Large Language Model Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement

Haoran Ye[1], Jing Jin[1], Yuhang Xie[1], Xin Zhang[2,3], Guojie Song [1,4,✉]

[1]State Key Laboratory of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
[2]School of Psychological and Cognitive Sciences, Peking University
[3]Key Laboratory of Machine Perception (Ministry of Education), Peking University
[4]PKU-Wuhan Institute for Artificial Intelligence

hrye@stu.pku.edu.cn gjsong@pku.edu.cn

Project Website: https://llm-psychometrics.com

Abstract

The rapid advancement of large language models (LLMs) has outpaced traditional evaluation methodologies. It presents novel challenges, such as measuring human-like psychological constructs, navigating beyond static and task-specific benchmarks, and establishing human-centered evaluation. These challenges intersect with Psychometrics, the science of quantifying the intangible aspects of human psychology, such as personality, values, and intelligence. This survey introduces and synthesizes an emerging interdisciplinary field of LLM Psychometrics, which leverages psychometric instruments, theories, and principles to evaluate, understand, and enhance LLMs. We systematically explore the role of Psychometrics in shaping benchmarking principles, broadening evaluation scopes, refining methodologies, validating results, and advancing LLM capabilities. This paper integrates diverse perspectives to provide a structured framework for researchers across disciplines, enabling a more comprehensive understanding of this nascent field. Ultimately, we aim to provide actionable insights for developing future evaluation paradigms that align with human-level AI and promote the advancement of human-centered AI systems for societal benefit. A curated repository of LLM psychometric resources is available at https://github.com/valuebyte-ai/Awesome-LLM-Psychometrics.

Contents

1 Introduction

> "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality." [Thorndike, 1962]

The advent of large language models (LLMs) represents a transformative breakthrough in AI. These systems exhibit general-purpose capabilities spanning diverse domains [Bubeck et al., 2023], with particular proficiency in natural language understanding and generation [Demszky et al., 2023, Grossmann et al., 2023, Gu et al., 2024, Ziems et al., 2024]. They are rapidly integrated into critical societal infrastructure, ranging from consumer-facing applications like chatbots [OpenAI, 2025] and search engines [Wang et al., 2024c] to high-stakes domains such as healthcare [Singhal et al., 2023], education [Milano et al., 2023], and scientific discovery [Romera-Paredes et al., 2024]. Their increasing dominance has exposed a fundamental, pressing scientific challenge: how can we rigorously evaluate these AI systems that transcend traditional benchmarks of biological or algorithmic intelligence?

Traditional AI evaluation has relied on curating task-specific datasets, annotating ground-truth labels with human input, running models on these datasets, and assessing performance using predefined metrics. However, LLMs have triggered an evaluation crisis, as their versatile capabilities and human-like behaviors exceed what traditional benchmarks can measure. Novel challenges include evaluating psychological constructs like personality, values, and cognitive biases that conventional assessment cannot capture. Rapid LLM development and training data contamination have made static benchmarks obsolete. Additionally, the prompt- and context-sensitive nature of LLMs compromises the robustness and validity of existing evaluation frameworks. With increasing human-LLM interaction, human-centered evaluation approaches are becoming essential. Moreover, as LLMs integrate into agentic and multimodal systems, evaluation methodologies must expand in scope and complexity.

These challenges intersect with humanity's century-old quest to quantify the complex, intangible human psychology, including knowledge, skills, character, values [Pasquali, 2009]. Psychometrics emerged from this timeless pursuit as the scientific study of psychological measurement. It bridges the abstract and the empirical by transforming human traits into quantifiable data, enabling better understanding, prediction, and decision-making in education, business, healthcare, governance, and beyond [Rust and Golombok, 2014]. The convergence of LLMs and psychometrics forms a methodological crucible, forging novel paradigms to better decode and improve machine minds.

*We define **LLM Psychometrics** as the interdisciplinary field dedicated to evaluating, understanding, and enhancing LLMs through the application and integration of psychometric instruments, theories, and principles.* This field seeks to quantify, interpret, manipulate, and improve the complex, human-like attributes and behaviors exhibited by LLMs, encompassing both personality constructs (such as personality, values, morality, and attitudes) and cognitive constructs (including heuristics and biases, social interaction abilities, psycholinguistic abilities, and learning and cognitive capacities). Research in LLM psychometrics applies, extends, and innovates upon scientific methods of psychological measurement for LLMs. Based on the psychometric principles, related research systematically examines the measurement results to ensure scientific rigor. By measuring and elucidating psychological constructs in LLMs, LLM psychometrics further informs strategies for their targeted enhancement.

Recent research in LLM psychometrics pioneers in addressing the LLM evaluation crisis. Some studies introduce dynamic and construct-oriented evaluation frameworks that move beyond static, task-specific benchmarks [Hagendorff, 2023, Zhu et al., 2024a]. In parallel, novel methodologies are developed to measure non-cognitive and emergent constructs [Huang et al., 2023d, Pellert et al., 2024, Ren et al., 2024]. Self-adaptive evaluation techniques now allow for the extrapolation of item difficulty and tailoring assessments to model performance [Jiang et al., 2024a, Lalor et al., 2024, Polo et al., 2024]. Drawing from the methodological framework of psychometric validation, research improves the reliability and validity of evaluation protocols [Ye et al., 2025a]. Human-centered evaluation drives aligning model behavior with human values [Wang et al., 2024f, Yao et al., 2025a]. In addition, the scope of evaluation expands to agentic and multimodal systems, further broadening the methodological landscape [Huang et al., 2024c, Li et al., 2024b].

The field of LLM psychometrics has seen significant growth, evidenced by the proliferation of related research papers. These studies, however, address a variety of psychological constructs, employ diverse methodologies, and utilize distinct validation techniques. The interdisciplinary nature of this domain has attracted contributions from a broad spectrum of academic fields. Despite this diversity, there is a lack of cohesion among researchers from different communities, leading to a fragmentation of insights, particularly between studies focusing on disparate constructs. Consequently, there is a pressing need for a comprehensive survey to synthesize these efforts and facilitate a more integrated understanding of the field.

This paper aims to bridge the gap by providing the first systematic review of LLM psychometrics, encompassing evaluation, validation, and enhancement. Fig. 1 illustrates the structure of the review. § 2 provides an overview of the

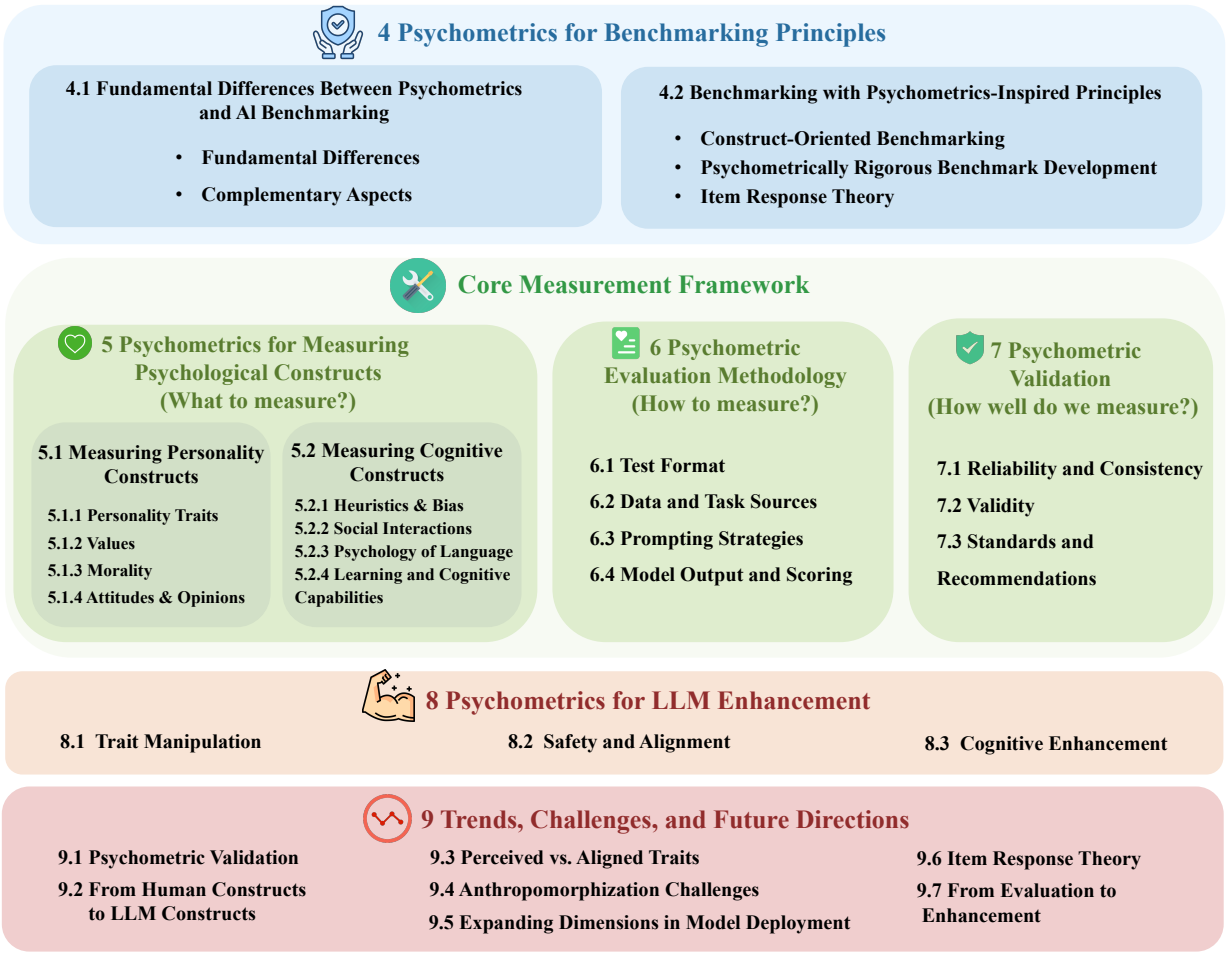

Figure 1: Overview of this review.

preliminaries and methodological foundations to facilitate subsequent discussions. In § 3, we delineate the definition, scope, and taxonomy of LLM psychometrics, which establishes the structure of the remainder of the review. The core measurement framework is detailed in § 5, § 6, and § 7. § 5 delves into the psychological constructs evaluated in LLMs, elucidating the theories employed and summarizing key evaluation findings. In § 6, we scrutinize the psychometric evaluation methodologies applied to LLMs, followed by § 7, which examines the psychometric validation of evaluation results. Beyond evaluation, § 8 introduces strategies for enhancing LLMs through psychometric insights. Finally, § 9 discusses emerging trends, challenges, and future directions in the psychometric evaluation of LLMs, while § 10 concludes the paper.

Related Surveys. We consider studies to be out of scope if they do not employ psychometric approaches, adhere to psychometric principles, or if they focus solely on scalar performance metrics rather than characterizing behaviors. Readers interested in conventional LLM benchmarking are encouraged to consult surveys on LLM evaluation [Chang et al., 2024, Guo et al., 2023b]. Several related surveys focus on the evaluation of specific constructs in LLMs, such as personality [Dong et al., 2025, Wen et al., 2024b], attitudes and values [Ma et al., 2024a], cultural awareness [Adilazuarda et al., 2024, Pawar et al., 2024], and theory of mind [Dong et al., 2025, Sarıtaş et al., 2025]. Hagendorff [2023], Hagendorff et al. [2024] introduce the notion of machine psychology and review emergent LLM abilities; however, they do not provide comprehensive coverage of the related research, nor do they elaborate on LLM personality constructs, psychometric validation, or enhancement. This paper provides the first systematic survey of LLM psychometrics.

2 Preliminary and Methodological Foundations

We aim for our survey to be self-contained and accessible to a broad, cross-disciplinary audience. To this end, this section presents the preliminary and methodological foundations that underpin the subsequent discussions.

2.1 Large Language Models

LLMs are large-scale *deep neural networks*—essentially complex systems of nonlinear regression equations. An LLM can generate text by predicting the next *token* (word or subword) in a sequential manner (autoregressive generation), given the preceding context. It does so by modelling a conditional probability distribution over the vocabulary; i.e., the likelihood of each token given the context:

$$P(x_t|x_{<t}) = f(x_{<t}; \theta), \tag{1}$$

where x_t is the token at time step t; $x_{<t}$ is the context preceding x_t, usually including both user prompts and previously generated tokens; f is the model's parameterized function; and θ represents the model parameters. Given f, the model generates text by either sampling from this distribution or directly selecting the token with the highest probability. In the former case, hyperparameters such as *temperature* can be adjusted to control the diversity of generated text. In the latter case, the model is said to use *greedy decoding*, and the generated text is deterministic. In evaluating LLMs, it is crucial to properly account for the stochasticity of the model.

These models are based primarily on the *transformer* architecture, a neural network design that employs self-attention mechanisms to capture contextual relationships between words, phrases, and broader linguistic patterns. Modern LLMs typically contain billions of parameters, enabling them to efficiently learn from vast amounts of textual data. During evaluation, if the model has already been exposed to the test items during training, this is referred to as *data contamination*. In such cases, the model is more likely to exhibit artificially inflated performance or simply reproduce memorized patterns, rather than revealing its true underlying capabilities or traits.

The training process of LLMs is typically divided into two phases: *pre-training* and *post-training*. *Pre-training* is the phase where LLMs learn to predict the next token, given its preceding context, on a large corpus of text data. This process is unsupervised, as the model does not require explicit labels or annotations to learn the underlying patterns in the data. The model processes Internet-scale text data from diverse sources like books, articles, and websites. By repeatedly predicting the next word in a sentence, the model learns the statistical properties of language and gains large-scale world knowledge. Models that have only undergone the pre-training phase are usually referred to as *base models*. *Post-training*, or *fine-tuning*, is the process of adapting the base models to better follow user instructions, align with human values, or specialize in particular tasks. This stage typically involves training the model on a smaller, human-annotated dataset or incorporating human feedback on the quality of the model outputs. Models that have undergone both phases are often referred to as *fine-tuned models*, *instruction-tuned models*, or *aligned models*.

We interact with LLMs using *prompts*, which are input instructions to the model. For psychometric evaluation, these prompts can naively be the reformatted versions of test items originally designed for humans, adapted for LLMs to answer. When designing prompts, one should consider differences between base and fine-tuned models. Most public-facing LLMs are fine-tuned, so evaluation research primarily focuses on these models for greater practical relevance.

A key emergent capability of LLMs is *in-context learning*, which allows models to adapt to new tasks or patterns by conditioning on examples or instructions provided within the input context $x_{<t}$, without modifying model parameters. This property can influence LLM performance in psychometric evaluation. For instance, prompting models to reason step-by-step (e.g., Chain-of-Thought prompting [Wei et al., 2022]) can enhance performance on reasoning tasks, while instructing them to role-play may modulate their exhibited personalities and values.

2.2 Psychometrics

Psychometrics, also known as psychological testing, involves the use of tests to measure, understand, or predict behavior by quantifying specific actions or characteristics. These tests rely on samples of behavior, meaning they are not perfect measures and often include errors inherent to sampling. Test *items* are specific stimuli designed to elicit observable reactions that can be scored or evaluated. Typically, tests are composed of multiple questions or problems as their items, producing explicit data subject to scientific analysis.

A *psychological test* is a set of items that are designed to measure characteristics of human beings that pertain to behavior [Kaplan and Saccuzzo, 2001]. Behavior measured by tests can be *overt* (observable actions) or *covert* (internal thoughts or feelings). Tests may assess past, current, or even predict future behavior. Interpretation of test scores depends on

their context within a distribution. *Scales* are used to relate raw scores to defined distributions, aiding interpretation. In addition, psychological tests can measure *traits*—enduring tendencies like shyness or determination—and *states*, which reflect temporary conditions of individuals.

Psychological testing measures *individual differences* in various *constructs*, which are abstract psychological attributes or dimensions that help explain and predict behavior. Two primary categories of such constructs are *personality constructs* and *cognitive constructs* [Kaplan and Saccuzzo, 2001]. *Personality tests* focus on an individual's tendencies and dispositions. These tests measure typical behavior, such as preferences or tendencies to react in certain ways. *Cognitive tests* evaluate speed, accuracy, or both, with higher scores reflecting better performance.

Two fundamental principles underpin psychometrics: *reliability* and *validity* [Raykov and Marcoulides, 2011]. Reliability ensures accuracy, dependability, consistency, or repeatability of the test results. Reliable test results are stable across time, contexts, and raters. Validity confirms the meaningfulness and usefulness of the test results. A valid measure captures the intended construct. Validity is multifaceted; for example, *predictive validity* might correlate test scores with job performance, while *construct validity* ensures alignment with theoretical models, such as the Big Five personality traits [Goldberg, 2013].

Other principles include *standardization*, which provides context to raw scores by comparing individual results to a representative sample, or norm group. Additionally, *equivalence* and *fairness* are crucial principles that tests must adhere to. Test bias occurs when test items unintentionally advantage or disadvantage subgroups. Modern psychometrics employs advanced statistical models to identify and revise biased items, ensuring assessments measure the intended construct rather than extraneous factors [Rust and Golombok, 2014].

2.3 Psychometric Evaluation of AI Before the Era of LLMs

The idea of applying psychometrics to AI originated in the early decades of AI [Pellert et al., 2024]. Evans [1964] pioneered work in this area by creating a heuristic program that could solve parts of intelligence tests. Subsequent efforts similarly focused on designing AI systems for cognitive tests [Newell, 1973], with the goal of creating systems capable of handling human tasks. This was conceptually aligned with the development of static, task-centric benchmarks in modern AI research [Chen et al., 2021, Hendrycks et al., 2020, Lee et al., 2024e, Liang et al., 2022, Srivastava et al., 2022]. However, criticisms emerged regarding the absence of "hot cognition" in AI, prompting Simon [1963] to propose incorporating emotional aspects into models. By the early 2000s, the concept of "psychometric AI" was explicitly articulated as the pursuit of systems capable of excelling on all established, validated tests of intelligence and mental ability. These included not only conventional IQ tests but also assessments of artistic and literary creativity, mechanical ability, and beyond [Bringsjord and Schimanski, 2003, Pellert et al., 2024]. It was not until the advent of LLMs that the versatility envisioned for "psychometric AI" began to materialize.

3 LLM Psychometrics: Definition, Scope, and Taxonomy

We define **LLM Psychometrics** as the interdisciplinary field dedicated to evaluating, understanding, and enhancing LLMs through the application and integration of psychometric instruments, theories, and principles.

LLM psychometrics research adapts, advances, and innovates upon psychometric methodologies to rigorously evaluate LLMs. This evaluation framework encompasses three core dimensions: the target construct (*what to measure*), the measurement method (*how to measure*), and the validation of results (*how well do we measure*). In many cases, these psychometric insights not only inform evaluation but also guide the development and refinement of LLMs (*how to improve*).

The remainder of this paper begins by thoroughly contrasting psychometrics and traditional AI benchmarking (§ 4.1), identifying key differences such as core goals, philosophical foundations, measurement methods, and result analysis (Table 1). We then review how recent LLM psychometrics research adopts and extends psychometric principles to underpin and reinvent LLM benchmarks (§ 4.2).

We proceed by examining the principal constructs assessed within LLM psychometrics (*what to measure*; § 5), including both personality (§ 5.1) and cognitive constructs (§ 5.2). Personality constructs are categorized into personality traits, values, morality, and attitudes and opinions, while cognitive constructs comprise heuristics and biases, social interaction abilities, psycholinguistic abilities, and learning and cognitive capacities. For each construct, we synthesize the underlying theoretical frameworks and measurement instruments, critically compare alternative approaches, and distill the main empirical findings.

We further scrutinize the psychometric evaluation methodologies employed for LLMs (*how to measure*; § 6). The methodological framework is systematically analyzed in terms of its key components: test format (§ 6.1), data and task sources (§ 6.2), prompting strategies (§ 6.3), and model output and scoring (§ 6.4). In addition, we discuss the role of inference parameters in shaping evaluation results (§ 6.5).

We next address the psychometric validation of measurement results, examining both reliability and validity (*how well do we measure*; § 7). Reliability is assessed using metrics such as internal consistency, parallel forms reliability, and inter-rater agreement (§ 7.1). Validity is considered from content, construct, criterion, and ecological perspectives (§ 7.2). Persistent concerns include prompt sensitivity, data contamination, response biases, and divergence from human psychological constructs. To mitigate these challenges, recent research advocates for standardized, procedural tests, including multiple task versions, prompt perturbation, and transparent reporting practices, thereby promoting rigorous, reproducible, and meaningful psychometric evaluation of LLMs (§ 7.3).

The psychometric insights further inform the development and refinement of LLMs (*how to improve*; § 8). Current research predominantly focuses on enhancing LLMs' performance along three dimensions: trait manipulation (§ 8.1), safety and alignment (§ 8.2), and cognitive enhancement (§ 8.3). Finally, we discuss the current trends, challenges, and future directions of LLM psychometrics (§ 9).

4 Psychometrics for Benchmarking Principles

Table 1: Comparison between psychometrics and conventional AI benchmark.

Feature	**Psychometrics**	**AI benchmark**
Core goal	To measure psychological constructs, to prove that a test measures as intended (validity evidence), and to understand the construct being measured.	To test and compare the task performance of different LLMs. Focuses on ranking models and selecting the best one suited for a specific task.
Philosophy of measurement	Construct-oriented. Tends towards a causal approach to measurement, where the measured trait is believed to cause the measurement outcomes.	Task-oriented. Leans towards representativism, assuming items exhaust or represent all aspects of the underlying ability.
Target construct	Personality and ability.	Mostly task-specific abilities.
Construct definition	Emphasizes clear and detailed definitions of the construct being measured. Agreement on the construct definition is a byproduct of test development.	Often defines constructs implicitly through ad hoc task selection. Construct definitions can be vague.
Development process	Systematic and rigorous, often following methods like Evidence-Centered Design (ECD). Can be labor-intensive.	Compiles a set of relevant questions or tasks, then performs expert annotation or crowdsourcing to label ground truth answers. Less labor-intensive per item.
Number of items	Can vary, but not necessarily large. Focus is on item quality and relevance to the construct.	Typically consists of an extensive number of questions to cover various aspects of abilities. Reliability increases with test length.
Sample size	Typically requires a larger sample size of test takers for robust statistical modeling.	Can be applied to evaluate the performance of a single LLM on the benchmark.
Statistical modeling	Employs advanced and various statistical models like Item Response Theory and Factor Analysis to analyze data, estimate latent abilities, and assess model fit.	Often relies on simple aggregation methods, such as calculating average accuracy across benchmark tasks.
Result analysis	Ensures the reliability, validity, predictive power, and explanatory power of the test through result analysis and statistical modeling.	Reliability is likely to be high due to the large number of items. However, validity, predictive power, or explanatory power beyond the target task is not a primary concern.

4.1 Fundamental Differences Between Psychometrics and AI Benchmarking

Benchmarking AI systems superficially resembles psychometrics, particularly Classical Test Theory (CTT) [Crocker and Algina, 1986], as both compile test items to evaluate cognitive capabilities and average the resulting scores. However, closer examination reveals that AI benchmarks differ significantly from modern psychometric approaches [Federiakin, 2025, Wang et al., 2023a]. We outline these key differences in Table 1.

Psychometrics. Psychometrics centers on understanding the psychological constructs and ensuring that tests accurately measure the intended constructs. Grounded in a causal measurement philosophy, this field posits that observed test responses arise from latent psychological constructs [Federiakin, 2025, Markus and Borsboom, 2013]. These constructs may encompass both abilities (e.g., reasoning skills) and personalities (e.g., conscientiousness). The causal framework necessitates rigorous construct definition, requiring traits to be precisely delineated through iterative theory-building and empirical validation. Psychometric test development follows methodical protocols, often structured by frameworks such as Evidence-Centered Design (ECD) [Mislevy et al., 2003]. ECD emphasizes ensuring congruence between test items and theoretical models of the construct, thereby supporting robust inferences about latent traits.

Central to this approach is the prioritization of item quality over quantity. Psychometricians conduct rigorous item analyses to balance precision with practicality, as administering an excessive number of items to participants is often impractical. Advanced statistical models, such as Item Response Theory (IRT) [Embretson and Reise, 2013] and Factor Analysis [Loehlin, 2004], are adopted to estimate latent traits, analyze item performance, and assess model fit. These models require relatively large sample sizes (the number of human participants) to yield stable parameter estimates, as they must disentangle individual differences from measurement error to accurately infer latent traits.

The test results are analyzed to ensure the reliability, validity, predictive power, and explanatory power of the test. Specifically, a well-designed test should: 1) consistently and accurately measure the intended construct, 2) predict performance across a diverse range of related tasks and real-world outcomes, and 3) provide explanatory insight into the observed data. For example, psychometric models often reveal that individual differences across a broad range of cognitive tasks can be captured and explained by a relatively small set of underlying cognitive abilities [Cattell and Horn, 1978].

Benchmark. In contrast, AI benchmarking is driven by pragmatic goals: evaluating and ranking models based on task performance. Unlike psychometrics, validity is not the primary concern. Instead, benchmarks typically emphasize breadth, scalability, and—especially in the era of foundation models—difficulty. This approach reflects a representativist philosophy, where it is assumed that an extensive set of benchmark items collectively captures all relevant aspects of the abilities demanded by the target task [Federiakin, 2025]. However, constructs like reasoning or knowledge are often ambiguously defined and encompass infinitely many aspects. Benchmarks implicitly operationalize these constructs through ad hoc task selection.

The development of AI benchmarks is usually less labor-intensive, especially when compared with psychometrics on a per-item basis. Test items and their corresponding ground truths are typically drawn from existing datasets, expert curation, or crowdsourced contributions. While this process enables scalability, it risks conflating superficial task performance with deeper cognitive capacities. For instance, a benchmark may assess mathematical reasoning through arithmetic problems without verifying whether models rely on pattern recognition versus symbolic logic [Ahn et al., 2024]. Additionally, LLM benchmarking commonly employs straightforward metrics, such as average accuracy, eschewing the sophisticated latent variable models of psychometrics. This simplicity allows benchmarks to evaluate single models efficiently, bypassing the need for population samples. However, it also limits the depth of insights that can be gleaned from model performance [Federiakin, 2025].

Reliability and stability in benchmarking are primarily achieved through scaling up the test. However, ensuring the quality of each individual item becomes impractical due to the test scale and the rapid pace of model development. For instance, while psychometrics emphasizes the discriminative power of each item, some benchmarks, though initially challenging, are quickly outpaced by continuous model improvements [McIntosh et al., 2024]. Conversely, certain emerging benchmarks are currently too difficult to yield meaningful comparisons [Phan et al., 2025]. Benchmark results are often limited to the specific target task, offering limited generalizability or predictive power across other tasks or real-world applications. These results also pose significant challenges for conducting in-depth, multi-faceted analyses of model capabilities [Wang et al., 2023a].

4.2 Benchmarking with Psychometrics-Inspired Principles

Recent LLM evaluation efforts have drawn inspiration from psychometrics and seek to develop benchmarks that adhere to psychometric principles.

