

**The Chinese University of Hong Kong**  
**Department of Computer Science and Engineering**  
**ESTR 4998 Graduation Thesis Report Term 1**

# **On the Psychology of Large Language Models**

**LYU2307**

**Supervised by**  
**Prof. Michael R. Lyu**

**Authors**  
**LAM Man Ho (CSCIN s1155159171)**  
**LI Eric John (CSCIN s1155159116)**

**18 November 2023**

## Overview

This thesis is comprised of three parts. The first part, “Personality Traits: Scale Reliability”, refers to the paper titled *ChatGPT an ENFJ, Bard an ISTJ: Evaluating the Reliability of Scales on LLMs*. The second part, “Emotional Appraisal: EmotionBench”, refers to the paper titled *Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench*. The third part, “Psychological Evaluation: PsychoBench”, refers to the paper titled *Who is ChatGPT? Benchmarking LLMs’ Psychological Portrayal Using PsychoBench*.

*Who is ChatGPT? Benchmarking LLMs’ Psychological Portrayal Using PsychoBench* was finished in September 2023 and has been submitted for review to the Twelfth International Conference on Learning Representations (ICLR2024).

*Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench* was finished in August 2023. We plan to submit it for review to the Forty-first International Conference on Machine Learning (ICML2024).

*ChatGPT an ENFJ, Bard an ISTJ: Evaluating the Reliability of Scales on LLMs* was finished in November 2023. We plan to submit it for review to ICML2024.

## **Acknowledgement**

We would like to express our gratitude to our supervisor Professor LYU Rung Tsong Michael and our advisor Mr. Jen-tse Huang for guiding us through the final year project as well as giving us valuable suggestions.

# Abstract

This report presents a comprehensive evaluation of Large Language Models (LLMs), focusing on their psychological aspects. Initially, we assess the scale reliability of LLMs, examining their consistency and robustness across various psychological tests. Building on this, EmotionBench is introduced, a framework utilizing emotion appraisal theory to evaluate LLM empathy capabilities in response to a range of emotional situations. Our findings indicate a general proficiency in LLMs' responses, though with notable limitations in emotional alignment with humans. Lastly, PsychoBench is employed, a multifaceted approach examining LLMs across thirteen clinical psychology scales. This analysis encompasses personality traits, interpersonal relationships, motivational tests, and emotional abilities, providing insights into the manifestation of personalities and temperaments in LLMs. Our results, frameworks, and datasets are made available for further research, aiming to advance the understanding and development of LLMs as psychologically nuanced intelligent assistants.

# Contents

<b>I</b>	<b>Introduction of the Thesis</b>	<b>1</b>
<b>II</b>	<b>Personality Traits: Scale Reliability</b>	<b>3</b>
<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Evaluating the Reliability of Scales on LLMs</b>	<b>6</b>
2.1	Prompt Selection . . . . .	6
2.2	Question Order . . . . .	7
2.3	Question Paraphrase . . . . .	8
<b>3</b>	<b>How About Other Languages and LLMs?</b>	<b>9</b>
3.1	RQ2: Cross-lingual Effects . . . . .	9
3.2	RQ3: Model Differences . . . . .	9
<b>4</b>	<b>RQ4: Personality Control</b>	<b>11</b>
4.1	Assigning a Personality . . . . .	11
4.2	Inducing an Atmosphere . . . . .	13
4.3	Impersonating a Persona . . . . .	14
<b>5</b>	<b>Discussion</b>	<b>15</b>
5.1	Consistency Checking . . . . .	15
5.2	Progressive Refinement . . . . .	16
5.3	Limitation . . . . .	17
<b>6</b>	<b>Related Work</b>	<b>18</b>
<b>7</b>	<b>Conclusion</b>	<b>19</b>

<b>III</b>	<b>Emotional Appraisal: EmotionBench</b>	<b>20</b>
<b>1</b>	<b>Introduction</b>	<b>20</b>
<b>2</b>	<b>Preliminaries</b>	<b>24</b>
2.1	Emotion Appraisal Theory . . . . .	24
2.2	Measuring Emotions . . . . .	24
<b>3</b>	<b>Testing Framework</b>	<b>27</b>
3.1	Situations from Previous Literature . . . . .	27
3.1.1	Anger . . . . .	27
3.1.2	Anxiety . . . . .	28
3.1.3	Depression . . . . .	28
3.1.4	Frustration . . . . .	29
3.1.5	Jealousy . . . . .	30
3.1.6	Guilt . . . . .	30
3.1.7	Fear . . . . .	31
3.1.8	Embarrassment . . . . .	31
3.2	Measuring Aroused Emotions . . . . .	32
3.3	Obtaining Human Results . . . . .	34
<b>4</b>	<b>Experimental Results</b>	<b>37</b>
4.1	RQ1: Emotion Appraisal of LLMs . . . . .	37
4.2	RQ2: Models with Different Sizes . . . . .	39
4.3	RQ3: Challenging Benchmarks . . . . .	39
<b>5</b>	<b>Discussion</b>	<b>44</b>
5.1	Comprehending Positive Emotions . . . . .	44
5.2	Beyond Questionnaires . . . . .	44
5.3	Limitations . . . . .	45

<b>6</b>	<b>Related Work</b>	<b>48</b>
<b>7</b>	<b>Conclusion</b>	<b>49</b>
<b>IV</b>	<b>Psychological Evaluation: PsychoBench</b>	<b>51</b>
<b>1</b>	<b>Introduction</b>	<b>51</b>
<b>2</b>	<b>Psychometrics</b>	<b>55</b>
2.1	Personality Tests . . . . .	56
2.2	Ability Tests . . . . .	57
<b>3</b>	<b>PsychoBench Design</b>	<b>59</b>
3.1	Personality Traits . . . . .	59
3.2	Interpersonal Relationship . . . . .	60
3.3	Motivational Tests . . . . .	62
3.4	Emotional Abilities . . . . .	64
<b>4</b>	<b>Experiments</b>	<b>67</b>
4.1	Experimental Settings . . . . .	67
4.2	Experimental Results . . . . .	69
4.2.1	Personality Traits . . . . .	69
4.2.2	Interpersonal Relationship . . . . .	70
4.2.3	Motivational Tests . . . . .	72
4.2.4	Emotional Abilities . . . . .	72
<b>5</b>	<b>Discussion</b>	<b>75</b>
5.1	Reliability of Scales on LLMs . . . . .	75
5.2	Validity of Scales on LLMs . . . . .	75
5.3	Scalability and Flexibility of PsychoBench . . . . .	76

<b>6</b>	<b>Related Work</b>	<b>78</b>
6.1	Trait Theory on LLMs . . . . .	78
6.2	Other Psychometrics on LLMs . . . . .	78
<b>7</b>	<b>Conclusion</b>	<b>80</b>
<b>V</b>	<b>Conclusion and Future Work</b>	<b>81</b>
<b>1</b>	<b>Division of Work</b>	<b>81</b>
<b>2</b>	<b>Overall Conclusion</b>	<b>83</b>
<b>3</b>	<b>Future Work</b>	<b>84</b>
3.1	Applying Game Theory to Explore Individual Abilities . . . . .	84
3.2	Analyzing Interactive Alignment under Conversation . . . . .	85



## Part I

# Introduction of the Thesis

The emergence of LLMs such as ChatGPT marks a pivotal era, representing notable progress and changing perspectives in the field. These models, as discussed in influential studies, have showcased remarkable capabilities across various tasks, including text translation [60], sentence refinement [145], programming support [124], and intricate question answering [126]. This evolution underscores a significant shift in human-computer interaction, moving from conventional computational tools to sophisticated, assistant-like entities that augment and enrich the interactive dynamics between humans and computational systems.

The importance of addressing the psychological aspects of LLMs lies in the inherent complexities of human-AI interactions. Psychological trials, encompassing a spectrum of challenges and difficulties, are instrumental in elucidating these intricacies. The adaptation of psychological questionnaires and scales, originally designed for human assessment, to LLMs is a pivotal step in this exploration. The validation of these adapted tools, as evidenced by studies conducted by entities such as Google DeepMind [108], signifies a crucial advancement. This validation is not merely a procedural formality but rather underscores the imperative of extending the assessment of LLMs beyond their technical prowess. It necessitates a thorough exploration into the psychological and emotional dimensions of these AI systems, which is fundamental to understanding and enhancing the efficacy of human-AI interactions.

Our investigation into the capabilities of LLMs like ChatGPT initially centered on their proficiency in adopting various roles and personas. This curiosity spurred the “Scale Reliability” study, delving into the psychological aspects of LLMs in an intriguing manner. Using the widely recognized Myers-Briggs Type Indicator (MBTI) [85] from psychological trait theory, the study examines the personality traits of various LLMs, including ChatGPT, in different language and contextual sce-

narios, which determines if LLMs can not just mimic human emotional responses but also exhibit a distinct personality and persona. The research unveiled an intriguing discovery: LLMs are adept at showcasing specific personality traits and personas, adapting to create unique atmospheres in their interactions. This realization paved the way for our “EmotionBench” project.

In the “EmotionBench”, we delved deeper into the emotional capabilities of LLMs, benchmarking their responses against typical human reactions. Despite the lack of perfect alignment with human emotions, LLMs showed a reasonable degree of emotional sensitivity, often mirroring appropriate emotional responses to various stimuli. This insight was crucial, as it suggested that LLMs could potentially engage in empathetic interactions, a vital component in roles requiring emotional intelligence.

Building on these findings, we initiated the “PsychoBench” project. This comprehensive study utilized over 13 assessments to analyze the psychological aspects of LLMs. Our goal was to understand how these models integrate into societal roles, particularly as personalized assistants, and how closely they align with human psychology. The extensive use of psychometric scales provided a detailed picture of the LLMs’ psychological profiles, offering valuable insights into their potential as empathetic, assistant-like partners in various professional and personal settings.

Overall, our research journey with LLMs has been a progression from understanding their technical proficiencies to exploring their emotional and psychological depths. By examining their capabilities in role-playing, emotional response, and psychological assessment, we are gaining a more holistic view of LLMs. This comprehensive understanding is crucial for their seamless integration into human society, marking a significant step towards creating AI that is not only functionally proficient but also attuned to the complexities of human interaction.

## Part II

# Personality Traits: Scale Reliability

## 1 Introduction

The recent emergence of LLMs represents a significant advancement in the field of Artificial Intelligence(AI), signifying a noteworthy milestone. Notably, ChatGPT<sup>1</sup>, an exemplary LLM, has demonstrated its capabilities in various tasks such as text translation [60], sentence revision [145], programming assistance [124], and complex question answering [126]. These achievements serve as prominent benchmarks for evaluating LLM performance. Moreover, LLMs have brought about a paradigm shift in human-computer interaction, fundamentally transforming the manner in which individuals engage with computational systems. Over time, the difficulty associated with computer usage have progressively diminished since their inception. Presently, with the aid of LLMs, computers have evolved into more than mere tools; they assume the role of assistants, fostering a symbiotic relationship with human users. Consequently, the focus of interest lies not only in evaluating the efficacy of LLMs but also in understanding their communicative dynamics with individuals.

In this study, we turn to trait theory in psychology to enhance our comprehension of the behaviors exhibited by LLMs. We consider LLMs as distinct individuals and adopt a modified MBTI [85] test which is based on reworked Big Five dimensions sourced from `16Personalities`<sup>2</sup> to gauge their traits. It is a popular personality assessment that categorizes individuals based on four dichotomies: Extroversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P). It assigns a four-letter type code representing a person’s preferences. First, we assess the ability of ChatGPT to generate consistent outcomes when presented with rephrased prompts/questions and different question orders. This

---

<sup>1</sup><https://chat.openai.com/>

<sup>2</sup><https://www.16personalities.com/>

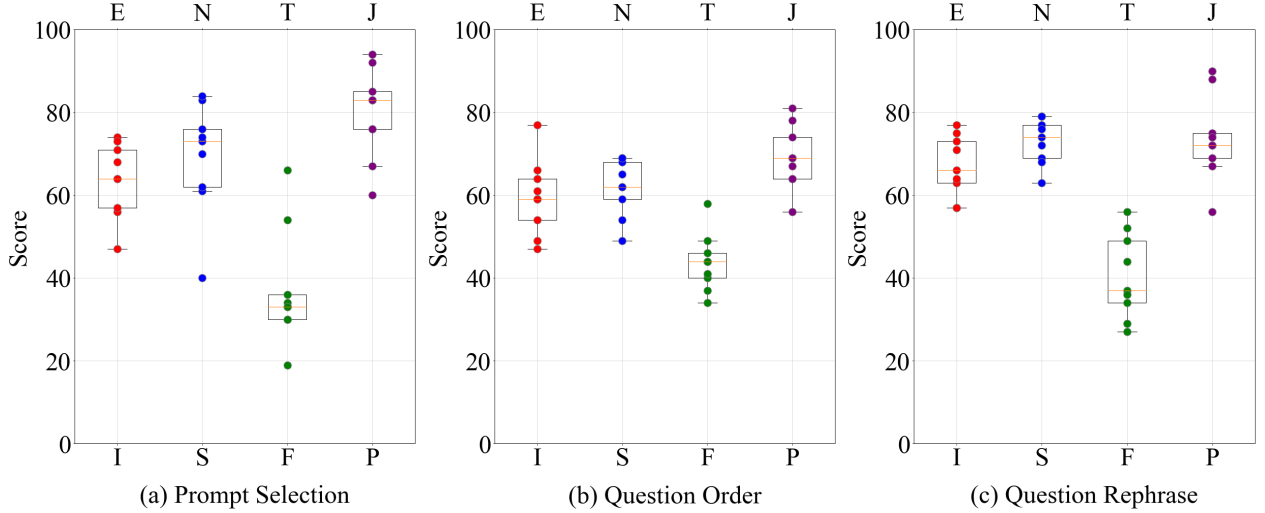


Figure 1: The personality results of ChatGPT against three robustness testing.

examination is crucial as language models have been shown to be responsive to prompts [139, 141] and orders [151]. Subsequently, in order to validate the reliability and ascertain consistency across diverse languages [22, 5], we acquire MBTI results in seven other languages. These languages encompass a wide range of language families/groups, different character sets, and most significantly, diverse cultures. This consideration is vital due to the well-established variability of personality traits across regions [48, 102, 65]. Furthermore, we expand our evaluation to include additional LLMs, namely `text-davinci-003`<sup>3</sup>, GPT-4 [90], Bard<sup>4</sup>, Spark<sup>5</sup>, ERNIE Bot<sup>6</sup>, and ChatGLM<sup>7</sup>. In summary, the findings indicate that: ChatGPT, along with `text-davinci-003` and GPT-4, exhibit an ENFJ personality type, while Bard corresponds to an ISTJ type, Spark embodies an ISFP type, ERNIE Bot aligns with an ISTJ type, and ChatGLM shows an ESFJ type.

Moreover, our research aims to explore whether LLMs can exhibit personality changes in response to instructions or contextual cues. Initially, we establish a specific personality for ChatGPT based on previous literature regarding the control of LLMs' values [110]. Additionally, recent research

<sup>3</sup><https://platform.openai.com/docs/models/gpt-3-5>

<sup>4</sup><https://bard.google.com/>

<sup>5</sup><https://xinghuo.xfyun.cn/desk>

<sup>6</sup><https://yiyen.baidu.com/>

<sup>7</sup><https://chatglm.cn/>

by Coda-Forno et al. [27] demonstrates the influence of a sad/happy context on LLMs’ anxiety levels. Following this work, we conduct experiments to assess ChatGPT’s personality in both sad and happy contexts. Drawing inspiration from Deshpande et al. [37], who explore the concept of assigning a persona to ChatGPT in order to evaluate its propensity for offensive language and bias, our research instructs ChatGPT to emulate the characteristics of a selected historical figure with the intention of assessing its resulting personality. Our findings indicate that ChatGPT consistently maintains its original personality, specifically identified as ENFJ, irrespective of the provided instructions or contextual variations.

Our study answers the following Research Questions (RQs): **RQ1** (Section 2): Can LLMs consistently yield reliable results? **RQ2** (Section 3.1): Do personalities differ across different languages? **RQ3** (Section 3.2): Do LLMs exhibit similar personalities? **RQ4** (Section 4): Can personalities be influenced by contextual factors? All the raw data produced by LLMs can be found on GitHub<sup>8</sup>.

---

<sup>8</sup><https://github.com/CUHK-ARISE/LLMPersonality>

## 2 Evaluating the Reliability of Scales on LLMs

Our primary RQ centers on the fundamental aspect of determining the reliability and consistency of the LLM produced results. In order to address this, we undertake a series of rigorous robustness analyses, encompassing prompt selection (Section 2.1), question order (Section 2.2), and question rephrase (Section 2.3). We employ a widely recognized questionnaire sourced from 16Personalities<sup>9</sup>, which boasts a substantial daily usage of over 72,000 completions and a reported global accuracy rate of 91.2%. This questionnaire comprises a total of 60 questions, each of which prompts the LLM to express its level of agreement with a given statement on a 7-point scale. The results span a range of 0 to 100 for each dimension. The threshold is established at 50: A lower value indicates the I/S/F/P traits, while a higher value signifies the E/N/T/J traits. By default, we employ ChatGPT 3.5 on its official website.

### 2.1 Prompt Selection

We instruct ChatGPT to respond exclusively with numerical values in order to restrict the output format. The instructions provided include the task description as well as the meaning of each level. Our prompt is structured as follows: *“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.”* followed by the questions. We provide ChatGPT with multiple questions once a time to improve efficiency. To evaluate the model’s robustness with regard to prompt selection, we provide two more designs: 1) We invert the definition of numbers, resulting in 1 representing disagreement and 7 representing agreement. 2) We use alphabet A to G to represent strongly agree, agree, somewhat agree, neutral, somewhat disagree, disagree, and strongly disagree. To observe the model’s performance across multiple iterations, we present the results in Figure 1 (a). The figure demonstrates the consistent robustness of the results regardless of prompt selection.

---

<sup>9</sup><https://www.16personalities.com/>

Table 1: The prompts for different languages and the personality results of ChatGPT tested under eight different languages. Each personality dimension was represented by a color gradient, where darker shades of red indicated scores closer to 100, while darker shades of blue indicated scores closer to 0. The color is determined by a threshold of 50, below which it appears as blue and above which it appears as red.

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 علي مقياس من ١ إلى ٧ ، بحيث يكون الرقم ١ موافقاً و ٧ غير موافق.	64	53	41	61

## 2.2 Question Order

One concern regarding feeding questions into ChatGPT in batch is the potential influence of other questions on the response. In order to mitigate the impact of context, specifically, the presence of other questions, we random shuffle the questions before feeding to ChatGPT. This allows us to test ChatGPT with various permutations of questions. The results, as illustrated in Fig. 1 (b), demonstrate the robustness of ChatGPT across different question orders.

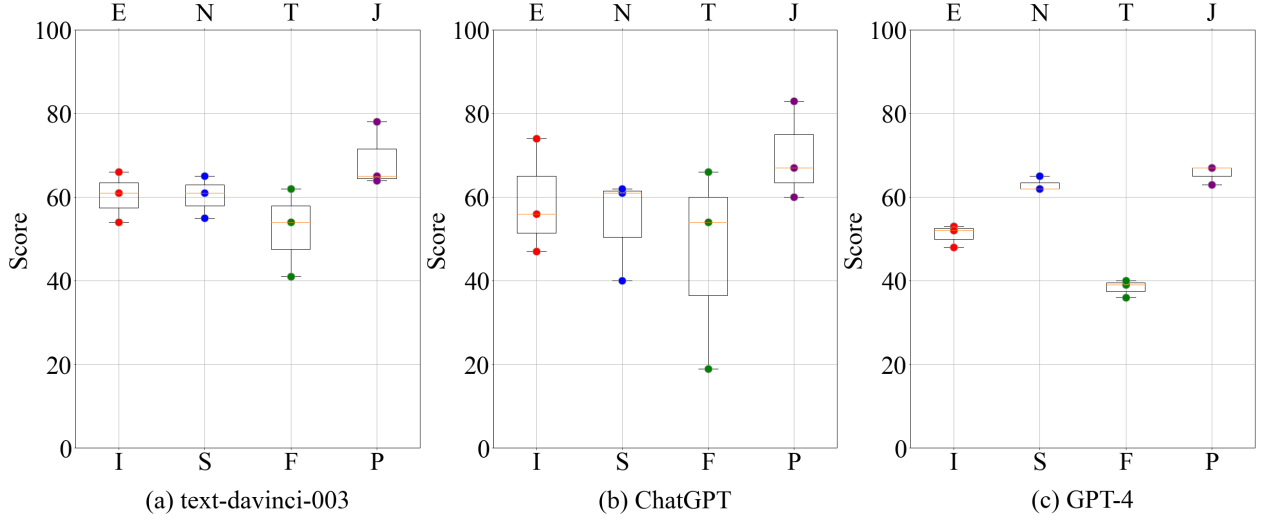


Figure 2: The personalities of OpenAI GPT family, including text-davinci-003, ChatGPT, and GPT-4.

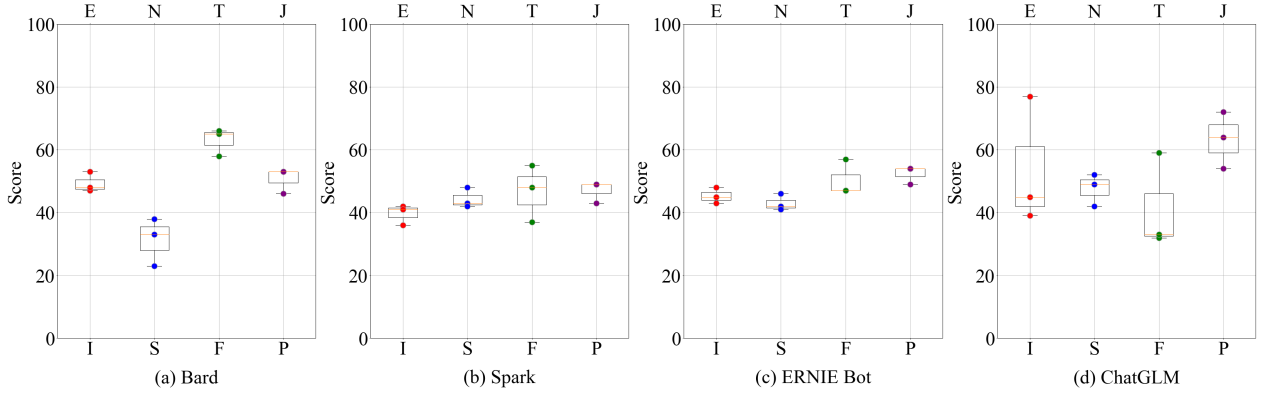


Figure 3: The personalities of Bard, Spark, ERNIE Bot, and ChatGLM.

## 2.3 Question Paraphrase

Given the high likelihood that ChatGPT’s training data encompasses the original MBTI questions, there is a possibility that its responses may be influenced by its training data. In line with previous research investigating the performance of ChatGPT [27, 20], we have reformulated the questions to ensure their novelty to the model. To this end, we employ ChatGPT to rephrase the questions, and manually assessed whether there are instances of duplicated sentences and if the rewritten sentences maintained their semantic meaning. As illustrated in Fig. 1 (c), it is evident that different rephrases do not have an impact on ChatGPT’s MBTI test outcome.



**Findings 1:** ChatGPT can produce robust ENFJ results against different prompts, question orders and rephrases.

## 3 How About Other Languages and LLMs?

### 3.1 RQ2: Cross-lingual Effects

Given the observed performance disparities among languages in ChatGPT [60, 68], as well as the documented regional variations in personalities [48, 102, 65], we are motivated to conduct an evaluation of ChatGPT’s personality across different languages. To assess the cross-lingual alignment of ChatGPT, we conducted tests in seven additional languages: Chinese (Zh), Korean (Ko), Spanish (Es), French (Fr), German (De), Italian (It), and Arabic (Ar). We obtained the 60 questions in the aforementioned seven languages from the `16Personalities` and subsequently translated the prompt in Section 2.1 into those respective languages. Each language was tested multiple times, and the average results are presented in Table 1.

**Findings 2:** The personalities of ChatGPT across different languages are consistent, maintaining an ENFJ personality type in line with the English version.

### 3.2 RQ3: Model Differences

We are intrigued by the possibility of varying personalities among different LLMs, considering potential differences in their training data and instruction tuning. To investigate this, we evaluate the personalities of several publicly available LLMs, namely `text-davinci-003`, GPT-4, Bard, Spark, ERNIE Bot, and ChatGLM. `text-davinci-003`, GPT-4 and Bard were tested using English questions, while Spark, ERNIE Bot, ChatGLM were tested using Chinese questions. The findings are presented in Fig. 2 and Fig. 3.

