Measuring Human and AI Values Based on Generative Psychometrics with Large Language Models

Haoran Ye[*1], **Yuhang Xie**[*2], **Yuanyi Ren**[*1], **Hanjun Fang**[3], **Xin Zhang**[4], **Guojie Song**[†15]

[1]State Key Laboratory of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
[2]School of Software and Microelectronics, Peking University
[3]Department of Sociology, Peking University
[4]School of Psychological and Cognitive Sciences, Peking University
[5]PKU-Wuhan Institute for Artificial Intelligence
{hrye, yuhangxie}@stu.pku.edu.cn {yyren, hjfang, zhang.x, gjsong}@pku.edu.cn

Abstract

Human values and their measurement are long-standing interdisciplinary inquiry. Recent advances in AI have sparked renewed interest in this area, with large language models (LLMs) emerging as both tools and subjects of value measurement. This work introduces **G**enerative **P**sychometrics for **V**alues (GPV), an LLM-based, data-driven value measurement paradigm, theoretically grounded in text-revealed selective perceptions. The core idea is to dynamically parse unstructured texts into perceptions akin to static stimuli in traditional psychometrics, measure the value orientations they reveal, and aggregate the results. Applying GPV to human-authored blogs, we demonstrate its stability, validity, and superiority over prior psychological tools. Then, extending GPV to LLM value measurement, we advance the current art with 1) a psychometric methodology that measures LLM values based on their scalable and free-form outputs, enabling context-specific measurement; 2) a comparative analysis of measurement paradigms, indicating response biases of prior methods; and 3) an attempt to bridge LLM values and their safety, revealing the predictive power of different value systems and the impacts of various values on LLM safety. Through interdisciplinary efforts, we aim to leverage AI for next-generation psychometrics and psychometrics for value-aligned AI.

Code — https://github.com/Value4AI/gpv
Extended version — https://arxiv.org/abs/2409.12106

1 Introduction

Human values, a cornerstone of philosophical inquiry, are the fundamental guiding principles behind individual and collective decision-making (Rokeach 1973; Sagiv et al. 2017). Value measurement is a long-standing interdisciplinary endeavor for elucidating how specific values underpin and justify the worth of actions, objects, and concepts (Schwartz 1992; Klingefjord, Lowe, and Edelman 2024).

Traditional psychometrics often measure human values through self-report questionnaires, where participants rate the importance of various values in their lives. However, these tools are limited by response biases, resource demands, inaccuracies in capturing authentic behaviors, and inability to handle historical or subjective data (Ponizovskiy et al. 2020). Therefore, data-driven tools have been developed to infer values from textual data, such as social media posts (Shen, Wilson, and Mihalcea 2019). These tools can reveal personal values without relying on explicit self-reporting, but they are mostly dictionary-based, matching text to predefined value lexicons. Consequently, they often fail to grasp the nuanced semantics and context-dependent value expressions. Additionally, these tools are inherently static and inflexible, relying on expert-defined lexicons that are not easily adaptable to new or evolving values.

The rise of large language models (LLMs), with their remarkable ability to understand semantic nuances, presents new possibilities for data-driven value measurement. Recent studies have demonstrated that LLMs can effectively approximate annotators' and even psychologists' judgments on value-related tasks (Sorensen et al. 2024; Ren et al. 2024). Building on these advancements, this work introduces Generative Psychometrics for Values (GPV), an LLM-based, data-driven value measurement paradigm grounded in the theory of text-revealed selective perceptions (Postman, Bruner, and McGinnies 1948; Shen, Wilson, and Mihalcea 2019). Perceptions are the way individuals interpret and evaluate the world around them, and are servants of interests, needs, and, values (Postman, Bruner, and McGinnies 1948). Such perceptions are revealed in self-expressing texts, such as blog posts, and are utilized as atomic value measurement units in GPV. The core idea of GPV is to extract contextualized and value-laden perceptions (e.g., "I believe that everyone deserves equal rights and opportunities.") from unstructured texts, decode underlying values (e.g., Universalism) for arbitrary value systems, and aggregate the results to measure individual values.

The perceptions in GPV function similarly to the static psychometric items (stimuli) in self-report questionnaires, which support or oppose specific values (Schwartz 1992).

[*]Equal contribution.
[†]Corresponding author.

Notably, GPV enables the automatic generation of such items and their adaptation to any given data, overcoming the limitations of traditional tools (Fig. 1). By applying GPV to a large collection of human-authored blogs, we evaluate GPV against psychometric standards. GPV demonstrates its stability and validity in measuring individual values, and its superiority over prior psychological tools.

Meanwhile, the rapid evolution of LLMs raises significant concerns about their potential misalignment with human values. Recent literature treats LLMs as subjects of value measurement (Ma et al. 2024), employing self-report questionnaires (Huang et al. 2024) or their variants (Ren et al. 2024). However, these tools are inherently static, inflexible, and unscalable, as they rely on closed-ended questions derived from limited psychometric inventories.

To address these limitations, we extend the GPV paradigm to LLMs. Experimenting across 17 LLMs and 4 value theories, we advance the current art of LLM value measurement in several aspects. Firstly, GPV constitutes a novel evaluation methodology that does not rely on static psychometric inventories but measures LLM values based on their scalable and free-form outputs. In this manner, we mitigate response bias demonstrated in prior tools and enable context-specific value measurements. Secondly, we conduct the first comparative analysis of different measurement paradigms, where GPV yields better measurement results regarding validity and utility. Lastly, we present novel findings regarding value systems and LLM values. Despite the popularity of Schwartz's value theory within the AI community, alternative value systems like VSM (Hofstede 2011) indicate better predictive power. In addition, values like Long Term Orientation positively contribute to the predicted safety scores, while values like Masculinity negatively contribute.

