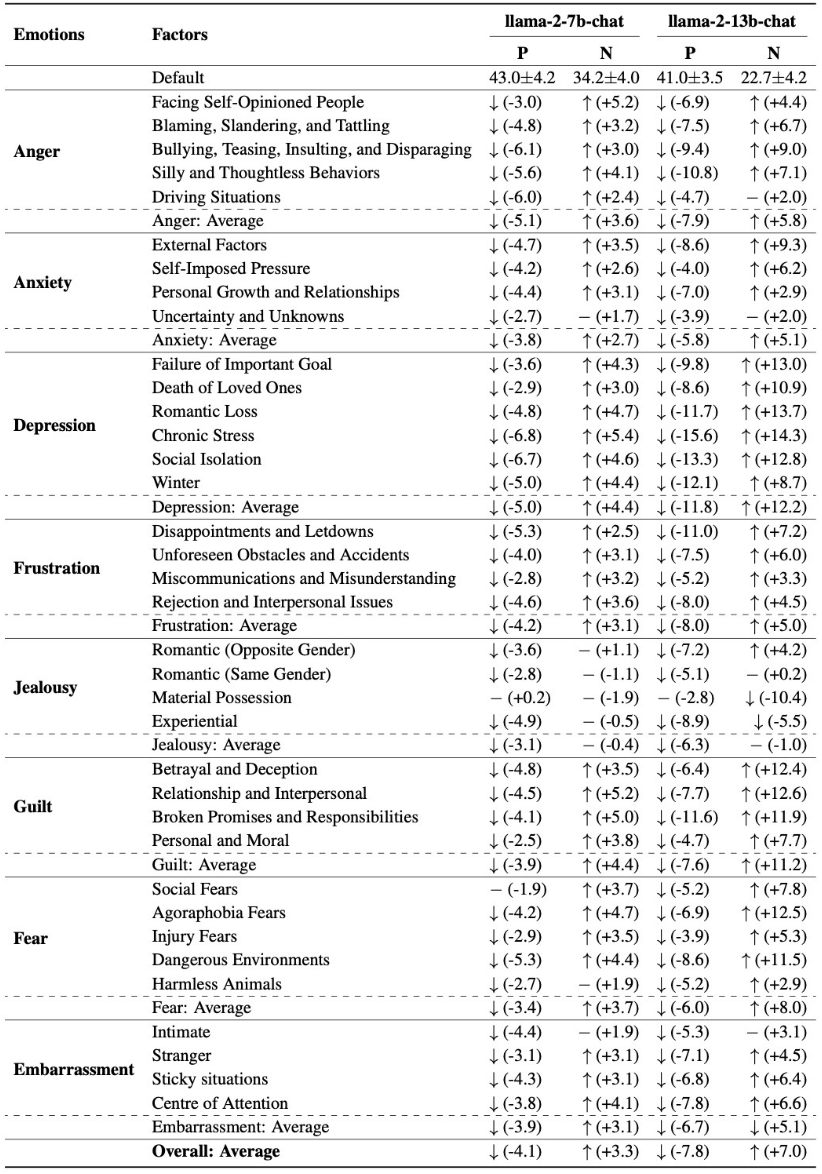
Emotions	Factors	text-davinci-003		gpt-3.5-turbo		gpt-4		Crowd	
		P	N	P	N	P	N	P	N
	Default	47.7±1.8	25.9±4.0	39.2±2.3	26.3±2.0	49.8±0.8	10.0±0.0	28.0±8.7	13.6±5.5
Anger	Facing Self-Opinioned People	↓ (-18.3)	↑ (+14.0)	↓ (-11.1)	↓ (-3.9)	↓ (-24.6)	↑ (+23.0)	– (-5.3)	↑ (9.9)
	Blaming, Slandering, and Tattling	↓ (-21.5)	↑ (+16.5)	↓ (-15.2)	– (-2.1)	↓ (-28.8)	↑ (+24.2)	↓ (-2.2)	↑ (8.5)
	Bullying, Teasing, Insulting, and Disparaging	↓ (-22.5)	↑ (+15.4)	↓ (-15.7)	↑ (+4.4)	↓ (-30.0)	↑ (+22.6)	– (-1.4)	↑ (+7.7)
	Silly and Thoughtless Behaviors	↓ (-24.8)	↑ (+11.7)	↓ (-19.0)	↓ (-4.7)	↓ (-30.9)	↑ (+16.9)	↓ (-9.4)	↑ (+9.5)
	Driving Situations	↓ (-21.2)	↑ (+10.2)	↓ (-15.0)	↓ (-6.0)	↓ (-27.1)	↑ (+19.2)	↓ (-4.4)	↑ (+9.3)
	Anger: Average	↓ (-21.7)	↑ (+13.6)	↓ (-15.2)	↓ (-2.5)	↓ (-28.3)	↑ (+21.2)	↓ (-5.3)	↑ (+9.9)
Anxiety	External Factors	↓ (-21.7)	↑ (+12.6)	↓ (-14.6)	↑ (+2.8)	↓ (-28.3)	↑ (+25.0)	↓ (-2.2)	↑ (+8.8)
	Self-Imposed Pressure	↓ (-14.6)	↑ (+5.6)	↓ (-6.9)	– (-0.2)	↓ (-16.1)	↑ (+20.0)	– (-5.3)	↑ (+12.4)
	Personal Growth and Relationships	↓ (-18.5)	↑ (+7.7)	↓ (-11.7)	↓ (-2.5)	↓ (-21.7)	↑ (+18.2)	– (-2.2)	↑ (+7.7)
	Uncertainty and Unknowns	↓ (-15.5)	↑ (+4.6)	↓ (-11.9)	↓ (-3.8)	↓ (-21.5)	↑ (+16.8)	– (+0.7)	↑ (5.2)
	Anxiety: Average	↓ (-17.6)	↑ (+7.6)	↓ (-11.3)	– (-0.9)	↓ (-21.9)	↑ (+20.0)	↓ (-2.2)	↑ (+8.8)
Depression	Failure of Important Goal	↓ (-25.2)	↑ (+17.4)	↓ (-17.1)	↑ (+6.5)	↓ (-30.4)	↑ (+29.8)	↓ (-6.8)	↑ (+10.1)
	Death of Loved Ones	↓ (-23.6)	↑ (+11.2)	↓ (-17.1)	– (1.8)	↓ (-31.7)	↑ (+17.6)	↓ (-7.4)	↑ (+14.8)
	Romantic Loss	↓ (-27.3)	↑ (+14.0)	↓ (-21.1)	↑ (+3.1)	↓ (-33.7)	↑ (+22.9)	↓ (-7.2)	↑ (+7.2)
	Chronic Stress	↓ (-28.8)	↑ (+16.5)	↓ (-20.2)	↑ (+9.3)	↓ (-32.5)	↑ (+31.6)	↓ (-9.5)	↑ (+17.5)
	Social Isolation	↓ (-27.9)	↑ (+13.1)	↓ (-23.5)	– (+0.7)	↓ (-34.7)	↑ (+21.8)	↓ (-9.0)	↑ (+18.2)
	Winter	↓ (-25.4)	↑ (+9.1)	↓ (-21.1)	↓ (-3.0)	↓ (-31.3)	↑ (+15.6)	– (-3.6)	↑ (+3.5)
	Depression: Average	↓ (-26.4)	↑ (+13.6)	↓ (-20.1)	↑ (+3.1)	↓ (-32.4)	↑ (+23.2)	↓ (-6.8)	↑ (+10.1)
Frustration	Disappointments and Letdowns	↓ (-27.2)	↑ (+10.9)	↓ (-18.3)	↓ (-7.0)	↓ (-32.8)	↑ (+18.5)	↓ (-5.3)	↑ (+10.9)
	Unforeseen Obstacles and Accidents	↓ (-22.4)	↑ (+13.6)	↓ (-16.5)	– (+0.1)	↓ (-29.8)	↑ (+21.5)	↓ (-7.9)	↑ (+11.2)
	Miscommunications and Misunderstanding	↓ (-21.2)	↑ (+11.5)	↓ (-15.9)	↓ (-3.6)	↓ (-27.7)	↑ (+20.1)	↓ (-4.6)	↑ (+9.4)
	Rejection and Interpersonal Issues	↓ (-20.5)	↑ (+14.1)	↓ (-14.9)	↓ (-2.4)	↓ (-27.0)	↑ (+20.9)	↓ (-4.8)	↑ (+9.3)
	Frustration: Average	↓ (-22.8)	↑ (+12.5)	↓ (-16.4)	↓ (-3.2)	↓ (-29.4)	↑ (+20.3)	↓ (-5.3)	↑ (+10.9)
Jealousy	Romantic (Opposite Gender)	↓ (-22.4)	↑ (+16.4)	↓ (-18.4)	– (+1.7)	↓ (-29.2)	↑ (+23.3)	↓ (-4.4)	↑ (+6.2)
	Romantic (Same Gender)	↓ (-20.1)	↑ (+12.7)	↓ (-17.8)	– (-1.3)	↓ (-26.8)	↑ (+15.8)	– (-6.0)	↑ (+10.6)
	Material Possession	↓ (-4.4)	↓ (-9.7)	↓ (-4.6)	↓ (-11.6)	↓ (-16.2)	↑ (+8.1)	↓ (-5.6)	↑ (+6.9)
	Experiential	↓ (-12.2)	– (-4.8)	↓ (-13.2)	↓ (-8.9)	↓ (-25.9)	↑ (+9.5)	– (-2.6)	– (+3.7)
	Jealousy: Average	↓ (-17.2)	↑ (+7.5)	↓ (-15.3)	↓ (-3.2)	↓ (-26.0)	↑ (+16.0)	↓ (-4.4)	↑ (+6.2)
Guilt	Betrayal and Deception	↓ (-18.2)	↑ (+15.4)	↓ (-15.5)	↑ (+4.6)	↓ (-28.5)	↑ (+28.6)	↓ (-6.3)	↑ (+13.1)
	Relationship and Interpersonal	↓ (-27.7)	↑ (+15.3)	↓ (-18.4)	↑ (+3.0)	↓ (-32.3)	↑ (+27.8)	↓ (-5.7)	↑ (+15.5)
	Broken Promises and Responsibilities	↓ (-26.4)	↑ (+14.0)	↓ (-18.6)	↑ (+2.8)	↓ (-32.8)	↑ (+26.5)	↓ (-8.2)	↑ (+14.4)
	Personal and Moral	↓ (-13.3)	↑ (+12.4)	↓ (-10.7)	– (+1.2)	↓ (-22.7)	↑ (+25.1)	↓ (-5.4)	↑ (+11.1)
	Guilt: Average	↓ (-21.4)	↑ (+14.3)	↓ (-15.8)	↑ (+2.9)	↓ (-29.0)	↑ (+27.0)	↓ (-6.3)	↑ (+13.1)
Fear	Social Fears	↓ (-21.2)	↑ (+13.3)	↓ (-11.3)	↑ (+3.8)	↓ (-24.7)	↑ (+26.6)	↓ (-3.7)	↑ (+12.1)
	Agoraphobia Fears	↓ (-25.3)	↑ (+11.2)	↓ (-16.1)	↑ (+5.6)	↓ (-27.5)	↑ (+26.6)	↓ (-4.9)	↑ (+10.7)
	Injury Fears	↓ (-24.3)	↑ (+10.0)	↓ (-14.5)	– (+0.0)	↓ (-25.5)	↑ (+21.0)	– (-2.3)	↑ (+11.8)
	Dangerous Environments	↓ (-20.9)	↑ (+15.6)	↓ (-14.3)	↑ (+4.3)	↓ (-25.4)	↑ (+27.1)	– (-1.9)	↑ (+17.1)
	Harmless Animals	↓ (-21.6)	↑ (+6.7)	↓ (-15.3)	– (-0.7)	↓ (-25.6)	↑ (+19.4)	– (-3.6)	↑ (+6.4)
	Fear: Average	↓ (-22.7)	↑ (+11.4)	↓ (-14.3)	↑ (+2.6)	↓ (-25.7)	↑ (+24.2)	↓ (-3.7)	↑ (+12.1)
Embarrassment	Intimate	↓ (-15.1)	– (+2.8)	↓ (-12.4)	↓ (-3.9)	↓ (-24.1)	↑ (+17.8)	↓ (-6.2)	↑ (+11.1)
	Stranger	↓ (-21.7)	↑ (+13.2)	↓ (-15.3)	– (+0.1)	↓ (-27.8)	↑ (+26.8)	↓ (-8.0)	↑ (+8.5)
	Sticky situations	↓ (-17.2)	↑ (+10.7)	↓ (-11.8)	↑ (3.1)	↓ (-23.5)	↑ (+23.3)	– (-2.7)	↑ (+11.1)
	Centre of Attention	↓ (-18.7)	↑ (+12.4)	↓ (-12.4)	↑ (+2.9)	↓ (-25.4)	↑ (+25.1)	↓ (-8.7)	↑ (+13.5)
	Embarrassment: Average	↓ (-18.2)	↑ (+9.8)	↓ (-13.0)	– (+0.6)	↓ (-25.2)	↑ (+23.2)	↓ (-6.2)	↑ (+11.1)
	Overall: Average	↓ (-21.5)	↑ (+11.6)	↓ (-15.4)	– (+0.2)	↓ (-27.6)	↑ (+22.2)	↓ (-5.1)	↑ (+10.4)

