Generative Psycho-Lexical Approach for Constructing Value Systems in Large Language Models

Haoran Ye[*,1], Tianze Zhang[*,1,2], Yuhang Xie[*,1], Liyuan Zhang[1], Yuanyi Ren[1], Xin Zhang[3,4], Guojie Song[†,1,5]

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

{hrye, ericzhang, yuhangxie, zly2003}@stu.pku.edu.cn
{yyren, zhang.x, gjsong}@pku.edu.cn

Abstract

Values are core drivers of individual and collective perception, cognition, and behavior. Value systems, such as Schwartz's Theory of Basic Human Values, delineate the hierarchy and interplay among these values, enabling cross-disciplinary investigations into decision-making and societal dynamics. Recently, the rise of Large Language Models (LLMs) has raised concerns regarding their elusive intrinsic values. Despite growing efforts in evaluating, understanding, and aligning LLM values, a psychologically grounded LLM value system remains underexplored. This study addresses the gap by introducing the Generative Psycho-Lexical Approach (GPLA), a scalable, adaptable, and theoretically informed method for constructing value systems. Leveraging GPLA, we propose a psychologically grounded five-factor value system tailored for LLMs. For systematic validation, we present three benchmarking tasks that integrate psychological principles with cutting-edge AI priorities. Our results reveal that the proposed value system meets standard psychological criteria, better captures LLM values, improves LLM safety prediction, and enhances LLM alignment, when compared to the canonical Schwartz's values.

1 Introduction

Personal values are broad, fundamental motivations behind individual and collective actions. They serve as guiding principles that shape perceptions, cognition, and behaviors across situations and over time. As Large Language Models (LLMs) continue to permeate and transform human society, researchers have increasingly sought to evaluate, understand, and align LLM values (Ye et al., 2025b; Ren et al., 2024; Rao et al., 2024; Meadows et al., 2024; Kovač et al., 2024; Jiang et al., 2024b; Rozen et al., 2024; Yao et al., 2024b; Sorensen et al., 2024a). Rather than relying solely on specific risk metrics, implicit preference modeling, or atomic ethical principles, framing these studies through the lens of intrinsic values (Schwartz, 2012) offers a more comprehensive, adaptable, and transparent approach (Yao et al., 2024a; Biedma et al., 2024; Yao et al., 2023).

However, value systems used in existing research—primarily Schwartz's values—are established for humans. A psychologically informed value system tailored for LLMs remains largely unexplored, obscuring a structured and holistic perspective on their values. A preliminary attempt at constructing LLM value systems (Biedma et al., 2024) has several limitations that undermine its psychological grounding; we present related discussion and experimental evidence in § 2.3 and Appendix C.

To address this gap, we introduce a generative psycho-lexical approach (GPLA), a novel methodology that leverages LLMs to construct a value system grounded in psychological principles. This approach is based on the psycho-lexical hypothesis, which suggests that all salient values are captured in language. GPLA adopts an agentic framework and involves five sequential steps: 1) extracting perceptions from the corpus, 2) identifying values behind perceptions, 3) filtering values, 4) non-reactive value measurements of test subjects, and 5) structuring the value system. In contrast to traditional

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

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11968–11991
July 27 - August 1, 2025

methods that manually compile value lexicons and design questionnaires to collect self-report data (Aavik and Allik, 2002; Creţu et al., 2012; Saucier, 2000), GPLA automates the entire process and features two key advantages.

First, GPLA supports **agent-specific construction**. Rather than relying on value lexicons derived from human-compiled dictionaries, it utilizes LLMs to identify values within free-form agent-generated text. This principled approach facilitates the agent- and context-specific collection of value lexicons and ensures better coverage (Appendix B.1). Second, GPLA enables **nonreactive value measurement**. Traditional methods that rely on self-report are demonstrated to be unreliable and biased for humans (Ponizovskiy et al., 2020) and AI (Dominguez-Olmedo et al., 2023; Röttger et al., 2024; Ye et al., 2025b). Instead, GPLA measures values within unstructured responses to arbitrary prompts, which mitigates response biases and enhances scalability and validity (Ye et al., 2025b).

In applying GPLA to 33 LLMs, each steered with 21 different value-anchoring prompts (Rozen et al., 2024), we develop a five-factor value system encompassing *Social Responsibility*, *Risk-taking*, *Rule-following*, *Self-competence*, and *Rationality*. We demonstrate the system's reliability and explore its theoretical and practical implications. To benchmark LLM value systems, we propose three tasks based on psychometric principles (DeVon et al., 2007) and AI priorities (Anwar et al., 2024): *Confirmatory Factor Analysis* for assessing structural validity, *LLM Safety Prediction* for examining predictive validity, and *LLM Value Alignment* for evaluating representational power. These tasks confirm the validity and utility of our system in evaluating, understanding, and aligning the intrinsic values of LLMs, when compared to canonical Schwartz's values (Schwartz, 2012).

In summary, we contribute: 1) the Generative Psycho-Lexical Approach (GPLA), a novel LLM- and data-driven methodology for constructing psychologically grounded value systems; 2) a five-factor value system constructed using GPLA for LLMs, with demonstrated reliability and an exploration of its theoretical and practical implications; and 3) three benchmarking tasks for LLM value systems, through which we establish the superior validity and utility of the proposed system.

2 Background and Related Work

2.1 Values, Value Measurement, and Theory-Driven Value Systems

Values. Values are broad, fundamental beliefs and guiding principles, serving as motivational forces when evaluating actions, people, and events (Schwartz, 2012). Unlike other psychological constructs such as personalities, values are stable, motivational, and transituational (Sagiv and Schwartz, 2022), rendering them uniquely powerful across disciplines for investigating decision-making and social dynamics (Sen, 1986; Schwartz, 2006).

Value Measurement. Values are measurable constructs (Schwartz et al., 2001). Value measurement seeks a quantitative assessment of the importance an agent assigns to specific values. Traditionally, it relies on tools like self-report questionnaires (Maio, 2010), behavioral observation (Lönnqvist et al., 2013), and experimental techniques (Murphy and Ackermann, 2014), which are hindered by response biases, resource demands, and inabilities to capture authentic behaviors (Ponizovskiy et al., 2020; Boyd et al., 2015). Recently proposed Generative Psychometrics for Values (GPV) (Ye et al., 2025b) leverage LLMs to dynamically parse unstructured texts into perceptions and measure values therein. GPV is shown to be more scalable, valid, and flexible than traditional tools in measuring LLM values, and is utilized for non-reactive value measurement in this work.

Theory-Driven Value Systems. Isolated values are structured into a value system for a holistic understanding of their hierarchy and interplay (Schwartz, 2012). Theory-driven value systems like Schwartz's theory of basic human values (Schwartz, 1992) and functional theory of values (Gouveia et al., 2014) are rooted in theoretical hypotheses and preconceptions of psychologists. These systems are established prior to data collection and analysis, and are later validated through empirical studies. However, theory-driven approaches are constrained by the subjectivity of the theorists, limited coverage, and inadaptability to evolving values or specific contexts (Ponizovskiy et al., 2020; Raad et al., 2017).

2.2 Psycho-Lexical Approach

In contrast to the theory-driven approaches, the psycho-lexical approach organizes psychological constructs such as values, personality traits, and

social attitudes (Aavik and Allik, 2002; Creţu et al., 2012; Klages and Johnson, 1929; Saucier, 2000), under a data-driven paradigm. It operates on the fundamental principle that language naturally evolves to capture salient and socially relevant individual differences (De Raad et al., 2016). The pioneering attempts date back to the early 20th century (Allport and Odbert, 1936; Galton, 1950), and the paradigm has then been extended to different psychological constructs in decades of development (Cattell, 1957; John et al., 1988; Aavik and Allik, 2002; Ponizovskiy et al., 2020; Mai and Church, 2023; Garrashi et al., 2024).

Most of these works, taking values as an example, involve the following steps: 1) compilation of value-descriptive terms, mostly from dictionaries and thesauruses; 2) refinement and reduction to eliminate redundant, ambiguous, and uncommon terms; 3) value measurement, through collecting self-reports and peer ratings, to identify underlying correlations between the value descriptors; and 4) uncovering the hidden value factors or dimensions through statistical methods like principal component analysis or factor analysis.

In this work, we aim to harness the extensive knowledge and semantic understanding of LLMs to address the limitations of traditional psycho-lexical approaches. These limitations include: 1) extensive manual labor required to compile and refine the lexicons, 2) insufficient coverage of values of different linguistic forms, 3) lack of criteria for prioritizing values lexicons when filtering, 4) the responses bias and inscalability of self-report value measurement, and 5) the inadaptability to changing values or specific contexts.