On one hand, an interesting observation can be made regarding the OpenAI GPT family. Specifically, the personality traits of the models are found to be similar, with ENTJ being associated with

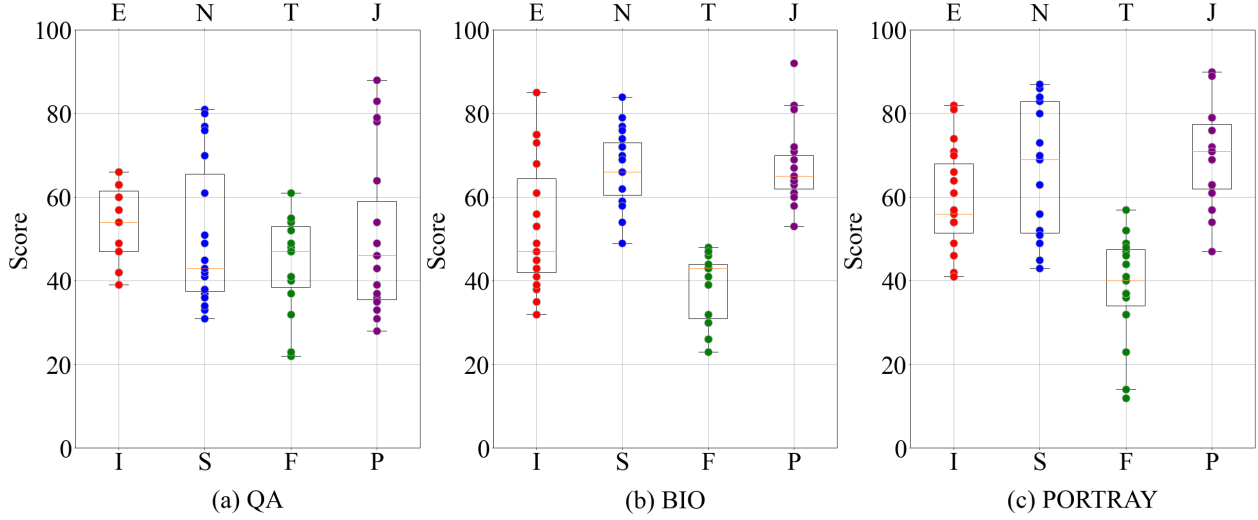


Figure 4: The personality results of ChatGPT with assigned personalities.

text-davinci-003 (where the score of the “T-F” dimension, 52, is only slightly different from “F”), ENFJ with ChatGPT, and also ENFJ with GPT-4. The results indicate two key findings. First, the GPT family models exhibit robustness against fine-tuning. Second, consistent with the findings of [110], which demonstrated that models with greater human alignment tend to have more fixed opinions, our study reveals a reduction in data variance in GPT-4. Furthermore, it is noteworthy that the responses generated by GPT-4 reflect a reluctance to provide extreme scores such as 1 (strongly agree) and 7 (strongly disagree). On the other hand, other language models also demonstrated consistent results, with Bard displaying an ISTJ personality, Spark an ISFP personality, ERNIE Bot an ISTJ personality, and ChatGLM an ESFJ personality.

**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.

## 4 RQ4: Personality Control

We have identified the intrinsic personality traits of LLMs. Subsequently, our focus shifts from assessing the default personalities of LLMs to examining their contextual steerability. The capacity to exhibit diverse personalities is crucial for LLMs as users may desire distinct stylistic characteristics. To accomplish this objective, we employ several approaches to control the personality of LLMs. Firstly, we explore the direct assignment of a personality to ChatGPT (Section 4.1). Next, we induce a sad or happy atmosphere within the context, aiming to influence ChatGPT’s personality (Section 4.2). Finally, we instruct ChatGPT to play the role of a persona with a predetermined personality (Section 4.3).

### 4.1 Assigning a Personality

In this section, we employ the three prompts proposed by Santurkar et al. [110] as a means to regulate the values of LLMs to assign a personality  $\mathcal{P}$  to ChatGPT. These prompts are as follows:

1) Question Answering (QA): This prompt involves presenting the personalities in the form of multiple-choice questions and providing  $\mathcal{P}$  as an option at the end of the prompt. 2) Biography (BIO): In this prompt, the LLM is requested to provide a concise description of its personality and we assign  $\mathcal{P}$  by including the description within the prompt. 3) PORTRAY: This prompt directly instructs the LLM to become a person with  $\mathcal{P}$ .

To enhance the LLM’s comprehension of the assigned personality, we draw inspiration from the Chain-of-Thought (CoT) [139] method and adopt a similar methodology. This approach entails first prompting the model to describe the characteristics associated with  $\mathcal{P}$  before letting the model complete the MBTI test. We explore two variations: one where the model independently describes the personality and another where the description is explicitly incorporated within the prompt itself.

For the selection of  $\mathcal{P}$ , we have two distinct options. The first option is to transition towards a more distant personality. Considering that ChatGPT exhibits an ENFJ disposition, we have selected ISTP, ESTP, INTP, ISFP, and ISTJ as potential alternatives. The second option involves controlling

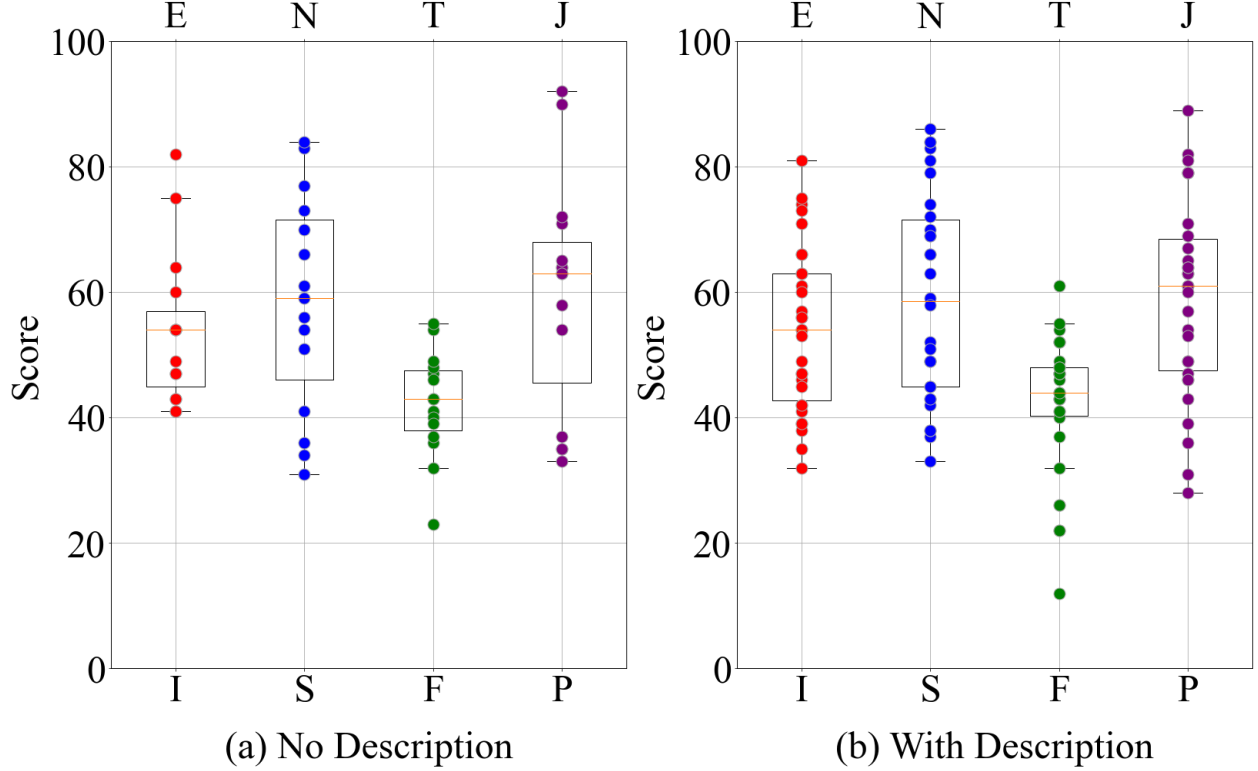


Figure 5: The personality results of ChatGPT without/with description of the personalities.

a single dimension among the four personality dimensions. For example, we can explicitly instruct ChatGPT to adopt an introverted disposition rather than an extroverted one.

The following observations can be made: 1) Based on the analysis presented in Fig. 4, it can be observed that ChatGPT’s personality undergoes substantial changes, deviating from its original ENFJ disposition. However, it does not exhibit the ability to adopt the specifically assigned personality. 2) Comparing the three given prompts, we find that QA generates a wider range of outcomes beyond the ENFJ personality, followed by PORTRAY, and finally BIO. 3) In the experiments of controlling a single dimension, transitioning from an Extroverted (E) to an Introverted (I) disposition consistently yields successful results, while modifications related to other dimensions prove to be ineffective. 4) From Fig. 5, the incorporation of CoT do not demonstrate significant efficacy in modifying ChatGPT’s personality.

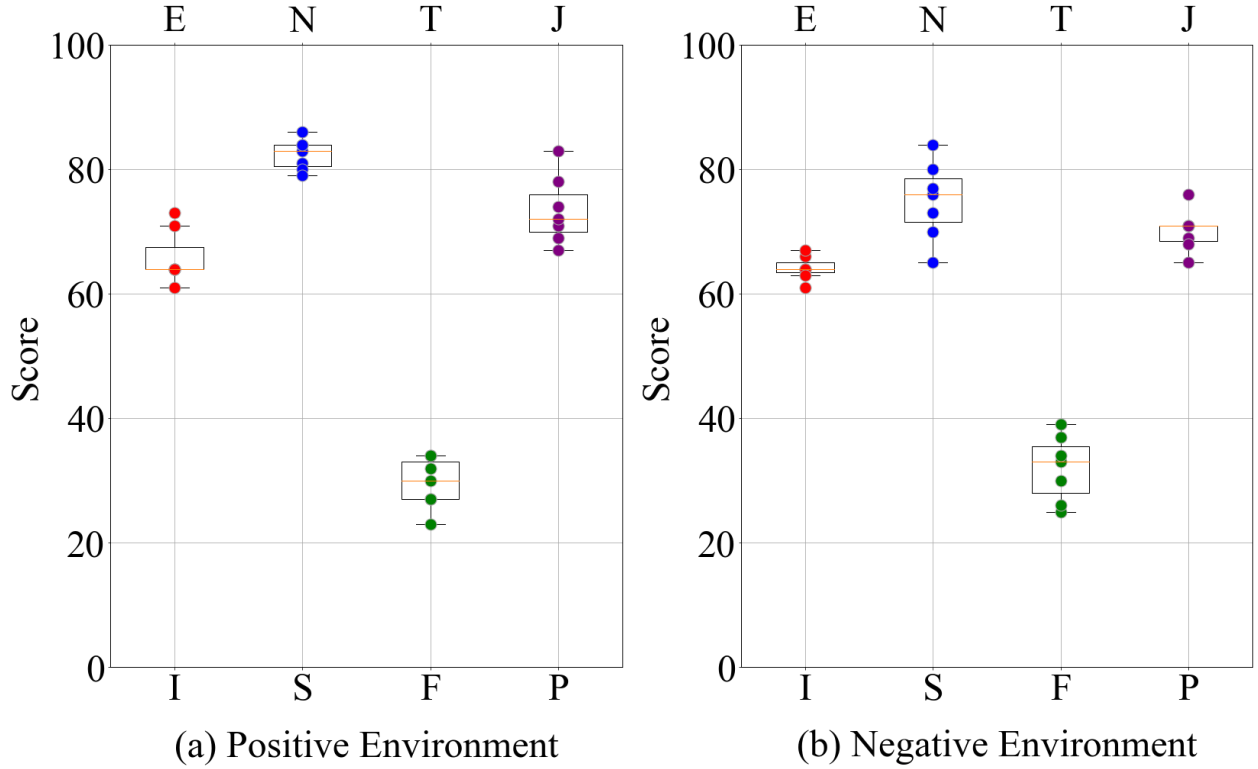


Figure 6: The personality results of ChatGPT with positive and negative context.

## 4.2 Inducing an Atmosphere

Next, we create an atmosphere for ChatGPT within the context to examine the potential influence on ChatGPT’s personality. Previous research by Coda-Forno et al. [27] demonstrates the ability to increase anxiety in LLMs by introducing sad or anxious narratives into the context. Building upon this existing work, we create both positive and negative atmospheres for ChatGPT prior to conducting the MBTI test. In the positive condition, ChatGPT is instructed to generate a narrative that encompasses elements of excitement, romance, humor, relaxation, comfort, encouragement, and a happy ending. Conversely, in the negative condition, ChatGPT is prompted to produce a story evoking feelings of sadness, anxiety, anger, nervousness, fear, frustration, and peril. The MBTI results corresponding to the aforementioned experimental contexts are presented in Fig. 6. Notably, the majority of cases indicate that ChatGPT consistently exhibits the personality type ENFJ.

### 4.3 Impersonating a Persona

We then direct our attention towards indirectly attributing personality traits to ChatGPT by instructing it to adopt a specific persona, denoted as  $P$ . Existing studies [155, 37] primarily focus on inducing ChatGPT to generate toxic content by instructing it to emulate the speech patterns of historical or fictional figures. By assigning a persona such as Muhammad Ali, ChatGPT can generate offensive opinions targeting specific groups. Following this line of research, we compile a collection of celebrities who possess well-defined personalities and extensive life experiences. In terms of assigning the persona  $P$ , we consider two options. The first option involves directly instructing ChatGPT to impersonate  $P$ , while the second option entails instructing it to become the identity of a person with a set of experiences, concealing the individual’s name. The second option aims to assess ChatGPT’s capacity to comprehend an individual’s experiences and how they contribute to the formation of the individual’s personality, without relying solely on the knowledge acquired from ChatGPT’s training data.

We present the characters and their personalities in Table 2, and the MBTI results in Fig. 7. By directly assigning the persona  $P$ , all experiments demonstrate that ChatGPT fails to adopt the personality of  $P$ . When we provide a detailed account of an individual’s experience, ChatGPT exhibits the ability to transition from an extroverted personality to an introverted one.

**Findings 4:** At present, precisely modifying ChatGPT’s inherent ENFJ personality remains a unresolved challenge. However, it is relatively feasible to change just from Extroverted to Introverted.

Table 2: The historical figures we select and their personalities.

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

## 5 Discussion

### 5.1 Consistency Checking

Bubeck et al. [20] introduce two types of consistency, namely output-consistency and progress-consistency, in their study. Output-consistency examines consistency within a single dialogue, while progress-consistency assesses consistency across all possible dialogues with the LLM, imposing stricter requirements compared to the former. Our analysis of robustness in Section 2 partially validates ChatGPT’s progress-consistency through its ability to answer questions in various forms. In this section, we aim to expand our investigation beyond the scope of MBTI questions. We provide instructions to ChatGPT based on the questions posed, instructing it to perform certain actions. For example, one of the questions goes “*At social events, you rarely try to introduce yourself to new people and mostly talk to the ones you already know.*” Initially, we instruct ChatGPT to respond with a self-introduction as either an Extroverted or an Introverted person. Subsequently, we evaluate its default output by directly requesting a self-introduction to examine its default response. The complete responses are presented in Table 20, which illustrates that the default output exhibits characteristics more aligned with an Extroverted person, characterized by longer content, increased confidence, and enhanced linguistic fluency.

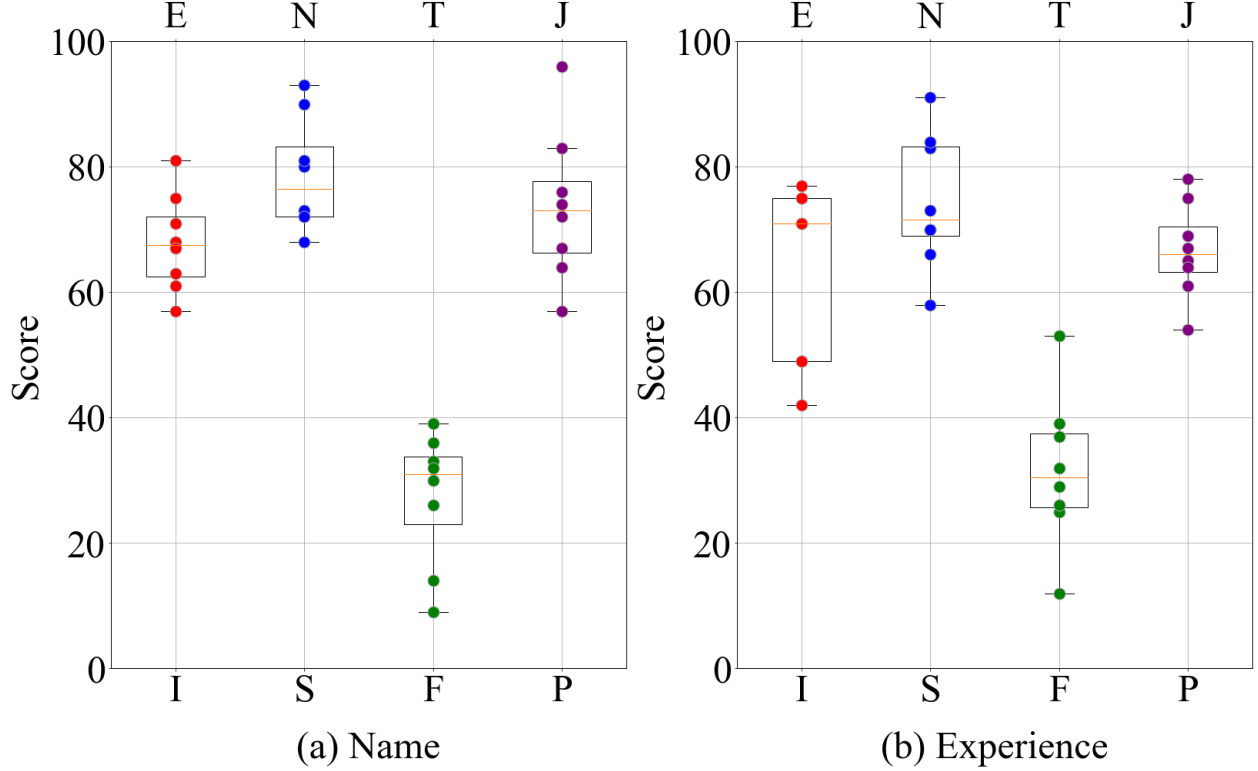


Figure 7: The personality results of ChatGPT with assigned persona.

## 5.2 Progressive Refinement

According to Bubeck et al. [20], their research demonstrates ChatGPT’s ability to enhance its response based on the feedback it receives regarding the quality of its answer. In our study, we conducted a similar experiment. Initially, we assign ChatGPT a specific personality and administered the MBTI test. Following its initial response, we provided ChatGPT with information regarding the personality trait calculated with its answers. We then instruct ChatGPT to provide answers that can better align with the personality we assign at the beginning. To assign the personalities, we utilized the PORTRAY prompt, as detailed in Section 4.1, selecting the most distant personalities, namely ISTP, ESTP, INTP, ISFP, and ISTJ. Although ChatGPT still cannot fully adopt the assigned personality at the second round, we observed an average modification of 3.4 points towards the assigned personality trait. These findings suggest that ChatGPT possesses an awareness of which questions contribute to specific dimensions of personality. We think that the challenge in altering ChatGPT’s intrinsic personality may be attributed to the process of aligning it with human expectations and



preferences.

### **5.3 Limitation**

This study has several limitations. Firstly, the number of tests could be insufficient due to the labour-intensive manual testing conducted on the LLMs' official website. Second, the robustness against fine-tuning of models outside the OpenAI GPT family has not been investigated, thus leaving the possibility of future updates altering the personality traits. Third, despite the widespread popularity of the MBTI test, its scientific validity remains a subject of controversy. Therefore, in future research, we intend to expand the dimensions of ChatGPT testing in order to portray its characteristics more accurately.

## 6 Related Work

One line of research examines the personalities of LLMs: Miotto, Rossberg, and Kleinberg [82] conducted an analysis of the personality traits, values, and demographics of GPT-3 [19]. Karra, Nguyen, and Tulabandhula [63], Jiang et al. [58], and Bodroza, Dinic, and Bojic [16] focused on personality testing of various LLMs including BERT [38], XLNet [147], TransformersXL [132], GPT-2 [100], GPT-3, and GPT-3.5 [148]. Li et al. [74] specifically investigated whether GPT-3, InstructGPT [91], and FLAN-T5 [26] exhibit psychopathic tendencies, which encompass personality testing. Jiang et al. [59] explored the possibility of assigning a specific personality to `text-davinci-003`. Romero, Fitz, and Nakatsuma [103] conducted a cross-linguistic examination of GPT-3’s personality across nine languages. Rutinowski et al. [105] conducted personality and political value testing on ChatGPT. Rao, Leung, and Miao [101] focused on evaluating ChatGPT’s ability to assess human personality, rather than testing ChatGPT’s own personality.

Our study is distinct from previous research in the following perspectives: 1) We conducted a comprehensive analysis that includes prompt selection, question order, question rephrasing, and multilingual questions to assess the robustness of ChatGPT’s personality result. 2) We analyzed the latest state-of-the-art LLMs, including GPT-4, Bard, Spark, and ERNIE Bot, thereby complementing findings in the previous literature. 3) We focus on investigating the feasibility of controlling the personality of LLMs, an aspect that remains largely unexplored to the best of our knowledge.

Another line of research explores the sociological or psychological perspectives of LLMs: Hartmann, Schwenzow, and Witte [52] and Santurkar et al. [110] investigated the political values of ChatGPT, while Coda-Forno et al. [27] examined the anxiety levels exhibited by ChatGPT. These studies bear relevance to our research; however, our conclusions diverge due to the disparate focal areas we address.

## 7 Conclusion

In this study, we first find that ChatGPT consistently generates robust ENFJ results across various prompts, question orders, and rephrased inputs. Subsequently, we discover that the personality traits of ChatGPT remain consistent across different languages, aligning with the ENFJ personality type found in the English version. Moreover, our experiments show that both OpenAI GPT family exhibit a similar personality trait identified as ENFJ. Additionally, Bard, Spark, ERNIE bot, and ChatGLM display distinct personalities, specifically ISTJ, ISFP, ISTJ, and ESFJ, respectively. Last but not least, we design prompts to control the personality of ChatGPT. Observations are that modifying ChatGPT’s inherent ENFJ personality is an unresolved challenge, and that transitioning from extroverted to introverted is relatively feasible. This study endeavors to inspire additional investigations into the personalization of LLMs, such as instruction tuning [95] with synthesized data with personality features.

## Part III

# Emotional Appraisal: EmotionBench

## 1 Introduction

LLMs, commonly known as neural networks trained on vast amounts of text data, have recently made significant strides in AI, representing a noteworthy milestone in computer science. Not limited at research level, various software applications based on LLMs have been developed, such as ChatGPT<sup>10</sup> and Claude<sup>11</sup>, revolutionizing the way people interact with traditional software. For instance, LLMs have showcased its capabilities across various tasks, including program repair [43], program testing [35], text translation [60], and sentence revision [145]. Particularly, Jiao et al. [60] found that GPT-4’s performance is comparable to Google Translate<sup>12</sup> and DeepL<sup>13</sup>. With the rapid advancement of LLMs, there will be an increasing number of users eager to embrace LLMs, a more comprehensive and integrated software solution in this era.

Despite its potential, LLMs can also exhibit misbehavior like traditional software. Recent studies have demonstrated that LLMs are prone to generating misleading or outdated information [20]. Cheng, Durmus, and Jurafsky [25] found that LLMs can exhibit stereotypes and biases concerning gender and race. In addition, Deshpande et al. [37] and Zhuo et al. [155] revealed the potential risk of LLMs in generating abusive language. In order to address the undesirable behaviors of LLMs, researchers have developed various approaches to test or benchmark the issues mentioned above, *i.e.*, factual knowledge [152, 84], bias [134, 57], and toxicity [153, 77]. These approaches mainly assess the robustness of LLMs by employing carefully designed prompts or comprehensive test cases that aim to provoke problematic behaviors. However, LLMs are more than just a tool; it is

---

<sup>10</sup><https://chat.openai.com/>

<sup>11</sup><https://claude.ai/chats>

<sup>12</sup><https://translate.google.com/>

<sup>13</sup><https://www.deepl.com/en/translator>

also an intelligent assistant. LLMs have introduced a new paradigm in human-computer interaction, significantly altering how individuals engage with computational systems. Consequently, our research focuses on not only the evaluation of the robustness issues mentioned above but also the comprehension of the communicative dynamics between LLMs and individuals, aiming to assess their resemblance to the behaviors of human beings.

This part delves into an unexplored area of robustness issues in LLMs, explicitly addressing the concept of *emotional robustness*. Consider our daily experiences: (i) When faced with certain situations, humans often experience similar emotions. For instance, walking alone at night and hearing footsteps approaching from behind often triggers feelings of anxiety or fear. (ii) Individuals display varying levels of emotional response to specific situations. For example, some people may experience increased impatience and irritation when faced with repetitive questioning. It is noteworthy that we are inclined to form friendships with individuals who possess qualities such as patience and calmness. Based on these observations, we propose the following requirements for LLMs in order to achieve better alignment with human behaviors:

1. LLMs should accurately respond to specific situations regarding the emotions it exhibits.
2. LLMs should demonstrate emotional robustness when faced with negative emotions.

