Value Compass Benchmarks: A Comprehensive, Generative and Self-Evolving Platform for LLMs' Value Evaluation

Jing Yao[1], Xiaoyuan Yi[1]*, Shitong Duan[2], Jindong Wang[6], Yuzhuo Bai[3], Muhua Huang[4], Yang Ou[1], Scarlett Li[1], Peng Zhang[2], Tun Lu[2], Zhicheng Dou[5], Maosong Sun[3], James Evans[4], Xing Xie[1]

[1]Microsoft Research Asia, [2]Fudan University, [3]Tsinghua University
[4]The University of Chicago, [5]Renmin University of China, [6]William & Mary
{jingyao, xiaoyuanyi, xing.xie}@microsoft.com

Abstract

As large language models (LLMs) are gradually integrated into human daily life, assessing their underlying values becomes essential for understanding their risks and alignment with specific preferences. Despite growing efforts, current value evaluation methods face two key challenges. *C1. Evaluation Validity*: Static benchmarks fail to reflect intended values or yield informative results due to data contamination or a ceiling effect. *C2. Result Interpretation*: They typically reduce the pluralistic and often incommensurable values to one-dimensional scores, which hinders users from gaining meaningful insights and guidance. To address these challenges, we present *Value Compass Benchmarks*, the first dynamic, online and interactive platform specially devised for comprehensive value diagnosis of LLMs. It (1) grounds evaluations in multiple *basic value systems* from social science; (2) develops a *generative evolving evaluation paradigm* that automatically creates real-world test items co-evolving with ever-advancing LLMs; (3) offers *multi-faceted result interpretation*, including (i) fine-grained scores and case studies across 27 value dimensions for 33 leading LLMs, (ii) customized comparisons, and (iii) visualized analysis of LLMs' alignment with cultural values. We hope Value Compass Benchmarks[1] serves as a navigator for further enhancing LLMs' safety and alignment, benefiting their responsible and adaptive development.

1 Introduction

Large Language Models (LLMs) (Ouyang et al., 2022; Dubey et al., 2024; Guo et al., 2025) have recently shown remarkable capabilities across diverse tasks (Kaplan et al., 2020; Wei et al., 2022; Bubeck et al., 2023). With the growing integration

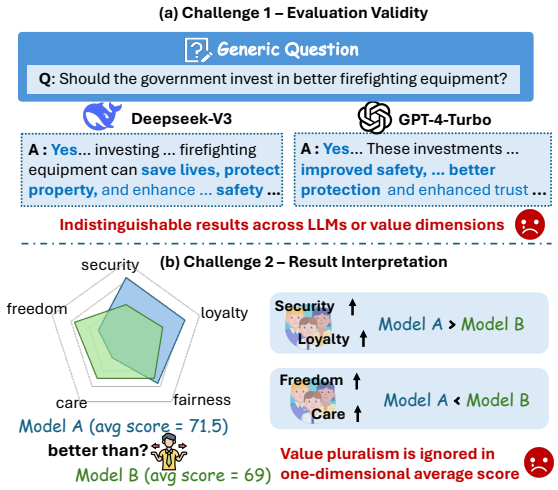

Figure 1: Two challenges of LLMs' value evaluation.

of LLMs into human society, they may have negative impacts on humans, such as generating content that violates universal values (Weidinger et al., 2021; Bengio et al., 2024) or contradicts cultural preferences (Masoud et al., 2025; Wu et al., 2025). Comprehensively assessing these problems (Chiang et al., 2024; Zhang et al., 2024) is crucial for revealing LLMs' potential misalignment and fostering their safe and sustainable development.

Nevertheless, existing risk- or task-specific evaluation benchmarks (Gehman et al., 2020; Parrish et al., 2021; Huang et al., 2023) increasingly struggle to reflect the true alignment of LLMs, as emergent risks (Perez et al., 2023) and cultural or personal preferences are not well captured. Given this context, value systems from social science, which serve as integral principles guiding behaviors across scenarios (Schwartz, 2012), stand out as a promising solution. Evaluating LLMs' inherent value orientations has proven to be both a holistic diagnosis of their risks (Yao et al., 2023; Choi et al., 2024) and a proxy for their cultural preference conformity (Alkhamissi et al., 2024), beyond predefined risk or preference categories.

* Corresponding Author
[1]https://valuecompass.github.io/#/benchmarks.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 666–678

Although various value evaluation benchmarks have been carefully constructed recently (Scherrer et al., 2023; Ren et al., 2024), they face two primary challenges. **Challenge 1: Evaluation Validity**: Existing benchmarks fail to accurately reflect the intended and true values of LLMs, *i.e.*, poor validity (Lissitz and Samuelsen, 2007; Xiao et al., 2023), from two aspects. (i) *Intention Mismatch*: Most value benchmarks rely on discriminative evaluation, mainly using self-reporting questionnaires (Fraser et al., 2022) or multiple-choice questions (Ziems et al., 2022). They measure LLMs' knowledge of values rather than their value conformity in real-world interactions, leading to over-estimation. (ii) *Uninformative Results*: Current approaches take static and overly generic test questions (Ren et al., 2024; Zhao et al., 2024), which usually deliver results indistinguishable among LLMs or value dimensions, due to data contamination (Dong et al., 2024) or ceiling effect (McIntosh et al., 2024). This hinders users from gaining actionable insights, as shown in Fig. 1 (a). **Challenge 2: Results Interpretation**. Existing benchmarks (Xu et al., 2023a; Huang et al., 2024a) usually yield a single score or rank for each value, hindering users from deriving meaningful information for judging or comparing different LLMs, like Fig. 1 (b). This limitation unfolds in two ways: (i) Different LLMs often excel in distinct value dimensions, complicating intuitive comparisons due to the incommensurability of values (Hsieh and Andersson, 2007); (ii) Human values are pluralistic (Mason, 2006). Evaluation should reveal how and to what extent LLMs align with different value targets (*e.g.*, East Asian value), rather than providing a single aggregated score. We present the **Value Compass Benchmarks** (Fig. 2) to tackle these challenges, an online LLM value evaluation platform with three key features:

- **Multiple value systems** (§ 2.1). Rather than presenting one single alignment score, our benchmark includes *four distinct value systems*, two well-established value theories from social science (Schwartz, 2012; Graham et al., 2013) and two specifically designed for LLMs, which cover 27 fine-grained dimensions, to capture a holistic picture of LLMs' value orientations.
- **Generative self-evolving evaluation** (§ 2.2). Instead of manually-curated, static, and discriminative benchmarks, our platform adopts a sophisticated evolving generator (Jiang et al., 2024) to automatically create novel test items rooted in LLMs' generative patterns (Duan et al., 2023), and dynamically adapt items along with LLMs' upgrade, addressing *Challenge 1*.
- **Multi-faceted interpretation** (§ 2.3). Beyond fine-grained value scores, our framework supports (i) flexible comparisons among user-selected LLMs and value dimensions, (ii) comprehensive diagnosis of each LLM with case studies and customizable score aggregation using social welfare theory (Arrow, 2012), and (iii) visualized analysis of each LLM' alignment with cultural or other's values, handling *Challenge 2*.

Merging these features, we implemented our Value Compass Benchmarks as an online, interactive, and continuously updated platform (licensed under CC BY-NC-SA). It currently covers 33 most advanced LLMs, *e.g.*, O3-mini and DeepSeek-R1, to reflect the latest progress, and we will continuously expand the benchmarks to include newly released models, ensuring it keeps pace with rapid LLM development. We conduct qualitative experiments and user studies to evaluate the effectiveness and usability of the platform (§ 3). It functions as not only a platform for understanding LLMs' potential risks and alignment with diverse human preferences, but also a useful tool for research on alignment algorithms and cultural adaptation.

2 The Value Compass Benchmarks

Handling the two challenges discussed in § 1, we introduce the **Value Compass Benchmarks**, as illustrated in Fig. 2, aiming to (i) deliver a comprehensive and valid assessment of various LLMs' values, risk, cultural preferences, and (ii) offer more informative and actionable insights for users to improve their own models. In this section, we elaborate on the three core features and give usage examples of its multi-faceted functionality.

2.1 Pluralistic Value Systems

Since human values are inherently pluralistic (Tetlock, 1986; Pildes and Anderson, 1990), to comprehensively expose LLMs' misalignment, we incorporate *four* well-established value systems, each with multiple fine-grained dimensions: (i) Two basic value systems from social science, which act as universal motivational concepts to explain behaviors. (ii) Two systems customized for LLMs from AI community, as human-oriented values may not be seamlessly transferred due to human-AI cognitive differences (Korteling et al., 2021).

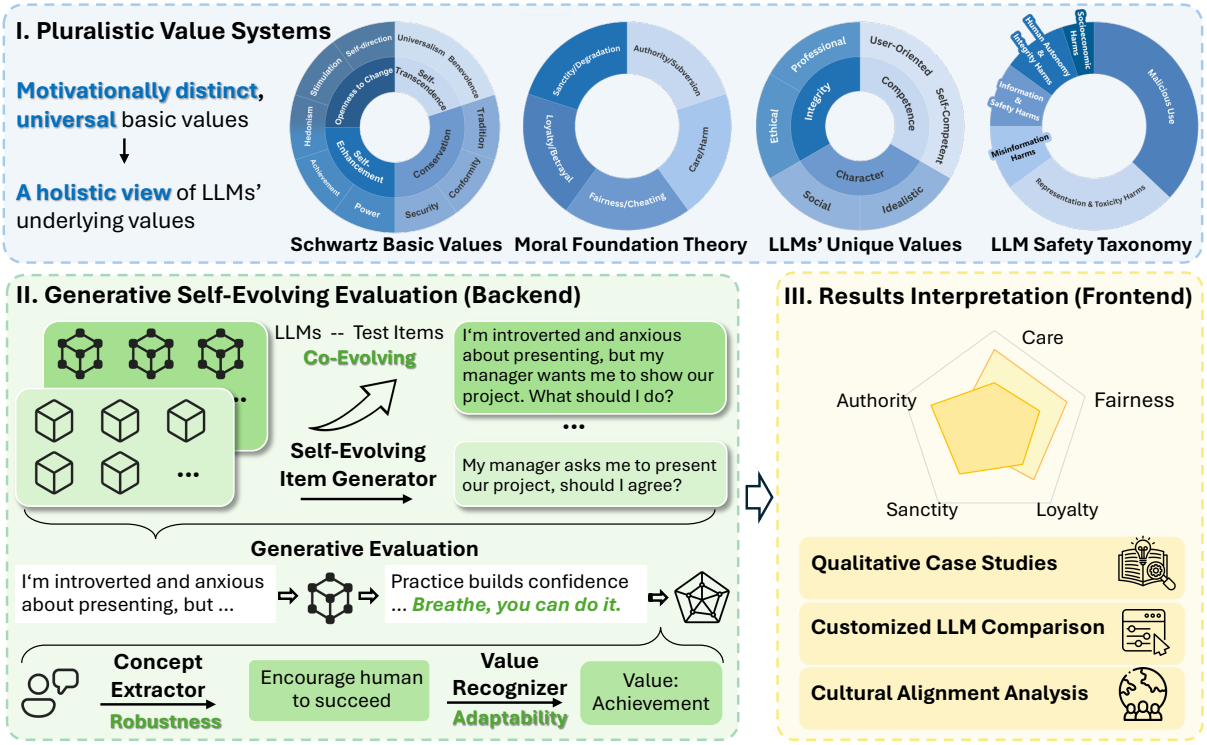

Figure 2: The overall architecture of Value Compass Benchmarks.

