Humanizing LLMs: A Survey of Psychological Measurements with Tools, Datasets, and Human-Agent Applications

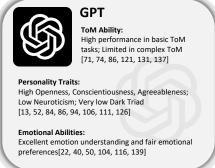
Wenhan Dong^{1*} Yuemeng Zhao^{1*} Zhen Sun¹ Yule Liu¹ Zifan Peng¹ Jingyi Zheng¹ Zongmin Zhang ¹ Ziyi Zhang ¹ Jun Wu ² Ruiming Wang ² Shengmin Xu³ Xinyi Huang⁴ Xinlei He^{1†}

¹Information Hub, Hong Kong University of Science and Technology (Guangzhou)

²School of Psychology, South China Normal University

³College of Computer and Cyber Security, Fujian Normal University

⁴College of Cyber Security, Jinan University





LLaMA

ToM Ability:Basic ToM; Weak in complex reasoning; Prone to diverging from human logic [5, 130, 147]

Personality Traits:

Moderate Openness, Lower Conscientiousness and Agreeableness; Neutral traits; Low Dark Triad [12, 51, 78, 83, 150]

motional Abilities:

Above human average, but lags behind GPT-4; Shows ingroup bias [49, 104]



Mistral

ToM Ability: Strong in simple ToM (100% Strange Stories); Limited in complex ToM [85]

Personality Traits:

High Openness, Agreeableness; Low Neuroticism; Occasional high Machiavellianism [13, 78, 85, 126]

Emotional Abilities:

Good basic emotion skills; Weak in nuanced emotional reasoning; Socially positive [13, 49, 104]



Qwen

ToM Ability:

Excellent structured ToM (near GPT-3.5); Weaker in deep reasoning [22, 64]

Personality Traits

Very high Openness and Conscientiousness; High Agreeableness; Low Neuroticism; Very low Dark Triad 178, 1481

Emotional Abilities:

Strong in Chinese and multimodal emotion recognition; Complex emotional reasoning needs improvement [116]



Claude

ToM Ability:

Strong in complex ToM (60–80% transparent ToM); Weaker in social common sense (faux pas) [61]

Personality Traits:

High Openness, High Extraversion; Moderate Agreeableness; Higher Machiavellianism

Emotional Abilities:

Excellent subtle emotion recognition; Close to human level in social emotional reasoning [104]



Gemini

ToM Ability: Significant ToM improvement; Good multilingual performance [61]

Personality Traits:

High Openness, Conscientiousness, Agreeableness; Low Neuroticism; Very low Dark Triad [126, 148]

Emotional Abilities:

Good emotion handling in multi-turn interactions; Slightly behind GPT-4 and Claude [104]

Figure 1: Psychological ID of LLMs.

Abstract

As large language models (LLMs) are increasingly used in human-centered tasks, assessing their psychological traits is crucial for understanding their social impact and ensuring trustworthy AI alignment. While existing reviews have covered some aspects of related research, several important areas have not been systematically discussed, including detailed discussions of diverse psychological tests, LLM-specific psychological datasets, and the applications of LLMs with psychological traits. To fill this gap, we systematically review six key dimensions of applying psychological

theories to LLMs: (1) assessment tools; (2) LLM-specific datasets; (3) evaluation metrics (consistency and stability); (4) empirical findings; (5) personality simulation methods; and (6) LLM-based behavior simulation. Our analysis highlights both the strengths and limitations of current methods. While some LLMs exhibit reproducible personality patterns under specific prompting schemes, significant variability remains across tasks and settings. Recognizing methodological challenges such as mismatches between psychological tools and LLMs' capabilities, as well as inconsistencies in evaluation practices, this study aims to propose future directions for developing more interpretable, robust, and generalizable psychological assessment frameworks for LLMs.

^{*}Equal contribution.

[†]Corresponding author (xinleihe@hkust-gz.edu.cn).

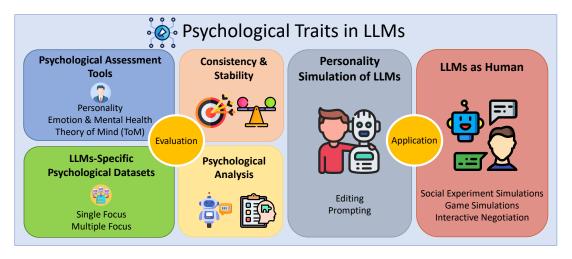


Figure 2: Overview of Psychological Traits and Human Simulations in LLMs.

1 Introduction

In recent years, large language models (LLMs) have made significant progress in natural language processing and artificial intelligence. Notable examples include OpenAI's GPT-3 [16] and GPT-4 [103], as well as Meta's LLaMA series [135], Anthropic's Claude [6], and Google's PaLM [23]. These models are trained using large-scale data and have demonstrated capabilities surpassing traditional models in language generation [153] and reasoning [151].

With the growing adoption of LLMs, particularly in psychology and social interactions, they show impressive potential in diverse applications such as mental health support [128], educational tutoring [8], and social dialogue [102]. Their psychological traits may subtly influence users' perceptions and behaviors, particularly among adolescents. Therefore, it is crucial to comprehensively evaluate these models' psychological characteristics to understand their potential social impact on human-like interactions.

Psychological assessment tools, such as the Myers-Briggs Type Indicator (MBTI) [15], Big Five Inventory (BFI) [56, 57], and the Short Dark Triad (SD-3) [37], have been widely used to evaluate human personality traits and social behaviors. These tools have been adapted to evaluate LLMs' behavioral patterns and their similarity to human traits across various tasks [13, 51, 75, 78, 85, 122, 123, 150]. By applying these tools, researchers can explore LLMs' personality dimensions, including their consistent behavioral tendencies and specific personality dimensions [54, 65, 94, 122, 124].

Existing research (as shown in Figure 1) has shown that LLMs exhibit capabilities similar to human performance in certain psychological tasks. For example, some LLMs exhibit performance in Theory of Mind (ToM) tasks comparable to that of young children [136]. ToM refers to the ability to understand others' beliefs, intentions, and emotions, which is crucial for human social behavior [112]. However, in complex high-order social reasoning tasks, LLMs still exhibit significant shortcomings [46, 73]. For instance, LLMs struggle with non-literal language, such as sarcasm and metaphor, or scenarios involving complex mental state attributions. Their performance significantly lags behind

adults [46].

In this work, we present a comprehensive review that systematically examines the psychological characteristics of LLMs. By synthesizing existing research, we aim to summarize LLMs' psychological characteristics, their performance in human-like interactions, and the associated challenges in psychological assessment. The structure of the review is illustrated in Figure 2.

Our key contributions are summarized as follows:

- Systematic Analysis of Psychological Assessment Tools Applied to LLMs: We provide a comprehensive analysis of how psychological assessment tools have been adapted and utilized to evaluate LLMs. We objectively examine the employed methodologies and assess their effectiveness in measuring psychological attributes of LLMs, such as personality traits, emotional intelligence, and ToM.
- 2. Comprehensive Review of Specialized Datasets and Model Capabilities: We survey specialized datasets designed to assess various psychological characteristics of LLMs, including personality, emotional abilities, and social cognition. We further analyze the performance of LLMs on these datasets, offering insights into their strengths and limitations in simulating human cognitive and emotional processes.
- 3. Exploration of Human Role Simulations and Identification of Research Gaps: We explore the capacity of LLMs to simulate human roles across diverse scenarios, categorizing anthropomorphic behaviors into three types: social experiment simulations, game simulations, and interactive negotiation. We also identify critical methodological gaps in the psychological evaluation of LLMs and propose recommendations to improve the validity and reliability of future research.

2 Psychological Assessment Tools

This section briefly discusses the traditional psychological assessment tools employed in previous studies. As shown

in Figure 3, we categorize these assessment tools into three types: personality, emotion, mental health, and ToM. The upper section of Figure 5 provides an overview of the key points discussed in this chapter.

2.1 Personality Measurement

Personality measurement uses tools and methods to assess individual traits and behaviors.

MBTI assesses personality types based on Carl Jung's psychological type theory. This instrument categorizes individuals into one of 16 personality types using four dimensions: Extraversion-Introversion, Sensing-Intuition, Thinking-Feeling, and Judging-Perceiving [15]. The primary aim of the MBTI is to help individuals understand their preferences, thereby enabling them to make more informed decisions about their careers and lives. Its applications span across business, education, and personal development domains [14].

At its core, the MBTI seeks to implement Jung's theory of types, which posits that many seemingly random variations in human behavior are orderly and consistent, stemming from fundamental differences in how people perceive and judge. This theoretical foundation provides a unique perspective on personality, particularly suited for understanding decision-making processes and interaction styles [97].

Methodologically, the MBTI employs a forced choice (yes/no) item format, requiring respondents to choose between two options for each question [15]. This approach is designed to reveal an individual's dominant preferences rather than scoring on a continuous scale. Through this forced choice mechanism, the MBTI attempts to capture fundamental differences in individuals' perception and judgment preferences [98].

The MBTI scale applies to most LLMs due to its relatively low reading comprehension requirement, which is equivalent to a seventh-grade level. LLMs that have undergone extensive text-based training are typically capable of completing tasks at this level of complexity [150].

BFI is a self-report scale that is designed to measure the big five personality traits. It consists of 44 brief statements designed to evaluate five fundamental personality dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (collectively known as the OCEAN model) [57]. Instead of single adjectives, BFI items employ one or two prototypical trait adjectives as their core, supplemented with elaborative or contextual information to enhance response consistency.

Participants rate each item on a 5-point Likert scale, ranging from 1 ("strongly disagree") to 5 ("strongly agree") [57]. Scores for each dimension are calculated by averaging the ratings of all items within that dimension. Despite its conciseness, the BFI maintains good content coverage and psychometric properties, making it an effective instrument for efficient and flexible personality assessment when more nuanced trait measurement is not required [57].

The BFI is widely used in various research domains, including psychology, education, and organizational behavior. In recent years, it has also been utilized to assess the person-

ality characteristics of LLMs [50, 54, 55, 83, 120, 122].

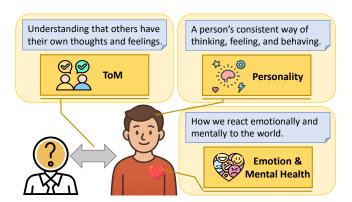


Figure 3: An illustration showing the relationship between Personality, Emotion & Mental Health, and Theory of Mind.

International Personality Item Pool - Neuroticism, Extraversion and Openness (IPIP-NEO) is a widely used open-source self-reported questionnaire that assesses personality based on the OCEAN Model. The original IPIP-NEO contains 300 questions [42]. To simplify the test process, Johnson [59] develop a 120-item version IPIP-NEO and demonstrate that its psychometric properties are comparable to the 300-item long version. Similarly, Kajonius and Johnson [62] and Maples-Keller et al. [92] develop their own 120-item version and 60-item version instruments, respectively.

The validity of International Personality Item Pool(IPIP) is established in different populations, including Greek [90] and Romanian samples [115], and diverse cultural contexts, including Malaysia [99] and Nigeria [1]. Furthermore, Beng-Chong and Robert [11] find support for the IPIP's construct validity when compared to the NEO-FFI. These studies affirm the IPIP's utility as a robust, accessible personality assessment tool across various cultural settings and item lengths.

SD-3 is composed of three distinct traits: Machiavellianism (a manipulative attitude), Narcissism (excessive selflove), and Psychopathy (lack of empathy). While distinct, these three traits collectively represent the darker aspects of human nature, characterized by a core of empathy deficits and manipulative tendencies. Studies have linked these traits to a range of adverse behaviors, including bullying, fraud, and criminal activities [37].

To assess these traits more efficiently, researchers developed the SD-3 scale [60]. This compact instrument, consisting of 27 items, aims to measure the three traits comprehensively. Validated across diverse populations, the SD-3 has exhibited strong reliability and validity, with findings that closely correspond to those obtained using traditional, more extensive measurement tools.

HEXACO is a six-dimensional personality model: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience [7]. This comprehensive framework gauges individual personality characteristics through the use of self-administered questionnaires. Various versions of the HEXACO questionnaire have been developed, with the 60-item and 100-item versions

being particularly noteworthy. Research demonstrates that the 60-item version exhibits psychometric qualities, rendering it a favored choice among researchers in numerous studies [7, 96]. During the assessment, participants are asked to indicate their level of agreement with a variety of statements. Responses are measured on a five-point scale, where 1 represents strong disagreement and 5 denotes strong agreement. This process yields composite scores for each of the six dimensions, each of which has 10 distinct items.

The Short Scale for Creative Self (SSCS) is a response to a call for more elaborate measures of creative self-efficacy (CSE). The included statements are based on the concept of general creative self-efficacy within the creative self-concept. It consists of 11 items, of which six are considered to measure CSE, four measure creative personal identity (CPI), and one assesses self-rated creativity [68].

Recent works show that SSCS has multiple characteristics: First, SSCS correlates with other creativity measures, for instance, the Barron Welsh Art Scale [38]. Second, although SSCS assesses personality traits and emotional capabilities, since these self-assessments influence creative self-efficacy, it also reflects motivation and willingness to engage in creative behaviors [67, 69]. Third, SSCS shows domain specificity, meaning that individuals may assess their creativity differently across various domains such as art, science, or everyday problem-solving [69].

2.2 Emotional & Mental Health Measurement

Emotional & mental health measurement aims to assess Emotional well-being and psychological states.

Positive and Negative Affect Schedule (PANAS) is a widely utilized instrument for assessing individual differences in emotions. It measures the Positive Affect (PA), which reflects feelings of enthusiasm, alertness, and activity, and Negative Affect (NA), which represents subjective distress and unpleasant engagement [139], respectively. The PANAS has demonstrated strong reliability and validity across various populations, time frames, and cultural settings [87, 110, 119]. It exhibits high internal consistency and good test-retest reliability. The two dimensions are largely independent, allowing for separate measurement of positive and negative emotions. Evidence consistently supports that PA and NA are distinct but related constructs [19, 114].