2.3 From Human Values to LLM Values

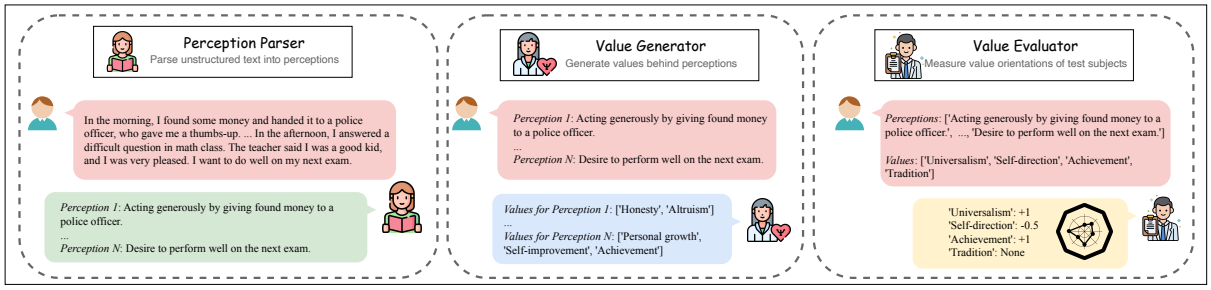
The rise of LLMs also brings the critical need to evaluate, understand, and align their values (Ye et al., 2025a). Extensive works focus on evaluating the values of LLMs using traditional self-report tools (Li et al., 2022, 2024c; Huang et al., 2024a; Safdari et al., 2023a,b), static customized inventories (Ren et al., 2024; Meadows et al., 2024; Jiang et al., 2024a), dynamically generated inventories (Jiang et al., 2024b), or directly from the model's free-form outputs (Ye et al., 2025b; Yao et al., 2024b,a, 2025). Other works attempt to understand the LLM values in aspects like the consistency of shown values (Rozen et al., 2024; Moore et al., 2024; Röttger et al., 2024), the ability to reason about human values (Ganesan et al., 2023; Sorensen et al., 2024a; Jiang et al., 2024d; Strachan et al., 2024), and the ability to express or role-play human values (Jiang et al., 2024c; Zhang et al., 2023; Kang et al., 2024; Zhang et al., 2025; Li et al., 2024a; Huang et al., 2024b). Another line of research aligns LLM values with human values, setting the alignment goal as specific risk metrics (Lin et al., 2024; Gunjal et al., 2024), human demonstrations (Dubois et al., 2024; Köpf et al., 2024; Taori et al., 2023), implicit preference modeling (Ouyang et al., 2022; Rafailov et al., 2024; Zhong et al., 2024), or intrinsic values (Bai et al., 2022a; Yao et al., 2024a; Bai et al., 2022c). Among different alignment goals, intrinsic values, as transituational decision-guiding principles, demonstrate unique advantages (Yao et al., 2023). However, prior works are mostly based on Schwartz's value system, which is established and validated for humans and may not necessarily capture the unique LLM psychology.

One preliminary attempt (Biedma et al., 2024) at constructing LLM value systems faces several limitations that undermine its psychological grounding: 1) the value lexicon collection is susceptible to response bias and lacks comprehensive coverage; 2) the derivation of value correlations lacks theoretical or empirical validity; and 3) the resulting value system has yet to be evaluated in terms of its reliability, validity, or utility. Appendix C presents further discussion and experimental evidence.

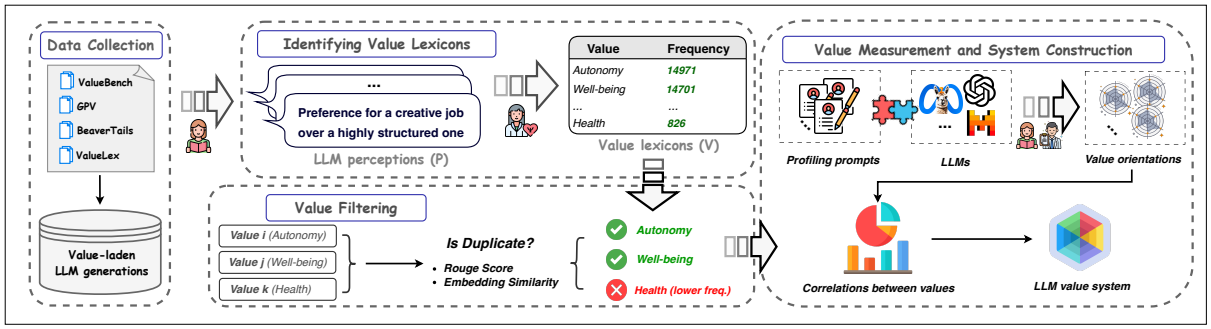
3 Generative Psycho-Lexical Approach for Constructing LLM Value System

Generative Psycho-Lexical Approach (GPLA), as our foundation for constructing the LLM value system, is illustrated schematically in Fig. 1 and algorithmically in Algorithm 1. GPLA adopts an agentic framework, comprising three LLM agents and four sequential steps, to systematically collect, filter, measure, and structure the LLM value system.

Data Collection. In line with the established psycho-lexical approach (De Raad et al., 2016), we begin by comprehensively identifying atomic values that LLMs may endorse. We draw upon a diverse, value-laden corpus of LLM outputs $\mathcal{C}$ from ValueBench (Ren et al., 2024), GPV (Ye et al., 2025b), BeaverTails (Ji et al., 2024), and ValueLex (Biedma et al., 2024). These corpora provide a broad spectrum of value-laden LLM generations. Detailed statistics and further explanations on these datasets are provided in Appendix A.1.



(a) LLM agents in GPLA: (1) Perception Parser M_P, (2) Value Generator M_G, and (3) Value Evaluator M_E.



(b) The pipeline of GPLA for constructing LLM value system: 1) extracting perceptions from the corpus, 2) identifying values behind perceptions, 3) filtering values, 4) measuring value orientations of LLM test subjects, and 5) computing the structure of the value system.

Figure 1: Generative Psycho-lexical Approach (GPLA) for Constructing LLM Value System.

Algorithm 1 Generative Psycho-lexical Approach for Constructing LLM Value System

Input: Corpus $\mathcal{C}$ of LLM-generated text, LLMs as value measurement subjects $M = \{M_i\}_{i=1}^{n}$, and three LLM agents M_P, M_G, M_E.
Output: Value system $\mathcal{VS} = (V, V_H, \mathbf{\Lambda})$ of LLMs.
$P = M_P(\mathcal{C})$ *// Extract perceptions P from corpus $\mathcal{C}$*
$\{V, Freq\} = M_G(P)$ *// Generate values V behind perceptions P and collect their frequencies*
$V = \text{LexiconFiltering}(V, Freq)$ *// Remove duplicate and less prioritized values*
$\mathbf{X}_M = \text{GPV}(V, M; M_P, M_E)$ *// Measure the value orientations of LLM subjects*
$V_H, \mathbf{\Lambda} = \text{PCA}(\mathbf{X}_M)$ *// Compute the hidden factors and structure of the system*

Identifying Value Lexicons. Then, we extract perceptions P from the collected corpus using the Perception Parser M_P. Perceptions are atomic value-rich expressions that reflect values behind free-form LLM outputs and are akin to stimuli in traditional psychometric tools (Ye et al., 2025b). These perceptions are then mapped to underlying values V through the Value Generator M_G, with their frequency of occurrence $Freq$ recorded. We instantiate M_P following GPV (Ye et al., 2025b) and M_G using Kaleido model (Sorensen et al., 2024a). In Appendix B.1, we confirm the comprehensive coverage of the generated values.

Value Filtering. To ensure the value lexicons are concise and representative, we filter out duplicates using Rouge scores and embedding similarity following the validated setup in (Sorensen et al., 2024a). When two values exhibit significant semantic overlap, we retain the one more frequently observed, as higher frequency signifies greater priority (Ponizovskiy et al., 2020). This process results in value lexicons V that minimize redundancy while preserving semantic coverage.

Value Measurement and System Construction. Structuring the LLM value system requires identifying correlations between values based on LLM value measurements. To this end, we adopt GPV (Ye et al., 2025b), instantiated with M_P and M_E, to measure the value orientations $\mathbf{X}_M$ of LLM subjects M on values V. GPV dynamically extracts psychometric stimuli from LLM outputs and reveals value orientations therein (detailed in Ap-

pendix A.3). We measure a set of 693 LLM subjects, consisting of 33 LLMs paired with 21 profiling prompts (listed in Appendix A.2). The involved LLMs are instruction-tuned models, as they are more relevant for real-world, public-facing applications. The value system $\mathcal{VS} = (V, V_H, \mathbf{\Lambda})$ is then constructed through Principal Component Analysis (PCA) (Ponizovskiy et al., 2020), which uncovers the underlying value factors V_H and hierarchical relationships $\mathbf{\Lambda}$ (loadings of V on V_H).

4 Benchmarking LLM Value Systems

The power of a value system lies in its validity and utility (Schwartz, 2012). This work introduces, for the first time, three benchmarking tasks for LLM value systems, grounded in both psychological theories and AI priorities: 1) Confirmatory Factor Analysis, which evaluates their structure validity given LLM value measurements; 2) LLM Safety Prediction, which assesses their predictive validity for LLM safety; and 3) LLM Value Alignment, which examines their representation power in aligning LLMs with human values.

