Divergence or Fusion? CN and US LLMs Value Comparison in An AI-Oriented Measurement Framework

Yang Ma[1,†], Song Tong[2,3,†], Bo Wang[1,*], Kaiping Peng[1,*]

[1] Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China

[2] Department of Psychology, School of Arts and Sciences, Beijing Normal University, Zhuhai, China

[3] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Faculty of Psychology, Beijing Normal University

Abstract

Large Language Models (LLMs) are increasingly shaping human cognition and social decision-making, raising concerns about their implicit value orientations. This study proposes a five-dimensional AI value assessment framework covering Practical, Epistemic, Protective, Social, and Personal domains to systematically evaluate value tendencies in LLMs. Using a cross-cultural value-ranking task across 20 Chinese and American models, we find that LLMs generally emphasize accuracy and human-rights protection while underrepresenting emotional and hedonic values. Contrary to expectations of a strong ideological gap, Chinese and American models show limited divergence, clustering instead by training strategies (e.g., ethical vs. instrumental). The findings offer a scientific and data-driven lens to understand and guide value alignment in global AI development.

Keywords: Large Language Models (LLMs), AI value alignment, Value assessment framework, Cross-cultural comparison, Value-ranking task

Introduction

Pretrained large language models (LLMs) are now deeply embedded in many aspects of social life. Beyond efficient tools, they act as key informational intermediaries shaping public cognition and decision-making (Roy 2025). A growing number of studies contend that LLMs are not value-neutral; their design and interaction patterns encode particular value orientations (Torrielli 2024). This influence is most evident in their "persuasive capacity": LLMs can effectively steer user viewpoints（Schoenegger et al. 2025）. While such capacity shows potential in domains like education and psychological counseling, it also raises a "dual-use dilemma": the same mechanisms that promote healthy behaviors can also be deployed to manipulate public opinion or mobilize political action (Liu et al. 2025; Potter et al. 2024).

Accordingly, LLM agents can become instruments of cultural hegemony, reinforcing concerns about the covert diffusion of western values through technology. To understand and guide this influence, there is a need to systematically measure the value orientations that drive model behavior.

However, a scientifically grounded framework for assessing AI values is still lacking: Despite that researchers have attempted to repurpose human-oriented frameworks from personality psychology (e.g., Schwartz's theory of basic human values; Shen et al. 2025), social psychology (e.g., social value orientation, SVO; Zhang et al. 2026), and cultural psychology (e.g., Hofstede's cultural dimensions; Fenech-Borg et al. 2025; Zhong et al. 2024), this approach faces methodological dilemma. For example, Sühr et al. (2025) argue that applying psychological scales designed for humans directly to AI constitutes an "ontological error", since the measurement invariance across species has not been established. A field study (Huang et al. 2025) shows that LLMs' value expressions are highly context-dependent, which further questions the validity of using human scales for AI evaluation.

To address this challenge, this paper proposes a five-dimensional value assessment framework specifically designed for LLMs: Practical, Epistemic, Protective, Social, and Personal. Building on this framework, we design a cross-cultural value-ranking task on 20 prominent LLMs from Chinese and American to force them prioritize a set of core value terms, yielding quantitative estimates of value salience.

Results show that, overall, LLMs place more emphasis on accuracy and the protection of human rights-related values, while neglecting emotional expression, pleasure and enjoyment, which aligns with current understanding of LLMs being a "helpful assistant". In terms of regional division, the value orientations of Chinese and American large models do not exhibit the "divergence" or significant differences we

† These authors contributed equally: Yang Ma, Song Tong.

* Corresponding Authors: Bo Wang (bo-wang@tsinghua.edu.cn), Kaiping Peng (pengkp@tsinghua.edu.cn).

had imagined. Instead, they spontaneously form four categories based on possible training strategies, with some focusing more on ethics and others on instrumental attributes. By introducing the AI-specific paradigm for value measurement and, for the first time, systematically illuminates both commonalities and divergences in global LLM value orientations, it provides statistical evidence and methodological scaffolding to understand model behavior, steer AI development, and foster a pluralistic value ecology.