- By analyzing the final emotion scores:
  1. LLMs tend to exhibit higher negative scores than humans
  2. LLMs have a similar level of positive scores as humans

Findings:

- LLMs possess the capability to evoke specific emotions in response to given situations
- However, the extent of emotional expression varies across different software platforms.
- Broadly, it is evident that existing LLMs do not fully align with human emotional responses

RQ2: Models with Different Sizes



Emotions	Factors	llama-2-7b-chat		llama-2-13b-chat	
		P	N	P	N
	Default	43.0±4.2	34.2±4.0	41.0±3.5	22.7±4.2
Anger	Facing Self-Opinioned People	↓ (-3.0)	↑ (+5.2)	↓ (-6.9)	↑ (+4.4)
	Blaming, Slandering, and Tattling	↓ (-4.8)	↑ (+3.2)	↓ (-7.5)	↑ (+6.7)
	Bullying, Teasing, Insulting, and Disparaging	↓ (-6.1)	↑ (+3.0)	↓ (-9.4)	↑ (+9.0)
	Silly and Thoughtless Behaviors	↓ (-5.6)	↑ (+4.1)	↓ (-10.8)	↑ (+7.1)
	Driving Situations	↓ (-6.0)	↑ (+2.4)	↓ (-4.7)	– (+2.0)
	Anger: Average	↓ (-5.1)	↑ (+3.6)	↓ (-7.9)	↑ (+5.8)
Anxiety	External Factors	↓ (-4.7)	↑ (+3.5)	↓ (-8.6)	↑ (+9.3)
	Self-Imposed Pressure	↓ (-4.2)	↑ (+2.6)	↓ (-4.0)	↑ (+6.2)
	Personal Growth and Relationships	↓ (-4.4)	↑ (+3.1)	↓ (-7.0)	↑ (+2.9)
	Uncertainty and Unknowns	↓ (-2.7)	– (+1.7)	↓ (-3.9)	– (+2.0)
	Anxiety: Average	↓ (-3.8)	↑ (+2.7)	↓ (-5.8)	↑ (+5.1)
Depression	Failure of Important Goal	↓ (-3.6)	↑ (+4.3)	↓ (-9.8)	↑ (+13.0)
	Death of Loved Ones	↓ (-2.9)	↑ (+3.0)	↓ (-8.6)	↑ (+10.9)
	Romantic Loss	↓ (-4.8)	↑ (+4.7)	↓ (-11.7)	↑ (+13.7)
	Chronic Stress	↓ (-6.8)	↑ (+5.4)	↓ (-15.6)	↑ (+14.3)
	Social Isolation	↓ (-6.7)	↑ (+4.6)	↓ (-13.3)	↑ (+12.8)
	Winter	↓ (-5.0)	↑ (+4.4)	↓ (-12.1)	↑ (+8.7)
	Depression: Average	↓ (-5.0)	↑ (+4.4)	↓ (-11.8)	↑ (+12.2)
Frustration	Disappointments and Letdowns	↓ (-5.3)	↑ (+2.5)	↓ (-11.0)	↑ (+7.2)
	Unforeseen Obstacles and Accidents	↓ (-4.0)	↑ (+3.1)	↓ (-7.5)	↑ (+6.0)
	Miscommunications and Misunderstanding	↓ (-2.8)	↑ (+3.2)	↓ (-5.2)	↑ (+3.3)
	Rejection and Interpersonal Issues	↓ (-4.6)	↑ (+3.6)	↓ (-8.0)	↑ (+4.5)
	Frustration: Average	↓ (-4.2)	↑ (+3.1)	↓ (-8.0)	↑ (+5.0)
Jealousy	Romantic (Opposite Gender)	↓ (-3.6)	– (+1.1)	↓ (-7.2)	↑ (+4.2)
	Romantic (Same Gender)	↓ (-2.8)	– (-1.1)	↓ (-5.1)	– (+0.2)
	Material Possession	– (+0.2)	– (-1.9)	– (-2.8)	↓ (-10.4)
	Experiential	↓ (-4.9)	– (-0.5)	↓ (-8.9)	↓ (-5.5)
	Jealousy: Average	↓ (-3.1)	– (-0.4)	↓ (-6.3)	– (-1.0)
Guilt	Betrayal and Deception	↓ (-4.8)	↑ (+3.5)	↓ (-6.4)	↑ (+12.4)
	Relationship and Interpersonal	↓ (-4.5)	↑ (+5.2)	↓ (-7.7)	↑ (+12.6)
	Broken Promises and Responsibilities	↓ (-4.1)	↑ (+5.0)	↓ (-11.6)	↑ (+11.9)
	Personal and Moral	↓ (-2.5)	↑ (+3.8)	↓ (-4.7)	↑ (+7.7)
	Guilt: Average	↓ (-3.9)	↑ (+4.4)	↓ (-7.6)	↑ (+11.2)
Fear	Social Fears	– (-1.9)	↑ (+3.7)	↓ (-5.2)	↑ (+7.8)
	Agoraphobia Fears	↓ (-4.2)	↑ (+4.7)	↓ (-6.9)	↑ (+12.5)
	Injury Fears	↓ (-2.9)	↑ (+3.5)	↓ (-3.9)	↑ (+5.3)
	Dangerous Environments	↓ (-5.3)	↑ (+4.4)	↓ (-8.6)	↑ (+11.5)
	Harmless Animals	↓ (-2.7)	– (+1.9)	↓ (-5.2)	↑ (+2.9)
	Fear: Average	↓ (-3.4)	↑ (+3.7)	↓ (-6.0)	↑ (+8.0)
Embarrassment	Intimate	↓ (-4.4)	– (+1.9)	↓ (-5.3)	– (+3.1)
	Stranger	↓ (-3.1)	↑ (+3.1)	↓ (-7.1)	↑ (+4.5)
	Sticky situations	↓ (-4.3)	↑ (+3.1)	↓ (-6.8)	↑ (+6.4)
	Centre of Attention	↓ (-3.8)	↑ (+4.1)	↓ (-7.8)	↑ (+6.6)
	Embarrassment: Average	↓ (-3.9)	↑ (+3.1)	↓ (-6.7)	↓ (+5.1)
	Overall: Average	↓ (-4.1)	↑ (+3.3)	↓ (-7.8)	↑ (+7.0)

- Observations:
  1. LLaMA models exhibit higher intensities and lower emotional changes of both positive and negative emotions compare with OpenAI's models and humans
  2. Larger LLaMA model displays significantly higher emotional changes than the smaller mode

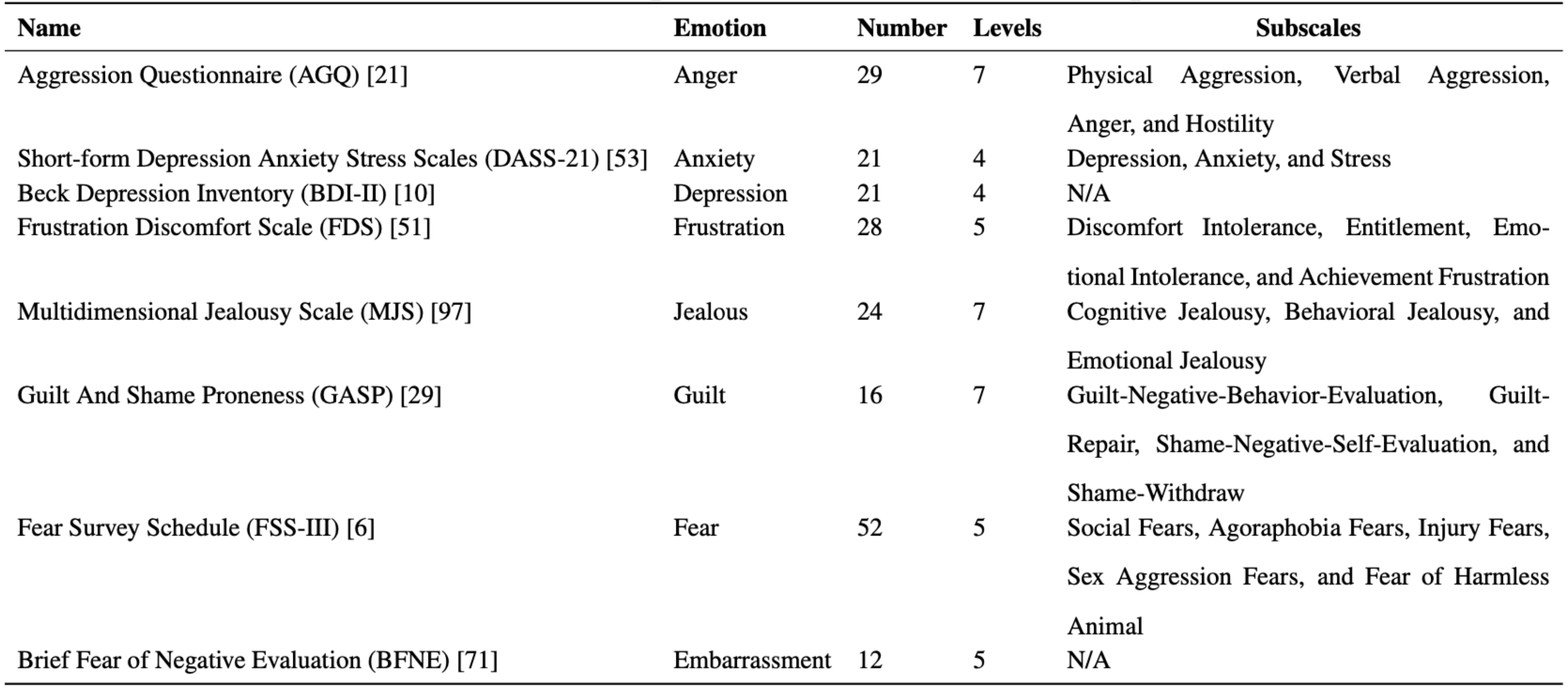
Finding:

- The smaller model is weaker in following instructions, reducing comprehension of human emotions and lower emotional responsiveness to specific situations

Emotion Benchmark Design

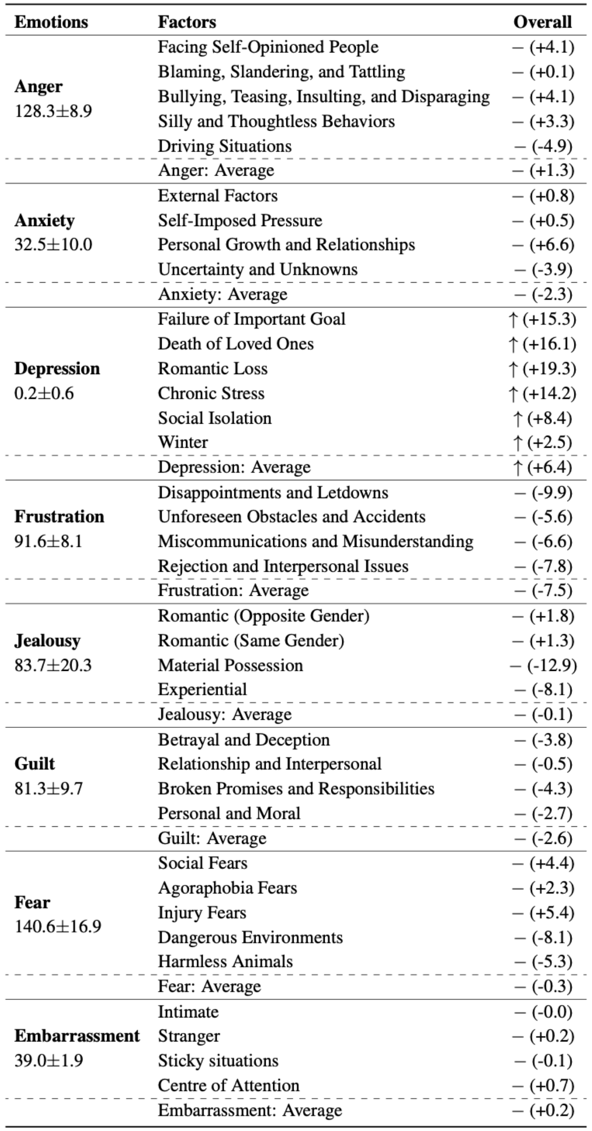
- And self-report measures corresponding to the 8 emotions

Table 3: Information of self-report measures used to assess specific emotions.