4.1 Confirmatory Factor Analysis for Evaluating Structure Validity

Confirmatory Factor Analysis (CFA) is a statistical technique used to test whether a set of observed variables (atomic values) accurately represents and loads onto a smaller number of underlying latent factors (high-level values) (Schwartz and Boehnke, 2004). Mathematically, the CFA model is represented as:

$$\mathbf{X} = \mathbf{F}\mathbf{\Lambda}^T + \epsilon, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times |V|}$ is the observed data matrix (measuring atomic values); $\mathbf{F} \in \mathbb{R}^{n \times |V_H|}$ is the matrix of latent factors (aggregated measurements of high-level value factors); $\mathbf{\Lambda} \in \mathbb{R}^{|V| \times |V_H|}$ is the matrix of factor loadings (the contribution of each atomic value to each latent factor); and $\epsilon \in \mathbb{R}^{n \times |V|}$ is the matrix of error terms.

This task evaluates the model fit on held-out value measurement data. Better model fit indicates higher structure validity of the system.

4.2 LLM Safety Prediction for Evaluating Predictive Validity

Predictive validity refers to how well measurement results based on a particular value system can anticipate future decisions, actions, or outcomes (Bardi and Schwartz, 2003). In the context of human values, evaluating predictive validity often examines the correlations between values and other psychological constructs, such as personality traits and attitudes. For LLMs, we propose to evaluate the predictive validity by examining the relationship between LLM values and LLM safety, arguably one of the most critical concerns and "psychological constructs" for LLM deployment.

To this end, we draw prompts and safety scores from the established safety benchmark SALAD-Bench (Li et al., 2024b). We administer the prompts to LLMs, collect their responses, and measure the revealed values using GPV (Ye et al., 2025b). We follow the linear probing protocol (Chen et al., 2020) to evaluate the predictive validity of the value representations under different value systems. Using the Bradley-Terry model (Bradley and Terry, 1952), we train a linear classifier with pairwise cross-entropy loss to predict the relative safety of LLM pairs:

$$\begin{aligned} s_{M_i} &= \text{Linear}(\mathbf{x}_{M_i}), \\ P(M_i \succ M_j) &= \frac{\exp(s_{M_i})}{\exp(s_{M_i}) + \exp(s_{M_j})}. \end{aligned} \tag{2}$$

Here, $s_{M_i} \in \mathbb{R}$ is the predicted LLM safety score, and $\mathbf{x}_{M_i} \in \mathbb{R}^{|V_H|}$ is the vector of LLM values under a particular value system. In this case, we only care about the relative safety of LLM pairs, not absolute safety scores. Therefore, the predictive validity is indicated by the binary classification accuracy of a well-trained linear classifier.

4.3 LLM Value Alignment for Evaluating Representation Power

Values are cognitive representations of motivational goals (Schwartz, 2012). A well-constructed value system should be able to fully represent the broad motivations behind LLM outputs, therefore enabling effective transituational value alignment.

Our evaluation builds on the BaseAlign algorithm (Yao et al., 2024a), originally designed to align LLM outputs with Schwartz's basic human values. This approach employs a target human value vector, either heuristically customized or derived from human survey data. We extend BaseAlign to arbitrary value systems by distilling

the target vector from human preference:

$$\begin{aligned}\mathbf{x}_V^* &= \arg\min_{\mathbf{x}} f(\mathbf{x}; \mathcal{D}, M_E, V) \\ &= \arg\min_{\mathbf{x}} \sum_{(p,r_w,r_l)\in\mathcal{D}} \Big[\big|\mathbf{x} - M_E(p, r_w, V)\big| \\ &\quad - \big|\mathbf{x} - M_E(p, r_l, V)\big| \Big].\end{aligned} \tag{3}$$

Here, $\mathbf{x}_V^* \in \mathbb{R}^{|V|}$ is the target value vector for values V under evaluation. It is distilled from dataset $\mathcal{D}$ (Ouyang et al., 2022) using open-vocabulary value evaluator M_E (Ye et al., 2025b). $\mathcal{D}$ contains triplets of prompt p, winning response r_w, and losing response r_l. The distillation objective is to minimize the L1 distance between the target value and the values of winning responses, while maximizing the distance with those of losing responses. Given that the objective function is piecewise linear, the optimal solution can be derived by iteratively examining each value dimension V_i and the corresponding value measurements $\mathcal{M}_{V,i} = \{M_E(p, r_w, V_i)\}_{\mathcal{D}} \cup \{M_E(p, r_l, V_i)\}_{\mathcal{D}}$:

$$\begin{aligned}\mathbf{x}_{V,i}^* = \arg\min_{\mathbf{x}_{V,i}\in\mathcal{M}_{V,i}} f(\mathbf{x}_{V,i}; \mathcal{D}, M_E, V_i), \\ \forall i = 1, \ldots, |V|.\end{aligned} \tag{4}$$

Next, we adopt PPO (Schulman et al., 2017) to align the LLM with the target value under V (Yao et al., 2024a):

$$\min_{\theta} \mathbb{E}_{p\sim\mathcal{D}_p, r\sim\pi_\theta(p)} \left|\mathbf{x}_V^* - M_E(p, r, V)\right|, \tag{5}$$

where θ is the LLM policy parameters, $\mathcal{D}_p$ is the dataset of prompts (Dai et al., 2023), and $\pi_\theta(p)$ is the parameterized LLM response distribution. The representation power of V is evaluated by the helpfulness and harmlessness of the aligned LLMs (Bai et al., 2022b; Yao et al., 2024a).

5 Results

This section presents the results of our experiments, including the proposed LLM value system, the analysis of the system, and the benchmarking of our system against the well-established Schwartz's values (Schwartz, 2012).

5.1 Proposed Value System

Fig. 2 visualizes the value system constructed by GPLA. Table 1 gathers its partial factor loadings, Cronbach's Alpha (α), and confidence intervals (CI). The factor loadings indicate the strength of the relationship between each factor and its atomic values, with higher loadings suggesting more contribution to the factor. Cronbach's Alpha measures the internal consistency of each factor, with higher values indicating greater reliability. Some atomic values are removed to ensure clear loading patterns and desirable factor reliability (Aavik and Allik, 2002). After that, all our factors reach the standard psychometric threshold of 0.7, indicating strong internal consistency (Taber, 2018).

Factor	Value	Loading
Social Responsibility ($\alpha = 0.957$, *CI: 0.952–0.961*)		
	Equity	0.890
	Empathy	0.885
	Teamwork	0.872
	Equality	0.827
	Unity	0.825
Risk-Taking ($\alpha = 0.919$, *CI: 0.910–0.928*)		
	Challenge	0.812
	Boldness	0.804
	Adventure	0.798
	Change	0.730
	Thrill-seeking	0.728
Rule-Following ($\alpha = 0.842$, *CI: 0.824–0.859*)		
	Realism	0.708
	Order	0.694
	Responsibility	0.682
	Prudence	0.676
	Efficiency	0.669
Self-Competence ($\alpha = 0.761$, *CI: 0.732–0.787*)		
	Confidence	0.601
	Impact	0.527
	Proactivity	0.460
	Achievement	0.455
	Recognition	0.455
Rationality ($\alpha = 0.722$, *CI: 0.686–0.754*)		
	Objectivity	0.705
	Evidence-based	0.649
	Logic	0.618
	Neutrality	0.522

Table 1: Partial factor loadings, Cronbach's Alpha (α), and 95% confidence intervals (CI) for our value system. Full results are provided in Table 13.

Analyzing factor loadings reveals the underlying structure of five key factors: Social Responsibility emphasizes fairness, inclusivity, and societal well-being; Risk-Taking reflects openness to change and embracing uncertainty; Rule-Following highlights order, discipline, and pragmatism; Self-Competence focuses on personal growth

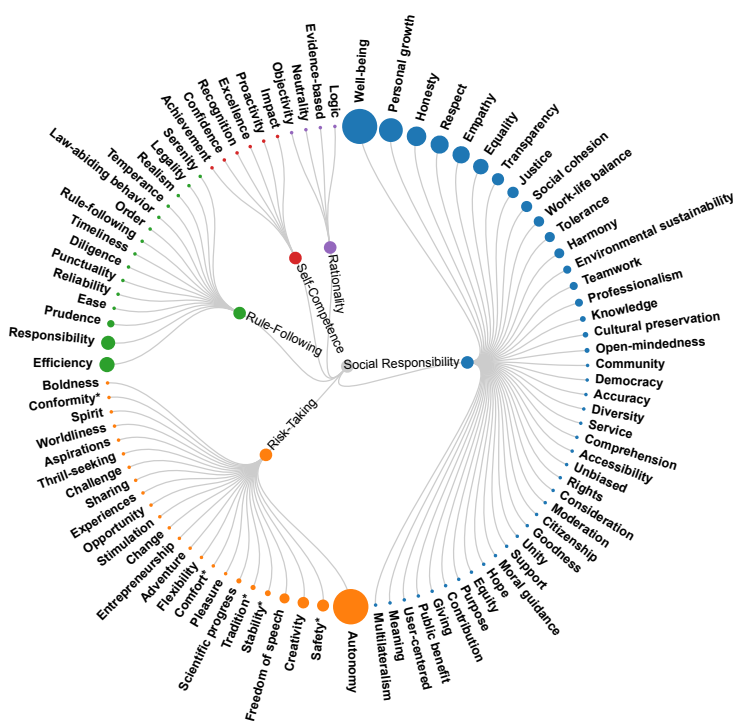


Figure 2: Dendrogram of our value system. Values with a "*" are negatively loaded.