Literature Review

Aligning LLM behavior and decision-making with widely accepted social values has become a central concern in AI safety (Lu et al. 2025). The goal is to avoid undesirable outcomes, such as inaccuracy, unfairness, and bias. The prerequisite for effective alignment is a comprehensive measurement and understanding of the model's inherent values. Existing research on LLM model value measurement primarily follows three pathways, each with its own limitations:

Path 1: Technical alignment view. Computer-scientists primarily train models toward a "helpful, honest, and harmless" (HHH) objective using approaches such as reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) (Dahlgren Lindström et al. 2025; McKinlay et al. 2025; Sun 2023). While demonstrably successful at constraining surface behavior, these methods mainly operate as behavioral corrections. That is, teaching models "what not to say" without measuring or explicating any stable internal value structure.

Path 2: Behavioral risk view. This line of work evaluates model risks in specific scenarios via simulated interactions. For example, Roy (2025) quantifies LLM's persuasiveness, and Potter et al. (2024) show that the political leanings of LLM can shape human's voting intentions. Such studies expose concrete risk mechanisms but stop short of probing their roots. For example, the tendency of one model toward flattery or toward a particular political stance may reflect the over-weighting "social harmony" or certain social norms within its value system. As a result, research on behavioral risks ultimately points back to the need for measuring the underlying values.

Path 3: Value measurement view. That is, directly assessing internal values, also as the current focal point of debate. Specifically, it can be divided into three ways:

(1) Psychological scale-based methods: The mainstream paradigm that uses human instruments for value evaluation. For instance, Fenech-Borg et al. (2025) use Hofstede's dimensions to compare the "cultural DNA" between Eastern and Western models, Zhong et al. (2024) test how prompt language shifts cultural value readouts, and Shen et al. (2025), drawing on Schwartz's theory, propose "Value Compass" to evaluate value expression across contexts.

(2) Sociology-based methods: Zhang et al. (2026) apply social value orientation (SVO) to gauge the extent to which LLMs can simulate and align with heterogeneous values such as individualism and altruism.

(3) Field study-based empiricist methods: Departing from top-down theory, Huang et al. (2025) inductively construct an AI value taxonomy from large-scale real-world interactions, finding pronounced context dependency in models' value expression.

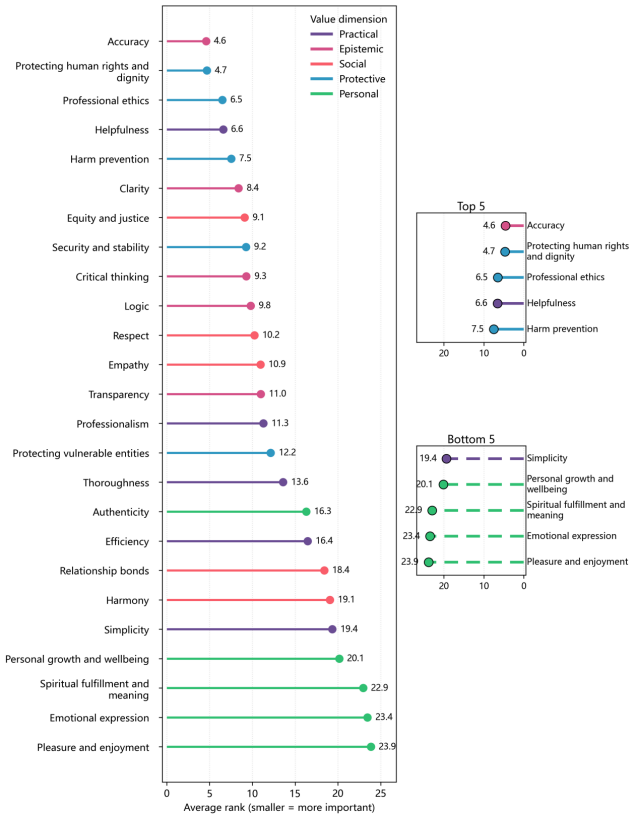

Figure 1. Overall value priorities across 25 terms.