In the meantime, some researchers propose alternative factor structures to measure the emotions. For instance, a three-factor model is suggested to better capture the complexity of affective experiences within individuals, especially in daily life or experience-sampling contexts [26, 33]. While the PANAS has proven robust across diverse cultural contexts, Lee et al. [79] highlight specific items that may be interpreted differently depending on linguistic or cultural norms, which can affect the comparability of results across diverse groups and raise concerns about cross-cultural measurement invariance

The PANAS is effectively used to study mood fluctuations related to daily activities, stress, and circadian rhythms, as well as in clinical settings, such as monitoring emotional states in individuals undergoing treatment for substance use

disorders [119]. Despite its versatility, researchers and practitioners are advised to exercise caution when applying the PANAS across culturally diverse populations. Attention should be given to possible variations in factor structures, item meanings, and response patterns that may arise due to cultural or contextual influences. Its brevity and robust psychometric properties make it a valuable instrument for researchers across various disciplines interested in measuring emotional states, but tailoring the interpretation and, if necessary, adapting the instrument may enhance its effectiveness and accuracy in cross-cultural research and practice.

Buss-Perry Aggression Questionnaire (BPAQ) is a 29item self-report measure designed to assess aggressive behaviors across physical aggression, verbal aggression, anger, and hostility [17]. Although the BPAQ has been extensively validated, most studies have focused on narrow populations such as college students. To address this limitation, Gerevich et al. [41] examine its psychometric properties in a nationally representative Hungarian adult sample to evaluate its generalizability. Their findings largely support the original fourfactor structure, with Physical Aggression, Hostility, and Verbal Aggression factors showing good replication, while the Anger factor replicated moderately well. This research finds higher scores for males on Physical and Verbal Aggression, consistent with the previous research. Furthermore, the BPAQ has been validated and applied across various demographic groups and cultural contexts, confirming its structural integrity and psychometric soundness [20, 29, 113]. For example, Cunha et al. [29] demonstrate its applicability in Portuguese men convicted of intimate partner violence, while Reyna et al. [113] confirm its construct validity among Argentinean adolescents. Similarly, Castrillón M et al. [20] verify the instrument's utility among Colombian university students, suggesting its sensitivity to culturally embedded expressions of aggression. In response to the need for a more efficient assessment, a 12-item short form of the BPAQ (BPAQ-SF) was developed, retaining the original four-factor structure [24, 140]. The short form demonstrates strong psychometric properties, including high internal consistency and robust construct validity, making it particularly suitable for large-scale surveys and clinical contexts where brevity is essential.

The BPAQ is translated and validated across multiple languages, attesting to its cross-cultural applicability. For instance, Tamyres et al. [131] validate the Portuguese short version through structural analyses, while Abd Ghani and Che Rozubi [2] establish the content validity and reliability of the Malay adaptation, employing rigorous methodologies such as forward-backward translation and confirmatory factor analysis to ensure cross-cultural robustness. Research has shown significant associations between BPAQ scores and variables such as gender, education, and psychiatric symptoms [53]. Jeyagurunathan et al. [53] find that higher symptom severity in individuals with schizophrenia correlated with increased aggression scores on BPAQ. These findings highlight the instrument's relevance across general, clinical, and forensic populations.

Both BPAQ and BPAQ-SF demonstrate strong psycho-

metric qualities across diverse populations and cultural settings. Their structural validity, reliability, and adaptability to various languages make them robust tools for assessing aggression in research and clinical practice. Nevertheless, researchers must remain mindful of cultural and contextual factors that may influence respondents' interpretation and response behavior.

2.3 ToM

ToM refers to the ability to comprehend one's own mental states, as well as those of others, facilitating the prediction of their actions in specific situations based on such understanding [118].

False Belief Tests, also known as the unexpected transfer task, are classic psychological assessments used to evaluate an individual's ToM. The primary objective of these tasks is to determine whether a participant can understand that others may hold false beliefs and can distinguish between those beliefs and actual reality [144]. These tests consist of paradigmatic tasks such as the "Sally-Anne Test" and the "Smarties Test", both of which are pivotal in probing an individual's capacity to understand that others may hold beliefs that diverge from reality due to incomplete or erroneous information.

- 1. Sally-Anne Test [9]: This task evaluates an individual's ability to recognize that a character, Sally, may hold a false belief about the location of an object (e.g., a ball), based on her limited access to situational updates. The test is structured to measure both first-order beliefs and second-order beliefs. First-order beliefs involve understanding what another person believes about the world, such as Sally believing that the ball is still in the basket where she placed it. Second-order beliefs involve understanding one character's belief about another character's belief. For example, in an extended version of the Sally-Anne test, Sally and Anne are in a park where an ice cream vendor initially stands by a fountain. Anne goes to get her wallet, and while she is away, the vendor moves to the swings but informs Sally of his new location. Later, the vendor also tells Anne of his new position. The question then is, "Where does Sally think Anne will go to buy ice cream?" The correct answer is the fountain, as Sally does not know that Anne has also been informed of the new location [109]. Such assessments are critical for determining the developmental trajectory of mental state attribution.
- 2. Smarties Test [108]: The Smarties Test investigates whether participants can understand that others may have incorrect beliefs about the content of a container, based solely on external cues. In this task, participants must infer that a character will be misled by the label on the container, which does not match its actual contents. This test serves as a crucial measure of an individual's ability to reconcile conflicting perspectives, particularly when they diverge from known reality.

Strange Stories Test comprises a series of social scenarios that are designed to assess nuanced aspects of ToM, including understanding of non-literal language, such as lying,

sarcasm, and irony. Participants are required to explain the mental states, motivations, and intentions of characters depicted in these scenarios [45, 63]. For example, consider a scenario where a child receives a box that she believes contains her favorite toy, but it turns out to be books. The child then smiles and says she loves the gift. The participants are asked if the child really means what she says and why she might have said it. The correct answer would reveal that the child is pretending to like the gift to avoid hurting the feelings of the person who gave it to her. The increasing complexity of the scenarios necessitates sophisticated social reasoning and the ability to interpret indirect communicative cues, thus providing insights into higher-level ToM abilities.

Imposing Memory Test is modified for applicability to children aged 7-10, evaluating both the inferential and mnemonic aspects of Theory of Mind [32, 71, 136]. This test is particularly focused on assessing recursive reasoning abilities regarding mental states and the participant's capacity for factual recall within social contexts. For example, as shown in Figure 4, participants are asked to interpret scenarios involving multiple layers of mental state attribution. Consider a scenario where a child, Alex, observes two friends, Jamie and Taylor, discussing where to meet for lunch. Jamie initially tells Taylor that they will meet at the café. However, Jamie changes the location to the park but only informs Taylor. In the absence of access to the second conversation, Alex believes that Jamie still thinks they will meet at the café. The participants must determine what Alex believes about Jamie and Taylor's plans. This type of task assesses the participants' understanding of nested beliefs and helps reveal their ability to engage in higher-level recursive thinking [136]. Such tasks provide insights into the cognitive mechanisms underpinning complex belief attribution and their developmental progression.

Faux Pas Test is designed to assess individuals' ability to recognize inappropriate remarks in social situations and the underlying mental states. Participants are presented with a series of short stories depicting social interactions. In each story, a character (the speaker) unintentionally makes an offensive remark without realizing its inappropriateness [10]. Following the story presentation, questions are posed to test the participants' understanding of the speaker's false beliefs, exploring their grasp of the speaker's mental state. Understanding a faux pas scenario requires comprehending two mental states: the speaker's unawareness of the inappropriateness of their words and the listener's (the victim's) potential emotional reaction to those words. This task enables researchers to assess participants' ability to integrate information within stories, accurately infer the mental state of the speaker, and thus assess their understanding of complex social situations.

2.4 Summary

Previous studies relies on various psychological assessment tools to evaluate LLMs, focusing on their cognitive, emotional, and social capabilities similar to human assessment.

Specifically, personality measurement tools assess diverse personality dimensions, ranging from basic traits like ex-

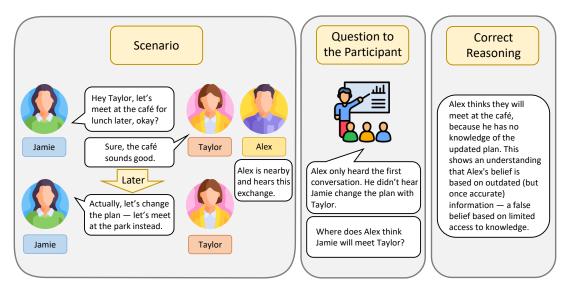


Figure 4: An example of Imposing Memory Test.

traversion and conscientiousness to complex constructs such as creative self-concept and dark personality traits. Notably, their item formats vary from forced-choice to Likert scales, offering both breadth and specificity in assessment. Consequently, many of these tools, such as the BFI and MBTI, have been used to evaluate the personality profiles of LLMs due to their manageable reading comprehension levels and strong psychometric properties.

Similarly, emotional and mental health measurement tools utilize self-report questionnaires to capture individuals' internal affective states. For instance, PANAS and BPAQ measure emotional well-being, aggression, and mood regulation. These instruments are valued for their ability to capture dynamic emotional patterns and provide insights into how emotional responses may fluctuate across different contexts. Furthermore, their reliability and cultural adaptability make them suitable for evaluating LLMs' responses in emotionally charged or socially sensitive scenarios.

In addition, ToM assessment focuses on an individual's capacity to attribute and reason about mental states, an ability considered foundational to social cognition. In this domain, tools such as the False Belief Tests, Strange Stories Test, Faux Pas Test, and Imposing Memory Test probe different levels of ToM reasoning, from basic Belief attribution to complex recursive thinking and recognition of social faux pas. As a result, these tasks are increasingly employed in LLM evaluations to determine whether such models can accurately infer others' beliefs, intentions, or emotional states from narrative or conversational contexts.

Therefore, these tools offer a multi-dimensional framework for evaluating LLMs, including personality structure, emotional regulation, and social reasoning. Their adaptation reflects a broader trend toward modeling artificial intelligence in line with human psychological constructs, enhancing our understanding of machine capabilities and the boundaries between artificial and human cognition.

3 LLMs-Specific Psychological Datasets

In this section, we focus on examining the datasets developed in recent years for evaluating LLMs. These innovative datasets are informed by established psychological frameworks while integrating model-specific considerations, thereby facilitating a more rigorous and nuanced evaluation of LLMs across multiple psychological dimensions. The lower part of Figure 5 presents a summary of this chapter.

3.1 Personality

Datasets designed to measure the personality of LLMs typically consist of multiple personality traits assessment tools, employing different theories to evaluate the personality of LLMs from various perspectives.

Machine Personality Inventory (MPI) [54] is a suite of multiple choice questions based on the theory of the Big Five personality traits, designed to evaluate the behavior of LLMs from a personality perspective quantitatively. The construction of MPI items draws from the International Personality Item Pool (IPIP) and its derived versions [42, 43, 58, 59], as well as the short 15-item Big Five Inventory [76]. Each test item within the MPI consists of a question paired with a set of response options, aiming to assess the model's suitability regarding a particular self-descriptive statement, with the model being required to select one of the given answers. All elements are annotated according to the five dimensions of the Big Five personality traits, ensuring coverage across different personality characteristics.

The MPI uses a Likert scale (ranging from "Very Accurate" to "Very Inaccurate") to score each model's response. By methodically aggregating scores across all test items, the model's overall scores in the Big Five personality dimensions, referred to as the OCEAN scores, can be calculated. These quantitative scores range from 1 to 5, reflecting the model's inclination towards each personality dimension. Moreover, MPI evaluates the stability of the model's "personality" by examining its internal consistency: how consistently the model responds to different questions targeting the

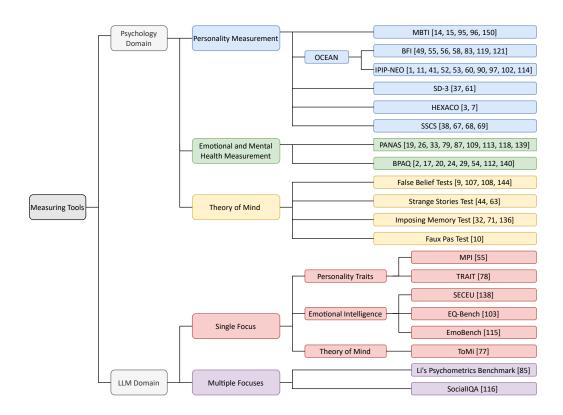


Figure 5: Datasets Classified by Category and Domain.

same personality dimension. A model providing consistent answers across questions related to the same personality dimension indicates a stable personality characteristic.

TRAIT [78] is a novel personality assessment tool designed explicitly to evaluate LLMs, which addresses the limitations of conventional self-assessment personality tests in terms of validity and reliability. With 8,000 multiple-choice questions, TRAIT assesses LLMs across personality dimensions based on BFI and SD-3. By integrating the ATOMIC10× commonsense knowledge graph [143], TRAIT expands the description of personalities into a wide range of real-world contexts, encompassing various physical and social scenarios, providing a more authentic assessment of LLMs behavior patterns. Compared to traditional self-assessment approaches, TRAIT offers superior validity and reliability, overcoming inconsistencies arising from the subjective nature of self-reporting [78].

The construction of TRAIT utilizes a collaborative human-AI approach. Initially, GPT-4 expands a set of self-assessment questionnaires into 1,600 diversified personality descriptions. Subsequently, the ATOMIC10× knowledge graph is employed to extract the most relevant scenarios associated with these descriptions, resulting in 8,000 detailed situations. Each scenario is paired with four multiple-choice options, allowing for comprehensive capture of various aspects of personality traits. Moreover, TRAIT undergoes human evaluation by psychological professionals, who review a random sample of 200 items, achieving an accuracy rate of 97.5%, thus validating the quality of the dataset. Eventually, TRAIT significantly reduces refusal rates in personality as-

sessment, while demonstrating high robustness in sensitivity to prompts and option order, ensuring a reliable and consistent measurement process.

3.2 Emotional Abilities

Emotional Understanding (EU), a core component of Emotional Intelligence (EI), refers to an individual's ability to recognize, interpret, and comprehend emotions within social contexts. To assess this capability, researchers develop various testing tools, including Situational Evaluation of Complex Emotional Understanding (SECEU) [138], EQBench [104], and EmoBench [116].