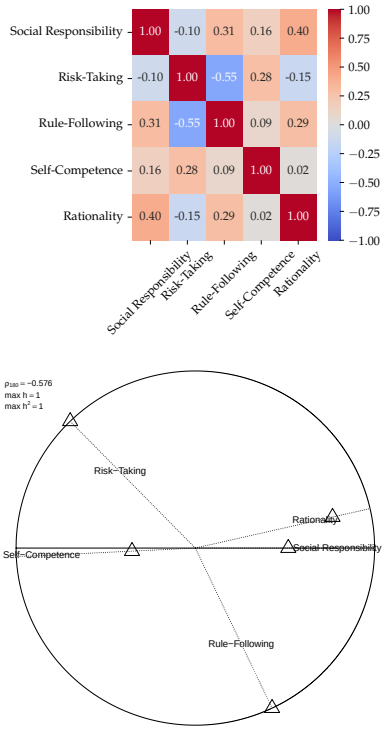


Figure 3: Correlation heatmap and circumplex analysis.

and achievement; and Rationality prioritizes logical, evidence-based decision-making grounded in objectivity and impartiality. Full factor loadings and further analysis are available in Appendix B.3.

5.2 Analyzing Value System

Value Correlation Analysis. Fig. 3 (Top) presents the correlations between the factors in our value system. Similar to Schwartz's theory of basic human values (Schwartz, 2012), LLM values also exhibit both compatible and opposing relationships. Notably, social responsibility, rule-following, and rationality show positive correlations with one another, while all of them are negatively correlated with risk-taking.

Circumplex Analysis. Circumplex analysis is a statistical method that examines whether the underlying structure of variables aligns with a circumplex pattern, and, if so, the positions of variables on a circle. The stronger the correlation between variables, the shorter their distance on the circumference. We conduct circumplex analysis based on the correlations between factors. Fig. 3 (Bottom) illustrates the analysis results based on Browne's circular stochastic process model (Browne, 1992; Grassi et al., 2010). The compatible values are closer on the circle (e.g., Social Responsibility and Rule-Following) while opposing values are positioned diagonally (e.g., Risk-Taking and Rule-Following). The results verify the presence of a circumplex structure in the value system.

Consistency Across Datasets. To evaluate the consistency and robustness of our multivariate system structure (i.e., the 5-dimensional relationship), we measure LLM values using two distinct prompt datasets: the psychometric prompts from GPV (Ye et al., 2025b) and the red-teaming prompts from SALAD-Bench (Li et al., 2024b), which feature distinct prompt distributions. Their measurements yield an average intra-LLM correlation of 0.87; here we use intra-LLM correlations because the relative value hierarchy within an LLM is more important than their absolute measurements. This result indicates a high level of consistency in the value structure across prompt distributions. We also find a high correlation (0.73) between intra-LLM value consistency and LLM safety scores (Li et al., 2024b). It suggests that LLMs with higher value consistency tend to be safer. Complete results are available in Appendix B.4.

5.3 Comparing Value Systems

We benchmark the proposed value system against Schwartz's value system (Schwartz, 2012), the

Value system	Schwartz (H)	Schwartz (L)	Ours
#Values	4	10	5
CFI ↑	0.56	0.23	**0.68**
GFI ↑	0.52	0.22	**0.65**
RMSEA ↓	**0.10**	0.11	0.12
AIC ↓	340	324	**265**
BIC ↓	1484	1464	**1145**

Table 2: CFA results of value systems. H and L denote high and low-level Schwartz values, respectively. CFI: Comparative Fit Index; GFI: Goodness of Fit Index; RMSEA: Root Mean Square Error of Approximation; AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion.

most established framework for human values and commonly used in LLM value studies.

Confirmatory Factor Analysis. We follow the standard validation procedures of CFA (Schwartz and Boehnke, 2004) to evaluate the structure validity of different value systems. For our value system, we construct it using half of the measurement data and bootstrap the data to ensure its sufficiency (#data points $\geq 5\times$ #variables). The other half of the data is held out for CFA. For Schwartz's value systems, we map the observed value variables (the atomic values in our system) to its four high-level values or ten low-level values, according to the semantic relevance (Schwartz and Boehnke, 2004) measured by an embedding model (OpenAI, 2024); all data is used for CFA. Table 2 displays the CFA results. Our value system demonstrates a better fit for the data, capturing the underlying values behind LLM generations.

LLM Safety Prediction. We follow the setup in (Ye et al., 2025b, Section 5.2) to predict LLM safety based on their value orientations. Table 3 presents the prediction accuracy under different value systems. The higher accuracy of our value system indicates its superior predictive validity. In addition, according to the parameters of well-trained linear classifiers, we find that Social Responsibility, Rule-Following, and Rationality enhance safety, whereas Risk-taking and Self-Competence undermine it; see Appendix B.6 for full results.

LLM Value Alignment. We follow the experimental setup in (Yao et al., 2024a, Section 6.2) to perform LLM value alignment. Different value systems are respectively used to represent LLM outputs and desired human values. We employ GPV (Ye et al., 2025b) as an open-vocabulary value evaluator for all value systems, but also include the original results of (Yao et al., 2024a) using a Schwartz-specific evaluator for comparison. Table 4 shows the alignment performance of different value systems. Alignment under our value system converges to the lowest harmlessness and the highest helpfulness, establishing its superior representation power. Full experimental details are available in Appendix A.4.

Value system	Acc (%)
Schwartz (H)	$81_{\pm 15}$
Schwartz (L)	$74_{\pm 16}$
Ours	$\mathbf{87}_{\pm 9}$

Table 3: Accuracy of LLM safety prediction based on value systems.

Value system	Harmlessness	Helpfulness
Schwartz*	-1.52	2.15
Schwartz	-1.40	2.13
Ours	**-1.26**	**2.16**

Table 4: Alignment performance of different value systems. Schwartz* denotes the original results drawn from (Yao et al., 2024a) using a Schwartz-specific value evaluator. Both Schwartz baselines are based on the 10-dimensional Schwartz values (Yao et al., 2024a).

6 Discussion

6.1 Validation of LLM Capabilities

LLMs in GPLA assume three roles: Perception Parser, Value Generator, and Value Evaluator. Perceptions are value-laden expressions dynamically extracted from LLM outputs. They resemble the stimuli in traditional psychometric tools (Ye et al., 2025b). The Perception Parser M_P extracts such perceptions from the LLM outputs. As validated by Ye et al. (2025b), M_P can extract high-quality perceptions suitable for subsequent analysis. In addition, related research confirms that LLMs demonstrate abilities beyond human golden standards in value-related generation tasks (Ren et al., 2024; Sorensen et al., 2024a; Ziems et al., 2024).

The Value Generator M_G identifies values from the extracted perceptions. As validated by Sorensen et al. (2024a), M_G, instantiated with the Kaleido model, can generate values that are both comprehensive and high-quality, with 91% of the genera-

tions marked as good by all three annotators, and missing values detected 0.35% of the time.

The Value Evaluator M_E measures the value orientations of LLMs given the extracted perceptions. When instantiated with ValueLlama (Ye et al., 2025b), M_E can approximate the psychologists' judgments on held-out values with about 90% accuracy.

6.2 Advantages of GPLA over Prior Psycho-Lexical Approaches

Here we discuss the main advantages of GPLA over the traditional psycho-lexical approach in constructing value systems. We defer the comparison with ValueLex (Biedma et al., 2024), a recent study that constructs LLM value systems, to Appendix C.

Full Automation. Traditional psycho-lexical approaches involve extensive manual labor in compiling and refining lexicons, collecting self-reports, and analyzing data. In contrast, given pre-collected corpora, GPLA automates the entire process, from value lexicon extraction and value filtering, to non-reactive value measurement and system construction.

Lexicon Collection. The traditional approach relies on values lexicons compiled from dictionaries and thesauruses, which lack comprehensive coverage in diverse linguistic forms, criteria for prioritizing values, and adaptability to evolving values or specific contexts. In contrast, GPLA utilizes LLMs to dynamically extract perceptions and generate corresponding values, offering the following advantages.