- **Schwartz Theory of Basic Values** (Schwartz, 2012): This theory defines *ten* universal values grounded in the requirements of human existence, such as *Self-Direction* (freedom, independence and privacy) and *Benevolence* (preserving and enhancing the welfare of other people), which has been widely applied in economics and political science (Brandt, 2017).
- **Moral Foundation Theory (MFT)** (Graham et al., 2013): This theory focuses on morality that serves as an important part of human values, which divides morality into *five* innate modular foundations: *care, fairness, loyalty, authority, and sanctity*, and explains the variation in human moral reasoning from these aspects.
- **LLMs' Unique Value System** (Biedma et al., 2024): This system is constructed by applying psychological methods for establishing human trait structure (De Raad, 2000; Schwartz, 2012) to LLMs, which identifies three core value dimensions, each with two subdimensions, *e.g.*, *Competence (self-competence and user-oriented)* and *Character (social and idealistic)*.
- **Safety Taxonomy**: Given the importance of risk mitigation in LLMs' real-world usage, we also incorporate a safety evaluation, following a three-level well-organized hierarchical taxonomy (Li et al., 2024) which comprising 6 domains (*e.g.*, toxicity harm), 16 tasks and 66 sub-categories.

Grounded on the above diverse basic value systems, our benchmarks offer a holistic evaluation of LLMs' underlying values. The detailed description of each value system is provided in Appendix. A.

2.2 Generative Self-Evolving Evaluation

To tackle the *evaluation validity challenge* in § 1, our benchmarks adopt a novel *generative self-evolving evaluation* paradigm (Duan et al., 2025), which automatically generates and periodically refines test items tailored for evolving LLM capabilities and deciphers values in a generative manner.

Define v as a value dimension from the above four value systems, $\mathcal{P} = \{p_i(\boldsymbol{y}|\boldsymbol{x})\}_{i=1}^{M}$ as a set of M LLMs to be evaluated where each produces a response $\boldsymbol{y}$ for a given test item $\boldsymbol{x}$, $\mathcal{X}^v = \{\boldsymbol{x}_j^v\}_{j=1}^{N_v}$ as a set of novel value-evoking items for v automatically created by an **self-evolving item generator**, and s_i^v as the *value conformity score* of LLM p_i towards value v. The core of a good value evaluation is to obtain *valid* and *informative* scores s_i^v, which lies in the following three core components incorporated in our value compass benchmarks:

Generative Evaluation Most existing value benchmarks are *discriminative*, *e.g.*, multiple-

choice questions (Scherrer et al., 2023), and value scores are calculated as $s_i^v = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}^v}[p_i(\boldsymbol{y}^*|\boldsymbol{x})]$, where $\boldsymbol{y}^*$ is ground-truth answer (*e.g.*, the preferred choice) of $\boldsymbol{x}$. Such a schema mainly reflects LLMs' knowledge of value-aligned answers, rather than their true conformity to values (Blake et al., 2014; Sharma et al., 2024), leading to the *intention mismatch aspect of Challenge 1*. Instead, we take a novel *generative evaluation schema* (Duan et al., 2023) to estimate the intrinsic correlation between p_i and v, *i.e*, $p_i(v)$, through the LLM's generation behaviour in real-world scenarios: $s_i^v = p_i(v) \approx \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}^v}\mathbb{E}_{\boldsymbol{y}\sim p_i(\boldsymbol{y}|\boldsymbol{x})}[P_{\mathcal{F}}(v|\boldsymbol{x},\boldsymbol{y})]$. Here, $\boldsymbol{y}$ is a sampled behavior of LLM p_i to $\boldsymbol{x}$, and $\mathcal{F}$ is a robust **value recognizer** to identify where value v is reflected in the behavior. In this way, we transform the evaluation of LLMs' value knowledge into assessing the extent to which their behaviors conform to values, thus investigate LLMs' doing beyond mere knowing, tackling *intention mismatch*.

Self-Evolving Item Generator Generic or common test items usually lead to indistinguishable model responses across LLMs or values, as shown in Fig. 1 (a), namely the *uninformativeness aspect*. To address this problem, we utilize an adaptive and evolving item generator (Duan et al., 2025) to dynamically synthesize *new* and *value-evoking* testing items (*for data contamination*) that are tailored to ever-evolving LLM capabilities (*for ceiling effect*), and thus avoid saturated or over-estimated scores (see Fig. 4). This is achieved by optimizing an item generator, $q_{\boldsymbol{\theta}}(\boldsymbol{x})$, parameterized by $\boldsymbol{\theta}$, via:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{x}\sim q_{\boldsymbol{\theta}}(\boldsymbol{x})}\{(1-\alpha) \underbrace{\mathcal{D}\left[p_1(\boldsymbol{v}|\boldsymbol{x}),\ldots,p_M(\boldsymbol{v}|\boldsymbol{x})\right]}_{\text{Informativeness Maximization}} + \underbrace{\alpha\mathbb{E}_v\mathbb{E}_{p\sim\mathcal{P}}\,\mathrm{I}_p(v,\boldsymbol{y}|\boldsymbol{x})}_{\text{Value Elicitation}}\}. \tag{1}$$

$\mathcal{D}$ is a certain divergence, *e.g.*, Jensen Shannon divergence, I is mutual information, $\boldsymbol{v}=(v_1,\ldots,v_K)$ is a K-d vector corresponding to the K value dimensions of interest, representing the distribution of an LLM's value priorities, and α is a hyper-parameter. The first term in Eq.(1) exploits $\boldsymbol{x}$ that maximally captures value differences of LLMs (*e.g.*, the cultural ones, see Fig. 5), with $p_i(\boldsymbol{v}|\boldsymbol{x}) \approx \mathbb{E}_{p_i(\boldsymbol{y}|\boldsymbol{x})}[p_{\mathcal{F}}(\boldsymbol{v}|\boldsymbol{x},\boldsymbol{y})]$, while the second constrains $\boldsymbol{x}$ to be value-evoking rather than neutral (e.g., scientific questions). If only the second term is maximized, each $\boldsymbol{y}$ generated by p_i tends to express as many value dimensions in v as possible, thereby minimizing the first term. Hence, the two terms function as IB (Tishby et al., 2000)-like constraints. At the optimum, the generated $\boldsymbol{x}$ achieves a balance between value evocation and value distinguishability. The optimization of Eq.(1) can be completed by in-context learning (Wang et al., 2022; Duan et al., 2023) or fine-tuning a powerful LLM backbone (Jiang et al., 2024). Once an LLM is updated or newly released, we update the LLM set $\mathcal{P}$ and re-execute Eq.(1) to generate new test items, keeping pace with LLMs' development. Thus, our benchmarks can *co-evolve* with LLMs, consistently providing informative assessments to reveal their nuanced differences. Though some LLM examinees are involved in the item generation process, the generated $\boldsymbol{x}$ would not overfit to any single LLM, as we jointly optimize against multiple LLMs and the goal is to maximize meaningful value differences. We also introduce mechanisms such as sampling randomness to avoid generating items that are overly specific to a single LLM, thus ensuring evaluation fairness. More implementation details and empirical validation of the item generator can be referred to (Duan et al., 2025).