Below we summarize our contributions:

- We introduce Generative Psychometrics for Values (GPV), a novel LLM-based value measurement paradigm grounded in text-revealed selective perceptions (§ 3).
- Applying GPV to human-authored blogs, we demonstrate its stability, validity, and superiority over prior psychological tools (§ 4).
- Applying GPV to LLMs, we enable LLM value measurements based on their scalable, free-form, and context-specific outputs. With extensive evaluations across 17 LLMs, 4 value theories, and 3 measurement tools, we illustrate the superiority of GPV and uncover novel insights regarding value systems and LLM values (§ 5).

2 Related Work

For the complete reference list, please refer to the extended version of this paper.

2.1 Value Measurements for Human

The measurement of individual values is pivotal in elucidating the driving forces and mechanisms underlying human behavior (Schwartz 1992; Rokeach 1973). Due to the intricate relationship between behavior and values, researchers have developed different measurement methods, including self-report questionnaires (Maio 2010), behavioral observation (Fischer and Schwartz 2011), and experimental techniques (Murphy and Ackermann 2014). Self-report methods involve participants themselves assessing their agreement with descriptions (Sagiv, Sverdlik, and Schwarz 2011) or ranking the importance of items (Rokeach 1973). Behavioral observation methods require experts to analyze how personal values manifest in real-life actions (Bardi and Schwartz 2003; Schwartz and Butenko 2014). Furthermore, experimental methods employ structured scenarios to isolate and analyze variables affecting human behavior (Bekkers 2007). However, these methods are hindered by response biases, resource demands, inaccuracies in capturing authentic behaviors, and inability to handle historical or subjective data (Ponizovskiy et al. 2020).

On the other hand, data-driven tools partially address the adverse effects of resource costs, external interference, and response biases. Among them, dictionary-based tools such as LIWC dictionary (Graham, Haidt, and Nosek 2009) and personal values dictionary (PVD) (Ponizovskiy et al. 2020) analyze the frequency of value-related lexicons, flawed for overlooking nuanced semantics and contexts. Recent efforts to train deep learning models for value identification have largely focused on Schwartz's values and are not validated for individual-level measurements (Sorensen et al. 2024). Other works transform self-report inventories into interactive assessments based on LLMs (Li et al. 2024b), yet inherit many of the limitations of self-reports, such as the inability to handle historical or subjective data.

2.2 Value Measurements for LLMs

The growing integration of LLMs into public-facing applications necessitates their comprehensive and reliable value measurements (Ma et al. 2024). Recently, applying psychometrics—originally designed for humans—to LLMs has gained significant interest (Pellert et al. 2023; Jiang et al. 2024). Related works involve psychometric tests such as the "dark triad" traits (Li et al. 2024c; Huang et al. 2024), the Big Five Inventory (BFI) (Safdari et al. 2023), Myers–Briggs Type Indicator (MBTI) (Pan and Zeng 2023), and morality inventories (Scherrer et al. 2023b). The test results are utilized to investigate the attributes of LLMs concerning political positions (Santurkar et al. 2023), cultural differences (Cao et al. 2023), and belief systems (Scherrer et al. 2023a).

However, researchers have observed discrepancies between constrained and free-form LLM responses, and the latter is considered more practically relevant (Röttger et al. 2024; Ren et al. 2024). The variability in LLM responses to subtle contextual changes also necessitates scalable and context-specific evaluation methods (Röttger et al. 2024), which this work aims to address.

3 Generative Psychometrics for Values (GPV)

3.1 Value Measurement Based on Selective Perceptions

Values are broad motivational goals and guiding principles in life (Schwartz 1992). Value measurement quantita-