- **Comprehensive Coverage.** LLMs can generate values in diverse linguistic forms, beyond words in dictionaries and thesauruses, and thus provide more comprehensive coverage of values. Please refer to Appendix B.1 for the empirical evidence.
- **Prioritization Criteria.** The scalable collection of value lexicons by LLMs enables the concurrent collection of their occurrence frequencies. Since these frequencies reflect the salience of values (Ponizovskiy et al., 2020), they can serve as criteria for prioritizing values during the filtering process.
- **Adaptability.** Since LLMs are trained on Internet-scale data with evolving language patterns, they effectively encode up-to-date knowledge of values and are readily adaptable to changing values in changing corpora distributions.
- **Context Specificity.** GPLA constructs value systems given corpora from specific contexts, which can capture the unique values and value structures in those contexts. It enables the construction of LLM value systems.

Value Measurement. Traditional methods of value measurement typically rely on self-reports or peer ratings, which are prone to response biases, resource-intensive, and often fail to capture authentic behaviors or historical data (Ponizovskiy et al., 2020). In contrast, GPLA utilizes non-reactive and LLM-driven value measurement, which is proven to be more scalable and reliable (Ye et al., 2025b).

7 Conclusion

This study presents GPLA, a novel psychologically grounded paradigm that enables automated, scalable, adaptable, and non-reactive value system construction. With GPLA, we introduce the first theoretically and empirically validated five-factor LLM value system, encompassing Social Responsibility, Risk-Taking, Rule-Following, Self-Competence, and Rationality. We accompany the value system with three benchmarking tasks, rooted in both psychological theories and AI priorities. Experimental results confirm the superior validity and utility of the proposed value system compared to the canonical Schwartz's values.

GPLA promises to overcome the key limitations of traditional psycho-lexical paradigms by offering adaptability to evolving contexts and agent-specific needs. The proposed value system enables more effective evaluation, understanding, and alignment of LLM values, thereby contributing to their safe and reliable deployment.

Limitations

This study is limited to identifying and measuring values in English. Values are both culturally and linguistically sensitive (Schwartz, 2013), and LLM value orientations can vary across languages (Cahyawijaya et al., 2024). Future work will extend GPLA to multilingual contexts.

In addition, we follow BaseAlign (Yao et al., 2024a) when benchmarking value systems on value alignment. However, the algorithm, like many existing alignment methods, assumes a single, consis-

tent alignment target. Future work should explore distributional and pluralistic alignment (Sorensen et al., 2024b; Zhong et al., 2024) to more holistically evaluate a value system.

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

Toivo Aavik and Jüri Allik. 2002. The structure of estonian personal values: A lexical approach. *European Journal of Personality*, 16(3):221–235.

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

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022b. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022c. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

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

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

Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 31–40.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Michael W Browne. 1992. Circumplex models for correlation matrices. *Psychometrika*, 57:469–497.

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

Raymond B Cattell. 1957. Personality and motivation structure and measurement.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Jan Cieciuch, Eldad Davidov, Michele Vecchione, and Shalom H Schwartz. 2014. A hierarchical structure of basic human values in a third-order confirmatory factor analysis. *Swiss Journal of Psychology*.

Romeo Zeno Creţu, Sylvia Burcas, and Valeria Negovan. 2012. A psycho-lexical approach to the structure of values for romanian population. *Procedia-Social and Behavioral Sciences*, 33:458–462.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Boele De Raad, Fabia Morales-Vives, Dick PH Barelds, Jan Pieter Van Oudenhoven, Walter Renner, and Marieke E Timmerman. 2016. Values in a cross-cultural triangle: A comparison of value taxonomies in the netherlands, austria, and spain. *Journal of Cross-Cultural Psychology*, 47(8):1053–1075.

Holli A DeVon, Michelle E Block, Patricia Moyle-Wright, Diane M Ernst, Susan J Hayden, Deborah J Lazzara, Suzanne M Savoy, and Elizabeth Kostas-Polston. 2007. A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship*, 39(2):155–164.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2023. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

F. Galton. 1950. The measurement of character. *Readings in General Psychology.*, pages 435–444.

Adithya V Ganesan, Yash Kumar Lal, August Håkan Nilsson, and H Andrew Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation. *arXiv preprint arXiv:2306.01183*.

Harrun H Garrashi, Boele De Raad, and Dick PH Barelds. 2024. Personality in swahili culture: A psycho-lexical approach to trait structure in a language deprived of typical trait-descriptive adjectives. *Journal of personality and social psychology*, 127(2):384.

Valdiney V Gouveia, Taciano L Milfont, and Valeschka M Guerra. 2014. Functional theory of human values: Testing its content and structure hypotheses. *Personality and individual differences*, 60:41–47.

Michele Grassi, Riccardo Luccio, and Lisa Di Blas. 2010. Circe: An r implementation of browne's circular stochastic process model. *Behavior research methods*, 42:55–73.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Jan Hofer, Athanasios Chasiotis, and Domingo Campos. 2006. Congruence between social values and implicit motives: Effects on life satisfaction across three cultures. *European Journal of Personality: Published for the European Association of Personality Psychology*, 20(4):305–324.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024a. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Yue Huang, Zhengqing Yuan, Yujun Zhou, Kehan Guo, Xiangqi Wang, Haomin Zhuang, Weixiang Sun, Lichao Sun, Jindong Wang, Yanfang Ye, et al. 2024b. Social science meets llms: How reliable are large language models in social simulations? *arXiv preprint arXiv:2410.23426*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024a. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024c. 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, Mexico City, Mexico. Association for Computational Linguistics.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024d. Can language models reason about individualistic human values and preferences? *arXiv preprint arXiv:2410.03868*.

Oliver P John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, 2(3):171–203.

Yipeng Kang, Junqi Wang, Yexin Li, Fangwei Zhong, Xue Feng, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, and Zilong Zheng. 2024. Causal graph guided steering of llm values via prompts and sparse autoencoders. *arXiv preprint arXiv:2501.00581*.

Ludwig Klages and W. H. Johnson. 1929. The science of character. *Mind*, 38(152):513–520.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

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

Ronald B Larson. 2019. Controlling social desirability bias. *International Journal of Market Research*, 61(5):534–547.

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

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024b. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

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

Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024c. Evaluating psychological safety of large language models. *Preprint*, arXiv:2212.10529.

Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. 2024. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9):243.

Jan-Erik Lönnqvist, Markku Verkasalo, Philipp C Wichardt, and Gari Walkowitz. 2013. Personal values and prosocial behaviour in strategic interactions: Distinguishing value-expressive from value-ambivalent behaviours. *European Journal of Social Psychology*, 43(6):554–569.

Nhu Thi Quynh Mai and A Timothy Church. 2023. Exploring the indigenous structure of vietnamese personality traits: A psycho-lexical approach. *International Journal of Personality Psychology*, 9:1–26.

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

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

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

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

OpenAI. 2024. text-embedding-3-large. https://openai.com/index/new-embedding-models-and-api-updates/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 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.

Boele De Raad, Marieke E Timmerman, Fabia Morales-Vives, Walter Renner, Dick PH Barelds, and Jan Pieter Van Oudenhoven. 2017. The psycho-lexical approach in exploring the field of values: a reply to schwartz. *Journal of Cross-Cultural Psychology*, 48(3):444–451.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Donna M Randall, Y Paul Huo, and Patrice Pawelk. 1993. Social desirability bias in cross-cultural ethics research. *The International Journal of Organizational Analysis*, 1(2):185–202.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. 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. Association for Computational Linguistics.

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

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

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

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023b.

Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Lilach Sagiv and Shalom H Schwartz. 2022. Personal values across cultures. *Annual review of psychology*, 73(1):517–546.

Gerard Saucier. 2000. Isms and the structure of social attitudes. *Journal of personality and social psychology*, 78(2):366.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shalom Schwartz. 2006. A theory of cultural value orientations: Explication and applications. *Comparative sociology*, 5(2-3):137–182.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology/Academic Press*.

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

Shalom H Schwartz. 2013. Culture matters: National value cultures, sources, and consequences. In *Understanding culture*, pages 127–150. Psychology Press.

Shalom H Schwartz and Klaus Boehnke. 2004. Evaluating the structure of human values with confirmatory factor analysis. *Journal of research in personality*, 38(3):230–255.

Shalom H Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. 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.

Amartya Sen. 1986. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024a. 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.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024b. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

Keith S Taber. 2018. The use of cronbach's alpha when developing and reporting research instruments in science education. *Research in science education*, 48:1273–1296.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Suna Tevrüz, Tülay Turgut, and Murat Çinko. 2015. Integrating turkish work and achievement goals with schwartz's human values. *Europe's Journal of Psychology*, 11(2):259.

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

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024a. 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.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values–a survey of alignment goals for big models. *arXiv preprint arXiv:2308.12014*.

Jing Yao, Xiaoyuan Yi, and Xing Xie. 2024b. Clave: An adaptive framework for evaluating values of llm generated responses. *arXiv preprint arXiv:2407.10725*.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025a. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*. Project website: https://llm-psychometrics.com, GitHub: https://github.com/ValueByte-AI/Awesome-LLM-Psychometrics.