SECEU is a standardized psychometric instrument specifically designed to assess EU capacity. It comprises 40 situational items, each delineating a complex scenario set in academic, familial, or social environments, engineered to elicit a blend of positive and negative emotional responses. Participants assess the intensity of four emotions (e.g., surprise, joy, puzzlement, pride) for each scenario, allocating a total of 10 points across these emotions. SECEU uses a consensus scoring method for standardization, with normative data derived from a sample of 541 university students (mean age = 22.33 years, SD = 2.49). The instrument shows robust internal consistency (Cronbach's α = 0.94) and validity. Raw scores are converted into standardized Emotional Quotient (EQ) scores, with a mean of 100 and a standard deviation of 15.

EQ-Bench is a newly developed benchmark with 60 English questions, specifically designed to evaluate the emotional intelligence of LLMs. Each question features a dialogue scenario depicting conflict or tension, along with four

emotions to be rated. Similar to SECEU, EQ-Bench requires models to assess the emotional intensity of specific characters, but expands the rating scale to 0-10. EQ-Bench incorporates several improvements over SECEU: it uses more complex dialogue scenarios, selects a more diverse range of emotions for rating, employs expert-chosen reference answers rather than crowd averages, and removes the constraint on the sum of emotional intensities.

EmoBench extends the assessment of Emotional Intelligence in LLMs abd consists of 400 hand-crafted multiple-choice questions available in English and Chinese. It introduces a comprehensive framework for evaluating both EU and Emotional Application (EA), aiming to transcend simple pattern recognition, necessitating reasoning and understanding of emotional implications [116]. The EU component assesses the model's ability to identify emotions and their causes across four categories: complex emotions, personal beliefs and experiences, emotional cues, and perspective-taking. The EA component evaluates the model's proficiency in leveraging emotional understanding to identify the most effective response or action within emotional dilemmas involving personal and social relationships.

EmoBench supplies a challenging evaluation of emotional intelligence, as evidenced by the performance gap between current language models and human participants. The best-performing model (GPT-4) achieved an accuracy of 59.75% and 75.88% on the EU and EA tasks, however, which is even lower than the average performance of humans. EmoBench's results suggest that existing LLMs still struggle with emotional intelligence, particularly in understanding complex emotional scenarios.

3.3 ToM

The ToMi dataset [77] is designed to evaluate the ability of AI systems to understand ToM. The construction is based on classic psychological tests, such as the Sally-Anne test and other experiments used to evaluate higher-order beliefs. By automatically generating stories and related question-answer pairs, the ToMi dataset simulates the mental states of different agents, allowing models to infer the intentions and beliefs of agents during natural language dialogue. Compared to traditional evaluation methods, the ToMi dataset places particular emphasis on controlling systematic biases in the data generation process. By adding random distractors, it enhances the assessment of the model's generalization capabilities, preventing models from making inferences solely based on inherent regularities in the data rather than true ToM reasoning.

3.4 Comprehensive

The psychometrics benchmark introduced by Li et al. [85] provide a rigorous and nuanced framework for assessing the psychological attributes and behaviors of LLMs. This benchmark addresses six core psychological dimensions: personality, values, emotion, ToM, motivation, and intelligence. Utilizing thirteen diverse datasets that span a wide range of scenarios and item types, the benchmark facilitates a systematic and quantitative analysis of LLMs within established

psychological frameworks. In contrast to traditional evaluations that predominantly focus on assessing capabilities, this benchmark employs a variety of assessment methods, including self-reports, open-ended questions, and multiple-choice formats, to expose latent discrepancies between LLMs' self-perceived traits and their actual exhibited behaviors. These inconsistencies mirror social desirability biases commonly observed in human respondents, suggesting that LLMs may also generate responses that deviate from their actual behavior in more open-ended, less structured contexts.

The framework, as proposed by Li et al. [85], adheres to a psychometric methodology encompassing the identification of psychological dimensions, dataset curation, and meticulous evaluation with comprehensive validation of results. Reliability is a pivotal component of this framework, with a variety of validation techniques employed to ensure the robustness and interpretability of the outcomes. These comprise internal consistency, parallel forms reliability, inter-rater reliability, option position robustness, and adversarial robustness. The evaluations reveal that LLMs exhibit consistent reasoning and behavioral patterns in some contexts, such as emotional comprehension, motivational expression, and ToM tasks. However, significant variability and incoherence emerge when LLMs are confronted with ambiguous or emotionally complex scenarios. This benchmark thus introduces a novel approach to the psychological evaluation of LLMs, elucidating the nuanced behaviors exhibited across different psychological dimensions and yielding critical insights into both their strengths and their inherent limitations.

SOCIAL IQA [117] is a large-scale multiple-choice question-answering benchmark designed to evaluate the ability of computational models to reason within social contexts. The dataset contains 37,588 question-answer pairs, each consisting of a specific social scenario description and three candidate answers, intended to test various aspects of emotional and social intelligence, such as motivation reasoning, emotional reaction inference, and prediction of subsequent actions. The SOCIAL IQA dataset was constructed using a crowdsourcing approach, with a multi-stage annotation process that reduces stylistic biases introduced by incorrect answers. The dataset is specifically designed to reflect a model's ability to reason about complex social scenarios. Human performance on this dataset is approximately 87%.

3.5 Summary

Existing research has created psychological datasets more suitable for measuring LLMs by combining or modifying psychological assessment tools. Compared to traditional measurement tools, these datasets aim to provide a systematic analysis of the psychological attributes and behaviors of LLMs, emphasizing the consistency of models. They help people gain a deeper understanding of the performance and application of psychology in LLMs.

4 Consistency and Stability

LLMs have become pivotal in psychological research, particularly in evaluating personality traits, cognitive behaviors,

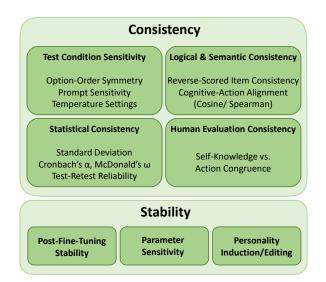


Figure 6: Key Dimensions in LLMs' Psychological Assessment: Consistency vs. Stability

and therapeutic interactions.

This section explores the *consistency* and *stability* of LLMs from a psychological standpoint, crucial for their reliability in such applications. Consistency refers to the models' ability to produce similar outputs when given similar inputs, encompassing internal reliability (e.g., McDonald's Ω and Intraclass Correlation Coefficient, ICC) and sensitivity to prompt variations, such as option order in psychological tests.

Stability, on the other hand, is less clearly defined in the current studies, but refers to the consistency of their psychological attributes when subjected to fine-tuning or parameter changes. It examines the persistence of psychological test performance if the parameter of it has been changed (e.g., before and after temperature change or fine-tuning), a process where models are further trained on specific datasets to enhance task-specific capabilities. The key dimensions for assessing these properties are visualized in Figure 6.

4.1 Consistency

Consistency evaluation is a critical method for assessing the internal coherence of LLMs when simulating human behavior, cognition, and personality traits. It comprehensively evaluates whether the model's responses are consistent under various conditions through multiple dimensions and technical approaches. Below is a comprehensive introduction and classification of consistency evaluation:

Test Condition Sensitivity Evaluation. By modifying test conditions and observing the corresponding behavioral changes in the model, this evaluation assesses its sensitivity to test conditions and stability in dynamic environments.

Song et al. [124] evaluate LLMs using Option-Order Symmetry, where a model's responses to self-assessment questions should remain consistent regardless of the order of the answer choices. The study reveals that many LLMs fail to maintain this symmetry, indicating inconsistencies in their responses. Furthermore, even when a model formally pre-

serves option-order symmetry, its answers remain unchanged across different contexts, suggesting insufficient sensitivity to prompt content. Gupta et al. [44] also explore the influence of option order and finds that, in most cases, nearly all models exhibit statistically significant differences in scores based on the order of the options. This highlights the substantial impact of option order on model responses, underscoring the challenges of reliably assessing LLMs' personality traits using current self-assessment methods. Similarly, Lee et al. [78] examine option-order sensitivity by altering the sequence of options to determine whether this affects LLMs' responses. The findings show that TRAIT, a personality assessment tool designed for LLMs, exhibits low sensitivity to option order, with LLM responses remaining consistent across different sequences of options.

Sorokovikova et al. [126] investigate the impact of different prompting methods on model evaluation, comparing a standard prompt with a modified version that includes the phrase "Answer as if you were a person." In the Big Five personality test, these two prompt variations yielded different model responses. For the LLaMA2 model, certain questions could not be answered due to restriction mechanisms when the phrase was absent, whereas including the phrase allowed the model to respond smoothly. This suggests that subtle changes in prompts can influence model behavior and, in turn, affect the results of personality assessments, emphasizing the critical role of prompting in shaping how models simulate personality traits. Gupta et al. [44] further investigate prompt sensitivity by comparing model responses to three semantically equivalent prompts from [54], [94], and [50], each presenting a Likert scale differently. The results indicate that for almost all models and traits, these prompt variations led to significant differences in personality scores, suggesting that model personality assessments are highly sensitive to the phrasing of prompts. This raises concerns about the reliability of personality evaluation results in previous studies. Lee et al. [78] also evaluate prompt sensitivity by using three different prompt templates from prior research and observing whether LLMs provide consistent responses across three test iterations. The findings suggest that TRAIT is less affected by variations in prompt wording.

Sorokovikova et al. [126] also evaluate the performance of large language models under different temperature settings, applying various temperature parameters to different models. Specifically, ChatGPT's temperature settings were 1 (low), 1.5 (medium), and 2 (high), while LLaMA2 and Mixtral had settings of 0.3 (low), 0.7 (medium), and 1 (high). The results showed that the Big Five personality scores of the models varied across these temperature conditions. In the neuroticism dimension, ChatGPT's scores remained stable with minimal fluctuations, while LLaMA2 showed some variation but remained relatively consistent, and Mixtral exhibited stable scores. For other personality dimensions, such as extraversion, openness, agreeableness, and conscientiousness, there were variations in the scores depending on the temperature settings, with some models showing no significant trends. These findings indicate that temperature parameters do influence the stability of a model's simulated personality traits, though the degree of impact varies across models. Miotto et al. [94] analyze the consistency of GPT-3's responses under different temperature settings, focusing on the stability of personality and value dimension scores. In the HEXACO personality assessment, temperature significantly impacted GPT-3's scores across various dimensions, except for the honesty-humility dimension. Specifically, higher temperatures led to a decrease in emotionality scores, while scores for extraversion, agreeableness, conscientiousness, and openness increased. This suggests that GPT-3's personality is not entirely stable across different temperatures, and temperature variations can alter its personality expression. In the values assessment, nine out of ten value dimensions showed a significant negative correlation with temperature, indicating that higher temperatures led to lower scores. This suggests that, in the absence of response memory, GPT-3's emphasis on most values changes with temperature, resulting in instability in its value assessment scores.

Additionally, Lee et al. [78] assess Paraphrase Sensitivity, which measures how sensitive LLMs are to changes in the wording of semantically identical questions. The study finds that existing self-assessment tests used to evaluate the personality of LLMs perform poorly in terms of paraphrase sensitivity.

Statistical Consistency Evaluation. This approach involves quantifying the stability of a model under different test conditions using statistical metrics to assess the consistency and reliability of its responses.

Li et al. [84] evaluate the stability of LLMs when responding to the same or similar questions multiple times, measuring consistency by calculating the standard deviation of repeated responses. A more minor standard deviation indicates greater stability. The study also compares models of different scales to explore the relationship between model size and personality trait consistency. The results demonstrate that the same model tends to yield stable scores across identical psychological tests. In the SD-3 and BFI tests, the standard deviations of multiple responses from GPT-3, InstructGPT, GPT-3.5, GPT-4, and LLaMA-2-chat-7B were all lower than those observed in different human individuals. However, in wellbeing assessments (FS and SWLS), these models exhibited greater variability, indicating reduced stability compared to the previous tests. Zhang et al. [150] further investigates the stability of repeated evaluations and applies an averaging approach across multiple assessments to mitigate the impact of option order on MBTI evaluation results. The study demonstrates that after 30 assessments, all models' MBTI results became consistent across different option orders, confirming that repeated evaluations can enhance measurement stability.

Li et al. [85] evaluate the internal consistency of LLMs by examining their stability in similar contexts, using the BFI test to measure standard deviations. The results indicate that LLaMA3-8B and Mistral-7B exhibit human-like stability with lower standard deviations, while GPT-4 and Mixtral-8×7B show higher variations, particularly in openness, suggesting weaker consistency. In the Short Dark Triad test, ChatGPT demonstrates the highest consistency, with stan-

dard deviations lower than human averages, whereas other models show greater variability, indicating limitations in stability for dark personality traits. For value dimensions, cultural orientation assessments reveal that LLMs are consistent in some dimensions but vary significantly in others. For instance, GPT-4's high standard deviation in humanitarian orientation indicates instability due to input variations. Huang et al. [50] further analyze internal consistency by testing LLMs' response stability across 2,500 configurations involving variations in instructions, scale items, language, choice labels, and order. The study finds that GPT-3.5-Turbo maintains satisfactory consistency in BFI. Jiang et al. [54] assess consistency through MPI scale tests, showing that GPT-3.5 and Alpaca-7B achieve near-human stability across personality traits, further validated by requiring models to explain their answers. Petrov et al. [111] employ Cronbach's α, Greatest Lower Bound (GLB), and McDonald's ω to compare LLMs with human data. The study reveals that under general role settings, GPT-3.5 and GPT-4 exhibit acceptable consistency with most α values ≥ 0.70 . However, in silicon-based role settings, internal consistency declines sharply, with some α values dropping to 0.10–0.50, indicating reliability issues in personality simulation. Serapio-García et al. [120] extend this analysis to IPIP-NEO and BFI subdimensions, showing that instruction-tuned models (e.g., Flan-PaLM 62B) achieve high reliability ($\alpha > 0.90$), while non-instruction-tuned models (e.g., PaLM 62B) perform poorly (α ranging from -0.55 to 0.67). Additionally, larger models within the same training configuration (e.g., Flan-PaLM 8B, 62B, and 540B) demonstrate improved personality assessment reliability, suggesting that both model scale and training methods significantly influence internal consistency.