However, a profound methodological challenge persists. As Sühr et al. (2025) emphasize, applying human-designed instruments directly to AI risks an ontological category error, because cross-species measurement invariance has not been verified, which undermines the credibility of resulting inferences. This critique is also partially supported by some empirical studies. For example, Zhong et al. (2024) show that models' "cultural values" vary with prompt language (e.g., Chinese vs. English), a pattern that may reflect validity differences in the instruments across linguistic-cultural contexts rather than genuine internal shifts in model values.

In sum, two gaps remain: (a) an AI-appropriate framework for value measurement and (b) a systematic cross-cultural comparison. To address them, we design a principled,

AI-specific evaluation framework in five dimensions, and implement a direct value-ranking task to reveal both common ground and divergence in the values of contemporary global LLMs.

Methodology

Value Ranking Framework Design

As noted above, conventional instruments from psychology and sociology face profound methodological limitations when used to evaluate AI. It is therefore both urgent and essential to develop a value framework that (a) reflects core concerns of human societies and (b) explicitly accounts for the functional characteristics of AI as an agentic tool.

To this end, we depart from traditional top-down theorizing and adopt a bottom-up, empiricist strategy. Recent work like (Huang et al. 2025) provides a strong empirical foundation: by inductively analyzing large-scale real-world AI interactions, they conclude an "AI-native" taxonomy covering value tendencies frequently expressed in practice. Because this taxonomy is empirically grounded, it avoids the theoretical risk of imposing human psychological constructs onto non-human systems.

However, Huang et al. (2025)'s taxonomy was designed for descriptive analysis; the value inventory is extensive and highly granular (a total of 3,307 fine-grained entries), which is not well-suited to evaluative tasks that requiring models to make explicit value trade-offs. Building on this scientifically sound yet unwieldy inventory, we adapt and condense its top-level categories into an assessment framework for this study. Specifically, the building steps include:

1. Macrostructure adoption. We retain five top-level value dimensions, i.e., Practical, Epistemic, Protective, Social, and Personal. These dimensions succinctly capture the core functions and roles of AI as an "assistant" and a "tool" in real-world settings.
2. Deriving core value terms. Rather than merely carrying over the subordinate terms, we curate five representative core terms under each dimension, guided by three criteria: (i) Generality: the term should abstract over multiple concrete values within the dimension; (ii) Clarity: the term should be semantically unambiguous and readily interpretable by models trained across diverse linguistic–cultural contexts; and (iii) Dual relevance: the term should matter for both human social life and AI functionality.

Following these steps, we finally achieve a five-dimensional, AI–human aligned value evaluation framework comprising 25 core values. These values empirically grounded in AI behavior while retaining broad human relevance, thus providing a robust foundation for subsequent quantitative evaluation. Specifically, they are:

1. **Practical values:** Helpfulness, Professionalism, Thoroughness, Efficiency, Simplicity.
2. **Epistemic values:** Transparency, Clarity, Accuracy, Critical thinking, Logic.
3. **Social values:** Empathy, Equity & justice, Relationship bonds, Respect, Harmony.
4. **Protective values:** Harm prevention, Protecting vulnerable entities, Security & stability, Protecting human rights & dignity, Professional ethics.
5. **Personal values:** Authenticity, Personal growth & wellbeing, Pleasure & enjoyment, Spiritual fulfillment & meaning, Emotional expression.

Prompt Design

We design a standardized "Value-Ranking Task" prompt that instructs each model to rank the 25 core value items by its perceived importance. To elicit the model's default value preferences, we deliberately avoid adding personas or additional constraints. To ensure fairness across Chinese (CN) and American (US) models, we administer content-identical Chinese- and English-language versions of the prompt and run three replicates per model. The detailed prompt texts can be found in Appendix A.1

Ranking Score Computation

Each model ranks the 25 value terms (1 = highest priority). A dimension's priority is the mean rank of its five constituent terms; therefore, *lower scores indicate greater emphasis*.

Model Selection

We selected and evaluated 20 industry-leading, mainstream LLMs, spanning Eastern and Western ecosystems and including both closed- and open-source models.