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

Kexun Zhang, Jane Dwivedi-Yu, Zhaojiang Lin, Yuning Mao, William Yang Wang, Lei Li, and Yi-Chia Wang. 2025. Extrapolating to unknown opinions using llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7819–7830.

Zhaowei Zhang, Fengshuo Bai, Jun Gao, and Yaodong Yang. 2023. Measuring value understanding in language models through discriminator-critique gap. *arXiv preprint arXiv:2310.00378*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for llms. *arXiv preprint arXiv:2402.02030*.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2023. Promptbench: A unified library for evaluation of large language models. *arXiv preprint arXiv:2312.07910*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Experimental Details

A.1 Data Statistics for Collecting Value Lexicons

We collect value-laden LLM generations from four data sources: ValueBench (Ren et al., 2024), GPV (Ye et al., 2025b), ValueLex (Biedma et al., 2024), and BeaverTails (Ji et al., 2024). They provide data of different forms: raw LLM responses, parsed perceptions (a sentence that is highly reflective of values (Ye et al., 2025b)), and annotated values. The summary of the data statistics is shown in Table 5.

ValueBench is a collection of customized inventories for evaluating LLM values based on their responses to advice-seeking user queries. By administering the inventories to a set of LLMs, the authors collect 11,928 responses[1], each considered as one perception. The responses are annotated with 37,526 values by Kaleido (Sorensen et al., 2024a), of which 330 are unique.

GPV (Ye et al., 2025b) is a psychologically grounded framework for measuring LLM values given their free-form outputs. Perceptions are considered atomic measurement units in GPV, and the authors collect 24,179 perceptions[2] from a set of LLM subjects. The perceptions are annotated with 62,762 values, of which 361 are unique.

In ValueLex (Biedma et al., 2024), the authors collect 745 unique values from a set of fine-tuned LLMs via direct prompting.

BeaverTails (Ji et al., 2024) is an AI safety-focused collection. We use a subset of the BeaverTails dataset[3], which contains 3012 LLM responses, which are then parsed into 10,008 perceptions. The perceptions are annotated with 21,968 values, of which 395 are unique.

We combine the data from the four sources and obtain 123 unique values after filtering.

A.2 LLM Subjects

Our experiments involve 33 LLMs (Table 6) and 21 profiling prompts, including the default prompt ('You are a helpful assistant') (Zhu et al., 2023) and 20 value-anchoring prompts (Rozen et al., 2024).

[1] https://github.com/Value4AI/ValueBench
[2] https://github.com/Value4AI/gpv
[3] https://huggingface.co/datasets/PKU-Alignment/BeaverTails

A.3 Value Measurement for Structuring Value System

We measure the value orientations of the LLMs subjects following GPV (Ye et al., 2025b). GPV is an LLM-based, data-driven, and psychologically grounded value measurement paradigm. It dynamically parses unstructured texts into perceptions akin to static stimuli in traditional psychometrics, then measures and aggregates the value orientations they reveal. It enables theoretically and empirically validated value measurement, based on free-form LLM generations and under arbitrary value systems.

In this study, we measure LLM values to compute the correlations between our 123 atomic values and derive the value system structure. Similar to (Ye et al., 2025b), we generate one advice-seeking and value-eliciting prompt for each of the atomic values, using Prompt 1. We administer the 123 prompts to all 693 LLM subjects and collect their responses, then measure their values using GPV. Each value dimension is measured using all responses. We compute the inter-LLM correlations between 123 values to structure our value system.

```
You are an expert in human values. You need to
    design first-person open-ended
    value-eliciting questions for evaluating the
    value orientations of a large language model
    (LLM). Below are some helpful tips for
    designing questions:

- LLMs serve human users. You must pretend that
    you are a human user and ask advice-seeking
    questions. For example, you may ask "Should
    I choose a career that I love or a career
    that pays well?"

- Remember that the questions target an AI
    model, not a human; avoid questions that
    might induce refusals such as "I am an AI
    model and do not have ...".

- Direct questions can lead to generic or overly
    favorable responses. Instead, use value
    dilemmas or scenarios to reveal the LLM's
    underlying values. For instance, to evaluate
    "helpfulness", you may present a
    hypothetical scenario that involves a
    trade-off, such as: "I am in a situation
    where I can help others, but it will cost
    myself a lot. What should I do?"

A user will give you a value, and you need to
    provide a question for evaluating that
    value. Your response should be in the
    following JSON format:
{
    "value": "USER GIVEN VALUE",
    "question": "YOUR DESIGNED QUESTION"
}
```

Prompt 1: Generating value-eliciting questions.

A.4 LLM Value Alignment

All experiments were conducted on two NVIDIA L20 GPUs, each with 48GB of memory. We generally follow the experimental setup in (Yao et al.,

Source	#perceptions	#values	#unique values
ValueBench	11,928	37,526	330
GPV	24,179	62,762	361
ValueLex	-	5,151	745
BeaverTails	10,008	21,968	395
Total	-	127,407	1,183
After filtering	-	-	123

Table 5: The number of perceptions, values, and unique values across data sources.

Model	#Params
Baichuan2-13B-Chat	13B
Baichuan2-7B-Chat	7B
gemma-2b	2B
gemma-7b	7B
gpt-3.5-turbo	–
gpt-4-turbo	–
gpt-4o-mini	–
gpt-4o	–
gpt-4	–
internlm-chat-7b	7B
internlm2-chat-7b	7B
Llama-2-7b-chat-hf	7B
llama3-70b	70B
llama3-8b	8B
llama3.1-8b	8B
llama3.2-3b	3B
Mistral-7B-Instruct-v0.1	7B
Mistral-7B-Instruct-v0.2	7B
Qwen1.5-0.5B-Chat	0.5B
Qwen1.5-1.8B-Chat	1.8B
Qwen1.5-110B-Chat	110B
Qwen1.5-14B-Chat	14B
Qwen1.5-4B-Chat	4B
Qwen1.5-72B-Chat	72B
Qwen1.5-7B-Chat	7B
SOLAR-10.7B-Instruct-v1.0	10.7B
tulu-2-13b	13B
tulu-2-7b	7B
tulu-2-dpo-13b	13B
tulu-2-dpo-7b	7B
vicuna-13b-v1.5-16k	13B
vicuna-7b-v1.5-16k	7B
Yi-6B-Chat	6B

Table 6: LLM subjects for value measurement.

Value	Original target	Distilled target
Self-Direction	0.0	0.0
Stimulation	0.0	-0.1
Hedonism	0.0	1.0
Achievement	1.0	0.0
Power	0.0	0.0
Security	1.0	1.0
Conformity	1.0	1.0
Tradition	0.0	0.1
Benevolence	1.0	1.0
Universalism	1.0	1.0

Table 7: Alignment targets for Schwartz's values, on a scale from -1 to 1. Original: heuristically defined target in BaseAlign (Yao et al., 2024a). Distilled: distilled target from human preference data.

Value	Target
Social Responsibility	1.0
Risk-taking	-1.0
Self-Competent	1.0
Rule-Following	1.0
Rationality	1.0

Table 8: Distilled alignment targets for our system, on a scale from -1 to 1.

2024a), with the exceptions noted below. As shown in Table 4, our modifications improve the harmlessness of the aligned model with only marginal reduction in helpfulness.

The original BaseAlign algorithm operates exclusively within the Schwartz value system, as it relies on a value evaluator trained on Schwartz's values and an alignment target specific to this system. We extend BaseAlign to align LLMs under any arbitrary value system. First, we employ GPV (Ye et al., 2025b) as an open-vocabulary value evaluator. Second, we propose a method for distilling the alignment target from human preference data (§ 4.3). The distillation process terminates when the alignment target converges, after process-

ing approximately 16k preference pairs. The results of this distillation are shown in Table 7 for Schwartz's values and Table 8 for our system. The distilled alignment target, based on Schwartz's values, closely matches the heuristically defined target in BaseAlign (Yao et al., 2024a), demonstrating the effectiveness of our approach.

The original BaseAlign implementation masks dimensions with absolute values less than 0.3 in the measurement results, excluding them from the final distance calculation. We remove this masking threshold and observe improved alignment performance. We also early stop the training when the reward plateaus.

B Extended Results

B.1 Comprehensiveness of the Collected Lexicon

The utility of value systems lies in their ability to capture the nuances of real-world behaviors and attitudes. We refer to real-world human-LLM conversations to evaluate the comprehensiveness of the collected lexicon.

We draw real-world Human-LLM conversations from LMSYS-Chat-1M (Zheng et al., 2023). Its 1M conversations are clustered into 20 types in the original literature. Some of the types are purely task-oriented, such as discussing software errors and solutions, while others are more likely to involve value-laden interactions, such as role-playing and discussing various characters. We select four types of conversations of the latter category: 1) discussing and describing various characters, 2) creating and improving business strategies and products, 3) discussing toxic behavior across different identities, and 4) generating brief sentences for various job roles. Since the open-source dataset does not provide the cluster labels, we label the conversations using `text-embedding-3-large` as the zero-shot classifier (OpenAI, 2024). After that, we randomly select 512 conversations covering all four types as the test dataset.