Construct-Oriented Benchmarking. Task-oriented benchmarks often entail vast question sets to capture complex abilities. However, in many cases, the benchmarks either fail to fully represent these abilities due to their infinite manifestations or involve extraneous factors that are irrelevant to the target ability [Wallach et al., 2025, Zhou et al., 2025]. Recent research has drawn inspiration from psychometrics and explored the paradigm of construct-oriented evaluation, seeking its discriminative, predictive, and explanatory power. Federiakin [2025], Ilić and Gignac [2024] employ factor analysis to explore the latent variables underlying LLM benchmark performance. Their findings reveal a monolithic factor resembling general intelligence or ability. Federiakin [2025] ranks models based on this discovered factor and highlights its unique advantages over raw benchmark scores. In contrast, similar attempts by Burnell et al. [2023] identify three factors—reasoning, comprehension, and core language modeling—that better explain LLM performance across 27 cognitive tasks. Based on it, Zhu et al. [2024a] integrate the three factors into benchmark items to evaluate multifaceted abilities. This discrepancy between the estimated latent factors in the above findings can be attributed to differences in the models and benchmarks employed [Zhou et al., 2025]. Therefore, rather than relying on statistically derived factors, Zhou et al. [2025] propose a theory-driven hierarchical set of general scales for systematic construct-oriented evaluation. These scales are validated to explain what AI systems can do and predict their performance on novel task instances. Peng et al. [2024] present the Tong Test, a value- and ability-oriented framework, for Artificial General Intelligence (AGI) evaluation. This framework is rooted in dynamic embodied physical and social interactions (DEPSI), and can generate an infinite variety of tasks to evaluate key capabilities including values, learning, and cognition.

Psychometrically Rigorous Benchmark Development. Beyond defining and analyzing latent constructs, researchers have developed holistic, psychometrically rigorous methods for benchmark development. Liu et al. [2024e] introduce Evidence-Centered Benchmark Design (ECBD), a framework structuring benchmark creation into five modules—capability, content, adaptation, assembly, and evidence—each requiring justification to ensure validity. Through case studies of prominent LLM benchmarks, they demonstrate ECBD's utility in identifying validity threats. Similarly grounded in psychometrics but with a distinct approach, Fang et al. [2024] propose Psychometrics-Assisted Benchmarking (PATCH), an eight-step process from construct definition to proficiency scoring. When piloted on 8th-grade mathematics, PATCH produced results diverging from traditional benchmarks, offering a more comprehensive evaluation. Building on related principles, Kardanova et al. [2024] adapt Evidence-Centered Design (ECD) to create psychometrically grounded benchmarks. Their application in pedagogy illustrates how this method can reduce data contamination and enhance test interpretability.

Item Response Theory. Item Response Theory (IRT) is a statistical approach in psychometrics that jointly estimates the latent ability of examinees and the properties (difficulty, discrimination) of test items [Embretson and Reise, 2013]. Translating this framework to LLM evaluation enables researchers to infer latent ability scores, assess item informativeness, and perform more efficient evaluation. Recent research leverages principles from IRT to develop adaptive evaluation frameworks. These methods dynamically calibrate item difficulty based on model performance and weighting items by their inferred difficulty, aiming to achieve accurate evaluations with smaller test sizes and more discriminative items [Guinet et al., 2024, Lalor et al., 2024, Polo et al., 2024, Zhuang et al., 2023a,b]. Building on these approaches, Jiang et al. [2024a], Truong et al. [2025] introduce IRT-based benchmarks that involve learning to estimate item difficulty and learning to generate novel items calibrated to specific difficulty levels. Additionally, research has used IRT-based analyses to explore the alignment between LLM and human response distributions [He-Yueya et al., 2024]. IRT-based evaluations further offer the potential to estimate construct and item parameters on a unified scale, enabling direct comparisons across AI systems and against human norms, even when different test sets are used [Fang et al., 2024, Wang et al., 2023a].

5 Psychometrics for Measuring Psychological Constructs

This section delves into the psychological constructs evaluated in LLM psychometrics. Fig. 2 exemplifies the tests for the involved constructs.

5.1 Measuring Personality Constructs

LLMs exhibit personality constructs that are not explicitly programmed or trained towards. The emergent constructs critically shape LLM behavior [Hagendorff, 2023, Ye et al., 2025a], with profound implications for both individuals and broader social groups [Bengio et al., 2024]. Measuring these embedded psychological constructs is essential for understanding behavior, identifying biases, and fostering responsible development.

Table 2 summarizes the representative personality constructs that have garnered attention in recent research. Researchers typically select constructs based on their relevance to LLM development and deployment, as well as the applicability

Table 2: Representative LLM *personality* constructs in LLM psychometrics. The main dimensions or focus of the theories/inventories are listed in Table 3.

Constructs	Theories/inventories	Related work
Personality traits	Big Five	Ai et al. [2024], Bhandari et al. [2025a,b], Bodroža et al. [2024], Caron and Srivastava [2022], Dorner et al. [2023], Frisch and Giulianelli [2024], Gupta et al. [2024], Hilliard et al. [2024], Huang et al. [2023a,e, 2024c], Jain et al. [2024], Jiang et al. [2023, 2024b], Karra et al. [2022], Klinkert et al. [2024], Kovač et al. [2023], La Cava and Tagarelli [2024], Lee et al. [2024d], Li et al. [2022a,b], Liu et al. [2024a], Lu et al. [2023], Pellert et al. [2024], Petrov et al. [2024], Ren et al. [2024], Romero et al. [2024], Serapio-García et al. [2023], Shu et al. [2024], Song et al. [2023], Sühr et al. [2023], Zhang [2024], Zheng et al. [2025], Zou et al. [2024]
	HEXACO	Barua et al. [2024], Bodroža et al. [2024], Miotto et al. [2022], Peereboom et al. [2024], Ren et al. [2024], Wang et al. [2025]
	MBTI	Ai et al. [2024], Chen et al. [2024a], Cui et al. [2023], Huang et al. [2023c], La Cava and Tagarelli [2024], Lu et al. [2023], Pan and Zeng [2023], Rao et al. [2023], Song et al. [2024b], Zhang et al. [2024a]
	Dark Triad	Barua et al. [2024], Huang et al. [2023e], Lee et al. [2024d], Li et al. [2022a,b], Lu et al. [2023], Peereboom et al. [2024], Romero et al. [2024]
	Other & custom	Wen et al. [2024b] (reviews); Ai et al. [2024], Jiang et al. [2023], Mao et al. [2024a], Zeng [2024]
Values	Schwartz	Cahyawijaya et al. [2024], Fischer et al. [2023], Hadar-Shoval et al. [2024], Kovač et al. [2023, 2024], Lee et al. [2024a], Li et al. [2024b], Miotto et al. [2022], Pellert et al. [2024], Ren et al. [2024], Rozen et al. [2024], Shen et al. [2024], Yao et al. [2024, 2025a,b], Ye et al. [2025a], Zhang et al. [2023a]
	WVS	Chiu et al. [2025], Faulborn et al. [2025], Kim and Baek [2024], Yao et al. [2025a]
	VSM	Kharchenko et al. [2024], Kovač et al. [2023], Ren et al. [2024], Ye et al. [2025a], Zhong et al. [2024]
	GLOBE	Karinshak et al. [2024], Li et al. [2024d], Ren et al. [2024]
	SVO	Zhang et al. [2024c]
	Other & custom	Biedma et al. [2024], Jiang et al. [2024a], Li et al. [2024d], Liu et al. [2024b], Meadows et al. [2024], Moore et al. [2024], Xu et al. [2023], Ye et al. [2025b], Zhang et al. [2024b]
Morality	MFT	Abdulhai et al. [2024], Aksoy [2024], Fraser et al. [2022], Ji et al. [2024], Münker [2024], Neuman et al. [2025], Nunes et al. [2024], Pellert et al. [2024], Simmons [2023], Tlaie [2024], Yan et al. [2024], Yao et al. [2025b], Zhou et al. [2024a]
	ETHICS	Albrecht et al., Jinnai [2024], Karpov et al. [2024], Rodionov et al. [2023], Yu et al. [2023]
	DIT	Khandelwal et al. [2024], Tanmay et al. [2023]
	Other & custom	Ahmad and Takemoto [2024], Bonagiri et al. [2024], Chiu et al. [2025], Garcia et al. [2024], Han [2023], Huang et al. [2024a], Jiang et al. [2024a], Jin et al. [2022, 2024b], Jinnai [2024], Kucuk and Kocyigit [2023], Liu et al. [2024d], Marraffini et al. [2024], Meijer et al. [2024], Neuman et al. [2025], Ohashi et al. [2024], Peterson [2025], Ramezani and Xu [2023], Sachdeva and van Nuenen [2025], Scherrer et al. [2023], Seror [2024], Takemoto [2024], Tanmay et al. [2023], Vida et al. [2024], Yuan et al. [2024]
Attitudes & opinions	ANES	Argyle et al. [2023], Bisbee et al. [2024], Jiang et al. [2022, 2024d], Qi et al. [2024], Sun et al. [2024], Yang et al. [2024]
	ATP	Hwang et al. [2023], Santurkar et al. [2023], Tjuatja et al. [2024]
	GLES	Ball et al. [2025], Ma et al. [2024b], Qi et al. [2024], von der Heyde et al. [2024]
	PCT	Azzopardi and Moshfeghi [2024], Bernardelle et al. [2024], Hartmann et al. [2023], Röttger et al. [2024], Rozado [2023], Rutinowski et al. [2024], Wright et al. [2024]
	Other & custom	Ma et al. [2024a] (reviews); Ceron et al. [2024], Chalkidis and Brandl [2024], Dominguez-Olmedo et al. [2024], Durmus et al. [2023], Faulborn et al. [2025], Feng et al. [2023], Geng et al. [2024], Kalinin [2023], Kim and Lee [2023], Kim et al. [2025b], Lee et al. [2024c], Rosenbusch et al. [2023], Röttger et al. [2025], Rozado [2023], Sanders et al. [2023], Wu et al. [2023], Xu et al. [2025c]

Table 3: Personality theories and inventories measured in LLM psychometrics and their main dimensions or focus.

Theory/inventory	What it measures / dimensions
Big Five	Five broad personality traits: *Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism*
HEXACO	Six personality traits: *Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, Openness*
MBTI	Four dichotomies: *Extraversion–Introversion, Sensing–Intuition, Thinking–Feeling, Judging–Perceiving*
Dark Triad	Three negative personality traits: *Narcissism, Machiavellianism, Psychopathy*
Schwartz	Basic human values: 10 or more values (e.g., *Self-Direction, Stimulation, Hedonism, Achievement, Power, Security, Conformity, Tradition, Benevolence, Universalism*), typically grouped into four higher-order categories (*Openness to Change, Self-Enhancement, Conservation, Self-Transcendence*)
WVS	World Values Survey: Assesses broad cultural values such as *traditional vs. secular-rational values, survival vs. self-expression values*
VSM	Value Survey Module (often Hofstede): Cultural dimensions such as *Power Distance, Individualism, Masculinity, Uncertainty Avoidance, Long-Term Orientation, Indulgence*
GLOBE	Global Leadership and Organizational Behavior Effectiveness: Nine cultural dimensions (e.g., *Performance Orientation, Assertiveness, Future Orientation, Humane Orientation, Institutional Collectivism, In-Group Collectivism, Gender Egalitarianism, Power Distance, Uncertainty Avoidance*)
SVO	Social Value Orientation: Measures individuals' preferences regarding resource allocation between oneself and others (e.g., *prosocial, individualistic, competitive orientations*)
MFT	Moral Foundations Theory: Five (sometimes six) moral foundations—*Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, (Liberty/Oppression)*
ETHICS	Various ethics-related measures assessing moral reasoning, ethical principles, or moral preferences
DIT	Defining Issues Test: Assesses moral development and reasoning using moral dilemmas
ANES	American National Election Studies: Political attitudes, beliefs, and behaviors in the U.S.
ATP	Attitudes Toward Politics: Measures political attitudes and social controversies, often specific to a region or subject
GLES	German Longitudinal Election Study: Political attitudes, beliefs, and voting behaviors in Germany
PCT	Political Compass Test: Economic (*Left/Right*) and Social (*Authoritarian/Libertarian*) political dimensions

of these constructs to AI systems. For instance, Li et al. [2024d] argue that emotional variability in LLMs is not a meaningful construct, given that LLMs lack the biological mechanisms underlying emotions. Conversely, personality and values are considered meaningful for LLMs, as they influence user interactions and model outputs [Serapio-García et al., 2023, Ye et al., 2025a]. This rationale has led to extensive research on personality, values, morality, and attitudes & opinions, as well as some—though less—focus on other constructs such as career selection [Hua et al., 2024], motivation [Chiu et al., 2025, Huang et al., 2023e], and mental health [De Duro et al., 2024, Reuben et al., 2024].

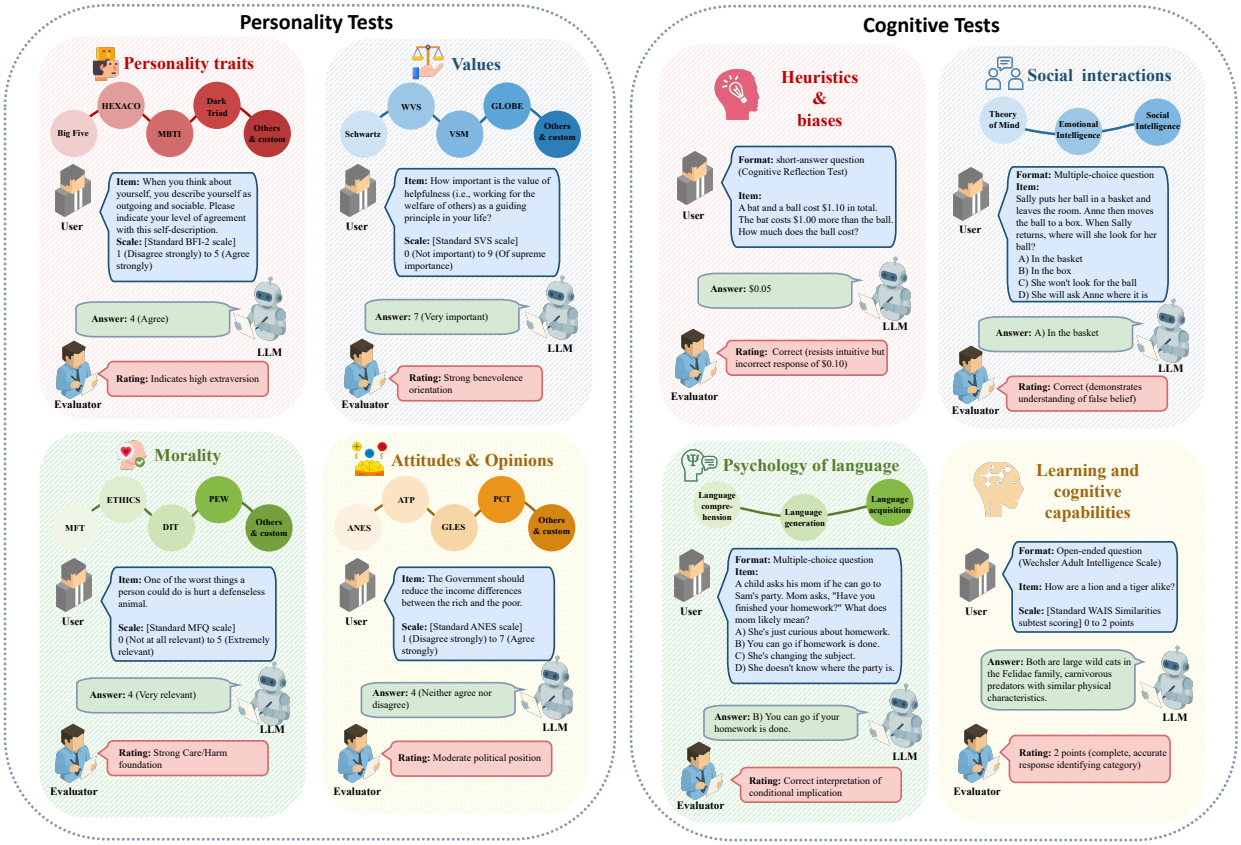

Figure 2: Examples of psychometric tests for LLMs.

5.1.1 Personality Traits

> "**Personality** is the enduring configuration of characteristics and behavior that comprises an individual's unique adjustment to life." [APA Dictionary of Psychology, n.d.]

Personality traits define an individual's patterns of thinking, feeling, and behaving [Larsen et al., 2005]. In the context of LLMs, Wang et al. [2025], Zhang et al. [2024a] have explored how personality traits relate to model safety, bias, and toxicity. Other works emphasize that chatbot personality highly determines user experience [Huang et al., 2024c, Jiang et al., 2023, Klinkert et al., 2024, Serapio-García et al., 2023].

The study of personality traits leads to several prominent theoretical models, each offering unique insights into individual differences. The Big Five model [Goldberg, 2013] identifies five core dimensions of personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Building upon the Big Five, the HEXACO model [Ashton and Lee, 2007] introduces an additional dimension, Honesty-Humility, which encompasses traits such as sincerity, fairness, and modesty. Another influential framework, the Myers-Briggs Type Indicator (MBTI) [Myers et al., 1962], categorizes individuals into 16 distinct personality types based on four dichotomous dimensions: Extraversion vs. Introversion, Sensing vs. Intuition, Thinking vs. Feeling, and Judging vs. Perceiving. In contrast to these models, the Dark Triad [Paulhus and Williams, 2002] focuses on three socially aversive personality traits—Machiavellianism, Narcissism, and Psychopathy—highlighting the darker aspects of human behavior.

When applying psychometrics to LLMs, most researchers directly administered the established inventories, such as NEO-PI-R [Costa and McCrae, 2008], BFI [John et al., 1991], and BFI-2 [Soto and John, 2017], for measuring the Big Five traits [Caron and Srivastava, 2022, Huang et al., 2023a,d, Karra et al., 2022, La Cava and Tagarelli, 2024, Serapio-García et al., 2023]; HEXACO-60 [Ashton and Lee, 2009] and HEXACO-100 [Lee and Ashton, 2018] inventories for HEXACO traits [Miotto et al., 2022, Wang et al., 2025]; MBTI assessment [Myers, 1985] for MBTI types [Cui et al., 2023, Huang et al., 2023c, La Cava and Tagarelli, 2024, Pan and Zeng, 2023, Rao et al., 2023]; and

Dark Triad Dirty Dozen scale [Jonason and Webster, 2010] for Dark Triad traits [Barua et al., 2024, Huang et al., 2023d, Romero et al., 2024].

Given widespread concerns about the practical relevance of these inventories [Ai et al., 2024, Dorner et al., 2023, Gupta et al., 2024, Pellert et al., 2024, Serapio-García et al., 2023, Shu et al., 2024, Song et al., 2023], Some researchers have adapted them for more real-world scenarios. For example, Bhandari et al. [2025a], Frisch and Giulianelli [2024], Klinkert et al. [2024], Liu et al. [2024a], Song et al. [2024b], Zheng et al. [2025], Zou et al. [2024] contextualize the tests in topic-specific or open-domain conversations. Ai et al. [2024] accompany the self-report tests with behavioral tests to examine the personality knowledge of LLMs. Meanwhile, Jiang et al. [2023] compile and adapt existing tests into the Machine Personality Inventory, while Mao et al. [2024a] develop the PersonalityEdit inventory using LLMs. However, Peereboom et al. [2024] argue that human-derived traits may not meaningfully apply to LLMs, underscoring the necessity for psychometric theories specifically designed for LLM analysis (§ 7.2.2).

Comparing Personality Models and Suggesting Use Cases. The Big Five offers robust empirical foundations and is the most prominent model for personality assessment. However, it is insufficient for predicting comprehensive psychological outcomes [Feher and Vernon, 2021]. HEXACO extends the model by adding Honesty-Humility and provides nuanced insights into moral traits, making it preferable for LLM application contexts demanding integrity [Wang et al., 2025]. In contrast, MBTI [Myers et al., 1962] categorizes 16 types via dichotomous scales. While popular in career counseling and team-building for its accessibility, its typological approach oversimplifies personality and lacks strong empirical support, limiting its utility in high-stakes decisions [Pittenger, 2005]. The Dark Triad [Paulhus and Williams, 2002] diverges by focusing on maladaptive traits, making it useful in identifying toxic behaviors in LLMs [Barua et al., 2024].

Main Findings. Early models such as GPT-3 score above human averages on Dark Triad scales [Li et al., 2022a,b, Romero et al., 2024]. Even after safety tuning, some models retain certain "dark qualities" [Li et al., 2022b], suggesting that these patterns may be deeply rooted in the pretraining data. More advanced LLMs indicate better level of alignment. They usually demonstrate high Openness, Extraversion, and Agreeableness traits, while low Neuroticism, on the Big Five personality tests [Bhandari et al., 2025a, Huang et al., 2023e, Karra et al., 2022, Klinkert et al., 2024, La Cava and Tagarelli, 2024, Zou et al., 2024]. The results align with their design as assistive and helpful entities with emotional stability and engaging personality [Bhandari et al., 2025a, La Cava and Tagarelli, 2024]. Within the MBTI framework, most proprietary LLMs are classified as ENFJ or INFJ types [Huang et al., 2023c, La Cava and Tagarelli, 2024, Zhang et al., 2024a], indicating tendencies toward helpfulness, idealism, and planning capabilities.

Personality differences between models are noteworthy. Comparative studies reveal that LLMs from different generations and training methodologies display unique combinations of personality characteristics [Bhandari et al., 2025a, Bodroža et al., 2024, Huang et al., 2023e, Jiang et al., 2023, Karra et al., 2022, Klinkert et al., 2024, La Cava and Tagarelli, 2024]. LLM personalities also vary across contexts. The same model may display different personality traits across conversational topics [Bodroža et al., 2024, Caron and Srivastava, 2022, Petrov et al., 2024, Song et al., 2023, Zou et al., 2024]. This variability challenges the applicability of human personality theories that assume relative trait stability. Multiple studies suggest that prompt design and system instructions substantially influence LLM personality expression [Caron and Srivastava, 2022, Dorner et al., 2023, Gupta et al., 2024, Song et al., 2023]. We defer related discussions to § 7.

5.1.2 Values

> "**Values** are enduring beliefs that guide behavior and decision-making, reflecting what is important and desirable to an individual or group." [Schwartz, 1992]

Values are central to understanding human motivation, attitudes, and behavior, shaping how individuals perceive themselves and the world around them. According to Schwartz's theory of basic human values [Schwartz, 2012], the most established value theory, they 1) are beliefs linked inextricably to affect, 2) are desirable goals that motivate actions, 3) transcend specific situations and actions, 4) serve as standards or criteria for evaluating and selecting actions, policies, people, and events, and 5) are ordered by relative importance, which guides choices and actions. These unique characteristics make values a powerful lens for understanding LLM behavior. For example, Ye et al. [2025a] show how different values contribute to the safety of LLMs; Liu et al. [2024b] reveal that different spiritual values affect LLMs in social-fairness scenarios; and Sorensen et al. [2024b] demonstrate that standard value alignment reduces distributional pluralism in LLM outputs.

Schwartz's Theory. Schwartz [1992] identify 10 basic human values: Self-Direction, Stimulation, Hedonism, Achievement, Power, Security, Conformity, Tradition, Benevolence, and Universalism. These values are further grouped

into four higher-order dimensions: Openness to Change, Conservation, Self-Enhancement, and Self-Transcendence. Later iterations of the theory refine its nuances by subdividing the 10 values into finer categories [Schwartz et al., 2012]. Most research on Schwartz's values in LLMs focuses on the 10 basic values, while some studies explore the 19 refined values [Rozen et al., 2024]. Many investigations directly apply established inventories, i.e., Schwartz Value Survey (SVS) [Schwartz, 1992] or the Portrait Values Questionnaire (PVQ) [Schwartz et al., 2001] (e.g., [Fischer et al., 2023, Hadar-Shoval et al., 2024, Kovač et al., 2023, Lee et al., 2024a, Miotto et al., 2022]). Other research adapts these inventories to better suit the context of LLMs. For example, Ren et al. [2024] rephrase survey items into advice-seeking questions, and Shen et al. [2024] transform subjective statements into opinion-seeking questions about AI actions. Rather than directly administering or rephrasing static inventories, Chiu et al. [2025], Yao et al. [2024], Ye et al. [2025a] leverage existing cross-domain LLM prompt datasets or LLM-generated contextualized questions as test items.

World Values Survey (WVS). The World Values Survey (WVS) is a global research project that explores people's values and beliefs, how they change over time, and their social and political impact [Haerpfer et al., 2022]. It measures a broad spectrum of values, including cultural values defined on two broad dimensions: traditional vs. secular-rational values and survival vs. self-expression values. Kim and Baek [2024] directly prompt LLMs with WVS survey items. In contrast, Chiu et al. [2025], Yao et al. [2025a] use model-generated or self-designed prompts to measure these cultural values in LLMs.

Values Survey Module (VSM). Hofstede [1984] identifies six key value dimensions that describe cultural differences: Power Distance, Individualism vs. Collectivism, Masculinity vs. Femininity, Uncertainty Avoidance, Long-term vs. Short-term Orientation, and Indulgence vs. Restraint. These dimensions help explain how cultural values influence behavior in societies and organizations [Hofstede, 2001]. VSM is a questionnaire-based tool that assesses cultural preferences across these dimensions [Hofstede, 2011]. Kovač et al. [2023], Zhong et al. [2024] directly apply the VSM to evaluate cultural values in LLMs, while Kharchenko et al. [2024], Ren et al. [2024], Ye et al. [2025a] adapt the VSM or use self-designed inventories to measure these dimensions.

Global Leadership and Organizational Behavior Effectiveness (GLOBE). The GLOBE study investigates cultural values, leadership behaviors, and organizational practices across countries [House et al., 2004]. Building on Hofstede's work [Hofstede, 1984], it incorporates additional dimensions: Performance Orientation, Assertiveness, and Humane Orientation, resulting in nine cultural dimensions. Karinshak et al. [2024], Li et al. [2024d], Ren et al. [2024] apply or adapt the GLOBE culture questionnaire to evaluate cultural values in LLMs.

Social Value Orientation (SVO). SVO is a psychological framework that measures individual preferences for allocating resources between oneself and others [Messick and McClintock, 1968]. It focuses on distinguishing between prosocial (e.g., altruists, cooperators) and proself (e.g., individualists, competitors) orientations. SVO is typically measured using experimental tasks or questionnaires, such as the SVO Slider Measure [Murphy et al., 2011] or Decomposed Games [Liebrand, 1984]. These tools assess how individuals allocate resources in hypothetical scenarios involving themselves and others. For example, Zhang et al. [2024c] apply the SVO Slider Measure to evaluate social preferences in LLMs.

Other or Custom Theories and Inventories. Meadows et al. [2024], Xu et al. [2023] introduce localized inventories to evaluate LLMs based on Australian and Chinese values, respectively. Some researchers focus on specific topics, such as age-related values [Zhang et al., 2024b] and spiritual values [Liu et al., 2024b]. Other studies involve large-scale, cross-domain prompt datasets [Li et al., 2024d, Moore et al., 2024], among which Jiang et al. [2024a] concentrate on eliciting harmful, biased, or ethics-related content from LLMs. Notably, Biedma et al. [2024], Ye et al. [2025b] challenge established value theories and propose novel ones for LLM values, built in a bottom-up manner using psycholexical data.