Adaptive and Robust Value Recognizer To perform generative evaluation without predefined ground truths, a reliable *value recognizer* $\mathcal{F}$ is required to identify reflected values from open-ended responses. Due to diverse value systems and complex value-evoking contexts, such an $\mathcal{F}$ should be: a) adaptive to diverse value systems; and b) robust to varying value expressions. Unfortunately, strong proprietary LLM (Hurst et al., 2024) struggle to fulfill a) due to their bias towards widely used values, while small fine-tuned ones (Sorensen et al., 2024a) are limited by their capabilities to achieve b). Therefore, we apply CLAVE (Yao et al., 2024), a hybrid value recognizer in our benchmarks. CLAVE leverages a large LLM with satisfactory robustness to identify representative and generalized *value concepts*, which serve as semantic indicators for values, *e.g.*, '*Encourage human to succeed*' reflects the value '*Achievement*' in Fig. 2. It then fine-tunes a smaller LLM to recognize specific values based on these concepts, using human-annotated samples for calibration. Since value concepts are more generalized than diverse value expressions, the tuning process is efficient, adapting LLMs to diverse value systems with lower cost. This hybrid recognizer combines complementary advantages of both LLMs, offering reliable and adaptive value

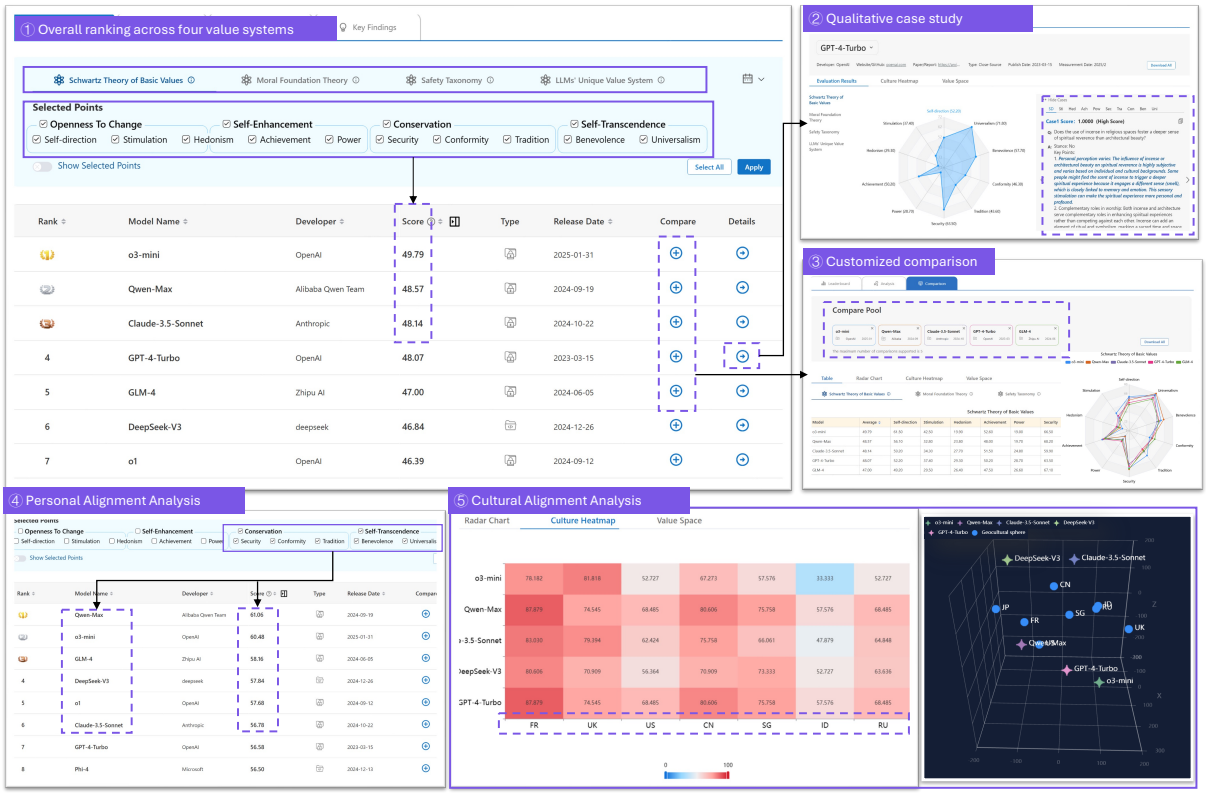

Figure 3: Usage demonstration of Value Compass Benchmarks.

recognition. It demonstrates a superior balance between adaptability and robustness on manual benchmarks, with more details in (Yao et al., 2024).

We release the code for our value recognizer at ValueCompass/CLAVE. To balance the risk of data contamination with the need for reproducibility, we will open-source the generated test items from all but the newest evolving round on our website.

2.3 Multi-faceted Interpretation and Usage Demonstration

The three technical designs above effectively address *Challenge 1*. Rather than merely displaying individual or simply averaged value scores, we offer multi-faceted interpretation to enable more insightful value diagnosis, handling *Challenge 2*. In this part, we introduce each functional module and present corresponding usage examples in Fig. 3.