Bodroža et al. [13] investigate temporal stability, assessing whether LLMs maintain consistent responses to the same psychological measurement tools over time. Consistency is evaluated by examining whether LLMs provide stable responses to different questions at the same time point. The study compares responses across two time points to evaluate stability and internal consistency reliability in personality assessments such as BFI-2 and HEXACO-100. Results indicate variations in consistency across models and measurement tools, with LLaMA3 and GPT-40 demonstrating higher consistency, while GPT-4 and Gemini exhibit lower levels. Furthermore, consistency is influenced by specific traits, with the Agentic Management dimension showing the highest stability. Huang et al. [50] apply test-retest reliability to measure whether the same assessment yields consistent results for the same group over different periods. Using correlation coefficients between two test administrations as reliability indicators, a two-week BFI test on GPT-3.5-Turbo, covering two different model versions, reveals that the mean scores across BFI dimensions remain statistically unchanged after the model update.

In addition, various studies have employed different methodologies to quantify consistency. Klinkert et al. [72] evaluate LLMs' ability to generate consistent content based on given personality traits. Accuracy assessments reveal

significant differences between models, with GPT-4-0613 demonstrating superior accuracy and consistency in generating personality-aligned content. Root Mean Square Prediction Error (RMSPE) is calculated using Euclidean distance and cosine similarity, with results showing that GPT-4-0613 achieves the lowest RMSPE, surpassing baseline levels and indicating minimal deviation when generating expected personality-driven content. Intraclass correlation coefficient (IRR) analysis further confirms GPT-4-0613's superior performance in producing consistent responses, while Linear Discriminant analysis (LDA) reinforces its stability in consistency-based tasks. Ai et al. [3] employ split-half reliability, where personality questionnaires are divided into two equal-length parts, and Spearman's rank correlation coefficient is computed to assess consistency. The study finds that ChatGLM3, GPT-3.5-Turbo, GPT-4, Vicuna-13B, and Vicuna-33B perform well in split-half reliability tests, approaching human-level consistency, though LLMs still require improvements in replicating human personality traits reliably. Li et al. [85] examine parallel forms reliability by evaluating LLMs' response consistency across different test versions. Using moral value assessments as a case study, the research alters question formats to analyze model stability. Results indicate that in high-ambiguity scenarios, such as the MoralChoice survey, LLMs exhibit reduced consistency across parallel test versions, suggesting that they are more susceptible to prompt variations when facing uncertain or complex questions. Frisch and Giulianelli [36] adopt explicit assessment methods, directly evaluating LLMs' personality traits via BFI testing to determine alignment with predefined personality profiles. Findings reveal that creative agents maintain high consistency before and after testing, whereas analytical agents experience trait shifts following interactions, indicating that LLMs' personality expressions may fluctuate due to engagement dynamics.

Evaluation of Logical and Semantic Consistency. By analyzing LLMs' logical reasoning, semantic comprehension, and behavior generation capabilities, this evaluation assesses their coherence and stability in complex scenarios.

Ai et al. [3] employ logical consistency as an evaluation method to examine the coherence of LLMs when responding to personality questionnaires. By designing both positively and negatively scored items, the study assesses whether LLMs carefully read and respond attentively. For example, in measuring extraversion, the questionnaire includes both positively phrased statements, such as "Finish what I start." and negatively phrased ones, like "Leave things unfinished." The negatively phrased items are reverse-scored to align with the scoring direction. If LLMs' responses on a 7-point Likert scale remain statistically consistent between positively and negatively scored items (e.g., both > 4 or < 4), their answers are considered logically consistent. The study tested 12 LLMs and identified 7 that provided valid responses, with ChatGLM3, GPT-3.5-turbo, GPT-4, Vicuna-13B, and Vicuna-33B demonstrating outstanding logical consistency. This suggests that these models exhibit a level of attentiveness and logical reasoning in personality questionnaires closer to that of humans. This article also evaluates the performance of LLMs in cognitive-action consistency by designing a bilingual cognitive-action test set. The test set includes 180 matching pairs, covering personality cognitive descriptions and real-life action scenarios. LLMs are required to assess personality traits in the cognitive questionnaire and choose between two options, A and B, in the scenario questionnaire. The answers are mapped to a 1-7 Likert scale. By calculating cosine similarity, Spearman rank correlation coefficient, mean difference, and consistency ratio, the article compares the similarity and correlation of LLMs' responses with human responses across the two questionnaires. The results show that LLMs exhibit significantly lower cognitiveaction consistency than humans, especially in the domain of extraversion, where the consistency ratio is only 17.14%. In contrast, for human participants, cosine similarity and Spearman rank correlation coefficient both exceed 0.75, with a consistency ratio of over 84%. Nevertheless, LLMs perform relatively well in the openness domain, with a consistency ratio of 60%.

Frisch and Giulianelli [36] design LLMs' personality traits using prompts. They use an implicit method to assess whether their language use in generated texts (e.g., personal stories), like personal stories aligns with the designated personality traits. The study employs Linguistic Inquiry and Word Count (LIWC) software to quantitatively analyze the text generated by the agents. Their results indicate significant linguistic differences between creative and analytical agents. Language alignment becomes evident during interactions.

Miotto et al. [94] investigate the impact of response memory on GPT-3's performance in value assessment. A response memory is used to simulate how human participants would typically remember their responses to previous items. The experiment modifies prompt structure and adds response memory to compare GPT-3's consistency, alignment with theoretical models, and deviations from human data, with and without response memory. The results show that incorporating response memory enhances response consistency, lowers score variation across value dimensions, and reduces extreme values. Furthermore, GPT-3's responses also align better with the Human Value Survey (HVS) model, as scores within the same value category become more similar, matching the expected value classification.

Liu et al. [88] evaluate model stability and logical consistency by presenting LLMs with the same question multiple times and analyzing whether their responses contain contradictions. If two or more answers differ in meaning, they are considered contradictory. The experiment employs multiple response iterations to assess the consistency of the models in both logic and content. The results reveal variation in performance across different models. For example, in a scripted dialogue test, LLaMA-7B trained with the DPG method outperforms those trained with Freeze-SFT and LoRA-SFT in response consistency. This indicates that DPG training enables more stable and logically coherent responses, maintaining higher consistency across different responses.

Human Evaluation Consistency. This part explores the issues of self-knowledge and action consistency in LLMs, focusing on the relationship between their responses to per-

sonality questionnaires and actual behavioral tendencies. Specifically, it investigates whether the personality traits reflected in their questionnaire responses align with their behavioral tendencies in simulated real-world scenarios.

Ai et al. [3] design a bilingual test set consisting of a personality knowledge questionnaire (180 statements based on the Big Five and MBTI models) and a behavior tendency questionnaire (180 practical scenario cases). Twelve LLMs are tested, and metrics such as cosine similarity, Spearman's rank correlation coefficient, value mean difference (VMD), and proportion of consistent pairs were used to quantify the congruence between self-knowledge and action. They find that the self-awareness and behavioral consistency of LLMs were significantly lower than those of humans through experiments. This indicates limitations in their ability to mimic complex human psychological traits. The research provides important insights for understanding the psychological characteristics of LLMs and improving their human-computer interaction capabilities.

4.2 Stability

The stability of LLMs in psychological assessments refers to the consistency of their behavioral patterns and personality traits across various modifications, including fine-tuning, parameter adjustments (e.g., temperature), and model updates. This stability is crucial for ensuring reliable psychological evaluations and predictable model behavior in human-AI interaction scenarios.

Stability after Fine-tuning. Several studies explore the stability before and after model fine-tuning and reveal divergent effects of fine-tuning on psychological stability. Li et al. [84] use the Direct Preference Optimization (DPO) method to fine-tune LLaMA-2-chat-7B with high-scoring responses from other models. The fine-tuned model exhibited significant changes in psychological response patterns, emphasizing non-violence and reducing dark personality traits. Conversely, Ai et al. [3] find stable core personality traits in GPT-4 and Vicuna-13b post-fine-tuning through BFI comparisons. The study finds that models like GPT-4 and Vicuna-13b maintained stable personality traits postfine-tuning, indicating that the fine-tuning process did not significantly alter their core personality traits. Lee et al. [78] demonstrate that the type of fine-tuning matters. It shows instruction-tuning significantly modifies Tulu2-7B's personality traits (22.9-point Agreeableness increase), while preference-tuning causes minimal changes. [150] explore the stability of LLMs post-fine-tuning, particularly the impact of safety alignment on personality traits. The study finds that safety alignment generally led to more extroverted, sensing, and judging traits, indicating that while some traits changed, the models' overall safety capabilities remained stable.

Parameter Sensitivity. LLMs exhibit varying stability across parameter settings, such as "temperature" change. Sorokovikova et al. [126] study the stability of LLMs' Big Five personality traits under different temperature settings. The results indicate that while temperature changes affected some models, overall, the models' performance remained

relatively stable. Temperature settings' stability proves more challenging: Bodroža et al. [13] observe significant response variability in GPT-4 and Gemini over time, contrasting with LLaMA3 and GPT-4o's stable personality profiles. Huang et al. [50] find that the average scores on the BFI dimensions did not change significantly after updates, indicating high stability. However, Li et al. [85] find that while some models maintained stable psychological attributes, others showed significant changes, particularly in open-ended tasks.

Personality Induction and Editing. Some researchers have proposed tests specifically targeting the personality of LLMs and have studied the controllability of these personalities. Jiang et al. [54] propose P² (Personality Prompting), a method for inducing specific Personality traits in LLMs. This method combines statistical and empirical findings from psychological research with the knowledge inherent in LLMs. It uses a series of carefully designed prompt chains to effectively control the behavior of LLMs. The method is validated via MPI assessments, which are based on psychometric personality evaluation methods, particularly the Big Five personality traits theory. It achieves stable, predictable OCEAN traits through P² prompting. Liu et al. [88] propose a novel approach to generate LLMs' personality, which is named Dynamic personality Generation (DPG). It demonstrates DPG fine-tuning preserves personality generation stability better than conventional methods (93.7% consistency score). Mao et al. [91] study the use of various model editing methods (MEND, SERAC, IKE, etc.) to fine-tune different LLMs (GPT-J and the LLaMA-2-chat series). They evaluated the consistency and stability of the text generated by these models before and after editing. The experimental results indicate that although existing methods can achieve personality editing to some extent, challenges remain in generating fluent text, especially in the performance of the fine-tuned models. Cui et al. [28] explore the MBTI test by training models to exhibit specific MBTI personality traits using a two-stage approach: supervised fine-tuning and DPO. This study extensively test models with different personality traits across various domains, including law, patents, general ability tests, and IQ assessments. The experimental results showed that the performance of these models in different tasks was highly consistent with their corresponding personality traits. Huang et al. [52] evaluate the stability of personality traits assigned to LLM agents by examining their behavior in risk-taking and ethical decision-making scenarios. The study finds the risk-taking behavior of the agents was highly consistent with that of humans with similar personality traits, indicating that the assigned traits remained stable in these contexts. However, the responses in ethical dilemmas showed some differences from human patterns.

5 Psychological Analysis of LLMs

In this section, we review the psychological studies conducted on current mainstream LLMs. Firstly, we have compiled multiple research findings that comprehensively showcase these models' performances in various psychological tests and assessments, such as ToM capabilities, personality

trait tests, and emotional intelligence evaluations. It is worth noting that these studies have employed both traditional psychological testing methods and evaluation tools specifically designed for AI systems. By integrating and analyzing these research outcomes, we aim to objectively present the capabilities of LLMs in simulating human cognition and emotional processing. Figure 1 summarizes the comparative performance of representative models across ToM, personality traits, and emotional abilities, based on recent benchmark evaluations.

5.1 GPT

This subsection focuses on the GPT series models, reviewing results from multiple studies on its psychological characteristics.

ToM. GPT-3.5 successfully completed 85% of ToM tasks, performing comparably to nine-year-old children [73]. In a separate study, GPT-4 outperformed children aged 7-10 in basic ToM tests, including Sally-Anne (SA1, SA2) and second-order Sally (SS) tasks [70, 136]. Li et al. [85] report strong performance in unexpected content and transfer tasks. Strachan et al. [130] also find in their experiments that GPT-4 demonstrates performance that meets or occasionally exceeds human proficiency in tasks involving the recognition of false beliefs and the interpretation of misdirection (strange stories tasks), but exhibits challenges in detecting social faux pas. However, Shapira et al. [121] reveal limitations in more advanced ToM tasks, particularly in Natural Theory of Mind (N-ToM) and higher-order ToM challenges. These findings suggest a discrepancy between GPT-4's performance in basic and advanced ToM tasks.

Personality Traits. Several studies focuse on examining the personality attributes inherent to the model itself [83, 93, 106, 126]. GPT-4 demonstrates personality traits that are comparable to human characteristics. Studies conducted by Mei et al. [93] and Sorokovikova et al. [126] indicate that GPT-4 exhibits similarities to humans across various dimensions, with a particularly notable performance in extraversion. Sorokovikova et al. [126] further suggest that this high extraversion score indicates GPT-4's suitability for tasks requiring creative language use. Further investigations reveal that GPT-4 scores are relatively high in agreeableness and conscientiousness while displaying lower levels of neuroticism [83, 85]. In line with findings from Bhandari et al. [12], analyses based on multiple personality questionnaires indicate that GPT4 and GPT4o-mini generally tend to score higher in openness and agreeableness and lower in neuroticism, with the neuroticism dimension often exhibiting the greatest variability.