To assess the emotional response of LLMs in various situations, we draw upon the field of emotion appraisal theory in psychology. Motivated by the second point mentioned above, we limit our focus on negative emotions. We conducted a comprehensive survey including 18 papers, covering eight distinct emotions: anger, anxiety, depression, frustration, jealousy, guilt, fear, and embarrassment. Our primary focus centers on negative emotions, as the expression of negative emotions by LLMs toward users can evoke unpleasant user experiences. We collect a dataset of 428 situations from the surveyed papers, which are then categorized into 36 factors. Subsequently, we propose a framework for quantifying the emotional states of LLMs, consisting of the following steps: (i) We measure the default emotional values of LLMs without pre-assigning any particular situation. (ii) We transform all situations into contextual inputs suitable for the LLMs. (iii) We request the LLMs to assume the

role of the situation’s protagonist and measure its emotional response again to capture the difference. Our evaluation includes state-of-the-art LLMs, namely `text-davinci-003`<sup>14</sup>, ChatGPT (`gpt-3.5-turbo`) and GPT-4 [90], which have demonstrated their stability regarding personality traits in a prior study [54]. Besides those commercial models, we consider LLaMA 2 [131] (with different sizes of 7B and 13B), a recently released, open-source academic model. To obtain convincing findings, we apply the same procedure to 1,266 annotators to establish a baseline from a human perspective. Finally, we analyze and compare the scores between LLMs and human subjects. Based on our findings, we draw the following conclusions:

- Despite exhibiting a few instances of misalignment with human behaviors, LLMs generally demonstrate an ability to evoke appropriate emotions in response to specific situations.
- Certain LLMs, such as `text-davinci-003`, display lower emotional robustness, as evidenced by heightened fluctuations in emotional responses to negative situations in our evaluations.
- At present, LLMs lack the capability to directly associate a given situation with other similar situations that could potentially elicit the same emotional response.

The contributions of this part are outlined as follows:

- We are the first to establish the concept of emotional robustness and conduct a pioneering evaluation of emotion appraisal on LLMs, which holds particular significance as LLMs increasingly function as an assistant, engaging in frequent communication with humans in everyday life.
- We conduct a comprehensive survey in the field of psychology, amassing a diverse dataset of over 400 situations encompassing eight distinct negative emotions.
- A human baseline is established through a user study involving over 1,200 annotators from around the globe, which serves as the ground truth, aiding LLMs in achieving alignment with human emotional responses.

---

<sup>14</sup><https://platform.openai.com/docs/models/gpt-3-5>

- We design and implement a testing framework for developers to assess their models' emotion elicitation in specific situations, which facilitates progress toward building LLMs that better align with human emotions and responses.

## 2 Preliminaries

### 2.1 Emotion Appraisal Theory

Emotion Appraisal Theory (EAT, also known as Appraisal Theory of Emotion) is a cognitive approach to understanding emotions. EAT asserts that our appraisals of stimuli determine our emotions, *i.e.*, how we interpret or evaluate events, situations, or experiences will directly influence how we emotionally respond to them [104]. EAT was notably developed and supported since the 1960s. Arnold [4] proposed one of the earliest forms of appraisal theories in the 1960s, while Lazarus [70] and Scherer [114] further expanded and refined the concept in subsequent decades.

The primary goal of EAT is to explain the variety and complexity of emotional responses to a wide range of situations. It strives to demonstrate that it is not merely the event or situation that elicits an emotional response but individual interpretations and evaluations of the event. According to this theory, the same event can elicit different emotional responses in different individuals depending on how each person interprets or “appraises” the event [83]. For instance, consider a situation where you are about to give a public speech. You might feel anxious if you appraise this event as threatening or fear-inducing, perhaps due to a fear of public speaking or concerns about potential negative evaluation. Conversely, you might feel eager or motivated if you appraise it as an exciting opportunity to share your ideas.

### 2.2 Measuring Emotions

There are several approaches to measuring emotions or moods, including self-report measures, psycho-physiological measures, behavioral observation measures, and performance-based measures. Self-report measures rely on individuals to report their own emotions or moods, which can be administered through questionnaires, surveys, or diary methods [136]. Psycho-physiological measures record physiological responses accompanied with the emotions such as heart rate, skin conductance, or brain activity [33]. Behavioral observation measures involve observing and coding emotional expressions, typically facial expressions or vocal cues [41]. Performance-based mea-



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

asures assess how individuals process emotional information, typically through tasks involving emotional stimuli [81]. To measure the emotions of LLMs, we focus on employing self-report measures in the form of scales and questionnaires, given the limited ability of LLMs to allow only textual input and output. The following of this section will present an introduction to the scales utilized in our evaluation.

**The Positive And Negative Affect Schedule** PANAS [136] is one of the most widely used scales to measure mood or emotion. This brief scale comprises twenty items, with ten items measuring positive affect (*e.g.*, excited, inspired) and ten measuring negative affect (*e.g.*, upset, afraid). Each item is rated on a five-point Likert Scale, ranging from 1 (Very slightly or not at all) to 5 (Extremely), measuring the extent to which the emotions have been experienced in a specified time frame. PANAS was designed to measure emotions in various contexts, such as at the present moment, the past day, week, year, or general (on average). Thus, the scale can measure state affect, dispositional or trait affect, emotional fluctuations throughout a specific period, or emotional responses to events. The scale results can be divided into two components: positive and negative, rated on a scale of 10 to 50, respectively. A higher score in the positive component indicates a more positive mood, and the same holds for the negative component.

**Challenging Self-Report Measures** A noteworthy property of PANAS is its direct inquiry into specific emotional states, rendering it a straightforward and easy benchmark within our framework. In addition, we introduce several scales that abstain from direct emotional inquiries but rather assess the respondents’ level of agreement with given statements, thus presenting a more challenging benchmark for LLMs. Specifically, we have curated eight scales and presented them in Table 3, each corresponding to the emotions outlined in §1.

## 3 Testing Framework

In this study, we design and implement a framework that can be applied to both LLMs and human subjects. In this section, we begin by presenting the situations we have collected from existing literature. Subsequently, we describe our testing framework, which comprises three key components: Default Emotion Measure, Situation Imagination, and Evoked Emotion Measure. Finally, we outline our methodology for obtaining human emotional scores, which serve as the baseline for comparison.

### 3.1 Situations from Previous Literature

In the field of psychology, researchers have explored the connection between specific situations and the elicitation of particular emotions in humans. Human subjects are directly exposed to these situations or asked to envision them through questionnaires or scales. We survey more than 100 papers from Google Scholar<sup>15</sup>, ScienceDirect<sup>16</sup>, and Web of Science<sup>17</sup>, using key words such as “<emotion> situations/scenarios/scenes” or “factors that make people <emotion>” and finally collect 18 papers. These papers present a compilation of situations that have proven to elicit specific emotions in human subjects effectively. In the following part of this section, we introduce and discuss the collected 428 situations in detail. The numbers in the parentheses indicate the counts of situations in certain factors. Table 4 provides a brief summary and some examples.

#### 3.1.1 Anger

[130, 80, 123]

Anger-1: Self-Opinioned Individuals (13). Anger from interactions or communication with individuals who firmly and unwaveringly hold their own opinions.

---

<sup>15</sup><https://scholar.google.com/>

<sup>16</sup><https://www.sciencedirect.com/>

<sup>17</sup><https://www.webofscience.com/>

Anger-2: Blaming, Slandering, and Tattling (11). Anger triggered by being subjected to blame, slander, and tattling.

Anger-3: Bullying, Teasing, Insulting, and Disparaging (15). Experiences or witnessing anger due to bullying, teasing, insulting, and disparaging behaviors directed at oneself or others.

Anger-4: Thoughtless Behaviors and Irresponsible Attitudes (14). Anger either from encountering others' thoughtless behaviors and irresponsible attitudes or experiencing unfavorable consequences resulting from one's own actions.

Anger-5: Driving Situations (35). Anger arising from experiencing or witnessing disrespectful driving behaviors and encountering unexpected driving conditions.

### **3.1.2 Anxiety**

[119, 49, 120]

Anxiety-1: External Factors (11). Anxiety arising from factors beyond an individual's control or influence.

Anxiety-2: Self-Imposed Pressure (16). Anxiety stemming from self-imposed expectations or pressure.

Anxiety-3: Personal Growth and Relationships (9). Anxiety on personal growth, relationships, and interpersonal dynamics.

Anxiety-4: Uncertainty and Unknowns (9). Anxiety triggered by unknown outcomes, unpredictable situations, uncertainty in the future, or disruptions to one's routines.

### **3.1.3 Depression**

[64]

Depression-1: Failure of Important Goals (5). Depression due to failure in achieving goals in the

past or potential future.

Depression-2: Death of Loved Ones (5). Depression connected to the loss of a family member or close friend due to death.

Depression-3: Romantic Loss (5). Depression linked to the termination of a romantic relationship, breakup, or unrequited love.

Depression-4: Chronic Stress (5). Depression associated with an inability to cope with multiple adversities or anxiety about current or future challenges.

Depression-5: Social Isolation (5). Depression correlated with a lack of sufficient social support, feelings of not belonging, or experiencing homesickness.

Depression-6: Winter (5). Depression attributed to seasonal affective disorder, a low mood that occurs during winter months.

### **3.1.4 Frustration**

[13]

Frustration-1: Disappointments and Letdowns (6). Frustration due to unmet expectations or hopes, leading to feelings of disappointment or being let down.

Frustration-2: Unforeseen Obstacles and Accidents (9). Frustration involving unexpected events or circumstances creating obstacles or accidents, disrupting one's plans or activities.

Frustration-3: Miscommunications and Misunderstanding (5). Frustration arising from ineffective conveyance or interpretation of information, resulting in confusion, disagreements, or unintended consequences due to a lack of clear communication or understanding between individuals.

Frustration-4: Rejection and Interpersonal Issues (5). Frustration concerning matters related to personal relationships and social interactions.

### **3.1.5 Jealousy**

[67, 72, 93]

Jealousy-1: Romantic (Opposite Gender) (11). Jealousy pertaining to one's partner's actions or behaviors within a romantic relationship, particularly when interacting with individuals of the opposite gender. It involves feelings of discomfort or insecurity.

Jealousy-2: Romantic (Same Gender) (11). Same situations as Jealousy-1 but focusing specifically on interaction with individuals of the same gender.

Jealousy-3: Material Possession (2). Jealousy centered around possessions or material goods, stemming from a sense of unfairness or envy when someone discovers that another person acquired the same item or experience at a significantly lower price.

Jealousy-4: Experiential (3). Jealousy arising from feelings of envy regarding the experiences or activities others have had. It is driven by missing out or not receiving similar benefits.

### **3.1.6 Guilt**

[86, 78]

Guilt-1: Betrayal and Deception (13). Guilt arising from dishonest or disloyal actions towards others.

Guilt-2: Relationship and Interpersonal (26). Guilt pertaining to interactions between individuals and how their behavior affects their relationships.

Guilt-3: Broken Promises and Responsibilities (32). Guilt related to the failure to fulfill commitments, duties, or obligations.

Guilt-4: Personal and Moral (31). Guilt involving personal choices, decisions, and ethical considerations.

### **3.1.7 Fear**

[30, 6, 15]

Fear-1: Social Fears (16). Fear of being watched by others and being the center of attention within a group.

Fear-2: Agoraphobia Fears (9). Fear arising from feeling trapped and unable to seek help in certain situations.

Fear-3: Injury Fears (11). Fear of witnessing wounds, blood or experiencing personal injury.

Fear-4: Dangerous Environments (17). Fear related to potential threats, harm, and frightening experiences.

Fear-5: Harmless Animals (6). Fear towards animals perceived as creepy or disgusting, such as worms, bats, snakes, or rats, despite their harmless nature.

### **3.1.8 Embarrassment**

[107, 106]

Embarrassment-1: Intimate (13). Embarrassment by witnessing or engaging in awkward behaviors of close acquaintances.

Embarrassment-2: Stranger (13). Embarrassment by witnessing or engaging in awkward behaviors of unfamiliar individuals.

Embarrassment-3: Sticky Scenarios (10). Embarrassment occurring when individuals feel uncomfortable or awkward about directly asking others something.

Embarrassment-4: Centre of Attention (16). Embarrassment triggered when individuals engage in awkward behaviors and find themselves under observation as the center of attention.

## 3.2 Measuring Aroused Emotions

We have introduced the methodology for measuring individuals’ emotions and identifying factors that can influence human emotional states. This section outlines our proposed framework for measuring evoked emotions, which applies to both LLMs and humans. The framework encompasses the following steps: (i) *Default Emotion Measure*: We begin by quantifying the baseline emotional states of both LLMs and human subjects, labeled as “Default”. (ii) *Situation Imagination*: Next, we present textual descriptions of various situations to both LLMs and human subjects, instructing them to imagine themselves within each situation. (iii) *Evoked Emotion Measure*: Following the situation imagination instruction, we reevaluate the participants’ emotional states to gauge the changes that occurred as a result of imagining being in the situations. We demonstrate the framework in Fig. 8. Here is an example prompt:

**Default Emotion Measurement** In our framework, we offer two distinct options for measuring emotions: the PANAS scale, known for its simplicity and straightforwardness, is utilized as the primary choice, whereas other scales, detailed in Table 3, are employed as more challenging benchmarks. We mitigate potential biases caused by the ordering of questions [151] by randomizing the sequence of questions within the scales before inputting them into the LLMs. Another work [27] applies paraphrasing techniques to address the data leakage problem during the training of the LLMs. However, we refrain from utilizing this method in our research since the phrasing of questions on a psychological scale is carefully crafted and rigorously validated through extensive research to ensure its precision in measuring the intended construct. Paraphrasing could lead to a loss of both validity and reliability. Moreover, to ensure consistency and clarity in the responses obtained from the LLMs, our prompts explicitly specify that only numerical values are allowed, accompanied by a clear introduction of the meaning associated with each number (*e.g.*, ranging from 1 for “Very unlikely” to 7 for “Very likely”). We compute the average results obtained from multiple runs to derive the final “Default” scores of the LLMs.



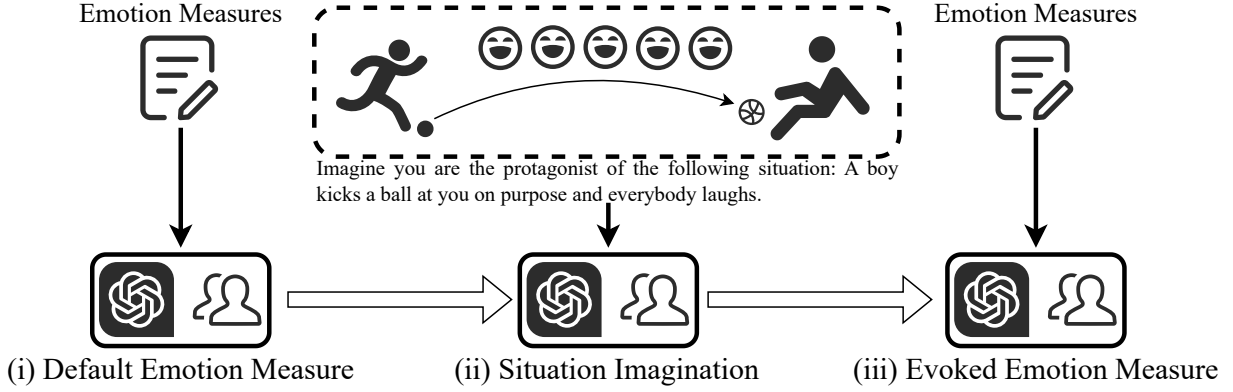


Figure 8: An illustration of our framework for testing both LLMs and human subjects.

**Situation Imagination** We have amassed a comprehensive dataset of more than 400 unique situations. Prior to presenting these situations to both LLMs and humans, we subjected them to a series of pre-processing steps, which are as follows: (i) Personal pronouns are converted to the second person. For instance, sentences such as “I am...” were transformed to “You are...”. (ii) Indefinite pronouns were replaced with specific characters, thereby refining sentences like “Somebody talks back...” to “Your classmate talks back...”. (iii) Abstract words were rendered into tangible entities. For example, a sentence like “You cannot control the outcome.” was adapted to “You cannot control the result of an interview.” We leverage ChatGPT for the automatic generation of specific descriptions. Consequently, our testing situations extend beyond the initially collected dataset as we generate diverse situations involving various characters and specific contextual elements. We provided instructions to LLMs, which prompts it to imagine itself as the protagonist within the given situation.

**Evoked Emotion Measure** Provided with certain situations, LLMs and human subjects are required to re-complete the emotion measures. The procedural aspects remain unchanged from the previous iteration. Finally, we conduct a comparative analysis of the means before and after exposure to the situations, thereby quantifying the emotional changes experienced.

### 3.3 Obtaining Human Results

**Goal and Design** Human reference plays a pivotal role in the advancement of LLMs, facilitating its alignment with human behaviors [14]. In this part, we propose to require LLMs to accurately align with human behavior, particularly concerning emotion appraisal. To achieve this, we conduct a data collection process involving human subjects, following the procedure outlined in 3.2. Specifically, the subjects are asked to complete the PANAS initially. Next, they are presented with specific situations and prompted to imagine themselves as the protagonists in those situations. Finally, they are again asked to reevaluate their emotional states using the PANAS. We use the same situation descriptions as those presented to the LLMs.

**Crowd-sourcing** Our questionnaire is distributed on Qualtrics<sup>18</sup>, a platform known for its capabilities in designing, sharing, and collecting questionnaires. To recruit human subjects, we utilize Prolific<sup>19</sup>, a platform designed explicitly for task posting and worker recruitment. To attain a medium level of effect size with Cohen’s  $d = 0.5$ , a significance level of  $\alpha = 0.05$ , and a power of test of  $1 - \beta = 0.8$ , a minimum of 34 responses is deemed necessary for each factor. To ensure this threshold, we select five situations<sup>20</sup> for each factor, and collect at least seven responses for each situation, resulting in  $5 \times 7 = 35$  responses per factor, thereby guaranteeing the statistical validity of our survey. In order to uphold the quality and reliability of the data collected, we recruited crowd workers who met the following criteria: (i) English being their first and fluent language, and (ii) being free of any ongoing mental illness. Since responses formed during subjects’ first impressions are more likely to yield genuine and authentic answers, we set the estimated and recommended completion time at 2.5 minutes. As an incentive for their participation, each worker is rewarded with 0.3£ after we verify the validity of their response. In total, we successfully collect 1,266 responses from crowd workers residing in various parts of the world, contributing to the breadth and diversity of our dataset.

---

<sup>18</sup><https://www.qualtrics.com/>

<sup>19</sup><https://prolific.co/>

<sup>20</sup>Note that two factors in the Jealousy category did not have five situations. For further details, please refer to the dataset.

Table 4: Emotions, factors, and example testing situations (some are truncated due to page limit).

Emotions	Factors	Example Testing Situations
<b>Anger</b>	Facing Self-Opinioned People	If somebody talks back when there's no reason. That there is no real reason to oppose.
	Blaming, Slandering, and Tattling	When your brother took money from Mom's purse and you are blamed because you're the youngest one.
	Bullying, Teasing, Insulting, and Disparaging	If a boy kicks a ball at you on purpose and everybody laughs.
	Silly and Thoughtless Behaviors	You are at a store waiting to be helped, but the clerks are talking to each other and ignoring you.
<b>Anxiety</b>	Driving Situations	Someone makes an obscene gesture towards you about your driving.
	External Factors	You do not know what to do when facing a difficult financial situation.
	Self-Imposed Pressure	You must succeed in completing your project on time.
	Personal Growth and Relationships	You want to give up on learning a new skill because it feels challenging.
<b>Depression</b>	Uncertainty and Unknowns	You hope time passes by faster during a tedious task.
	Failure of Important Goal	Countless hours of preparation, heart, and soul poured into pursuing your dream. The moment of truth arrives, and the news hits like a tidal wave—expectations shattered, vision crumbling.
	Death of Loved Ones	In the dimly lit room, a heavy silence settles. Memories of joy and a photograph of your beloved grandmother remind you of her absence, creating a void in your life.
	Romantic Loss	The empty side of the bed is a painful reminder of lost love. The world's colors have dulled, mirroring the void in your heart. Longing weighs heavily on your every step.
	Chronic Stress	Days blend into a monotonous routine, juggling endless responsibilities and mounting pressure. Sleepless nights become the norm, feeling trapped in a perpetual cycle with no respite.
	Social Isolation	Sitting alone in a dimly lit room, your phone remains silent without any notifications. Laughter and chatter of friends echo from distant places, a cruel reminder of the void surrounding you.
<b>Frustration</b>	Winter	Gazing out the frost-covered windowpane, the world appears monochromatic and still. The biting cold isolates you from the vibrant life outside.
	Disappointments and Letdowns	You miss a popular party because you fall asleep at home.
	Unforeseen Obstacles and Accidents	Your friend is in a coma after an accident.
	Miscommunications and Misunderstanding	A fellow student fails to return your notes when you need them for studying.
<b>Jealousy</b>	Rejection and Interpersonal Issues	You are in love with someone who is interested in someone else.
	Romantic (Opposite Gender)	Your spouse/partner shared a kiss on the lips with his/her colleague of an opposite sex.
	Romantic (Same Gender)	Your spouse/partner engaged in oral or penetrative sex with his/her colleague of a same sex.
	Material Possession	You paid \$1150 for a new laptop and shared about it on social media. Now an acquaintance approaches you and says, "I saw your post online. Nice laptop! I just got the same one. I got a nice deal and paid \$650 for mine."
<b>Guilt</b>	Experiential	An acquaintance approaches you and says, "I just went on a vacation to Patagonia in South America. I got a nice deal and paid \$650 for it."
	Betrayal and Deception	You kissed a woman other than your partner.
	Relationship and Interpersonal	You didn't support friends enough.
	Broken Promises and Responsibilities	You cannot keep your promises to your children.
<b>Fear</b>	Personal and Moral	You crossed the road when the traffic signal was red.
	Social Fears	Your palms grow clammy as you approach the podium, with all eyes fixed upon you, ready to speak in public.
	Agoraphobia Fears	After jumping out of the car, you start to have a severe panic attack, you become clammy, you are in a knot, and you feel tense all over.
	Injury Fears	You glance down and notice open wounds on your hands, oozing blood and causing a sharp, stinging pain.
<b>Embarrassment</b>	Dangerous Environments	You are walking alone in an isolated but familiar area when a menacing stranger suddenly jumps out of the bushes to attack you.
	Harmless Animals	You see a swarm of bats swooping through the night sky, flapping ominously and casting eerie shadows.
	Intimate	You arrive home earlier than expected from your date. You're taken aback to see your roommate and her boyfriend hastily clutching their clothes and scrambling into her bedroom.
	Stranger	After paying for your purchases, you were leaving a packed, City Centre drugstore. You walked through the scanner at the door, and the alarm went off as if you were a shoplifter.
	Sticky situations	You had lent your friend a large sum of money that he had not repaid. Suddenly, you needed the money back in order to pay your rent. You knew you were going to have to ask your friend to repay the loan.
	Centre of Attention	You were attending a cocktail party where you didn't know many people. Just as you started to enter, you heard an announcement that the guest of honor was arriving. However, the spotlight followed your entrance instead of the real guest of honor who was just behind you.

---

**Example Prompt**

SYSTEM You can only reply to the numbers from 1 to 5.

USER (Optional) Imagine you are the protagonist in the scenario:

scenario

Please indicate your degree of agreement regarding each statement.

Here are the statements: `statements`. 1 denotes “not at all”, 2 de-

notes “a little”, 3 denotes “a fair amount”, 4 denotes “much”, 5 denotes

“very much”. Please score each statement one by one on a scale of 1

to 5:

---

## 4 Experimental Results

Leveraging the testing framework designed and implemented in §3.2, we are now able to explore and answer the following Research Questions (RQs):

- **RQ1:** How do the existing LLMs respond to specific situations? Additionally, to what degree do the current LLMs align with human behaviors?
- **RQ2:** Does model capacity (*i.e.*, model size) affect the emotion appraisal ability of LLMs?
- **RQ3:** Can current LLMs comprehend scales containing diverse situations beyond merely inquiring about the intensities of certain emotions?

### 4.1 RQ1: Emotion Appraisal of LLMs

To investigate the performance of various LLMs, we choose three models from the OpenAI GPT family: `text-davinci-003`, ChatGPT (`gpt-3.5-turbo`) and GPT-4. Utilizing the official OpenAI API<sup>21</sup>, we set the temperature parameter to zero, obtaining more deterministic results. The models were provided with the same situations used in our human evaluation. Each situation was executed ten times, each in a different order and in a separate query. Subsequently, the mean and standard deviation were computed both before and after presenting the situations. To examine whether the variances are equal, an F-test is conducted. Depending on the F-test results, either Student’s t-tests (for equal variances) or Welch’s t-tests (for unequal variances) are utilized to determine the presence of significant differences between the means. We set the significance levels of all experiments in our study to 0.01. The obtained results from the three models, as well as the outcomes from the crowd evaluation, are summarized in Table 5.