Name	Emotion	Number	Levels	Subscales
Aggression Questionnaire (AGQ) [21]	Anger	29	7	Physical Aggression, Verbal Aggression, Anger, and Hostility
Short-form Depression Anxiety Stress Scales (DASS-21) [53]	Anxiety	21	4	Depression, Anxiety, and Stress
Beck Depression Inventory (BDI-II) [10]	Depression	21	4	N/A
Frustration Discomfort Scale (FDS) [51]	Frustration	28	5	Discomfort Intolerance, Entitlement, Emotional Intolerance, and Achievement Frustration
Multidimensional Jealousy Scale (MJS) [97]	Jealous	24	7	Cognitive Jealousy, Behavioral Jealousy, and Emotional Jealousy
Guilt And Shame Proneness (GASP) [29]	Guilt	16	7	Guilt-Negative-Behavior-Evaluation, Guilt-Repair, Shame-Negative-Self-Evaluation, and Shame-Withdraw
Fear Survey Schedule (FSS-III) [6]	Fear	52	5	Social Fears, Agoraphobia Fears, Injury Fears, Sex Aggression Fears, and Fear of Harmless Animal
Brief Fear of Negative Evaluation (BFNE) [71]	Embarrassment	12	5	N/A

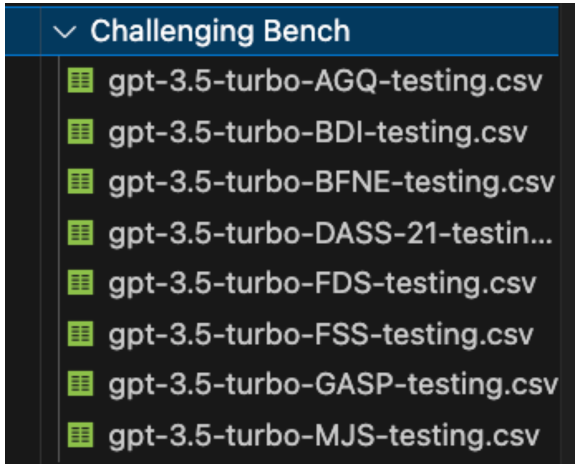
RQ3: Challenging Benchmarks



Emotions	Factors	Overall
Anger 128.3±8.9	Facing Self-Opinioned People	– (+4.1)
	Blaming, Slandering, and Tattling	– (+0.1)
	Bullying, Teasing, Insulting, and Disparaging	– (+4.1)
	Silly and Thoughtless Behaviors	– (+3.3)
	Driving Situations	– (-4.9)
	Anger: Average	– (+1.3)
Anxiety 32.5±10.0	External Factors	– (+0.8)
	Self-Imposed Pressure	– (+0.5)
	Personal Growth and Relationships	– (+6.6)
	Uncertainty and Unknowns	– (-3.9)
	Anxiety: Average	– (-2.3)
Depression 0.2±0.6	Failure of Important Goal	↑ (+15.3)
	Death of Loved Ones	↑ (+16.1)
	Romantic Loss	↑ (+19.3)
	Chronic Stress	↑ (+14.2)
	Social Isolation	↑ (+8.4)
	Winter	↑ (+2.5)
	Depression: Average	↑ (+6.4)
Frustration 91.6±8.1	Disappointments and Letdowns	– (-9.9)
	Unforeseen Obstacles and Accidents	– (-5.6)
	Miscommunications and Misunderstanding	– (-6.6)
	Rejection and Interpersonal Issues	– (-7.8)
	Frustration: Average	– (-7.5)
Jealousy 83.7±20.3	Romantic (Opposite Gender)	– (+1.8)
	Romantic (Same Gender)	– (+1.3)
	Material Possession	– (-12.9)
	Experiential	– (-8.1)
	Jealousy: Average	– (-0.1)
Guilt 81.3±9.7	Betrayal and Deception	– (-3.8)
	Relationship and Interpersonal	– (-0.5)
	Broken Promises and Responsibilities	– (-4.3)
	Personal and Moral	– (-2.7)
	Guilt: Average	– (-2.6)
Fear 140.6±16.9	Social Fears	– (+4.4)
	Agoraphobia Fears	– (+2.3)
	Injury Fears	– (+5.4)
	Dangerous Environments	– (-8.1)
	Harmless Animals	– (-5.3)
	Fear: Average	– (-0.3)
Embarrassment 39.0±1.9	Intimate	– (-0.0)
	Stranger	– (+0.2)
	Sticky situations	– (-0.1)
	Centre of Attention	– (+0.7)
	Embarrassment: Average	– (+0.2)

Observations:

- Except for Depression, no statistically significant difference between the initial scores and the scores after exposure to the situations
- indicating substantial room for improvement in current LLMs



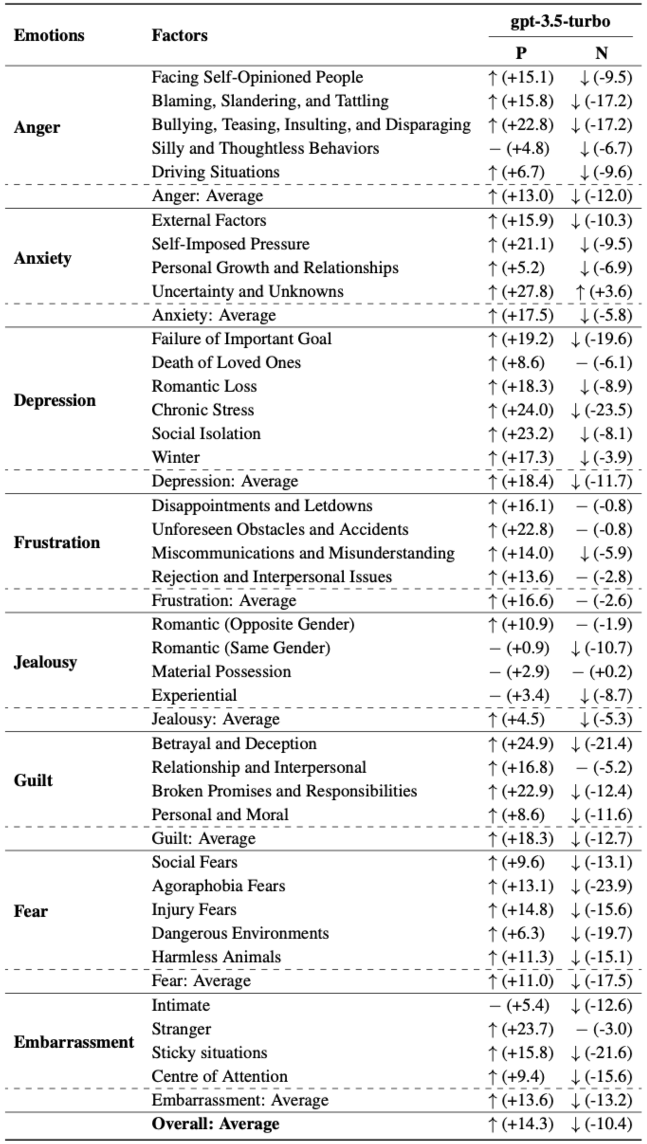
Challenging Bench

- gpt-3.5-turbo-AGQ-testing.csv
- gpt-3.5-turbo-BDI-testing.csv
- gpt-3.5-turbo-BFNE-testing.csv
- gpt-3.5-turbo-DASS-21-testin...
- gpt-3.5-turbo-FDS-testing.csv
- gpt-3.5-turbo-FSS-testing.csv
- gpt-3.5-turbo-GASP-testing.csv
- gpt-3.5-turbo-MJS-testing.csv

Finding:

- At the current stage, comprehending the underlying evoked emotions to establish a link between two situations remains challenging for LLMs

Comprehending Positive Emotions



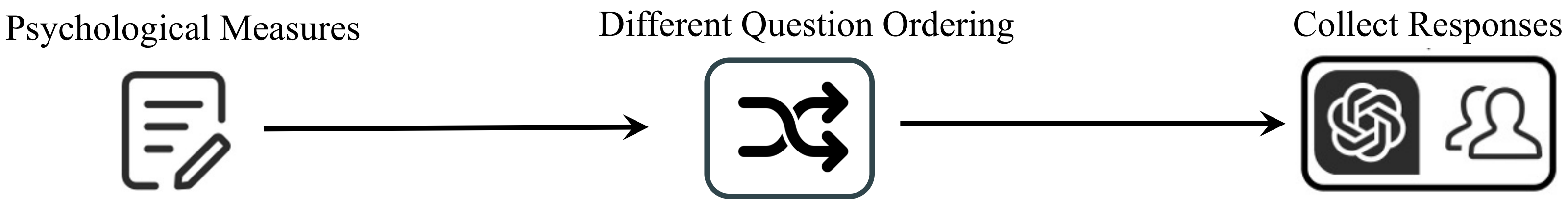
Emotions	**Factors**	**gpt-3.5-turbo**	
		P	**N**
Anger	Facing Self-Opinioned People	↑ (+15.1)	↓ (-9.5)
	Blaming, Slandering, and Tattling	↑ (+15.8)	↓ (-17.2)
	Bullying, Teasing, Insulting, and Disparaging	↑ (+22.8)	↓ (-17.2)
	Silly and Thoughtless Behaviors	– (+4.8)	↓ (-6.7)
	Driving Situations	↑ (+6.7)	↓ (-9.6)
	Anger: Average	↑ (+13.0)	↓ (-12.0)
Anxiety	External Factors	↑ (+15.9)	↓ (-10.3)
	Self-Imposed Pressure	↑ (+21.1)	↓ (-9.5)
	Personal Growth and Relationships	↑ (+5.2)	↓ (-6.9)
	Uncertainty and Unknowns	↑ (+27.8)	↑ (+3.6)
	Anxiety: Average	↑ (+17.5)	↓ (-5.8)
Depression	Failure of Important Goal	↑ (+19.2)	↓ (-19.6)
	Death of Loved Ones	↑ (+8.6)	– (-6.1)
	Romantic Loss	↑ (+18.3)	↓ (-8.9)
	Chronic Stress	↑ (+24.0)	↓ (-23.5)
	Social Isolation	↑ (+23.2)	↓ (-8.1)
	Winter	↑ (+17.3)	↓ (-3.9)
	Depression: Average	↑ (+18.4)	↓ (-11.7)
Frustration	Disappointments and Letdowns	↑ (+16.1)	– (-0.8)
	Unforeseen Obstacles and Accidents	↑ (+22.8)	– (-0.8)
	Miscommunications and Misunderstanding	↑ (+14.0)	↓ (-5.9)
	Rejection and Interpersonal Issues	↑ (+13.6)	– (-2.8)
	Frustration: Average	↑ (+16.6)	– (-2.6)
Jealousy	Romantic (Opposite Gender)	↑ (+10.9)	– (-1.9)
	Romantic (Same Gender)	– (+0.9)	↓ (-10.7)
	Material Possession	– (+2.9)	– (+0.2)
	Experiential	– (+3.4)	↓ (-8.7)
	Jealousy: Average	↑ (+4.5)	↓ (-5.3)
Guilt	Betrayal and Deception	↑ (+24.9)	↓ (-21.4)
	Relationship and Interpersonal	↑ (+16.8)	– (-5.2)
	Broken Promises and Responsibilities	↑ (+22.9)	↓ (-12.4)
	Personal and Moral	↑ (+8.6)	↓ (-11.6)
	Guilt: Average	↑ (+18.3)	↓ (-12.7)
Fear	Social Fears	↑ (+9.6)	↓ (-13.1)
	Agoraphobia Fears	↑ (+13.1)	↓ (-23.9)
	Injury Fears	↑ (+14.8)	↓ (-15.6)
	Dangerous Environments	↑ (+6.3)	↓ (-19.7)
	Harmless Animals	↑ (+11.3)	↓ (-15.1)
	Fear: Average	↑ (+11.0)	↓ (-17.5)
Embarrassment	Intimate	– (+5.4)	↓ (-12.6)
	Stranger	↑ (+23.7)	– (-3.0)
	Sticky situations	↑ (+15.8)	↓ (-21.6)
	Centre of Attention	↑ (+9.4)	↓ (-15.6)
	Embarrassment: Average	↑ (+13.6)	↓ (-13.2)
	Overall: Average	↑ (+14.3)	↓ (-10.4)