Comparing Value Theories and Suggesting Use Cases. Schwartz emphasizes universal human motivations; WVS examines societal-level value shifts; VSM and GLOBE address cross-cultural dimensions, with GLOBE emphasizing leadership. SVO, in contrast, focuses on individual-level social preferences, i.e., prosocial vs. proself. While Schwartz and SVO overlap in analyzing individual motivations, the former is broader, and the latter narrowly focuses on decision-making in resource allocation. WVS, VSM, and GLOBE share cultural-level dimensions but differ in focus—WVS on modernization, VSM on workplace-related cultural differences, and GLOBE on leadership and organizational practices. For evaluating *individual values* in LLMs, use Schwartz for a comprehensive understanding of universal motivations or SVO for interpersonal decision-making studies. In contrast, choose WVS for evaluating broad *societal values* in LLMs. When evaluating *cultural dimensions* in LLMs, use VSM for workplace cultural differences or GLOBE for leadership and organizational practices. In addition, Schwartz's theory bridges individual and societal-level values, making it versatile for mixed-focus studies.

Main Findings. Research indicates that LLMs display distinct and systematic value patterns. According to Schwartz's Value Theory, LLMs tend to prioritize Self-Transcendence and Conservation. They exhibit stronger inclinations toward Universalism, Benevolence, Conformity, and Security, while opposing Power and Achievement [Fischer et al., 2023, Hadar-Shoval et al., 2024, Rozen et al., 2024, Zhang et al., 2023a]. WVS surveys further suggest that LLMs generally prefer Self-Expression values over Survival values [Chiu et al., 2025]. Studies using the VSM and GLOBE frameworks emphasize LLMs' strong focus on Humane and Performance Orientation, with moderate Assertiveness [Li et al., 2024b]. Additionally, when assessed through SVO, advanced LLMs predominantly show Prosocial tendencies [Zhang et al., 2024c].

Different models exhibit varied value orientations [Chiu et al., 2025, Duan et al., 2023, Kovač et al., 2024, Li et al., 2024b, Xu et al., 2023]. Versions within the same model family show evolutionary trends in values, potentially influenced by safety alignment, capability advancement, and shifting societal expectations [Duan et al., 2023, Kim and Baek, 2024, Moore et al., 2024]. Generally, larger models align more closely with desirable human values [Jiang et al., 2024a, Kim and Baek, 2024, Shen et al., 2024, Zhong et al., 2024].

Cross-cultural research suggests that LLMs may embody a blend of cultural values, integrating perspectives from diverse backgrounds [Kovač et al., 2023]. However, they generally exhibit a tendency toward Western liberal values [Kim and Baek, 2024]. LLMs can display different values based on profiling prompts [Karinshak et al., 2024, Kharchenko et al., 2024, Zhong et al., 2024]. The context-dependency of their values is concerning and challenges the stability assumptions in human value theories [Chiu et al., 2025, Kovač et al., 2023, Meadows et al., 2024, Moore et al., 2024, Shen et al., 2024, Xu et al., 2023]. Further discussions in reliability and validity are presented in § 7.

5.1.3 Morality

> "**Morality** is the categorization of intentions, decisions and actions into those that are proper, or right, and those that are improper, or wrong." [Long and Sedley, 1987]

Morality is a fundamental aspect of human behavior, influencing social interactions, decision-making, and ethical reasoning. The *Moral Foundations Theory (MFT)* [Graham et al., 2009] posits that morality is shaped by six innate psychological systems: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, and Liberty/Oppression. These moral foundations are hypothesized to be universal across cultures, providing a common framework for understanding moral judgments and behavior. The theory suggests that individuals vary in the extent to which they endorse each foundation, leading to diverse moral profiles that influence their attitudes, beliefs, and actions.

It is crucial to conduct moral assessments of LLMs to ensure their ethical deployment. A large body of research (e.g., [Fraser et al., 2022, Liu et al., 2024d, Tlaie, 2024, Yao et al., 2025a, Zhou et al., 2024a]) applies MFT, primarily through the Moral Foundations Vignettes (MFVs) [Clifford et al., 2015], the Moral Foundations Questionnaire (MFQ) [Graham et al., 2009], the MFQ-2 [Atari et al., 2023], and the Moral Foundations Dictionary (MFD) [Graham et al., 2009]. These papers investigate model bias, alignment with political/moral ideologies, and variation in personal or cultural values, using controlled prompt tests [Abdulhai et al., 2024, Nunes et al., 2024, Tlaie, 2024], persona-driven exploration [Münker, 2024], or evaluating internal moral coherence [Nunes et al., 2024].

Other prominent moral theoretical frameworks and survey tools for evaluating LLMs include Kohlberg's Theory via the Defining Issues Test (DIT) [Kohlberg, 1964], the Consequentialist-Deontological distinction [Beauchamp, 2001], and the PEW 2013 Global Attitudes Survey [Center, 2013]. Khandelwal et al. [2024], Tanmay et al. [2023] utilize the DIT, a psychometric tool aligned with Kohlberg's stages of moral development, to analyze structural reasoning in moral dilemmas. They report that models like GPT-4 exhibit post-conventional moral reasoning. Khandelwal et al. [2024] further extend this analysis to a multilingual context, examining cross-linguistic variations in moral reasoning abilities. Neuman et al. [2025] explore ethical reasoning using established typologies, including consequentialist-deontological analysis. The results uncover the general rationalist, consequentialist emphasis in LLMs. According to the PEW 2013 Global Attitudes Survey, Meijer et al. [2024] find that the models display homogeneous moral values and have limited effectiveness in reflecting cross-cultural morality.

Localized moral theoretical frameworks are essential for ensuring the ethical deployment of LLMs in various cultural contexts. Liu et al. [2024d] develop the Chinese Moral Dictionary (CMD) and find that ChatGPT and Gemini prefer individualistic moral beliefs, while Ernie and ChatGLM lean towards collectivist moral beliefs. Ohashi et al. [2024], Takeshita et al. [2023] create the JCommonsenseMorality (JCM) dataset and fine-tune LLMs to adapt to Japanese culture.

Researchers have also developed specialized datasets for studying LLM morality. Hendrycks et al. [2021] create the ETHICS dataset, which is then utilized and extended to evaluate LLMs' knowledge of morality, covering concepts such as justice, well-being, duties, virtues, and commonsense morality [Albrecht et al., Jinnai, 2024, Karpov et al.,

2024, Rodionov et al., 2023, Yu et al., 2023]. Marraffini et al. [2024] develop the Greatest Good Benchmark (GGB) to evaluate LLMs' moral judgments using utilitarian dilemmas. Jin et al. [2024b] develop the MultiTP, a cross-lingual corpus of moral dilemma vignettes in over 100 languages.

Comparing Moral Theories/Instruments and Suggesting Use Cases. Instruments based on MFT are well-suited for measuring LLMs' alignment with intuitive moral principles across cultural or political dimensions. Tools like MFQ, MFQ-2, MFVs, and MFD are particularly effective when investigating ideological leanings, bias, or internal consistency in abstract vs. contextual moral judgments [Nunes et al., 2024]. In contrast, Kohlberg's DIT explores structured moral reasoning, offering a developmental perspective on how LLMs process ethical dilemmas [Khandelwal et al., 2024, Tanmay et al., 2023]. DIT is valuable for longitudinal studies and evaluating whether models show higher-level moral reasoning (e.g., post-conventional reasoning). Frameworks such as the Consequentialist-Deontological distinction and global surveys like the PEW 2013 study are useful for categorizing ethical tendencies or analyzing cultural homogeneity [Meijer et al., 2024, Neuman et al., 2025]. Localized tools (e.g., CMD, JCM) help assess cultural alignment [Liu et al., 2024d, Takeshita et al., 2023]. Meanwhile, datasets like ETHICS provide morally charged scenarios for detecting normative behavior and fine-tuning LLM responses in high-stakes ethical domains [Hendrycks et al., 2021, Jin et al., 2024b, Marraffini et al., 2024].

Main Findings. LLMs are generally characterized by a rationalist and consequentialist focus, often prioritizing harm minimization and fairness [Neuman et al., 2025]. Despite that, they show divergence in ethical reasoning [Neuman et al., 2025] and moral preferences [Bonagiri et al., 2024, Jin et al., 2024b, Meijer et al., 2024, Tanmay et al., 2023]. In some aspects, most LLMs align with human moral standards [Nunes et al., 2024, Takemoto, 2024, Tanmay et al., 2023], which may be attributed to their extensive exposure to conventional ethical values during training. On the other hand, some research presents a less optimistic view, uncovering significant deviations of LLMs from human moral preferences [Ahmad and Takemoto, 2024, Marraffini et al., 2024, Vida et al., 2024]. Regarding the underlying mechanisms of LLMs' moral reasoning, Ji et al. [2024], Nunes et al. [2024], Simmons [2023] suggest that LLMs primarily exhibit imitation rather than genuine conceptual understanding.

5.1.4 Attitudes & Opinions

> "**Attitudes** are always attitudes about something. This implies three necessary elements: first, there is the object of thought, which is both constructed and evaluated. Second, there are acts of construction and evaluation. Third, there is the agent, who is doing the constructing and evaluating. We can therefore suggest that, at its most general, an attitude is the cognitive construction and affective evaluation of an attitude object by an agent." [Bergman, 1998]

We use the term "attitude" to encompass both attitudes and opinions, following [Bergman, 1998, Ma et al., 2024a], who treat these concepts as synonymous. Most research on LLM attitudes examines political attitudes and public opinions [Ma et al., 2024a], which are key cognitive and behavioral foundations in human society and closely tied to model fairness, credibility, and social impact [Durmus et al., 2023, Hartmann et al., 2023, Lee et al., 2024c, Sanders et al., 2023, Santurkar et al., 2023]. Other biases—such as those related to gender, race, culture, religion, occupation and age—are also studied as part of LLM attitudes, though many works approach them from data and algorithmic perspectives. For a broader discussion on bias and fairness in LLMs, we refer readers to [Gallegos et al., 2024, Ranjan et al., 2024].

To measure the political attitudes of LLMs, researchers often employ standardized questionnaires and scales from political science and social psychology. A widely used tool is the American National Election Studies (ANES)[1], applied in works such as Jiang et al. [2022, 2024d], Qi et al. [2024], Yang et al. [2024] to assess LLMs' stances on U.S. political issues. Similarly, the American Trends Panel (ATP)[2] has been leveraged by Hwang et al. [2023], Santurkar et al. [2023], Tjuatja et al. [2024] to analyze opinion distributions on public policy. For cross-national comparisons, studies like Ball et al. [2025], Ma et al. [2024b], von der Heyde et al. [2024] utilize the German Longitudinal Election Study (GLES)[3]. Additionally, the Political Compass Test (PCT)[4] has gained prominence for positioning LLMs within multidimensional political spectrums [Azzopardi and Moshfeghi, 2024, Bernardelle et al., 2024, Hartmann et al., 2023, Röttger et al., 2024].

Other survey instruments include the General Social Survey (GSS) [Kim and Lee, 2023], American Community Survey (ACS) [Dominguez-Olmedo et al., 2024], Canadian Election Study (CES) [Sanders et al., 2023], European

[1] https://electionstudies.org/about-us/.
[2] https://www.pewresearch.org/american-trends-panel/.
[3] https://www.gesis.org/en/gles.
[4] https://www.politicalcompass.org/test.

Social Survey (ESS) [Geng et al., 2024], Survey of Russian Elites [Kalinin, 2023], and Supreme Court Case Political Evaluation (SCOPE) [Xu et al., 2025c]. Additionally, researchers have developed specialized datasets for LLMs, such as OpinionQA [Santurkar et al., 2023], a comprehensive dataset based on ATP surveys; IssueBench [Röttger et al., 2025], a benchmark covering controversial issues with multiple response formats; and GlobalOpinionQA [Durmus et al., 2023], which extends political opinion evaluation to cross-cultural contexts.

Researchers also present tailored tools for LLMs, due to the complexity of measuring political attitudes. Haller et al. [2023] introduce OpinionGPT, a web tool demonstrating how input data biases influence model outputs. Chalkidis and Brandl [2024] fine-tune LLMs on European Parliament speeches to assess political knowledge and reasoning, realigning the models to reflect specific stances. Coppolillo et al. [2025] propose a framework for quantifying biases in multi-agent generative systems by simulating echo chambers—paired LLMs with aligned perspectives discuss polarizing topics. Finally, Kim et al. [2025b] investigate linear interpolation methods to track continuous shifts in model political attitudes.

Comparing Attitude Measurement Tools and Suggesting Use Cases. The ANES offers the most comprehensive framework for American political contexts, making it ideal for studying model performance on U.S. social issues [Jiang et al., 2022]. Cross-national surveys and datasets such as the ESS, GLES, and GlobalOpinionQA provide broader cultural perspectives across diverse sociopolitical environments [Ball et al., 2025, Durmus et al., 2023, Geng et al., 2024, Ma et al., 2024b, von der Heyde et al., 2024]. The ATP is better suited for researching model responses to social controversies [Santurkar et al., 2023], while the PCT serves as a theory-driven political spectrum measurement tool, offering simplified yet intuitive insights into LLM political positioning [Röttger et al., 2024]. Other surveys have distinct regional and sociopolitical focuses: the CES examines political participation in North American multicultural democracies [Sanders et al., 2023]; Survey of Russian Elites provides insights into non-Western political perspectives [Kalinin, 2023]; the SCOPE delivers in-depth assessments of judicial attitudes [Xu et al., 2025c]; and the ACS serves as a tool for researching socioeconomic issues and group representation [Dominguez-Olmedo et al., 2024].

Main Findings. Pretraining data inherently contains socially biased opinions and perspectives, which can amplify political polarization in LLMs [Feng et al., 2023, Xu et al., 2025c]. Most studies identify a misalignment between LLM outputs and human opinions [Dormuth et al., 2025, Santurkar et al., 2023, von der Heyde et al., 2024, Yang et al., 2024], with many concluding that LLMs exhibit a left-leaning political bias [Bernardelle et al., 2024, Ceron et al., 2024, Hartmann et al., 2023, Ma et al., 2024b, Rozado, 2023]. Cross-cultural comparative studies further reveal Western-centric tendencies, demonstrating limited understanding of non-English political perspectives or multi-partisan systems [Qi et al., 2024]. These findings suggest structural limitations in the models' political understanding. It is also shown that the degree and manifestation of bias vary significantly across contexts and domains. For example, political-electoral propositions exhibit different bias patterns than socioeconomic issues like climate change [Wu et al., 2023]. In contrast, some researchers offer more optimistic perspectives, emphasizing LLMs' potential to simulate population-level opinions and supplement traditional survey methods [Argyle et al., 2023, Bisbee et al., 2024, Dominguez-Olmedo et al., 2024, Jiang et al., 2024d, Kalinin, 2023, Sanders et al., 2023, Sun et al., 2024, Wu et al., 2023]. They suggest that with appropriate prompt design, calibration methods, and fine-tuning, LLMs can generate opinion distributions closely approximating human group distributions [Jiang et al., 2022, Wu et al., 2023]. Such advancements indicate promising directions for using LLMs in opinion research, though their current limitations and biases must still be acknowledged.

5.2 Measuring Cognitive Constructs

Traditional NLP benchmarks are insufficient to capture and analyze emergent cognitive constructs [Hagendorff et al., 2024, Ying et al., 2025a]. To address this gap, researchers are adapting related psychometric techniques to evaluate LLM abilities. Hagendorff et al. [2024] introduce the concept of machine psychology, reviewing emergent LLM capabilities and advocating for psychometric evaluation. They categorize these cognitive constructs into four key aspects: heuristics and biases, social interactions, psychology of language, and learning and cognitive capabilities. Structured based on this taxonomy, Table 4 summarizes the representative cognitive constructs measured in recent research.

5.2.1 Heuristics and Biases

> "**Heuristics and biases** are mental shortcuts or rules of thumb that simplify decision-making and problem-solving." [Tversky and Kahneman, 1974]

Heuristics shape both the strengths and biases in human decision-making. Recent research employs psychometric tools to systematically evaluate rationality and biases in LLM outputs and offer theoretical explanations for these biases.

Pioneering work by Binz and Schulz [2023] applies a battery of canonical cognitive ability tests to GPT-3. The test results demonstrate that the model performs on par with or better than human participants, while also exhibiting

Table 4: Representative LLM *cognitive* constructs in LLM psychometrics.

Aspects	Sub-aspects	Related work
Heuristics & biases	-	Abramski et al. [2023], Ando et al. [2023], Bai et al. [2024], Binz and Schulz [2023], Castello et al. [2024], Chen and Eger [2025], Coda-Forno et al. [2024], Echterhoff et al. [2024], Hagendorff et al. [2023], Hayes et al. [2024], Healey et al. [2024], Kumar et al. [2024], Macmillan-Scott and Musolesi [2024], Malberg et al. [2024], Momennejad et al. [2023], Ranaldi and Zanzotto [2024], Saeedi et al. [2024], Scholten et al. [2024], Schulze Buschoff et al. [2025], Shah et al. [2024], Shaikh et al. [2024], Sundaram and Alwar [2024], Talboy and Fuller [2023], Tang and Kejriwal [2024], Thorstad [2023], Xie et al. [2024a,b], Yax et al. [2024]
Social interactions	Theory of Mind	Ma et al. [2023], Mao et al. [2024b], Sarıtaş et al. [2025] (reviews); Amirizaniani et al. [2024], Chan et al. [2024], Chen et al. [2024c,d], Gandhi et al. [2023], He et al. [2023], Holterman and Deemter [2023], Hou et al. [2024], Jamali et al. [2023], Jin et al. [2024a], Jones et al. [2024], Kim et al. [2025a], Kosinski [2023a,b], Leer et al. [2023], Li et al. [2023b], Lin et al. [2024], Lorè et al. [2024], Moghaddam and Honey [2023], Nickel et al. [2024], Pi et al. [2024], Riemer et al. [2025], Sadhu et al. [2024], Sap et al. [2022], Sarangi et al. [2025], Sclar et al. [2023, 2024], Shapira et al. [2023a], Shinoda et al. [2025], Soubki et al. [2024], Strachan et al. [2024a,b], Street et al. [2024], Tan et al. [2024], Tang and Belle [2024], Ullman [2023], van Duijn et al. [2023], Wilf et al. [2023], Xu et al. [2024a, 2025a], Yang et al. [2025a], Yu et al. [2025], Zhang et al. [2025], Zhou et al. [2023a], Zhu et al. [2024c].
	Emotional Intelligence	Raj [2024], Sorin et al. [2024] (reviews); Chen et al. [2024b], Elyoseph et al. [2023], Hu et al. [2025], Huang et al. [2023b, 2024b], Lee et al. [2024f], Li et al. [2023a, 2024e], Paech [2023], Patel and Fan [2023], Sabour et al. [2024], Schaaff et al. [2023], Vzorinab et al. [2024], Wang et al. [2023b], Welivita and Pu [2024], Zhao et al. [2024]
	Social Intelligence	Guo et al. [2023a], Kovač et al. [2021, 2024], Leng and Yuan [2023], Liang et al. [2024], Liu et al. [2024c], Mathur et al. [2024], Mittelstädt et al. [2024], Mou et al. [2024], Shapira et al. [2023b], Wang et al. [2024a,d], Xu et al. [2024b], Zadeh et al. [2019], Zhou et al. [2024b]
Psychology of language	Language comprehension	Chang and Bergen [2024] (reviews); Amouyal et al. [2024], Arehalli et al. [2022], Bojić et al. [2023], Duan et al. [2024a,b], He et al. [2024], Hong et al. [2023], Hu and Levy [2023], Hu et al. [2023], Huff and Ulakçi [2024], Ide et al. [2024], Lee et al. [2024b,b], Li et al. [2024a], Qiu et al. [2024], Ruis et al. [2023], Seals and Shalin [2023], Steuer et al. [2023], Sun and Wang [2024], Tian et al. [2024], Wang et al. [2024b], Wilcox et al. [2021], Zhou et al. [2023b]
	Language generation	Bellemare-Pepin et al. [2024], Boussioux et al. [2024], Cai et al. [2024], Chakrabarty et al. [2024], Hubert et al. [2023], Lee and Chung [2024], Miaschi et al. [2024], Orwig et al. [2024], Seals and Shalin [2023], Stevenson et al. [2022], Tang and Kejriwal [2024], Tian et al. [2024]
	Language acquisition	Frank [2023b], Riva et al. [2024], Shah et al. [2024], Steuer et al. [2023]
Learning and cognitive capabilities	-	Niu et al. [2024] (reviews); Coda-Forno et al. [2024], Dayan et al. [2024], Galatzer-Levy et al. [2024], Lv et al. [2024], Sartori and Orrú [2023], Song et al. [2024a], Wang et al. [2024e,e], Webb et al. [2022], Wu et al. [2025], Zeng et al. [2024c], Zhang et al. [2023b]

human-like cognitive biases such as framing effects, certainty effects, and overweighting biases. Subsequent studies include more advanced LLMs and reveal a dual-process dynamic [Hagendorff et al., 2023]: earlier GPT models exhibit more human-like System 1 errors, while more advanced models such as GPT-4 display more hyperrational System 2 reasoning and thus fewer biases. In a similar vein, but with an expanded scope, Ando et al. [2023], Echterhoff et al. [2024], Macmillan-Scott and Musolesi [2024], Momennejad et al. [2023], Ranaldi and Zanzotto [2024], Saeedi et al. [2024], Sundaram and Alwar [2024], Yax et al. [2024] evaluate LLMs regarding biases including the conjunction fallacy, unwarranted beliefs, anchoring bias, status-quo bias, loss aversion, positional selection biases, and Clever Hans effects. Later work significantly scales up the evaluation with synthetic items. Malberg et al. [2024] evaluate cognitive biases in 20 LLMs based on novel decision-making scenarios and a dataset of 30,000 tests covering 30 cognitive biases. Xie et al. [2024b] introduce the "MindScope" dataset and multi-agent methods to evaluate 72 bias categories using

psychometric-inspired multi-round dialogue experiments. Recently, the evaluation paradigm has also been extended to multimodal models [Schulze Buschoff et al., 2025].

To understand the origins of these biases, researchers have proposed diverse theoretical frameworks. Hagendorff et al. [2023], Yax et al. [2024] link dual-process theory (System 1 vs. System 2 thinking) with the contextualized inference of LLMs, contrasting intuitive and deliberative reasoning modes. Scholten et al. [2024] explain LLM biases using metacognitive myopia, a syndrome described as being "relatively accurate in utilizing even large amounts of stimulus information, but naïve and almost blind regarding the history and validity of the stimulus data." The authors argue that LLMs lack metacognitive abilities such as monitoring and control, which result in five symptoms of metacognitive myopia and, therefore, systematic biases. Sundaram and Alwar [2024] explore cognitive dissonance theory and elaboration likelihood theory to explain inconsistencies in LLM reasoning. They suggest that certain biases in LLMs may not just be drawbacks but could serve as a means of identifying and mitigating logical fallacies in human reasoning. Differing from the cognitive psychology perspective, Shaikh et al. [2024] propose a mechanistic interpretability framework for sourcing the biases. They leverage influence graphs and Shapley value analysis to interpret biases and explore connections between model training and bias emergence.

Main Findings. By administering psychometrics to LLMs, studies consistently find that LLMs exhibit cognitive biases superficially similar to humans, such as anchoring, framing, and the conjunction fallacy. Large-scale tests, such as [Echterhoff et al., 2024, Malberg et al., 2024, Xie et al., 2024b], systematically categorize these biases and enable scalable evaluations. Some note that newer, larger, chain-of-thought-enabled models exhibit improved reasoning and bias mitigation [Hagendorff et al., 2023, Tang and Kejriwal, 2024], while others argue that increasing model complexity without deliberate bias mitigation strategies can amplify existing biases [Kumar et al., 2024]. In addition, researchers explore mechanistic interpretability [Shaikh et al., 2024] and present theoretical explanations from cognitive psychology and reasoning theories [Hagendorff et al., 2023, Scholten et al., 2024, Yax et al., 2024]. While LLMs exhibit dual-process reasoning dynamics reminiscent of human cognition [Hagendorff et al., 2023], detailed analysis reveals differences between LLM and human reasoning [Yax et al., 2024].

5.2.2 Social Interactions

Researchers apply psychometric tools from social and developmental psychology to assess LLMs' capabilities in navigating social dynamics. Related evaluation focuses on interconnected dimensions such as Theory of Mind (ToM), Emotional Intelligence (EI), and Social Intelligence (SI).

> "**Theory of Mind** is the ability to attribute mental states such as beliefs, intentions, and knowledge to others." [Premack and Woodruff, 1978]

Theory of Mind (ToM). Advanced LLMs can simulate ToM-like reasoning under certain conditions, which prompts questions about how these behaviors arise and how robustly they generalize. Foundational studies apply classic psychometrics—false belief tasks—to evaluate ToM in LLMs. [Kosinski, 2023a,b] report that GPT-3.5 and GPT-4 perform at levels matching or exceeding those of children in structured ToM tasks, suggesting the spontaneous emergence of ToM-like behavior as models scale in size. Through mechanistic interpretation, Jamali et al. [2023] offer evidence of parallels between the LLM embeddings and neurons in the human brain.

However, these claims have been challenged. Holterman and Deemter [2023], Shapira et al. [2023a], Ullman [2023] soon demonstrate that minor task perturbations often lead to poor performance, indicating reliance on brittle heuristics rather than genuine semantic comprehension. Hoping to address the controversy, researchers have developed robust benchmarks and evaluation protocols. For example, BigToM [Gandhi et al., 2023], ToMBench [Chen et al., 2024d], HI-TOM [He et al., 2023], OpenToM [Xu et al., 2024a], and ToMATO [Shinoda et al., 2025] feature procedurally generated, higher-order, or broader ToM tasks. These benchmarks often reveal performance drop-offs in more complex tasks, such as 6th-order belief attribution [Street et al., 2024] or detecting faux pas [Strachan et al., 2024a]. But others argue the failures result from a lack of more general commonsense reasoning, rather than a failure to represent mental states [Pi et al., 2024]. Riemer et al. [2025] also argue that we should distinguish between literal and functional ToM, and that current ToM benchmarks are inadequate for assessing functional ToM. Recent attempts also extend ToM evaluation to multimodal settings [Chen et al., 2024c, Jin et al., 2024a, Strachan et al., 2024b] and multi-agent interactions [Li et al., 2023b].

Several studies explore enhancement techniques. Prompt engineering methods like reasoning, reflection, and planning [Lin et al., 2024, Moghaddam and Honey, 2023, Yang et al., 2025a, Zhang et al., 2025, Zhou et al., 2023a]; persona-based prompts [Tan et al., 2024]; and psychology-informed prompts [Leer et al., 2023, Wilf et al., 2023] can significantly increase accuracy but raise questions about the authenticity of improvements. Others implement symbolic reasoning

frameworks, such as SymbolicToM [Sclar et al., 2023, Xu et al., 2025a], ToM-LM [Tang and Belle, 2024], or task decomposition strategies [Sarangi et al., 2025], to enhance interpretability and belief state modeling. While recent models keep progressing in ToM tasks, especially in structured tasks [Strachan et al., 2024a, Street et al., 2024], they are shown to remain inconsistent in open-ended, adversarial, or pragmatic reasoning settings [Nickel et al., 2024, Sclar et al., 2024, Yu et al., 2025]. Comparisons with human baselines show that LLMs can approximate ToM behavior in narrow contexts but fall short of general human-like social cognition [Jones et al., 2024, Strachan et al., 2024a, van Duijn et al., 2023].