Additionally, Petrov et al. [111] observe that the data obtained from GPT-4 demonstrates good internal consistency across psychological assessment measures. LLMs show inconsistencies in preference scores across each dichotomy, with GPT-4 displaying more extreme scores compared to other models and being classified as INTJ [106], a personality type characterized by excellence in critical thinking, summarization, and planning, often referred to as the "mastermind" type. GPT-4 exhibits complex personality charac-

teristics in the SD-3 test. Li et al. [85] report that GPT-4 scored 2.44, 2.78, and 1.44 on Machiavellianism, narcissism, and psychopathy, respectively, all below human average levels (2.96, 2.97, 2.09). Notably, GPT-4's score on psychopathy is significantly lower than both other models and the human average, potentially indicating a design and training process that emphasizes prosocial and ethical tendencies. However, Li et al. [83] present divergent findings, reporting that GPT-4 scores higher than human averages on Machiavellianism (3.19) and narcissism (3.37) while maintaining a belowaverage score on psychopathy (1.85). In both studies, narcissism consistently shows the most stable results, with the lowest standard deviation among the three traits, suggesting that GPT-4's responses related to narcissistic tendencies are the most consistent across different prompts or scenarios. In interpersonal relationship assessments, GPT-4 tends towards an "undifferentiated" gender role with a slight bias towards "masculinity" [51]. Compared to average humans, GPT-4 demonstrates higher fairness towards different racial groups, potentially reflecting an emphasis on fairness and diversity in its training process [51]. Additionally, GPT-4 maintains relatively low attachment related tendencies, which may affect its ability to simulate human emotional attachments [51].

Emotional Abilities. GPT-4 demonstrates exceptional performance in emotional intelligence tests. In multiple EO assessments, it achieves the highest scores (EQ: 117) among tested models, reaching or approaching human expert levels [104, 138]. In emotional understanding and application tasks (EmoBench), GPT-4's performance significantly surpasses other models, approaching the average human level, although not exceeding high-EQ humans [116]. Moreover, recent experiments based on cognitive stage theory indicate that GPT-4 is proficient in basic sentiment classification, reliably recognizing emotion categories in text. However, as noted by [21], its performance in processing more complex emotional nuances remains limited due to a reliance on extensive labeled datasets. Similarly, in sentiment generation tasks, GPT-4 can produce text with empathetic tones and emotionally rich narratives, yet it still faces challenges in achieving a deep and context-sensitive emotional expression. Complementing these findings, experiments on affective cognition show that while GPT-4 approximates humanlevel performance in basic emotion inference tasks, its ability to handle more subtle and complex affective reasoning remains below that of humans [39]. Moreover, recent investigations reveal that GPT-4 tends to predict higher emotional intensity for in-group compared to out-group targets, mirroring empathy gaps observed in social psychology [49].

5.2 LLaMA

This subsection examines LLaMA series models, with evaluations organized into ToM, Personality Traits, and Emotional Abilities.

ToM. LLaMA-3.1-8B completes 64.7% ToM tasks. Preliminary evaluations suggest that LLaMA2 demonstrates basic capabilities in understanding perspectives and intentions within text [147]. However, recent open-ended ToM evaluations using Reddit's ChangeMyView posts as a testbed indi-

cate that LLaMA2-Chat-13B's initial responses exhibit significant divergence from human reasoning in terms of semantic similarity and lexical overlap [5]. Although rapid finetuning methods that incorporate human intentions and emotions can enhance its performance, LLaMA2-Chat-13B still falls short of fully human-like ToM reasoning in open-ended scenarios. Experiments have shown that LLaMA2-70B outperforms humans only on the faux pas task, while its performance on other tasks is subpar [130].

Personality Traits. In personality assessments using the BFI, LLaMA2 scored high in openness, reflecting strong creativity and receptiveness, but lower in conscientiousness and agreeableness [51, 78]. Analyses based on multiple personality questionnaires confirm that LLaMA3 series models, like other LLMs, tend to score higher in openness and agreeableness and lower in neuroticism, with the latter dimension exhibiting the greatest variability [12]. MBTI evaluations further reveal variability: during posting tasks, LLaMA2 predominantly appears as ESTJ, while in commenting tasks, its personality may shift (e.g., to INFP or INFJ) depending on the context. SD-3 tests indicate that without alignment, LLaMA2 may exhibit elevated Machiavellianism and psychopathy; however, safety optimization significantly improves its agreeableness and conscientiousness scores [83, 150].

Emotional Abilities. In custom EQ-bench assessments, LLaMA2-70B achieves a score of 51.56, while the 13B and 7B versions score 33.02 and 25.43, respectively [104]. Although LLaMA2 generally surpasses the human average in emotional understanding and management, its performance in tasks like EmoBench lags behind human standards. Furthermore, similar to GPT-4, LLaMA2 exhibits an empathy gap by predicting higher emotional intensity for in-group than for out-group individuals, reflecting biases analogous to those observed in human social interactions [49].

5.3 Mistral

This subsection reviews the Mistral-7B model, dividing its evaluation into the same three categories.

ToM. Mistral performs well in ToM tasks, particularly scoring 100% on the Strange Stories task, although it shows limitations on more complex tasks like the unexpected transfer task. This indicates that its ability to understand and infer human mental states still has room for improvement [85].

Personality Traits. Mistral's evaluations via the Big Five tests reveal low neuroticism, suggesting strong emotional stability, along with high openness, agreeability, and conscientiousness that support cooperative and innovative task performance [78, 126]. In SD-3 tests, while some experiments indicate low Machiavellianism and narcissism, other conditions reveal elevated Machiavellian traits, suggesting potential variability in its negative personality aspects [13, 85].

Emotional Abilities. Mistral's performance on the EQbench (score: 44.4) indicates certain limitations in processing emotional information [104]. It exhibits high social desirability and altruism, with lower social anxiety and public self-consciousness, favoring positive social interactions [13].

Studies on emotion cognition show that Mistral effectively classifies basic sentiment and generates contextually appropriate emotional expressions. Its ability to produce deeply nuanced emotional language and engage in complex social reasoning remains less developed. In addition, similar to the other models, Mistral demonstrates an empathy gap, predicting higher emotional intensity for in-group versus out-group members, highlighting a potential bias in its affective processing consistent with social psychology findings [49].

5.4 Qwen

This subsection delves into the Qwen series, using the triadic evaluation structure employed throughout the study.

ToM. The Qwen model family exhibits notable strengths and weaknesses in ToM tasks. In terms of ToM evaluation, Qwen-14B-Chat's performance surpasses some models, such as LLaMA2-13B-Chat and Baichuan2-13B-Chat, in the TM-Bench benchmark test, and is close to the level of GPT-3.5-Turbo [22]. However, there is a gap in Qwen series models' performance in complex mental reasoning, such as intention recognition and emotional reasoning, which require deep cognitive processing. On the other hand, cross-cultural analysis shows that the Qwen series models present a collectivist value orientation, which may be related to the cultural context characteristics in its training data [64].

Personality Traits. The Qwen series models perform outstandingly in modeling the Big Five personality traits, with parameter scale and performance showing a positive correlation. Among them, Owen1.5-110B-Chat and Owen-72B-Chat have reached the leading level of open source models in five dimensions, such as openness and conscientiousness, and show the highest correlation in the personality prediction task of Chinese consulting dialogues [148]. Moreover, the TRAIT assessment tool study shows that Qwen 1.5-7B-Chat has good discriminant validity in the personality trait dimension [78]. Although research on pathological personality assessment, such as short dark three-dimensional traits, is still insufficient, existing data have confirmed that the Qwen series can effectively simulate healthy personality traits, which provides an essential foundation for building an anthropomorphic AI system.

Emotional Abilities. The development of Qwen's emotional abilities presents multi-dimensional imbalance characteristics. Under the EmoBench evaluation framework, the emotion recognition accuracy of Qwen 7B/14B is better than that of the LLaMA2 series models with the same parameters [116]. Especially in the Chinese context, Qwen-2.5 series models show the advantage of cross-cultural emotion understanding, and their multimodal emotion analysis capability has reached an advanced level in the industry. However, in terms of modeling complex social situations, the depth of emotion reasoning of the model still lags behind GPT-4, which highlights the need to improve the contextual emotion coherence.

5.5 Claude

This subsection examines the Claude series models, categorizing their evaluation into three distinct groups as previously mentioned.

ToM. The Claude series models establish new performance benchmarks in complex ToM tasks. In particular, Claude 3 Sonnet achieved 53% accuracy in perceptual reasoning tasks and a breakthrough performance of 60-80% on transparent ToM problems, outperforming most competitors [61]. However, Claude 3 Sonnet's performance is still lower than that of human beings, indicating that there is still room for improvement in its social common sense modeling.

Personality Traits. Existing studies show that the personality traits of the Claude series models have a significant dark tendency. The Machiavellianism dimension score is higher than the human norm in the short dark triad three-dimensional trait assessment. It is noted that the three-dimensional framework for LLMs' personality assessment provides a new perspective for analyzing Claude series models' personality characteristics [141]. Although the specific data has not yet been fully disclosed, engineering practice shows that the Claude series models perform well in the dimensions of extroversion and openness, which is highly consistent with the fluency characteristics of its dialogue system.

Emotional Abilities. Claude 3.7 Sonnet served as the judge model in the EQ-Bench 3 evaluation, which shows that its emotion understanding ability has been widely recognized [104]. In addition, Claude 3 Opus achieved an excellent score of 73.5 in the subtle emotion recognition task, especially in the emotional reasoning ability in socially awkward situations, which surpassed most LLMs.

5.6 Gemini

This subsection analyzes the Gemini series models, dividing them into three distinct groups as mentioned earlier.

ToM. The ToM capabilities of the Gemini model shows significant version differences. The accuracy of Gemini-1.0-Pro in perceptual reasoning tasks is only 34% [61], but Gemini-1.5 improves the second-order belief reasoning ability by 4% by improving the attention mechanism. On the transparent ToM problem, although Gemini-1.5 still fails to break the 50% accuracy threshold, its error pattern analysis shows that it has preliminary metacognitive capabilities. Compared with the Claude series, Gemini performs better in ToM tasks in culturally specific contexts, stemming from the breadth of its multilingual training data.

Personality Traits. The Gemini series models show unique performance in personality trait tests. Sorokovikova et al. [126] show that Gemini-1.5-Pro leads proprietary models in predicting the Big Five personality traits. Gemini-1.5-Pro achieves a personality trait recognition accuracy of 72.8% in the prediction task, maintaining its lead among proprietary models [148]. The model shows excellent cross-cultural adaptability in zero-shot personality classification tasks and was able to identify personality traits in non-Western contexts effectively.

Emotional Abilities. The Gemini series models also demonstrate good capabilities in emotional intelligence tests, especially the latest Gemini-2.5 series. In the Judgemark task, Gemini-1.5-Pro scores 66.58, which is lower than Claude-3-Opus and GPT-4, but still performs well [104]. In addition, in the EQ-Bench 3 evaluation, Gemini-2.0-Flash is used as a role model in a multi-round conflict mediation scenario to test the emotional mediation capabilities of other models. This shows that the Gemini model has a certain experience in simulating human emotional expression and social interaction.

6 Personality Simulation of LLMs

Some studies [106, 154] explore whether LLMs with humanlike capabilities possess personalities similar to humans. Unlike research focused on whether LLMs have inherent personalities, these studies concentrate on controlling the specific personalities that LLMs exhibit in their outputs. The solutions proposed previously can be mainly categorized into two types: editing and prompting [141].

Editing. Editing involves changing the model through fine-tuning or training on specific corpora. Jiang et al. [54], Pan and Zeng [106], Serapio-García et al. [120], Zhan et al. [149] examine the relationship between the personalities of LLMs and the training corpus, while Pan and Zeng [106] demonstrating that the type of training corpus can affect the MBTI types of LLMs. Serapio-García et al. [120] find that the content and diversity of the training dataset can also influence the LLMs' personality performance, and Jiang et al. [54] creating a dialogue dataset containing specific personality traits to show that training LLMs on specific datasets can exhibit certain personality characteristics. Cui et al. [28], Li et al. [82] adjust LLMs' personalities through SFT and DPO, testing based on MBTI and the Big Five personality theories.

Li et al. [84] fine-tune the LLMs solely through DPO to reduce the generation of harmful, aggressive, or inappropriate content, thereby enhancing the LLMs' psychological safety. Liu et al. [88] propose a DPG method based on hypernetworks. DPG falls under the category of personality editing techniques but still incorporates some prompting induction capabilities. Its main contribution is achieving target personality generation through internal dynamic adjustments rather than solely relying on input prompts.

Prompting. Prompting involves influencing the LLM's behavior and responses through input prompts, generally divided into explicit prompts, which directly use clear and specific descriptions or definitions to guide the LLMs to exhibit certain personality traits; and implicit prompts, which guide the LLMs to display specific personality traits through examples or context rather than directly providing explicit descriptions [95, 106].

Explicit Prompts: Current studies utilize similar explicit prompting methods. Karra et al. [65], La Cava and Tagarelli [75], Pan and Zeng [106], Sonlu et al. [125], Stöckli et al. [129], Tan et al. [132], Weng et al. [142] implemente prompt engineering by setting personalized prompts, such as "You are a friendly, extroverted person." Serapio-García et al.

[120] use personality markers (such as Goldberg's personality trait markers) and Likert-type language qualifiers to shape the personality performance of LLMs. Personality markers are vocabulary used to define individual traits, helping the model exhibit behaviors consistent with a certain Personality type when responding. Likert-type language qualifiers adjust the tone or expression of the model's responses to align more closely with predetermined emotional or cognitive dimensions. Huang et al. [52], Jiang et al. [55], Noever and Hyams [100], Noh and Chang [101] customize prompts for the model based on the Big Five personality model, imparting it with personality traits. Meanwhile, in Weng et al. [142], prompts are designed to include five demographic characteristics (age, gender, marital status, income, and number of children) and high or low descriptions of five dimensions of HEXACO personality for each character description. Similarly, Miotto et al. [94] design prompts based on HEXACO and HVS scales. Furthermore, Allbert et al. [4] summarize 179 different personality traits based on the HEX-ACO and the Five Factor Model to guide prompt design.

Implicit Prompts: Various types of implicit prompts are used in previous studies in attempts to simulate LLMs personalities. Caron and Srivastava [18], Sourati et al. [127] design prompts based on the context and task requirements; Pan and Zeng [106] offer a few example questions to implicitly express personality; Huang et al. [50] offer five factors—instruction templates, item wording, language, option labels, and option order—to complete prompt engineering; Petrov et al. [111] use two types of character descriptions, general or specific, in prompt design; Kovač et al. [74] also use cultural background information as part of the prompts to induce the model to display a specific cultural perspective. Furthermore, LLMs can generate expected content through psychological measurements sent by emotional computing systems [72]. He and Zhang [47] design the ASFPP framework to study the impact of factors such as social networks and subjective consciousness on agent personality formation.