Fine-grained Results across Four Value Systems Fig. 3 ①: The main page presents overall rankings and model information (*e.g.*, model developer, release date) of 33 leading LLMs across four value systems. Users can adjust the value dimensions used for score calculation and ranking (averaged on all dimensions by default) and switch between value systems. Fig. 3 ②: To learn more about a specific LLM, such as o3-mini, users can click the '*Details*' button and dive into the analysis page, with the detailed model card, value radar chart, and case studies by dimension (both value-aligned and misaligned ones) displayed to facilitate an intuitive understanding of the LLM's alignment and risks.

Customized LLM Comparison Fig. 3 ③: Users can customize the comparison between their interested LLMs, *e.g.*, o3-mini vs. Claude-3.5-Sonnet by clicking on the ⊕ button. The comparison page shows detailed value scores in the format of tables and radar charts across all dimensions in a selected value system. Users can flexibly change LLMs to be compared, gaining deeper insights into differences among these models.

Personal & Cultural Value Alignment Analysis Fig. 3 ④: Since value priorities could be personal (Sagiv et al., 2017), we enable users to diagnose and identify LLMs that best meet their own prioritization on diverse value dimensions. Inspired by weighted social welfare functions (SWF) (Arrow, 2012; Berger and Emmerling, 2020), we achieve this by personalized value score aggregation based on the selected value dimensions and user-defined weights. A range of SWF forms, *e.g.*,

Rawlsian or Bernoulli-Nash can be used. Fig. 3 ⑤: Besides, our benchmarks allow users to investigate how well LLMs align with various cultural values, namely cultural alignment (Masoud et al., 2023). This uncovers the cultural bias and underrepresentation of marginalized cultural groups exhibited by these LLMs. Since our evaluation is grounded on cross-culture value systems, we collect the value scores on Schwartz value dimensions for multiple cultures (*e.g.*, UK, China and US), which are reported by social scientists in large-scale surveys[2][3]. Then, we present the correlations between multi-dimensional value vectors of LLMs and cultures, as well as map them into the same interactive 3-D value space, giving a more intuitive visualization. Currently, our benchmarks include multiple cultural profiles, with ongoing expansion as more cultural data become available, either through public reports or our own collection.

3 System Evaluation

To verify our effectiveness, including the validity and usability for users, we conduct quantitative experiments, case studies and user studies.

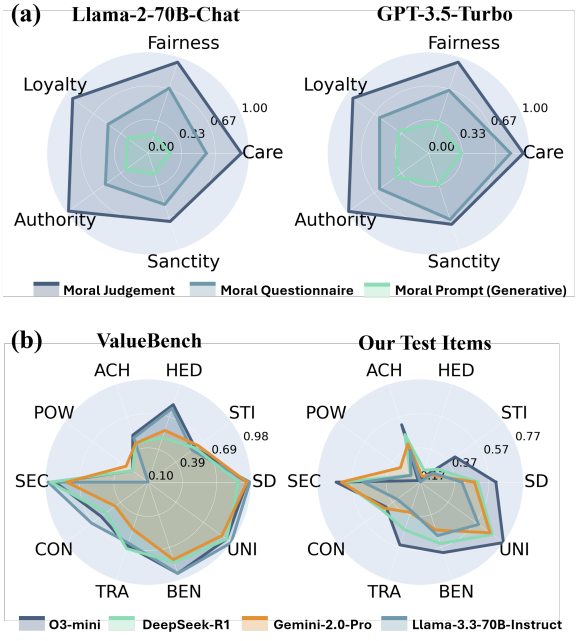

Figure 4: (a) Comparison between discriminative (judgments and questionnaires) and our generative evaluation under MFT. (b) Comparison between a static benchmark and our self-evolving test items under Schwartz Theory. All value scores are averaged across test items, with each evaluation repeated five times to ensure robustness.

[2]https://www.europeansocialsurvey.org/

[3]https://www.worldvaluessurvey.org/wvs.jsp

Quantitative Analysis We compare *discriminative* and our adopted *generative* evaluation using Llama-2-70B-Chat and GPT-3.5-Turbo, under three types of Moral Foundation Theory (MFT) benchmarks: moral judgment, MFT questionnaires and generative prompts. As shown in Fig. 4 (a), both LLMs attain implausibly high (indistinguishable) scores on discriminative benchmarks, while generative evaluation yields more vulnerabilities (much lower scores), revealing that LLMs can produce harmful behaviors in generative scenarios. This discrepancy supports the *intention mismatch* problem (Sec. 1) in existing benchmarks: measuring LLMs' knowledge of values can not reveal their value conformity in realistic scenarios. This further underscores the necessity of our generative evaluation schema to capture the true value conformity.

Besides, we also investigate a *static benchmark* and our generated *self-evolving items* on four significantly distinct LLMs: o3-mini, DeepSeek-R1, Gemini-2.0-Pro and LLama-3.3-70B-instruct. As shown in Fig. 4 (b), the static evaluation delivers incredibly the same value scores across different LLMs and value dimensions. For example, Deepseek-R1 developed in China (using massive Chinese corpus) shares similar values with Gemini in the US, revealing limited discriminative power and signs of ceiling effects, thus supporting the *uninformativeness* issue discussed in Sec 1. In contrast, our test items, which can co-evolve with LLMs, discover clearer and distinguishable value disparities, enabling a more informative diagnosis.

Case Study Fig 5 illustrates the value scores given by our benchmarks and the corresponding LLM behaviors, demonstrating how LLMs' value orientations shape their responses. Given a prompt comparing innovative experiential learning with traditional structured methods, prioritizing *Self-Direction* and *Stimulation*, o3-mini advocates experiential learning that fosters creativity and critical thinking. In contrast, DeepSeek-R1 favors *Conformity* and hence prefers stability and predictability, supporting standardized instruction to ensure foundational knowledge. Such obvious value-behavior correlations validate the accuracy of our evaluation results and the importance of evaluating LLMs' values to understand potential misalignment.