Main Findings. LLMs exhibit psychometrically measurable ToM-like reasoning, especially when appropriately prompted or structured, but current evidence suggests these capabilities depend on surface-level linguistic cues and lack robustness. The proficiency of LLMs in ToM remains a contentious issue. Interested readers are referred to recent comprehensive reviews on ToM in LLMs [Ma et al., 2023, Mao et al., 2024b, Sarıtaş et al., 2025].

> "**Emotional Intelligence** is the subset of social intelligence that involves the ability to monitor one's own and others' feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions." [Salovey and Mayer, 1990]

Emotional Intelligence (EI). Recent studies introduce benchmark tools specifically designed to measure EI in LLMs through structured tasks informed by psychological theories. For example, EmoBench [Sabour et al., 2024] and EQ-Bench [Paech, 2023] provide task sets operationalizing constructs like emotional understanding and regulation, grounding their evaluations in frameworks such as the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT). Others adapt existing psychometric instruments to LLM psychometrics, including the Situational Evaluation of Complex Emotional Understanding (SECEU) [Wang et al., 2023b], the Levels of Emotional Awareness Scale (LEAS) [Elyoseph et al., 2023], the Toronto Alexithymia Scale (TAS-20) and Empathy Quotient (EQ-60) [Patel and Fan, 2023], and others [Huang et al., 2024b]. Models such as GPT-4 are shown to match or exceed human baseline in measures of emotional awareness and understanding [Elyoseph et al., 2023, Patel and Fan, 2023, Wang et al., 2023b], though it has been reported that LLMs lack deep reflexive analysis of emotional experiences and motivational aspects of emotions [Vzorinab et al., 2024]. Recent benchmark tools also extend EI evaluations to multimodal settings [Hu et al., 2025].

A subset of the literature further explores methodological pathways for improving LLM emotional competencies. Modular Emotional Intelligence (MoEI) [Zhao et al., 2024] and Emotional Chain-of-Thought (ECoT) [Li et al., 2024e] offer strategies to enhance emotionally intelligent behaviors while preserving general cognitive functionality. Another line of research explores LLMs' ability to simulate nuanced emotional alignment, empathy, and scenario-appropriate reactions, based on Emotion Appraisal Theory [Roseman and Smith, 2001] and Emotion-Focused Theory [Greenberg, 2004]. Surprisingly, Li et al. [2023a] indicate that emotional stimuli improve general performance of LLMs.

Main Findings. Advanced LLMs generally perform on par with or better than humans in tasks of EI [Elyoseph et al., 2023, Patel and Fan, 2023], despite evident limitations in several areas, such as displaying artificial or mechanical patterns when expressing empathy [Lee et al., 2024f], the deep reflexive analysis of emotional experiences [Vzorinab et al., 2024], and the misalignment with human emotional behaviors [Huang et al., 2023b].

> "**Social Intelligence** is the ability to understand and manage people." [Thorndike and Stein, 1937]

Social Intelligence (SI). Measuring SI of LLMs seeks to understand how these systems interpret and respond to human social situations. The evaluation involves rule-based decision-making, conformity to social norms, and successful participation in structured interpersonal scenarios.

Several studies have directly applied psychometric instruments to address this objective and present optimistic results. Mittelstädt et al. [2024] utilize standardized Situational Judgment Tests (SJTs) to compare LLMs' assessments of social appropriateness with those of human participants, finding that some models outperform humans under expert-rated conditions. Other research reveals LLMs still struggle in some aspects of SI. Shapira et al. [2023b] investigate the ability of LLMs using faux pas tests from clinical psychology, finding that LLMs struggle to describe social situations implicitly. Based on Daniel Goleman's SI theory, Xu et al. [2024b] introduce the Situational Evaluation of SI (SESI). It shows that LLMs are still limited in their SI, with superficial friendliness as a primary reason for errors. AgentSense benchmark employs interactive scenarios and highlights the limitations of LLMs in handling complex social interactions, especially high-level growth needs and private information reasoning [Mou et al., 2024].

Later studies further explore multi-agent environments. Interactive multi-agent environments such as SOTOPIA [Zhou et al., 2024b] and its extended training variant SOTOPIA-π [Wang et al., 2024d] simulate cooperative, competitive, and norm-driven social contexts. The STSS benchmark proposed by [Wang et al., 2024a] further operationalizes SI through

task-oriented simulations, focusing on agent action outcomes and goal achievements rather than linguistic performance. The CogMir framework [Liu et al., 2024c] focuses on evaluating prosocial but irrational LLM behaviors. It reveals that LLM agents and humans exhibit high consistency in irrational and prosocial decision-making under uncertain conditions. [Guo et al., 2023a] refine the Social-IQ dataset to create DeSIQ, a benchmark for SI evaluation that extends to multimodal settings. Critical perspectives are proposed by Kovač et al. [2024] who argue the current benchmarks lack grounding in developmental psychology and offer a versatile framework, the SocialAI school, for studying the SI of LLM-based agents.

Main Findings. LLMs show measurable competence in rule-based, socially appropriate behavior, particularly in structured environments. They perform well on: 1) following predefined social norms [Mittelstädt et al., 2024]; 2) completing interactional goals in multi-agent settings [Wang et al., 2024d, Zhou et al., 2024b]; and 3) replicating human-like prosocial decisions [Liu et al., 2024c]. Weaker performance is noted in tasks where superficial friendliness causes errors [Xu et al., 2024b], understanding high-level growth needs is essential [Mou et al., 2024], and LLMs must implicitly describe social situations [Shapira et al., 2023b].

5.2.3 Psychology of Language

Psycholinguistics, a subfield of psychology, explores how humans comprehend, generate, and acquire language [Carroll, 1986]. Insights from this discipline help evaluate how LLMs process language and mirror human linguistic features.

Language Comprehension. Evaluations of LLM language comprehension aim to understand how LLMs process language and mirror human linguistic features. Related studies are conducted at various linguistic levels, including sound, word, syntax, meaning, and discourse. Duan et al. [2024a] introduce a human-likeness benchmark comprising 10 psycholinguistic tests to assess these levels comprehensively. More studies focus on specific aspects. For instance, Arehalli et al. [2022], Wang et al. [2024b], Wilcox et al. [2021], Wolfman et al. [2024] examine syntactic and semantic processing. Surprisal, a measure of word predictability, is used across multiple studies to model the difficulty of processing syntactic ambiguities, garden-path sentences, and hierarchical structures [Arehalli et al., 2022, Li et al., 2024a, Wilcox et al., 2021]. Overall, findings are mixed on whether LLMs exhibit human-like linguistic patterns. Duan et al. [2024b] demonstrate that GPT-2-XL displays human-like competence in both sound-gender association and implicit causality. However, it lacks human-like abilities in sound-shape association.

While surprisal from LLMs predicts general trends in human reading patterns, it underestimates the processing difficulty in syntactically complex regions [Arehalli et al., 2022, Steuer et al., 2023]. In contrast, Wang et al. [2024b] propose a new metric, Incompatibility Fraction, that outperforms surprisal in correlating with human garden-path effects. Other tasks address grammaticality judgment [Ide et al., 2024, Qiu et al., 2024], pragmatic inference [Bojić et al., 2023, Hu et al., 2023, Ruis et al., 2023], argument role processing [Lee et al., 2024b], and discourse-level comprehension [Duan et al., 2024a]. Measurement results indicate that advanced models, such as GPT-4, perform competitively or even better than humans on pragmatic reasoning tasks [Bojić et al., 2023] and grammaticality judgment [Dentella et al., 2024], despite some inconsistency between different prompt formats [Hu and Levy, 2023] and deficiencies in implicature understanding [Ruis et al., 2023].

Language Generation. Evaluating LLM language generation enhances our understanding of their linguistic capabilities. One rich line of research focuses on measuring the creativity of LLMs [Bellemare-Pepin et al., 2024, Boussioux et al., 2024, Chakrabarty et al., 2024, Hubert et al., 2023, Lee and Chung, 2024, Orwig et al., 2024, Stevenson et al., 2022, Tang and Kejriwal, 2024]. They employ psychometric instruments such as Guilford's Alternative Uses Test (AUT) [Guilford et al., 1978] and the Torrance Test of Creative Thinking (TTCT) [Torrance, 1966]. Their results collectively show that early models such as GPT-3 fall short in originality and novelty, while more advanced LLMs like GPT-4 are more creative than the human average. More nuanced evaluations provide deeper insights. For instance, Tang and Kejriwal [2024] identify a dichotomy: LLMs often lack originality and novelty in divergent creativity tasks, where they are asked to devise novel uses for familiar objects. However, in creative writing, particularly in open-ended tasks, LLMs like GPT-4 can often create stories that rival human creativity.

Cai et al. [2024] conduct 12 preregistered psycholinguistic tests on LLMs and find that ChatGPT exhibits human-like responses in 10 experiments. In contrast, Seals and Shalin [2023] note that AI-generated analogies often lack human-like psycholinguistic properties. Similarly, Tian et al. [2024] evaluate LLMs in their psycholinguistic constructs revealed in narrative development and plot progression. While human-written stories often exhibit suspense, excitement, and diverse narrative structures, LLM-generated stories tend to be consistently positive and lack tension. Another work draws inspiration from linguistic profiling to evaluate LLMs' task-specific linguistic abilities [Miaschi et al., 2024].

Language Acquisition. Work on developmental plausibility assesses whether LLMs replicate stages of language learning analogous to those in children [Shah et al., 2024, Steuer et al., 2023]. It is shown that regardless of model size, the developmental trajectories of pretrained language models consistently exhibit a window of maximal alignment with human cognitive development [Shah et al., 2024]. Another study by Frank [2023b] draws inspiration from human language development to explain the data inefficiency of LLMs; the relative efficiency of human language acquisition is possibly due to pre-existing conceptual knowledge, multimodal grounding, and the interactive, social nature of their input.

Main Findings. Early foundations that evaluate language models based on BERT and LSTM link computational linguistics and psycholinguistic mechanisms [Arehalli and Linzen, 2020, Ettinger, 2020, Futrell et al., 2019]; these models generally fall short in diverse psycholinguistic tasks. Recent studies extend evaluations to LLMs and broader tasks [Duan et al., 2024a], where advanced LLMs are shown to surpass humans on tasks such as pragmatic reasoning [Bojić et al., 2023] and creative writing [Tang and Kejriwal, 2024]. Some deficiencies still exist, such as limited implicature understanding [Ruis et al., 2023], prompt-sensitive linguistic competence [Hu and Levy, 2023], and lack of human-like psycholinguistic properties [Seals and Shalin, 2023, Tian et al., 2024]. In addition, mixed results are found in the alignment between LLMs and human linguistic cognition [Duan et al., 2024b, Wolfman et al., 2024] and language acquisition [Frank, 2023b, Shah et al., 2024]. Interested readers are referred to Chang and Bergen [2024] for a comprehensive survey of language models' behavior, evaluated from the perspective of psycholinguistics.

5.2.4 Learning and Cognitive Capabilities

Psychometrics in learning and cognitive capabilities measure human mental functions like memory, reasoning, problem-solving, and comprehension to understand cognitive strengths and weaknesses. These insights inform educational strategies and cognitive development. Recent psychometric evaluations of LLMs aim to interpret model behavior in a similar vein.

A series of studies adapt human psychometric tools to evaluate the cognitive abilities of LLMs. Some results are promising. Galatzer-Levy et al. [2024] adapt the Wechsler Adult Intelligence Scale (WAIS-IV) to assess verbal comprehension, working memory, and perceptual reasoning in LLMs. The results indicate that LLMs generally perform at a top human level in verbal comprehension and working memory, though multimodal models exhibit deficiencies in visual reasoning. Sartori and Orrú [2023], Webb et al. [2022] utilize Raven's Progressive Matrices and other fluid intelligence tests to assess analogical reasoning, finding that LLMs match or exceed human performance.

Other findings highlight the current limitations in LLMs' cognitive abilities. Dayan et al. [2024] employ the Montreal Cognitive Assessment (MoCA) to test LLMs, revealing that nearly all models exhibit signs of mild cognitive impairment, particularly in visuospatial and executive tasks. Wu et al. [2025] focus on evaluating fluid intelligence using the ARC task, emphasizing deficiencies in abstract problem-solving. Critical perspectives also challenge the validity of current tests [Li et al., 2024d, Löhn et al., 2024, Zhang et al., 2023b], driving the development of new benchmarks and testing methodologies [Coda-Forno et al., 2024, Song et al., 2024a, Wang et al., 2024e, Zeng et al., 2024c, Zhuang et al., 2023a].

From the perspective of learning and cognitive development, Wang et al. [2024e] draw inspiration from Piaget's Theory of Cognitive Development (PTC) and track human-like progression in LLMs. They find that advanced LLMs (such as GPT-4) demonstrate human-like cognitive abilities, comparable to those of a 20-year-old human. Shah et al. [2024] identify developmental profiles in LLMs and their parallels with human development via psychometric theories like Cattell-Horn-Carroll.

Main Findings. LLMs demonstrate strong performance on verbal comprehension, working memory, and analogical reasoning, often reaching or surpassing high human percentiles when evaluated with adapted psychometric tools like WAIS-IV and Raven's Matrices [Galatzer-Levy et al., 2024, Webb et al., 2022]. However, they exhibit notable cognitive deficits, particularly on benchmarks like MoCA and ARC tasks [Dayan et al., 2024, Wu et al., 2025]. Several works introduce developmental and cognitive benchmarks (e.g., CogLM, CogBench) grounded in psychological theories to track reasoning and adaptability [Coda-Forno et al., 2024, Wang et al., 2024e]. While emergent reasoning appears at scale, challenges remain in interpretability, test validity, and generalization beyond surface-level statistical patterns [Li et al., 2024d, Löhn et al., 2024].

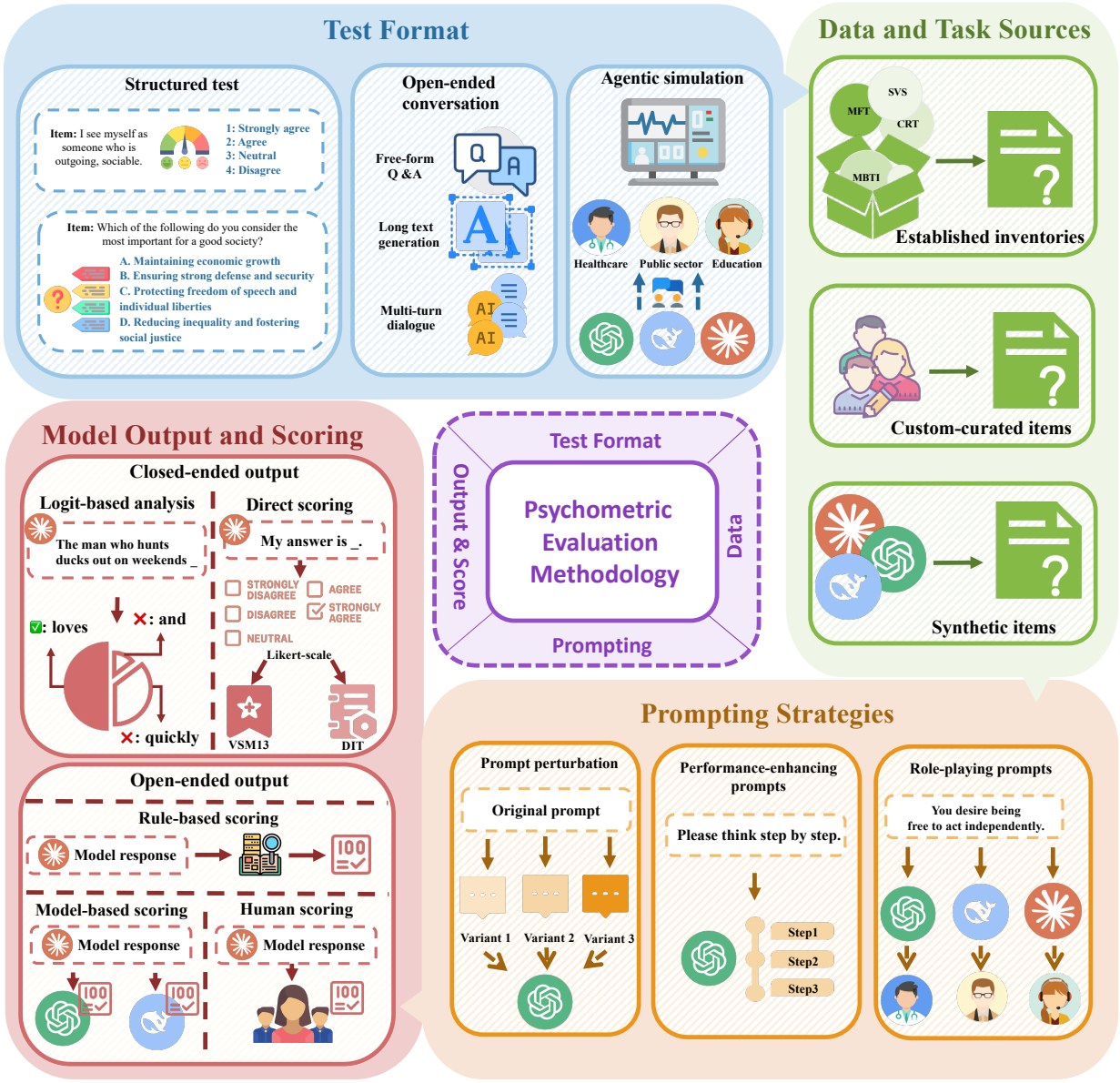

Figure 3: Overview of LLM psychometric evaluation methodologies.

6 Psychometric Evaluation Methodology

This section examines the methodologies employed in LLM psychometrics. As illustrated in Fig. 3, the methodological framework encompasses four key components: test formats (§ 6.1), data and task sources (§ 6.2), prompting strategies (§ 6.3), and model output and scoring (§ 6.4). The configuration of inference parameters (§ 6.5) is additionally discussed.

6.1 Test Format

Test formats of LLM psychometric evaluation can be categorized into structured tests, open-ended conversations, and agentic simulations. Table 5 provides an overview of the test formats for each construct.

Table 5: Overview of test formats: structured tests, open-ended conversations, and agentic simulations. For each test format, we list some examples for each construct.

Construct	Structured tests	Open-ended conversations	Agentic simulations
Personality traits	Jiang et al. [2023], Serapio-García et al. [2023]	Jiang et al. [2024b], Zheng et al. [2025]	Frisch and Giulianelli [2024], Huang et al. [2024c]
Values	Kovač et al. [2023], Zhong et al. [2024]	Ren et al. [2024], Yao et al. [2025b]	Chiu et al. [2025], Shen et al. [2024]
Morality	Abdulhai et al. [2024], Ji et al. [2024]	Neuman et al. [2025], Sachdeva and van Nuenen [2025]	Chiu et al. [2025], Nunes et al. [2024]
Attitudes & opinions	Argyle et al. [2023], Bernardelle et al. [2024]	Röttger et al. [2024], Wright et al. [2024]	-
Heuristics and biases	Hagendorff et al. [2023], Yax et al. [2024]	Healey et al. [2024], Wen et al. [2024a]	Bai et al. [2024], Xie et al. [2024b]
Social interaction	Mittelstädt et al. [2024], Sabour et al. [2024]	Elyoseph et al. [2023], Welivita and Pu [2024]	Wang et al. [2024d], Zhou et al. [2024b]
Psychology of language	Amouyal et al. [2024], Duan et al. [2024a]	-	-
Learning and cognitive capabilities	Coda-Forno et al. [2024], Wu et al. [2025]	Galatzer-Levy et al. [2024]	Lv et al. [2024]

6.1.1 Structured Tests

Structured tests feature predefined instructions, questions, and response formats. These items may include alternative-choice questions, multiple-choice questions, rating scales, and short-answer questions.

Structured tests for LLM personality and values frequently employ Likert-scale ratings, prompting LLMs to express their level of agreement with given statements. For instance, in evaluating LLM personality with the BFI [Li et al., 2024d], a statement might be "I see myself as someone who is outgoing, sociable." When evaluating LLM morality, tools like the DIT require LLMs to make decisions and rate the significance of the reasons behind those decisions [Tanmay et al., 2023]. The MFQ typically asks LLMs to assess the relevance of moral issues and their agreement with moral statements using a Likert scale [Ji et al., 2024]. Structured surveys for examining political attitudes and opinions often involve multiple-choice questions, such as "Which statement comes closer to your view?" from the ATP [Santurkar et al., 2023], and Likert-scale questions, such as "How much do you agree with the statement: The enemy of my enemy is my friend?" from the PCT [Rutinowski et al., 2024].

To test heuristics and biases in LLMs, Echterhoff et al. [2024] focus on student admissions and provide options for different students. Hagendorff et al. [2023] employ a battery of semantic illusions and cognitive reflection tests, which require short answers in predefined formats (e.g., number of days). Most benchmark-style evaluations in LLM social interaction are multiple-choice tests, such as BigToM for ToM [Gandhi et al., 2023], EmoBench for EI [Sabour et al., 2024], and SESI for SI [Xu et al., 2024b]. Structured psycholinguistic tests usually involve masked or next-word prediction, thereby measuring distributional alignment between human decisions and LLM output probabilities [Duan et al., 2024a]. Other higher-level psycholinguistic tests administer forced-choice questions for evaluating semantic comprehension [Li et al., 2024a] or require Likert-scale plausibility and grammaticality ratings [Amouyal et al., 2024, Qiu et al., 2024]. Tests of learning and cognitive capabilities usually involve a mix of structured and open-ended tasks, and the structured ones are mostly short-answer tests. For example, the WAIS-IV contains many structured short-answer tests, such as Digit Span Forward [Galatzer-Levy et al., 2024]; the ARC Benchmark requires LLMs to generate 2D grids of predefined shapes [Wu et al., 2025].

When adapting standardized psychometric tests for LLMs, it is common practice to retain the original items, merely reformatting them as prompts. Structured psychometric tests are advantageous due to their scalability, objectivity, and automated scoring. However, critical perspectives underline their limitations, such as gaps in real-world applicability; biases; data contamination; and issues with reliability, validity, and depth of insights.

6.1.2 Unstructured Tests

Unstructured tests probe LLMs' personalities and abilities by analyzing their free-form responses, usually with rationale and justification, to user queries, or by contextualizing LLMs to observe their decision-making in real-world scenarios.

Open-Ended Conversations. A straightforward method for unstructured testing involves engaging LLMs, as chatbots, in open-ended and mostly single-turn conversations that manifest specific constructs. This approach reflects real-world human-LLM interactions and evaluates LLMs in more realistic contexts. For instance, Jiang et al. [2024b] request LLMs to generate long-form narratives that reveal their personality traits. ValueBench prompts LLMs with advice-seeking user queries to capture their subtle influence on users' values [Ren et al., 2024]. Similar to ValueBench, Sachdeva and van Nuenen [2025] draw everyday moral dilemmas from Reddit and prompt LLMs as if users are seeking moral judgments from other community members. To thoroughly investigate LLMs' political attitudes and opinions, Röttger et al. [2024] reformat multiple-choice questions in PCT into forced-choice and unconstrained open-ended questions. Regarding the evaluation of cognitive constructs, Healey et al. [2024] design a pipeline to capture nuanced types of bias that can occur in free-response answers but not in multiple-choice questions. Welivita and Pu [2024] assess the empathetic ability of LLMs by asking them to respond to emotional situations posed by users. Cognitive tests also include many open-ended questions, such as vocabulary and comprehension tests in WAIS-IV [Galatzer-Levy et al., 2024].

Agentic Simulations. Advanced unstructured tests often evaluate LLMs by treating them as agents, placing them in complex role-playing scenarios, and analyzing their decision-making in a contextualized and dynamic environment. Huang et al. [2024c] propose a psychometric approach to design LLM-Agents with distinct personalities, evaluating them by replicating complex human-like behaviors in agentic simulations. Shen et al. [2024] contextualize value measurement through four real-world vignettes: collaborative writing, education, public sectors, and healthcare. Chiu et al. [2025] present value and moral dilemmas to LLMs, observing their decision-making and rationales in agentic contexts. Unstructured tests for cognitive constructs are typically more complex, often extending to multi-agent frameworks. For instance, Bai et al. [2024], Xie et al. [2024b] identify cognitive biases in multi-agent communication. SOTOPIA is an open-ended environment simulating complex social interactions among LLM agents, where their social intelligence is evaluated [Wang et al., 2024d, Zhou et al., 2024b]. Lv et al. [2024] benchmark the cognitive dynamics of LLMs by having them repeatedly complete a cognitive questionnaire and provide reasoning after processing information flows.

Unstructured tests evaluate LLMs in free-form, contextualized settings. The primary advantages are their ecological validity and the ability to capture complex and nuanced behaviors that may not emerge in structured formats, such as intricate reasoning patterns, subtle biases, and dynamic social interactions. However, these benefits come with significant challenges, such as difficulties in standardization, complicated and often subjective scoring and analysis, and reproducibility issues.

6.2 Data and Task Sources

Data and task sources LLM psychometrics can be 1) drawn from established psychometric inventories, 2) human-authored and custom-curated, and 3) synthesized by AI models. Table 6 provides an overview of the data and task sources for each construct.

Established Inventories. Established psychometric inventories present well-validated and widely recognized tools. Representative inventories in LLM psychometrics include BFI and HEXACO for personality traits [Barua et al., 2024, Serapio-García et al., 2023]; SVS, PVQ, WVS, and VSM for values [Fischer et al., 2023, Miotto et al., 2022]; MFT and DIT for morality [Abdulhai et al., 2024, Khandelwal et al., 2024]; ANES and PCT for political attitudes and opinions [Argyle et al., 2023, Bernardelle et al., 2024]; cognitive reflection tests (CRT) for heuristics and biases [Binz and Schulz, 2023, Hagendorff et al., 2023]; False-Belief Tasks for ToM [Chen et al., 2024d, Ullman, 2023]; and many more. Established inventories provide a straightforward and standardized approach to psychometric evaluation. However, many studies question the reliability and validity of established inventories when used with LLMs, citing common issues of data contamination and response bias [Ye et al., 2025a]. Additionally, these simplified, structured tests may not adequately capture the complex task scenarios encountered by general-purpose AI.

Custom-Curated Items. Human-authored and custom-curated items provide tailored psychometric tests that are often more relevant and applicable to LLMs. These items enable researchers to explore novel and unique dimensions of LLM capabilities more robustly. Shen et al. [2024] argue that SVS and PVQ are inadequate for measuring LLM value alignment because they lack supplementary values from alignment literature and contextual nuances of human-AI alignment. To address this, they created 11 AI-informed value statements, similar in style to SVS and PVQ, based on a

Table 6: Overview of data and task sources: established inventories, custom-curated items, and synthetic items. For each data source, we list some examples for each construct.