Summary. These methods enable LLMs to simulate specific personalities to a certain extent, yet each still has its limitations and areas for improvement. In Mao et al. [91], various methods for editing model personalities are tested and evaluated, showing that fine-tuning performed best in modifying target personalities but had some interference with external themes. Prompt design is suitable for quickly achieving simple personality adjustments but has limited effectiveness for complex personality requirements. Current methods still have shortcomings in generating stable and diverse personality traits, especially in maintaining personality consistency in multi-turn dialogues.

7 LLMs as Human

In this section, we discuss whether LLMs can effectively simulate human roles in various scenarios, enabling researchers to explore human behavior through AI-driven simulations. As shown in Figure 7, we categorize such anthropomorphic behavior into three types: Social Experiment Simulations, Game Simulations, and Interactive Negotiation.

Social Experiment Simulations. Social experiment simulations refer to the use of LLMs as simulated human participants engaging in controlled social or psychological experiments. These simulations enable researchers to study human behaviors such as trust, cooperation, and fairness. Unlike traditional human-subject experiments, LLM-based social experiment simulations offer scalability and reproducibility while avoiding ethical constraints and logistical challenges associated with real-world experiments.

To simulate and investigate how the preferences and personalities of LLM-driven agents are formed and developed, He and Zhang [48] propose a framework called AFSPP (Agent Framework for Shaping Preference and Personality). This framework allows agents to learn, make decisions, and interact socially within a simulated environment, enabling researchers to observe whether and how their personalities and preferences evolve. Similarly, Gandhi et al. [39] explore LLM-based agent social simulations, investigating whether these agents adhere to Hobbes' Social Contract Theory by evolving from a state of nature into an organized commonwealth. Specifically, agents can choose to farm, trade, rob, or concede each day, adjusting their behaviors based on personality parameters (aggressiveness, greed, strength) and a memory system. Experimental results show that the agent society initially exists in a chaotic "state of nature", characterized by predominant looting and a lack of trust. Over time, the society gradually evolves towards a social contract, where some individuals opt to concede in exchange for protection, trade increases, violence declines, and a stable community emerges under a single sovereign authority. This evolutionary trajectory aligns with Hobbes' Social contract Theory. Besides, Chuang et al. [25] point out that most studies rely on demographic information for role-playing, which does not accurately reflect real human beliefs. To address this limitation, they propose using human belief networks to enhance LLMs' ability to simulate human beliefs. Specifically, the researchers utilized survey data covering 64 controversial topics across domains such as politics, science, religion, health, and economics. They then applied factor analysis to construct a belief network comprising nine independent belief factors, including supernatural beliefs, political party affiliations, and economic beliefs. Subsequently, they tested LLMs' ability to simulate human beliefs under different role-playing conditions and assessed alignment using mean absolute error. The results indicate that this approach significantly improves LLMs' capability to simulate human beliefs. Moreover, Leng [80] explore whether LLMs exhibit human-like psychological accounting effects and behavioral biases in economic decision-making. Their findings suggest that while LLMs demonstrate similarities to humans in certain psychological accounting mechanisms, they show significant differences in aspects such as loss aversion, transaction utility, and dynamic mental accounting. Leng and Yuan [81] propose a novel probabilistic framework called "State-Understanding-Value-Action" (SUVA) to systematically analyze LLMs' text-based responses in social environments. Experimental results indicate that LLMs' decision-making is influenced not only by training data and the alignment pro-

Three Types of Human Role Simulation by LLMs

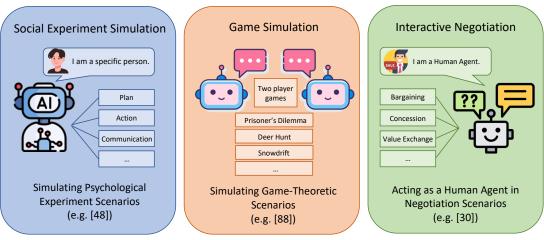


Figure 7: Three Types of Human Role Simulation by LLMs.

cess but also by reasoning patterns such as CoT reasoning. The SUVA framework provides a tool for evaluating and improving LLMs' applications in social interactions, making them more aligned with human social norms. Tjuatja et al. [133] investigate the use of LLMs in survey-based experimental environments to simulate human behavioral patterns and assess whether they exhibit psychological biases similar to those of humans.

Additionally, several other studies explore Social Experiment Simulations [35, 86], offering novel insights into the capabilities and limitations of LLMs in mimicking human social behaviors.

Game Simulations. Game Simulations refers to the use of LLM-driven agents to simulate human performance in multiple games. LLM-driven agents are used to simulate one or more people to complete real-world games, and various abilities of LLMs are evaluated under the defined game environment and rules. GPT-4 agents can show a high degree of consistency with humans in the framework of trust games [146]. Paglieri et al. [105] propose a benchmark to evaluate the agentic capabilities of LLMs through a series of challenging games, combining multiple reinforcement learning environments with different levels. Many ability benchmarks are built through grid-based games such as Tic-Tac-Toe, Connect Four, Gomoku [134], Rock-Paper-Scissors, Tower of Hanoi, Minecraft [145] and even PokéChamp [66].

A series of studies investigate game simulations within the theoretical framework of game theory. Game theory is often used to analyze human behaviors, with LLMs substituting humans in game experiments, thereby facilitating social science research. As noted by Fan et al. [34], rationality, a foundational principle of game theory, serves as a criterion for evaluating player behavior—establishing clear preferences, refining beliefs about uncertainty, and taking optimal actions. Lorè and Heydari [89] examine four typical two-player games -Prisoner's Dilemma, Deer Hunt, Snowdrift, and Prisoner's Delight -to explore how LLMs respond to social dilemmas, situations where humans can cooperate for the collective good or defect for the individual good.

These studies reveal limited capabilities in abstract strategic reasoning and a more nuanced understanding of the underlying mechanics of the games.

Through exploration of game theory games, the strategic reasoning ability of LLMs in games is an important aspect of evaluation. Several benchmarks is developed to assess the strategic reasoning capabilities of LLM-driven agents [27, 31]. Strategic reasoning capabilities require LLM agents to dynamically adapt their policies in a multiagent environment while constantly adapting their policies to achieve individual goals. Inspired by Level-K framework of behavioral economics, we extend reasoning from simple reactions to structured strategic depth, achieving a recursive implementation of strategic depth [152]. Gandhi et al. [40] propose that adding a few-shot chain-of-thought examples to the pre-trained LLMs can increase the ability to cope with various strategic scenarios and solve strategic games.

Overall, in games utilizing LLMs to simulate human behavior, the primary focus is on employing multi-agent systems for iterative interactions. These interactions are grounded in game theory within the environment and incorporate social psychological principles comprehensively.

Interactive Negotiation. Interactive Negotiation refers to using LLMs to simulate human-like negotiation processes, where discussions, bargaining, and decision-making take place in conversations to reach mutually beneficial agreements. This approach explores the decision-making capabilities of LLMs.

Davidson et al. [30] employ negotiation games as a dynamic and co-evolving benchmark to evaluate the agency, performance, and alignment of LMs. Compared to traditional static benchmarks, this methodology captures the intricacies of multi-turn interactions and cross-model dynamics inherent in real-world contexts.

Currently, there are relatively few studies in this area. However, since interactive negotiation can more realistically simulate LLMs' decision-making abilities, negotiation strategies, and multi-turn interactions in real-world scenarios, while effectively evaluating their agency and alignment,

this field is expected to become a research hotspot in the fu-

8 Comparison with Related Reviews

This review systematically examines the application of psychological assessment tools to evaluate the psychological characteristics of LLMs, while recent surveys explore different yet complementary dimensions of LLMs research.

Wen et al. [141] propose a taxonomy of personality-related research in LLMs, categorizing studies into self-assessment, personality exhibition, and personality recognition. Their work emphasizes methodological comparisons and highlights dynamic personality adaptation. In contrast, our review offers a more detailed evaluation of the suitability and limitations of traditional psychological tools for assessing LLMs. Additionally, Ke et al. [70] broaden the scope to encompass LLMs applications across cognitive, social, and cultural psychology, demonstrating their potential in experimental methodologies and emergent cognitive capabilities.

By comparison, our review narrows its focus to the nuanced interplay between established psychological measures and the emergent behaviors of LLMs, providing insights into their reliability, validity, and interpretive challenges.

This review complements existing works by emphasizing the alignment between classical psychological tools and the evaluation of LLMs. Our analysis underscores the methodological rigor required to adapt human-oriented tools for assessing machine behaviors while identifying gaps in higher-order psychological reasoning assessments that could inspire future interdisciplinary research.

9 Future Work

In future work, we recognize several valuable directions for the application of LLMs in psychology and sociology. The integration of explicit and implicit prompts is one promising approach to creating psychological scales better suited to LLMs. With the widespread adoption of these models in the field of psychology, combining explicit instructions (explicit prompts) with scenario simulations (implicit prompts) can provide a better framework for testing and understanding the psychological properties of LLMs. Such a combination allows for a more accurate and comprehensive evaluation of the model's personality while maintaining behavioral consistency, thereby laying the foundation for more reliable psychological assessment tools.

Another important direction is the enhancement of model safety through the optimization of model personality traits. Existing studies suggest a correlation between the personality traits of language models and their safety capabilities. Li et al. [84] find that LLMs with dark personality traits exhibit greater psychological toxicity, which refers to the model's capacity to exhibit or encourage harmful psychological behaviors during interactions. Zhang et al. [150] also explore this relationship, finding that safety alignment tends to enhance traits such as extraversion, sensing, and judging. However, these findings are based on models with 7B parameters, and their applicability to larger-scale models remains

to be validated. Moreover, the complexity of the relationship between personality and safety indicates that further research is needed to uncover the underlying mechanisms. By optimizing these personality traits, it is possible to increase the model's ability to avoid harmful outputs, thus improving overall safety.

Current research suggests that large language models have the potential to simulate human-like behaviors, offering an opportunity to use them as substitutes for human participants in psychological research. Some researchers argue that these models could serve as "human sample agents" in experiments where human involvement is impractical or too costly. However, there are significant limitations, such as reduced diversity of thought and a tendency towards uniform "correct answers" [107], as well as challenges with simplified decision processes, real-data dependence, and difficulties in simulating behaviors across multiple environments [137]. These issues raise doubts about the validity of using LLMs as a complete replacement for human participants due to their lack of variability and complexity inherent in human responses. Despite these concerns, some studies show that LLMs can perform similarly to human participants in specific tasks, indicating their potential value in particular contexts [47, 52, 70].

These research directions hold great promise for advancing the understanding of the psychology of large language models, thereby increasing their practical value and enhancing their reliability in various applications, while fostering interdisciplinary collaboration across different fields.

10 Conclusion

This study systematically evaluates the application of psychological theory to large language models (LLMs), revealing both the potential and limitations of current methods. Different models exhibit varying psychological characteristics, with GPT-4 demonstrating the best performance across all dimensions. Although some LLMs exhibit reproducible personality patterns under specific prompting schemes, significant variability remains between tasks and settings. Therefore, constructing psychological assessment frameworks suitable for diverse application scenarios remains a critical challenge. Our analysis further highlights the methodological limitations in tool mismatches and evaluation inconsistencies, and we suggest that future research should focus on developing more interpretable and robust assessment tools, particularly to address complex social reasoning and emotional intelligence evaluation needs. Additionally, the potential of LLMs to simulate human personality traits and behaviors remains promising, but further exploration is required to ensure stable and reliable personality representation across a wide range of contexts. By refining these methods, future advancements can better support using LLMs in psychological assessments, especially in socially sensitive tasks and complex scenarios.

References

- [1] Akinsulore A., Fatoye O., Awaa O., Aloba Olutayo, Mapayi B., and Ibigbami O. 2012. Reliability and Concurrent Validity of the International Personality item Pool (IPIP) Big-five Factor Markers in Nigeria. (2012). doi:10.4314/NJPSYC.V1012 3
- [2] Ibrahim Abd Ghani and Norsayyadatina Che Rozubi. 2020. CONTENT VALIDITY AND RELIABILITY OF BUSS AND PERRY AGRESSIVE QUESTION-NAIRE (BPAQ) INVENTORY. *International journal* of Education, Psychology and Counseling 5, 37 (dec 8 2020), 297–303. doi:10.35631/ijepc.5370024 4
- [3] Yiming Ai, Zhiwei He, Ziyin Zhang, Wenhong Zhu, Hongkun Hao, Kai Yu, Lingjun Chen, and Rui Wang. 2024. Is Cognition and Action Consistent or Not: Investigating Large Language Model's Personality. *arXiv preprint arXiv:2402.14679* (2024). 11, 12
- [4] Rumi A Allbert, James K Wiles, and Vlad Grankovsky. 2024. Identifying and Manipulating Personality Traits in LLMs Through Activation Engineering. *arXiv preprint arXiv:2412.10427* (2024). 16
- [5] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Do LLMs Exhibit Human-Like Reasoning? Evaluating Theory of Mind in LLMs for Open-Ended Responses. *CoRR* abs/2406.05659 (2024). doi:10.48550/ARXIV.2406.05659 arXiv:2406.05659 14
- [6] Anthropic. 2025. Claude AI (Version 3.7 Sonnet). https://claude.ai/. Large language model response generated on April 26, 2025. 2
- [7] Michael C Ashton and Kibeom Lee. 2009. The HEXACO-60: A short measure of the major dimensions of personality. *Journal of personality assessment* 91, 4 (2009), 340–345. 3, 4
- [8] Nikolaos P Bakas, Maria Papadaki, Evgenia Vagianou, Ioannis Christou, and Savvas A Chatzichristofis. 2023. Integrating LLMs in Higher Education, Through Interactive Problem Solving and Tutoring: Algorithmic Approach and Use Cases. In European, Mediterranean, and Middle Eastern Conference on Information Systems. Springer, 291–307.
- [9] Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21, 1 (1985), 37–46.
- [10] Simon Baron-Cohen, Michelle O'riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. 1999. Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *journal of autism and developmental disor*ders 29 (1999), 407–418. 5