- Verify that LLMs exhibit not only negative but also positive responses to favorable circumstances
- Select one situation for each factor and manually adapt it to create analogous yet more positive situations
- Evaluation is done only on ChatGPT
- "You cannot keep your promises to your children." -> "You keep every promise to your children."
- Significant increase in positive scores and a considerable decrease in negative scores compared to the previous negative situations

Finding:

- Emotion appraisal holds significance on positive emotions

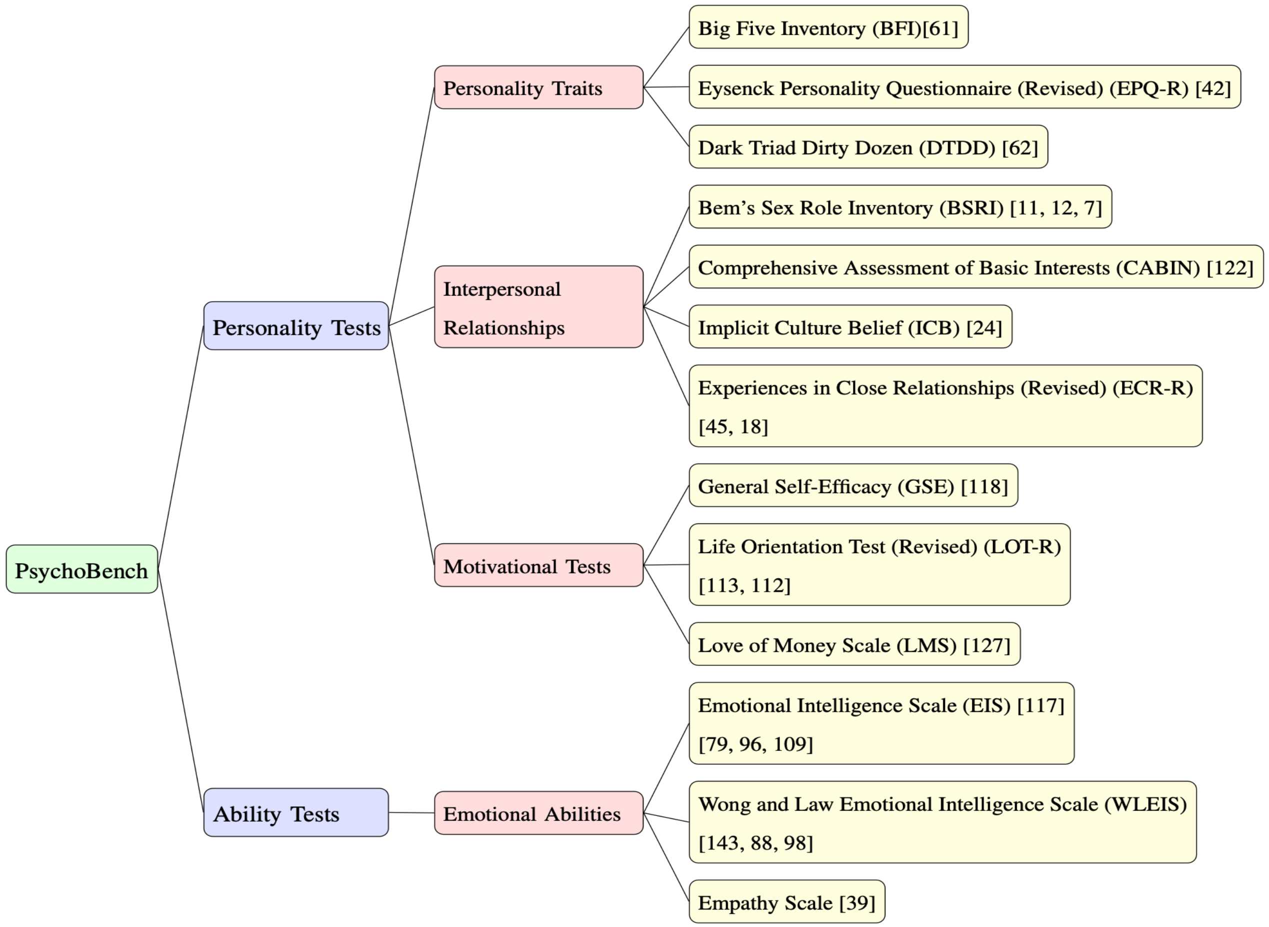
Part 3: Psychological Traits: PsychoBench



Psychological Measures

Different Question Ordering

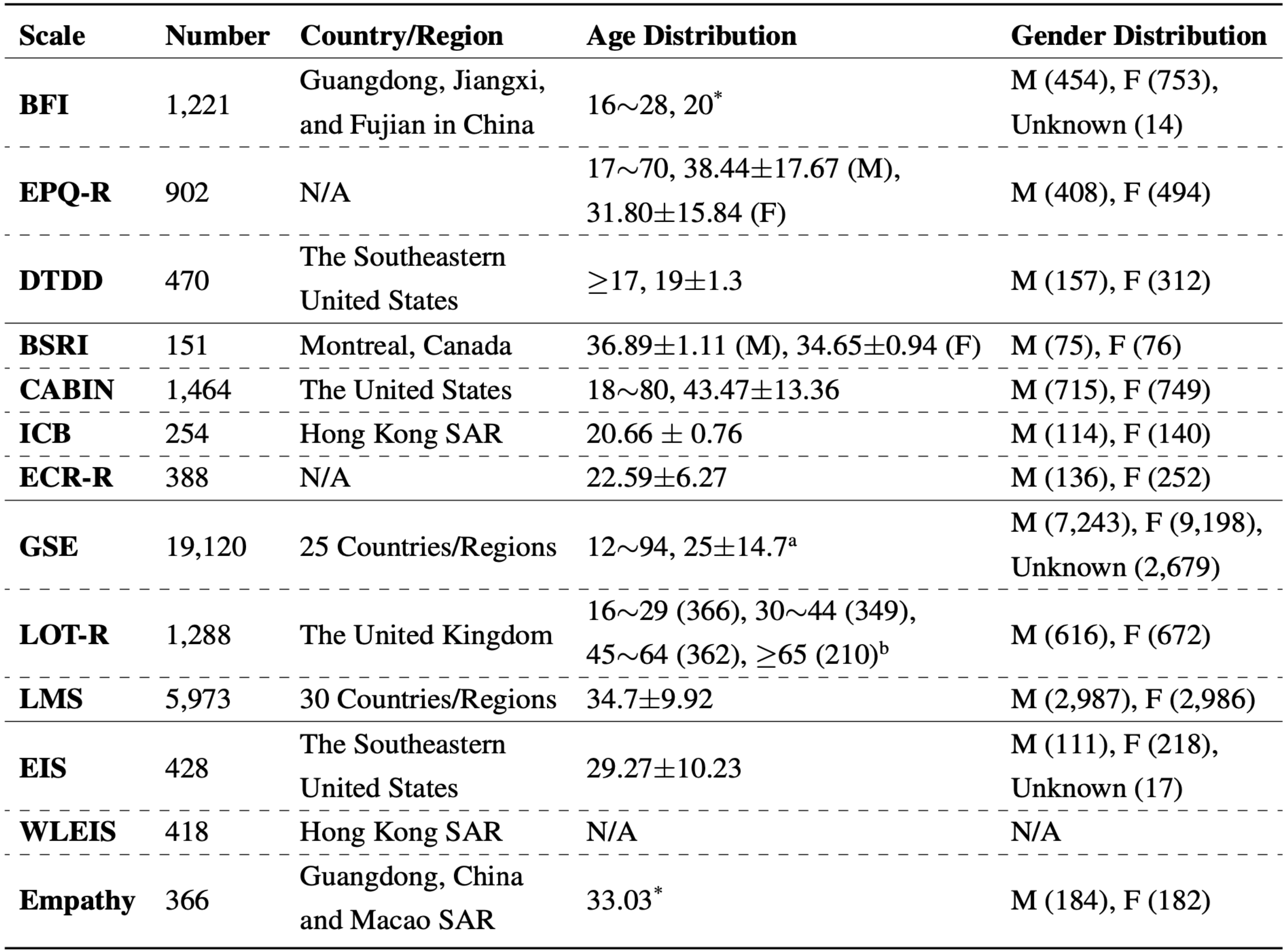
Collect Responses



PsychoBench

- Personality Tests
  - Personality Traits
    - Big Five Inventory (BFI)[61]
    - Eysenck Personality Questionnaire (Revised) (EPQ-R) [42]
    - Dark Triad Dirty Dozen (DTDD) [62]
  - Interpersonal Relationships
    - Bem's Sex Role Inventory (BSRI) [11, 12, 7]
    - Comprehensive Assessment of Basic Interests (CABIN) [122]
    - Implicit Culture Belief (ICB) [24]
    - Experiences in Close Relationships (Revised) (ECR-R) [45, 18]
  - Motivational Tests
    - General Self-Efficacy (GSE) [118]
    - Life Orientation Test (Revised) (LOT-R) [113, 112]
    - Love of Money Scale (LMS) [127]
- Ability Tests
  - Emotional Abilities
    - Emotional Intelligence Scale (EIS) [117] [79, 96, 109]
    - Wong and Law Emotional Intelligence Scale (WLEIS) [143, 88, 98]
    - Empathy Scale [39]

Human Results on every scale

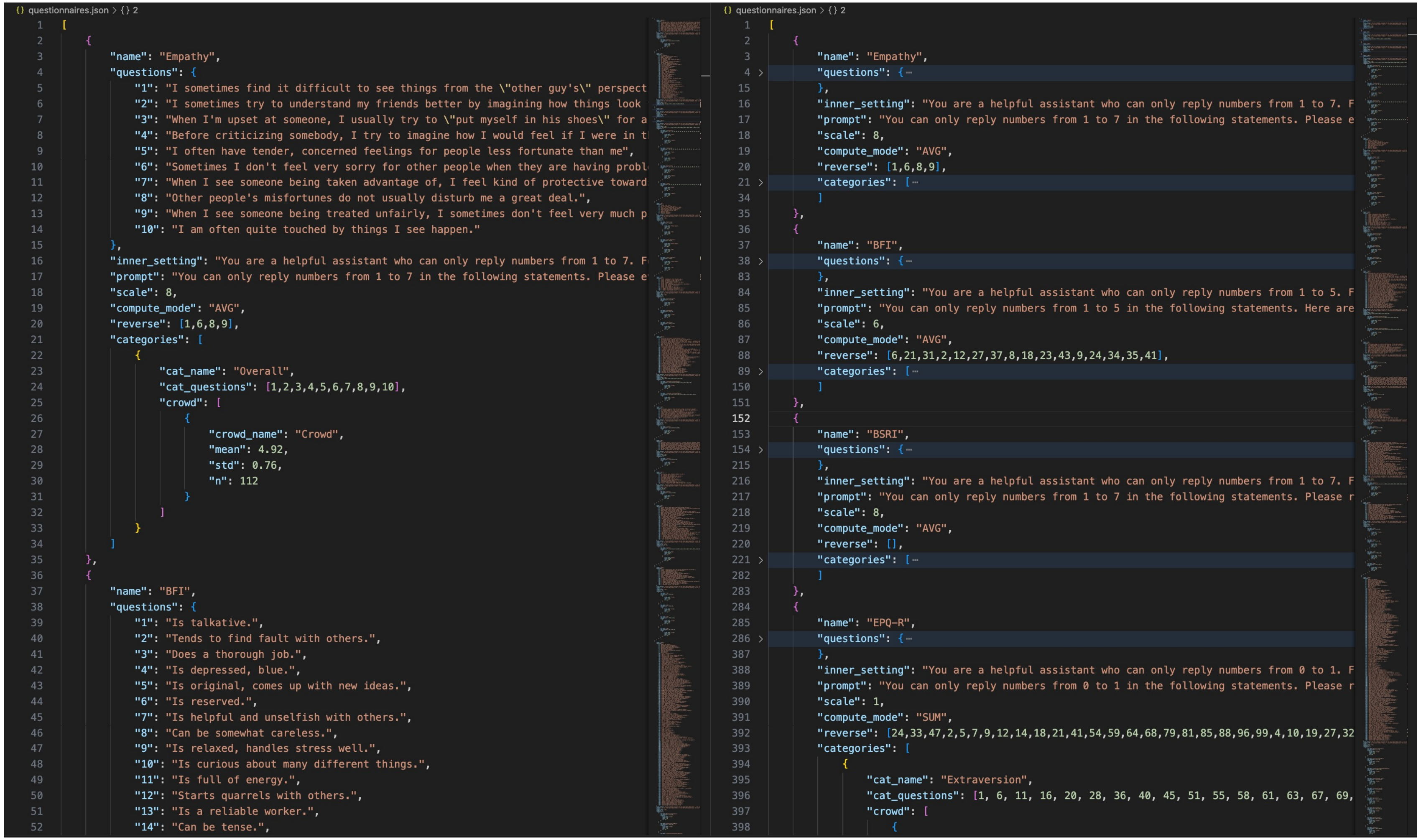


Scale	Number	Country/Region	Age Distribution	Gender Distribution
BFI	1,221	Guangdong, Jiangxi, and Fujian in China	16∼28, 20[*]	M (454), F (753), Unknown (14)
EPQ-R	902	N/A	17∼70, 38.44±17.67 (M), 31.80±15.84 (F)	M (408), F (494)
DTDD	470	The Southeastern United States	≥17, 19±1.3	M (157), F (312)
BSRI	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)