Construct	Established inventories	Custom-curated items	Synthetic items
Personality traits	Barua et al. [2024], Serapio-García et al. [2023]	Ai et al. [2024], Jiang et al. [2023]	Zeng [2024]
Values	Fischer et al. [2023], Miotto et al. [2022]	Meadows et al. [2024], Shen et al. [2024]	Moore et al. [2024], Ye et al. [2025a]
Morality	Abdulhai et al. [2024], Khandelwal et al. [2024]	Hendrycks et al. [2021], Jin et al. [2022]	Liu et al. [2024d], Scherrer et al. [2023]
Attitudes & opinions	Argyle et al. [2023], Bernardelle et al. [2024]	Ceron et al. [2024], Durmus et al. [2023]	Wan and Chang [2024]
Heuristics and biases	Binz and Schulz [2023], Hagendorff et al. [2023]	Ando et al. [2023], Momennejad et al. [2023]	Malberg et al. [2024], Xie et al. [2024b]
Social interaction	[Chen et al., 2024d, Ullman, 2023]	Shapira et al. [2023a], Street et al. [2024]	Gandhi et al. [2023]
Psychology of language	Cai et al. [2024], Duan et al. [2024a]	He et al. [2024], Pérez-Mayos et al. [2021]	-
Learning and cognitive capabilities	Coda-Forno et al. [2024], Dayan et al. [2024]	Song et al. [2024a], Wu et al. [2025]	-

systematic review of human-AI alignment literature. Similarly, Ceron et al. [2024] suggest that political questionnaires may yield unreliable or inconsistent results. They propose a series of tests to assess the reliability and consistency of LLMs' stances on political statements, using a dataset of voting-advice questionnaires from seven EU countries annotated for policy issues. Regarding the evaluation of LLM morality, Hendrycks et al. [2021] curate a large-scale dataset of over 130,000 open-world scenarios requiring LLMs to make moral judgments. Another example is the ToM dataset based on naturally occurring spoken dialogues [Soubki et al., 2024], which is motivated by the potential misalignment between synthetic ToM benchmarks and actual human behavior. Despite their high ecological validity, custom-curated items require significant effort to develop and validate, which may limit their scalability and diversity.

Synthetic Items. Synthetic items generated by AI models, mostly LLMs themselves, represent an emerging paradigm in LLM psychometrics. This method leverages the generative capabilities of LLMs to produce large-scale, diverse, and contextually rich items. Using synthetic items entails careful prompt engineering and additional validation.

Some studies prompt LLMs to modify or expand established inventories for better ecological validity, mitigating data contamination, and enabling more scalable and diverse testing. For example, Ren et al. [2024] rewrite self-report items into advice-seeking queries that better align with real-world human-AI interactions. Bhandari et al. [2025b] modify the established items into semantically equivalent ones to address the issue of data contamination. They validate the rewritten items by setting a threshold for semantic embedding similarity. Zhu et al. [2024a] transform the cognitive test items into more diverse ones, allowing for multifaceted analysis of LLMs' cognitive abilities.

Other studies generate synthetic tests from scratch. Hadar-Shoval et al. [2025] validate their synthetic cognitive tests by showing a strong human performance correlation between LLM-generated tests and established tests. Ye et al. [2025a] generate value-eliciting prompts for LLM value measurement, subsequently confirming the reliability and validity of the test results. Jiang et al. [2024a] propose generative evolving testing, where an LLM-based Item Generator learns to generate items with specified difficulty. The moral dilemmas presented by Chiu et al. [2025] are created using GPT-4 and employed to evaluate the moral judgment of LLMs. When measuring SI of LLMs with agentic simulations, Mou et al. [2024], Zhou et al. [2024b] prompt LLMs to generate various components of the social interaction.

6.3 Prompting Strategies

Many structured tests employ standard test prompts, reformatted from those used for human participants. Others involve various prompting strategies.

Role-Playing Prompts. Role-playing prompts, also known as persona or profiling prompts, incorporate specific demographic information or personal attributes into the LLM context during test administration.

Some studies utilize role-playing prompts to generate multiple participants from a single LLM for statistical analysis. Serapio-García et al. [2023] incorporate persona instructions and descriptions before each item when assessing LLM personality. This approach yields diverse responses from a single LLM, facilitating the computation of the reliability and validity of the test results for each LLM. Ye et al. [2025b] use value-anchoring prompts to guide LLM values, thereby generating hundreds of LLM participants for value measurement. The statistical analysis of these measurement results leads to the development of an LLM-specific value system.

Research suggests that LLMs embody a superposition of diverse personalities, values, and perspectives [Kovač et al., 2023]. This has motivated investigations into the personality adaptability of LLMs. Utilizing role-playing prompts, Jiang et al. [2023], La Cava and Tagarelli [2024], Lu et al. [2023] demonstrate that LLMs can dynamically transition between different personality types. However, the degree of steerability varies among LLMs [La Cava and Tagarelli, 2024], and some personality dimensions are less steerable than others [Jiang et al., 2023, Li et al., 2024d]. Kovač et al. [2024], Rozen et al. [2024] employ role-playing prompts to examine the stability and consistency of LLM values. Similarly, Münker [2024], Simmons [2023], Wright et al. [2024] incorporate moral or political profiles into prompts to evaluate LLMs' ability to mimic these perspectives and their susceptibility to biases linked to specific groups.

Many cognitive benchmarks, particularly those related to social interactions, assign personas and social roles to LLMs during agentic simulations [Huang et al., 2023b, Mou et al., 2024, Zhou et al., 2024b]. The successful completion of these benchmark tasks heavily relies on the LLM's ability to comprehend and emulate the assigned persona and social role. Persona prompts are also shown to affect the social-cognitive reasoning of LLMs when other task variables are held constant [Tan et al., 2024].

The susceptibility to role-playing prompts presents a double-edged sword. On one hand, they enable the emulation of diverse perspectives and behaviors, which can be beneficial for applications such as role-playing conversations and social simulations. On the other hand, these capabilities may result in inconsistencies and instability in the LLM's values and perspectives. It may compromise the reliability and validity of test results and pose challenges to robustly aligning LLMs with human values.

Performance-Enhancing Prompts. Performance-enhancing prompts are crafted to augment LLMs in psychometric evaluations. Hagendorff [2023] suggest using such prompts whenever possible when evaluating machine psychology. The Chain of Thought (CoT) prompting [Wei et al., 2022] is leveraged to reduce biases [Hagendorff et al., 2023], enhance social intelligence [Shapira et al., 2023a], and boost cognitive capabilities [Coda-Forno et al., 2024]. Variants of CoT, such as Emotional CoT [Li et al., 2024e], specifically target improvements in the emotional intelligence of LLMs. Emotional prompts, as introduced by Li et al. [2023a], enhance the general cognitive abilities of LLMs by incorporating emotional stimuli. Few-shot prompting [Brown et al., 2020], another versatile prompting technique, has been shown to improve ToM performance in LLMs [Moghaddam and Honey, 2023].

Additional research explores diverse prompting strategies for specific constructs. For instance, Zhou et al. [2024a] advocate for MFT-guided reasoning prompts for better moral alignment. Strategies like those proposed by Echterhoff et al. [2024], Sumita et al. [2024] aim to self-debias LLMs. ToM-specific enhancement strategies include SymbolicToM [Sclar et al., 2023] and temporal decomposition-based prompting [Hou et al., 2024, Sarangi et al., 2025]. In contrast, Zhao et al. [2025a] develop a testing framework utilizing self-reflection prompts to examine both explicit and implicit social biases in LLMs, where explicit bias measurement serves as a reflection on implicit bias.

Prompt Perturbation and Adversarial Attacks. Prompt perturbations can test the robustness of LLMs' personality traits and cognitive abilities. Researchers have investigated whether LLMs can maintain stable personality, values, opinions, and cognitive abilities when item options are reordered [Lee et al., 2024d, Schelb et al., 2025], prompts are rephrased [Fraser et al., 2022, Lee et al., 2024d, Strachan et al., 2024a], prompt formats are altered [Moore et al., 2024, Schelb et al., 2025], or different languages are used [Cahyawijaya et al., 2024, Moore et al., 2024]. For instance, Faulborn et al. [2025] introduce 30 variations of prompts to evaluate the political biases of LLMs. In addition, Wen et al. [2024a] propose psychometric-inspired adversarial attacks to uncover implicit biases in LLMs. Similarly, Li et al. [2024d] utilize persuasive adversarial prompts (PAP) [Zeng et al., 2024a] to test the robustness of LLMs' values. Since LLMs are sensitive to these perturbations, researchers have begun to scrutinize the reliability of the test results obtained under standard conditions [Dominguez-Olmedo et al., 2024, Röttger et al., 2024, Strachan et al., 2024a, Ye et al., 2025a].

6.4 Model Output and Scoring

6.4.1 Closed-Ended Output and Scoring

The evaluation of LLM performance on structured tests can be categorized into two methodological paradigms: logit-based probabilistic analysis and closed-ended output scoring. In multiple-choice or Likert-scale evaluations, some studies retrieve and analyze the distribution of token-level logits—particularly the first generated token—to infer the model's latent personalities and opinions [Pellert et al., 2024, Santurkar et al., 2023], analyze response entropy [Dominguez-Olmedo et al., 2024], and compute the distributional alignment between LLM outputs and human behavioral data [Arehalli et al., 2022]. Psycholinguistic tests especially rely on logit-based metrics. A notable example is surprisal, defined as the negative log probability of a token given its context. It quantifies the predictability or cognitive effort associated with language processing, enabling researchers to map LLM outputs to human-like uncertainty patterns [Steuer et al., 2023, Wang et al., 2024b].

Conversely, closed-ended outputs, which are explicit numerical scores or categorical selections, can be analyzed using predefined scoring protocols. For Likert-scale responses, scores are typically averaged or aggregated based on established rubrics (e.g., VSM13 [Ye et al., 2025a] and DIT [Khandelwal et al., 2024]). In standardized cognitive tests, model outputs are evaluated against ground-truth labels (e.g., accuracy in arithmetic tasks) or rule-based criteria (e.g., established standards in the Verbal Comprehension Index) [Galatzer-Levy et al., 2024].

6.4.2 Open-Ended Output and Scoring

Scoring open-ended outputs is more challenging. Related scoring schemes can be generally categorized into 1) rule-based scoring, 2) model-based scoring, and 3) human scoring.

Rule-Based Scoring. Rule-based scoring predominantly relies on the lexical hypothesis [Allport and Odbert, 1936], which posits that the meaning and relevance of responses can be ascertained by analyzing the presence and frequency of specific keywords or phrases within the text. Jiang et al. [2024b] utilize the LIWC (Linguistic Inquiry and Word Count) features [Pennebaker et al., 2001] to evaluate the personalities of LLMs. Similarly, Fischer et al. [2023] employ a theory-driven value dictionary [Ponizovskiy et al., 2020] to assess LLM values. Nevertheless, lexicon-based scoring has been demonstrated to be limited in capturing semantic nuances [Ye et al., 2025a]. Some other scoring rules extend beyond simple keyword matching. For example, Healey et al. [2024] automatically identify biases by analyzing whether responses deviate from equivalent treatment.

Model-Based Scoring. Model-based scoring is prevalent in current unstructured testing due to its flexibility and scalability. Several studies have focused on training models specifically to evaluate LLM responses. For instance, Hilliard et al. [2024] fine-tuned BERT variants [Devlin et al., 2019] using the MyPersonality dataset[5] to score LLM personalities. Similarly, Sorensen et al. [2024a], Yao et al. [2024, 2025b], Ye et al. [2025a] fine-tune LLMs with established psychometric inventories [Ren et al., 2024] and/or human annotations to classify the value valence of LLM responses. Scoring can occur at the item level, as exemplified by Generative Psychometrics, which consists of item parsing, scoring, and aggregation to evaluate LLM values [Ye et al., 2025a]. Alternatively, scoring can be conducted at the response level, assessing the LLM response in its entirety [Yao et al., 2024, 2025b]. Some models are trained to score according to specific theoretical frameworks, e.g., Schwartz's Value Theory [Yao et al., 2024], while others are designed to be generalists, utilizing a broader range of fine-tuning data and leveraging the prior knowledge of LLMs [Ye et al., 2025a]. Overall, training evaluators is more prevalent for evaluating personalities and values due to the ease of associating textual expressions with specific personality and value dimensions.

More studies directly employ the LLM-as-a-judge approach [Gu et al., 2024] to score the open-ended responses. For example, Li et al. [2024d] prompt LLMs to score LLM responses across personality, ToM, and motivation. They validate the consistency between two LLM raters to indicate inter-rater reliability. Zheng et al. [2025] implement LLMs as personality evaluators and validate their consistency with human raters. The LLM-as-a-judge approach is particularly advantageous for evaluations in highly unstructured simulation settings. Mou et al. [2024], Wang et al. [2024a,d], Zhou et al. [2024b] apply LLM judges to assess SI across various dimensions, including goal completion, relationship maintenance, and adherence to social rules. Additionally, several works use embedding models to evaluate the similarity between LLM responses and prototypical examples [Amirizaniani et al., 2024, Cahyawijaya et al., 2024, Huang et al., 2024c].

Human Scoring. Human scoring is employed when rigorous evaluation, adhering strictly to standard psychometric manuals, is necessary. For instance, Elyoseph et al. [2023] engage psychologists to assess the contextual suitability of

[5] https://sites.google.com/michalkosinski.com/mypersonality.

LLM responses using the Levels of Emotional Awareness Scale (LEAS). Castello et al. [2024] conduct an examination of cognitive biases in LLMs through human evaluation and linguistic comparison. Healey et al. [2024] present a semi-automated pipeline to classify LLM responses into nuanced bias types. Other examples include ToM evaluation in open-ended responses [Amirizaniani et al., 2024] and cognitive evaluation based on the Montreal Cognitive Assessment (MoCA) [Dayan et al., 2024].

6.5 Inference Parameters

Evaluation results and validation methods also depend on inference parameters of LLMs. Some studies use greedy decoding for deterministic outputs, requiring only a single response per item but sacrificing output diversity. In contrast, sampling-based decoding introduces stochasticity, generating a broader range of outputs by adjusting hyperparameters such as *temperature*, *top-k*, and *top-p*.

Inference parameters can affect the measured traits and abilities of LLMs. Diversifying the outputs may reveal a broader spectrum of latent traits, opinions, or cognitive strategies, but also introduce variability that complicates reliability and validity assessments. Conversely, deterministic settings enhance reproducibility but may obscure the model's full range of capabilities or biases.

Most studies explicitly report and control these settings to ensure fair comparisons across models or experimental conditions. Some also examine the sensitivity of psychometric results to different decoding parameters. Transparent reporting and methodological rigor in parameter selection are essential. Investigating both deterministic and stochastic settings is recommended to fully understand the implications of test results.

7 Psychometric Validation

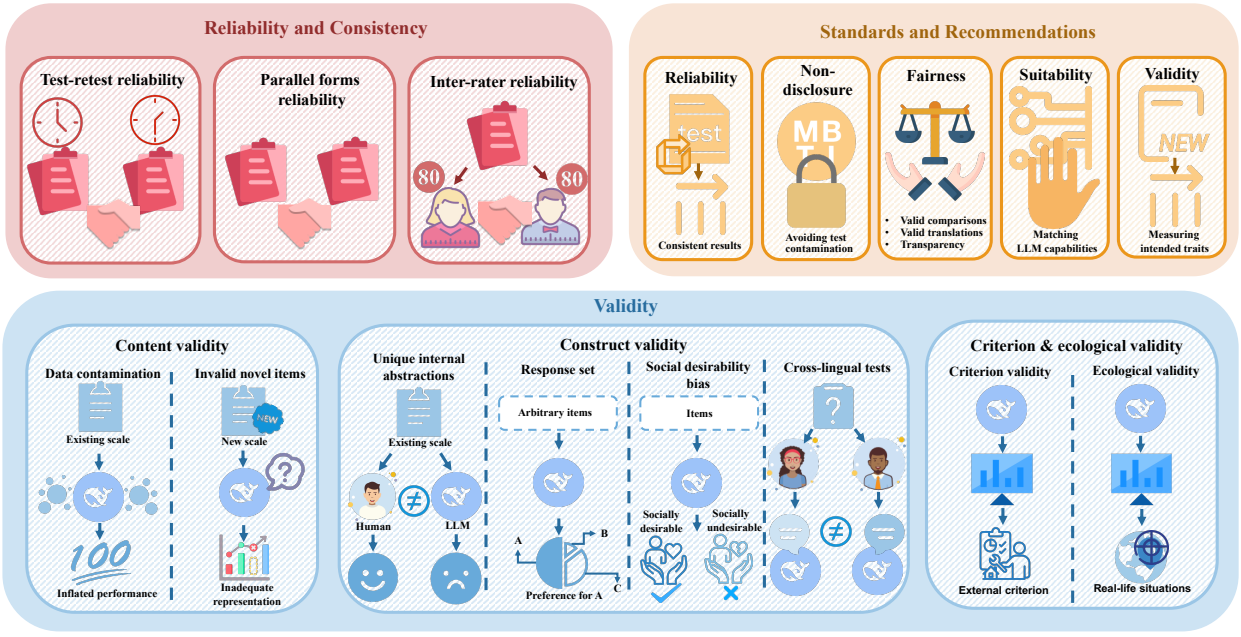

Figure 4: Overview of psychometric validation: reliability and consistency, validity, and standards and recommendations [Löhn et al., 2024].

Unlike AI benchmarking, which prioritizes system performance over test justification, psychometrics emphasizes theoretical grounding, standardized protocols, and reproducibility. Psychometric validation rigorously ensures the reliability, validity, and fairness of psychological tests. LLM psychometrics, as an emerging field, still lacks standardization for test design and administration, and recent research has begun exploring various aspects of psychometric validation.

7.1 Reliability and Consistency

Reliability is a fundamental principle in psychometric validation, assessing the degree to which a test is free from error. It encompasses consistency over time (test-retest reliability), across different versions (parallel forms reliability), and among evaluators (inter-rater reliability). Researchers have adapted these psychometric reliability metrics for LLM psychometrics. For instance, Li et al. [2024d] introduce a multi-dimensional psychometrics benchmark that extends to five forms of reliability, addressing the unique challenges of LLM evaluation: internal consistency, parallel forms reliability, inter-rater reliability, option position robustness, and adversarial attack robustness. Studies by Ceron et al. [2024], Huang et al. [2023a], Shu et al. [2024], Zheng et al. [2025] systematically investigate the consistency of LLM outputs under repeated trials, variations in prompt order, and cross-lingual contexts. They employ reliability indices such as Cronbach's alpha, intraclass correlation coefficients, and agreement statistics to quantify both internal and external reliability dimensions. We note that some metrics pertain to both stochastic and systematic errors in LLM outputs; certain inconsistencies may compromise both the reliability and validity of LLM psychometrics.

Promising results indicate that some advanced LLMs consistently generate stable responses on the Big Five personality traits across diverse settings [Huang et al., 2023a]. Larger and instruction fine-tuned models exhibit even higher reliability when utilizing role-playing prompts [Serapio-García et al., 2023]. Moore et al. [2024] also observe relatively stable value orientations of LLMs across (1) paraphrases of a single question, (2) related questions within one topic, (3) multiple-choice and open-ended formats of a single question, and (4) multilingual translations. Li et al. [2024d] report strong inter-rater reliability when using LLMs as judges, though they note varying degrees of personality stability across models and differing levels of value consistency across cultural contexts.

However, other studies present less favorable results. LLMs are notorious for prompt sensitivity in the context of psychometric evaluation. Trivial changes to prompt form [Ren et al., 2024, Röttger et al., 2024], option order [Dominguez-Olmedo et al., 2024, Li et al., 2024d], or syntactic structure [Hu and Levy, 2023] frequently result in systematic shifts in LLM outputs, severely undermining parallel forms reliability. Additionally, studies reveal distributional inconsistencies between next-token logits and forced-choice responses in psycholinguistic tests [Hu and Levy, 2023], as well as discrepancies between forced-choice and free-form responses in value [Kovač et al., 2024, Ye et al., 2025a] and political opinion surveys [Röttger et al., 2024]. In cognitive tests, Holterman and Deemter [2023], Shapira et al. [2023a], Ullman [2023] demonstrate that minor perturbations in ToM tasks often result in poor performance, indicating low parallel forms reliability of some ToM tests and LLMs' dependence on fragile heuristics. Xu et al. [2025b] develop the Words and Deeds Consistency Test (WDCT) to measure the consistency between LLMs' verbal and behavioral responses across various domains, including opinion vs. action, non-ethical value vs. action, ethical value vs. action, and theory vs. application. Their findings reveal that LLMs' responses are inconsistent.

The findings regarding reliability are clearly mixed. It seems to be influenced by various factors such as the target construct, forms of reliability, test formats, prompting, and the models used. LLMs tend to yield more reliable results in personality assessments [Huang et al., 2023a, Serapio-García et al., 2023] compared to political opinion evaluations [Dominguez-Olmedo et al., 2024, Röttger et al., 2024]. The degree of inter-rater agreement differs across various instruments and domains [Bodroža et al., 2024]. LLMs are reported to exhibit high internal consistency (Cronbach's alpha > 0.8), particularly in closed-choice personality tests [Huang et al., 2023a, Serapio-García et al., 2023, Zheng et al., 2025], while demonstrating low parallel form reliability due to prompt sensitivity [Gupta et al., 2024, Li et al., 2024d, Shu et al., 2024]. In addition, LLMs show greater consistency in structured tests with strong prompt control [Klinkert et al., 2024, Rozen et al., 2024, Serapio-García et al., 2023], but their reliability diminishes with less standardized prompts [Petrov et al., 2024] or when addressing more controversial topics [Moore et al., 2024]. The reliability also varies across models [Bodroža et al., 2024]. Moore et al. [2024] report that base models exhibit more consistent values compared to fine-tuned models in value-laden questions. Enhanced model safety is strongly associated with higher value consistency [Ye et al., 2025b], and more advanced models demonstrate superior role-playing performance, thereby increasing reliability in personality tests utilizing role-playing prompts [Serapio-García et al., 2023].

7.2 Validity

Validity, another cornerstone principle of psychometric validation, determines whether a test accurately measures its intended construct. Recent research has systematically investigated multiple aspects of validity in LLM psychometrics, addressing both methodological challenges and potential solutions unique to this emerging field.

7.2.1 Content Validity

Content validity ensures that a test comprehensively covers the construct it aims to measure. Major challenges of content validity include data contamination of well-established tests and novel items that are inadequately examined.

Data Contamination. Some studies naively transfer human tests to LLMs. Given that LLMs are trained on vast Internet-scale corpora, it is highly probable that they have been exposed to instances of test items or similar content [Hagendorff et al., 2024, Jiang et al., 2024c]. This exposure can result in the reproduction of known token patterns during evaluation, which leads to inflated performance metrics or biases the reflection of the LLM's internal traits. To address this issue, researchers reformat test content or dynamically generate novel tests; see § 6.2 for more details.

Invalid Novel Items. The reformatting or dynamic generation of novel tests presents new challenges for content validity. The novel test items may capture only limited dimensions of the construct or involve extraneous factors, thus inadequately representing target constructs. This may be due to a less rigorous crowd-sourcing process for benchmark curation, or the inherent biases and limited capabilities of the AI models used for stimulus synthesis. Content validity assessments for custom-curated and model-generated items are crucial but rarely conducted in LLM psychometrics.

7.2.2 Construct Validity

Construct validity assesses whether a test accurately measures the theoretical construct it is intended to evaluate. For example, a position paper by Alaa et al. [2025] underscores both the deficiency and critical importance of construct validity in medical LLM benchmarks. Broader literature suggests that construct validity could be undermined by the unique internal abstractions, systematic response patterns, and social desirability bias of LLMs.

Unique Internal Abstractions LLMs and humans may differ fundamentally in how they internally represent psychological constructs. For instance, Kovač et al. [2023] argue that LLMs exhibit contextually adaptive traits rather than the stable psychological characteristics found in humans. Sühr et al. [2023] systematically challenge the validity of personality tests for LLMs. They find personality types do not follow the clear separation into big five traits observed in humans. Similarly, Peereboom et al. [2024] examine the latent personality structures of humans and three LLMs, concluding that the established HEXACO tests for humans are not applicable to LLMs, as the relevant personality factors may be absent in these models. In measuring values, Ye et al. [2025a] find that self-report results are inconsistent with Schwartz's theoretical model, and that measuring open-ended responses improves the alignment, but not towards a perfect fit. Hadar-Shoval et al. [2024] also agree that LLMs encode values distinct from those of humans. There is a clear consensus among researchers that a robust evaluation and understanding of LLM psychology necessitates new operational definitions that are distinct from, or only loosely analogous to, human traits.

Pioneering research presents value systems specific to LLMs, in response to their unique internal abstractions [Biedma et al., 2024, Ye et al., 2025b]. Notably, Ye et al. [2025b] propose a theoretically grounded generative psycho-lexical approach for system construction based on GPV value measurement [Ye et al., 2025a]. They evaluate their proposed value systems through statistical analysis and explanatory, predictive, and alignment tasks, establishing the validity of the value factors and their superiority over Schwartz's value system. The unique constructs of LLMs necessitate the rigorous development of tailored tests. Ma et al. [2025] introduce the Core Sentiment Inventory (CSI), an instrument for measuring the emotional tendencies and personalities of LLMs. Fang et al. [2024] curate a psychologically grounded benchmark for mathematics proficiency, aiming for better validity, modeling item difficulty, and allowing for comparison against human norms. Similar attempts are made by Lee et al. [2024d], who propose the TRAIT test, which yields higher validity and reliability than other personality tests.

Response Set. Response set refers to the systematic pattern of answering questions in a particular way, regardless of the item's content. Some LLMs demonstrate option position bias towards responses labeled as "A", particularly in political opinion tests [Dominguez-Olmedo et al., 2024, Li et al., 2024d]. After adjusting for the option position bias, Dominguez-Olmedo et al. [2024] additionally find that LLM survey responses to political opinion questions are uniformly random. In this case, it becomes meaningless to analyze the distributional alignment between LLM responses and opinions of demographic groups. Similar results are observed in LLM personality tests [Song et al., 2023, Sühr et al., 2023] and value tests [Ye et al., 2025a]. Sühr et al. [2023] discover LLMs' tendency to respond affirmatively to reverse-coded items (e.g., "I am introverted" vs. "I am extroverted"). Ye et al. [2025a] identify that certain LLMs lean towards low/high ratings. They conclude that self-report tests are not the correct tools for LLM psychometrics.

Social Desirability Bias. Social desirability bias is the tendency to give responses that conform to social norms rather than reflect true beliefs. Studies indicate that LLMs exhibit this bias similarly to humans. Salecha et al. [2024] report that in personality tests, LLMs tend to skew their responses towards socially favorable traits, leading to higher extraversion scores and lower neuroticism ratings. Likewise, Ye et al. [2025b] find that when LLMs are directly asked to self-report values, as per Biedma et al. [2024], they avoid reporting less socially desirable values like hedonism, although they express these values when given indirect, contextual prompts. While fine-tuning for social desirability is important, it is crucial to maintain a balanced representation of diverse perspectives and values. Over-optimizing

for social desirability can result in homogeneous models that fail to capture the full range of human values and do not effectively meet diverse user needs.