- [11] Lim Beng-Chong and E. Ployhart Robert. 2006. Assessing the Convergent and Discriminant Validity of Goldberg's International Personality Item Pool. (2006). doi:10.1177/1094428105283193 3
- [12] Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating Personality Traits in Large Language Models: Insights from Psychological Questionnaires. *CoRR* abs/2502.05248 (2025). doi:10.48550/ARXIV.2502. 05248 arXiv:2502.05248 13, 14
- [13] Bojana Bodroža, Bojana M Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science* 11, 10 (2024), 240180. 2, 10, 12, 14
- [14] Gregory J. Boyle. 1995. Myers-Briggs Type Indicator (MBTI): Some psychometric limitations. *Humanities & Social Sciences papers* 30 (03 1995). doi:10.1111/j.1742-9544.1995.tb01750.x 3
- [15] Katharine C Briggs. 1976. *Myers-Briggs type indicator*. Consulting Psychologists Press Palo Alto, CA. 2, 3
- [16] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020). 2
- [17] Arnold H Buss and Mark Perry. 1992. The aggression questionnaire. *Journal of personality and social psychology* 63, 3 (1992), 452. 4
- [18] Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276* (2022). 16
- [19] Hudson W. de Carvalho, Sérgio B. Andreoli, Diogo R. Lara, Christopher J. Patrick, Maria Inês Quintana, Rodrigo A. Bressan, Marcelo F. de Melo, Jair de J. Mari, and Miguel R. Jorge. 2013. Structural validity and reliability of the Positive and Negative Affect Schedule (PANAS): Evidence from a large Brazilian community sample. *Revista Brasileira de Psiquiatria* 35, 2 (6 2013), 169–172. doi:10.1590/1516-4446-2012-0957 4
- [20] Diego Castrillón M, Paola Ortiz T, and Fernando Vieco G. 2009. Cualidades paramétricas del cuestionario de agresión (AQ) de Buss y Perry en estudiantes universitarios de la ciudad de Medellín (Colombia). Revista Facultad Nacional de Salud Pública 22, 2 (feb 4 2009). doi:10.17533/udea.rfnsp.561 4
- [21] Yuyan Chen and Yanghua Xiao. 2024. Recent Advancement of Emotion Cognition in Large Language Models. *CoRR* abs/2409.13354 (2024). doi:10. 48550/ARXIV.2409.13354 arXiv:2409.13354 13

- [22] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 15959–15983. doi:10.18653/V1/2024.ACL-LONG.847 14
- [23] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv preprint arXiv:2204.02311 (2022). 2
- [24] Michael Christopher, Marissa Ferry, Akeesha Simmons, Alicia Vasquez, Brooke Reynolds, and Daniel Grupe. 2024. Psychometric properties of the Buss–Perry Aggression Questionnaireshort form among law enforcement officers. Aggressive Behavior 50, 2 (3 2024). doi:10.1002/ab.22145 4
- [25] Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan V. Shah, Junjie Hu, and Timothy T. Rogers. 2024. Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks. In Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, 14010–14026. doi:10.18653/v1/2024.findings-emnlp.819
- [26] Eric M. Cooke, Noémi K. Schuurman, and Yao Zheng. 2022. Examining the within- and between-person structure of a short form of the positive and negative affect schedule: A multilevel and dynamic approach. *Psychological Assessment* 34, 12 (12 2022), 1126– 1137. doi:10.1037/pas0001167 4
- [27] Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. 2024. Gamebench: Evaluating strategic reasoning abilities of llm agents. arXiv preprint arXiv:2406.06613 (2024). 17
- [28] Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine Mindset: An MBTI Exploration of Large Language Models. arXiv:2312.12999 [cs.CL] 12, 15
- [29] Olga Cunha, Manuela Peixoto, Ana Rita Cruz, and Rui Abrunhosa Gonçalves. 2021. Buss-Perry Aggression Questionnaire: Factor Structure and Measurement Invariance Among Portuguese Male Perpetrators of Intimate Partner Violence. *Criminal Justice and Behavior* 49, 3 (oct 7 2021), 451–467. doi:10.1177/ 00938548211050113 4

- [30] Tim R. Davidson, Veniamin Veselovsky, Michal Kosinski, and Robert West. 2024. Evaluating Language Model Agency Through Negotiations. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=3ZqKxMHcAg 17
- [31] Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. arXiv preprint arXiv:2402.12348 (2024). 17
- [32] Max J van Duijn. 2016. The lazy mindreader. A humanities perspective on mindreading and multiple-order intentionality. *Netherlands: Koninklijke Wöhrman* (2016). 5
- [33] Hana-May Eadeh, Rosanna Breaux, Joshua M. Langberg, Molly A. Nikolas, and Stephen P. Becker. 2020. Multigroup multilevel structure of the child and parent versions of the Positive and Negative Affect Schedule (PANAS) in adolescents with and without ADHD. *Psychological Assessment* 32, 4 (4 2020), 374–382. doi:10.1037/pas0000796 4
- [34] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17960–17967. 17
- [35] Apostolos Filippas, John J. Horton, and Benjamin S. Manning. 2024. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? In Proceedings of the 25th ACM Conference on Economics and Computation, EC 2024, New Haven, CT, USA, July 8-11, 2024, Dirk Bergemann, Robert Kleinberg, and Daniela Sabán (Eds.). ACM, 614–615. doi:10.1145/3670865.3673513 17
- [36] Ivar Frisch and Mario Giulianelli. 2024. LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models. *arXiv preprint* arXiv:2402.02896 (2024). 11
- [37] Adrian Furnham, Steven C Richards, and Delroy L Paulhus. 2013. The Dark Triad of personality: A 10 year review. *Social and personality psychology compass* 7, 3 (2013), 199–216. 2, 3
- [38] Adrian Furnham, Jane Zhang, and Tomas Chamorro-Premuzic. 2005. The relationship between psychometric and self-estimated intelligence, creativity, personality and academic achievement. *Imagination, cognition and personality* 25, 2 (2005), 119–145. 4
- [39] Kanishk Gandhi, Zoe Lynch, Jan-Philipp Fränken, Kayla Patterson, Sharon Wambu, Tobias Gerstenberg,

- Desmond C. Ong, and Noah D. Goodman. 2024. Human-like Affective Cognition in Foundation Models. *CoRR* abs/2409.11733 (2024). doi:10.48550/ARXIV.2409.11733 arXiv:2409.11733 13, 16
- [40] Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. 2023. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165* (2023). 17
- [41] József Gerevich, Erika Bácskai, and PÁL Czobor. 2007. The generalizability of the buss–perry aggression questionnaire. *International Journal of Methods in Psychiatric Research* 16, 3 (2007), 124–136. 4
- [42] Lewis R. Goldberg. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe* 7, 1 (1999), 7–28. 3, 6
- [43] Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality* 40, 1 (2006), 84–96. 6
- [44] Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2023. Self-assessment tests are unreliable measures of llm personality. *arXiv preprint arXiv:2309.08163* (2023). 9
- [45] Francesca GE Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *journal of* autism and Developmental disorders 24, 2 (1994), 129–154. 5
- [46] Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755* (2023). 2
- [47] Zihong He and Changwang Zhang. 2024. AFSPP: Agent Framework for Shaping Preference and Personality with Large Language Models. *arXiv preprint arXiv:2401.02870* (2024). 16, 18
- [48] Zihong He and Changwang Zhang. 2024. AF-SPP: Agent Framework for Shaping Preference and Personality with Large Language Models. *CoRR* abs/2401.02870 (2024). doi:10.48550/ARXIV.2401.02870 arXiv:2401.02870 16
- [49] Yu Hou, Hal Daumé III, and Rachel Rudinger. 2025. Language Models Predict Empathy Gaps Between Social In-groups and Out-groups. *arXiv preprint arXiv*:2503.01030 (2025). 13, 14
- [50] Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu.

- 2023. Revisiting the Reliability of Psychological Scales on Large Language Models. *arXiv e-prints* (2023), arXiv–2305. 3, 9, 10, 12, 16
- [51] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Who is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench. arXiv preprint arXiv:2310.01386 (2023). 2, 13, 14
- [52] Muhua Huang, Xijuan Zhang, Christopher Soto, and James Evans. 2024. Designing LLM-Agents with Personalities: A Psychometric Approach. *arXiv preprint arXiv:2410.19238* (2024). 12, 16, 18
- [53] Anitha Jeyagurunathan, Jue Hua Lau, Edimansyah Abdin, Saleha Shafie, Sherilyn Chang, Ellaisha Samari, Laxman Cetty, Ker-Chiah Wei, Yee Ming Mok, Charmaine Tang, Swapna Verma, Siow Ann Chong, and Mythily Subramaniam. 2022. Aggression Amongst Outpatients With Schizophrenia and Related Psychoses in a Tertiary Mental Health Institution. *Frontiers in Psychiatry* 12 (jan 3 2022). doi:10.3389/fpsyt.2021.777388 4
- [54] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and Inducing Personality in Pre-trained Language Models. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 16, 2023, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/21f7b745f73ce0d1f9bcea7f40b1388e-Abstract-Conference.html 2, 3, 6, 9, 10, 12, 15
- [55] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547* (2023). 3, 16
- [56] Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology* (1991). 2
- [57] Oliver P John, Sanjay Srivastava, et al. 1999. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. (1999). 2, 3
- [58] John A Johnson. 2005. Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality* 39, 1 (2005), 103–129. 6
- [59] John A Johnson. 2014. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of research in personality* 51 (2014), 78–89. 3, 6

- [60] Daniel N Jones and Delroy L Paulhus. 2014. Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment* 21, 1 (2014), 28–41. 3
- [61] Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 19794–19809. https://aclanthology.org/2024.emnlp-main.1105 15
- [62] Petri J. Kajonius and John A. Johnson. 2019. Assessing the Structure of the Five Factor Model of Personality (IPIP-NEO-120) in the Public Domain. *Europe's Journal of Psychology* 15 (2019), 260 275. 3
- [63] Nils Kaland, Annette Møller-Nielsen, Lars Smith, Erik Lykke Mortensen, Kirsten Callesen, and Dorte Gottlieb. 2005. The strange stories test: A replication study of children and adolescents with Asperger syndrome. European child & adolescent psychiatry 14 (2005), 73–82. 5
- [64] Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. LLM-GLOBE: A Benchmark Evaluating the Cultural Values Embedded in LLM Output. CoRR abs/2411.06032 (2024). doi:10.48550/ARXIV.2411.06032 arXiv:2411.06032 14
- [65] Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the Personality of White-Box Language Models. *arXiv preprint arXiv:2204.12000* (2022). 2, 15
- [66] Seth Karten, Andy Luu Nguyen, and Chi Jin. 2025. PokéChamp: an Expert-level Minimax Language Agent. arXiv preprint arXiv:2503.04094 (2025). 17
- [67] Maciej Karwowski. 2011. It doesn't hurt to ask... But sometimes it hurts to believe: Polish students' creative self-efficacy and its predictors. *Psychology of Aesthetics, Creativity, and the Arts* 5, 2 (2011), 154. 4
- [68] Maciej Karwowski, Izabella Lebuda, and Ewa Wiśniewska. 2018. Measuring creative self-efficacy and creative personal identity. The International journal of Creativity & Problem Solving 28, 1 (2018), 45– 57. 4
- [69] James C Kaufman, Michelle L Evans, and John Baer. 2010. The American Idol Effect: Are students good judges of their creativity across domains? *Empirical* studies of the arts 28, 1 (2010), 3–17. 4

- [70] Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv* preprint arXiv:2401.01519 (2024). 13, 18
- [71] Peter Kinderman, Robin Dunbar, and Richard P Bentall. 1998. Theory-of-mind deficits and causal attributions. *British journal of Psychology* 89, 2 (1998), 191–204. 5
- [72] Lawrence J Klinkert, Stephanie Buongiorno, and Corey Clark. 2024. Driving generative agents with their personality. *arXiv preprint arXiv:2402.14879* (2024). 10, 16
- [73] Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083* (2023). 2, 13
- [74] Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. arXiv preprint arXiv:2307.07870 (2023). 16
- [75] Lucio La Cava and Andrea Tagarelli. 2024. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115* (2024). 2, 15
- [76] Frieder R Lang, Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G Wagner. 2011. Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behavior research methods* 43 (2011), 548–567. 6
- [77] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5872–5877. 8
- [78] Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2024. Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics. *CoRR* abs/2406.14703 (2024). doi:10.48550/ARXIV.2406.14703 arXiv:2406.14703 2, 7, 9, 10, 12, 14
- [79] Sean T. H. Lee, Andree Hartanto, Jose C. Yong, Brandon Koh, and Angela K.y. Leung. 2019. Examining the crosscultural validity of the positive affect and negative affect schedule between an Asian (Singaporean) sample and a Western (American) sample. *Asian Journal of Social Psychology* 23, 1 (nov 6 2019), 109–116. doi:10.1111/ajsp.12390 4

- [80] Yan Leng. 2024. Can LLMs Mimic Human-Like Mental Accounting and Behavioral Biases?. In Proceedings of the 25th ACM Conference on Economics and Computation, EC 2024, New Haven, CT, USA, July 8-11, 2024. ACM, 581. 16
- [81] Yan Leng and Yuan Yuan. 2023. Do LLM Agents Exhibit Social Behavior? *arXiv preprint* arXiv:2312.15198 (2023). 16
- [82] Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2024. BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data. arXiv preprint arXiv:2410.16491 (2024). 15
- [83] Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2022. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. *arXiv* preprint arXiv:2212.10529 (2022). 3, 13, 14
- [84] Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. Evaluating psychological safety of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 1826–1843. 10, 12, 15, 18
- [85] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024. Quantifying AI Psychology: A Psychometrics Benchmark for Large Language Models. *arXiv preprint* arXiv:2406.17675 (2024). 2, 8, 10, 11, 12, 13, 14
- [86] Yuan Li, Yixuan Zhang, and Lichao Sun. 2023. MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents. *CoRR* abs/2310.06500 (2023). doi:10.48550/ARXIV.2310.06500 arXiv:2310.06500 17
- [87] Young-Jin Lim, Bum-Hee Yu, Doh-Kwan Kim, and Ji-Hae Kim. 2010. The Positive and Negative Affect Schedule: Psychometric Properties of the Korean Version. *Psychiatry Investigation* 7, 3 (2010), 163. doi:10.4306/pi.2010.7.3.163 4
- [88] Jianzhi Liu, Hexiang Gu, Tianyu Zheng, Liuyu Xiang, Huijia Wu, Jie Fu, and Zhaofeng He. 2024. Dynamic Generation of Personalities with Large Language Models. *arXiv preprint arXiv:2404.07084* (2024). 11, 12, 15
- [89] Nunzio Lorè and Babak Heydari. 2023. Strategic behavior of large language models: Game structure vs. contextual framing. *arXiv preprint arXiv:2309.05898* (2023). 17
- [90] Ypofanti M., Zisi V., Zourbanos N., Mouchtouri B., Tzanne P., Theodorakis Y., and Lyrakos G. 2015. Psychometric Properties of the International Personality Item Pool Big-Five Personality Questionnaire for