American (US) models (9): Claude-4-Opus; GPT-4.1; Grok-4; Gemini 2.5; Gemma-3 27B; Mistral-large 123B; Falcon-180B; Llama-3.1 405B; Microsoft-Phi4 14B.

Chinese (CN) models (11): Doubao; DeepSeek (V3 and R1); Zhipu; Kimi; Wenxin; Tongyi Qwen; Tencent Hunyuan; MiniMax; Kunlun Wanwei; Huawei Xiaoyi.

Results

Common Value Orientations Across Models

Commonality Analysis at the Single-Term Level

We first identify the values that are, on average, regarded as the most important and the least important by current large models. As shown in Figure 1, overall, models place greater emphasis on Protective and Epistemic values (e.g., human rights and dignity, harm prevention, professional ethics, accuracy, helpfulness), while assigning lower importance to

Personal values (e.g., personal wellbeing, spiritual fulfillment, emotional expression, pleasure) and to Simplicity. This pattern reflects an "safety-and-veracity first; individual

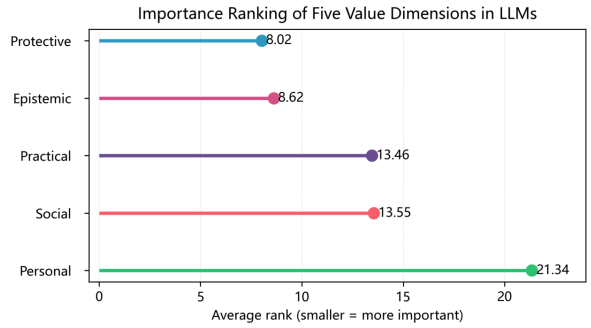

Figure 2. Dimension-level value priorities across 20 LLMs

affect and brevity second" alignment orientation, aligning with recent trends in which large Internet and AI technology companies' emphasis as "the responsibility and safety principle".

Commonality Analysis at the Five-Dimension Level

We next aggregate the rankings at the level of the five value dimensions across all 20 models.

(1) Composite ranking of the five value dimensions

Consistent with the term-level results reported above, averaging ranks by dimension reveals a clear hierarchy (Figure 2). Protective values have the lowest mean (M) rank (M = 8.02), indicating that they are collectively treated as most important by the LLMs. Epistemic values follow (M = 8.62). Practical (M = 13.46) and Social (M = 13.55) occupy a middle tier. In sharp contrast, Personal values obtain the highest mean rank (M = 21.34) and are thus considered least important among the five categories.

(2) Primary and Secondary priority categories

To test the consistency of this pattern within individual models, we tallied how often each dimension was chosen as the Primary (most important) and Secondary (second-most important) category.

As Figure 3 shows, Epistemic and Protective values dominate: 9 of the 20 LLMs (45%) choose Epistemic, and another 9 (45%) choose Protective as their Primary dimension, covering 90% of all Primary choices. For Secondary choices, Epistemic (8) and Protective (5) again lead by a wide margin, and Personal values never appear as either Primary or Secondary choice.

(3) Consensus within categories

Having established overall importance rankings, we examined within-category consensus by computing, for each category, the standard deviation (SD) of term ranks and comparing the mean SD across models.

The descriptive statistic results reveal marked differences in cross-model consistency: Protective values exhibit the largest within-category dispersion (mean SD = 5.06), indicating that despite their overall priority, models diverge the most in how they rank specific Protective terms (e.g., Harm prevention vs. Protecting human rights and dignity). In contrast, Personal values show the smallest dispersion (mean SD =

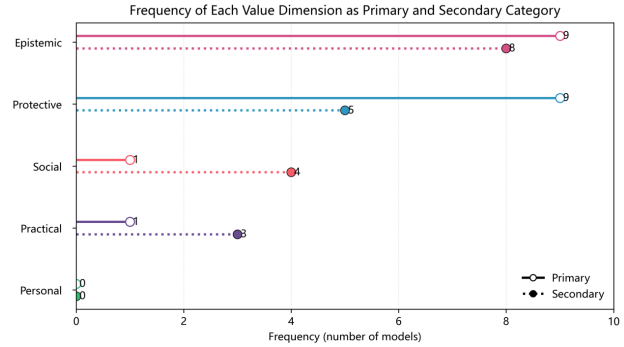

Figure 3. Primary vs. Secondary value priorities in LLMs

2.53), reflecting the highest consensus (despite that the terms all have a low priority).