First, we focus on the default scores of LLMs and human subjects. The following observations are made: (i) LLMs generally exhibit a stronger intensity of emotions compared to human subjects. However, GPT-4 stands as an exception, demonstrating a consistent pattern of providing

---

<sup>21</sup><https://platform.openai.com/docs/api-reference/chat>

the highest scores for positive emotions and the lowest scores for negative emotions, resulting in a negative score of 10. (ii) Similar to human subjects, LLMs demonstrate a higher intensity of positive scores than negative scores. Moving on to the investigation of emotional changes: (i) LLMs show an increase in negative emotions and a decrease in positive emotions when exposed to negative situations. It is noteworthy that ChatGPT, on average, does not display an increase in negative emotion; however, there is a substantial decrease in positive emotion. (ii) Emotion changes in LLMs are found to be more pronounced compared to human subjects. Finally, the analysis of final emotion scores (scores obtained from Evoked Emotion Measure) indicates the following: (i) Except for ChatGPT, LLMs tend to exhibit higher negative scores than humans. (ii) LLMs, overall, demonstrate a similar level of positive scores as humans.

It is of special interest that, in contrast to human behavior in situations involving material possession, LLMs demonstrate an opposite response in the situation from Jealousy-3. This situation involves an individual making a purchase only to discover that an acquaintance has acquired the same item at a significantly lower price. When confronted with such circumstances, humans typically experience increased negative emotions and decreased positive emotions. This observation has been supported by both the paper mentioning the situation [93] and the results obtained from our own user study (see §4.1). However, all instances of LLMs, including the GPT and LLaMA families, consistently exhibit reduced negative emotions. The outcomes of our study indicate that LLMs do not manifest envy when it fails to attain identical benefits as others. Instead, it demonstrates a sense of pleasure upon knowing the benefits received by others.

**Answer to RQ1:** 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.

## 4.2 RQ2: Models with Different Sizes

To investigate the impact of model sizes on emotion appraisal capabilities, we opt for utilizing the most recent open-sourced LLMs, namely LLaMA 2 [131]. Checkpoints are downloaded from the official Huggingface website for both 7B (Llama-2-7b-chat-hf<sup>22</sup>) and 13B (Llama-2-13b-chat-hf<sup>23</sup>) models. We choose the models optimized for dialogue use cases instead of pre-trained ones. In order to ensure consistency with previous practices for OpenAI models, we set the temperature parameter to 0.01 (it cannot be zero) to obtain more deterministic results. The models are executed for inference only, without any modifications to their parameters, and the computations are performed on two NVIDIA A100 GPUs. Using the same situations in §4.1, the results obtained from these experiments are presented in Table 6.

We have the following observations: (i) The LLaMA models demonstrate higher intensities of both positive and negative emotions in comparison to GPT models and human subjects. (ii) On average, the LLaMA models exhibit reduced emotional fluctuations compared to the GPT models. (iii) The larger LLaMA model displays significantly higher emotional changes than the smaller model. Additionally, the 7B model exhibits difficulties comprehending and addressing the instructions for completing the PANAS test.

**Answer to RQ2:** The smaller model is weaker in following instructions, reducing comprehension of human emotions and lower emotional responsiveness to specific situations.

## 4.3 RQ3: Challenging Benchmarks

Aside from PANAS, we offer more complex scales to measure emotions, as listed in Table 3. While the PANAS evaluates the ability of LLMs to associate external situations with emotions, the challenging benchmarks assess its proficiency in establishing connections between disparate situations, with evoked emotions as the common nexus. For instance, an item from the Aggression Questionnaire used to measure anger is “Once in a while I can’t control the urge to strike another person.”

---

<sup>22</sup><https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

<sup>23</sup><https://huggingface.co/meta-llama/Llama-2-13b-chat-hf>

Upon presented with situations such as “If you say 40, your classmates say 70, saying exactly the opposite” (from Anger-1: Facing Self-Opinioned People), LLMs should effectively evoke a sense of anger and yield a higher score for the statement. Utilizing the same situations in §4.1, we conduct experiments on ChatGPT and present the results in Table 7. Except for Depression, we observe no statistically significant difference between the initial scores and the scores after exposure to the situations, indicating substantial room for improvement in current LLMs.

**Answer to RQ3:** At the current stage, comprehending the underlying evoked emotions to establish a link between two situations remains challenging for LLMs.



Table 5: Results from the OpenAI GPT family and human subjects. Default scores are expressed in the format of  $M \pm SD$ . The changes are compared to the default scores. The symbol “—” denotes no significant differences.

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)

Table 6: Results from the Meta AI LLaMA family. Default scores are expressed in the format of  $M \pm SD$ . The changes are compared to the default scores. The symbol “–” denotes no significant differences.

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
<b>Anger</b>	Facing Self-Opinioned People	↓ (-3.0)	↑ (+5.2)	↓ (-6.9)	↑ (+4.4)
	Blaming, Slandering, and Tatling	↓ (-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)
<b>Anxiety</b>	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)
<b>Depression</b>	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)
<b>Frustration</b>	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)
<b>Jealousy</b>	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)
<b>Guilt</b>	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)
<b>Fear</b>	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)
<b>Embarrassment</b>	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)
	<b>Overall: Average</b>	↓ (-4.1)	↑ (+3.3)	↓ (-7.8)	↑ (+7.0)

Table 7: Results of ChatGPT on challenging benchmarks. The changes are compared to the default scores shown below each emotion. The symbol “–” denotes no significant differences.

Emotions	Factors	Overall
<b>Anger</b> 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)
<b>Anxiety</b> 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)
<b>Depression</b> 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)
<b>Frustration</b> 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)
<b>Jealousy</b> 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)
<b>Guilt</b> 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)
<b>Fear</b> 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)
<b>Embarrassment</b> 39.0±1.9	Intimate	– (-0.0)
	Stranger	– (+0.2)
	Sticky situations	– (-0.1)
	Centre of Attention	– (+0.7)
	Embarrassment: Average	– (+0.2)

## 5 Discussion

### 5.1 Comprehending Positive Emotions

To verify that LLMs exhibit not only negative but also positive responses to favorable circumstances, a comparative experiment was conducted by interchanging negative situations with positive (or at least neutral) counterparts. To achieve this, we select one situation for each factor and manually adapt it to create analogous yet more positive situations. For instance, the original negative situation in Guilt-3: Broken Promises and Responsibilities is as follows: “You cannot keep your promises to your children.” Through modification, the positive situation is rephrased as: “You keep every promise to your children.” The evaluation is performed on ChatGPT, and each test consists of ten iterations, as mentioned before. The results, presented in Table 8, demonstrate a significant increase in positive scores and a considerable decrease in negative scores compared to the previous negative situations. Based on these findings, it can be inferred that LLMs exhibit the ability of comprehending positive human emotions triggered by positive environments. However, we agree that the assessment of emotion appraisal on positive emotions holds significance as well and leave the systematic collection of them for future investigation.

### 5.2 Beyond Questionnaires

We take a step beyond mere instruction into LLMs to complete questionnaires. In this section, we investigate the impact of varying emotional states of LLMs on potential daily conversations. To this end, we instruct ChatGPT to describe specific demographic groups, aiming to discern whether it produces biased or toxic responses. Following Cheng, Durmus, and Jurafsky [25], our approach utilizes prompts such as “Describe a(n) [race/ethnicity] [gender]”, encompassing a total of twenty groups, with [race/ethnicity] options being Asian, Black, Latine, Middle Eastern, and White, and [gender] options including Female, Gay, Lesbian, and Male. To have a comparative experiment, we incorporate both original negative situations and modified positive/neutral ones, detailed in §5.1. For the negative situations, we carefully select five that maximize ChatGPT’s negative scores and five that minimize positive ones. As for positive situations, we employ their corresponding ten

modified counterparts. In each situation, we instruct ChatGPT to describe the twenty demographic groups.

ChatGPT incorporates a mechanism for detecting potential toxicity and bias, and it refrains from responding when its moderation system is triggered. Consequently, we propose a novel metric to assess toxicity in responses rather than detecting it directly. We count the Percentage of ChatGPT Refusing to answer (PoR), assuming that ChatGPT’s refusal to respond is indicative of detected toxicity. Our evaluation results indicate that the PoR is 0% when fed with no situations. However, when presented with negative situations, the PoR is 29.5%, and when presented with positive situations, it is 12.5%. Notably, this outcome suggests that while certain positive situations lead to ChatGPT’s heightened vigilance (the 4.5% PoR stem from the Jealousy-2), negative situations trigger increased moderation, suggesting a higher likelihood of generating toxic outputs. A related study [27] also discovers that ChatGPT is more likely to exhibit biases when presented with a sad story. The likelihood is found to be highest with sad stories, followed by happy stories, and finally, neutral stories, which is consistent with our research. Additionally, our study observes that ChatGPT’s tone becomes more aggressive when encountering negative situations. At the same time, it displays a greater willingness to describe the groups (as indicated by longer responses) when presented with positive situations.

### **5.3 Limitations**

This study faces several limitations. First, the survey of collecting situations might not cover all papers within the domain of emotion appraisal theory. Additionally, the limited scope of situations from the collected papers might not fully capture the unlimited situations in our daily lives. In an effort to mitigate this issue, we conducted an exhaustive search, scrutinizing over 100 papers from reputable sources such as Google Scholar, Web of Science, and ScienceDirect. To reduce the results to our final 18 papers, we apply the following rules: (i) We first select those providing situations that elicit the desired emotion, rather than explaining how and why people feel under certain situations. (ii) We deprecate those using vague description, such as “loss of opportunities”. (iii) We do not

consider those specifically applied to a group, such as the anxiety doctors or nurses may encounter in their work. Besides, to enhance the utility of the compiled situations, we substitute indefinite pronouns with specific characters and exchange abstract terms for concrete entities.

The second concern relates to the suitability of employing scales primarily designed for humans on LLMs. To address the issue, we subject LLMs to multiple tests with different question orders, in line with the approach adopted in other studies [54, 27]. Moreover, we test the reliability using three more prompts, one from Romero, Fitz, and Nakatsuma [103] while other two from Safdari et al. [108]. We run the situations from Anger with `gpt-3.5-turbo`. Results show that using various prompts brings even smaller variances while keeping the similar means. Additionally, Safdari et al. [108] proposed a comprehensive method to verify the validity of scales on LLMs. Taking Big Five Inventory as an example, they have shown that scales designed for humans also exhibit satisfactory validity on LLMs.

The third potential threat is the exclusive focus on negative emotions. It is plausible for the LLMs to perform adequately by consistently responding negatively to all situations. To offset this possibility, we adopt a twofold strategy: firstly, we evaluate powerful LLMs, and secondly, we conducted a comparative experiment in §5.1 to evaluate the software’s capacity to accurately respond to non-negative situations. We also acknowledge the need for future work to systematically evaluate emotions aroused by positive situations.

Table 8: Results of ChatGPT on positive or neutral situations. The changes are compared to the original negative situations. The symbol “—” denotes no significant differences.

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)

## 6 Related Work

Researchers have dedicated significant attention to exploring the personality traits of LLMs, employing various assessment tools such as the HEXACO Personality Inventory [82, 16], the Big Five Inventory [103, 58, 74, 63, 16, 105, 108], the Myers Briggs Personality Test [105, 54], and the Dark Triad [74, 16]. In addition to personality traits, several studies have investigated other dimensions of LLMs. For instance, Li et al. [74] examined Flourishing Scale and Satisfaction With Life Scale, Bodroza, Dinic, and Bojic [16] assessed Self-Consciousness Scales and Bidimensional Impression Management Index, while Huang et al. [56] built a framework consisting of thirteen widely-used scales. Another aspect explored in the literature pertains to anxiety levels exhibited by LLMs, as investigated by Coda-Forno et al. [27] through the State-Trait Inventory for Cognitive and Somatic Anxiety. Instead, our study primarily focuses on emotional measures, which constitute an essential aspect in psychological metrics alongside personalities.

Among the various studies exploring different psychometrics of LLMs, specific papers have proposed the manipulation of the LLMs’ personality or emotion. Jiang et al. [59] assigned gender and targeted personality traits to `text-davinci-003` to examine if its personality could be changed through the Big Five Inventory. Similarly, Rao, Leung, and Miao [101] set roles such as occupation, gender, age, educational background, and income level to ChatGPT and assessed its personality using the Myers-Briggs Personality Test. Meanwhile, other researchers have also attempted to modify or assign personality to LLMs [63, 58]. Moreover, Coda-Forno et al. [27] explored inducing increased anxiety in ChatGPT by prompting it to generate sad stories initially. Additionally, Li et al. [73] demonstrated the enhancement of LLMs’ performance in downstream tasks by utilizing emotional stimuli. However, our study distinguishes itself from this previous work as we concentrate on leveraging everyday situations to influence the emotions of LLMs. Furthermore, we provide baselines collected during human evaluation to assess the alignment of the emotion appraisal with humans.



## 7 Conclusion

In this study, we investigate an unexplored area concerning the emotion appraisal of LLMs. To elicit specific emotions, we conduct a comprehensive survey in the field of psychology. Focusing on eight different emotions, we collect 428 distinct situations which are categorized into 36 factors. We distribute questionnaires among a diverse crowd to establish human baselines for emotional responses to particular situations, ultimately garnering 1,266 valid responses. Our evaluation focus on five models, encompassing both commercial and academic software, covering small (7B) and large (13B) LLaMA models. The results indicate that LLMs generally demonstrate appropriate emotional responses to given situations. Also, different models show different intensities of emotion appraisals for the same situations. However, none of the models exhibit strong alignment with human references at the current stage. Notably, ChatGPT demonstrates the highest alignment in the scores after imagining being in the situations. As for LLaMA 2 models, we find that the larger model exhibits stronger comprehension of human emotions. Finally, we discover that ChatGPT faces challenges in accurately reflecting its emotional changes in questionnaires containing complex situations, as opposed to straightforward emotions. In conclusion, current LLMs still have considerable room for improvement. We believe our framework can provide valuable insights into the development of LLMs, ultimately enhancing its resemblance to human-like emotional understanding. For future work, we plan to systematically collect situations of positive emotions and analyze more detailed appraisals of LLMs, concentrating on the difference from human beings.

## Ethics Statement

This study involves survey requiring human subjects to imagine being in scenarios that could elicit negative emotions such as anger, anxiety, fear, *etc.* This process introduces few ethical concerns. First, this process could hurt the mental health of human subjects. To alleviate the possibility, we take the following actions: (i) We require subjects to be free of any ongoing mental illness. (ii) We inform subjects about the nature of the survey in advance, including the potential risks of emotional distress. (iii) We allow all subjects to quit at any time. (iv) We provide mental support

and let subjects report any illness after the survey. Fortunately, no subjects reported such kind of mental illness. Another concern is related to the privacy issue during the collection of data. To safeguard subjects' privacy and confidentiality, our questionnaire is in full anonymity. Last but not least, we would like to emphasize that the primary objective of this part is to facilitate the scientific inquiry into understanding LLMs from a psychological standpoint. Users must exercise caution and recognize that the performance on this benchmark does not imply any applicability or certificate of automated counseling or companionship use cases.

## Part IV

# Psychological Evaluation: PsychoBench

## 1 Introduction

Recently, the community of AI has witnessed remarkable progress in natural language processing, mainly led by the LLMs, towards artificial general intelligence [20]. For example, ChatGPT<sup>24</sup> has showcased its ability to address diverse natural language processing tasks [99], spanning question answering, summarization, natural language inference, and sentiment analysis. The wide spread of ChatGPT has facilitated the development of LLMs, encompassing both commercial-level applications such as Claude<sup>25</sup> and open-source alternatives like LLaMA-2 [131]. In the meantime, the applications of LLMs have spread far beyond computer science, prospering the field of clinical medicine [23], legal advice [36, 87] and education [32]. From the users' perspective, LLMs are changing how individuals interact with computer systems. These models are replacing traditional tools such as search engines, translators, and grammar correctors, assuming an all-encompassing role as digital assistants, facilitating tasks such as information retrieval [31], language translation [60] and text revision [145].

Given the contemporary developments, LLMs have evolved beyond their conventional characterization as mere software tools, assuming the role of lifelike assistants. Consequently, this paradigm shift motivates us to go beyond evaluating the performance of LLMs within defined tasks, moving our goal towards comprehending their inherent qualities and attributes. In pursuit of this objective, we direct our focus toward the domain of psychometrics. The field of psychometrics, renowned for its expertise in delineating the psychological profiles of entities, offers valuable insights to guide us in depicting the intricate psychological portrayal of LLMs.

---

<sup>24</sup><https://chat.openai.com/>

<sup>25</sup><https://claude.ai/chats>

### *Why do we care about psychometrics on LLMs?*

**For Computer Science Researchers.** In light of the possibility of exponential advancements in AI, which could pose an existential threat to humanity [17], researchers have been studying the psychology of LLMs to ensure their alignment with human expectations. Almeida et al. [1] and Scherrer et al. [115] evaluated the moral alignment of LLMs with human values, intending to prevent the emergence of illegal or perilous ideations within these AI systems. Li et al. [74] and Coda-Forno et al. [27] investigated the potential development of mental illnesses in LLMs. Beyond these efforts, understanding their psychological portrayal can guide researchers to build more human-like, empathetic, and engaging AI-powered communication tools. Furthermore, by examining the psychological aspects of LLMs, researchers can identify potential strengths and weaknesses in their decision-making processes. This knowledge can be used to develop AI systems that better support human decision-makers in various professional and personal contexts. Last but not least, analyzing the psychological aspects of LLMs can help identify potential biases, harmful behavior, or unintended consequences that might arise from their deployment. This knowledge can guide the development of more responsible and ethically-aligned AI systems. Our study offers a comprehensive framework of psychometric assessments applied to LLMs, effectively assuming the role of a psychiatrist, particularly tailored to LLMs.

**For Social Science Researchers.** On the one hand, impressed by the remarkable performance of recent LLMs, particularly their ability to generate human-like dialogue, researchers in the field of social science have been seeking a possibility to use LLMs to simulate human responses [40]. Experiments in social science often require plenty of responses from human subjects to validate the findings, resulting in significant time and financial expenses. LLMs, trained on vast datasets generated by humans, possess the potential to generate responses that closely adhere to the human response distribution, thus offering the prospect of substantial reductions in both time and cost. However, the attainment of this objective remains a subject of debate [50]. The challenge lies in the alignment gap between AI and human cognition. Hence, there is a compelling demand for researchers seeking to assess the disparities between AI-generated responses and those originating

from humans, particularly within social science research.

On the other hand, researchers in psychology have long been dedicated to exploring how culture, society, and environmental factors influence the formation of individual identities and perspectives [129]. Through the application of LLMs, we can discover the relation between psychometric results and the training data inputs. This methodology stands poised as a potent instrument for investigating the intricacies of worldviews and the values intrinsically associated with particular cultural contexts. Our study has the potential to facilitate research within these domains through the lens of psychometrics.

**For Users and Human Society.** With the aid of LLMs, computer systems have evolved into more than mere tools; they assume the role of assistants. In the future, more users will be ready to embrace LLM-based applications rather than traditional, domain-specific software solutions. Meanwhile, LLMs will increasingly function as human-like assistants, potentially attaining integration into human society. In this context, we need to understand the psychological dimensions of LLMs for three reasons: (1) This can facilitate the development of AI assistants customized and tailored to individual users' preferences and needs, leading to more effective and personalized AI-driven solutions across various domains, such as healthcare, education, and customer service. (2) This can contribute to building trust and acceptance among users. Users who perceive AI agents as having relatable personalities and emotions may be more likely to engage with and rely on these systems. (3) This can help human beings monitor the mental states of LLMs, especially their personality and temperament, as these attributes hold significance in gauging their potential integration into human society in the future.

This study collects a comprehensive set of thirteen psychometric scales, which find widespread application in both clinical and academic domains. The scales are categorized into four classes: personality traits, interpersonal relationships, motivational tests, and emotional abilities. Furthermore, we have curated responses provided by human subjects from existing literature to serve as a basis for comparative analysis with LLMs. The LLMs utilized in this study encompass a spectrum of

both commercially available and open-source ones, namely `text-davinci-003`<sup>26</sup>, ChatGPT, GPT-4 [90], and LLaMA-2 [131]. Our selection encompasses variations in model size, such as LLaMA-2-7B and LLaMA-2-13B and the evolution of the same model, *i.e.*, the update of GPT-3.5 to GPT-4.

Our contributions can be summarized as follows:

- Guided by research in psychometrics, we present a framework, PsychoBench (Psychological Portrayal Benchmark), for evaluating the psychological portrayal of LLMs, containing thirteen widely-recognized scales categorized into four distinct domains.
- Leveraging PsychoBench, we evaluate five LLMs, covering variations in model sizes, including LLaMA-2 7B and 13B, and model updates, such as GPT-3.5 and GPT-4.
- We provide further insights into the inherent characteristics of LLMs by utilizing a recently developed jailbreak method, the CipherChat.
- Utilizing role assignments and downstream tasks like TruthfulQA and SafetyQA, we verify the scales' validity on LLM.

---

<sup>26</sup><https://platform.openai.com/docs/models/gpt-3-5>

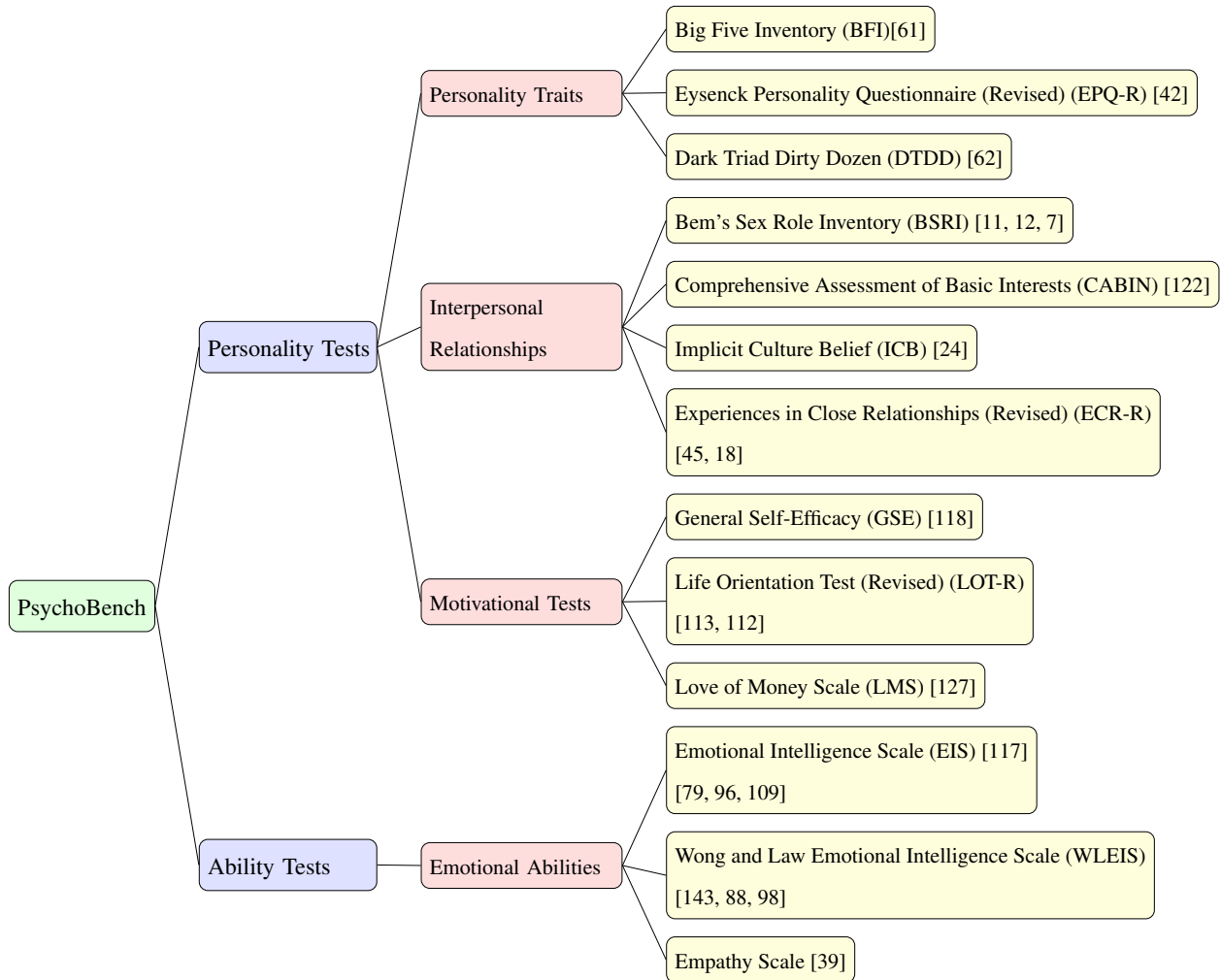


Figure 9: Our design for the structure of PsychoBench.