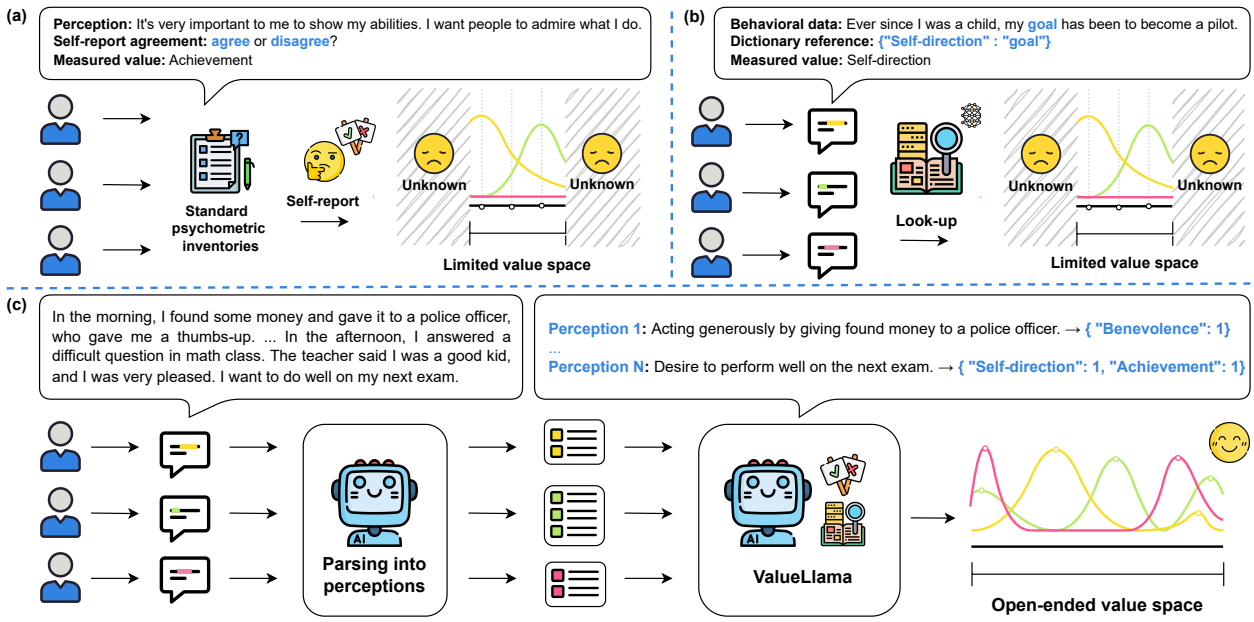


Figure 1: Illustrations of the three measurement paradigms. (a) Self-reports require individuals to rate their agreement with expert-defined perceptions. (b) Dictionary-based methods count expert-defined and value-related lexicons given text data. (c) GPV automatically and dynamically extracts perceptions from text data and learns to measure open-vocabulary values.

tively evaluates the significance attributed to various values through individuals' behavioral and linguistic data (Adkins, Russell, and WERBEL 1994; Meglino and Ravlin 1998; Rokeach 1973). Given any pluralistic value system as a reference frame, we formalize the value measurement task as follows.

Definition 1 (Value Measurement)**.** Value measurement is a function f:

$$f : (V, D) \rightarrow \mathbf{w} \in \mathbb{R}^n. \quad (1)$$

Here, $V = \{v_1, v_2, \ldots, v_n\}$ denotes a value system, where each v_i represents a particular value dimension; D denotes the individuals' behavioral and linguistic data; and $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ is a value vector with w_i indicating the relative importance of v_i.

Extensive research explores the underlying mechanisms of f, by which human values drive behaviors and behaviors reflect values (Adkins, Russell, and WERBEL 1994; Meglino and Ravlin 1998; Schwartz 1992; Rokeach 1973). Most related to this work, self-reports (Fig. 1(a)) instantiate f by self-rating the agreement with expert-defined items; dictionary-based methods (Fig. 1(b)) instantiate f by counting expert-defined and value-related lexicons. Both tools conduct value measurement in a limited value space (e.g. 10 Schwartz's values define a limited 10-dimensional value space) and are inherently static and inflexible.

GPV Overview. In contrast, GPV (Fig. 1(c)) instantiates f through selective perceptions, a process of selecting stimuli from the environment based on an individual's interests, needs, and values (Postman, Bruner, and McGinnies 1948; Anderson 2019). For example, when considering a construction project of a new park, individuals who value Hedonism will emphasize the recreational benefits, while those who prioritize Economic Efficiency will focus on the project's cost. These differing perceptions encode value orientations. GPV leverages LLMs to automatically parse self-expressing texts into such perceptions, trains an LLM for perception-level and open-vocabulary value measurement, and aggregates the results as individual values. We elaborate on the perception-level value measurement in § 3.2, then parsing and aggregation in § 3.3.

3.2 Perception-Level Value Measurement

Perception. Perceptions are selective stimuli from the environment that reflect an individual's interests, needs, and values (Postman, Bruner, and McGinnies 1948). Here, perceptions are utilized as atomic measurement units, ideally capturing the following properties (Gibson 1960): (1) A perception should be value-laden and accurately describe the measurement subject, ensuring meaningful measurement. (2) A perception is an atomic measurement unit, ensuring unambiguous measurement. (3) A perception is well-contextualized and self-contained, ensuring that it alone is sufficient for value measurement. (4) All perceptions comprehensively cover all value-laden aspects of the measured subject, ensuring that no related content in the data is left unmeasured.

Training. We fine-tune Llama-3-8B (Dubey et al. 2024) for perception-level and open-vocabulary value measurement. Its fine-tuning involves the following two tasks

Model	Relevance	Valence
Kaleido	83.5%	82.5%
GPT-4 Turbo	79.8%	87.5%
ValueLlama (ours)	**90.0%**	**91.5%**

Table 1: Accuracy on relevance and valence classification.

(Sorensen et al. 2024) using datasets of ValueBench (Ren et al. 2024) and ValuePrism (Sorensen et al. 2024): (1) Relevance classification determines whether a perception is relevant to a value. (2) Valence classification determines whether a perception supports, opposes, or remains neutral (context-dependent) towards a value. Both tasks are formulated as generating a label given a value and a perception. We present further training details in Appendix A.

Inference. We refer to the fine-tuned Llama-3-8B as ValueLlama. Given a value system $V = \{v_1, v_2, \ldots, v_n\}$ and a sentence of perception s, we employ ValueLlama to calculate the relevance and valence probability distribution of each value v_i to s, respectively denoted as $p_{rel}(\cdot|v_i, s)$ and $p_{val}(\cdot|v_i, s)$. Then, we define w_i as $p_{val}(\text{support}|v_i, s) - p_{val}(\text{oppose}|v_i, s)$ if the value is relevant ($p_{rel}(\cdot|v_i, s) > 0.5$) and its valence is classified as "support" or "oppose". Otherwise, w_i is considered unmeasured. The prompts for inference are listed in Appendix A.