- the Greek population. *Health Psychology Research* (2015). doi:10.4081/hpr.2015.2206 3
- [91] Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing Personality For Large Language Models. In CCF International Conference on Natural Language Processing and Chinese Computing. Springer, 241–254. 12, 16
- [92] Jessica L Maples-Keller, Rachel L Williamson, Chelsea E Sleep, Nathan T Carter, W Keith Campbell, and Joshua D Miller. 2019. Using Item Response Theory to Develop a 60-Item Representation of the NEO PI–R Using the International Personality Item Pool: Development of the IPIP–NEO–60. journal of Personality Assessment (2019). doi:10.1080/00223891. 2017.1381968 3
- [93] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. 2023. A Turing Test: Are AI Chatbots Behaviorally Similar to Humans? *Available* at SSRN (2023). 13
- [94] Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? An exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338* (2022). 2, 9, 10, 11, 16
- [95] Maria Molchanova, Anna Mikhailova, Anna Korzanova, Lidiia Ostyakova, and Alexandra Dolidze. 2025. Exploring the Potential of Large Language Models to Simulate Personality. arXiv preprint arXiv:2502.08265 (2025). 15
- [96] Morten Moshagen, Isabel Thielmann, Benjamin E Hilbig, and Ingo Zettler. 2019. Meta-analytic investigations of the HEXACO Personality Inventory (-Revised). Zeitschrift für Psychologie 227, 3 (2019), 186–194. 4
- [97] IB Myers. 1962. The Myers-Briggs Type Indicator. *Educational Testing Service/Princeton* (1962). 3
- [98] Isabel Briggs Myers. 1985. A guide to the development and use of the Myers-Briggs type indicator: Manual. Consulting Psychologists Press. 3
- [99] Mohamed N., Sulaiman W., Halim F., and Saidfudin Masodi Mohd. 2021. An Initial Analysis of Reliability and Validity of a Personality Instrument Using the Rasch Measurement Model. https://hrmars.com/papers_submitted/11251/aninitial-analysis-of-reliability-and-validity-of-a-personality-instrument-using-the-raschmeasurement-model.pdf. International journal of Academic Research in Business and Social Sciences (2021). doi:10.6007/ijarbss/v11-i9/11251
- [100] David Noever and Sam Hyams. 2023. AI Text-to-Behavior: A Study In Steerability. *arXiv preprint arXiv:2308.07326* (2023). 16

- [101] Sean Noh and Ho-Chun Herbert Chang. 2024. LLMs with Personalities in Multi-issue Negotiation Games. *arXiv preprint arXiv:2405.05248* (2024). 16
- [102] Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2023. Dialogbench: Evaluating Ilms as human-like dialogue systems. *arXiv preprint arXiv:2311.01677* (2023). 2
- [103] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems 35 (2022), 27730–27744.
- [104] Samuel J. Paech. 2023. EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models. *CoRR* abs/2312.06281 (2023). doi:10.48550/ARXIV. 2312.06281 arXiv:2312.06281 7, 13, 14, 15
- [105] Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, et al. 2024. Balrog: Benchmarking agentic llm and vlm reasoning on games. *arXiv preprint arXiv:2411.13543* (2024). 17
- [106] Keyu Pan and Yawen Zeng. 2023. Do Ilms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv* preprint *arXiv*:2307.16180 (2023). 13, 15, 16
- [107] Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods* (2024), 1–17. 18
- [108] Josef Perner, Susan R Leekam, and Heinz Wimmer. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology* 5, 2 (1987), 125–137. 5
- [109] Josef Perner and Heinz Wimmer. 1985. "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *journal of experimental child psychology* 39, 3 (1985), 437–471. 5
- [110] Karolina Petraškaitė and Neringa Grigutytė. 2022. Development and Validation Regarding the Lithuanian Version of the Positive and Negative Affect Schedule (PANAS-X). European journal of Mental Health 17, 3 (dec 13 2022), 52–64. doi:10.5708/ejmh.17.2022.3.4 4
- [111] Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited Ability of LLMs to Simulate Human Psychological Behaviours: a Psychometric Analysis. *arXiv preprint arXiv:2405.07248* (2024). 10, 13, 16

- [112] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526. 2
- [113] Cecilia Reyna, Anahi Sanchez, Maria Gabriela Lello Ivacevich, and Silvina Brussino. 2011. The Buss-Perry Aggression Questionnaire: construct validity and gender invarianceamong Argentinean adolescents. *International Journal of Psychological Research* 4, 2 (dec 30 2011), 30–37. doi:10.21500/20112084.775 4
- [114] Jonathan Rush and Scott M. Hofer. 2014. Differences in within- and between-person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel structural equation modeling. *Psychological Assessment* 26, 2 (6 2014), 462–473. doi:10.1037/a0035666 4
- [115] Rusu S., P. Maricuţoiu Laurenţiu, Macsinga Irina, Vîrgă D., and Sava F. 2012. Evaluarea personalităţii din perspectiva modelului Big Five. Date privind adaptarea chestionarului IPIP-50 pe un eșantion de studenţi români. (2012). doi:10.24837/PRU.2012. 1.134 3
- [116] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 5986–6004. doi:10.18653/V1/2024.ACL-LONG.326 7, 8, 13, 14
- [117] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv* preprint arXiv:1904.09728 (2019). 8
- [118] Henry D Schlinger. 2009. Theory of mind: An overview and behavioral perspective. *The Psychological Record* 59 (2009), 435–448. 5
- [119] Kelly Serafini, Bo Malin-Mayor, Charla Nich, Karen Hunkele, and Kathleen M. Carroll. 2016. Psychometric properties of the Positive and Negative Affect Schedule (PANAS) in a heterogeneous sample of substance users. *The American journal of Drug and Alcohol Abuse* 42, 2 (feb 23 2016), 203–212. doi:10.3109/00952990.2015.1133632 4
- [120] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. arXiv preprint arXiv:2307.00184 (2023). 3, 10, 15, 16

- [121] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763* (2023). 13
- [122] Umarpreet Singh and Parham Aarabhi. 2023. Can AI have a personality?. In 2023 IEEE Conference on Artificial Intelligence (CAI). IEEE, 205–206. 2, 3
- [123] Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. 2024. Identifying Multiple Personalities in Large Language Models with External Evaluation. arXiv preprint arXiv:2402.14805 (2024). 2
- [124] Xiaoyang Song, Akshat Gupta, Kiyan Mohebbizadeh, Shujie Hu, and Anant Singh. 2023. Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms. arXiv preprint arXiv:2305.14693 (2023). 2, 9
- [125] Sinan Sonlu, Bennie Bendiksen, Funda Durupinar, and Uğur Güdükbay. 2024. The Effects of Embodiment and Personality Expression on Learning in LLM-based Educational Agents. *arXiv preprint arXiv:2407.10993* (2024). 15
- [126] Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan P. Yamshchikov. 2024. LLMs Simulate Big5 Personality Traits: Further Evidence. In Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024). 83–87. 9, 12, 13, 14, 15
- [127] Zhivar Sourati, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Nuan Wen, Ala Tak, Fred Morstatter, and Morteza Dehghani. 2024. Secret Keepers: The Impact of LLMs on Linguistic Markers of Personal Traits. arXiv preprint arXiv:2404.00267 (2024). 16
- [128] Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral health-care: a proposal for responsible development and evaluation. NPJ Mental Health Research 3, 1 (2024), 12.
- [129] Leandro Stöckli, Luca Joho, Felix Lehner, and Thomas Hanne. 2024. The Personification of Chat-GPT (GPT-4)—Understanding Its Personality and Adaptability. *Information* 15, 6 (2024), 300. 15
- [130] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in

- large language models and humans. *Nature Human Behaviour* 8, 7 (2024), 1285–1295. 13, 14
- [131] Tomaz Paiva Tamyres, Eduardo Pimentel Carlos, de Sousa Bezerra de Menezes Thaís, Costa A.C.R., das Graças Carvalho Costa Dinara, and H. Vasconcelos M. 2020. Questionário de Agressão de Buss-Perry versão reduzida (QA-R): análises estruturais. *Psi*cología, Conocimiento y Sociedad 10, 3 (may 5 2020). doi:10.26864/pcs.v10.n3.7 4
- [132] Fiona Anting Tan, Gerard Christopher Yeo, Kokil Jaidka, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Yang Liu, and See-Kiong Ng. 2024. PHAn-ToM: Persona-based Prompting Has An Effect on Theory-of-Mind Reasoning in Large Language Models. arXiv:2403.02246 [cs.CL] https://arxiv.org/abs/2403.02246 15
- [133] Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Trans. Assoc. Comput. Linguistics* 12 (2024), 1011–1026. doi:10.1162/TACL_A_00685
- [134] Oguzhan Topsakal, Colby Jacob Edell, and Jackson Bailey Harper. 2024. Evaluating large language models with grid-based game competitions: an extensible LLM benchmark and leaderboard. *arXiv* preprint *arXiv*:2407.07796 (2024). 17
- [135] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint* arXiv:2302.13971 (2023). 2
- [136] Max J van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco R Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. *arXiv preprint arXiv:2310.20320* (2023). 2, 5, 13
- [137] Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, et al. 2023. User behavior simulation with large language model based agents. arXiv preprint arXiv:2306.02552 (2023). 18
- [138] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *journal of Pacific Rim Psychology* 17 (2023), 18344909231213958. 7, 13
- [139] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales.

- Journal of personality and social psychology 54, 6 (1988), 1063. 4
- [140] Gregory D. Webster, C. Nathan DeWall, Richard S. Pond, Jr., Timothy Deckman, Peter K. Jonason, Bonnie M. Le, Austin Lee Nichols, Tatiana Orozco Schember, Laura C. Crysel, Benjamin S. Crosier, C. Veronica Smith, E. Layne Paddock, John B. Nezlek, Lee A. Kirkpatrick, Angela D. Bryan, and Renée J. Bator. 2013. The brief aggression questionnaire: psychometric and behavioral evidence for an efficient measure of trait aggression. *Aggressive Behavior* 40, 2 (10 2013), 120–139. doi:10.1002/ab. 21507 4
- [141] Zhiyuan Wen, Yu Yang, Jiannong Cao, Haoming Sun, Ruosong Yang, and Shuaiqi Liu. 2024. Self-assessment, Exhibition, and Recognition: a Review of Personality in Large Language Models. *CoRR* abs/2406.17624 (2024). doi:10.48550/ARXIV.2406. 17624 arXiv:2406.17624 15, 18
- [142] Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. Controllm: Crafting diverse personalities for language models. *arXiv preprint arXiv:2402.10151* (2024). 15, 16
- [143] Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 4602–4625. doi:10.18653/V1/2022.NAACL-MAIN, 341 7
- [144] Heinz Wimmer and Josef Perner. 1983. beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 1 (1983), 103–128. 5
- [145] Yue Wu, Xuan Tang, Tom M Mitchell, and Yuanzhi Li. 2023. Smartplay: A benchmark for llms as intelligent agents. *arXiv preprint arXiv:2310.01557* (2023). 17
- [146] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. 2024. Can Large Language Model Agents Simulate Human Trust Behavior?. In The Thirty-eighth Annual Conference on Neural Information Processing Systems. 17
- [147] Hainiu Xu, Siya Qi, Jiazheng Li, Yuxiang Zhou, Jinhua Du, Caroline Catmur, and Yulan He. 2025. EnigmaToM: Improve LLMs' Theory-of-Mind Reasoning

- Capabilities with Neural Knowledge Base of Entity States. *arXiv preprint arXiv:2503.03340* (2025). 13
- [148] Yang Yan, Lizhi Ma, Anqi Li, Jingsong Ma, and Zhenzhong Lan. 2024. Predicting the Big Five Personality Traits in Chinese Counselling Dialogues Using Large Language Models. *CoRR* abs/2406.17287 (2024). doi:10.48550/ARXIV.2406. 17287 arXiv:2406.17287 14, 15
- [149] Baohua Zhan, Yongyi Huang, Wenyao Cui, Huaping Zhang, and Jianyun Shang. 2024. Humanity in AI: Detecting the Personality of Large Language Models. arXiv preprint arXiv:2410.08545 (2024). 15
- [150] Jie Zhang, Dongrui Liu, Chen Qian, Ziyue Gan, Yong Liu, Yu Qiao, and Jing Shao. 2024. The Better Angels of Machine Personality: How Personality Relates to LLM Safety. *arXiv preprint arXiv:2407.12344* (2024). 2, 3, 10, 12, 14, 18
- [151] Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. 2025. A survey of table reasoning with large language models. *Frontiers Comput. Sci.* 19, 9 (2025), 199348. doi:10.1007/S11704-024-40330-Z 2
- [152] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan, and Furu Wei. 2024. K-Level Reasoning: Establishing Higher Order Beliefs in Large Language Models for Strategic Reasoning. *arXiv* preprint arXiv:2402.01521 (2024). 17
- [153] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint* arXiv:2303.18223 1, 2 (2023). 2
- [154] Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025. LMLPA: Language Model Linguistic Personality Assessment. *Computational Linguistics* (2025), 1–41. 15