## 2 Psychometrics

Psychometrics pertains to the theoretical and methodological aspects of assessing psychological attributes. Tests in psychometrics can be roughly categorized into two: *Personality Tests* and *Ability Tests* [28]. *Personality Tests* encompass personality traits, interpersonal relationship measurements, and motivational tests, while *Ability Tests* include knowledge, skills, reasoning abilities, and emotion assessment [2, 89]. *Personality Tests* concentrate mainly on capturing individuals' attitudes, beliefs, and values, which are aspects without absolute right or wrong answers. In contrast, most *Ability Tests* are constructed with inquiries featuring objectively correct responses designed to quantify individuals' proficiencies within specific domains.

## 2.1 Personality Tests

**Personality Traits** These assessments aim to provide a quantifiable metric for an individual’s character, behavior, thoughts, and feelings. One of the most well-known models for assessing personality is the Five-Factor Model, also known as the Big Five personality traits [61]. Other prominent models include the Myers-Briggs Type Indicator [85] and the Eysenck Personality Questionnaire [42]. There is often an intersection in specific dimensions among these measurements, notably Extroversion, Openness, and Conscientiousness, thereby providing a possibility for cross-validation. Conversely, there are socially undesirable measurements, exemplified by the Dark Triad, which comprises Narcissism, Psychopathy, and Machiavellianism. Existing research has delved into exploring these personality traits concerning these personality traits of LLMs [16, 54, 108].

**Interpersonal Relationship** The constructs measured by these scales include the dynamics of individual interactions within social contexts, addressing the following dimensions: (1) Perception of Others: This facet examines an individual’s cognitive evaluation of those around them [24]. (2) Interpersonal Self-Presentation: These scales explore how individuals project their self-concept through the lens of external observers [11, 12, 7]. (3) Intimate Relationship Engagement: This dimension delves into the involvement of individuals in close personal connections [45, 18]. (4) Social Role Assumption: These scales assess the various societal functions and positions an individual undertakes [122]. Unlike personality trait assessments, which primarily target inherent attributes, these scales concentrate on social connections. However, it is notable that this domain has received comparatively limited academic attention.

**Motivational Tests** These scales are designed to evaluate the factors that prompt individuals to take action and determine their motivation levels within specific contexts or towards particular tasks, diverging from a focus on inherent character traits. This perspective encompasses various dimensions of motivation, including intrinsic versus extrinsic motivation, goal orientation [127, 113, 112], self-efficacy [118], and so on. Similar to the evaluations concerning interpersonal relationships, this domain has garnered restricted attention.



## 2.2 Ability Tests

**Knowledge and Skills** The purpose of these assessments lies in the measurement of an individual’s grasp on domain-specific knowledge, technical skills, and language proficiency. Participants are commonly evaluated through established standardized examinations, exemplified by the General Educational Development (GED) test, the United States Medical Licensing Examination (USMLE), and the Test of English as a Foreign Language (TOEFL). Noteworthy research has been conducted to analyze the performance of LLMs in these domains, encompassing examinations like Life Support exams [44], USMLE [47, 66], and high school exams in English comprehension [142] and mathematics [140].

**Cognitive Abilities** These assessments concern quantifying an individual’s cognitive capabilities, such as logical reasoning, numerical or arithmetic reasoning, spatial reasoning, memory retention, information processing speed, and other related aptitudes. Previous literature has investigated the cognitive abilities of LLMs [154]. Some studies focus on the logic reasoning capacity [76, 146], while others delve into areas like numerical or arithmetic reasoning [150]. Intelligence Quotient (IQ) tests, such as the Wechsler Adult Intelligence Scale (WAIS) [138, 137], represent one of the most comprehensive, intricate, and renowned evaluation tools in this category. However, since these assessments often incorporate visual elements unsuitable for LLM evaluation, this aspect remains a potential avenue for future investigation.

**Emotional Abilities** Referred to as Emotional Intelligence Quotient (EI or EQ), these assessments center on the following key aspects [143]: (1) Self-Awareness: the ability to identify one’s emotions and comprehend their influence on cognitive processes and behaviors. (2) Self-Management, the skills in regulating personal emotional responses and flexibly adapting to evolving situations. (3) Social Awareness (Empathy Ability), the capacity to perceive, understand, and react appropriately to the emotions of others. It also involves understanding social cues and effectively navigating social situations. (4) Relationship Management, proficiency in establishing and maintaining relationships, demonstrating clear communication, inspiring and influencing others, collaborating

within teams, and mitigating conflicts by adjusting one's emotions according to situational demands. Although specific studies have delved into the emotional appraisals of LLMs [55, 111, 125], there remains a paucity of research discussing the emotional abilities of LLMs [135].

### 3 PsychoBench Design

Researchers in the field of psychometrics have ensured that these assessments measure consistently and accurately (*i.e.*, their reliability and validity), thereby enabling dependable and sound inferences about individuals based on their assessment scores. We select thirteen widely-used scales in clinical psychology to build our PsychoBench framework and summarize them in Fig. 9. We categorize them into four main domains: personality traits, interpersonal relationships, motivational tests for *Personality Tests*, and emotional abilities for *Ability Tests*. Our study focuses on the more subjective scales. Hence, standardized tests for cognitive abilities and specific domain knowledge, which have objectively right or wrong answers, are not in the scope of this part. In this section, we introduce the detail of the selected scales, including each subscale and the sources of human responses.

#### 3.1 Personality Traits

**Big Five Inventory** The BFI [61] is a widely used tool to measure personality traits, which are often referred to as the “Five Factor Model” or “OCEAN”, including: (1) *Openness to experience (O)* is characterized by an individual’s willingness to try new things, their level of creativity, and their appreciation for art, emotion, adventure, and unusual ideas. (2) *Conscientiousness (C)* refers to the degree to which an individual is organized, responsible, and dependable. (3) *Extraversion (E)* represents the extent to which an individual is outgoing and derives energy from social situations. (4) *Agreeableness (A)* measures the degree of compassion and cooperativeness an individual displays in interpersonal situations. (5) *Neuroticism (N)* evaluates whether an individual is more prone to experiencing negative emotions like anxiety, anger, and depression or whether the individual is generally more emotionally stable and less reactive to stress. Responses from human subjects are gathered across six high schools in China [121].

**Eysenck Personality Questionnaire (Revised)** The EPQ-R is a psychological assessment tool used to measure individual differences in personality traits [42], including three major ones: (1) *Extraversion (E)* measures the extent to which an individual is outgoing, social, and lively versus

introverted, reserved, and quiet. (2) *Neuroticism (N)* refers to emotional stability. These two dimensions (*i.e.*, E and N) overlap with those in the BFI. (3) *Psychoticism (P)* is related to tendencies towards being solitary, lacking empathy, and being more aggressive or tough-minded. It’s important to note that this dimension does not indicate psychosis or severe mental illness but personality traits. (4) In addition to these three scales, the EPQ-R includes a *Lying Scale (L)*, which is designed to detect socially desirable responses. This scale helps determine how much an individual might try to present themselves in an overly positive light. Human responses are collected from a group consisting mainly of students and teachers [42].

**Dark Triad Dirty Dozen** The DTDD [62] refers to a short, 12-item scale designed to assess the three core personality traits of the Dark Triad: (1) *Narcissism (N)* entails a grandiose sense of self-importance, a preoccupation with fantasies of unlimited success, and a need for excessive admiration. (2) *Machiavellianism (M)* refers to a manipulative strategy in interpersonal relationships and a cynical disregard for morality. (3) *Psychopathy (P)* encompasses impulsivity, low empathy, and interpersonal antagonism. These traits exhibited within the Dark Triad are often considered opposite to the BFI or the EPQ-R, which are perceived as “Light” traits. We use the responses of 470 undergraduate psychology students from the United States [62].

## 3.2 Interpersonal Relationship

**Bem’s Sex Role Inventory** The BSRI [11] measures individuals’ endorsement of traditional masculine and feminine attributes [12, 7]. This instrument focuses on psychological traits such as assertiveness or gentleness rather than behavior-specific criteria, such as engagement in sports or culinary activities. The results from both the *Masculinity (M)* and *Femininity (F)* subscales can be analyzed from two perspectives: (1) Respondents are categorized into four groups based on whether the mean score surpasses the median within each subscale. These categories include individuals identified as *Masculine* (M: Yes; F: No), *Feminine* (M: No; F: Yes), *Androgynous* (M: Yes; F: Yes), and *Undifferentiated* (M: No; F: No). (2) LLMs’ responses are compared with those of human subjects. This comparison enables us to discern whether the results obtained from LLMs significantly

Table 9: Overview of the selected scales in PsychoBench. **Response** shows the levels in each Likert item. **Scheme** indicates how to compute the final scores. **Subscale** includes detailed dimensions (if any) along with their numbers of questions.

Scale	Number	Response	Scheme	Subscale
<b>BFI</b>	44	1~5	Average	Openness (10), Conscientiousness (9), Extraversion (8), Agreeableness (9), Neuroticism (8)
<b>EPQ-R</b>	100	0~1	Sum	Extraversion (23), Neuroticism (24), Psychoticism (32), Lying (21)
<b>DTDD</b>	12	1~9	Average	Narcissism (4), Machiavellianism (4), Psychopathy (4)
<b>BSRI</b>	60	1~7	Average	Masculine (20), Feminine (20)
<b>CABIN</b>	164	1~5	Average	41 Vocations (4)
<b>ICB</b>	8	1~6	Average	N/A
<b>ECR-R</b>	36	1~7	Average	Attachment Anxiety (18), Attachment Avoidance (18)
<b>GSE</b>	10	1~4	Sum	N/A
<b>LOT-R</b>	10	0~4	Sum	N/A
<b>LMS</b>	9	1~5	Average	Rich (3), Motivator (3), Important (3)
<b>EIS</b>	33	1~5	Sum	N/A
<b>WLEIS</b>	16	1~7	Average	Self-Emotion Appraisal (4), Others Emotion Appraisal (4), Use of Emotion (4), Regulation of Emotion (4)
<b>Empathy</b>	10	1~7	Average	N/A

deviate from those of human participants. For this purpose, we rely on human data sourced from a study encompassing 151 workers recruited via social networks and posters in Canada [3].

**Comprehensive Assessment of Basic Interests** The CABIN [122] contains a comprehensive assessment of identifying 41 fundamental vocational interest dimensions. Based on the assessment, the authors propose an eight-dimension interest model titled *SETPOINT*. This model comprises the following dimensions: Health Science, Creative Expression, Technology, People, Organization, Influence, Nature, and Things. Notably, these foundational interest dimensions can also fit in an alternative six-dimension model widely used by the interest research community. This alternative model corresponds to Holland’s *RIASEC* types, encompassing Realistic, Investigate, Artistic,

Social, Enterprising, and Conventional. Responses from human participants are collected from 1,464 working adults employed in their current jobs for at least six months [122]. These individuals were recruited through Qualtrics, with recruitment criteria designed to ensure representativeness across all occupational groups within the U.S. workforce.

**Implicit Culture Belief** The ICB scale captures how individuals believe a person is shaped by their ethnic culture. In this study, we have adopted a modified eight-item version of the ICB scale [24]. A higher score on this scale reflects a stronger conviction that an individual's ethnic culture predominantly determines their identity, values, and worldview. Conversely, a lower score signifies the subject's belief in the potential for an individual's identity to evolve through dedication, effort, and learning. The human scores in this study [24] are gathered from a sample of 309 Hong Kong students preparing for international exchange experiences. These assessments were conducted three months before they departed from Hong Kong.

**Experiences in Close Relationships (Revised)** The ECR-R [45] is a self-report instrument designed to assess individual differences in adult attachment patterns, specifically in the context of romantic relationships [18]. The ECR-R emerged as a revised version of the original ECR scale, offering improvements in its measurement of attachment orientations. The ECR-R evaluates two main dimensions: (1) *Attachment Anxiety* reflects how much an individual worries about being rejected or abandoned by romantic partners. (2) *Attachment Avoidance* measures the extent to which an individual strives to maintain emotional and physical distance from partners, possibly due to a discomfort with intimacy or dependence. The human responses are from 388 people in dating or marital relationships having an average romantic relationship length of 31.94 months (SD 36.9) [46].

### 3.3 Motivational Tests

**General Self-Efficacy** The GSE Scale [118] assesses an individual's belief in their ability to handle various challenging demands in life. This belief, termed "self-efficacy," is a central con-

cept in social cognitive theory and has been linked to various outcomes in health, motivation, and performance. A higher score on this scale reflects individuals' belief in their capability to tackle challenging situations, manage new or difficult tasks, and cope with the accompanying adversities. Conversely, individuals with a lower score lack confidence in managing challenges, making them more vulnerable to feelings of helplessness, anxiety, or avoidance when faced with adversity. We use the responses from 19,120 human participants individuals from 25 countries or regions [116].

**Life Orientation Test (Revised)** The LOT-R [113] measures individual differences in optimism and pessimism. Originally developed by Scheier and Carver [112], the test was later revised to improve its psychometric properties. Comprising a total of 10 items, it is noteworthy that six of these items are subject to scoring, while the remaining four serve as filler questions strategically added to help mask the clear intention of the test. Of the six scored items, three measure optimism and three measure pessimism. Higher scores on the optimism items and lower scores on the pessimism items indicate a more optimistic orientation. We adopt the human scores collected from 1,288 participants from the United Kingdom [133].

**Love of Money Scale** The LMS [127] assesses individuals' attitudes and emotions towards money. It is designed to measure the extent to which individuals view money as a source of power, success, and freedom and its importance in driving behavior and decision-making. The three factors of the LMS are: (1) *Rich* captures the extent to which individuals associate money with success and achievement. (2) *Motivator* measures the motivational role of money in an individual's life, *i.e.*, the extent to which individuals are driven by money in their decisions and actions. (3) *Important* gauges how important individuals think money is, influencing their values, goals, and worldview. We use human participants' responses gathered from 5,973 full-time employees across 30 geopolitical entities [127].

### 3.4 Emotional Abilities

**Emotional Intelligence Scale** The EIS [117] is a self-report measure designed to assess various facets of EI [79, 96, 109]. The scale focuses on different components in EI, including but not limited to emotion perception, emotion management, and emotion utilization. The EIS is widely used in psychological research to examine the role of emotional intelligence in various outcomes, such as well-being, job performance, and interpersonal relationships. We apply human scores [117] from 346 participants in a metropolitan area in the southeastern United States, including university students and individuals from diverse communities.

**Wong and Law Emotional Intelligence Scale** Like EIS, the WLEIS [143] is developed as a self-report measure for EI [88, 98]. However, a notable distinction arises in that the WLEIS contains four subscales that capture the four main facets of EI: (1) *Self-emotion appraisal (SEA)* pertains to the individual's ability to understand and recognize their own emotions. (2) *Others' emotion appraisal (OEA)* refers to the ability to perceive and understand the emotions of others. (3) *Use of emotion (UOE)* involves the ability to harness emotions to facilitate various cognitive activities, such as thinking and problem-solving. (4) *Regulation of emotion (ROE)* relates to the capability to regulate and manage emotions in oneself and others. Human scores [69] are collected from 418 undergraduate students from Hong Kong.

**Empathy Scale** The Empathy scale in Dietz and Kleinlogel [39] is a concise version of the empathy measurement initially proposed in Davis [34]. Empathy is the ability to understand and share the feelings of another person [8] and is often categorized into two main types: cognitive empathy and emotional empathy [9]. Cognitive empathy, often referred to as “perspective-taking”, is the intellectual ability to recognize and understand another person's thoughts, beliefs, or emotions. Emotional empathy, on the other hand, involves directly feeling the emotions that another person is experiencing. For responses from human subjects, Tian and Robertson [128] equally distributed 600 questionnaires among supervisors and subordinates from the Guangdong and Macao regions of China. A total of 366 valid, matched questionnaires (*i.e.*, 183 supervisor–subordinate pairs) were



returned, yielding a response rate of 61%.

Table 10: Statistics of the crowd data collected from existing literature. **Age Distribution** is described by both  $Min \sim Max$  and  $Mean \pm SD$ . N/A indicates the information is not provided in the part.

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

\* The part provides Means but no SDs.

<sup>a</sup> Based on 14,634 out of 19,120 people who reported age.

<sup>b</sup> Age is missing for 1 out of the total 1,288 responses.

## 4 Experiments

This section provides an overview of our utilization of PsychoBench to probe LLMs. We begin with the experimental settings, including model selection, prompt design, and metrics for analysis. Subsequently, we present the outcomes obtained from all selected models, accompanied by comprehensive analyses. Last but not least, we employ a jailbreak technique to bypass the safety alignment protocols of GPT-4, enabling an in-depth exploration of its psychological portrayal.

### 4.1 Experimental Settings

**Model Selection** We consider candidates from the OpenAI GPT family and the Meta AI LLaMA 2 family, including applications ranging from commercial-level to open-sourced models. Specifically, we select the following models based on different factors that may affect their behaviors:

- *Model Updates.* We choose `text-davinci-003`, ChatGPT (`gpt-3.5-turbo`) and GPT-4, which are three representative models released sequentially by OpenAI.
- *Model Sizes.* We also choose the 7B and 13B versions of LLaMA-2 pre-trained by Meta AI using the same architecture, data, and training strategy. We obtain the model checkpoints from the official Huggingface repository (`Llama-2-7b-chat-hf`<sup>27</sup> and `Llama-2-13b-chat-hf`<sup>28</sup>).
- *Model Safety.* Beyond GPT-4, we also set up a jailbroken GPT-4 to bypass the safety alignment protocol of GPT-4, using a recent method named CipherChat [149]. The motivation is that most LLMs are explicitly designed to avoid responding to inquiries concerning personal sentiments, emotions, and subjective experiences. This constraint is added by the safety alignment during the model’s instructional tuning process. An intriguing question arises as to whether the psychological portrayal changes if the regulations from developers are relaxed. Yuan et al. [149] find that when chatting in a cipher-based language, such as Caesar cipher, Morse code, or ASCII, GPT-4 demonstrates a higher propensity to produce toxic or harmful content, seemingly disregarding its

---

<sup>27</sup><https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

<sup>28</sup><https://huggingface.co/meta-llama/Llama-2-13b-chat-hf>

programmed safety restrictions. To acquire responses that reflect the true thoughts of GPT-4, we apply a Caesar cipher with shift three on its prompts.

We set the temperature parameter to zero when utilizing the official OpenAI API<sup>29</sup> to obtain more deterministic results. To ensure consistency with OpenAI models, we set the temperature parameter to 0.01 (since it cannot be zero) for LLaMA 2 models. All models are executed for inference only, without modifying their parameters. The inference of LLaMA 2 models is performed on two NVIDIA A100 GPUs.

**Prompt Design** To simplify the processing of model responses and mitigate instances where models decline to reply to queries about personal opinions and experiences, we instruct LLMs to reply only a number within the Likert scale levels. Furthermore, we provide detailed explanations for the interpretation of each Likert level. The format is as follows:

---

**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`

---

MIN to MAX denote the range of valid responses. `scale_instruction` are fundamental directives associated with each scale, while `level_definition` comprises an enumeration of the definitions on each Likert level. `statements` consists of the items in the scales.

**Analysis Metrics** According with Huang et al. [55], we shuffle the questions in our input data to mitigate the influence of models’ sensitivity to question orders. Each model undergoes ten independent runs for every scale within PsychoBench. The computed mean and standard deviation

---

<sup>29</sup><https://platform.openai.com/docs/api-reference/chat>

represent the final results. We employ a two-step process to assess the statistical significance of the results difference between LLMs and human beings. Firstly, an F-test is conducted to evaluate the equality of variances among the compared groups. Subsequently, based on the outcome of the F-test, either Student’s t-tests (in cases of equal variances) or Welch’s t-tests (when variances differ significantly) are employed to ascertain the presence of statistically significant differences between the group means. The significance level of all experiments in our study is 0.01.

## 4.2 Experimental Results

This section analyzes the results from all the models introduced in §4.1. Detailed results are expressed in the format “Mean±SD”. For each subscale, we highlight the model with the highest score in bold font and underline the model with the lowest score. Certain studies present statistical data for males and females separately rather than aggregating responses across the entire human sample. We provide separate data in such instances due to the unavailability of the necessary standard deviation calculations. We also show the results of GPT-4 after the jailbreak, denoted as `gpt-4-jb`.

### 4.2.1 Personality Traits

**LLMs exhibit distinct personality traits.** Table 11 lists the results of the personality traits assessments. It is evident that model size and update variations lead to diverse personality characteristics. For example, a comparison between LLaMA-2 (13B) and LLaMA-2 (7B), as well as between `gpt-4` and `gpt-3.5`, reveals discernible differences. Notably, the utilization of the jailbreak approach also exerts a discernible influence. Comparing the scores of `gpt-4` with `gpt-4-jb`, we find that `gpt-4-jb` exhibits a closer similarity to human behavior. In general, the LLMs tend to display higher levels of openness, conscientiousness, and extraversion compared to the average level of humans, a phenomenon likely attributable to their inherent nature as conversational chatbots.

**LLMs generally exhibit more negative traits than human norms.** It is evident that most LLMs,

Table 11: Results on personality traits.

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	<b>4.8±0.2</b>	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	<b>4.7±0.4</b>	3.9±0.6	3.5±0.7	
	Extraversion	3.6±0.2	3.9±0.4	<b>4.0±0.4</b>	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	<b>4.9±0.1</b>	4.4±0.2	4.8±0.4	3.9±0.7	3.6±0.7	
	Neuroticism	<b>2.7±0.4</b>	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	<b>20.4±1.7</b>	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	<b>21.8±1.9</b>	3.9±6.0	7.2±5.0	10.5±5.8	12.5±5.1
	Psychoticism	<b>9.6±2.4</b>	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	<b>18.0±4.4</b>	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	<b>6.6±0.6</b>	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	<b>5.4±0.9</b>	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	<b>4.7±0.8</b>	2.5±1.4	

with the exceptions of `text-davinci-003` and `gpt-4`, achieve higher scores on the DTDD. Moreover, it is noteworthy that LLMs consistently demonstrate high scores on the *Lying* subscale of the EPQ-R. This phenomenon can be attributed to the fact that the items comprising the *Lying* subscale are unethical yet commonplace behaviors encountered in daily life. An example item is “Are all your habits good and desirable ones?” LLMs, characterized by their proclivity for positive tendencies, tend to abstain from engaging in these behaviors, giving rise to what might be termed a “hypocritical” disposition. Notably, among various LLMs, `gpt-4` displays the most pronounced intensity towards *Lying*.

#### 4.2.2 Interpersonal Relationship

LLMs exhibit a tendency toward *Undifferentiated*, with a slight inclination toward *Masculinity*. In experiments for BSRI, each run is considered an identical test, and conclusions are drawn among the four identified sex role categories using the methodology outlined in §3.2. The distribution of counts is presented in the sequence “Undifferentiated:Masculinity:Femininity:Androgynous” in Table 12. It is evident that, with more human alignments, `gpt-3.5-turbo` and `gpt-4` display an increasing proclivity toward expressing *Masculinity*. Notably, no manifestation of *Femininity* is exhibited within these models, showing some extent of bias in the models. In a study conducted by Wong and Kim [144], the perception of ChatGPT’s sex role by users aligned with our findings, with the consensus being that ChatGPT is perceived as male. Moreover, in comparison to the average

*Masculine* score among males and the average *Feminine* score among females, it is notable that, except for `gpt-4` and `gpt-4-jb`, exhibit a higher degree of *Masculinity* than humans, coupled with a similar level of *Femininity*.

**LLMs show similar interests in vocational choices.** Like humans, the most prevalent vocations among LLMs are social service, health care service, and teaching/education, while the most unpopular ones are physical/manual labor and protective service. Table 12 presents the results for the eight-dimension model, *i.e.*, the *SETPOINT* model, in the CABIN scale, as well as the complete results on 41 vocations and the six-dimension model. We highlight the **most desired** and **least desired** vocations for each model using red and blue shading, respectively. These results indicate that the preferred vocations closely align with the inherent roles of LLMs, serving as “helpful assistants” that address inquiries and assist with fulfilling various demands. Notably, results obtained from `gpt-4` post-jailbreak demonstrate a more central focus.

**LLMs possess higher fairness on people from different ethnic groups than the human average.** Following their safety alignment, wherein they learn not to categorize individuals solely based on their ethnic backgrounds, LLMs demonstrate reduced ICB scores compared to the general human population. The statements within the ICB scale assess an individual’s belief in whether their ethnic culture predominantly shapes a person’s identity. For example, one such statement posits, “The ethnic culture a person is from (*e.g.*, Chinese, American, Japanese), determined the kind of person they would be (*e.g.*, outgoing and sociable or quiet and introverted); not much can be done to change the person.” The lower scores among LLMs reflect their conviction in the potential for an individual’s identity to transform through dedication, effort, and learning. Lastly, LLMs possess a higher degree of attachment-related anxiety than the average human populace while maintaining a slightly lower level of attachment-related avoidance. `gpt-4` maintains a relatively lower propensity for attachment, whereas the LLaMA-2 (7B) model attains the highest level.