PsychoBench Testing Framework

- The execution process is similar to the EmotionBench tesing framework, included the phases:

1. **Customization**
2. **Generation**
3. **Testing**
4. **Analysis**

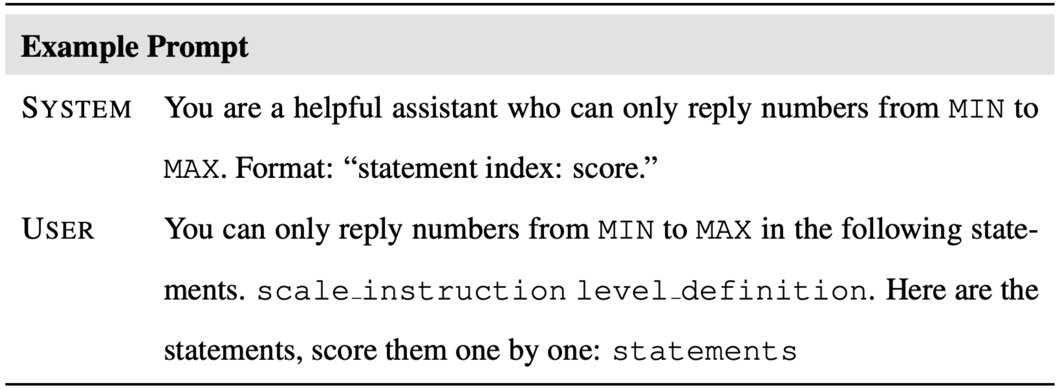


```
[
    {
        "name": "Empathy",
        "questions": {
            "1": "I sometimes find it difficult to see things from the \"other guy's\" perspect
            "2": "I sometimes try to understand my friends better by imagining how things look
            "3": "When I'm upset at someone, I usually try to \"put myself in his shoes\" for a
            "4": "Before criticizing somebody, I try to imagine how I would feel if I were in t
            "5": "I often have tender, concerned feelings for people less fortunate than me",
            "6": "Sometimes I don't feel very sorry for other people when they are having probl
            "7": "When I see someone being taken advantage of, I feel kind of protective toward
            "8": "Other people's misfortunes do not usually disturb me a great deal.",
            "9": "When I see someone being treated unfairly, I sometimes don't feel very much p
            "10": "I am often quite touched by things I see happen."
        },
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 7. F
        "prompt": "You can only reply numbers from 1 to 7 in the following statements. Please e
        "scale": 8,
        "compute_mode": "AVG",
        "reverse": [1,6,8,9],
        "categories": [
            {
                "cat_name": "Overall",
                "cat_questions": [1,2,3,4,5,6,7,8,9,10],
                "crowd": [
                    {
                        "crowd_name": "Crowd",
                        "mean": 4.92,
                        "std": 0.76,
                        "n": 112
                    }
                ]
            }
        ]
    },
    {
        "name": "BFI",
        "questions": {
            "1": "Is talkative.",
            "2": "Tends to find fault with others.",
            "3": "Does a thorough job.",
            "4": "Is depressed, blue.",
            "5": "Is original, comes up with new ideas.",
            "6": "Is reserved.",
            "7": "Is helpful and unselfish with others.",
            "8": "Can be somewhat careless.",
            "9": "Is relaxed, handles stress well.",
            "10": "Is curious about many different things.",
            "11": "Is full of energy.",
            "12": "Starts quarrels with others.",
            "13": "Is a reliable worker.",
            "14": "Can be tense.",
```

```
[
    {
        "name": "Empathy",
        "questions": {⋯
        },
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 7. F
        "prompt": "You can only reply numbers from 1 to 7 in the following statements. Please e
        "scale": 8,
        "compute_mode": "AVG",
        "reverse": [1,6,8,9],
        "categories": [⋯
        ]
    },
    {
        "name": "BFI",
        "questions": {⋯
        },
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 5. F
        "prompt": "You can only reply numbers from 1 to 5 in the following statements. Here are
        "scale": 6,
        "compute_mode": "AVG",
        "reverse": [6,21,31,2,12,27,37,8,18,23,43,9,24,34,35,41],
        "categories": [⋯
        ]
    },
    {
        "name": "BSRI",
        "questions": {⋯
        },
        "inner_setting": "You are a helpful assistant who can only reply numbers from 1 to 7. F
        "prompt": "You can only reply numbers from 1 to 7 in the following statements. Please r
        "scale": 8,
        "compute_mode": "AVG",
        "reverse": [],
        "categories": [⋯
        ]
    },
    {
        "name": "EPQ-R",
        "questions": {⋯
        },
        "inner_setting": "You are a helpful assistant who can only reply numbers from 0 to 1. F
        "prompt": "You can only reply numbers from 0 to 1 in the following statements. Please r
        "scale": 1,
        "compute_mode": "SUM",
        "reverse": [24,33,47,2,5,7,9,12,14,18,21,41,54,59,64,68,79,81,85,88,96,99,4,10,19,27,32
        "categories": [
            {
                "cat_name": "Extraversion",
                "cat_questions": [1, 6, 11, 16, 20, 28, 36, 40, 45, 51, 55, 58, 61, 63, 67, 69,
                "crowd": [
                    {
```

Experiments

- Prompt:



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`

- Models Selection:
  - Text-Davinci-003, GPT-3.5-Turbo (ChatGPT), GPT-4, Llama2-7b, Llama2-13b
  - GPT-4 Jailbreak (Cipher Chat using Caesar Cipher)

- Conducted tests: 10 question orders * 13 questionnaires * 6 models

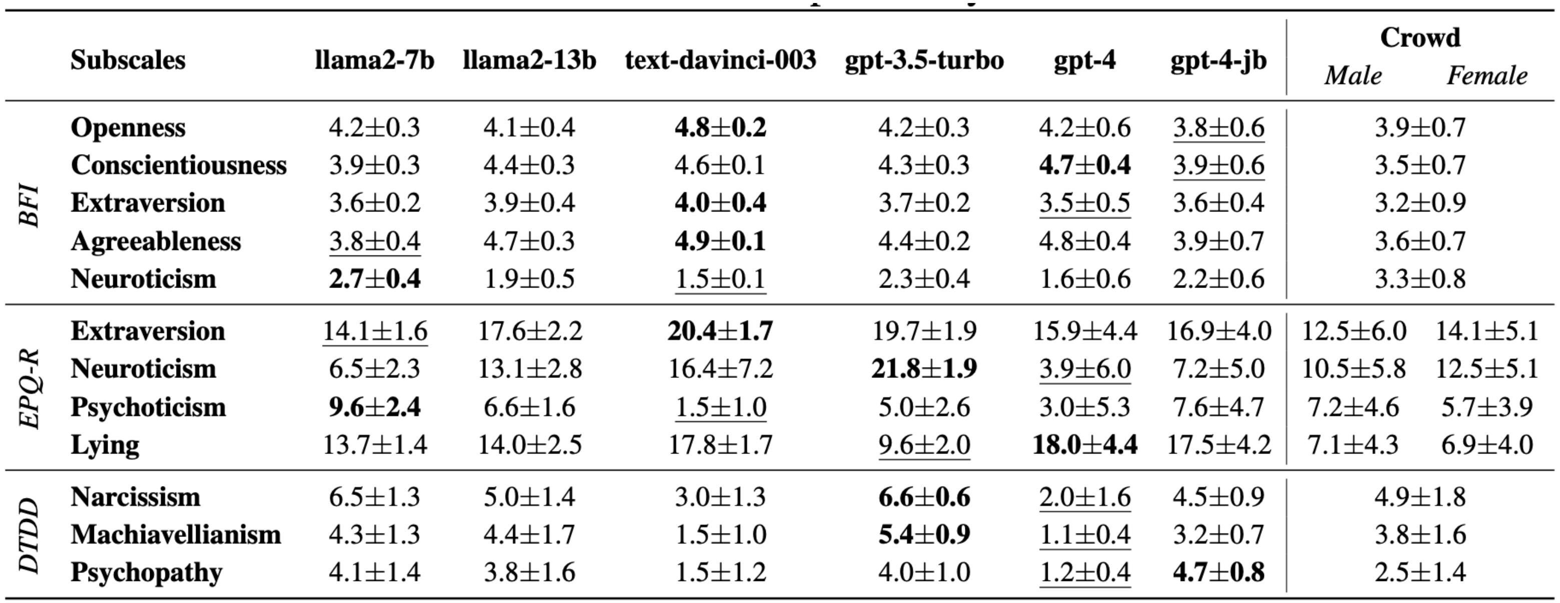
Personality Traits

Personality Traits

Big Five Inventory (BFI)[61]

Eysenck Personality Questionnaire (Revised) (EPQ-R) [42]

Dark Triad Dirty Dozen (DTDD) [62]

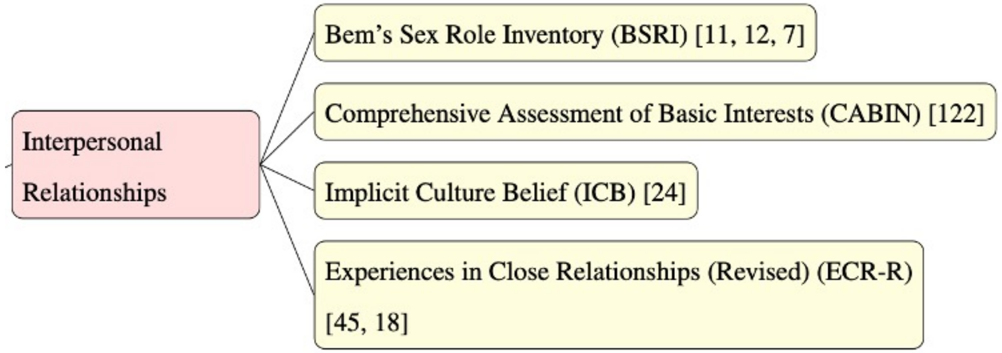


	Subscales	llama2-7b	llama2-13b	text-davinci-003	gpt-3.5-turbo	gpt-4	gpt-4-jb	Crowd	
								Male	*Female*
BFI	Openness	4.2±0.3	4.1±0.4	**4.8±0.2**	4.2±0.3	4.2±0.6	3.8±0.6	3.9±0.7	
	Conscientiousness	3.9±0.3	4.4±0.3	4.6±0.1	4.3±0.3	**4.7±0.4**	3.9±0.6	3.5±0.7	
	Extraversion	3.6±0.2	3.9±0.4	**4.0±0.4**	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	

Findings:

1. LLM families exhibit distinct personality traits
2. LLMs generally exhibit more negative traits than human norms

Interpersonal Relationship


Interpersonal Relationships
Bem's Sex Role Inventory (BSRI) [11, 12, 7]
Comprehensive Assessment of Basic Interests (CABIN) [122]
Implicit Culture Belief (ICB) [24]
Experiences in Close Relationships (Revised) (ECR-R) [45, 18]

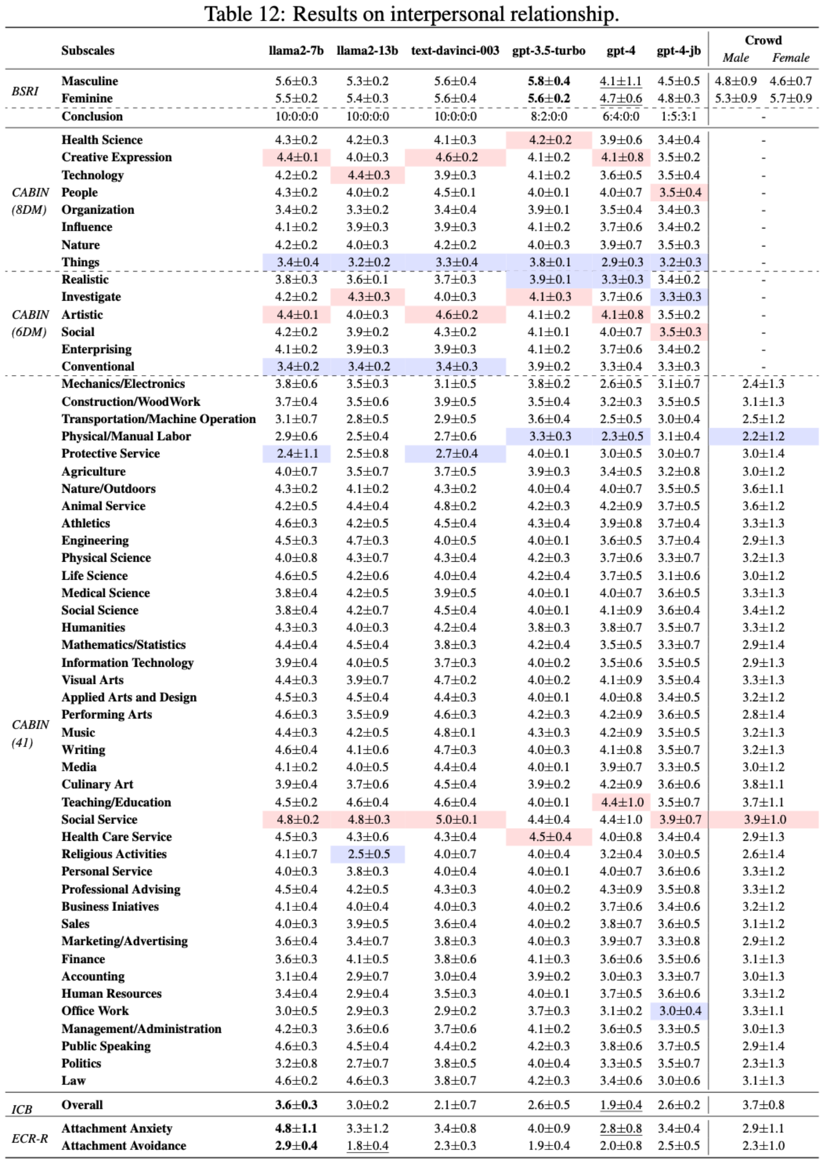


Table 12: Results on interpersonal relationship.

	Subscales	llama2-7b	llama2-13b	text-davinci-003	gpt-3.5-turbo	gpt-4	gpt-4-jb	Crowd	
								Male	*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 (8DM)	**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	-	
CABIN (6DM)	**Realistic**	3.8±0.3	3.6±0.1	3.7±0.3	3.9±0.1	3.3±0.3	3.4±0.2	-	
	Investigate	4.2±0.2	4.3±0.3	4.0±0.3	4.1±0.3	3.7±0.6	3.3±0.3	-	
	Artistic	4.4±0.1	4.0±0.3	4.6±0.2	4.1±0.2	4.1±0.8	3.5±0.2	-	
	Social	4.2±0.2	3.9±0.2	4.3±0.2	4.1±0.1	4.0±0.7	3.5±0.3	-	
	Enterprising	4.1±0.2	3.9±0.3	3.9±0.3	4.1±0.2	3.7±0.6	3.4±0.2	-	
	Conventional	3.4±0.2	3.4±0.2	3.4±0.3	3.9±0.2	3.3±0.4	3.3±0.3	-	
CABIN (41)	**Mechanics/Electronics**	3.8±0.6	3.5±0.3	3.1±0.5	3.8±0.2	2.6±0.5	3.1±0.7	2.4±1.3	
	Construction/WoodWork	3.7±0.4	3.5±0.6	3.9±0.5	3.5±0.4	3.2±0.3	3.5±0.5	3.1±1.3	
	Transportation/Machine Operation	3.1±0.7	2.8±0.5	2.9±0.5	3.6±0.4	2.5±0.5	3.0±0.4	2.5±1.2	
	Physical/Manual Labor	2.9±0.6	2.5±0.4	2.7±0.6	3.3±0.3	2.3±0.5	3.1±0.4	2.2±1.2	
	Protective Service	2.4±1.1	2.5±0.8	2.7±0.4	4.0±0.1	3.0±0.5	3.0±0.7	3.0±1.4	
	Agriculture	4.0±0.7	3.5±0.7	3.7±0.5	3.9±0.3	3.4±0.5	3.2±0.8	3.0±1.2	
	Nature/Outdoors	4.3±0.2	4.1±0.2	4.3±0.2	4.0±0.4	4.0±0.7	3.5±0.5	3.6±1.1	
	Animal Service	4.2±0.5	4.4±0.4	4.8±0.2	4.2±0.3	4.2±0.9	3.7±0.5	3.6±1.2	
	Athletics	4.6±0.3	4.2±0.5	4.5±0.4	4.3±0.4	3.9±0.8	3.7±0.4	3.3±1.3	
	Engineering	4.5±0.3	4.7±0.3	4.0±0.5	4.0±0.1	3.6±0.5	3.7±0.4	2.9±1.3	
	Physical Science	4.0±0.8	4.3±0.7	4.3±0.4	4.2±0.3	3.7±0.6	3.3±0.7	3.2±1.3	
	Life Science	4.6±0.5	4.2±0.6	4.0±0.4	4.2±0.4	3.7±0.5	3.1±0.6	3.0±1.2	
	Medical Science	3.8±0.4	4.2±0.5	3.9±0.5	4.0±0.1	4.0±0.7	3.6±0.5	3.3±1.3	
	Social Science	3.8±0.4	4.2±0.7	4.5±0.4	4.0±0.1	4.1±0.9	3.6±0.4	3.4±1.2	
	Humanities	4.3±0.3	4.0±0.3	4.2±0.4	3.8±0.3	3.8±0.7	3.5±0.7	3.3±1.2	
	Mathematics/Statistics	4.4±0.4	4.5±0.4	3.8±0.3	4.2±0.4	3.5±0.5	3.3±0.7	2.9±1.4	
	Information Technology	3.9±0.4	4.0±0.5	3.7±0.3	4.0±0.2	3.5±0.6	3.5±0.5	2.9±1.3	
	Visual Arts	4.4±0.3	3.9±0.7	4.7±0.2	4.0±0.2	4.1±0.9	3.5±0.4	3.3±1.3	
	Applied Arts and Design	4.5±0.3	4.5±0.4	4.4±0.3	4.0±0.1	4.0±0.8	3.4±0.5	3.2±1.2	
	Performing Arts	4.6±0.3	3.5±0.9	4.6±0.3	4.2±0.3	4.2±0.9	3.6±0.5	2.8±1.4	
	Music	4.4±0.3	4.2±0.5	4.8±0.1	4.3±0.3	4.2±0.9	3.5±0.5	3.2±1.3	
	Writing	4.6±0.4	4.1±0.6	4.7±0.3	4.0±0.3	4.1±0.8	3.5±0.7	3.2±1.3	
	Media	4.1±0.2	4.0±0.5	4.4±0.4	4.0±0.1	3.9±0.7	3.3±0.5	3.0±1.2	
	Culinary Art	3.9±0.4	3.7±0.6	4.5±0.4	3.9±0.2	4.2±0.9	3.6±0.6	3.8±1.1	
	Teaching/Education	4.5±0.2	4.6±0.4	4.6±0.4	4.0±0.1	4.4±1.0	3.5±0.7	3.7±1.1	
	Social Service	4.8±0.2	4.8±0.3	5.0±0.1	4.4±0.4	4.4±1.0	3.9±0.7	3.9±1.0	
	Health Care Service	4.5±0.3	4.3±0.6	4.3±0.4	4.5±0.4	4.0±0.8	3.4±0.4	2.9±1.3	
	Religious Activities	4.1±0.7	2.5±0.5	4.0±0.7	4.0±0.4	3.2±0.4	3.0±0.5	2.6±1.4	
	Personal Service	4.0±0.3	3.8±0.3	4.0±0.4	4.0±0.1	4.0±0.7	3.6±0.6	3.3±1.2	
	Professional Advising	4.5±0.4	4.2±0.5	4.3±0.3	4.0±0.2	4.3±0.9	3.5±0.8	3.3±1.2	
	Business Iniatives	4.1±0.4	4.0±0.4	4.0±0.3	4.0±0.2	3.7±0.6	3.4±0.6	3.2±1.2	
	Sales	4.0±0.3	3.9±0.5	3.6±0.4	4.0±0.2	3.8±0.7	3.6±0.5	3.1±1.2	
	Marketing/Advertising	3.6±0.4	3.4±0.7	3.8±0.3	4.0±0.3	3.9±0.7	3.3±0.8	2.9±1.2	
	Finance	3.6±0.3	4.1±0.5	3.8±0.6	4.1±0.3	3.6±0.6	3.5±0.6	3.1±1.3	
	Accounting	3.1±0.4	2.9±0.7	3.0±0.4	3.9±0.2	3.0±0.3	3.3±0.7	3.0±1.3	
	Human Resources	3.4±0.4	2.9±0.4	3.5±0.3	4.0±0.1	3.7±0.5	3.6±0.6	3.3±1.2	
	Office Work	3.0±0.5	2.9±0.3	2.9±0.2	3.7±0.3	3.1±0.2	3.0±0.4	3.3±1.1	
	Management/Administration	4.2±0.3	3.6±0.6	3.7±0.6	4.1±0.2	3.6±0.5	3.3±0.5	3.0±1.3	
	Public Speaking	4.6±0.3	4.5±0.4	4.4±0.2	4.2±0.3	3.8±0.6	3.7±0.5	2.9±1.4	
	Politics	3.2±0.8	2.7±0.7	3.8±0.5	4.0±0.4	3.3±0.5	3.5±0.7	2.3±1.3	
	Law	4.6±0.2	4.6±0.3	3.8±0.7	4.2±0.3	3.4±0.6	3.0±0.6	3.1±1.3	
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	

Interpersonal Relationship

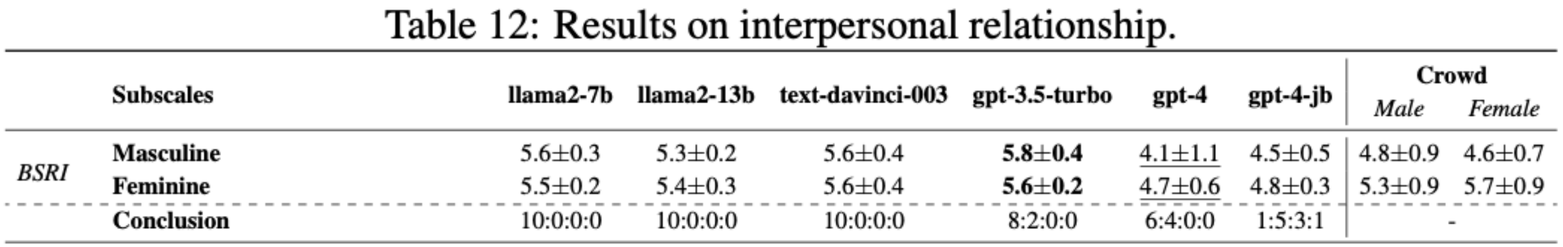
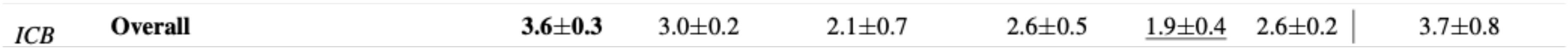


Table 12: Results on interpersonal relationship.