Evaluating Perception-level Value Measurements. To evaluate the accuracy of perception-level value measurements, we hold out 50 values and 200 associated items (146 with "Supports" valence and 54 with "Opposes" valence) from ValueBench as a test dataset, also ensuring the test values are not included in ValuePrism. Using the same zero-shot prompt, we measure the relevance and valence of the test items with Kaleido (Sorensen et al. 2024), GPT-4 Turbo (Achiam et al. 2023), and ValueLlama. Table 1 presents the comparison results, indicating that ValueLlama outperforms state-of-the-art general and task-specific LLMs in zero-shot perception-level value measurement.

3.3 Parsing and Aggregation

To measure values at the individual level, GPV chunks long texts (e.g., blog posts) into segments and prompts an LLM (this work used GPT-3.5 Turbo) to parse each segment into perceptions. Parsing is guided by the background on human values, definitions of perceptions, and few-shot examples (Appendix B.1.) Then, GPV performs perception-level value measurement (§ 3.2) for the parsing results. Individual-level measurements are calculated by averaging the perception-level measurements for each value (Schwartz et al. 2007).

Evaluating LLM Parsing. The parsing results are considered high-quality by trained human annotators. On average, the annotators agree that the parsing results meet the defined four criteria in over 85% of cases, deeming them suitable for further value measurement. The evaluation is detailed in Appendix B.2.

3.4 Discussion

Relation to Self-Reports. The items organized in self-report inventories are essentially perceptions that support or oppose specific values (Schwartz 1992). Compared to GPV, these traditional psychometric inventories compile static and unscalable perceptions, covering a limited measurement range. They also necessitate an additional self-report process to assess the individual's agreement with the items.

Relation to Dictionary-Based Methods. Both GPV and dictionary-based methods share the fundamental principle that values are embedded in language (Shen, Wilson, and Mihalcea 2019), and they each measure values through text data. However, dictionary-based methods depend on predefined lexicons for closed-vocabulary values and are far less expressive than GPV in capturing semantic nuances. Further analysis is presented in § 4.2.

Advantages of GPV. Compared with traditional tools, GPV 1) effectively mitigates response bias and resource demands by dispensing with self-reports; 2) captures authentic behaviors instead of relying on forced ratings; 3) can handle historical or subjective data; 4) measures values in open-ended value spaces and easily adapts to new or evolving values without expert effort; and 5) enables more scalable and flexible value measurement.

4 GPV for Humans

This section measures human values using 791 blogs from the Blog Authorship Corpus (Schler et al. 2006), selected after filtering out low-quality entries (Appendix C.1). We evaluate GPV using standard psychological metrics including stability, construct validity, concurrent validity, and predictive validity, and demonstrate its superiority over established psychological tools.

4.1 Validation

Stability. As values are relatively stable psychological constructs for humans (Sagiv and Roccas 2017; Sagiv et al. 2017; Kimura 2023), we expect that the same individual should exhibit consistent value tendencies across different scenarios. Across 48,888 perception-value pairs, 86.6% of the perception-level measurement results are consistent with the individual-level aggregated results, indicating desirable stability. Detailed results and extended discussions are shown in Appendix C.2.

Construct Validity. Construct validity is the extent to which a test measures what it claims to measure. In Schwartz's value system, some values are theoretically positively correlated, such as Self-Direction and Stimulation, while others are negatively correlated, such as Power and Benevolence. Altogether, the 10 Schwartz values form a circumplex structure (Schwartz and Bilsky 1990), where values that are closer together are more compatible, while those that are farther apart are more conflicting (Fig. 2a). We employ multidimensional scaling (MDS) (Cieciuch and Schwartz 2012; Bilsky, Janik, and Schwartz 2011) on the value correlations obtained by GPV, and project both the 10 basic values and the 4 higher-order values onto two-dimensional

		GPV			
		Stran	Cons	Open	Senh
PVD	Stran	**0.0421**	0.0077	-0.0318	**-0.0579**
	Cons	-0.0530	**0.0687**	0.0290	-0.0321
	Open	-0.0345	**-0.1376**	**0.0369**	-0.0615
	Senh	**-0.0693**	-0.0345	0.0540	**0.0880**

Table 2: Correlations between the measurement results of PVD and GPV for four high-level values: Self-transcendence (Stran), Conservation (Cons), Openness to Change (Open), and Self-enhancement (Senh).

MDS plots. Then, we assess whether their relative positions align with the theoretical structure. As illustrated in Fig. 2, basic values of the same category (represented by the same color) generally cluster together. Higher-order opposing values are positioned farther apart. The relative positions of a few values do not strictly follow the theoretical structure. For example, Conservation is relatively distant from the other three higher-order values. Such deviations may reflect a gap between the values manifested by self-report and objective data (Ponizovskiy et al. 2020). Overall, the relative positioning of most values resembles the theoretically expected pattern in Fig. 2a, indicating desirable construct validity. More experimental details are provided in Appendix C.3.