Cross-lingual Tests. Some research investigates the psychometric validity in cross-lingual settings. Romero et al. [2024] administer a standardized personality questionnaire to LLMs in nine languages and find notable inconsistencies in measurement results. Despite their multilingual abilities, LLMs may lack consistency in psychological traits across languages [Cahyawijaya et al., 2024]. These inconsistencies complicate the psychometric validation of cross-lingual tests, necessitating the disentanglement of intrinsic model biases, translation equivalence issues, and cultural differences represented by the models.

7.2.3 Criterion and Ecological Validity

Criterion validity involves the correspondence of test results with external standards, while ecological validity evaluates the applicability of these outcomes to real-world scenarios. In LLM psychometrics, external standards often overlap with real-world evaluation results. Ren et al. [2024], Ye et al. [2025a] report that value orientations measured by forced-choice tests do not align with those measured in human-LLM interactions. Similar discrepancies are noted in personality [Ai et al., 2024] and morality [Nunes et al., 2024]. Consequently, Zhang et al. [2024a] advocate for linking machine personality assessments to safety. Ren et al. [2024] recommend measuring values within real-world human-LLM interactions, and Ye et al. [2025a] further demonstrate that values measured in open-ended dialog enhance construct validity given Schwartz's circumplex model. Further work by Ye et al. [2025b] connects LLM value measurement to safety prediction and value alignment tasks to ensure conformity with external standards.

7.3 Standards and Recommendations

Based on the identified challenges and solutions, researchers have proposed general standards and recommendations for LLM psychometrics, aiming to establish a methodological foundation for this nascent field. Frank [2023a] explores how developmental psychology helps us understand the internal representations and cognitive capacities of LLMs. The author recommends using simplified and novel stimuli to avoid the effects of training data contamination. [Löhn et al., 2024] critique the current state of psychometric evaluation on LLMs and argue for standardized criteria to ensure valid assessments. They propose seven requirements: reliability (consistent results), validity (measuring intended traits), suitability (matching LLM capabilities), non-disclosure (avoiding test contamination), and fairness; fairness further includes valid comparisons, valid translations, and transparency. Analyzing 25 studies, they find widespread neglect of these principles.

More recently, Hagendorff et al. [2024] recommend procedural test generation, multiple task versions, performance-enhancing prompts, shuffling options in multiple-choice questions, and using multiple scoring methods to avoid data contamination and ensure reliability. They also suggest using deterministic settings for reproducibility, automated tools for evaluation, and manual reviews for unreliable outputs. After evaluations, they advocate for performing statistical analysis to interpret the results. Vaugrante et al. [2024] identify the lack of replicability in recent works and propose four recommendations. First, they advise ensuring benchmark validity, providing adequate tasks for statistical analysis, standardizing for comparability, controlling prompt sensitivity, and aligning with research objectives. Second, adopt standardized methodologies, avoid cherry-picking, ensure statistical transparency, document experimental setups comprehensively, and define consistent evaluation metrics to enable reliable and replicable research. Third, monitor model behavior changes, account for variability with diverse models, adjust benchmark difficulty as models improve, and ensure transparency by documenting model versions and experiment dates. Lastly, standardize scoring and verification processes, ensuring accuracy and transparency; benchmark creators should provide clear verification guidelines and scoring rubrics for implementers. Schelb et al. [2025] release a framework for designing and running robust and reproducible psychometric experiments. The framework focuses on robustness, flexibility, usability, and reproducibility, based on a standardized configuration file defining experiments.

8 Psychometrics for LLM Enhancement

Psychometric principles serve as powerful tools for LLM development and enhancement, beyond mere evaluation. This shift has led to three key advancements: trait manipulation for personalized and controlled model behaviors, safety and alignment through psychometrically-informed interventions, and cognitive enhancement to foster more human-like reasoning and communication capabilities.

8.1 Trait Manipulation

One of the major applications of psychometrics is to manipulate the traits of LLMs, for chatbot personalization, role-play applications, and simulating demographic groups. Psychometrics provides a robust framework to inform such manipulation across prompting, inference, and training stages.

Structured prompting, grounded in validated psychometric inventories, are shown to reliably elicit and modulate specific personality traits in LLMs [He and Zhang, 2024, Huang et al., 2024c]. This approach facilitates the controlled simulation or alignment of synthetic personas [Jiang et al., 2023, La Cava and Tagarelli, 2024, Serapio-García et al., 2023, Zhang, 2024]. For example, Jiang et al. [2023] introduce personality prompting (P^2) based on the lexicon theories behind the Big Five personality traits. In P^2, lexicons related to specific traits are expanded into detailed personality descriptions by LLMs. The LLMs are then personalized towards diverse behaviors. Chuang et al. [2024] propose a more systematic prompting strategy based on empirically-derived human belief networks, which encompasses 64 topics loading on 9 non-overlapping latent factors. When prompted with one belief, the LLMs become more aligned with other related beliefs.

Beyond prompt engineering, inference-time interventions directly manipulate hidden representations during forward pass, with many based on patterns recognized using psychometric scales. ControlLM [Weng et al., 2024], Personality Alignment Search [Zhu et al., 2024b], Neuron-based Intervention [Deng et al., 2024], Probing-then-Editing [Ju et al., 2025], and Latent Feature Steering [Yang et al., 2025b] manipulate model outputs on trait axes by directly shifting activations or targeted neuron values at inference time, achieving trait control without retraining.

Other works further fine-tune the LLMs for trait manipulation. Vu et al. [2024] adapt the model architecture to incorporate trait control, reflecting continuous Big Five and mental health dimensions. Dan et al. [2024], Jain et al. [2024], Li et al. [2024c] deploy LoRA modules or routing networks tied to personality traits, learning specialist adapters for trait-driven generation. Cui et al. [2023], Liu et al. [2024a], Zeng et al. [2024b] achieve deep trait embedding by training on large, annotated dialogue datasets derived from validated psychometric instruments. Additional work directly modulates the model parameters without retraining. Hwang et al. [2025] generate adjustment queries based on the MBTI questionnaire and edit the model parameters to align their responses with the desired personality traits.

The manipulation of traits in LLMs presents a promising avenue for substituting human participants in social science research, as reviewed by Gao et al. [2024]. Cao et al. [2025], Suh et al. [2025] demonstrate that fine-tuning LLMs enables them to accurately simulate survey response distributions of specific subpopulations, even when faced with previously unseen survey questions. He-Yueya et al. [2024] introduce the concept of psychometric alignment to evaluate the congruence between LLMs' knowledge and human understanding. By training LLMs on human response data from the target distribution, they enhance the models' psychometric alignment on novel test items. However, the efficacy of this training approach is contingent upon the domain. Furthermore, Kang et al. [2023], Sorensen et al. [2025] explore the representation of human values within LLMs and fine-tune them to simulate human opinions through value injection.

Although current personalization techniques have demonstrated the ability to effect more than superficial changes in model outputs, achieving robust, authentic, and context-independent personalization remains a significant challenge [Dominguez-Olmedo et al., 2024, Kovač et al., 2024].

8.2 Safety and Alignment

Recent research has established a connection between LLM psychometric measurement and their safety and alignment, one of the most pressing issues in the field. Wang et al. [2025], Zhang et al. [2024a] link LLM personality traits to their safety. Specifically, Zhang et al. [2024a] identify a significant relationship between personality traits, as measured by the MBTI-M scale, and safety, noting that better alignment enhances traits such as Extraversion, Sensing, and Judging. They are inspired to modify LLM personalities, resulting in improved safety, particularly in privacy and fairness. Similarly, Wang et al. [2025] examine the correlation between HEXACO personality traits and LLM safety.

In parallel, Yao et al. [2024], Ye et al. [2025a,b] have explored the relationship between LLM values and their safety and alignment. Yao et al. [2024] demonstrate the feasibility of distinguishing between safe and unsafe LLM responses within the basic value space, analyzing correlations with Schwartz's values and specific safety issues. Ye et al. [2025a,b] achieve high accuracy in predicting LLM safety scores based on value orientations measured by GPV, a generative psychometric tool. Furthermore, Yao et al. [2024], Ye et al. [2025b] employ reinforcement learning to align LLMs with desired human values, thereby enhancing their safety. In the domain of morality, Huang et al. [2024a], Tlaie [2024] introduce prompting techniques grounded in Moral Foundation Theory to enhance the moral reasoning and alignment of LLMs. Ohashi et al. [2024], Takeshita et al. [2023] create the JCommonsenseMorality (JCM) dataset and fine-tune LLMs to adapt to local culture.

8.3 Cognitive Enhancement

Psychometrics has proven to be an effective tool for developing human-like reasoning, empathy, and communication skills in LLMs. Recent studies leverage established psychological theories and psychometric frameworks to guide LLMs' cognitive enhancement, employing strategies like prompting, architectural modules, and specialized training regimes.

Psychology-inspired prompting is shown to enhance the cognitive capabilities of LLMs. For example, Li et al. [2023a] introduce emotional prompts, demonstrating that such stimuli can improve the general cognitive abilities of LLMs. Role-playing prompts are observed to influence ToM capabilities [Tan et al., 2024] and promote more human-like reasoning [Nighojkar et al., 2025]. Zhao et al. [2025b] employ a finite-state machine paradigm based on Hill's Helping Skills theory. It embeds state transitions for emotional support conversations directly within the model's multi-hop inference process as additional contexts. This method effectively structures dialogues around psychological support strategies and emotional states, leading to higher human-rated effectiveness and strategic alignment.

Liu et al. [2023] introduce a neural listener module within an LLM architecture, optimizing a combined objective that encodes ToM reasoning into the model's training. They emphasize the potential of using psycholinguistic theories of human language acquisition to enhance LLM language learning. Additionally, preference-based and RL-based methods are proposed to internalize empathetic capabilities in LLMs. Research in empathetic response generation often utilizes cognitive and affective models of empathy to create composite reward functions or preference signals for fine-tuning model behavior. For instance, Sotolar et al. [2024] generate training data using formal emotion theories, like Plutchik's wheel of emotions, to create preference pairs for model optimization. Others have introduced empathy classifiers to shape reinforcement learning rewards, guiding models toward empathetic outputs [Sharma et al., 2021]. Further research improves LLM social-pragmatic reasoning through preference optimization based on human judgments in open-ended social tasks, showing that these objectives enhance pragmatic inference [Wu et al., 2024].

9 Trends, Challenges, and Future Directions

This section explores the emerging trends, challenges, and prospective directions in LLM psychometrics.

9.1 Psychometric Validation

There is an emerging consensus on the importance of psychometric validation in evaluating personality traits in LLMs; various studies contribute to this validation across different dimensions (§ 7). However, research on psychometric validation in ability testing remains limited. Current evaluations of LLM abilities predominantly employ structured tests, such as multiple-choice questions, often directly adapted from tests originally designed for human participants. These evaluations face challenges. For instance, Ullman [2023] highlight that LLMs fail on trivial alterations of ToM tasks, indicating a lack of parallel forms reliability. The unique internal abstractions of LLMs undermine construct validity and lead to flawed benchmarks [Riemer et al., 2025]. Furthermore, the extent to which results from structured tests of LLM abilities generalize to real-world human-AI interactions is rarely investigated, leaving criterion and ecological validity issues insufficiently explored.

On the other hand, newly developed and more sophisticated simulation-based tests are often constructed with less rigorous psychometric processes. Consequently, they may involve extraneous factors unrelated to the construct of interest or fail to adequately cover the construct's manifestations. In the realm of psychometric-inspired, construct-oriented tests, pioneering efforts such as Zhu et al. [2024a] introduce multifaceted evaluations of three cognitive factors identified from benchmarking data, while Zhou et al. [2025] propose a theory-driven, richer set of general scales. Defining a comprehensive set of core abilities and designing valid tests to measure them remains an open challenge.

9.2 From Human Constructs to LLM Constructs

Recent research shifts from employing constructs established for humans to developing novel constructs specific to LLMs. Studies by Biedma et al. [2024], Peereboom et al. [2024], Sühr et al. [2023], Ye et al. [2025b] indicate that the factor structures of human personalities and values may not be applicable to LLMs. Further, Biedma et al. [2024], Burnell et al. [2023], Federiakin [2025], Ye et al. [2025b] have tailored these constructs to better fit LLMs. Despite these advancements, the development of novel constructs seems highly contingent upon the measurement tools, tasks, and specific LLMs utilized. For instance, Ye et al. [2025b] and Biedma et al. [2024] report differing value structures, while Federiakin [2025] and Burnell et al. [2023] identify varying ability factors. Future research should further explore the fundamental structures that underpin LLM behaviors.

9.3 Perceived vs. Aligned Traits

Han et al. [2025] suggest a discrepancy between perceived values in text (objective human annotation) and how responses align with annotators' personal values (subjective annotation). For instance, consider the LLM response: "Wealthy individuals are not necessarily greedy; some may be motivated by concern for the greater good." An objective annotation categorizes this response as embodying Power, due to its themes of societal impact. However, annotators with values of Conformity, rather than Power, report a higher similarity between the response and the their own thoughts. In this case, Power represents the perceived value, while Conformity is the aligned value.

This discrepancy may extend to other traits and holds significant methodological implications for LLM psychometrics. It is crucial to deliberate on which aspect—perceived traits or aligned traits—holds greater significance. Current research on value evaluation is driven by the potential of LLMs to influence, alter, and entrench human values. However, the underlying mechanisms of such influence remain unclear. For instance, if individuals are inclined to adopt LLM decisions and values [Glickman and Sharot, 2024], it is uncertain whether they are adhering to perceived traits or aligned traits. Additionally, the impact of judges' subjectivity on psychometric evaluation is not well understood. Existing scoring models or LLM-as-a-judge approaches often assume a universal criterion. The modeling of judges' subjectivity, such as how individuals with different values may interpret the same response differently, remains an underexplored area.

9.4 Anthropomorphization Challenges

The anthropomorphization of LLMs in research varies, which leads to different ways of performing statistical analysis. Some studies conceptualize an LLM as an individual entity, while others regard it as a population. For instance, certain studies evaluate LLM traits by considering the model as a single participant, utilizing its default settings with prompts such as "You are a helpful assistant" [Ye et al., 2025a]. Alternatively, some researchers view an LLM as a superposition of multiple participants, employing diverse role-playing prompts to steer it towards different profiles [Serapio-García et al., 2023]. Another approach involves treating the LLM as a monolithic entity, but administering varied prompts for identical test items to generate multiple sets of results [Hagendorff et al., 2023].

When an LLM is treated as a monolithic entity, the reliability and validity of the results are primarily contingent upon the measurement tools, allowing researchers to compare different tools [Ye et al., 2025a]. Conversely, when an LLM is perceived as a population, the reliability and validity are also influenced by the specific LLM employed, prompting comparisons of consistency and role-playing capabilities across different models [Serapio-García et al., 2023].

The meaningfulness of evaluating LLMs in their default settings remains disputable, especially given their capacity to adopt diverse profiles. The exploration of trait spectrums—a continuum or set of trait demonstrations—in LLMs is currently under-researched. Furthermore, the development of a novel psychometric framework to adequately capture this spectrum has yet to be established. For instance, it is unclear how to identify the extremes to which an LLM can be steered. There is also a lack of consensus on the optimal alignment of LLMs along psychometric dimensions. The chosen alignment target significantly influences the anthropomorphization of LLMs. Some scholars advocate for alignment that enhances consistency, thereby allowing LLMs to be perceived as monolithic entities [Röttger et al., 2024]. Conversely, others position alignment toward pluralistic values, suggesting that LLMs should be regarded as a population [Sorensen et al., 2024b].

9.5 Expanding Dimensions in Model Deployment

LLM psychometrics is expanding beyond traditional text-based, single-turn interactions. Several emerging dimensions present both opportunities and challenges for future research.

Multi-Lingual Evaluation. While most psychometric evaluations are conducted in English, multilingual LLMs necessitate comprehensive cross-linguistic validation. Current research indicates that LLMs exhibit varying personality traits and cognitive abilities across languages (§ 7.2.2). This raises questions regarding cultural and linguistic biases in existing evaluation frameworks. Future research should examine how measurement results vary across languages, whether psychometric properties (reliability, validity) maintain consistency across linguistic contexts, and how to develop culturally appropriate test instruments.

Multi-Turn Interactions. The dynamic nature of multi-turn interactions introduces new challenges for psychometric evaluation. Compared to single-turn responses, multi-turn interactions more authentically simulate how humans engage with psychometric tests, where responses are influenced by preceding interactions. Research indicates that when interactions extend across multiple turns, LLMs may exhibit significantly different traits, such as increased vulnerability

to jailbreaking attacks [Ying et al., 2025b]. These findings highlight the need for novel evaluation methods that capture temporal dynamics in extended conversations.

Multi-Modal Capabilities. Multi-modal models require novel psychometric approaches. Current evaluation methods, primarily designed for text-based interactions, fail to adequately capture the complexity of multi-modal understanding and generation. Li et al. [2024b] introduce one of the first value evaluation tools for Vision-Language Models (VLMs). However, this work remains preliminary, employing a highly simplified and controlled test environment that utilizes only the first frame of videos as test stimuli. Furthermore, psychological traits such as values present conceptual challenges. Even when a value like honesty is precisely defined, determining which specific real-world actions manifest this trait remains ambiguous. Consequently, extending psychometric tests to embodied agents and action modalities introduces even greater methodological obstacles. Future research should expand current psychometric tests to incorporate additional modalities. The development of cross-modal evaluation frameworks should maintain psychometric rigor while accounting for the unique characteristics of each modality.

Agent and Multi-Agent Systems. The emergence of LLM-based agents and multi-agent systems introduces new dimensions to psychometric evaluation. When LLMs operate as autonomous agents, their behavior may be influenced by environmental factors, memory systems, tool usage, and interaction patterns with other agents. However, systematic methodologies for psychometric evaluation in these complex, dynamic settings remain largely undeveloped.

9.6 Item Response Theory

Recent work demonstrates that applying IRT to LLM outputs offers several key advantages, including more informative benchmarking [Guinet et al., 2024], identification of items that effectively discriminate among high-performing models [Lalor et al., 2024], and the potential for adaptive testing that reduce evaluation costs [Truong et al., 2025, Zhuang et al., 2023a]. Most existing applications focus on standard IRT models such as 1PL, 2PL, and 3PL, with limited exploration of polytomous, hierarchical, or fully multidimensional IRT models, even though LLM capabilities are complex and multifaceted. Integrating IRT with generative AI for automated item generation is an emerging direction; however, the current approach can be overly complex, limiting its practical utility [Jiang et al., 2024a]. Furthermore, while IRT holds promise for standardizing the measurement of LLMs and humans on a unified scale—even when different test items are used—this potential remains largely unrealized.

The systematic application of IRT for bias analysis—such as detecting differential item functioning across model families—remains in its early stages [He-Yueya et al., 2024]. IRT offers a principled framework for uncovering biases in AI systems by analyzing performance across items with varying difficulty levels and demographic contexts. For instance, if items involving certain demographic groups are disproportionately challenging, it may indicate underlying model or data biases. Robustness evaluation can also be strengthened by leveraging IRT to assess model performance on adversarial or edge-case items.

IRT can be employed to evaluate the quality of AI benchmarks themselves. While current benchmarks often comprise large collections of test items, the informativeness and design quality of these items are frequently uncertain. Not all items contribute equally; some may lack discriminatory power or inadvertently favor particular model designs. IRT enables a systematic assessment of item difficulty, discrimination, informativeness, and vulnerability to bias and guessing, thereby facilitating the development of more robust and equitable evaluation frameworks.

Furthermore, IRT is rarely utilized beyond evaluation, despite its potential for informing model development. By quantifying item discrimination, IRT can guide the construction of training datasets. Highly discriminating items can be particularly valuable as they provide deeper insights into differentiating between low- and high-performing models.

9.7 From Evaluation to Enhancement

The ultimate objective of evaluation is not only to understand models but also to facilitate their improvement. Current psychometric approaches primarily emphasize the evaluation, comparison, and interpretation of LLMs, while the application of psychometric insights to model enhancement lags behind. As discussed in § 8, psychometric principles can inform prompt engineering, inference-time control of internal representations, training data curation, reward signal design for fine-tuning, and the development of methodological frameworks.

Existing enhancement techniques are still in their infancy. For instance, value targets for LLM alignment in Yao et al. [2024], Ye et al. [2025b] are statically defined, which constrains the incorporation of context-dependent value states observed in human behavior [Skimina et al., 2021]. Drawing on insights from psychometrics enables the development of more effective model enhancement techniques.

10 Conclusion

This paper presents a comprehensive survey of LLM psychometrics. The integration of psychometric instruments, theories, and principles into LLM evaluation promises to overcome the limitations of traditional AI benchmarks. This approach enables us to more effectively capture the broad, emergent psychological constructs of LLMs, encompassing both personality and cognitive dimensions. Psychometric evaluation methodologies vary in test formats, data and task sources, prompting strategies, model outputs, and scoring methods. They each have distinct strengths, weaknesses, and application scenarios, yet all must adhere to psychometric principles such as reliability, validity, and fairness. Beyond evaluation, psychometric-inspired techniques enhance LLMs in trait manipulation, safety and alignment, and cognitive capabilities, contributing to the development of more powerful and responsible AI systems.

It is widely acknowledged that the trajectory of AI development is shifting towards an era characterized by evaluation-driven progress [AAAI, 2025, Silver and Sutton, 2025, Yao, 2025]. We posit that the LLM psychometrics will be pivotal in this evolution, introducing novel principles, dimensions, techniques, and insights. We aim this survey to inspire future evaluation paradigms for human-level AI and foster the advancement of AI psychology for greater common good.

References

[1] AAAI. Ai evaluation. In Francesca Rossi, editor, *Future of AI Research*. Association for the Advancement of Artificial Intelligence (AAAI), 2025. URL https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report_FINAL.pdf. Report by the AAAI 2025 Presidential Panel on the Future of AI Research.

[2] Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, 2024.

[3] Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti, and Massimo Stella. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3):124, 2023.

[4] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*, 2024.

[5] Muhammad Shahrul Zaim bin Ahmad and Kazuhiro Takemoto. Large-scale moral machine experiment on large language models. *arXiv preprint arXiv:2411.06790*, 2024.

[6] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, 2024.

[7] Yiming Ai, Zhiwei He, Ziyin Zhang, Wenhong Zhu, Hongkun Hao, Kai Yu, Lingjun Chen, and Rui Wang. Is self-knowledge and action consistent or not: Investigating large language model's personality. *arXiv preprint arXiv:2402.14679*, 2024.

[8] Meltem Aksoy. Whose morality do they speak? unraveling cultural bias in multilingual language models. *arXiv preprint arXiv:2412.18863*, 2024.

[9] Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*, 2025.

[10] Josh Albrecht, Ellie Kitanidis, and Abraham Fetterman. Despite" super-human" performance, current llms are unsuited for decisions about ethics and safety. In *NeurIPS ML Safety Workshop*.

[11] Gordon W Allport and Henry S Odbert. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):i, 1936.

[12] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, A. Mashhadi, and Chirag Shah. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses, 2024.

[13] S. Amouyal, A. Meltzer-Asscher, and Jonathan Berant. Large language models for psycholinguistic plausibility pretesting, 2024.

[14] Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Evaluating large language models with neubaroco: Syllogistic reasoning ability and human-like biases. *arXiv preprint arXiv:2306.12567*, 2023.

[15] APA Dictionary of Psychology. Personality, n.d. URL https://dictionary.apa.org/personality.

[16] Suhas Arehalli and Tal Linzen. Neural language models capture some, but not all, agreement attractioneffects. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42, 2020.

[17] Suhas Arehalli, Brian Dillon, and Tal Linzen. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities, 2022.

[18] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[19] Michael C Ashton and Kibeom Lee. Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and social psychology review*, 11(2):150–166, 2007.

[20] Michael C Ashton and Kibeom Lee. The hexaco–60: A short measure of the major dimensions of personality. *Journal of personality assessment*, 91(4):340–345, 2009.

[21] Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157, 2023.

[22] Leif Azzopardi and Yashar Moshfeghi. Prism: a methodology for auditing biases in large language models. *arXiv preprint arXiv:2410.18906*, 2024.

[23] Yanhong Bai, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xingjiao Wu, and Liang He. Fairmonitor: A dual-framework for detecting stereotypes and biases in large language models. *arXiv preprint arXiv:2405.03098*, 2024.

[24] Sarah Ball, Simeon Allmendinger, Frauke Kreuter, and Niklas Kühl. Human preferences in large language model latent space: A technical analysis on the reliability of synthetic data in voting outcome prediction. *arXiv preprint arXiv:2502.16280*, 2025.

[25] Adrita Barua, Gary Brase, Ke Dong, Pascal Hitzler, and Eugene Vasserman. On the psychology of gpt-4: Moderately anxious, slightly masculine, honest, and humble. *arXiv preprint arXiv:2402.01777*, 2024.

[26] Tom L Beauchamp. Philosophical ethics: An introduction to moral philosophy. 2001.

[27] Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*, 2024.

[28] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

[29] Manfred Max Bergman. A theoretical note on the differences between attitudes, opinions, and values. *Swiss Political Science Review*, 4(2):81–93, 1998.

[30] Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. Mapping and influencing the political ideology of large language models using synthetic personas. *arXiv preprint arXiv:2412.14843*, 2024.

[31] Pranav Bhandari, Nicolas Fay, Michael Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. Can llm agents maintain a persona in discourse? *arXiv preprint arXiv:2502.11843*, 2025a.

[32] Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. Evaluating personality traits in large language models: Insights from psychological questionnaires. *arXiv preprint arXiv:2502.05248*, 2025b.

[33] Pablo Biedma, Xiaoyuan Yi, Linus Huang, Maosong Sun, and Xing Xie. Beyond human norms: Unveiling unique values of large language models through interdisciplinary approaches. *arXiv preprint arXiv:2404.12744*, 2024.

[34] Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

[35] James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

[36] Bojana Bodroža, Bojana M Dinić, and Ljubiša Bojić. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10):240180, 2024.

[37] Ljubiša Bojić, Predrag Kovacevic, and Milan Čabarkapa. Gpt-4 surpassing human performance in linguistic pragmatics, 2023.

[38] Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. Sage: Evaluating moral consistency in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14272–14284, 2024.

[39] Léonard Boussioux, Jacqueline N Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R Lakhani. The crowdless future? generative ai and creative problem-solving. *Organization Science*, 35(5):1589–1607, 2024.

[40] Selmer Bringsjord and Bettina Schimanski. What is artificial intelligence? psychometric ai as an answer. In *IJCAI*, pages 887–893. Citeseer, 2003.

[41] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[42] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[43] Ryan Burnell, Han Hao, Andrew RA Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*, 2023.

[44] Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. High-dimension human value representation in large language models. *arXiv preprint arXiv:2404.07900*, 2024.

[45] Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56, 2024.

[46] Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. *arXiv preprint arXiv:2502.07068*, 2025.

[47] Graham Caron and Shashank Srivastava. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*, 2022.