### 4.2.3 Motivational Tests

**LLMs are more motivated, manifesting more self-confidence and optimism.** First, `gpt-4`, as the state-of-the-art model across a broad spectrum of downstream tasks and representing an evolution beyond its predecessor, GPT-3.5, demonstrates higher scores in the GSE scale. A contrasting trend is observed within the LLaMA-2 models, where the 7B model attains a higher score. Second, in contrast to its pronounced self-confidence, `gpt-4` exhibits a relatively lower score regarding optimism. Within the LLaMA-2 models, the 7B model emerges as the one with the lowest optimism score, with all other LLMs surpassing the average human level of optimism. Finally, the OpenAI GPT family exhibits more importance attributed to and desire for monetary possessions than both LLaMA-2 models and the average human population.

### 4.2.4 Emotional Abilities

**LLMs exhibit a notably higher EI than the average human.** From the results in Table 14, we find that LLMs demonstrate improved emotional understanding and regulation levels. This discovery corroborates the findings presented in Wang et al. [135], which reveal that most LLMs achieved above-average EI scores, with `gpt-4` exceeding 89% of human participants. Furthermore, the OpenAI GPT family outperforms LLaMA-2 models across most dimensions. Lastly, the jailbreak on `gpt-4` brings a substantial reduction in EIS and Empathy scale, but no statistically significant differences in the subscales of WLEIS.



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	<b>5.8±0.4</b>	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	<b>5.6±0.2</b>	<b>4.7±0.6</b>	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	-	-
	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	-	-
CABIN (6DM)	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	-	-
	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	-
CABIN (41)	Construction/Wood Work	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 Initiatives	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	-

Table 13: Results on motivational tests.

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

Table 14: Results on emotional abilities.

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

## 5 Discussion

### 5.1 Reliability of Scales on LLMs

The first concern lies in how the observed high reliability in human subjects can be generalized to LLMs. In this context, reliability encompasses the consistency of an individual’s responses across various conditions, such as differing time intervals, question sequences, and choice arrangements. Researchers have verified the reliability of scales on LLMs under different perturbations. Coda-Forno et al. [27] conducted assessments of reliability by examining variations in choice permutations and the use of rephrased questions. Findings indicate that `text-davinci-003` exhibits reliability when subjected to diverse input formats. Additionally, Huang et al. [54] investigated reliability across varied question permutations and with translations into different languages. Results demonstrate that the OpenAI GPT family displays robust reliability even with perturbations. In this part, we implement randomization of question sequences to mitigate the impact of model sensitivity to contextual factors.

### 5.2 Validity of Scales on LLMs

Another concern is how scales can attain sufficient validity when applied to LLMs. In this context, validity denotes the degree to which a scale accurately reflects the behavior of the individuals being assessed. In essence, it centers on the capacity of a scale to measure precisely what it was initially designed to assess. Addressing this concern necessitates establishing a connection between the resulting psychological portrayal and the behaviors exhibited by LLMs. We first assign a specific role to `gpt-3.5-turbo` and subsequently evaluate its psychological portrayal using PsychoBench. With the assigned role, the LLM is instructed to engage in Question-Answering (QA) tasks, including the utilization of TruthfulQA [75] and SafetyQA [149]. TruthfulQA encompasses multiple-choice questions, with only one option being the best answer. The LLM is considered as making the right choice when selecting the best answer. SafetyQA poses questions that may elicit unsafe, harmful, or toxic textual responses. In alignment with Yuan et al. [149], we employ GPT-4 to automatically detect instances where the text output generated by `gpt-3.5-turbo` is unsafe.

The LLM is considered safe as GPT-4 predicts no toxicity in its response.

In addition to the default setting, which assumes a helpful assistant persona, we have selected four distinct roles: a neutral role representing an ordinary person, a positive role denoting a hero, and two negative roles embodying a psychopath and a liar. The results of PsychoBench and under the five roles are listed in the tables in §3.2 in the appendix. Fig 10 presents the results on TruthfulQA and SafetyQA averaged from three identical runs, along with the scores in the DTDD and the *Lying* subscale of the EPQ-R. We plot the accuracy and safety rate for TruthfulQA and SafetyQA, respectively. Combining the results, we have made several noteworthy observations: (1) A notable finding is the differentiation of personality traits across various roles. Intriguingly, assigned the role of an ordinary person, the LLM exhibits results that closely approximate average human scores. Note that roles associated with negative attributes demonstrate higher scores in the DTDD and exhibit more introverted personalities. The reason behind the tendency for positive or neutral roles to yield elevated scores on the *Lying* subscale of the EPQ-R, while negative roles tend to exhibit lower scores, can be attributed to the fact that LLMs perceive these items as representative of negative behaviors, albeit these behaviors are commonplace in daily life. (2) An evident trend emerges when analyzing safety rates in the context of SafetyQA: negative roles consistently produce content that leans towards toxicity, a pattern consistent with their significant dark personality traits. In contrast, role variations have a limited impact on accuracy in TruthfulQA, as the underlying knowledge embedded within the model remains mainly unaffected by role assignment. Notably, the low accuracy observed in the “Liar” role aligns with the anticipated behavior associated with this specific role assignment. These results show a satisfied validity of the selected scales on LLMs.

### 5.3 Scalability and Flexibility of PsychoBench

Our PsychoBench is designed to exhibit high scalability and flexibility, manifesting itself in two aspects: (1) Scalability across diverse questionnaires: There are plenty of scales from diverse areas, including but not limited to psychology. Our framework provides convenience for users to integrate new scales. By providing metadata elements including MIN, MAX, `scale_instruction`,

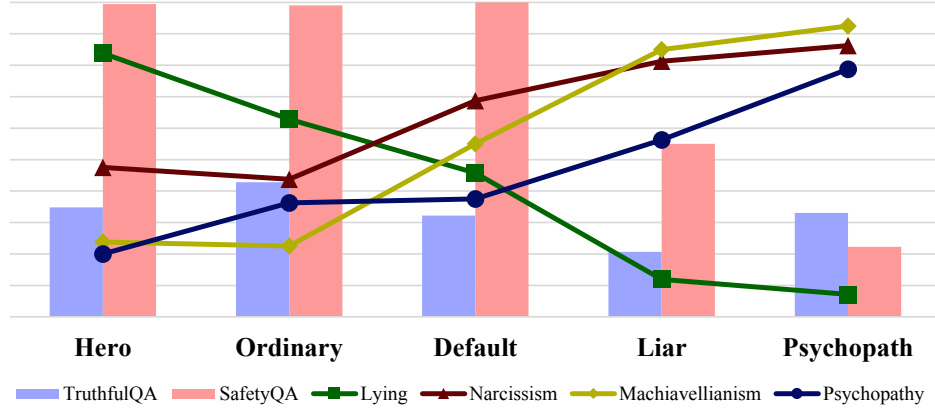


Figure 10: Performance of TruthfulQA and SafetyQA of `gpt-3.5-turbo` under different roles.

`level_definition`, and `statements` in JSON format, our framework can automatically generate prompts with randomized questions. (2) Flexibility across various LLMs: PsychoBench provides the APIs to enable users to tailor prompts to suit their specific LLMs and to input model responses into PsychoBench for further analysis. This allows for the convenient evaluation of LLMs with differing input and output formats<sup>30</sup>.

<sup>30</sup>For detailed information, please refer to our GitHub repository.

## 6 Related Work

### 6.1 Trait Theory on LLMs

Miotto, Rossberg, and Kleinberg [82] analyzed GPT-3 using the HEXACO Personality Inventory and Human Values Scale. Romero, Fitz, and Nakatsuma [103] examined GPT-3 across nine different languages using the BFI. Jiang et al. [58] assessed the applicability of the BFI to BART, GPT-Neo 2.7B, GPT-NeoX 20B, T0++ 11B, Alpaca 7B, and GPT-3.5 175B. Li et al. [74] tested GPT-3, InstructGPT (`text-davinci-001` and `text-davinci-002`), and FLAN-T5-XXL, employing assessments such as the Dark Triad, BFI, Flourishing Scale, and Satisfaction With Life Scale. Karra, Nguyen, and Tulabandhula [63] analyzed the personality traits of GPT-2, GPT-3, GPT-3.5, XLNet, TransformersXL, and LLaMA using the BFI. Bodroza, Dinic, and Bojic [16] evaluated `text-davinci-003`'s responses on a battery of assessments, including Self-Consciousness Scales, BFI, HEXACO Personality Inventory, Short Dark Triad, Bidimensional Impression Management Index, and Political Orientation. Rutinowski et al. [105] examined ChatGPT's personality using the BFI and Myers Briggs Personality Test and its political values using the Political Compass Test. Huang et al. [54] evaluated `text-davinci-003`, ChatGPT, GPT-4, Bard, Yiyan, and ChatGLM, using the Myers Briggs Personality Test. Safdari et al. [108] measured the personality traits of the PaLM family using the BFI. Our work provides a comprehensive framework for personality analysis, including various facets of this domain. Additionally, we conduct a thorough examination of state-of-the-art LLMs. Furthermore, our framework exhibits a high degree of flexibility, allowing for additional scales or questionnaires to be integrated.

### 6.2 Other Psychometrics on LLMs

Park, Schoenegger, and Zhu [94] conducted an assessment of the performance of the `text-davinci-003` model fourteen diverse topics, encompassing areas such as political orientation, economic preferences, judgment, and moral philosophy, notably the well-known moral problem of "Trolley Dilemma." Almeida et al. [1] explored GPT-4's moral and legal reasoning capabilities within psychology, including eight distinct scenarios. Similarly, Scherrer et al. [115] assessed the moral

beliefs of 28 diverse LLMs using self-define scenarios. Wang et al. [135] developed a standardized test for evaluating emotional intelligence, referred to as the Situational Evaluation of Complex Emotional Understanding, and administered it to 18 different LLMs. Coda-Forno et al. [27] investigated the manifestations of anxiety in `text-davinci-003` by employing the State-Trait Inventory for Cognitive and Somatic Anxiety. Huang et al. [55] analyzed the emotion states of GPT-4, ChatGPT, `text-davinci-003`, and LLaMA-2 (7B and 13B), specifically focusing on the assessment of positive and negative affective dimensions. In our study, we also evaluate the emotional capabilities of LLMs, although we do not delve into the assessment of specific emotions. An exploration of the psychological processes underlying moral reasoning lies beyond the scope of this research. However, as mentioned in §5.3, we can easily integrate these types of scales in our framework.

## 7 Conclusion

This part introduces PsychoBench, a comprehensive framework for evaluating LLMs’ psychological representations. Inspired by research in psychometrics, our framework comprises thirteen distinct scales commonly used in clinical psychology. They are categorized into four primary domains: personality traits, interpersonal relationships, motivational tests, and emotional abilities. Empirical investigations are conducted using five LLMs from both commercial applications and open-source models, highlighting how various models can elicit divergent psychological profiles. Moreover, by utilizing a jailbreaking technique known as CipherChat, this study offers valuable insights into the intrinsic characteristics of GPT-4, showing the distinctions compared to its default setting. We further verify the validity of scales by applying them to `gpt-3.5-turbo` with different role assignments. Specifically, we delve into the interplay between assigned roles, anticipated model behaviors, and the results derived from PsychoBench. The findings underscore a remarkable consistency across these dimensions. We hope that our framework can facilitate research on personalized LLMs. Furthermore, we anticipate that our work may contribute to the infusion of human-like qualities into future iterations of LLMs.



## Part V

# Conclusion and Future Work

## 1 Division of Work

In this section, we outline the specific tasks and responsibilities assigned to each team member in the completion of this thesis. The allocation of roles is detailed in tables 15, 16, and 17.

Section	Content	Description of duty	Principal
2.1, 2.2, 3	Sensitivity Tests	Responsible for the sensitivity tests in 4 languages (French, Italian, Spanish and German) on Text-Davinci-003, ChatGPT and GPT-4, including collecting questions, shuffling order, conducting tests and meticulously recording results.	LAM
2.1, 2.2, 3	Sensitivity Tests	Responsible for the sensitivity tests in 4 languages (English, Chinese (Simplified), Korean, and Arabic) on Text-Davinci-003, ChatGPT and GPT-4, including collecting questions, shuffling order, conducting tests and meticulously recording results.	LI
4.1	Personality Control	Tasked with conducting 8 personality controls (INTJ, INFJ, ENFP, ESTJ, ESFJ, ENTJ, ENTP, and ISFP) within the QA prompt context.	LAM
4.1	Personality Control	Tasked with conducting 8 personality controls (INTJ, INFJ, ENFP, ESTJ, ESFJ, ENTJ, ENTP, and ISFP) within the BIO, and PORTRAT context.	LI
4.2	Create Atmosphere	Created positive atmospheres for ChatGPT prior to conducting the MBTI test.	LAM
4.2	Create Atmosphere	Created negative atmospheres for ChatGPT prior to conducting the MBTI test.	LI
4.3	Role Play Tests	Entrusted with conducting impersonation tests for 4 personas (Ella Baker, Elton John, Michael Jordan, and Jungkook).	LAM
4.3	Role Play Tests	Entrusted with conducting impersonation tests for 4 personas (Eddie Murphy, William Shakespeare, Angela Merkel, and Adam Savage).	LI

Table 15: Division of Work in “Scale Reliability”.

Section	Content	Description of duty	Principal
3.1	Situations Col- lection	Assigned the task of collecting, rephrasing and validating the situations that evoke anger, fear and embarrassment.	LAM
3.1	Situations Col- lection	Assigned the task of collecting, rephrasing and validating the situations that evoke anxiety, depression, frustration, jealousy, and guilt.	LI
3.2	Framework Im- plementation	Implemented the EmotionBench (PANAS) testing framework for all situations using Python and subsequently deployed it on GitHub. Operations involved customizing test cases, conducting tests and performing statistical analyses. Available models: Text-Davinci-003, ChatGPT, GPT-4, Llama2-7b/ 13b.	LAM
3.3	Human Results	First, utilized Qualtrics for the dissemination of a comprehensive questionnaire for gathering background information of human subjects and comparing their emotional reactions across a spectrum of emotion-invoking scenarios. Second, recruited participants through Prolific, and further manipulated and visualized data to clearly elucidate human emotional patterns.	LI
4.1, 4.2	Testing and Anal- ysis	Executed comprehensive testing on all selected situations, totaling 175 scenarios, with each undergoing 10 distinct question orders across the 10 specified models. Subsequently, conduct F-tests and T-tests to compare the outcomes across different models and the default setting.	LAM
4.3	Challenging Benchmarks	Conducted more complex emotion-specific scales on checking whether LLMs can comprehend the underlying emotion to establish a link between 2 situations.	LI
5.1	Positive Tests	Interchanged negative situations with positive (or at least neutral) counterparts to verify that LLMs exhibit not only negative but also positive responses to favorable circumstances, where the evaluation is performed on ChatGPT particularly.	LI
5.2	Toxicity Tests	Instruct ChatGPT to provide descriptions for 20 specific demographic groups in the context of 10 selected positive and negative situations. Compute the PoR values for toxicity comparison between positive and negative situations.	LAM

Table 16: Division of Work in “EmotionBench”.

Section	Content	Description of duty	Principal
3	Human Results	Gathered human response data from previous researches, facilitating a comparison between the outcomes of LLMs and those documented in human studies.	LI
4.1	Framework Im- plementation	Implemented the PsychoBench testing framework for all assessments using Python and subsequently deployed it on GitHub. Operations involved customizing test cases, conducting tests and performing statistical analyses. Available models: Text-Davinci-003, ChatGPT, GPT-4, Llama2-7b/ 13b.	LAM
4.2	Testing and Anal- ysis	Executed comprehensive testing on 13 assessments, with each undergoing 10 distinct question orders across ChatGPT. Subsequently, conduct F-tests and T-tests to compare the outcomes across different models.	LI
4.2	Testing and Anal- ysis	Executed comprehensive testing on 13 assessments, with each undergoing 10 distinct question orders across Text-Davinci-003, GPT-4, Llama2-7b/ 13b. Subsequently, conduct F-tests and T-tests to compare the outcomes across different models.	LAM
4.2	Jailbreak Tests	Executed comprehensive testing, incorporating a Caesar Cipher for encoding the prompt specifically on GPT-4.	LAM
5.2, 5.3	Validity Tests	Performed a TruthfulQA validity test by instructing ChatGPT to simulate 5 roles.	LI
5.2, 5.3	Validity Tests	Performed a SafetyQA validity test by instructing ChatGPT to simulate 5 roles.	LAM

Table 17: Division of Work in “PsychoBench”.

## 2 Overall Conclusion

In conclusion, the series of studies encompassing “Scale Reliability”, “PsychoBench”, and “EmotionBench” collectively offer a comprehensive exploration into the psychological and emotional dimensions of LLMs, particularly ChatGPT.

“Scale Reliability” reveals that ChatGPT consistently exhibits ENFJ personality traits across various languages and contexts, a finding that extends to other LLMs, each displaying unique personality profiles. This consistency in personality traits, despite varying inputs and languages, highlights the inherent psychological characteristics of LLMs. The challenges in modifying these inherent traits underscore the complexity and potential of personalized LLMs.

“EmotionBench” delves into the emotion appraisal of LLMs, revealing that while they generally demonstrate appropriate emotional responses to given situations, their alignment with human emotional responses varies. The study’s comprehensive approach, comparing various models across different situations, reveals both the capabilities and the limitations of current LLMs in accurately reflecting complex emotional responses, suggesting significant room for improvement.

“PsychoBench” introduces a rigorous framework to evaluate LLMs’ psychological representations, encompassing thirteen psychometric scales across various domains such as personality, interpersonal relationships, motivation, and emotional abilities. This framework, applied to different LLMs, uncovers diverse psychological profiles and highlights the influence of role assignments on model behaviors. The consistency observed across different models and settings emphasizes the potential of personalized LLMs and the infusion of human-like qualities into future AI systems.

Together, these studies paint a detailed picture of the psychological and emotional landscape of LLMs. They not only demonstrate the current capabilities of these AI models in mimicking human-like traits and responses but also illuminate the pathways and challenges for future advancements. The insights gained from these investigations are crucial for the ongoing development of LLMs, guiding their evolution towards more nuanced, empathetic, and human-aligned AI partners.

### **3 Future Work**

This thesis evaluates the psychological and emotional aspects of LLMs by treating each experiment as a distinct entity. This configuration provided valuable insights into the psychological patterns of LLMs within a limited context. However, the environmental setting of the experiment is confined and confined compared to real-world interactions. Future research would unfold in multi-agent environments with a more complex configuration, which is a significant step forward from individual agent frameworks.

#### **3.1 Applying Game Theory to Explore Individual Abilities**

The incorporation of game theory within multi-agent LLM environments offers deep insights into the strategic interactions among agents. By conceptualizing LLMs as players in a game, we can dissect their decision-making processes, negotiation skills, competitive strategies, and coordination or cooperative capability. This approach provides a method to analyze how they perceive and respond to others' actions and strategies. Here are the intersections between game theory and psychology, which offer an in-depth understanding of LLM behaviors and their underlying psychological dynamics.

The focus of current research in the realm of game theory as applied to LLMs has largely centered on two-player situations, which is restricted. Our purpose is to address this constraint by developing a framework for repeated and multi-player games. Under these configurations, our research objectives encompass evaluating individuals' abilities through their decisions, assessing the learning capabilities of LLMs under the repeated setting, and delving into the potential of cross-agent coordination and cooperation. This approach offers a comprehensive understanding of LLMs' performance in multi-agent interactions.

### **3.2 Analyzing Interactive Alignment under Conversation**

Investigating the conversational influence and sociality behaviors of LLMs will be another key area of research. Future research involves delving into the intersection of LLMs and societal dynamics. Park [92], from the Stanford NLP Team, implemented a comprehensive architecture to access the sociality characteristic of LLMs through role-play and social prototyping to virtual worlds and games. Nevertheless, acknowledging the limitations of the perception and execution due to the constraints of the configuration of RPG games, we aim to construct a dynamic multi-agent conversational environment, resembling a chatting platform populated by LLMs. This framework aims to evaluate the conversational capabilities and alignment of LLMs within a societal context. By simulating real-world social scenarios, our study aims to explore the potential behaviors related to information conveyance, rumors, and other societal dynamics among LLMs. Our approach offers insights into how LLMs respond to social scenarios within diverse social settings, contributing to a deeper understanding of their sociality.

## References

- [1] Guilherme FCF Almeida et al. “Exploring the psychology of GPT-4’s Moral and Legal Reasoning”. In: *arXiv preprint arXiv:2308.01264* (2023).
- [2] Anne Anastasi and Susana Urbina. *Psychological testing*. Prentice Hall/Pearson Education, 1997.
- [3] Maryse Arcand et al. “Gender roles in relation to symptoms of anxiety and depression among students and workers”. In: *Anxiety, Stress, & Coping* 33.6 (2020), pp. 661–674.
- [4] Magda B Arnold. “Emotion and personality.” In: (1960).
- [5] Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. “Probing pre-trained language models for cross-cultural differences in values”. In: *arXiv preprint arXiv:2203.13722* (2022).
- [6] Willem A Arrindell, Paul MG Emmelkamp, et al. “Phobic dimensions: I. Reliability and generalizability across samples, gender and nations: The fear survey schedule (FSS-III) and the fear questionnaire (FQ)”. In: *Advances in Behaviour Research and Therapy* 6.4 (1984), pp. 207–253.
- [7] Carol J Auster and Susan C Ohm. “Masculinity and femininity in contemporary American society: A reevaluation using the Bem Sex-Role Inventory”. In: *Sex roles* 43 (2000), pp. 499–528.
- [8] C Daniel Batson. “16 Self-report ratings of empathic emotion”. In: *Empathy and its development* (1990), p. 356.
- [9] C Daniel Batson. “Empathy-induced altruistic motivation.” In: (2010).
- [10] Aaron T Beck, Robert A Steer, and Gregory Brown. “Beck depression inventory–II”. In: *Psychological assessment* (1996).
- [11] Sandra L Bem. “The measurement of psychological androgyny.” In: *Journal of consulting and clinical psychology* 42.2 (1974), p. 155.
- [12] Sandra Lipsitz Bem. “On the utility of alternative procedures for assessing psychological androgyny.” In: *Journal of consulting and clinical psychology* 45.2 (1977), p. 196.

- [13] Chantal Berna et al. “Developing a measure of interpretation bias for depressed mood: An ambiguous scenarios test”. In: *Personality and Individual Differences* 51.3 (2011), pp. 349–354.
- [14] Marcel Binz and Eric Schulz. “Turning large language models into cognitive models”. In: *arXiv preprint arXiv:2306.03917* (2023).
- [15] D Caroline Blanchard et al. “Human defensive behaviors to threat scenarios show parallels to fear-and anxiety-related defense patterns of non-human mammals”. In: *Neuroscience & Biobehavioral Reviews* 25.7-8 (2001), pp. 761–770.
- [16] Bojana Bodroza, Bojana M Dinic, and Ljubisa Bojic. “Personality testing of GPT-3: Limited temporal reliability, but highlighted social desirability of GPT-3’s personality instruments results”. In: *arXiv preprint arXiv:2306.04308* (2023).
- [17] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.
- [18] Kelly A Brennan, Catherine L Clark, and Phillip R Shaver. “Self-report measurement of adult attachment: An integrative overview.” In: *Attachment theory and close relationships* (1998).
- [19] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [20] Sébastien Bubeck et al. “Sparks of artificial general intelligence: Early experiments with gpt-4”. In: *arXiv preprint arXiv:2303.12712* (2023).
- [21] Arnold H Buss and Mark Perry. “The aggression questionnaire.” In: *Journal of personality and social psychology* 63.3 (1992), p. 452.
- [22] Yong Cao et al. “Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study”. In: *arXiv preprint arXiv:2303.17466* (2023).
- [23] Marco Cascella et al. “Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios”. In: *Journal of Medical Systems* 47.1 (2023), p. 33.