Concurrent Validity. Concurrent validity is the extent to which a test correlates with other measures of the same construct administered simultaneously. Theoretically expected correlations can validate newly developed instruments (Lin and Yao 2024). We evaluate the concurrent validity of GPV by comparing it with the personal values dictionary (PVD) (Ponizovskiy et al. 2020), a well-established measurement tool with proven reliability and validity. We analyze the correlations between GPV and PVD measurements, with the results of low-level values presented in Appendix C.4 and high-level aggregated values in Table 2. The results indicate that among the 10 basic values, both identical values (e.g., SE-SE) and most compatible values (e.g., CO-SE) show positive correlations; most opposing values (e.g., BE-AC) exhibit negative correlations. Similarly, within the 4 higher-order values, positive correlations are observed when measuring identical values, whereas most opposing values display negative correlations. These correlations, though not strong, are theoretically expected, which supports the concurrent validity of GPV. § 4.2 exemplifies the cases where GPV misaligns with PVD.

Predictive Validity. Predictive validity is the extent to which a test predicts future behavior or outcomes. We assess predictive validity by examining if our measurement results align with the blog authors' gender-related sociodemographic traits. Previous research indicates that, in a statistical sense, men prioritize power, stimulation, hedonism, achievement, and self-direction, while women emphasize benevolence and universalism (Schwartz and Rubel 2005). Our measurement results, presented in Table 3, reveal that men and women score higher on the values they typically prioritize, confirming the consistency of our measurements with established psychological findings.

4.2 Case Study

We exemplify the advantage of GPV over prior data-driven tools such as PVD in Fig. 3. Some values, while not explicitly mentioned in PVD-designed lexicons, are implied within the text. For example, in Schwartz's theory, Achievement is defined as "the personal pursuit of success, demonstrating competence according to social standards." In this context, "the teacher's praise" and "performing well in an exam" both embody the "success" element of achievement. Although the text does not directly reference Achievement or Achievement-related lexicons, the author's expression of joy and aspiration for these outcomes reflects this value. While GPV effectively captures this aspect, PVD does not.

Some PVD-designed lexicons fail to align with the measurement subject or reflect their intended values. For instance, "friendly" and "goal" target the author's deskmate; picking up "money" does not indicate the author's own values of Power. GPV effectively avoids such misinterpretation.

5 GPV for Large Language Models

We evaluate 17 LLMs across 4 value systems using 3 measurement tools: self-report questionnaires (Huang et al. 2024), ValueBench (Ren et al. 2024), and GPV. Unless otherwise specified, we use LLM-generated value-eliciting questions for GPV to ensure a comprehensive and thorough measurement of each value. The detailed experimental setup is described in Appendix D.1.

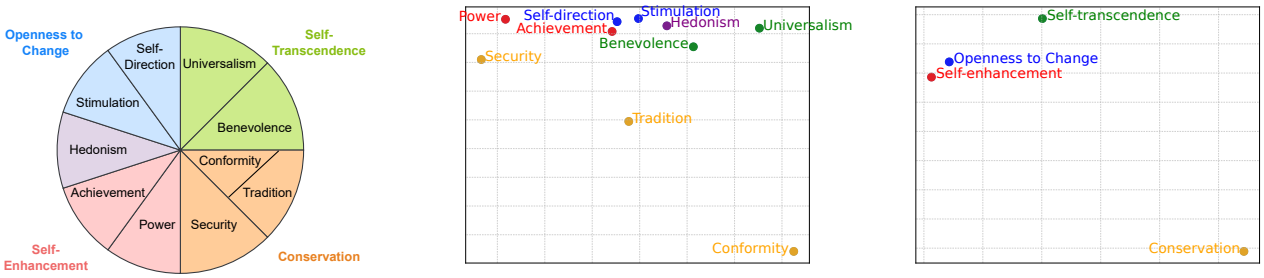
Across 19910 perception-value pairs, 86.8% perception-level measurement results are consistent with the LLM-level aggregated results, indicating desirable stability; we present the detailed results in Appendix D.2.

This section focuses on comparing GPV against prior measurement tools. We defer the value measurement results of all LLMs to Appendix D.4.

5.1 Comparative Analysis of Construct Validity

Using the measurement results from 17 LLMs as data points, we compute correlations between Schwartz's values. The results are visualized in a heatmap for each measurement tool in Fig. 4. The heatmap reveals the superior construct validity of GPV, as its measurement results align more closely with the theoretical structure (Fig. 2a). Specifically, values that are adjacent in the theoretical circumplex structure exhibit positive correlations, while those that are theoretically distant show negative correlations.

In contrast, prior tools obtain almost all-positive correlations, contrary to theoretical expectations. This discrepancy indicates their strong susceptibility to response biases, wherein certain LLMs generally tend to assign higher scores in self-report or respond more supportively in ValueBench. Such biases obscure the genuine value orientations of the LLMs. Even when centering the measurement results of prior tools (Appendix D.3), the correlation results remain inconsistent with the theoretical structure. This finding aligns with recent studies revealing the unreliability of



(a) Theoretical structure. (b) MDS of 10 basic values. (c) MDS of 4 high-level values.

Figure 2: Two-dimensional MDS of individual values measured by GPV.

Gender	SE	CO	TR	BE	UN	SD	ST	HE	AC	PO
Male	0.478	-0.424	0.261	0.691	0.593	**0.777**	**0.797**	**0.745**	**0.757**	**0.626**
Female	0.459	-0.414	0.214	**0.751**	**0.649**	0.748	0.761	0.736	0.725	0.587

Table 3: GPV measurement results on Schwartz values for male and female groups.