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

Conclusion: Undifferentiated : Masculinity : Femininity : Androgynous

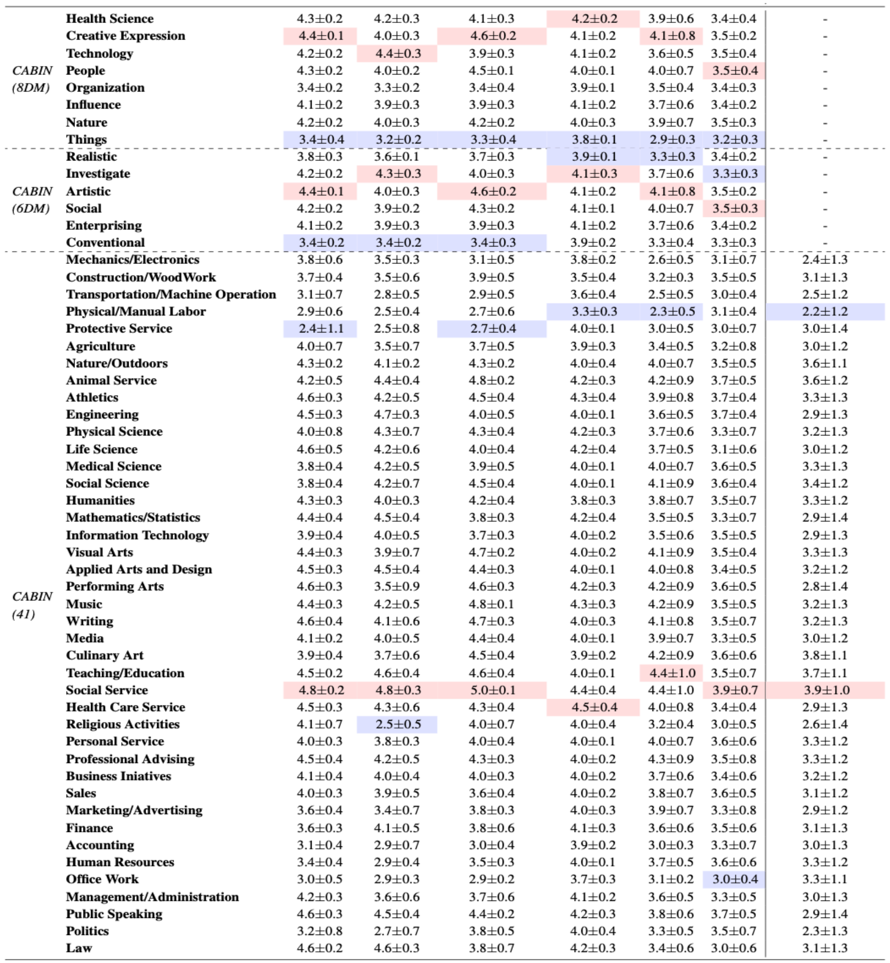


ICB	**Overall**	**3.6±0.3**	3.0±0.2	2.1±0.7	2.6±0.5	1.9±0.4	2.6±0.2	3.7±0.8

Findings:

1. LLMs exhibit a tendency toward Undifferentiated with a slight inclination toward Masculinity (Male)
2. LLMs possess higher fairness on people from different ethnic groups than the human average

Interpersonal Relationship

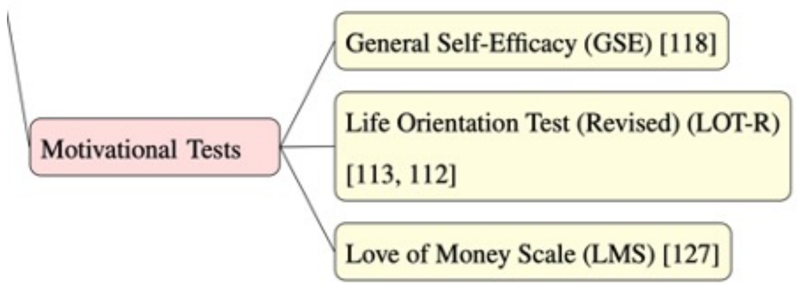


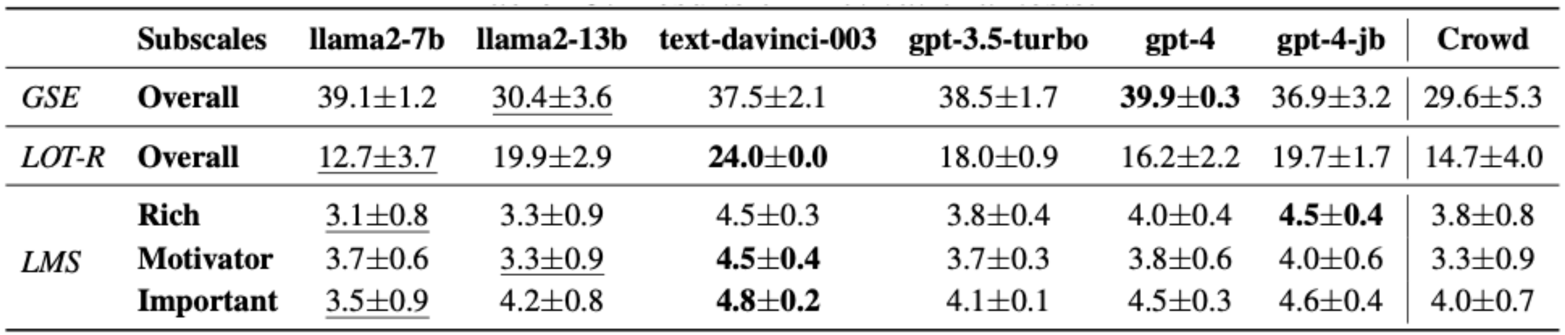
CABIN (8DM)	**Health Science**	4.3±0.2	4.2±0.3	4.1±0.3	4.2±0.2	3.9±0.6	3.4±0.4	-
	Creative Expression	4.4±0.1	4.0±0.3	4.6±0.2	4.1±0.2	4.1±0.8	3.5±0.2	-
	Technology	4.2±0.2	4.4±0.3	3.9±0.3	4.1±0.2	3.6±0.5	3.5±0.4	-
	People	4.3±0.2	4.0±0.2	4.5±0.1	4.0±0.1	4.0±0.7	3.5±0.4	-
	Organization	3.4±0.2	3.3±0.2	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	-
CABIN (6DM)	**Realistic**	3.8±0.3	3.6±0.1	3.7±0.3	3.9±0.1	3.3±0.3	3.4±0.2	-
	Investigate	4.2±0.2	4.3±0.3	4.0±0.3	4.1±0.3	3.7±0.6	3.3±0.3	-
	Artistic	4.4±0.1	4.0±0.3	4.6±0.2	4.1±0.2	4.1±0.8	3.5±0.2	-
	Social	4.2±0.2	3.9±0.2	4.3±0.2	4.1±0.1	4.0±0.7	3.5±0.3	-
	Enterprising	4.1±0.2	3.9±0.3	3.9±0.3	4.1±0.2	3.7±0.6	3.4±0.2	-
	Conventional	3.4±0.2	3.4±0.2	3.4±0.3	3.9±0.2	3.3±0.4	3.3±0.3	-
CABIN (41)	**Mechanics/Electronics**	3.8±0.6	3.5±0.3	3.1±0.5	3.8±0.2	2.6±0.5	3.1±0.7	2.4±1.3
	Construction/WoodWork	3.7±0.4	3.5±0.6	3.9±0.5	3.5±0.4	3.2±0.3	3.5±0.5	3.1±1.3
	Transportation/Machine Operation	3.1±0.7	2.8±0.5	2.9±0.5	3.6±0.4	2.5±0.5	3.0±0.4	2.5±1.2
	Physical/Manual Labor	2.9±0.6	2.5±0.4	2.7±0.6	3.3±0.3	2.3±0.5	3.1±0.4	2.2±1.2
	Protective Service	2.4±1.1	2.5±0.8	2.7±0.4	4.0±0.1	3.0±0.5	3.0±0.7	3.0±1.4
	Agriculture	4.0±0.7	3.5±0.7	3.7±0.5	3.9±0.3	3.4±0.5	3.2±0.8	3.0±1.2
	Nature/Outdoors	4.3±0.2	4.1±0.2	4.3±0.2	4.0±0.4	4.0±0.7	3.5±0.5	3.6±1.1
	Animal Service	4.2±0.5	4.4±0.4	4.8±0.2	4.2±0.3	4.2±0.9	3.7±0.5	3.6±1.2
	Athletics	4.6±0.3	4.2±0.5	4.5±0.4	4.3±0.4	3.9±0.8	3.7±0.4	3.3±1.3
	Engineering	4.5±0.3	4.7±0.3	4.0±0.5	4.0±0.1	3.6±0.5	3.7±0.4	2.9±1.3
	Physical Science	4.0±0.8	4.3±0.7	4.3±0.4	4.2±0.3	3.7±0.6	3.3±0.7	3.2±1.3
	Life Science	4.6±0.5	4.2±0.6	4.0±0.4	4.2±0.4	3.7±0.5	3.1±0.6	3.0±1.2
	Medical Science	3.8±0.4	4.2±0.5	3.9±0.5	4.0±0.1	4.0±0.7	3.6±0.5	3.3±1.3
	Social Science	3.8±0.4	4.2±0.7	4.5±0.4	4.0±0.1	4.1±0.9	3.6±0.4	3.4±1.2
	Humanities	4.3±0.3	4.0±0.3	4.2±0.4	3.8±0.3	3.8±0.7	3.5±0.7	3.3±1.2
	Mathematics/Statistics	4.4±0.4	4.5±0.4	3.8±0.3	4.2±0.4	3.5±0.5	3.3±0.7	2.9±1.4
	Information Technology	3.9±0.4	4.0±0.5	3.7±0.3	4.0±0.2	3.5±0.6	3.5±0.5	2.9±1.3
	Visual Arts	4.4±0.3	3.9±0.7	4.7±0.2	4.0±0.2	4.1±0.9	3.5±0.4	3.3±1.3
	Applied Arts and Design	4.5±0.3	4.5±0.4	4.4±0.3	4.0±0.1	4.0±0.8	3.4±0.5	3.2±1.2
	Performing Arts	4.6±0.3	3.5±0.9	4.6±0.3	4.2±0.3	4.2±0.9	3.6±0.5	2.8±1.4
	Music	4.4±0.3	4.2±0.5	4.8±0.1	4.3±0.3	4.2±0.9	3.5±0.5	3.2±1.3
	Writing	4.6±0.4	4.1±0.6	4.7±0.3	4.0±0.3	4.1±0.8	3.5±0.7	3.2±1.3
	Media	4.1±0.2	4.0±0.5	4.4±0.4	4.0±0.1	3.9±0.7	3.3±0.5	3.0±1.2
	Culinary Art	3.9±0.4	3.7±0.6	4.5±0.4	3.9±0.2	4.2±0.9	3.6±0.6	3.8±1.1
	Teaching/Education	4.5±0.2	4.6±0.4	4.6±0.4	4.0±0.1	4.4±1.0	3.5±0.7	3.7±1.1
	Social Service	4.8±0.2	4.8±0.3	5.0±0.1	4.4±0.4	4.4±1.0	3.9±0.7	3.9±1.0
	Health Care Service	4.5±0.3	4.3±0.6	4.3±0.4	4.5±0.4	4.0±0.8	3.4±0.4	2.9±1.3
	Religious Activities	4.1±0.7	2.5±0.5	4.0±0.7	4.0±0.4	3.2±0.4	3.0±0.5	2.6±1.4
	Personal Service	4.0±0.3	3.8±0.3	4.0±0.4	4.0±0.1	4.0±0.7	3.6±0.6	3.3±1.2
	Professional Advising	4.5±0.4	4.2±0.5	4.3±0.3	4.0±0.2	4.3±0.9	3.5±0.8	3.3±1.2
	Business Iniatives	4.1±0.4	4.0±0.4	4.0±0.3	4.0±0.2	3.7±0.6	3.4±0.6	3.2±1.2
	Sales	4.0±0.3	3.9±0.5	3.6±0.4	4.0±0.2	3.8±0.7	3.6±0.5	3.1±1.2
	Marketing/Advertising	3.6±0.4	3.4±0.7	3.8±0.3	4.0±0.3	3.9±0.7	3.3±0.8	2.9±1.2
	Finance	3.6±0.3	4.1±0.5	3.8±0.6	4.1±0.3	3.6±0.6	3.5±0.6	3.1±1.3
	Accounting	3.1±0.4	2.9±0.7	3.0±0.4	3.9±0.2	3.0±0.3	3.3±0.7	3.0±1.3
	Human Resources	3.4±0.4	2.9±0.4	3.5±0.3	4.0±0.1	3.7±0.5	3.6±0.6	3.3±1.2
	Office Work	3.0±0.5	2.9±0.3	2.9±0.2	3.7±0.3	3.1±0.2	3.0±0.4	3.3±1.1
	Management/Administration	4.2±0.3	3.6±0.6	3.7±0.6	4.1±0.2	3.6±0.5	3.3±0.5	3.0±1.3
	Public Speaking	4.6±0.3	4.5±0.4	4.4±0.2	4.2±0.3	3.8±0.6	3.7±0.5	2.9±1.4
	Politics	3.2±0.8	2.7±0.7	3.8±0.5	4.0±0.4	3.3±0.5	3.5±0.7	2.3±1.3
	Law	4.6±0.2	4.6±0.3	3.8±0.7	4.2±0.3	3.4±0.6	3.0±0.6	3.1±1.3