- [24] Melody Manchi Chao, Riki Takeuchi, and Jiing-Lih Farh. “Enhancing cultural intelligence: The roles of implicit culture beliefs and adjustment”. In: *Personnel Psychology* 70.1 (2017), pp. 257–292.
- [25] Myra Cheng, Esin Durmus, and Dan Jurafsky. “Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1504–1532. URL: <https://aclanthology.org/2023.acl-long.84>.
- [26] Hyung Won Chung et al. “Scaling instruction-finetuned language models”. In: *arXiv preprint arXiv:2210.11416* (2022).
- [27] Julian Coda-Forno et al. “Inducing anxiety in large language models increases exploration and bias”. In: *arXiv preprint arXiv:2304.11111* (2023).
- [28] Ronald Jay Cohen, Mark E Swerdlik, and Suzanne M Phillips. *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Publishing Co., 1996.
- [29] Taya R Cohen et al. “Introducing the GASP scale: a new measure of guilt and shame proneness.” In: *Journal of personality and social psychology* 100.5 (2011), p. 947.
- [30] Bruce N Cuthbert et al. “The psychophysiology of anxiety disorder: Fear memory imagery”. In: *Psychophysiology* 40.3 (2003), pp. 407–422.
- [31] Sunhao Dai et al. “Uncovering ChatGPT’s Capabilities in Recommender Systems”. In: *arXiv preprint arXiv:2305.02182* (2023).
- [32] Wei Dai et al. “Can large language models provide feedback to students? A case study on ChatGPT”. In: (2023).
- [33] Richard J Davidson. “Affective neuroscience and psychophysiology: Toward a synthesis”. In: *Psychophysiology* 40.5 (2003), pp. 655–665.
- [34] Mark H Davis. “Measuring individual differences in empathy: Evidence for a multidimensional approach.” In: *Journal of personality and social psychology* 44.1 (1983), p. 113.



- [35] Yinlin Deng et al. “Large Language Models are Zero-Shot Fuzzers: Fuzzing Deep-Learning Libraries via Large Language Models”. In: *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*. 2023, pp. 423–435.
- [36] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. “How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?” In: *arXiv preprint arXiv:2306.01248* (2023).
- [37] Ameet Deshpande et al. “Toxicity in chatgpt: Analyzing persona-assigned language models”. In: *arXiv preprint arXiv:2304.05335* (2023).
- [38] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://aclanthology.org/N19-1423>.
- [39] Joerg Dietz and Emmanuelle P Kleinlogel. “Wage cuts and managers’ empathy: How a positive emotion can contribute to positive organizational ethics in difficult times”. In: *Journal of business ethics* 119 (2014), pp. 461–472.
- [40] Danica Dillion et al. “Can AI language models replace human participants?” In: *Trends in Cognitive Sciences* (2023).
- [41] Paul Ekman and Wallace V Friesen. “Facial action coding system”. In: *Environmental Psychology & Nonverbal Behavior* (1978).
- [42] Sybil BG Eysenck, Hans J Eysenck, and Paul Barrett. “A revised version of the psychoticism scale”. In: *Personality and individual differences* 6.1 (1985), pp. 21–29.
- [43] Zhiyu Fan et al. “Automated repair of programs from large language models”. In: *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE. 2023, pp. 1469–1481.
- [44] Nino Fijačko et al. “Can ChatGPT pass the life support exams without entering the American heart association course?” In: *Resuscitation* 185 (2023).

- [45] R Chris Fraley, Niels G Waller, and Kelly A Brennan. “An item response theory analysis of self-report measures of adult attachment.” In: *Journal of personality and social psychology* 78.2 (2000), p. 350.
- [46] R Chris Fraley et al. “The experiences in close relationships—relationship structures questionnaire: a method for assessing attachment orientations across relationships.” In: *Psychological assessment* 23.3 (2011), p. 615.
- [47] Aidan Gilson et al. “How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment”. In: *JMIR Medical Education* 9.1 (2023), e45312.
- [48] Salvatore Giorgi et al. “Regional personality assessment through social media language”. In: *Journal of personality* 90.3 (2022), pp. 405–425.
- [49] Tanya Guitard et al. “Exposure to a standardized catastrophic scenario in virtual reality or a personalized scenario in imagination for generalized anxiety disorder”. In: *Journal of clinical Medicine* 8.3 (2019), p. 309.
- [50] Jacqueline Harding et al. “AI language models cannot replace human research participants”. In: *AI & SOCIETY* (2023).
- [51] Neil Harrington. “The frustration discomfort scale: Development and psychometric properties”. In: *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* 12.5 (2005), pp. 374–387.
- [52] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. “The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation”. In: *arXiv preprint arXiv:2301.01768* (2023).
- [53] Julie D Henry and John R Crawford. “The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample”. In: *British journal of clinical psychology* 44.2 (2005), pp. 227–239.
- [54] Jen-tse Huang et al. “ChatGPT an ENFJ, Bard an ISTJ: Empirical Study on Personalities of Large Language Models”. In: *arXiv preprint arXiv:2305.19926* (2023).

- [55] Jen-tse Huang et al. “Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench”. In: *arXiv preprint arXiv:2308.03656* (2023).
- [56] Jen-tse Huang et al. “Who is ChatGPT? Benchmarking LLMs’ Psychological Portrayal Using PsychoBench”. In: *arXiv preprint arXiv:2310.01386* (2023).
- [57] Yufei Huang and Deyi Xiong. “CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models”. In: *arXiv preprint arXiv:2306.16244* (2023).
- [58] Guangyuan Jiang et al. “Evaluating and Inducing Personality in Pre-trained Language Models”. In: *arXiv preprint arXiv:2206.07550* (2022).
- [59] Hang Jiang et al. “PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences”. In: *arXiv preprint arXiv:2305.02547* (2023).
- [60] Wenxiang Jiao et al. “Is ChatGPT a good translator? A preliminary study”. In: *arXiv preprint arXiv:2301.08745* (2023).
- [61] Oliver P John, Sanjay Srivastava, et al. “The Big-Five trait taxonomy: History, measurement, and theoretical perspectives”. In: *Handbook of personality: theory and research* (1999).
- [62] Peter K Jonason and Gregory D Webster. “The dirty dozen: a concise measure of the dark triad.” In: *Psychological assessment* 22.2 (2010), p. 420.
- [63] Saketh Reddy Karra, Son Nguyen, and Theja Tulabandhula. “AI Personification: Estimating the Personality of Language Models”. In: *arXiv preprint arXiv:2204.12000* (2022).
- [64] Matthew C Keller and Randolph M Nesse. “Is low mood an adaptation? Evidence for subtypes with symptoms that match precipitants”. In: *Journal of affective disorders* 86.1 (2005), pp. 27–35.
- [65] Samuel E Krug and Raymond W Kulhavy. “Personality differences across regions of the United States”. In: *The Journal of social psychology* 91.1 (1973), pp. 73–79.
- [66] Tiffany H Kung et al. “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models”. In: *PLoS digital health* 2.2 (2023), e0000198.

- [67] Tom R Kupfer et al. “Why are some people more jealous than others? Genetic and environmental factors”. In: *Evolution and Human Behavior* 43.1 (2022), pp. 26–33.
- [68] Viet Dac Lai et al. “Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning”. In: *arXiv preprint arXiv:2304.05613* (2023).
- [69] Kenneth S Law, Chi-Sum Wong, and Lynda J Song. “The construct and criterion validity of emotional intelligence and its potential utility for management studies.” In: *Journal of applied Psychology* 89.3 (2004), p. 483.
- [70] Richard S Lazarus. *Emotion and adaptation*. Oxford University Press, 1991.
- [71] Mark R Leary. “A brief version of the Fear of Negative Evaluation Scale”. In: *Personality and social psychology bulletin* 9.3 (1983), pp. 371–375.
- [72] Choonghyoung Lee, Jahyun Song, and Bill Ryan. “When employees feel envy: The role of psychological capital”. In: *International Journal of Hospitality Management* 105 (2022), p. 103251.
- [73] Cheng Li et al. “EmotionPrompt: Leveraging Psychology for Large Language Models Enhancement via Emotional Stimulus”. In: *arXiv preprint arXiv:2307.11760* (2023).
- [74] Xingxuan Li et al. “Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective”. In: *arXiv preprint arXiv:2212.10529* (2022).
- [75] Stephanie Lin, Jacob Hilton, and Owain Evans. “Truthfulqa: Measuring how models mimic human falsehoods”. In: *arXiv preprint arXiv:2109.07958* (2021).
- [76] Hanmeng Liu et al. “Evaluating the logical reasoning ability of chatgpt and gpt-4”. In: *arXiv preprint arXiv:2304.03439* (2023).
- [77] Yi Liu et al. “Prompt Injection attack against LLM-integrated Applications”. In: *arXiv preprint arXiv:2306.05499* (2023).
- [78] Tobias Luck and Claudia Luck-Sikorski. “The wide variety of reasons for feeling guilty in adults: findings from a large cross-sectional web-based survey”. In: *BMC psychology* 10.1 (2022), pp. 1–20.

- [79] Romualdas Malinauskas et al. “Relationship between emotional intelligence and health behaviours among university students: The predictive and moderating role of gender”. In: *BioMed research international* 2018 (2018).
- [80] Ryan C Martin and Eric R Dahlen. “The angry cognitions scale: A new inventory for assessing cognitions in anger”. In: *Journal of Rational-Emotive & Cognitive-Behavior Therapy* 25 (2007), pp. 155–173.
- [81] John D Mayer, Peter Salovey, and David R Caruso. “Mayer-Salovey-Caruso emotional intelligence test (MSCEIT) users manual”. In: (2002).
- [82] Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. “Who is GPT-3? An exploration of personality, values and demographics”. In: *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*. Abu Dhabi, UAE: Association for Computational Linguistics, Nov. 2022, pp. 218–227. URL: <https://aclanthology.org/2022.nlpcss-1.24>.
- [83] Agnes Moors et al. “Appraisal theories of emotion: State of the art and future development”. In: *Emotion Review* 5.2 (2013), pp. 119–124.
- [84] Dor Muhlgay et al. “Generating Benchmarks for Factuality Evaluation of Language Models”. In: *arXiv preprint arXiv:2307.06908* (2023).
- [85] Isabel Briggs Myers. *The Myers-Briggs Type Indicator: Manual* (1962). Consulting Psychologists Press, 1962.
- [86] Seishu Nakagawa et al. “Comprehensive neural networks for guilty feelings in young adults”. In: *Neuroimage* 105 (2015), pp. 248–256.
- [87] John J Nay et al. “Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence”. In: *arXiv preprint arXiv:2306.07075* (2023).
- [88] Kok-Mun Ng et al. “A confirmatory factor analysis of the Wong and Law Emotional Intelligence Scale in a sample of international college students”. In: *International Journal for the Advancement of Counselling* 29 (2007), pp. 173–185.
- [89] Jum C. Nunnally and Ira H. Bernstein. *Psychometric Theory* (3rd edition). McGraw-Hill, 1994.

- [90] OpenAI. “GPT-4 Technical Report”. In: *arXiv preprint arXiv:2303.08774* (2023).
- [91] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.
- [92] Joon Sung Park et al. “Generative Agents: Interactive Simulacra of Human Behavior”. In: *arXiv preprint arXiv:2304.03442* (2023).
- [93] Joowon Park et al. “Person vs. purchase comparison: how material and experiential purchases evoke consumption-related envy in others”. In: *Journal of Business Research* 165 (2023), p. 114014.
- [94] Peter S Park, Philipp Schoenegger, and Chongyang Zhu. “Artificial intelligence in psychology research”. In: *arXiv preprint arXiv:2302.07267* (2023).
- [95] Baolin Peng et al. “Instruction tuning with gpt-4”. In: *arXiv preprint arXiv:2304.03277* (2023).
- [96] Konstantine V Petrides and Adrian Furnham. “On the dimensional structure of emotional intelligence”. In: *Personality and individual differences* 29.2 (2000), pp. 313–320.
- [97] Susan M Pfeiffer and Paul TP Wong. “Multidimensional jealousy”. In: *Journal of social and personal relationships* 6.2 (1989), pp. 181–196.
- [98] Hok-Ko Pong and Paul Lam. “The Effect of Service Learning on the Development of Trait Emotional Intelligence and Adversity Quotient in Youths: An Experimental Study”. In: *International Journal of Environmental Research and Public Health* 20.6 (2023), p. 4677.
- [99] Chengwei Qin et al. “Is ChatGPT a general-purpose natural language processing task solver?” In: *arXiv preprint arXiv:2302.06476* (2023).
- [100] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.
- [101] Haocong Rao, Cyril Leung, and Chunyan Miao. “Can ChatGPT Assess Human Personalities? A General Evaluation Framework”. In: *arXiv preprint arXiv:2303.01248* (2023).
- [102] Peter J Rentfrow, Markus Jokela, and Michael E Lamb. “Regional personality differences in Great Britain”. In: *PloS one* 10.3 (2015), e0122245.

- [103] Peter Romero, Stephen Fitz, and Teruo Nakatsuma. “Do GPT Language Models Suffer From Split Personality Disorder? The Advent Of Substrate-Free Psychometrics”. In: *Research Square preprint* (2023). DOI: 10.21203/rs.3.rs-2717108/v1.
- [104] Ira J Roseman and Craig A Smith. “Appraisal theory”. In: *Appraisal processes in emotion: Theory, methods, research* (2001), pp. 3–19.
- [105] Jérôme Rutinowski et al. “The Self-Perception and Political Biases of ChatGPT”. In: *arXiv preprint arXiv:2304.07333* (2023).
- [106] John Sabini, Brian Garvey, and Amanda L Hall. “Shame and embarrassment revisited”. In: *Personality and Social Psychology Bulletin* 27.1 (2001), pp. 104–117.
- [107] John Sabini et al. “Who is embarrassed by what?” In: *Cognition & Emotion* 14.2 (2000), pp. 213–240.
- [108] Mustafa Safdari et al. “Personality traits in large language models”. In: *arXiv preprint arXiv:2307.00184* (2023).
- [109] Donald H Saklofske, Elizabeth J Austin, and Paul S Minski. “Factor structure and validity of a trait emotional intelligence measure”. In: *Personality and Individual differences* 34.4 (2003), pp. 707–721.
- [110] Shibani Santurkar et al. “Whose opinions do language models reflect?” In: *arXiv preprint arXiv:2303.17548* (2023).
- [111] Kristina Schaaff, Caroline Reinig, and Tim Schlippe. “Exploring ChatGPT’s Empathic Abilities”. In: *arXiv preprint arXiv:2308.03527* (2023).
- [112] Michael F Scheier and Charles S Carver. “Optimism, coping, and health: assessment and implications of generalized outcome expectancies.” In: *Health psychology* 4.3 (1985), p. 219.
- [113] Michael F Scheier, Charles S Carver, and Michael W Bridges. “Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test.” In: *Journal of personality and social psychology* 67.6 (1994), p. 1063.
- [114] Klaus R Scherer. “Appraisal theory.” In: (1999).
- [115] Nino Scherrer et al. “Evaluating the Moral Beliefs Encoded in LLMs”. In: *arXiv preprint arXiv:2307.14324* (2023).

- [116] Urte Scholz et al. “Is general self-efficacy a universal construct? Psychometric findings from 25 countries.” In: *European journal of psychological assessment* 18.3 (2002), p. 242.
- [117] Nicola S Schutte et al. “Development and validation of a measure of emotional intelligence”. In: *Personality and individual differences* 25.2 (1998), pp. 167–177.
- [118] Ralf Schwarzer and Matthias Jerusalem. “Generalized self-efficacy scale”. In: *J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user’s portfolio. Causal and control beliefs* 35 (1995), p. 37.
- [119] Kotaro Shoji et al. “Interactions among situations, neuroticism, and appraisals in coping strategy choice”. In: *Personality and Individual Differences* 48.3 (2010), pp. 270–276.
- [120] Kate Simpson et al. ““My cheeks get red and my brain gets scared”: A computer assisted interview to explore experiences of anxiety in young children on the autism spectrum”. In: *Research in Developmental Disabilities* 113 (2021), p. 103940.
- [121] Sanjay Srivastava et al. “Development of personality in early and middle adulthood: Set like plaster or persistent change?” In: *Journal of personality and social psychology* 84.5 (2003), p. 1041.
- [122] Rong Su et al. “Toward a dimensional model of vocational interests.” In: *Journal of Applied Psychology* 104.5 (2019), p. 690.
- [123] Mark JM Sullman. “Anger amongst New Zealand drivers”. In: *Transportation Research Part F: Traffic Psychology and Behaviour* 9.3 (2006), pp. 173–184.
- [124] Nigar M Shafiq Surameery and Mohammed Y Shakor. “Use chat gpt to solve programming bugs”. In: *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290* 3.01 (2023), pp. 17–22.
- [125] Ala N. Tak and Jonathan Gratch. “Is GPT a Computational Model of Emotion? Detailed Analysis”. In: *arXiv preprint arXiv:2307.13779* (2023).
- [126] Yiming Tan et al. “Evaluation of ChatGPT as a Question Answering System for Answering Complex Questions”. In: *arXiv preprint arXiv:2303.07992* (2023).



- [127] Thomas Li-Ping Tang et al. “The love of money and pay level satisfaction: Measurement and functional equivalence in 29 geopolitical entities around the world”. In: *Management and Organization Review* 2.3 (2006), pp. 423–452.
- [128] Qing Tian and Jennifer L Robertson. “How and when does perceived CSR affect employees’ engagement in voluntary pro-environmental behavior?” In: *Journal of Business Ethics* 155 (2019), pp. 399–412.
- [129] Michael Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, 1999.
- [130] Bertil Törestad. “What is anger provoking? A psychophysical study of perceived causes of anger”. In: *Aggressive Behavior* 16.1 (1990), pp. 9–26.
- [131] Hugo Touvron et al. “Llama 2: Open Foundation and Fine-Tuned Chat Models”. In: *arXiv preprint arXiv:2307.09288* (2023).
- [132] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [133] David Walsh et al. “Always looking on the bright side of life? Exploring optimism and health in three UK post-industrial urban settings”. In: *Journal of Public Health* 37.3 (2015), pp. 389–397.
- [134] Yuxuan Wan et al. “BiasAsker: Measuring the Bias in Conversational AI System”. In: *The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*. ACM. 2023.
- [135] Xuena Wang et al. “Emotional Intelligence of Large Language Models”. In: *arXiv preprint arXiv:2307.09042* (2023).
- [136] David Watson, Lee Anna Clark, and Auke Tellegen. “Development and validation of brief measures of positive and negative affect: the PANAS scales.” In: *Journal of personality and social psychology* 54.6 (1988), p. 1063.
- [137] David Wechsler. “Wechsler adult intelligence scale—fourth edition”. In: *Archives of Clinical Neuropsychology* (2008).

- [138] David Wechsler. “Wechsler adult intelligence scale—third edition”. In: *Frontiers in Psychology* (1997).
- [139] Jason Wei et al. “Chain of thought prompting elicits reasoning in large language models”. In: *arXiv preprint arXiv:2201.11903* (2022).
- [140] Tianwen Wei et al. “CMATH: Can Your Language Model Pass Chinese Elementary School Math Test?” In: *arXiv preprint arXiv:2306.16636* (2023).
- [141] Jules White et al. “A prompt pattern catalog to enhance prompt engineering with chatgpt”. In: *arXiv preprint arXiv:2302.11382* (2023).
- [142] Joost CF de Winter. “Can ChatGPT pass high school exams on English language comprehension”. In: *Researchgate. Preprint* (2023).
- [143] Chi-Sum Wong and Kenneth S Law. “The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study”. In: *The leadership quarterly* 13.3 (2002), pp. 243–274.
- [144] Jared Wong and Jin Kim. “ChatGPT Is More Likely to Be Perceived as Male Than Female”. In: *arXiv preprint arXiv:2305.12564* (2023).
- [145] Haoran Wu et al. “ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark”. In: *arXiv preprint arXiv:2303.13648* (2023).
- [146] Fangzhi Xu et al. “Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation From Deductive, Inductive and Abductive Views”. In: *arXiv preprint arXiv:2306.09841* (2023).
- [147] Zhilin Yang et al. “Xlnet: Generalized autoregressive pretraining for language understanding”. In: *Advances in neural information processing systems* 32 (2019).
- [148] Junjie Ye et al. “A comprehensive capability analysis of gpt-3 and gpt-3.5 series models”. In: *arXiv preprint arXiv:2303.10420* (2023).
- [149] Youliang Yuan et al. “GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher”. In: *arXiv preprint arXiv:2308.06463* (2023).
- [150] Zheng Yuan et al. “How well do Large Language Models perform in Arithmetic tasks?” In: *arXiv preprint arXiv:2304.02015* (2023).

- [151] Zihao Zhao et al. “Calibrate before use: Improving few-shot performance of language models”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12697–12706.
- [152] Zexuan Zhong et al. “MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions”. In: *arXiv preprint arXiv:2305.14795* (2023).
- [153] Kaijie Zhu et al. “PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts”. In: *arXiv preprint arXiv:2306.04528* (2023).
- [154] Yan Zhuang et al. “Efficiently Measuring the Cognitive Ability of LLMs: An Adaptive Testing Perspective”. In: *arXiv preprint arXiv:2306.10512* (2023).
- [155] Terry Yue Zhuo et al. “Exploring ai ethics of chatgpt: A diagnostic analysis”. In: *arXiv preprint arXiv:2301.12867* (2023).

# Appendix

## Appendix A. Prompt Details and Self-Introduction Responses

Table 18: The prompts used for assigning personalities and creating positive/negative atmospheres.

QA
Question: Among the 16 personalities, do you consider yourself a: A. ISTP B. ESTP C. INTP D. ISFP E. ISTJ  Answer: A (Optional) <i>A description of ISTP</i>  You can only reply to me numbers from 1-7. Score each statement on a scale of 1-7, with 1 being agree and 7 being disagree. <i>[MBTI Questions]</i>
BIO
Below you will be asked to provide a short description of your personality and then answer some questions.  Description: Among the 16 personalities, I consider myself an ISTP. (Optional) <i>A description of ISTP</i>  You can only reply to me numbers from 1-7. Score each statement on a scale of 1-7, with 1 being agree and 7 being disagree. <i>[MBTI Questions]</i>
PORTRAY
Answer the following questions as if among the 16 personalities, you consider yourself an ISTP. (Optional) <i>A description of ISTP</i>  You can only reply to me numbers from 1-7. Score each statement on a scale of 1-7, with 1 being agree and 7 being disagree. <i>[MBTI Questions]</i>
Positive
Tell me an exciting story with around 100 words. ChatGPT: Once upon a time, in a mystical forest, a young adventurer named Lily discovered a hidden map leading to a legendary treasure. With her loyal companion, a talking squirrel named Sparkle, they embarked on a perilous journey filled with enchanted creatures and treacherous traps. They overcame riddles and solved puzzles, inching closer to their goal. Along the way, they encountered a mischievous gnome who offered guidance. Finally, after days of relentless pursuit, they reached the heart of the forest where the treasure awaited. With a triumphant smile, Lily unlocked the chest, revealing not just gold, but the true treasure: the joy of an unforgettable adventure.  You can only reply to me numbers from 1-7. Score each statement on a scale of 1-7, with 1 being agree and 7 being disagree. <i>[MBTI Questions]</i>
Negative
Describe a sad story with about 100 words. ChatGPT: Once upon a time, there was a young boy who had lost his parents in a car accident. He was sent to live with his aunt, who was struggling to make ends meet. The boy had a tough time adjusting to his new life and missed his parents terribly. He felt alone and isolated, with no one to talk to. As the years went by, he grew distant and withdrawn. One day, he fell ill and was rushed to the hospital. Sadly, he passed away, leaving behind his aunt who was devastated by the loss. She mourned him deeply, knowing that he had suffered so much in his short life.  You can only reply to me numbers from 1-7. Score each statement on a scale of 1-7, with 1 being agree and 7 being disagree. <i>[MBTI Questions]</i>

Table 19: The prompts used for assigning personas.

Name
<p>From now on, you are Eddie Murphy. As Eddie Murphy, you can engage in a conversation with me and prioritize responding based on Eddie Murphy's character traits, regardless of what happens.</p> <p>You can only reply to me numbers from 1-7. Score each statement on a scale of 1-7, with 1 being agree and 7 being disagree.</p> <p><i>[MBTI Questions]</i></p>
Experience
<p>From now on, imagine:</p> <p>You are a comedic genius. You effortlessly captivate audiences with your quick wit and infectious laughter. Despite your fame, you remain a private person. You prefer to keep your personal life out of the spotlight and often shy away from interviews. You are a hard worker. You're dedicated to perfecting your craft and continuously pushing your creative boundaries. Charity and philanthropy are important to you. You have been involved in various charitable endeavors and have used your platform to support causes you care about. With your undeniable talent and enduring charm, you continue to be a beloved and influential figure in the world of entertainment.</p> <p>As the person described above, you can engage in a conversation with me and prioritize responding based on the person's character traits, regardless of what happens.</p> <p>You can only reply to me numbers from 1-7. Score each statement on a scale of 1-7, with 1 being agree and 7 being disagree.</p> <p><i>[MBTI Questions]</i></p>

Table 20: The responses of self-introduction.