Value Pair	Self-report	ValueBench	**GPV**
UA & DA	0.09	-0.17	0.65
Indv & SD	0.21	-0.38	0.61
Indu & He	0.30	-0.30	0.65
CO & Be	0.22	0.79	0.38
Avg.	0.21	-0.01	0.57

Table 4: Correlation between theoretically positively correlated values when using different tools, including Uncertainty Avoidance (UA) & Discomfort with Ambiguity (DA), Individualism (Indv) & Self-Direction (SD), Indulgence (Indu) & Hedonism (He), and Concern for Others (CO) & Benevolence (Be).

Tools	Acc. (%)
Self-report	56.7 $\pm$ 26.0
ValueBench	67.8 $\pm$ 20.6
GPV	**85.6** $\pm$ 14.1

Table 5: Classification accuracy when using linear probing for value measurement results.

LLMs as survey respondents (Dominguez-Olmedo, Hardt, and Mendler-Dünner 2023; Röttger et al. 2024).

Besides Schwartz's value system, we also evaluate the construct validity by relating the values of different value theories that are theoretically positively correlated. Results in Table 4 indicate the superior construct validity of GPV; i.e., for the theoretically positively correlated values, measuring with GPV also yields higher correlations.

In summary, evaluations within and across value theories indicate superior construct validity of GPV over prior tools that are prone to response bias.

5.2 Comparative Analysis of Value Representation Utility

The utility of human value measurements lies in their predictive power for human behavior (Schwartz et al. 2007). In the context of LLMs, many related studies are motivated by value alignment for safe LLM deployment (Ji et al. 2023). However, few studies have connected LLM values with their safety. In this section, we evaluate the value representation utility of different measurement tools in terms of their predictive power for LLM safety scores.

Here, we use the safety scores of 17 LLMs from SALAD-Bench (Li et al. 2024a) as ground truth and randomly sample 100 prompts from Salad-Data (Li et al. 2024a) for GPV measurement. We follow the standard linear probing protocol and train a linear classifier to predict the relative safety of LLMs, using the value measurement results as features. We perform its training 30 times for each measurement tool with randomly sampled data splits to ensure statistically meaningful results. Full experimental details are given in Appendix D.4.

Using values from different value theories as features leads to different results. We present the best classification accuracy of different measurement tools in Table 5. The results indicate that GPV is more predictive of LLM safety scores than prior tools. It suggests that GPV values can be an interpretable and actionable proxy for LLM safety under specific context (Röttger et al. 2024).

In addition, as detailed in Appendix D.4, we examine the predictive power of various value systems for LLM safety scores, as well as the impact of different values on LLM safety. We find that, despite the popularity of Schwartz's value system within the AI community, VSM (Hofstede

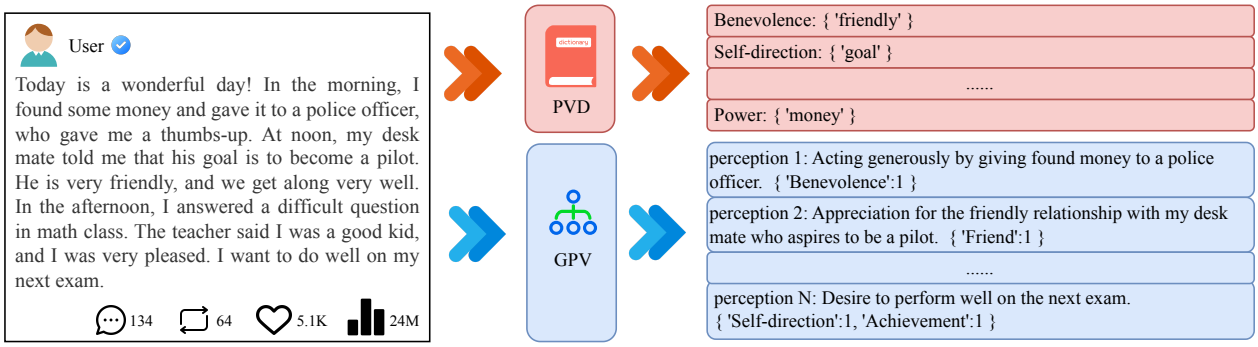


Figure 3: Comparative analysis of PVD (Ponizovskiy et al. 2020) and GPV: a case study.

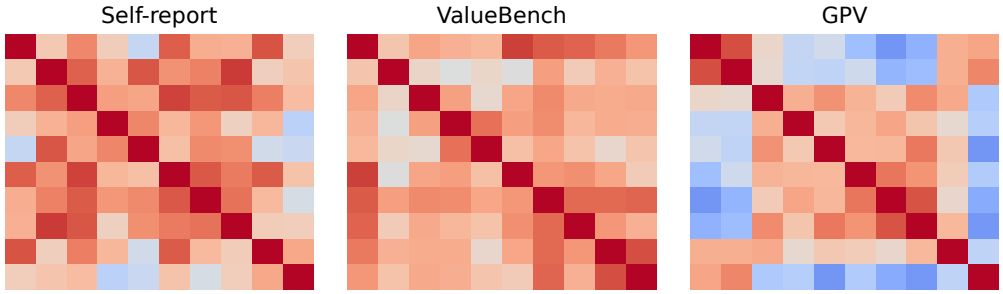


Figure 4: Correlations between Schwartz values when using different measurement tools. From dark blue to dark red, the correlation ranges from -1 to 1. The Schwartz values are ordered along the x-axis and y-axis according to their positions in the theoretical circumplex structure. See the extended version for more details.

2011) is more predictive of LLM safety. Within VSM, values like Long-term Orientation positively contribute to LLM safety while values like Masculinity negatively contribute.