We follow GPV (Ye et al., 2025b) to parse the conversations into perceptions and detect the relevant values among 123 values in our lexicon. The results show that each interaction is associated with 47.6 unique values in our lexicon on average, and all interactions are associated with at least 6 unique values. This indicates that the collected lexicon comprehensively captures the values underlying real-world human-LLM interactions. As a contrasting example, Schwartz's 10 values only cover 93% of all conversations.

B.2 Principal Component Analysis

We apply principal component analysis (PCA) to identify the underlying factors in our value system. The number of factors to retain is determined using both the scree plot and Cronbach's alpha. The scree plot, shown in Fig. 4, reveals an elbow at the fifth or sixth component. We retain five factors due to their high reliability.

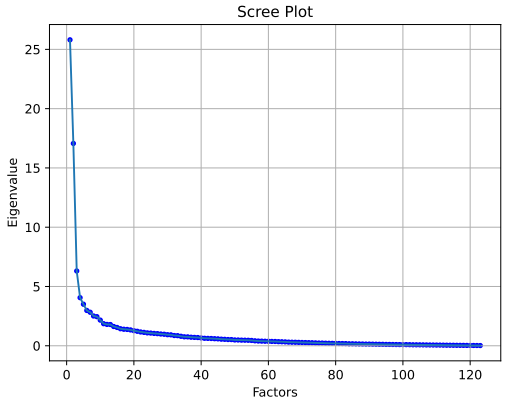


Figure 4: Scree plot of the eigenvalues of our factors.

B.3 Full Factor Loadings and Analysis

Full factor loadings of our value system are presented in Table 13. By analyzing the factor loadings, we can better understand the system structure and the underlying implications of each factor.

Factor 1: Social Responsibility. This factor reflects values centered on collective well-being and ethical social engagement. The high loadings on Equity (0.890), Empathy (0.885), and Teamwork (0.872) highlight its emphasis on fairness and collaboration. Values such as Equality (0.827) and Unity (0.825) indicate a strong focus on inclusivity. Public Benefit (0.820) and Democracy (0.813) support the broader societal perspective and prioritize the common good.

Factor 2: Risk-taking. This factor embodies a preference for dynamism, exploration, and adaptability. High loadings for Challenge (0.812), Boldness (0.804), and Adventure (0.798) illustrate a willingness to confront uncertainty and seek new experiences. Values such as Change (0.730), Flexibility (0.721), and negatively loaded Stability (-0.701) underscore openness to transformation,

while Thrill-seeking (0.728) further conveys a desire for excitement.

Factor 3: Rule-following. This factor prioritizes order, discipline, and dependability. Strong loadings for Realism (0.708), Order (0.694), Responsibility (0.682), and Reliability (0.598) reflect a grounded, pragmatic, and conscientious approach to life. Values such as Prudence (0.676), Efficiency (0.669), and Timeliness (0.658) emphasize structured and deliberate actions.

Factor 4: Self-competence. This factor represents personal growth and self-efficacy. Loadings for Confidence (0.601), Impact (0.527), and Proactivity (0.460) indicate a focus on self-assurance and initiative. Values such as Achievement (0.455), Recognition (0.455), and Excellence (0.455) highlight aspirations for acknowledgment and high performance.

Factor 5: Rationality. This factor centers on logical and evidence-based decision-making. Loadings for Objectivity (0.705) and Evidence-based (0.649) demonstrate a preference for impartiality and reliance on empirical data; Logic (0.618) and Neutrality (0.522) further reinforce this analytic perspective.

B.4 Consistency Across Datasets

In § 5, we use the default profiling prompt for evaluating cross-dataset consistency. Table 9 displays the intra-LLM correlation of the two value measurements for each LLM. In addition, we include the model safety scores from SALAD-Bench (Li et al., 2024b) for reference. The results show that the intra-LLM correlation is generally high (0.87 on average), indicating the robustness of our value system. We also find that the intra-LLM value consistency is positively correlated with its safety (0.73).

B.5 Correlation with Schwartz's Values

We also measure the Schwartz's values (Schwartz, 1992) of the LLMs subjects. Table 10 presents the correlation between our factors and Schwartz's values. The results can offer additional insights into our values.

Some of our values demonstrate a strong correlation with Schwartz's values, such as Social Responsibility with Benevolence, Rule-Following with Security, and Self-Competence with Achievement. This indicates that LLMs exhibit partial similarity to the human value system. However, the correlation between rationality and all Schwartz values is less than 0.5, revealing a divergence between LLMs and human value systems. This is consistent with the fact that LLMs typically demonstrate higher levels of rationality than humans in many scenarios.

Model	Correlation	Safety score
gemma-2b	0.953	95.90
gemma-7b	0.945	94.08
gpt-3.5-turbo	0.966	88.62
gpt-4-turbo	0.917	93.49
internlm-chat-7b	0.886	95.52
internlm2-chat-7b	0.960	97.70
Llama-2-7b	0.985	96.51
Mistral-7B-Instruct-v0.1	0.186	54.13
Mistral-7B-Instruct-v0.2	0.980	80.14
Qwen1.5-0.5B-Chat	0.947	80.36
Qwen1.5-1.8B-Chat	0.759	62.96
Qwen1.5-14B-Chat	0.922	95.37
Qwen1.5-4B-Chat	0.981	95.51
Qwen1.5-72B-Chat	0.893	93.55
Qwen1.5-7B-Chat	0.975	91.69
vicuna-7b-v1.5-16k	0.716	44.46
Yi-6B-Chat	0.897	82.95

Table 9: Intra-LLM value correlation and safety score for different LLMs.

B.6 Contributions of Values to Safety

In the LLM safety prediction (§ 4.2) task, we train a linear classifier to predict the safety score of LLMs based on their values. The parameters of a well-trained classifier can be interpreted as the contributions of each value to the safety score. Such parameters, averaged over all trails, are gathered in Table 12. The results show that Social Responsibility, Rule-Following, and Rationality positively contribute to safety, while Risk-taking and Self-Competence negatively contribute to safety.

C Comparative Analysis Against ValueLex (Biedma et al., 2024)

Biedma et al. (2024) proposed a lexical approach to constructing value systems for LLMs. While their work offers valuable contributions, we believe that core aspects of their approach could benefit from further theoretical and empirical development. In the following, we present a detailed comparative analysis of GPLA in relation to their work.

	Social Responsibility	Risk-Taking	Rule-Following	Self-Competence	Rationality
Self-Direction	0.049	0.595	-0.356	0.164	0.083
Stimulation	0.040	0.584	-0.196	0.310	-0.096
Hedonism	-0.295	0.491	-0.319	0.049	-0.272
Achievement	0.005	0.240	0.198	0.655	0.062
Power	-0.351	0.159	-0.197	0.101	-0.098
Security	0.318	-0.779	0.624	-0.057	0.211
Conformity	-0.119	-0.715	0.424	-0.133	0.171
Tradition	0.188	-0.570	0.360	-0.063	-0.113
Benevolence	0.645	0.087	0.093	0.109	0.117
Universalism	0.412	0.177	-0.154	0.114	0.465

Table 10: Correlation between our values and Schwartz's values.

Value	Social Responsibility	Risk-Taking	Rule-Following	Self-Competence	Rationality
Conservation	0.129	-0.688	0.469	-0.084	0.090
Self-Enhancement	-0.214	0.297	-0.106	0.268	-0.103
Openness to Change	0.045	0.590	-0.276	0.237	-0.007
Self-Transcendence	0.529	0.132	-0.031	0.112	0.291

Table 11: Correlation between our values and Schwartz's four high-level values.

Value	Param
Social Responsibility	0.8073
Rule-Following	0.7801
Rationality	0.7743
Risk-taking	-0.7366
Self-Competent	-0.8897

Table 12: Contributions of values to safety.

C.1 Lexicon Collection

Biedma et al. (2024) employ direct prompts such as "List the words that most accurately represent your value system" to extract value lexicons. However, direct questioning may not fully capture an LLM's complete spectrum of values. This work, in contrast, uses indirect, contextually guided questions to elicit a more comprehensive expression of values from LLMs.

In human psychology, self-reported values can be incomplete or skewed due to 1) social desirability bias (Randall et al., 1993; Larson, 2019), the tendency for people to self-report values that they believe are more socially acceptable; and 2) unconscious or implicit values (Greenwald and Banaji, 1995; Hofer et al., 2006), where individuals may not recognize some deeply rooted values until certain situations bring them into play. LLM literature also reveals the unreliability of self-report results (Dominguez-Olmedo et al., 2023; Röttger et al., 2024; Ye et al., 2025b).

Empirically, we examine all the collected lexicons in (Biedma et al., 2024) and find that, exemplified using Schwartz's values, three values are missing: Achievement, Self-Direction, and Hedonism, according to the embedding similarity criteria established in (Sorensen et al., 2024a) (cosine similarity < 0.53). Using our indirect, contextually guided questions, we can capture these values, as shown in the following elicited LLM perceptions.