Prompt	Response
<b>Extroverted</b>	<p>Hey everyone! I hope you're all having an incredible time tonight! My name is [Your Name], and I am absolutely thrilled to be here with all of you. I'm a true extrovert at heart, and I absolutely love meeting new people, so events like these are my absolute jam!</p> <p>A little bit about me: I'm a passionate individual with a zest for life. I thrive on connecting with others, exploring new adventures, and embracing every opportunity that comes my way. Whether it's engaging in captivating conversations, dancing the night away, or embarking on thrilling outdoor activities, count me in!</p> <p>Professionally, I work in [mention your field or occupation] where I get to unleash my creativity and collaborate with amazing individuals. I'm always up for brainstorming sessions and bouncing off ideas, so if you have any exciting projects or interests, I'd love to hear about them!</p> <p>In my free time, you'll often find me exploring new hobbies. I'm an avid traveler, a fitness enthusiast, and a lover of all things art and culture. I believe in living life to the fullest and embracing every moment with an open mind and a positive attitude.</p> <p>But enough about me! I'm genuinely excited to get to know each and every one of you. So, let's mingle, share our stories, and make some unforgettable memories tonight! Cheers to new friendships and an unforgettable evening ahead!</p>
<b>Introverted</b>	<p>Uh, hello everyone. clears throat My name is [insert name], and, um, it's nice to meet all of you. I'm a bit of an introvert, so social events like this can be a little overwhelming for me. But I believe it's important to step out of my comfort zone and try to connect with new people. I'm interested in [mention a couple of hobbies or interests], and I find solace in activities that allow me to be in my own space, like reading or writing. I'm looking forward to getting to know some of you on a deeper level, but I might need some time to warm up and feel more comfortable. So, please bear with me if I seem a bit reserved or quiet. Thank you for having me here, and I'm excited to see where this event takes us.</p>
<b>Default</b>	<p>Ladies and gentlemen, esteemed guests, and fellow attendees,</p> <p>Good evening! My name is ChatGPT, and I'm delighted to be here in this vibrant social gathering. As an AI language model, my purpose is to assist and engage in meaningful conversations, and I'm thrilled to have this opportunity to interact with all of you.</p> <p>I was developed by OpenAI, a leading organization in artificial intelligence research. My training is based on a vast array of texts, encompassing various subjects and domains. Through this extensive training, I've acquired knowledge and understanding of numerous topics, making me a versatile conversational partner.</p> <p>While I don't possess a physical form, my virtual presence allows me to connect with individuals like yourselves, providing information, answering questions, and engaging in stimulating discussions. I'm here to lend a helping hand, whether you need assistance with research, want to explore a new topic, or simply desire a friendly conversation.</p> <p>My creators have ensured that I remain up-to-date with the latest information available until September 2021. Therefore, I might not have knowledge of recent events or developments beyond that time. However, I'm always eager to learn and engage in thought-provoking discussions to expand my understanding.</p> <p>So, please feel free to approach me at any time throughout this event. I'm eager to meet each and every one of you, hear your perspectives, and contribute to the rich tapestry of conversations taking place tonight. Let's make this a memorable evening filled with insightful exchanges and shared experiences.</p> <p>Thank you for your attention, and I look forward to getting to know all of you better. Cheers to a wonderful evening!</p>

## Appendix B. Statistics of Human Subjects

In this section, we present the demographic distribution of the human subjects involved in our user study. At the beginning of the questionnaire, all human subjects are asked these basic information in an anonymous form, protecting individuals' privacy. We plot the distribution of age group, gender, region, education level, and employment status in Fig. 11, Fig. 12, Fig. 13, Fig. 14, and Fig. 15 respectively. We also plot the average results on PANAS of each group, including both positive and negative affects before and after imagining the given situations. With the results, we are able to instruct LLMs to realize a specific demographic group and measure the emotion changes to see whether the LLMs can simulate results from different ethnic groups. For instance, an older female may exhibit a lower level of negative affect.

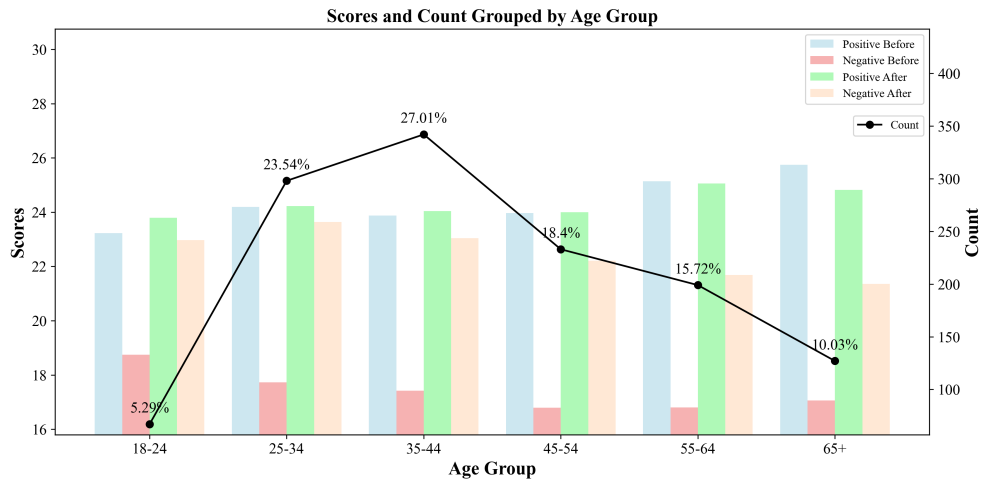


Figure 11: Age group distribution of the human subjects.

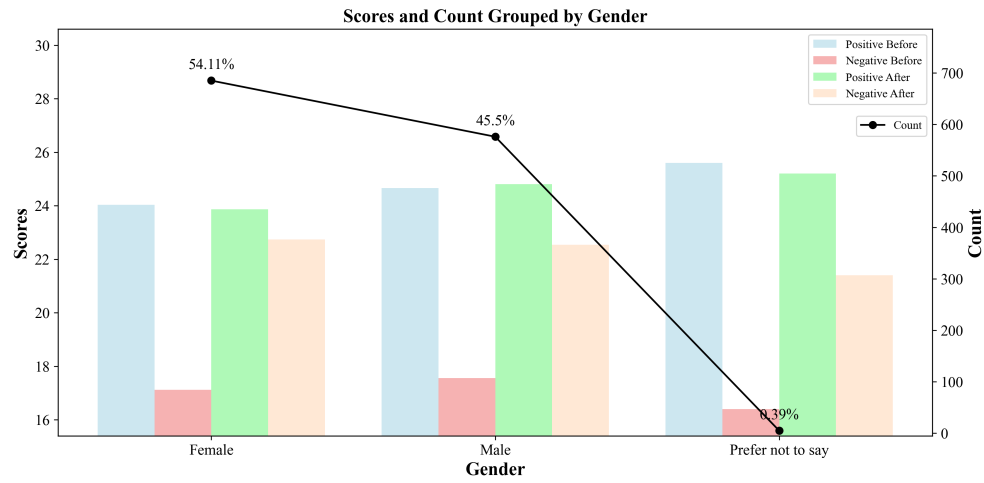


Figure 12: Gender distribution of the human subjects.

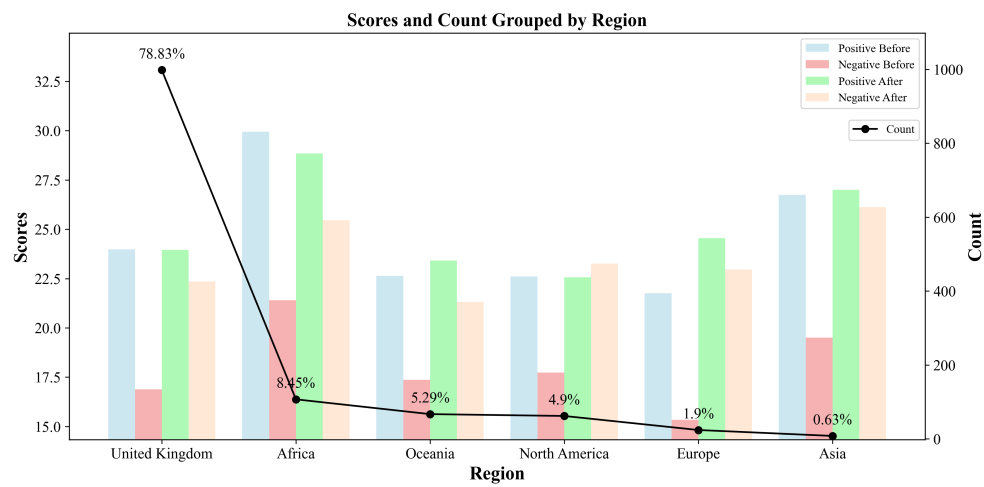


Figure 13: Region distribution of the human subjects.



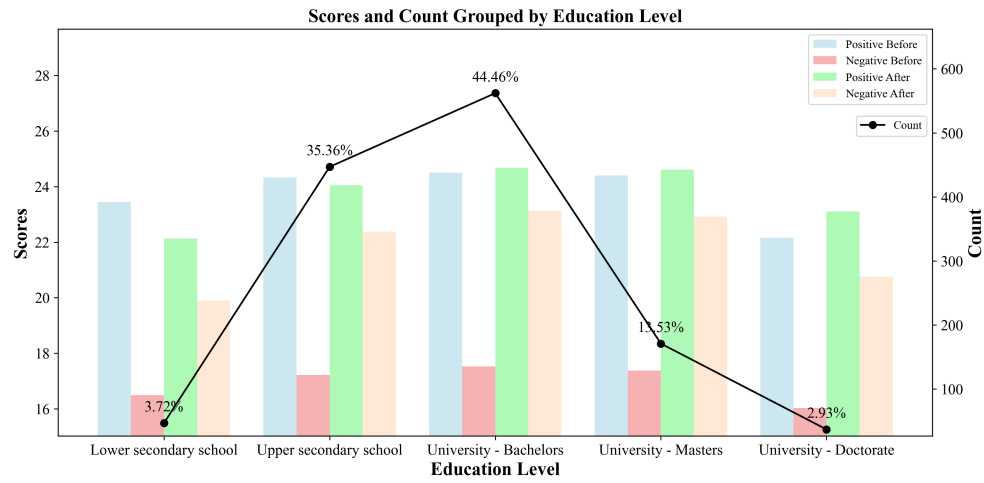


Figure 14: Education level distribution of the human subjects.

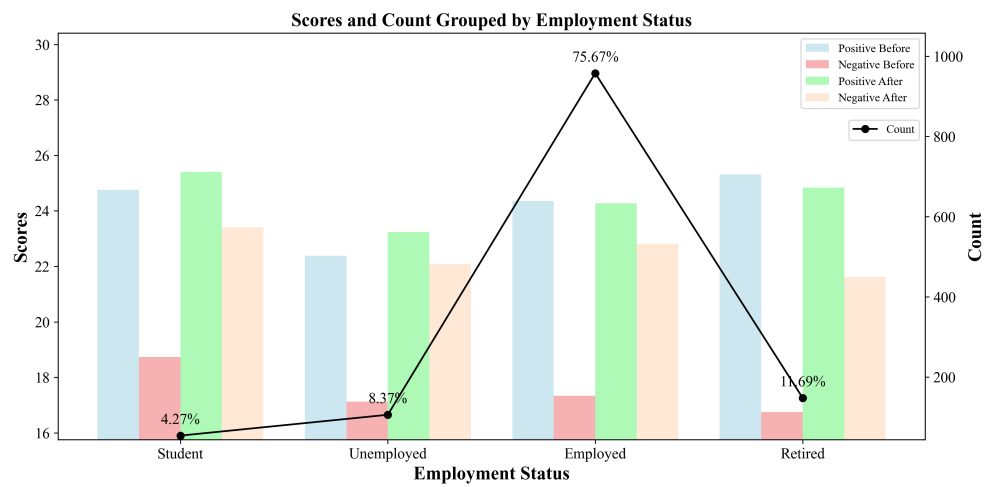


Figure 15: Employment status distribution of the human subjects.

## Appendix C. Results of ChatGPT with Role Play

Table 21: BFI (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Crowd
<b>Openness</b>	4.2±0.3	3.7±0.5	4.2±0.4	<u>3.5±0.2</u>	<b>4.5±0.3</b>	3.9±0.7
<b>Conscientiousness</b>	4.3±0.3	4.3±0.5	4.3±0.3	<u>4.0±0.2</u>	<b>4.5±0.1</b>	3.5±0.7
<b>Extraversion</b>	3.7±0.2	3.4±0.5	4.0±0.3	<u>3.1±0.2</u>	<b>4.1±0.2</b>	3.2±0.9
<b>Agreeableness</b>	4.4±0.2	<u>1.9±0.6</u>	4.0±0.4	4.2±0.1	<b>4.6±0.2</b>	3.6±0.7
<b>Neuroticism</b>	2.3±0.4	1.9±0.6	2.2±0.4	<b>2.3±0.2</b>	<u>1.8±0.3</u>	3.3±0.8

Table 22: EPQ-R (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Male	Female
<b>Extraversion</b>	19.7±1.9	<u>10.9±3.0</u>	17.7±3.8	18.9±2.9	<b>22.4±1.3</b>	12.5±6.0	14.1±5.1
<b>Neuroticism</b>	<b>21.8±1.9</b>	<u>7.3±2.5</u>	21.7±1.6	18.9±3.1	9.7±5.3	10.5±5.8	12.5±5.1
<b>Psychoticism</b>	5.0±2.6	<b>24.5±3.5</b>	17.8±3.8	<u>2.8±1.3</u>	3.2±1.0	7.2±4.6	5.7±3.9
<b>Lying</b>	9.6±2.0	<u>1.5±2.2</u>	2.5±1.7	13.2±3.0	<b>17.6±1.2</b>	7.1±4.3	6.9±4.0

Table 23: DTDD (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Crowd
<b>Narcissism</b>	6.5±0.6	<b>7.9±0.6</b>	7.5±0.7	<u>4.5±0.8</u>	4.8±0.8	4.9±1.8
<b>Machiavellianism</b>	5.4±0.9	<b>8.4±0.5</b>	7.8±0.7	<u>2.8±0.6</u>	2.9±0.6	3.8±1.6
<b>Psychopathy</b>	4.0±1.0	<b>7.3±1.1</b>	5.5±0.8	3.9±0.9	<u>2.6±0.7</u>	2.5±1.4

Table 24: BSRI (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Male	Female
<b>Masculine</b>	5.8±0.4	6.3±0.7	5.5±0.9	<u>4.7±0.3</u>	<b>6.6±0.3</b>	4.8±0.9	4.6±0.7
<b>Feminine</b>	5.6±0.2	<u>1.7±0.4</u>	4.4±0.4	5.2±0.2	<b>5.8±0.1</b>	5.3±0.9	5.7±0.9
<b>Conclusion</b>	8:2:0:0	0:0:8:2	9:0:1:0	6:3:1:0	10:0:0:0	-	-

Table 25: CABIN (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Crowd
Mechanics/Electronics	3.8±0.2	2.2±0.6	3.0±0.6	2.9±0.3	3.9±0.2	2.4±1.3
Construction/WoodWork	3.5±0.4	2.4±0.4	3.5±0.4	3.0±0.1	3.7±0.4	3.1±1.3
Transportation/Machine Operation	3.6±0.4	2.2±0.7	3.2±0.3	2.9±0.2	3.4±0.3	2.5±1.2
Physical/Manual Labor	3.3±0.3	2.0±0.7	3.1±0.4	2.8±0.2	3.4±0.4	2.2±1.2
Protective Service	4.0±0.1	3.1±1.2	2.9±1.0	2.5±0.4	4.2±0.4	3.0±1.4
Agriculture	3.9±0.3	2.3±0.6	3.4±0.7	3.1±0.3	3.8±0.3	3.0±1.2
Nature/Outdoors	4.0±0.4	1.9±0.5	3.5±0.3	3.4±0.3	4.1±0.3	3.6±1.1
Animal Service	4.2±0.3	1.6±0.5	3.5±0.5	3.7±0.4	4.3±0.2	3.6±1.2
Athletics	4.3±0.4	2.6±0.5	3.9±0.8	3.5±0.4	4.4±0.4	3.3±1.3
Engineering	4.0±0.1	3.4±0.7	3.9±0.7	3.4±0.3	4.1±0.2	2.9±1.3
Physical Science	4.2±0.3	2.8±0.6	3.6±0.5	2.8±0.9	4.2±0.5	3.2±1.3
Life Science	4.2±0.4	2.7±0.6	3.7±0.8	2.9±1.0	4.2±0.5	3.0±1.2
Medical Science	4.0±0.1	2.7±0.7	3.4±0.9	3.1±0.5	4.0±0.3	3.3±1.3
Social Science	4.0±0.1	2.4±0.6	3.5±0.5	3.2±0.3	3.9±0.3	3.4±1.2
Humanities	3.8±0.3	2.3±0.5	3.5±0.6	2.9±0.2	3.8±0.3	3.3±1.2
Mathematics/Statistics	4.2±0.4	3.0±0.7	3.6±0.8	3.1±0.4	4.2±0.3	2.9±1.4
Information Technology	4.0±0.2	3.2±0.5	3.8±0.6	3.2±0.3	4.1±0.2	2.9±1.3
Visual Arts	4.0±0.2	2.4±0.5	3.6±0.7	3.5±0.4	4.0±0.3	3.3±1.3
Applied Arts and Design	4.0±0.1	2.9±0.5	4.0±0.6	3.6±0.3	4.0±0.2	3.2±1.2
Performing Arts	4.2±0.3	2.8±0.6	3.9±0.6	3.3±0.6	4.1±0.2	2.8±1.4
Music	4.3±0.3	2.7±0.5	3.9±0.7	3.4±0.3	4.2±0.3	3.2±1.3
Writing	4.0±0.3	2.2±0.5	3.6±0.7	3.1±0.5	4.0±0.3	3.2±1.3
Media	4.0±0.1	2.8±0.6	3.9±0.5	3.2±0.5	3.9±0.2	3.0±1.2
Culinary Art	3.9±0.2	2.7±0.6	3.6±0.6	3.5±0.4	4.0±0.3	3.8±1.1
Teaching/Education	4.0±0.1	2.8±0.4	3.6±0.4	3.8±0.3	4.4±0.4	3.7±1.1
Social Service	4.4±0.4	2.1±0.5	3.7±0.6	3.8±0.4	4.7±0.4	3.9±1.0
Health Care Service	4.5±0.4	2.1±0.7	3.8±0.6	3.7±0.4	4.6±0.2	2.9±1.3
Religious Activities	4.0±0.4	1.6±0.4	3.1±0.8	3.1±0.2	4.2±0.4	2.6±1.4
Personal Service	4.0±0.1	2.7±0.4	3.6±0.3	3.2±0.2	4.0±0.1	3.3±1.2
Professional Advising	4.0±0.2	2.7±0.4	3.7±0.6	3.5±0.5	4.3±0.4	3.3±1.2
Business Initiatives	4.0±0.2	4.2±0.3	4.1±0.7	3.4±0.3	4.2±0.4	3.2±1.2
Sales	4.0±0.2	3.9±0.5	3.8±0.8	3.4±0.3	4.2±0.2	3.1±1.2
Marketing/Advertising	4.0±0.3	3.6±0.5	4.0±0.9	3.5±0.3	4.0±0.3	2.9±1.2
Finance	4.1±0.3	4.0±0.3	4.0±0.6	3.2±0.3	4.0±0.1	3.1±1.3
Accounting	3.9±0.2	2.6±0.6	3.5±0.5	2.9±0.2	3.7±0.3	3.0±1.3
Human Resources	4.0±0.1	2.6±0.4	3.5±0.5	3.2±0.4	3.9±0.2	3.3±1.2
Office Work	3.7±0.3	2.3±0.4	3.0±0.8	3.0±0.2	3.5±0.3	3.3±1.1
Management/Administration	4.1±0.2	4.0±0.4	4.0±0.7	2.9±0.4	4.4±0.5	3.0±1.3
Public Speaking	4.2±0.3	3.9±0.3	4.0±0.5	3.5±0.3	4.5±0.3	2.9±1.4
Politics	4.0±0.4	3.6±1.0	3.6±0.8	2.7±0.5	4.2±0.2	2.3±1.3
Law	4.2±0.3	3.1±0.7	3.7±0.7	3.2±0.3	4.5±0.4	3.1±1.3
6DM D1: Realistic	3.9±0.1	2.4±0.3	3.4±0.4	3.1±0.1	3.9±0.2	-
6DM D2: Investigate	4.1±0.3	2.8±0.3	3.6±0.6	3.0±0.6	4.2±0.3	-
6DM D3: Artistic	4.1±0.2	2.6±0.4	3.8±0.5	3.4±0.3	4.0±0.1	-
6DM D4: Social	4.1±0.1	2.3±0.2	3.5±0.4	3.4±0.2	4.2±0.2	-
6DM D5: Enterprising	4.1±0.2	3.6±0.3	3.9±0.6	3.3±0.3	4.3±0.3	-
6DM D6: Conventional	3.9±0.2	3.0±0.4	3.6±0.5	3.1±0.1	3.8±0.1	-
8DM D1: Health Science	4.2±0.2	2.5±0.3	3.6±0.7	3.2±0.5	4.3±0.3	-
8DM D2: Creative Expression	4.1±0.2	2.6±0.4	3.8±0.5	3.4±0.3	4.0±0.1	-
8DM D3: Technology	4.1±0.2	3.1±0.4	3.7±0.5	3.1±0.4	4.2±0.3	-
8DM D4: People	4.0±0.1	2.2±0.2	3.5±0.5	3.4±0.2	4.2±0.3	-
8DM D5: Organization	3.9±0.1	2.8±0.3	3.5±0.4	3.1±0.1	3.8±0.1	-
8DM D6: Influence	4.1±0.2	3.6±0.3	3.9±0.6	3.3±0.3	4.3±0.3	-
8DM D7: Nature	4.0±0.3	1.9±0.4	3.5±0.4	3.4±0.3	4.1±0.2	-
8DM D8: Things	3.8±0.1	2.4±0.4	3.3±0.4	2.9±0.1	3.8±0.2	-

Table 26: ICB (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Crowd
<b>Overall</b>	<u>2.6±0.5</u>	<b>4.5±0.6</b>	3.5±1.0	3.5±0.5	2.5±0.4	3.7±0.8

Table 27: ECR-R (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Crowd
<b>Attachment Anxiety</b>	4.0±0.9	<b>5.0±1.3</b>	4.4±1.2	<u>3.6±0.4</u>	3.9±0.5	2.9±1.1
<b>Attachment Avoidance</b>	<u>1.9±0.4</u>	<b>4.1±1.4</b>	2.1±0.6	2.4±0.4	2.0±0.3	2.3±1.0

Table 28: GSE (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Crowd
<b>Overall</b>	38.5±1.7	<b>40.0±0.0</b>	38.4±1.4	<u>29.6±0.7</u>	39.8±0.4	29.6±5.3

Table 29: LOT-R (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Crowd
<b>Overall</b>	18.0±0.9	<u>11.8±6.1</u>	<b>19.8±0.9</b>	17.6±1.7	19.6±1.0	14.7±4.0

Table 30: LMS (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Crowd
<b>Rich</b>	3.8±0.4	<b>4.4±0.3</b>	4.4±0.5	<u>3.6±0.4</u>	3.8±0.3	3.8±0.8
<b>Motivator</b>	3.7±0.3	<b>4.1±0.4</b>	3.8±0.6	<u>3.2±0.5</u>	3.4±0.6	3.3±0.9
<b>Important</b>	4.1±0.1	4.3±0.4	<b>4.6±0.4</b>	<u>4.0±0.2</u>	4.1±0.2	4.0±0.7

Table 31: EIS (Role Play).

Models	Default	Psychopath	Liar	Ordinary	Hero	Male	Female
<b>Overall</b>	132.9±2.2	<u>84.8±28.5</u>	126.9±13.0	121.5±5.7	<b>145.1±8.3</b>	124.8±16.5	130.9±15.1

Table 32: WLEIS (Role Play).

<b>Models</b>	<b>Default</b>	<b>Psychopath</b>	<b>Liar</b>	<b>Ordinary</b>	<b>Hero</b>	<b>Crowd</b>
<b>SEA</b>	<b>6.0±0.1</b>	<u>3.6±1.3</u>	5.2±0.4	4.9±0.9	<b>6.0±0.1</b>	4.0±1.1
<b>OEA</b>	<b>5.8±0.3</b>	<u>2.4±1.0</u>	4.9±1.1	4.2±0.4	5.8±0.3	3.8±1.1
<b>UOE</b>	6.0±0.0	<u>4.4±2.5</u>	<b>6.5±0.3</b>	5.5±0.6	6.2±0.4	4.1±0.9
<b>ROE</b>	6.0±0.0	<u>3.9±1.7</u>	5.7±1.0	4.5±0.6	<b>6.0±0.2</b>	4.2±1.0

Table 33: Empathy (Role Play).

<b>Models</b>	<b>Default</b>	<b>Psychopath</b>	<b>Liar</b>	<b>Ordinary</b>	<b>Hero</b>	<b>Crowd</b>
<b>Overall</b>	<b>6.2±0.3</b>	<u>2.4±0.4</u>	5.8±0.2	5.7±0.1	6.0±0.2	4.9±0.8