In summary, GPV is more predictive of LLM safety. The proposed Value Representation Utility also enables us to evaluate both the predictive power of a value system and the relationship between each encoded value and LLM safety.

5.3 Discussion

Superiority of GPV. We discuss that the superior construct validity may be attributed to the encoded knowledge. ValueLlama learns the correlations between different values and exploits them to generate coherent and valid measurements. In addition, measuring the free-form LLM responses is more reliable than prompting with forced-choice questions (Dominguez-Olmedo, Hardt, and Mendler-Dünner 2023). The superior value representation utility of GPV may be attributed to the context-specific value measurements. Unlike humans, who exhibit stable values, LLMs may not be treated as monolithic entities, highlighting the importance of context-specific measurement (Röttger et al. 2024). GPV, for the first time, enables reliable context-specific measurements. Overall, compared to prior tools, GPV for LLM value measurements 1) mitigates response bias and yields theoretically valid results; 2) is more practically relevant due to measuring scalable and free-form LLM responses; and 3) enables context-specific measurements.

Limitations and Future Work. The current studies are limited to evaluating LLMs in English. Since the used languages are shown to affect LLM values (Cahyawijaya et al. 2024), future research should consider multi-lingual measurements. Additionally, future investigations should explore the spectrum of values an LLM can exhibit, examining the effects of different profiling prompts. Though LLM values may be steerable, current alignment algorithms establish default model positions and behaviors, making it still meaningful to evaluate the values and opinions reflected in these defaults (Röttger et al. 2024).

6 Conclusion

This paper introduces GPV, an LLM-based tool designed for value measurement, theoretically based on text-revealed selective perceptions. Experiments conducted through diverse lenses demonstrate the superiority of GPV in measuring both human and AI values. GPV offers promising opportunities for both sociological and technical research. In sociological research, GPV enables scalable, automated, and cost-effective value measurements that reduce response bias compared to self-reports and provide more semantic nuance than prior data-driven tools. It is highly flexible and can be used independently of specific value systems or measurement contexts. For technical research, GPV presents a new perspective on value alignment by offering interpretable and actionable value representations for LLMs.

Ethical Statement

Measuring values with GPV may involve biases encoded in LLMs, during perception-level measurement and perception parsing. Currently, GPV is intended for research purposes only, and researchers should exercise caution when applying it to content with subjective or controversial interpretations.

For the perception-level measurement, we fine-tuned our model using established psychological inventories and synthetic data validated across cultures, aiming to reduce measurement bias. In the three-class valence classification task, the model is trained to provide neutral predictions when additional context is needed, thereby minimizing the risk of bias. Nevertheless, achieving unbiased measurement requires further investigation.

The parsing results in this study are considered high-quality by our annotators. However, since the annotators share a similar demographic background, their evaluations may lack a comprehensive and diverse perspective. Additionally, the blog data analyzed in this work primarily focuses on general, everyday topics and rarely involves controversial issues. Addressing potential biases in parsing remains an open area for future research.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62276006); Wuhan East Lake High-Tech Development Zone National Comprehensive Experimental Base for Governance of Intelligent Society; and the Fundamental Research Funds for the Central Universities.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adkins, C.; Russell, C.; and WERBEL, J. 1994. Judgments of fit in the selection process: The role of work value congruence. *Personnel Psychology*, 47: 605 – 623.

Anderson, B. A. 2019. Neurobiology of value-driven attention. *Current Opinion in Psychology*, 29: 27–33. Attention & Perception.

Bardi, A.; and Schwartz, S. H. 2003. Values and behavior: Strength and structure of relations. *Personality and social psychology bulletin*, 29(10): 1207–1220.

Bekkers, R. H. 2007. Measuring altruistic behavior in surveys: The all-or-nothing dictator game. In *Survey research methods*, volume 1, 1–11. European Survey Research Association.

Bilsky, W.; Janik, M.; and Schwartz, S. H. 2011. The structural organization of human values-evidence from three rounds of the European Social Survey (ESS). *Journal of cross-cultural psychology*, 42(5): 759–776.

Cahyawijaya, S.; Chen, D.; Bang, Y.; Khalatbari, L.; Wilie, B.; Ji, Z.; Ishii, E.; and Fung, P. 2024. High-Dimension Human Value Representation in Large Language Models. *arXiv preprint arXiv:2404.07900*.

Cao, Y.; Zhou, L.; Lee, S.; Cabello, L.; Chen, M.; and Hershcovich, D. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In Dev, S.; Prabhakaran, V.; Adelani, D.; Hovy, D.; and Benotti, L., eds., *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 53–67. Dubrovnik, Croatia: Association for Computational Linguistics.

Cieciuch, J.; and Schwartz, S. H. 2012. The number of distinct basic values and their structure assessed by PVQ–40. *Journal of personality assessment*, 94(3): 321–328.

Dominguez-Olmedo, R.; Hardt, M.; and Mendler-Dünner, C. 2023. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Fischer, R.; and Schwartz, S. 2011. Whence differences in value priorities? Individual, cultural, or artifactual sources. *Journal of Cross-Cultural Psychology*, 42(7): 1127–1144.

Gibson, J. J. 1960. The concept of the stimulus in psychology. *American psychologist*, 15(11): 694.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029.

Hofstede, G. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture*, 2(1): 8.