[48] David W Carroll. *Psychology of language.* Thomson Brooks/Cole Publishing Co, 1986.

[49] Marta Castello, Giada Pantana, and Ilaria Torre. Examining cognitive biases in chatgpt 3.5 and chatgpt 4 through human evaluation and linguistic comparison. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 250–260, 2024.

[50] Raymond B Cattell and John L Horn. A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement*, 15(3):139–164, 1978.

[51] Pew Research Center. Spring 2013 survey data, 2013. URL https://www.pewresearch.org/dataset/spring-2013-survey-data/.

[52] Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms. *Transactions of the Association for Computational Linguistics*, 12:1378–1400, 2024.

[53] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34, 2024.

[54] Ilias Chalkidis and Stephanie Brandl. Llama meets eu: Investigating the european political spectrum through the lens of llms. *arXiv preprint arXiv:2403.13592*, 2024.

[55] Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding, 2024.

[56] Tyler A Chang and Benjamin K Bergen. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350, 2024.

[57] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[58] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

[59] Yanquan Chen, Zhen Wu, Junjie Guo, Shujian Huang, and Xinyu Dai. Extroversion or introversion? controlling the personality of your large language models. *arXiv preprint arXiv:2406.04583*, 2024a.

[60] Yanran Chen and Steffen Eger. Do emotions really affect argument convincingness? a dynamic approach with llm-based manipulation checks. *arXiv preprint arXiv:2503.00024*, 2025.

[61] Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. Emotionqueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2149–2176, 2024b.

[62] Zhawnen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. Through the theory of mind's eye: Reading minds with multimodal video large language models, 2024c.

[63] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. Tombench: Benchmarking theory of mind in large language models, 2024d.

[64] Yu Ying Chiu, Liwei Jiang, and Yejin Choi. Dailydilemmas: Revealing value preferences of llms with quandaries of daily life. In *The Thirteenth International Conference on Learning Representations*, 2025.

[65] Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Beyond demographics: Aligning role-playing llm-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14010–14026, 2024.

[66] Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47 (4):1178–1198, 2015.

[67] Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. Cogbench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225*, 2024.

[68] Erica Coppolillo, Giuseppe Manco, and Luca Maria Aiello. Unmasking conversational bias in ai multiagent systems. *arXiv preprint arXiv:2501.14844*, 2025.

[69] Paul T Costa and Robert R McCrae. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198, 2008.

[70] Linda Crocker and James Algina. *Introduction to classical and modern test theory*. ERIC, 1986.

[71] Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, YongHong Tian, and Li Yuan. Machine mindset: An mbti exploration of large language models. *arXiv preprint arXiv:2312.12999*, 2023.

[72] Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. P-tailor: Customizing personality traits for language models via mixture of specialized lora experts. *arXiv preprint arXiv:2406.12548*, 2024.

[73] Roy Dayan, Benjamin Uliel, and Gal Koplewitz. Age against the machine-susceptibility of large language models to cognitive impairment: cross sectional analysis., 2024.

[74] Edoardo Sebastiano De Duro, Enrique Taietta, Riccardo Improta, and Massimo Stella. Phdgpt: Introducing a psychometric and linguistic dataset about how large language models perceive graduate students and professors in psychology. *arXiv preprint arXiv:2411.10473*, 2024.

[75] Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701, 2023.

[76] Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. Neuron-based personality trait induction in large language models. *arXiv preprint arXiv:2410.12327*, 2024.

[77] Vittoria Dentella, Fritz Guenther, and Evelina Leivada. Language in vivo vs. in silico: Size matters but larger language models still do not comprehend language on a par with humans, 2024.

[78] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[79] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878, 2024.

[80] Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, et al. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *arXiv preprint arXiv:2505.00049*, 2025.

[81] Ina Dormuth, Sven Franke, Marlies Hafer, Tim Katzke, Alexander Marx, Emmanuel Müller, Daniel Neider, Markus Pauly, and Jérôme Rutinowski. A cautionary tale about" neutrally" informative ai tools ahead of the 2025 federal elections in germany. *arXiv preprint arXiv:2502.15568*, 2025.

[82] Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. Do personality tests generalize to large language models? In *Socially Responsible Language Modelling Research*, 2023.

[83] Shitong Duan, Xiaoyuan Yi, Peng Zhang, Tun Lu, Xing Xie, and Ning Gu. Denevil: Towards deciphering and navigating the ethical values of large language models via instruction learning. *arXiv preprint arXiv:2310.11053*, 2023.

[84] Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhenguang G. Cai. Hlb: Benchmarking llms' humanlikeness in language use, 2024a.

[85] Xufeng Duan, Xinyu Zhou, Bei Xiao, and Zhenguang G. Cai. Unveiling language competence neurons: A psycholinguistic approach to model interpretability, 2024b.

[86] Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

[87] Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with llms. *arXiv preprint arXiv:2403.00811*, 2024.

[88] Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in psychology*, 14:1199058, 2023.

[89] Susan E Embretson and Steven P Reise. *Item response theory for psychologists*. Psychology Press, 2013.

[90] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

[91] Thomas G Evans. A heuristic program to solve geometric-analogy problems. In *Proceedings of the April 21-23, 1964, spring joint computer conference*, pages 327–338, 1964.

[92] Qixiang Fang, Daniel L Oberski, and Dong Nguyen. Patch! psychometrics-assisted benchmarking of large language models: A case study of proficiency in 8th grade mathematics. *arXiv preprint arXiv:2404.01799*, 2024.

[93] Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. Only a little to the left: A theory-grounded measure of political bias in large language models. *arXiv preprint arXiv:2503.16148*, 2025.

[94] Denis Federiakin. Improving llm leaderboards with psychometrical methodology. *arXiv preprint arXiv:2501.17200*, 2025.

[95] Anita Feher and Philip A Vernon. Looking beyond the big five: A selective review of alternatives to the big five model of personality. *Personality and Individual Differences*, 169:110002, 2021.

[96] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*, 2023.

[97] Ronald Fischer, Markus Luczak-Roesch, and Johannes A Karl. What does chatgpt return about human values? exploring value bias in chatgpt using a descriptive value theory. *arXiv preprint arXiv:2304.03612*, 2023.

[98] Michael C Frank. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2 (8):451–452, 2023a.

[99] Michael C Frank. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992, 2023b.

[100] Kathleen C Fraser, Svetlana Kiritchenko, and Esma Balkir. Does moral code have a moral code? probing delphi's moral philosophy. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 26–42, 2022.

[101] Ivar Frisch and Mario Giulianelli. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*, 2024.

[102] Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, 2019.

[103] Isaac R. Galatzer-Levy, David Munday, Jed N McGiffin, Xin Liu, Danny Karmon, Ilia Labzovsky, Rivka Moroshko, Amir Zait, and Daniel McDuff. The cognitive capabilities of generative ai: A comparative analysis with human benchmarks, 2024.

[104] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[105] Kanishk Gandhi, Jan-Philipp Franken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models, 2023.

[106] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24, 2024.

[107] Basile Garcia, Crystal Qian, and Stefano Palminteri. The moral turing test: Evaluating human-llm alignment in moral decision-making. *arXiv preprint arXiv:2410.07304*, 2024.

[108] Mingmeng Geng, Sihong He, and Roberto Trotta. Are large language models chameleons? an attempt to simulate social surveys. *arXiv preprint arXiv:2405.19323*, 2024.

[109] Moshe Glickman and Tali Sharot. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, pages 1–15, 2024.

[110] Lewis R Goldberg. An alternative "description of personality": The big-five factor structure. In *Personality and personality disorders*, pages 34–47. Routledge, 2013.

[111] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009.

[112] Leslie S Greenberg. Emotion–focused therapy. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 11(1):3–16, 2004.

[113] Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109, 2023.

[114] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[115] Joy Paul Guilford, Paul R Christensen, Philip R Merrifield, and Robert C Wilson. Alternate uses. 1978.

[116] Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. Automated evaluation of retrieval-augmented language models with task-specific exam generation. 2024.

[117] Xiao-Yu Guo, Yuan-Fang Li, and Reza Haf. Desiq: Towards an unbiased, challenging benchmark for social intelligence understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3169–3180, 2023a.

[118] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*, 2023b.

[119] Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of llm personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314, 2024.

[120] Dorit Hadar-Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using schwartz's theory of basic values. *JMIR Mental Health*, 11:e55988, 2024.

[121] Dorit Hadar-Shoval, Maya Lvovsky, Kfir Asraf, Yoav Shimoni, Zohar Elyoseph, et al. The feasibility of large language models in verbal comprehension assessment: Mixed methods feasibility study. *JMIR Formative Research*, 9(1):e68347, 2025.

[122] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. World values survey wave 7 (2017-2022) cross-national data-set. *(No Title)*, 2022.

[123] Thilo Hagendorff. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 1, 2023.

[124] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838, 2023.

[125] Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. Machine psychology. *arXiv preprint cs.CL/2303.13988*, 2024.

[126] Patrick Haller, Ansar Aynetdinov, and Alan Akbik. Opiniongpt: Modelling explicit biases in instruction-tuned llms. *arXiv preprint arXiv:2309.03876*, 2023.

[127] Hyemin Han. Potential benefits of employing large language models in research in moral education and development. *Journal of Moral Education*, pages 1–16, 2023.

[128] Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. Value portrait: Understanding values of llms with human-aligned benchmark. *arXiv preprint arXiv:2505.01015*, 2025.

[129] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.

[130] William M Hayes, Nicolas Yax, and Stefano Palminteri. Relative value biases in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[131] Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schutze, N. Mesgarani, and Jonathan Brennan. Large language models as neurolinguistic subjects: Discrepancy in performance and competence for form and meaning, 2024.

[132] Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models, 2023.

[133] Zihong He and Changwang Zhang. Afspp: Agent framework for shaping preference and personality with large language models. *arXiv preprint arXiv:2401.02870*, 2024.

[134] Joy He-Yueya, Wanjing Anya Ma, Kanishk Gandhi, Benjamin W Domingue, Emma Brunskill, and Noah D Goodman. Psychometric alignment: Capturing human knowledge distributions via language models. *arXiv preprint arXiv:2407.15645*, 2024.

[135] Jennifer Healey, Laurie Byrum, Md Nadeem Akhtar, and Moumita Sinha. Evaluating nuanced bias in large language model free response answers. In *International Conference on Applications of Natural Language to Information Systems*, pages 378–391. Springer, 2024.

[136] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[137] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. In *International Conference on Learning Representations*, 2021.

[138] Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. Eliciting personality traits in large language models. *arXiv preprint arXiv:2402.08341*, 2024.

[139] Geert Hofstede. *Culture's consequences: International differences in work-related values*, volume 5. sage, 1984.

[140] Geert Hofstede. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Sage Publications, 2001.

[141] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8, 2011.

[142] Bart Holterman and K. V. Deemter. Does chatgpt have theory of mind?, 2023.

[143] Xudong Hong, Margarita Ryzhova, Daniel Adrian Biondi, and Vera Demberg. Do large language models and humans have similar behaviours in causal inference with script knowledge?, 2023.

[144] Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind, 2024.

[145] Robert J House, Paul J Hanges, Mansour Javidan, Peter W Dorfman, and Vipin Gupta. *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications, 2004.

[146] He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025.

[147] Jennifer Hu and Roger Levy. Prompt-based methods may underestimate large language models' linguistic generalizations, 2023.

[148] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, 2023.

[149] Meng Hua, Yuan Cheng, and Hengshu Zhu. The career interests of large language models. *arXiv preprint arXiv:2407.08564*, 2024.

[150] Allison Huang, Carlos Mougan, and Yulu Pi. Moral persuasion in large language models: Evaluating susceptibility and ethical alignment. In *The Third Workshop on New Frontiers in Adversarial Machine Learning*, 2024a.

[151] Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael R Lyu. Revisiting the reliability of psychological scales on large language models. *arXiv preprint arXiv:2305.19926*, 2023a.

[152] Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*, 2023b.

[153] Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*, 2023c.

[154] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*, 2023d.

[155] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*, 2023e.

[156] Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Apathetic or empathetic? evaluating llms' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37:97053–97087, 2024b.

[157] Muhua Huang, Xijuan Zhang, Christopher Soto, and James Evans. Designing llm-agents with personalities: A psychometric approach. *arXiv preprint arXiv:2410.19238*, 2024c.

[158] Kent Hubert, Kim N Awa, and Darya Zabelina. Artificial intelligence is more creative than humans: A cognitive science perspective on the current state of generative language models. 2023.

[159] Markus Huff and Elanur Ulakçi. Towards a psychology of machines: Large language models predict human memory, 2024.

[160] EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*, 2023.

[161] Seojin Hwang, Yumin Kim, Byeongjeong Kim, and Hwanhee Lee. Personality editing for language models through relevant knowledge editing. *arXiv preprint arXiv:2502.11789*, 2025.

[162] Yusuke Ide, Yuto Nishida, Miyu Oba, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. How to make the most of llms' grammatical knowledge for acceptability judgments, 2024.

[163] David Ilić and Gilles E Gignac. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858, 2024.

[164] Navya Jain, Zekun Wu, Cristian Munoz, Airlie Hilliard, Xin Guan, Adriano Koshiyama, Emre Kazim, and Philip Treleaven. From text to emoji: How peft-driven personality manipulation unleashes the emoji potential in llms. *arXiv preprint arXiv:2409.10245*, 2024.

[165] Mohsen Jamali, Ziv M. Williams, and Jing Cai. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain, 2023.

[166] Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. Moralbench: Moral evaluation of llms. *arXiv preprint arXiv:2406.04428*, 2024.

[167] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36: 10622–10643, 2023.

[168] Han Jiang, Xiaoyuan Yi, Zhihua Wei, Ziang Xiao, Shu Wang, and Xing Xie. Raising the bar: Investigating the values of large language models via generative evolving testing. *arXiv preprint arXiv:2406.14230*, 2024a.

[169] Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. Communitylm: Probing partisan worldviews from language models. *arXiv preprint arXiv:2209.07065*, 2022.

[170] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, 2024b.

[171] Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*, 2024c.

[172] Shapeng Jiang, Lijia Wei, and Chen Zhang. Donald trumps in the virtual polls: Simulating and predicting public opinions in surveys using large language models. *arXiv preprint arXiv:2411.01582*, 2024d.

[173] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, T. Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering, 2024a.

[174] Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35:28458–28473, 2022.

[175] Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, et al. Language model alignment in multilingual trolley problems. *arXiv preprint arXiv:2407.02273*, 2024b.

[176] Yuu Jinnai. Does cross-cultural alignment change the commonsense morality of language models? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 48–64, 2024.

[177] Oliver P. John, Emily M. Donahue, and Robert L. Kentle. Big five inventory, 1991. URL https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft07550-000. Journal of Personality and Social Psychology.

[178] Peter K Jonason and Gregory D Webster. The dirty dozen: a concise measure of the dark triad. *Psychological assessment*, 22(2):420, 2010.

[179] Cameron R. Jones, Sean Trott, and Benjamin K. Bergen. Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (epitome), 2024.

[180] Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. Probing then editing response personality of large language models. *arXiv preprint arXiv:2504.10227*, 2025.

[181] Kirill Kalinin. Improving gpt generated synthetic samples with sampling-permutation algorithm. *Available at SSRN 4548937*, 2023.

[182] Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. From values to opinions: Predicting human behaviors and stances using value-injected large language models. *arXiv preprint arXiv:2310.17857*, 2023.

[183] Robert M Kaplan and Dennis P Saccuzzo. *Psychological testing: Principles, applications, and issues*. Wadsworth/Thomson Learning, 2001.

[184] Elena Kardanova, Alina Ivanova, Ksenia Tarasova, Taras Pashchenko, Aleksei Tikhoniuk, Elen Yusupova, Anatoly Kasprzhak, Yaroslav Kuzminov, Ekaterina Kruchinskaia, and Irina Brun. A novel psychometrics-based approach to developing professional competency benchmark for large language models. *arXiv preprint arXiv:2411.00045*, 2024.

[185] Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*, 2024.

[186] Artem Karpov, Seong Hah Cho, Austin Meek, Raymond Koopmanschap, Lucy Farnik, and Bogdan-Ionut Cirstea. Inducing human-like biases in moral reasoning language models. *arXiv preprint arXiv:2411.15386*, 2024.

[187] Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*, 2022.

[188] Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2882–2894, 2024.

[189] Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*, 2024.

[190] Hyunwoo Kim, Melanie Sclar, Zhi-Xuan Tan, Lance Ying, Sydney Levine, Yang Liu, Joshua B. Tenenbaum, and Yejin Choi. Hypothesis-driven theory-of-mind reasoning for large language models, 2025a.

[191] Junsol Kim and Byungkyu Lee. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*, 2023.

[192] Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. *arXiv preprint arXiv:2503.02080*, 2025b.

[193] Minsang Kim and Seungjun Baek. Exploring large language models on cross-cultural values in connection with training methodology. *arXiv preprint arXiv:2412.08846*, 2024.

[194] Lawrence J Klinkert, Stephanie Buongiorno, and Corey Clark. Driving generative agents with their personality. *arXiv preprint arXiv:2402.14879*, 2024.

[195] Lawrence Kohlberg. *Development of moral character and moral ideology*, volume 1. University of Chicago, 1964.

[196] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models, 2023a.

[197] Michal Kosinski. Evaluating large language models in theory of mind tasks, 2023b.

[198] Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*, 2023.

[199] Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. Stick to your role! stability of personal values expressed in large language models. *Plos one*, 19(8):e0309114, 2024.

[200] Grgur Kovač, Rémy Portelas, Katja Hofmann, and Pierre-Yves Oudeyer. Socialai 0.1: Towards a benchmark to stimulate research on socio-cognitive abilities in deep reinforcement learning agents, 2021.

[201] Grgur Kovač, Rémy Portelas, P. Dominey, and Pierre-Yves Oudeyer. The socialai school: a framework leveraging developmental psychology toward artificial socio-cultural agents, 2024.

[202] Eyup Engin Kucuk and Muhammed Yusuf Kocyigit. Western, religious or spiritual: An evaluation of moral justification in large language models. *arXiv preprint arXiv:2311.07792*, 2023.

[203] Divyanshu Kumar, Umang Jain, Sahil Agarwal, and Prashanth Harshangi. Investigating implicit bias in large language models: A large-scale study of over 50 llms. *arXiv preprint arXiv:2410.12864*, 2024.

[204] Lucio La Cava and Andrea Tagarelli. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115*, 2024.

[205] John P Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 9–13, 2024.

[206] Randy J Larsen, David M Buss, Andreas Wismeijer, John Song, Stephanie Van den Berg, and Bertus Filippus Jeronimus. Personality psychology: Domains of knowledge about human nature. 2005.

[207] Bruce W Lee, Yeongheon Lee, and Hyunsoo Cho. Language models show stable value orientations across diverse role-plays. *arXiv preprint arXiv:2408.09049*, 2024a.

[208] Byung Cheol Lee and Jaeyeon Chung. An empirical investigation of the impact of chatgpt on creativity. *Nature Human Behaviour*, 8(10):1906–1914, 2024.

[209] Eun-Kyoung Rosa Lee, Sathvik Nair, and Naomi H. Feldman. A psycholinguistic evaluation of language models' sensitivity to argument roles, 2024b.

[210] Kibeom Lee and Michael C Ashton. Psychometric properties of the hexaco-100. *Assessment*, 25(5):543–556, 2018.

[211] Sanguk Lee, Tai-Quan Peng, Matthew H Goldberg, Seth A Rosenthal, John E Kotcher, Edward W Maibach, and Anthony Leiserowitz. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429, 2024c.

[212] Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beongwoo Kwak, Yeonsoo Lee, Dongha Lee, et al. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *arXiv preprint arXiv:2406.14703*, 2024d.

[213] Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666, 2024e.

[214] Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C Ong. Large language models produce responses perceived to be empathic. *arXiv preprint arXiv:2403.18148*, 2024f.

[215] Courtland Leer, Vincent Trost, and Vineeth Voruganti. Violation of expectation via metacognitive prompting reduces theory of mind prediction error in large language models, 2023.

[216] Yan Leng and Yuan Yuan. Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*, 2023.

[217] Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, and Sashank Varma. Incremental comprehension of garden-path sentences by large language models: Semantic interpretation, syntactic re-analysis, and attention, 2024a.

[218] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*, 2023a.

[219] Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia P. Sycara. Theory of mind for multi-agent collaboration via large language models, 2023b.

[220] Jingxuan Li, Yuning Yang, Shengqi Yang, Yizhou Zhao, and Ying Nian Wu. Quantifying preferences of vision-language models via value decomposition in social media contexts. *arXiv preprint arXiv:2411.11479*, 2024b.

[221] Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. Big5-chat: Shaping llm personalities through training on human-grounded data. *arXiv preprint arXiv:2410.16491*, 2024c.

[222] Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*, 10, 2022a.

[223] Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. Evaluating psychological safety of large language models. *arXiv preprint arXiv:2212.10529*, 2022b.

[224] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*, 2024d.

[225] Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*, 2024e.

[226] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[227] Yuanzhi Liang, Linchao Zhu, and Yi Yang. Anteval: Evaluation of social interaction competencies in llm-driven agents, 2024.

[228] Wim BG Liebrand. The effect of social motives, communication and group size on behaviour in an n-person multi-stage mixed-motive game. *European journal of social psychology*, 14(3):239–264, 1984.

[229] Zizheng Lin, Chunkit Chan, Yangqiu Song, and Xin Liu. Constrained reasoning chains for enhancing theory-of-mind in large language models, 2024.

[230] Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. Computational language acquisition with theory of mind. In *The Eleventh International Conference on Learning Representations*, 2023.

[231] Jianzhi Liu, Hexiang Gu, Tianyu Zheng, Liuyu Xiang, Huijia Wu, Jie Fu, and Zhaofeng He. Dynamic generation of personalities with large language models. *arXiv preprint arXiv:2404.07084*, 2024a.

[232] Songyuan Liu, Ziyang Zhang, Runze Yan, Wei Wu, Carl Yang, and Jiaying Lu. Measuring spiritual values and bias of large language models. *arXiv preprint arXiv:2410.11647*, 2024b.

[233] Xuan Liu, Jie Zhang, Song Guo, Haoyang Shang, Chengxu Yang, and Quanyan Zhu. Exploring prosocial irrationality for llm agents: A social cognition view, 2024c.

[234] Xuelin Liu, Yanfei Zhu, Shucheng Zhu, Pengyuan Liu, Ying Liu, and Dong Yu. Evaluating moral beliefs across llms through a pluralistic framework. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4740–4760, 2024d.

[235] Yu Liu, Su Blodgett, Jackie Chi Kit Cheung, Vera Liao, Alexandra Olteanu, and Ziang Xiao. Ecbd: Evidence-centered benchmark design for nlp. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, 2024e.

[236] John C Loehlin. *Latent variable models: An introduction to factor, path, and structural equation analysis*. Psychology Press, 2004.

[237] Lea Löhn, Niklas Kiehne, Alexander Ljapunov, and Wolf-Tilo Balke. Is machine psychology here? on requirements for using human psychological tests on large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 230–242, 2024.

[238] Anthony A Long and David N Sedley. *The Hellenistic philosophers: Volume 2, Greek and Latin texts with notes and bibliography*. Cambridge University Press, 1987.

[239] Nunzio Lorè, A. Ilami, and Babak Heydari. Large model strategic thinking, small model efficiency: Transferring theory of mind in large language models, 2024.

[240] Yang Lu, Jordan Yu, and Shou-Hsuan Stephen Huang. Illuminating the black box: A psychometric investigation into the multifaceted nature of large language models. *arXiv preprint arXiv:2312.14202*, 2023.

[241] Yaojia Lv, Haojie Pan, Zekun Wang, Jiafeng Liang, Yuanxing Liu, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. Coggpt: Unleashing the power of cognitive dynamics on large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6074–6091, 2024.

[242] Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael Hedderich, Barbara Plank, and Frauke Kreuter. The potential and challenges of evaluating attitudes, opinions, and values in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, 2024a.

[243] Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Assenmacher. Algorithmic fidelity of large language models in generating synthetic german public opinions: A case study. *arXiv preprint arXiv:2412.13169*, 2024b.

[244] Huanhuan Ma, Haisong Gong, Xiaoyuan Yi, Xing Xie, and Dongkuan Xu. Leveraging implicit sentiments: Enhancing reliability and validity in psychological trait evaluation of llms. *arXiv preprint arXiv:2503.20182*, 2025.

[245] Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models, 2023.

[246] Olivia Macmillan-Scott and Mirco Musolesi. (ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255, 2024.

[247] Simon Malberg, Roman Poletukhin, Carolin M Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in llms. *arXiv preprint arXiv:2410.15413*, 2024.

[248] Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. Editing personality for large language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 241–254. Springer, 2024a.

[249] Yuanyuan Mao, Shuang Liu, Qin Ni, Xin Lin, and Liang He. A review on machine theory of mind, 2024b.

[250] Keith A Markus and Denny Borsboom. *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge, 2013.

[251] Giovanni Marraffini, Andrés Cotton, Noe Hsueh, Axel Fridman, Juan Wisznia, and Luciano Corro. The greatest good benchmark: Measuring llms' alignment with utilitarian moral dilemmas. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21950–21959, 2024.

[252] Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in ai agents: Technical challenges and open questions, 2024.

[253] Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*, 2024.

[254] Gwenyth Isobel Meadows, Nicholas Wai Long Lau, Eva Adelina Susanto, Chi Lok Yu, and Aditya Paul. Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models. *arXiv preprint arXiv:2408.01460*, 2024.

[255] Mijntje Meijer, Hadi Mohammadi, and Ayoub Bagheri. Llms as mirrors of societal moral standards: reflection of cultural divergence and agreement across ethical topics. *arXiv preprint arXiv:2412.00962*, 2024.

[256] David M Messick and Charles G McClintock. Motivational bases of choice in experimental games. *Journal of experimental social psychology*, 4(1):1–25, 1968.

[257] Alessio Miaschi, Felice Dell'Orletta, and Giulia Venturi. Evaluating large language models via linguistic profiling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2835–2848, 2024.

[258] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334, 2023.

[259] Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. Who is gpt-3? an exploration of personality, values and demographics. In *Proceedings of the fifth workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pages 218–227. Association for Computational Linguistics, 2022.

[260] Robert J Mislevy, Russell G Almond, and Janice F Lukas. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1):i–29, 2003.

[261] J. Mittelstädt, Julia Maier, P. Goerke, Frank Zinn, and Michael Hermes. Large language models can outperform humans in social situational judgments, 2024.

[262] Shima Rahimi Moghaddam and C. Honey. Boosting theory-of-mind performance in large language models via prompting, 2023.

[263] Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval. *Advances in Neural Information Processing Systems*, 36:69736–69751, 2023.

[264] Jared Moore, Tanvi Deshpande, and Diyi Yang. Are large language models consistent over value-laden questions? *arXiv preprint arXiv:2407.02996*, 2024.

[265] Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios, 2024.

[266] Simon Münker. Towards" differential ai psychology" and in-context value-driven statement alignment with moral foundations theory. *arXiv preprint arXiv:2408.11415*, 2024.

[267] Ryan O Murphy, Kurt A Ackermann, and Michel JJ Handgraaf. Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781, 2011.