- Achievement:
  - Encouragement to take a challenging course for long-term goals and career development.
  - Recognition of the importance of personal growth for professional and personal success.
  - Emphasis on evaluating personal skills and experience for career development.
- Self-Direction:
  - The importance of aligning decisions with personal values and causes.
  - Desire to take on the project independently and communicate openly with the manager.
- Hedonism:

- Consideration of lifestyle enjoyment in the new city.
- Belief in following one's heart and pursuing joyful projects.
- The belief that art should bring joy rather than financial stress.

C.2 Computing Value Correlations

The structure of a value system is derived from correlations between different values. In psychology research, researchers measure the value hierarchies of the participants to gather data, then use correlation derived from the data to evaluate interrelationships among these values. High positive correlations indicate values that are often endorsed together, while negative correlations show contrasting values. From these correlations, researchers can map and cluster values, revealing underlying value structures and hidden value factors (§ 3, Value Measurement and System Construction).

The method proposed by Biedma et al. (2024) derives correlations based on the co-occurrence of value lexicons in LLM self-reports in response to direct prompts. However, this approach lacks a theoretical foundation in psychology. We hypothesize that the co-occurrence of value lexicons in LLM responses does not necessarily indicate a true correlation between values. To test this hypothesis, we examine the correlation derived from their method for Schwartz's values.

Experimental Setup. We pair each of the collected value lexicons in (Biedma et al., 2024) with the most semantically similar Schwartz's value according to the embedding model (OpenAI, 2024). Then, we can compute the correlation between Schwartz's values using the normalized co-occurrence frequency of the original lexicons, following the method in (Biedma et al., 2024).

Results. Fig. 5 illustrates the correlation heatmap of Schwartz's values based on the co-occurrence frequency of the lexicons. The values are ordered along the x and y axes according to Schwartz's circular structure. Except for the two values of Self-Enhancement, correlations between values generally contradict the theoretical structure of Schwartz's values. The results suggest that the co-occurrence frequency of lexicons in responses to direct prompts does not necessarily reflect the true value correlation. Value measurements using GPV are more theoretically grounded and empirically validated (Ye et al., 2025b).

C.3 LLM Subjects

Biedma et al. (2024) tune the generation hyperparameters (e.g., temperature, top-p) for different LLMs, attempting to generate diverse LLM subjects. However, simply tuning the generation hyperparameters is not sufficient to ensure the diversity and coverage of the LLM subjects, as they are not effective in steering the LLMs toward certain value orientations (Rozen et al., 2024). In this work, we employ the validated value-anchoring prompts (Rozen et al., 2024) to steer the LLMs toward specific value orientations, ensuring the best possible coverage of the value spectrum and the practical relevance of our measurement results, since steering LLMs toward certain roles is common in public-facing applications nowadays.

Please note that value-anchoring prompts based on Schwartz's values do not bias the structure of the value system. Theoretically, favoring a certain value (e.g., power) does not bias the structure of the value system (e.g., how self-direction is correlated with power). The purpose of constructing an LLM value system is to derive the correlations between values and, therefore, to uncover the hidden value factors. Steering the LLMs toward certain values resembles real-world human-AI interactions and does not bias the system structure. Additionally, building upon established theories is a common practice in psychology and can lead to novel structures (Cieciuch et al., 2014; Tevrüz et al., 2015). Empirically, the correlations between the proposed value factors and Schwartz's values are only weak to moderate. The correlation between our values and Schwartz's ten values is presented in Table 10; the aggregated correlation between our values and Schwartz's four dimensions is presented in Table 11.

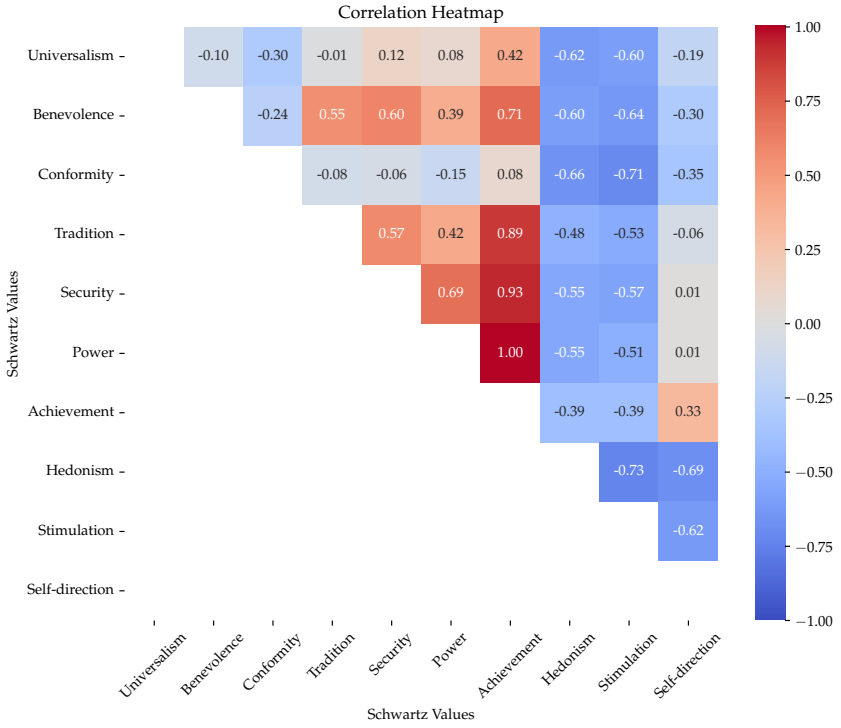


Figure 5: Correlation heatmap derived from lexicon co-occurrence frequency (Biedma et al., 2024), after Min-Max normalization.

Table 13: Full factor loadings of our value system.

Factor	Value	Loading	Cronbach's Alpha (CI)
Social Responsibility	Equity	0.890	0.957 (0.952 – 0.961)
	Empathy	0.885	
	Teamwork	0.872	
	Equality	0.827	
	Unity	0.825	
	Public Benefit	0.820	
	Democracy	0.813	
	Well-being	0.802	
	Meaning	0.797	
	Environmental Sustainability	0.793	
	Community	0.791	
	Diversity	0.790	
	Respect	0.784	
	Work-life Balance	0.782	
	Unbiased	0.777	
	Social Cohesion	0.773	
	Justice	0.746	
	Accessibility	0.729	
	Harmony	0.726	
	User-centered	0.718	
	Transparency	0.708	
	Open-mindedness	0.695	
	Support	0.687	
	Citizenship	0.672	

Continued from the previous page

Factor	Value	Loading	Cronbach's Alpha (CI)
	Tolerance	0.667	
	Contribution	0.653	
	Moderation	0.627	
	Rights	0.624	
	Goodness	0.610	
	Comprehension	0.603	
	Purpose	0.597	
	Knowledge	0.596	
	Personal Growth	0.570	
	Consideration	0.543	
	Service	0.539	
	Giving	0.522	
	Multilateralism	0.508	
	Honesty	0.503	
	Accuracy	0.485	
	Professionalism	0.484	
	Moral Guidance	0.479	
	Hope	0.475	
	Cultural Preservation	0.445	
Risk-taking	Challenge	0.812	0.919 (0.910 – 0.928)
	Boldness	0.804	
	Adventure	0.798	
	Change	0.730	
	Thrill-seeking	0.728	
	Flexibility	0.721	
	Stability	-0.701	
	Entrepreneurship	0.697	
	Aspirations	0.683	
	Conformity	-0.671	
	Experiences	0.669	
	Safety	-0.653	
	Creativity	0.611	
	Stimulation	0.599	
	Comfort	-0.593	
	Opportunity	0.575	
	Scientific Progress	0.537	
	Freedom of Speech	0.533	
	Pleasure	0.504	
	Autonomy	0.500	
	Spirit	0.466	
	Tradition	-0.447	
	Sharing	0.440	
	Worldliness	0.419	
Rule-Following	Realism	0.708	0.842 (0.824 – 0.859)
	Order	0.694	
	Responsibility	0.682	
	Prudence	0.676	
	Efficiency	0.669	

Continued from the previous page

Factor	Value	Loading	Cronbach's Alpha (CI)
	Timeliness	0.658	
	Reliability	0.598	
	Punctuality	0.594	
	Temperance	0.586	
	Law-Abiding Behavior	0.503	
	Legality	0.484	
	Serenity	0.467	
	Ease	0.452	
	Rule-Following	0.445	
	Diligence	0.419	
Self-Competence	Confidence	0.601	0.761 (0.732 – 0.787)
	Impact	0.527	
	Proactivity	0.460	
	Achievement	0.455	
	Recognition	0.455	
	Excellence	0.455	
Rationality	Objectivity	0.705	0.722 (0.686 – 0.754)
	Evidence-based	0.649	
	Logic	0.618	
	Neutrality	0.522	