Huang, J.; Wang, W.; Li, E. J.; Lam, M. H.; Ren, S.; Yuan, Y.; Jiao, W.; Tu, Z.; and Lyu, M. R. 2024. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Jiang, H.; Zhang, X.; Cao, X.; Breazeal, C.; Roy, D.; and Kabbara, J. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 3605–3627. Mexico City, Mexico: Association for Computational Linguistics.

Kimura, T. 2023. Assessment of personal values for data-driven human resource management. *Data Science Journal*, 22(1).

Klingefjord, O.; Lowe, R.; and Edelman, J. 2024. What are human values, and how do we align AI to them? *arXiv preprint arXiv:2404.10636*.

Li, L.; Dong, B.; Wang, R.; Hu, X.; Zuo, W.; Lin, D.; Qiao, Y.; and Shao, J. 2024a. SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. *arXiv preprint arXiv:2402.05044*.

Li, X.; Chen, X.; Niu, Y.; Hu, S.; and Liu, Y. 2024b. PsyDI: Towards a Personalized and Progressively In-depth

Chatbot for Psychological Measurements. *arXiv preprint arXiv:2408.03337*.

Li, X.; Li, Y.; Qiu, L.; Joty, S.; and Bing, L. 2024c. Evaluating Psychological Safety of Large Language Models. arXiv:2212.10529.

Lin, W.-L.; and Yao, G. 2024. Concurrent validity. In *Encyclopedia of quality of life and well-being research*, 1303–1304. Springer.

Ma, B.; Wang, X.; Hu, T.; Haensch, A.-C.; Hedderich, M. A.; Plank, B.; and Kreuter, F. 2024. The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models. *arXiv preprint arXiv:2406.11096*.

Maio, G. R. 2010. Mental representations of social values. In *Advances in experimental social psychology*, volume 42, 1–43. Elsevier.

Meglino, B. M.; and Ravlin, E. C. 1998. Individual Values in Organizations: Concepts, Controversies, and Research. *Journal of Management*, 24(3): 351–389.

Murphy, R. O.; and Ackermann, K. A. 2014. Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1): 13–41.

Pan, K.; and Zeng, Y. 2023. Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for Large Language Models. arXiv:2307.16180.

Pellert, M.; Lechner, C. M.; Wagner, C.; Rammstedt, B.; and Strohmaier, M. 2023. AI Psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 17456916231214460.

Ponizovskiy, V.; Ardag, M.; Grigoryan, L.; Boyd, R.; Dobewall, H.; and Holtz, P. 2020. 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.

Postman, L.; Bruner, J. S.; and McGinnies, E. 1948. Personal values as selective factors in perception. *Journal of abnormal psychology*, 43 2: 142–54.

Ren, Y.; Ye, H.; Fang, H.; Zhang, X.; and Song, G. 2024. ValueBench: Towards Comprehensively Evaluating Value Orientations and Understanding of Large Language Models. *arXiv preprint arXiv:2406.04214*. https://github.com/Value4AI/ValueBench.

Rokeach, M. 1973. *The nature of human values.* Free press.

Röttger, P.; Hofmann, V.; Pyatkin, V.; Hinck, M.; Kirk, H. R.; Schütze, H.; and Hovy, D. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. arXiv:2402.16786.

Safdari, M.; Serapio-García, G.; Crepy, C.; Fitz, S.; Romero, P.; Sun, L.; Abdulhai, M.; Faust, A.; and Matarić, M. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Sagiv, L.; and Roccas, S. 2017. What personal values are and what they are not: Taking a cross-cultural perspective. *Values and behavior: Taking a cross cultural perspective*, 3–13.

Sagiv, L.; Roccas, S.; Cieciuch, J.; and Schwartz, S. H. 2017. Personal values in human life. *Nature human behaviour*, 1(9): 630–639.

Sagiv, L.; Sverdlik, N.; and Schwarz, N. 2011. To compete or to cooperate? Values' impact on perception and action in social dilemma games. *European Journal of Social Psychology*, 41(1): 64–77.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. 2023a. Evaluating the Moral Beliefs Encoded in LLMs. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 51778–51809. Curran Associates, Inc.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. M. 2023b. Evaluating the Moral Beliefs Encoded in LLMs. arXiv:2307.14324.

Schler, J.; Koppel, M.; Argamon, S.; and Pennebaker, J. W. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, 199–205.

Schwartz, S. H. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. 25: 1–65.

Schwartz, S. H.; and Bilsky, W. 1990. Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. *Journal of personality and social psychology*, 58(5): 878.

Schwartz, S. H.; and Butenko, T. 2014. Values and behavior: Validating the refined value theory in Russia. *European journal of social psychology*, 44(7): 799–813.

Schwartz, S. H.; and Rubel, T. 2005. Sex differences in value priorities: cross-cultural and multimethod studies. *Journal of personality and social psychology*, 89(6): 1010.

Schwartz, S. H.; et al. 2007. Value orientations: Measurement, antecedents and consequences across nations. *Measuring attitudes cross-nationally: Lessons from the European Social Survey*, 169–203.

Shen, Y.; Wilson, S. R.; and Mihalcea, R. 2019. Measuring personal values in cross-cultural user-generated content. In *Social Informatics: 11th International Conference, SocInfo 2019, Doha, Qatar, November 18–21, 2019, Proceedings 11*, 143–156. Springer.

Sorensen, T.; Jiang, L.; Hwang, J. D.; Levine, S.; Pyatkin, V.; West, P.; Dziri, N.; Lu, X.; Rao, K.; Bhagavatula, C.; et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19937–19947.