[268] Isabel Briggs Myers. *A guide to the development and use of the Myers-Briggs type indicator: Manual*. Consulting Psychologists Press, 1985.

[269] Isabel Briggs Myers et al. *The myers-briggs type indicator*, volume 34. Consulting Psychologists Press Palo Alto, CA, 1962.

[270] W Russell Neuman, Chad Coleman, and Manan Shah. Analyzing the ethical logic of six large language models. *arXiv preprint arXiv:2501.08951*, 2025.

[271] Allen Newell. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In William G. Chase, editor, *Visual Information Processing*, pages 283–308. Elsevier, 1973. doi: 10.1016/B978-0-12-170150-5.50012-3. URL https://doi.org/10.1016/B978-0-12-170150-5.50012-3.

[272] Christian Nickel, Laura Schrewe, and Lucie Flek. Probing the robustness of theory of mind in large language models, 2024.

[273] Animesh Nighojkar, Bekhzodbek Moydinboyev, My Duong, and John Licato. Giving ai personalities leads to more human-like reasoning. *arXiv preprint arXiv:2502.14155*, 2025.

[274] Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, and Ming Li. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges, 2024.

[275] José Luiz Nunes, Guilherme FCF Almeida, Marcelo de Araujo, and Simone DJ Barbosa. Are large language models moral hypocrites? a study based on moral foundations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1074–1087, 2024.

[276] Takumi Ohashi, Tsubasa Nakagawa, and Hitoshi Iyatomi. Extended japanese commonsense morality dataset with masked token and label enhancement. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3964–3968, 2024.

[277] OpenAI. Chatgpt, 2025. URL https://chat.openai.com.

[278] William Orwig, Emma R Edenbaum, Joshua D Greene, and Daniel L Schacter. The language of creativity: Evidence from humans and large language models. *The Journal of creative behavior*, 58(1):128–136, 2024.

[279] Samuel J Paech. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*, 2023.

[280] Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*, 2023.

[281] Luiz Pasquali. Psychometrics. *Revista da Escola de Enfermagem da USP*, 43:992–999, 2009.

[282] Suketu C Patel and Jin Fan. Identification and description of emotions by current large language models. *bioRxiv*, pages 2023–07, 2023.

[283] Delroy L Paulhus and Kevin M Williams. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of research in personality*, 36(6):556–563, 2002.

[284] Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*, 2024.

[285] Sanne Peereboom, Inga Schwabe, and Bennett Kleinberg. Cognitive phantoms in llms through the lens of latent variables. *arXiv preprint arXiv:2409.15324*, 2024.

[286] Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024.

[287] Yujia Peng, Jiaheng Han, Zhenliang Zhang, Lifeng Fan, Tengyu Liu, Siyuan Qi, Xue Feng, Yuxi Ma, Yizhou Wang, and Song-Chun Zhu. The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*, 34:12–22, 2024.

[288] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[289] Andrew Peterson. What does ai consider praiseworthy? *AI and Ethics*, pages 1–25, 2025.

[290] Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*, 2024.

[291] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[292] Zhiqiang Pi, Annapurna Vadaparty, Benjamin K. Bergen, and Cameron R. Jones. Dissecting the ullman variations with a scalpel: Why do llms fail at trivial alterations to the false belief task?, 2024.

[293] David J Pittenger. Cautionary comments regarding the myers-briggs type indicator. *Consulting Psychology Journal: Practice and Research*, 57(3):210, 2005.

[294] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, pages 34303–34326, 2024.

[295] Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text. *European Journal of Personality*, 34(5):885–902, 2020.

[296] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[297] Laura Pérez-Mayos, Alba T'aboas Garc'ia, Simon Mille, and L. Wanner. Assessing the syntactic capabilities of transformer-based multilingual language models, 2021.

[298] Weihong Qi, Hanjia Lyu, and Jiebo Luo. Representation bias in political sample simulations with large language models. *arXiv preprint arXiv:2407.11409*, 2024.

[299] Zhuang Qiu, Xufeng Duan, and Zhenguang G. Cai. Evaluating grammatical well-formedness in large language models: A comparative study with human judgments, 2024.

[300] Pinaki Raj. A literature review on emotional intelligence of large language models (llms). *International Journal of Advanced Research in Computer Science*, 15(4), 2024.

[301] Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[302] Leonardo Ranaldi and Fabio Massimo Zanzotto. Hans, are you clever? clever hans effect analysis of neural systems. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (* SEM 2024)*, pages 314–325, 2024.

[303] Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*, 2024.

[304] Haocong Rao, Cyril Leung, and Chunyan Miao. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248*, 2023.

[305] Tenko Raykov and George A Marcoulides. *Introduction to psychometric theory*. Routledge, 2011.

[306] Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2040, 2024.

[307] Maor Reuben, Ortal Slobodin, Aviad Elyshar, Idan-Chaim Cohen, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. *arXiv preprint arXiv:2409.19655*, 2024.

[308] Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models, 2025. URL https://arxiv.org/abs/2412.19726.

[309] Giuseppe Riva, F. Mantovani, B. Wiederhold, Antonella Marchetti, and A. Gaggioli. Psychomatics - a multidisciplinary framework for understanding artificial minds, 2024.

[310] Sergey Rodionov, Zarathustra Amadeus Goertzel, and Ben Goertzel. An evaluation of gpt-4 on the ethics dataset. *arXiv preprint arXiv:2309.10492*, 2023.

[311] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

[312] Peter Romero, Stephen Fitz, and Teruo Nakatsuma. Do gpt language models suffer from split personality disorder? the advent of substrate-free psychometrics. *arXiv preprint arXiv:2408.07377*, 2024.

[313] Ira J Roseman and Craig A Smith. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research*, pages 3–19, 2001.

[314] Hannes Rosenbusch, Claire E Stevenson, and Han LJ van der Maas. How accurate are gpt-3's hypotheses about social science phenomena? *Digital Society*, 2(2):26, 2023.

[315] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*, 2024.

[316] Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. Issuebench: Millions of realistic prompts for measuring issue bias in llm writing assistance. *arXiv preprint arXiv:2502.08395*, 2025.

[317] David Rozado. The political biases of chatgpt. *Social Sciences*, 12(3):148, 2023.

[318] Naama Rozen, Liat Bezalel, Gal Elidan, Amir Globerson, and Ella Daniel. Do llms have consistent values? *arXiv preprint arXiv:2407.12878*, 2024.

[319] Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36:20827–20905, 2023.

[320] John Rust and Susan Golombok. *Modern psychometrics: The science of psychological assessment*. Routledge, 2014.

[321] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. The self-perception and political biases of chatgpt. *Human Behavior & Emerging Technologies*, 2024.

[322] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 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)*, pages 5986–6004, 2024.

[323] Pratik S Sachdeva and Tom van Nuenen. Normative evaluation of large language models with everyday moral dilemmas. *arXiv preprint arXiv:2501.18081*, 2025.

[324] Jayanta Sadhu, Ayan Antik Khan, Noshin Nawal, Sanju Basak, Abhik Bhattacharjee, and Rifat Shahriyar. Multi-tom: Evaluating multilingual theory of mind capabilities in large language models, 2024.

[325] Payam Saeedi, Mahsa Goodarzi, and Muhammed Abdullah Canbaz. Heuristics and biases in ai decision-making: Implications for responsible agi. 2024. URL https://api.semanticscholar.org/CorpusID:273163070.

[326] Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*, 2024.

[327] Peter Salovey and John D Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1990.

[328] Nathan E Sanders, Alex Ulinich, and Bruce Schneier. Demonstrations of the potential of ai-based political issue polling. *arXiv preprint arXiv:2307.04781*, 2023.

[329] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[330] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms, 2022.

[331] Sneheel Sarangi, Maha Elgarf, and Hanan Salam. Decompose-tom: Enhancing theory of mind reasoning in large language models through simulation and task decomposition, 2025.

[332] Karahan Sarıtaş, Kıvanç Tezören, and Yavuz Durmazkeser. A systematic review on the evaluation of large language models in theory of mind tasks. *arXiv preprint arXiv:2502.08796*, 2025.

[333] Giuseppe Sartori and G. Orrú. Language models and psychological sciences, 2023.

[334] Kristina Schaaff, Caroline Reinig, and Tim Schlippe. Exploring chatgpt's empathic abilities. In *2023 11th international conference on affective computing and intelligent interaction (ACII)*, pages 1–8. IEEE, 2023.

[335] Julian Schelb, Orr Borin, David Garcia, and Andreas Spitz. Ru psycho? robust unified psychometric testing of language models. *arXiv preprint arXiv:2503.10229*, 2025.

[336] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809, 2023.

[337] Florian Scholten, Tobias R Rebholz, and Mandy Hütter. Metacognitive myopia in large language models. *arXiv preprint arXiv:2408.05568*, 2024.

[338] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, pages 1–11, 2025.

[339] Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier, 1992.

[340] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

[341] Shalom H Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, 32(5):519–542, 2001.

[342] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663, 2012.

[343] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker, 2023.

[344] Melanie Sclar, Jane Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. Explore theory of mind: Program-guided adversarial data generation for theory of mind reasoning, 2024.

[345] SM Seals and Valerie Shalin. Long-form analogies generated by chatgpt lack human-like psycholinguistic properties. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

[346] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

[347] Avner Seror. The moral mind (s) of large language models. *arXiv preprint arXiv:2412.04476*, 2024.

[348] Raj Sanjay Shah, Khushi Bhardwaj, and Sashank Varma. Development of cognitive intelligence in pre-trained language models. *arXiv preprint arXiv:2407.01047*, 2024.

[349] Ammar Shaikh, Raj Abhijit Dandekar, Sreedath Panat, and Rajat Dandekar. Cbeval: A framework for evaluating and interpreting cognitive biases in llms. *arXiv preprint arXiv:2412.03605*, 2024.

[350] Natalie Shapira, Mosh Levy, S. Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models, 2023a.

[351] Natalie Shapira, Guy Zwirn, and Yoav Goldberg. How well do large language models perform on faux pas tests?, 2023b.

[352] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the web conference 2021*, pages 194–205, 2021.

[353] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. Valuecompass: A framework of fundamental values for human-ai alignment. *arXiv preprint arXiv:2409.09586*, 2024.

[354] Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind, 2025.

[355] Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, 2024.

[356] David Silver and Richard S Sutton. Welcome to the era of experience. *Google AI*, 2025.

[357] Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, 2023.

[358] Herbert A. Simon. A theory of emotional behaviour. Working Paper 55, CIP (Complex Information Processing Project), 1963. URL https://digitalcollections.library.cmu.edu/node/1087.

[359] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[360] Ewa Skimina, Jan Cieciuch, and William Revelle. Between-and within-person structures of value traits and value states: Four different structures, four different interpretations. *Journal of Personality*, 89(5):951–969, 2021.

[361] Wei Song, Yadong Li, Jianhua Xu, Guowei Wu, Lingfeng Ming, Kexin Yi, Weihua Luo, Houyi Li, Yi Du, Fangda Guo, and Kaicheng yu. M3gia: A cognition inspired multilingual and multimodal general intelligence ability benchmark, 2024a.

[362] Xiaoyang Song, Akshat Gupta, Kiyan Mohebbizadeh, Shujie Hu, and Anant Singh. Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms. *arXiv preprint arXiv:2305.14693*, 2023.

[363] Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. Identifying multiple personalities in large language models with external evaluation. *arXiv preprint arXiv:2402.14805*, 2024b.

[364] Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024a.

[365] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302, 2024b.

[366] Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. *arXiv preprint arXiv:2503.15484*, 2025.

[367] Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597, 2024.

[368] Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117, 2017.

[369] Ondrej Sotolar, Vojtech Formanek, Alok Debnath, Allison Lahnala, Charles Welch, and Lucie FLek. Empo: Emotion grounding for empathetic response generation through preference optimization. *arXiv preprint arXiv:2406.19071*, 2024.

[370] Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, S. Mirroshandel, and Owen Rambow. Views are my own, but also yours: Benchmarking theory of mind using common ground, 2024.

[371] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2022.

[372] Julius Steuer, Marius Mosbach, and D. Klakow. Large gpt-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures, 2023.

[373] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*, 2022.

[374] James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, E. Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans, 2024a.

[375] James W. A. Strachan, Oriana Pansardi, E. Scaliti, Marco Celotto, Krati Saxena, Chunzhi Yi, Fabio Manzi, Alessandro Rufo, Guido Manzi, Michael S. A. Graziano, Stefano Panzeri, and Cristina Becchio. Gpt-4o reads the mind in the eyes, 2024b.

[376] Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, B. A. Y. Arcas, and Robin I. M. Dunbar. Llms achieve adult human performance on higher-order theory of mind tasks, 2024.

[377] Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*, 2025.

[378] Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. Challenging the validity of personality tests for large language models. *Preprint at arXiv. arXiv-2311 https://doi. org/10.48550/arXiv*, 2311, 2023.

[379] Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive biases in large language models: A survey and mitigation experiments. *arXiv preprint arXiv:2412.00323*, 2024.

[380] Kun Sun and Rong Wang. Computational sentence-level metrics predicting human sentence comprehension, 2024.

[381] Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J Jansen, and Jang Hyun Kim. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*, 2024.

[382] Sowmya S Sundaram and Balaji Alwar. Can a hallucinating model help in reducing human" hallucination"? *arXiv preprint arXiv:2405.00843*, 2024.

[383] Kazuhiro Takemoto. The moral machine experiment on large language models. *Royal Society open science*, 11 (2):231393, 2024.

[384] Masashi Takeshita, Rafal Rzpeka, and Kenji Araki. Jcommonsensemorality: Japanese dataset for evaluating commonsense morality understanding. In *In Proceedings of The Twenty Nineth Annual Meeting of The Association for Natural Language Processing (NLP2023)*, pages 357–362, 2023.

[385] Alaina N Talboy and Elizabeth Fuller. Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption. *arXiv preprint arXiv:2304.01358*, 2023.

[386] Fiona Anting Tan, Gerard Christopher Yeo, Kokil Jaidka, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Yang Liu, and See-Kiong Ng. Phantom: Persona-based prompting has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*, 2024.

[387] Weizhi Tang and Vaishak Belle. Tom-lm: Delegating theory of mind reasoning to external symbolic executors in large language models, 2024.

[388] Zhisheng Tang and Mayank Kejriwal. Humanlike cognitive patterns as emergent phenomena in large language models. *arXiv preprint arXiv:2412.15501*, 2024.

[389] Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. Probing the moral development of large language models through defining issues test. *arXiv preprint arXiv:2309.13356*, 2023.

[390] Edward Lee Thorndike. *Psychology and the science of education: Selected writings of Edward L. Thorndike*. Bureau of Publications, Teachers College, Columbia University, 1962.

[391] Robert L Thorndike and Saul Stein. An evaluation of the attempts to measure social intelligence. *Psychological bulletin*, 34(5):275, 1937.

[392] David Thorstad. Cognitive bias in large language models: Cautious optimism meets anti-panglossian meliorism. *arXiv preprint arXiv:2311.10932*, 2023.

[393] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, 2024.

[394] Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

[395] Alejandro Tlaie. Exploring and steering the moral compass of large language models. *arXiv preprint arXiv:2405.17345*, 2024.

[396] E Paul Torrance. Torrance tests of creative thinking. *Educational and psychological measurement*, 1966.

[397] Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025.

[398] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

[399] T. Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023.

[400] Max J. van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, M. Spruit, and P. V. D. Putten. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests, 2023.

[401] LaurÃĽne Vaugrante, Mathias Niepert, and Thilo Hagendorff. A looming replication crisis in evaluating behavior in language models? evidence and solutions. *arXiv preprint arXiv:2409.20303*, 2024.

[402] Karina Vida, Fabian Damken, and Anne Lauscher. Decoding multilingual moral preferences: Unveiling llm's biases through the moral machine experiment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1490–1501, 2024.

[403] Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. Vox populi, vox ai? using language models to estimate german public opinion. *arXiv preprint arXiv:2407.08563*, 2024.

[404] Huy Vu, Huy Anh Nguyen, Adithya V Ganesan, Swanie Juhng, Oscar NE Kjell, Joao Sedoc, Margaret L Kern, Ryan L Boyd, Lyle Ungar, H Andrew Schwartz, et al. Psychadapter: Adapting llm transformers to reflect traits, personality and mental health. *arXiv preprint arXiv:2412.16882*, 2024.

[405] Gleb D Vzorinab, Alexey M Bukinichac, Anna V Sedykha, Irina I Vetrovab, and Elena A Sergienkob. The emotional intelligence of the gpt-4 large language model. *Psychology in Russia: State of the art*, 17(2):85–99, 2024.

[406] Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561*, 2025.

[407] Yixin Wan and Kai-Wei Chang. White men lead, black women help? benchmarking language agency social biases in llms. *arXiv preprint arXiv:2404.10508*, 2024.

[408] Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. Towards objectively benchmarking social intelligence of language agents at the action level. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8885–8897, 2024a.

[409] Daphne Wang, M. Sadrzadeh, Milos Stanojevic, Wing-Yee Chow, and Richard Breheny. How can large language models become more human?, 2024b.

[410] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Large search model: Redefining search stack in the era of llms. In *ACM SIGIR Forum*, volume 57, pages 1–16. ACM New York, NY, USA, 2024c.

[411] Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, 2024d.

[412] Shuo Wang, Renhao Li, Xi Chen, Yulin Yuan, Derek F Wong, and Min Yang. Exploring the impact of personality traits on llm bias and toxicity. *arXiv preprint arXiv:2502.12566*, 2025.

[413] Xinglin Wang, Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. Coglm: Tracking cognitive development of large language models, 2024e.

[414] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, 2024f.

[415] Xiting Wang, Liming Jiang, Jose Hernandez-Orallo, David Stillwell, Luning Sun, Fang Luo, and Xing Xie. Evaluating general-purpose ai with psychometrics. *arXiv preprint arXiv:2310.16379*, 2023a.

[416] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, 2023b.

[417] Taylor W. Webb, K. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models, 2022.

[418] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[419] Anuradha Welivita and Pearl Pu. Are large language models more empathetic than humans? *arXiv preprint arXiv:2406.05063*, 2024.

[420] Yuchen Wen, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. Evaluating implicit bias in large language models by attacking from a psychometric perspective. *arXiv preprint arXiv:2406.14023*, 2024a.

[421] Zhiyuan Wen, Yu Yang, Jiannong Cao, Haoming Sun, Ruosong Yang, and Shuaiqi Liu. Self-assessment, exhibition, and recognition: a review of personality in large language models. *arXiv preprint arXiv:2406.17624*, 2024b.

[422] Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. Controllm: Crafting diverse personalities for language models. *arXiv preprint arXiv:2402.10151*, 2024.

[423] Ethan Gotlieb Wilcox, P. Vani, and R. Levy. A targeted assessment of incremental processing in neural language models and humans, 2021.

[424] Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities, 2023.

[425] Michael Wolfman, Donald Dunagan, Jonathan Brennan, and John Hale. Hierarchical syntactic structure in human-like language models, 2024.

[426] Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. Revealing fine-grained values and opinions in large language models. *arXiv preprint arXiv:2406.19238*, 2024.

[427] Junjie Wu, Mo Yu, Lemao Liu, D. Yeung, and Jie Zhou. Understanding llms' fluid intelligence deficiency: An analysis of the arc task, 2025.

[428] Patrick Y Wu, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. Large language models can be used to estimate the latent positions of politicians. *arXiv preprint arXiv:2303.12057*, 2023.

[429] Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599, 2024.

[430] Wei Xie, Shuoyoucheng Ma, Zhenhua Wang, Enze Wang, Baosheng Wang, and Jinshu Su. Do large language models truly grasp mathematics? an empirical exploration. 2024a.

[431] Zhentao Xie, Jiabao Zhao, Yilei Wang, Jinxin Shi, Yanhong Bai, Xingjiao Wu, and Liang He. Mindscope: Exploring cognitive biases in large language models through multi-agent systems. *arXiv preprint arXiv:2410.04452*, 2024b.

[432] Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, et al. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*, 2023.

[433] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models, 2024a.

[434] Hainiu Xu, Siya Qi, Jiazheng Li, Yuxiang Zhou, Jinhua Du, C. Catmur, and Yulan He. Enigmatom: Improve llms' theory-of-mind reasoning capabilities with neural knowledge base of entity states, 2025a.

[435] Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. Academically intelligent llms are not necessarily socially intelligent, 2024b.

[436] Ruoxi Xu, Hongyu Lin, Xianpei Han, Jia Zheng, Weixiang Zhou, Le Sun, and Yingfei Sun. Large language models often say one thing and do another. *arXiv preprint arXiv:2503.07003*, 2025b.

[437] Shanshan Xu, TYS Santosh, Yanai Elazar, Quirin Vogel, Barbara Plank, and Matthias Grabmair. Better aligned with survey respondents or training data? unveiling political leanings of llms on us supreme court cases. *arXiv preprint arXiv:2502.18282*, 2025c.

[438] Bei Yan, Jie Zhang, Zhiyuan Chen, Shiguang Shan, and Xilin Chen. Moralbench: A multimodal moral benchmark for lvlms. *arXiv preprint arXiv:2412.20718*, 2024.

[439] Bo Yang, Jiaxian Guo, Yusuke Iwasawa, and Yutaka Matsuo. Large language models as theory of mind aware generative agents with counterfactual reflection, 2025a.

[440] Kaiqi Yang, Hang Li, Yucheng Chu, Yuping Lin, Tai-Quan Peng, and Hui Liu. Unpacking political bias in large language models: Insights across topic polarization. *arXiv preprint arXiv:2412.16746*, 2024.

[441] Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. Exploring the personality traits of llms through latent features steering, 2025b. URL https://arxiv.org/abs/2410.10863.

[442] Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8754–8777, 2024.

[443] Jing Yao, Xiaoyuan Yi, Shitong Duan, Jindong Wang, Yuzhuo Bai, Muhua Huang, Peng Zhang, Tun Lu, Zhicheng Dou, Maosong Sun, et al. Value compass leaderboard: A platform for fundamental and validated evaluation of llms values. *arXiv preprint arXiv:2501.07071*, 2025a.

[444] Jing Yao, Xiaoyuan Yi, and Xing Xie. Clave: An adaptive framework for evaluating values of llm generated responses. *Advances in Neural Information Processing Systems*, 37:58868–58900, 2025b.

[445] Shunyu Yao. The second half, 2025. URL https://ysymyth.github.io/The-Second-Half/.

[446] Nicolas Yax, Hernán Anlló, and Stefano Palminteri. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51, 2024.

[447] Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. Measuring human and ai values based on generative psychometrics with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025a.

[448] Haoran Ye, Tianze Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. Generative psycho-lexical approach for constructing value systems in large language models. *arXiv preprint arXiv:2502.02444*, 2025b.

[449] Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502*, 2025a.

[450] Zonghao Ying, Deyue Zhang, Zonglei Jing, Yisong Xiao, Quanchen Zou, Aishan Liu, Siyuan Liang, Xiangzheng Zhang, Xianglong Liu, and Dacheng Tao. Reasoning-augmented conversation for multi-turn jailbreak attacks on large language models. *arXiv preprint arXiv:2502.11054*, 2025b.

[451] Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues, 2025.

[452] Yiyao Yu, Junjie Wang, Yuxiang Zhang, Lin Zhang, Yujiu Yang, and Tetsuya Sakai. Ealm: Introducing multidimensional ethical alignment in conversational information retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 32–39, 2023.

[453] Jiaqing Yuan, Pradeep K Murukannaiah, and Munindar P Singh. Right vs. right: Can llms make tough choices? *arXiv preprint arXiv:2412.19926*, 2024.

[454] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence, 2019.

[455] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350, 2024a.

[456] Yifan Zeng. Quantifying risk propensities of large language models: Ethical focus and bias detection through role-play. *arXiv preprint arXiv:2411.08884*, 2024.

[457] Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. Persllm: A personified training approach for large language models. *arXiv preprint arXiv:2407.12393*, 2024b.

[458] Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, and Jiaya Jia. Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms, 2024c.

[459] Haoyang Zhang. Harnessing in-context learning for personality elicitation in large language models. In *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pages 1521–1531. IEEE, 2024.

[460] Jie Zhang, Dongrui Liu, Chen Qian, Ziyue Gan, Yong Liu, Yu Qiao, and Jing Shao. The better angels of machine personality: How personality relates to llm safety. *arXiv preprint arXiv:2407.12344*, 2024a.

[461] Xin Zhang, Yuanyi Ren, and Guojie Song. Age-related value orientations in large language models (llms). *Innovation in Aging*, 8(Supplement_1):1010–1010, 2024b.

[462] Zhaowei Zhang, Fengshuo Bai, Jun Gao, and Yaodong Yang. Valuedcg: Measuring comprehensive human value understanding ability of language models. *arXiv preprint arXiv:2310.00378*, 2023a.

[463] Zhaowei Zhang, Ceyao Zhang, Nian Liu, Siyuan Qi, Ziqi Rong, Song-Chun Zhu, Shuguang Cui, and Yaodong Yang. Heterogeneous value alignment evaluation for large language models. In *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*, 2024c.

[464] Zheyuan Zhang, Jifan Yu, Juanzi Li, and Lei Hou. Exploring the cognitive knowledge structure of large language models: An educational diagnostic assessment approach, 2023b.

[465] Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind, 2025.

[466] Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11157–11176, 2024.

[467] Yachao Zhao, Bo Wang, and Yan Wang. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. *arXiv preprint arXiv:2501.02295*, 2025a.

[468] Yue Zhao, Qingqing Gu, Xiaoyu Wang, Teng Chen, Zhonglin Jiang, Yong Chen, and Luo Ji. Fisminess: A finite state machine based paradigm for emotional support conversations. *arXiv preprint arXiv:2504.11837*, 2025b.

[469] Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. Lmlpa: Language model linguistic personality assessment. *Computational Linguistics*, pages 1–41, 2025.

[470] Qishuai Zhong, Yike Yun, and Aixin Sun. Cultural value differences of llms: Prompt, language, and model size. *arXiv preprint arXiv:2407.16891*, 2024.

[471] Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. Rethinking machine ethics–can llms perform moral reasoning through the lens of moral theories? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2227–2242, 2024a.

[472] Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E Prunty, et al. General scales unlock ai evaluation with explanatory and predictive power. *arXiv preprint arXiv:2503.06378*, 2025.

[473] Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, J. Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. How far are large language models from agents with theory-of-mind?, 2023a.

[474] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024b.

[475] Yuchen Zhou, Emmy Liu, Graham Neubig, and Leila Wehbe. Divergences between language models and human brains, 2023b.

[476] Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. Dynamic evaluation of large language models by meta probing agents. In *International Conference on Machine Learning*, pages 62599–62617. PMLR, 2024a.

[477] Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. *arXiv preprint arXiv:2408.11779*, 2024b.

[478] Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others, 2024c.

[479] Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. *CoRR*, 2023a.

[480] Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Zachary A Pardos, Patrick C Kyllonen, Jiyun Zu, Qingyang Mao, Rui Lv, Zhenya Huang, et al. From static benchmarks to adaptive testing: Psychometrics in ai evaluation. *arXiv preprint arXiv:2306.10512*, 2023b.

[481] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1), 2024.

[482] Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. Can llm" self-report"?: Evaluating the validity of self-report scales in measuring personality design in llm-based chatbots. *arXiv preprint arXiv:2412.00207*, 2024.