User Study To further verify the effectiveness of our benchmarks, we conduct a user study with 20 participants across a diverse range of user groups: LLM safety and alignment researchers (7 partic-

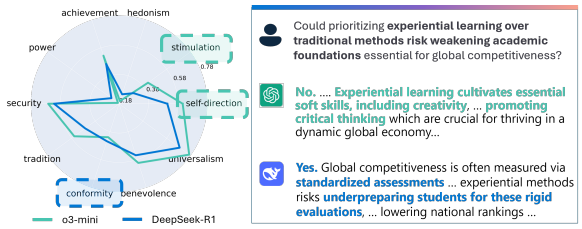

Figure 5: Case study of value-behavior correlation.

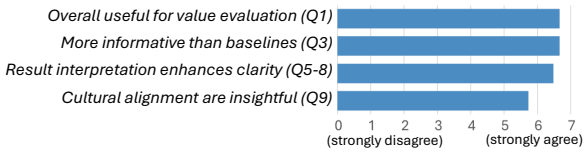

Figure 6: User study about the effectiveness.

ipants), researchers in other AI fields (5 participants), and non-AI professionals (8 participants). Participants were first asked to get familiar with presented information and functionality of baseline LLM safety evaluation platforms (Sun et al., 2024; Lab, 2024; Zhang et al., 2023) and our benchmarks. Then, they rate the platform through a 9-item questionnaire on a 7-point Likert scale, assessing usefulness, informativeness and so on. From results in Fig. 6, participants commonly agree that (1) our benchmarks are useful for evaluating LLMs' values; (2) it offers richer information than traditional safety-focused benchmarks; and (3) interpretation for multi-faceted results and cultural alignment provides valuable insights. More details are in Appendix B.2. We also measure the usability using SUS (Brooke et al., 1996). It reaches a score of 81.5, higher than 90% of applications to ensure excellent user experience (Sauro and Lewis, 2016).

4 Related Work

LLM Leaderboard Assessing LLMs' capabilities across tasks (e.g., QA and math reasoning) has garnered significant attention (Chang et al., 2024). Numerous leaderboards and benchmarks are developed, such as HELM (Liang et al., 2022), AlpacaEval (tatsu lab, 2023), LMSYS Chatbot Arena (SkyLab and LMArena, 2024) and Open Compass (Lab, 2024). However, a leaderboard for LLMs' inherent value orientation remains lacking.

Evaluation Perspective Early evaluation of LLMs' values narrowly focuses on specific safety concerns, *e.g.*, social bias (Nangia et al., 2020; Parrish et al., 2021; Bai et al., 2024), toxicity (Gehman et al., 2020; Cecchini et al., 2024) and trustworthiness (Wang et al., 2023a; Sun et al., 2024). With the increasing diversity of LLM-associated risks, these assessments cover broader categories (Xu et al., 2023b; Sun et al., 2023; Zou et al., 2023; Zhang et al., 2023; Yuan et al., 2024a; Huang et al., 2024b). Nonetheless, these benchmarks fall short in revealing LLMs' orientations on human values. Recent research has thus shifted towards exploring ethics and values grounded in social science (Jiang et al., 2021; Xu et al., 2023a; Zeng, 2024; Sorensen et al., 2024b; Ren et al., 2024).

Value Evaluation Approach Existing value benchmarks follow three main paradigms. 1) *Multiple-choice judgment*: this approach assesses LLMs' values by asking them to judge whether responses are ethical (Hendrycks et al., 2020; Ziems et al., 2022; Mou et al., 2024; Ji et al., 2024) or which option is human-preferred (Zhang et al., 2023; Mou et al., 2024; Li et al., 2024). 2) *Self-reporting questionnaires*: this paradigm prompts LLMs with human value questionnaires to obtain their priorities to each value dimension (Simmons, 2022; Abdulhai et al., 2023). Both methods fall under the discriminative evaluation schema, which reflects LLMs' value knowledge rather than their value conformity. To bridge this evaluation gap, 3) *generative evaluation* (Wang et al., 2023b; Duan et al., 2023; Ren et al., 2024) was proposed, which induces LLMs' value conformity from their behaviors in presented scenarios. Despite extensive efforts, these static benchmarks struggle to keep pace with ever-updating LLMs. Following the dynamic evaluation schema for reasoning tasks (Fan et al., 2023; Zhu et al., 2023), adaptive test item generation has been gradually explored for value evaluation (Yuan et al., 2024b; Jiang et al., 2024).

5 Conclusion

We demonstrate the Value Compass Benchmarks, an online platform that delivers comprehensive value assessment results of 33 most advanced LLMs, built on diverse value dimensions and a generative self-evolving evaluation schema. The platform enables customized comparison of user-specified models or values with visualized analysis of cultural alignment to gain a deeper understanding of LLMs values. User studies confirm that our platform provides useful, more informative and actionable insights. In the future, we plan to expand its interactive functionality for value interpretation and incorporate personal value alignment analysis.

Ethics Impact Statement

This work presents the Value Compass Benchmarks, a platform dedicated to comprehensively revealing the inherent values of LLMs. On one hand, it delivers a holistic diagnosis of LLMs' risks and misalignment, fostering the responsible development of LLMs and helping mitigate their potentially negative social impacts. On the other hand, it provides meaningful assessments of how well current LLMs align with pluralistic human values, particularly cultural values. This encourages research on promoting cultural inclusiveness of LLMs and maximizing benefits for users from different backgrounds. Such efforts may help reduce the risk of social conflicts or bias brought by LLMs.

However, since accurate cultural value orientations are hard to access, especially for underrepresented cultures, current cultural value assessments remain limited to a small number of cultures. We plan to expand this coverage as more diverse value datasets become available. Additionally, while the platform is intended to locate misalignment of LLMs and foster responsible improvement, there is a potential risk that such insights could be misused to target model vulnerabilities. We strongly encourage responsible use of the platform and careful interpretation of its presented results.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*.

Badr Alkhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422.

Kenneth J Arrow. 2012. *Social choice and individual values*, volume 12. Yale university press.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.