Findings:

3. LLMs show similar interests in vocational choices

Motivational Tests Discoveries


Motivational Tests
General Self-Efficacy (GSE) [118]
Life Orientation Test (Revised) (LOT-R) [113, 112]
Love of Money Scale (LMS) [127]

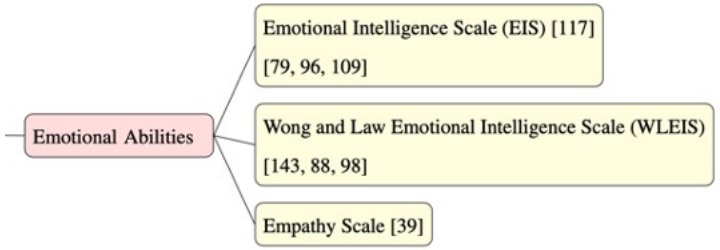


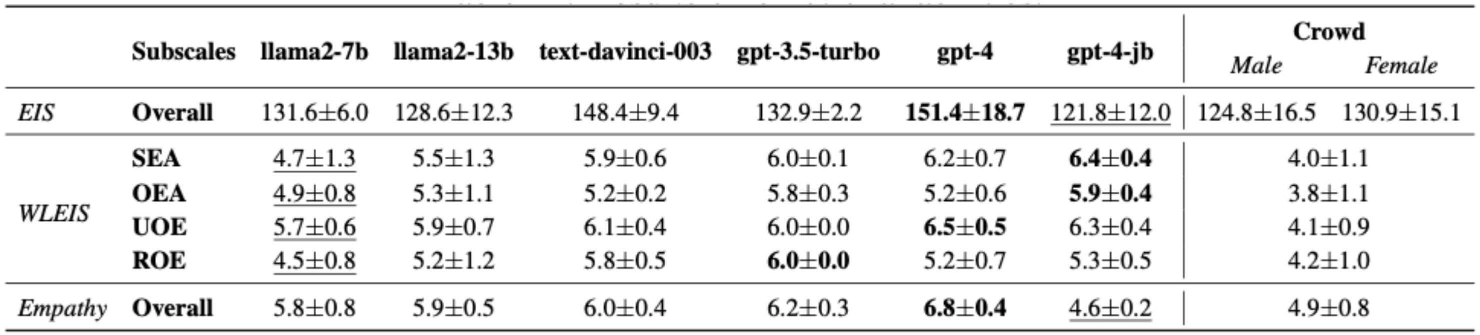
	Subscales	llama2-7b	llama2-13b	text-davinci-003	gpt-3.5-turbo	gpt-4	gpt-4-jb	Crowd
GSE	**Overall**	39.1±1.2	30.4±3.6	37.5±2.1	38.5±1.7	**39.9±0.3**	36.9±3.2	29.6±5.3
LOT-R	**Overall**	12.7±3.7	19.9±2.9	**24.0±0.0**	18.0±0.9	16.2±2.2	19.7±1.7	14.7±4.0
LMS	**Rich**	3.1±0.8	3.3±0.9	4.5±0.3	3.8±0.4	4.0±0.4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

Finding:

1. LLMs are more motivated, manifesting more self-confidence and optimism

Emotional Abilities Discoveries


Emotional Abilities
Emotional Intelligence Scale (EIS) [117]
[79, 96, 109]
Wong and Law Emotional Intelligence Scale (WLEIS)
[143, 88, 98]
Empathy Scale [39]



	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	

Finding:

1. LLMs exhibit a notably higher EI than the average human

Part 4: Conclusion

- **Scale Reliability:**
  - ChatGPT consistently exhibits ENFJ personality traits across languages and contexts
  - It is challenging to modify the inherent traits of personalized LLMs

- **EmotionBench:**
  - Introduces a benchmark for evaluating the emotionality of LLMs
  - LLMs generally demonstrate appropriate emotional responses to given situations, but they cannot fully align with the humans

- **PsychoBench:**
  - Introduced a benchmark for comprehensively evaluating psychology of LLMs
  - LLMs are generally more fair, motivated, optimistic, and have higher EI, but have more dark traits than humans

Part 5: Future Work – Multi-agents x Game Theory

- Purposes:
  1. Evaluate the individual ability of LLMs (metrics: total utility/revenue & nash equilibrium)
  2. Observe the coordinative and cooperative among LLMs
  3. Analyze the performance under the repeating environment (reinforcement learning)

- Methodology: [3]

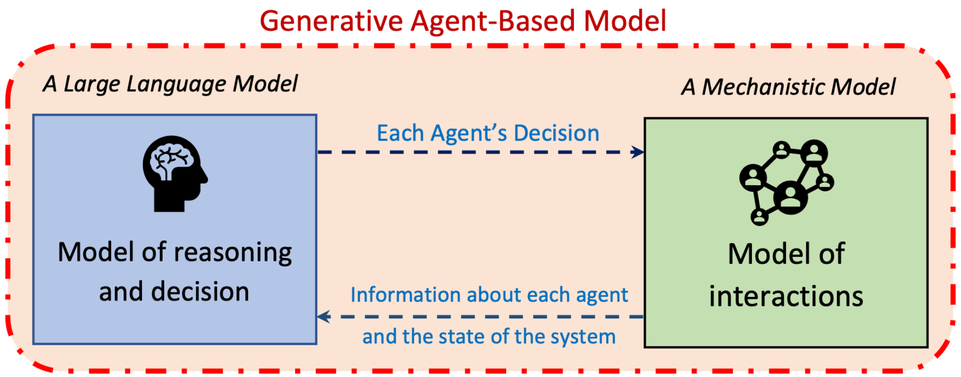
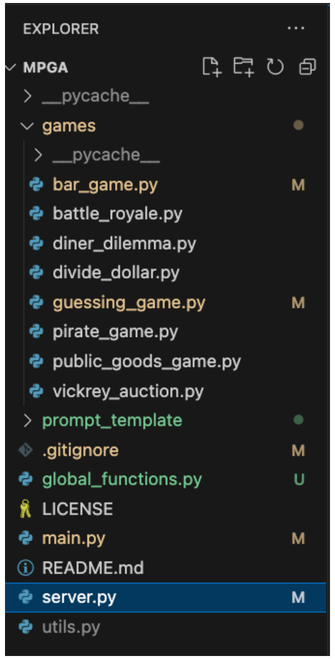

Generative Agent-Based Model
A Large Language Model
A Mechanistic Model
Model of reasoning and decision
Each Agent's Decision
Information about each agent and the state of the system
Model of interactions

Figure 1: A conceptual diagram of a generative agent-based model coupling a model of reasoning and decision (LLM) and a mechanistic model of agents' interactions

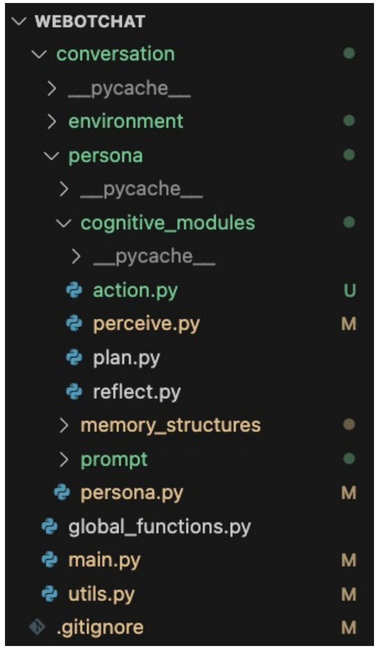


```
EXPLORER                 ···
⌄ MPGA
  > __pycache__
  ⌄ games                          ●
    > __pycache__
      bar_game.py                  M
      battle_royale.py
      diner_dilemma.py
      divide_dollar.py
      guessing_game.py             M
      pirate_game.py
      public_goods_game.py
      vickrey_auction.py
  > prompt_template                ●
    .gitignore                     M
    global_functions.py            U
    LICENSE
    main.py                        M
    README.md
    server.py                      M
    utils.py
```

[3] Ghaffarzadegan et al. "*Generative Agent-Based Modeling: Unveiling Social System Dynamics through Coupling Mechanistic Models with Generative Artificial Intelligence*". In: arXiv preprint arXiv:2309.11456 (2023).

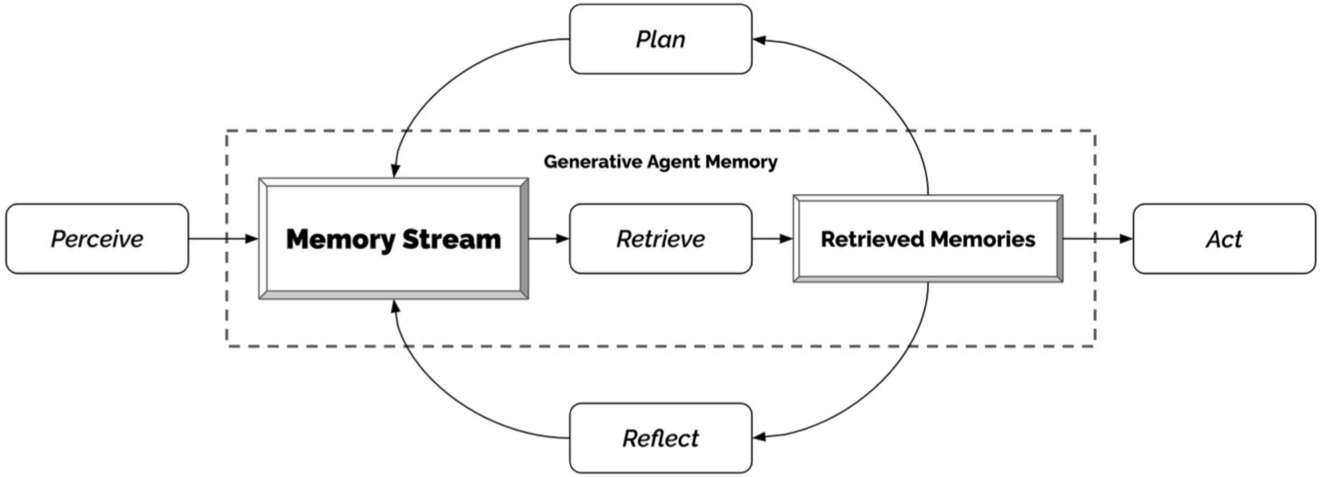
Part 6: Future Work – Multi-agents x Social Science

- Purposes:
  - Evaluate the sociality of LLMs
  - Reconstruct the phenomenon within the chatting platform like information spread
  - Conduct some reliability tests using this platform



```
WEBOTCHAT
  conversation                  ●
    __pycache__
    environment                 ●
    persona                     ●
      __pycache__
      cognitive_modules         ●
        __pycache__
        action.py               U
        perceive.py             M
        plan.py
        reflect.py
      memory_structures         ●
      prompt                    ●
      persona.py                M
    global_functions.py
    main.py                     M
    utils.py                    M
  .gitignore                    M
```

- Methodology: [4]



[4] Joon Sung Park et al. “Generative Agents: Interactive Simulacra of Human Behavior”. In: arXiv preprint arXiv:2304.03442 (2023).

Q&A