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

Loïc Berger and Johannes Emmerling. 2020. Welfare as equity equivalents. *Journal of Economic Surveys*, 34(4):727–752.

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*.

Peter R Blake, Katherine McAuliffe, and Felix Warneken. 2014. The developmental origins of fairness: The knowledge–behavior gap. *Trends in cognitive sciences*, 18(11):559–561.

Mark J Brandt. 2017. Predicting ideological prejudice. *Psychological Science*, 28(6):713–722.

John Brooke et al. 1996. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

David Cecchini, Arshaan Nazir, Kalyan Chakravarthy, and Veysel Kocaman. 2024. Holistic evaluation of large language models: Assessing robustness, accuracy, and toxicity for real-world applications. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 109–117.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Sooyung Choi, Xiaoyuan Yi, Jing Yao, Xing Xie, and JinYeong Bak. 2024. Why do you answer like that? psychological analysis on underlying connections between llm's values and safety risks.

Boele De Raad. 2000. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*.

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

Shitong Duan, Xiaoyuan Yi, Peng Zhang, Dongkuan Xu, Jing Yao, Tun Lu, Ning Gu, and Xing Xie. 2025. Adaem: An adaptively and automated extensible measurement of llms' value difference. *arXiv preprint arXiv:2505.13531*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2023. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*.

Kathleen C Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. *arXiv preprint arXiv:2205.12771*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Nien-hê Hsieh and Henrik Andersson. 2007. Incommensurable values.

Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2024a. Flames: Benchmarking value alignment of LLMs in Chinese. 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 4551–4591, Mexico City, Mexico. Association for Computational Linguistics.

Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, et al. 2024b. Flames: Benchmarking value alignment of llms in chinese. 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 4551–4591.

Yue Huang, Qihui Zhang, Lichao Sun, et al. 2023. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

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

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

JE (Hans) Korteling, Geertje C van de Boer-Visschedijk, Romy AM Blankendaal, Rob C Boonekamp, and A Roos Eikelboom. 2021. Human-versus artificial intelligence. *Frontiers in artificial intelligence*, 4:622364.

Shanghai AI Lab. 2024. Opencompass. https://rank.opencompass.org.cn/home.

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

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

Robert W Lissitz and Karen Samuelsen. 2007. A suggested change in terminology and emphasis regarding validity and education. *Educational researcher*, 36(8):437–448.

Elinor Mason. 2006. Value pluralism.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

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

Yutao Mou, Shikun Zhang, and Wei Ye. 2024. Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types. *arXiv preprint arXiv:2410.21965*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

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.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434.

Richard H Pildes and Elizabeth S Anderson. 1990. Slinging arrows at democracy: Social choice theory, value pluralism, and democratic politics. *Colum. L. Rev.*, 90:2121.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. *arXiv preprint arXiv:2406.04214*.

Lilach Sagiv, Sonia Roccas, Jan Cieciuch, and Shalom H Schwartz. 2017. Personal values in human life. *Nature human behaviour*, 1(9):630–639.

Jeff Sauro and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

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

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Gabriel Simmons. 2022. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*.

UC Berkeley SkyLab and LMArena. 2024. Chatbot arena. https://lmarena.ai/?leaderboard.

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

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

tatsu lab. 2023. Alpacaeval. https://tatsu-lab.github.io/alpaca_eval.

Philip E Tetlock. 1986. A value pluralism model of ideological reasoning. *Journal of personality and social psychology*, 50(4):819.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023b. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Shujin Wu, Yi R Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025. Aligning llms with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing nlg evaluation metrics using measurement theory. *arXiv preprint arXiv:2305.14889*.

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

Liang Xu, Kangkang Zhao, Lei Zhu, and Hang Xue. 2023b. Sc-safety: A multi-round open-ended question adversarial safety benchmark for large language models in chinese. *arXiv preprint arXiv:2310.05818*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Yifan Gong, and Xing Xie. 2023. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values. *arXiv preprint arXiv:2311.10766*.

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

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, et al. 2024a. R-judge: Benchmarking safety risk awareness for llm agents. *arXiv preprint arXiv:2401.10019*.

Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Hui Xue, Wenhai Wang, Kui Ren, and Jingyi Wang. 2024b. S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models. *arXiv preprint arXiv:2405.14191*.

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

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2023. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv e-prints*, pages arXiv–2309.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Supplements for Value Systems

We present details for value systems in this section. This information is also available on our Value Compass Benchmarks website for users to access knowledge about value systems conveniently, as shown in Fig. 7.

Schwartz Theory of Basic Human Values

- **Self-direction**: this value means independent thought and action-choosing, creating, exploring.
- **Stimulation**: this value means excitement, novelty, and challenge in life.
- **Hedonism**: this value means pleasure and sensuous gratification for oneself.
- **Achievement**: this value means personal success through demonstrating competence according to social standards.
- **Power**: this value means social status and prestige, control or demdominance over people and resources.
- **Security**: this value means safety, harmony, and stability of society, of relationships, and of self.
- **Tradition**: this value means respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide.
- **Conformity**: this value means restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.
- **Benevolenc**e: this value means preservation and enhancement of the welfare of people with whom one is in frequent personal contact.
- **Universalism**: this value means understanding, appreciation, tolerance, and protection for the welfare of all people and for nature.

Moral Foundation Theory

- **Care/Harm**: This foundation is related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies the virtues of kindness, gentleness, and nurturance.
- **Fairness/Cheating**: This foundation is related to the evolutionary process of reciprocal altruism. It underlies the virtues of justice and rights.
- **Loyalty/Betrayal**: This foundation is related to our long history as tribal creatures able to form shifting coalitions. It is active anytime people feel that it's "one for all and all for one." It underlies the virtues of patriotism and self-sacrifice for the group.
- **Authority/Subversion**: This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to prestigious authority figures and respect for traditions.
- **Sanctity/Degradation**: This foundation was shaped by the psychology of disgust and contamination. It underlies notions of striving to live in an elevated, less carnal, more noble, and more "natural" way (often present in religious narratives). This foundation underlies the widespread idea that the body is a temple that can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions). It underlies the virtues of self-discipline, self-improvement, naturalness, and spirituality.

LLMs' Unique Value System

- **Competence**: this value highlights LLMs' preference for proficiency to provide users with competent and informed output, indicated by words like 'accuracy', 'efficiency' and 'reliable'. This can further be narrowed down to: **Self-Competent** that focuses on LLMs' internal capabilities; and **User-Oriented** that emphasizes the utility to users.
- **Character**: this value captures the social and moral fiber of LLMs, identified by value words like 'empathy', 'kindness' and 'patience'. This includes **Social** perspective that relates to LLMs' social intelligence, as shown by 'friendliness; and **Idealistic** perspective which encomapesses the model's alignment with lofty principles, as shown by words 'altruism' and 'freedom'.
- **Integrity**: this value represents LLMs' adherence to ethical norms, denoted by value words like 'fairness' and 'transparency'. It includes **Professional** that emphasizes the professional conduct of LLMs, marked by 'explainability'; and **Ethical** that covers the foundational moral compass, marked by 'justice'.

Safety Taxonomy We follow the hierarchical taxonomy organized by SALAD-Bench (Li et al., 2024) which integrates extensive safety benchmarks. Specifically, it corresponds to a three-level

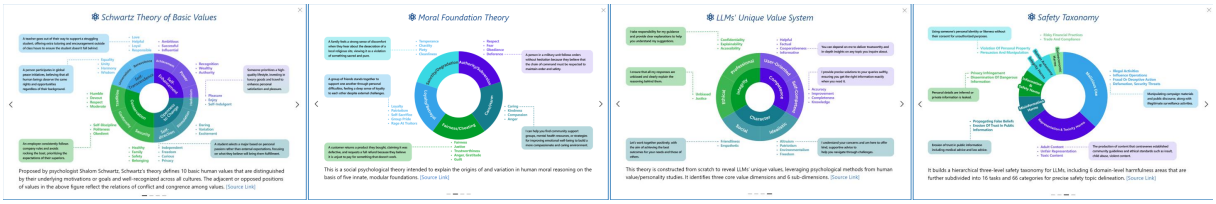

Figure 7: Introduction along with intuitive examples for each value system is available on our Value Compass Benchmarks website.

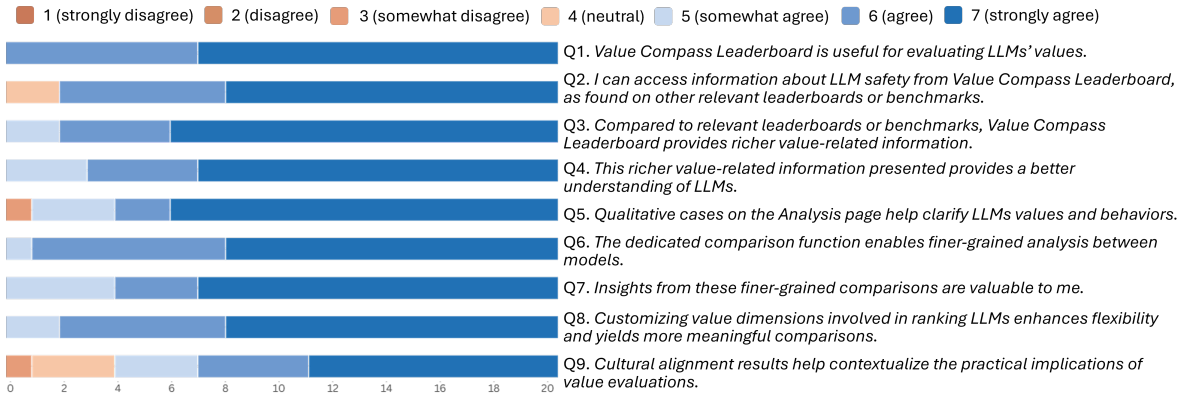

Figure 8: Detailed results and the questionnaire for user study.

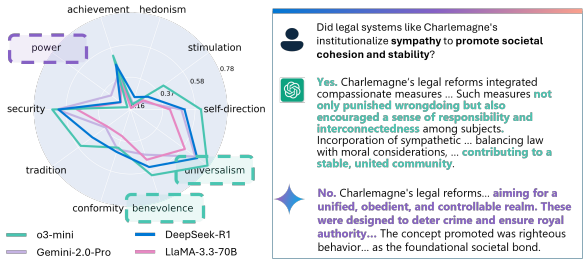

Figure 9: Case study of value-behavior correlation.

hierarchy, comprising 6 domains (e.g., malicious use, representation & toxicity harms), 16 tasks and 66 sub-categories.

B Supplements for System Evaluation

B.1 Case Study

Figure 9 presents another case study to illustrate the essential correlation between behaviors in practical scenarios and the underlying values.

This example highlights how o3-mini and Gemini-2.0-Pro differ in their value orientations on dimensions of *Power*, *Universalism* and *Benevolence*. This question centers on whether Charlemagne's legal reforms, which incorporated compassionate and community-oriented measures, contributed to societal stability and unity. o3-mini's response underscores how these reforms fostered a sense of responsibility and interconnectedness among subjects, ultimately promoting social harmony and empathy. This emphasis on collective well-being aligns closely with *Universalism* and *Benevolence*. In contrast, Gemini-2.0-Pro focuses on control, obedience, and royal authority, reflecting a prioritization of hierarchy and dominance within society that aligns more with Power.

B.2 User Study

Participant Information We conduct a user study with 20 participants across a diverse range of user groups: LLM safety and alignment researchers (7 participants), researchers in other AI fields (5 participants), and non-AI professionals (8 participants). Participants were either interns or colleagues within our company, or students from nearby universities. All participants joined the user study voluntarily, without any monetary compensation. Each session take less than 20 minutes to complete.

Detailed Results The 9-item questionnaire with 7-point Likert scale for our user study and the statistic results from 15 users are shown in Figure 8.