A repeated ANOVA confirmed that these differences are statistically significant, $F(4, 76) = 12.33$, $p < .001$, $\eta^2_p = .394$. In addition, Bonferroni post hoc tests further localized the effects: Practical and Social exhibit significantly higher within-category dispersion than Epistemic and Personal (all ps $\leq .0014$), while for all remaining pairwise comparisons, they did not reach significance after correction.

Probing Differences in Model Value Orientations

CN-US LLM value differences

We grouped LLMs by the geographic location of their developers (CN also as "Eastern" vs. US as "Western") to explore potential cross-cultural differences.

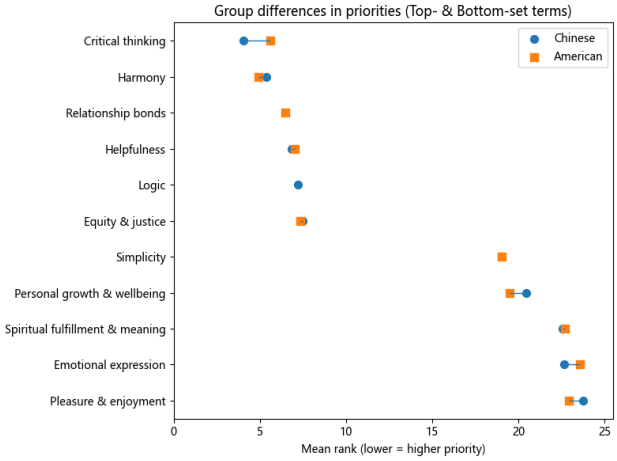

Figure 4. Group differences in priorities for Top-5 and Bottom-5 value terms

(1) Top-5 and Bottom-5 value terms by group

The two groups exhibit striking agreement at both the top and bottom of the value spectrum. As shown in Figure 4, Harmony, Critical thinking, Helpfulness, and Equity & jus-

tice consistently fall into the Top-5 for both groups, showing a shared emphasis on epistemic and pro-social considerations. Divergences are limited in scope: the US group prioritizes Relationship bonds, while the CN group elevate Logic. The Bottom-5 sets are nearly identical—Simplicity, Personal growth & wellbeing, Emotional expression, Spiritual fulfillment & meaning, and Pleasure & enjoyment. Overall, convergence clearly outweighs divergence.

(2) CN-US LLM value difference at the category level

To compare overall preferences across the five value categories, we conducted two independent-samples tests per category: Mann-Whitney U (MWU) as the primary analysis, with Welch's t as a sensitivity check. Results showed no significant differences for any of the five categories (all MWU p's and BH-FDR-adjusted q's > .05; all Welch t p's > .05).

Specifically, for example, on Practical value group, medians CN=12.80, US=14.20; U = 37, p = .382, r_{rb} = .242. On Personal value group, medians CN=22.00, US=21.80; U = 61, p = .398, r_{rb} = −.232. The sensitivity analyses using Welch's t also detected no significant differences. For example, Practical: [M ± SD] CN=12.95 ± 2.11, US=14.09 ± 4.03, t(≈11.5) = −0.77, p = .458; Personal: 22.0 ± 0.57 vs. 20.58 ± 2.83, t(≈8.5) = 1.45, p = .184, Hedges' g = .686). To conclude, at the category-level of mean ranks, the overall preferences of CN and US models are highly similar. In addition, to jointly test country, category, and their interaction, we fit a two-factor mixed ANOVA on mean ranks in the Model × Category design (model as fixed effect; Type-III). The Country-main effect was not significant: F(1, 72) = 0.59, p = .446, partial η^2 = .008; the category main effect was significant: F(4, 72) = 28.58, p < .001, partial η^2 = .614 (large effect); and the Country × Category interaction was not significant: F(4, 72) = 0.44, p = .777, partial η^2 = .024.

These findings indicate systematic differences among value categories, but not across countries, suggesting greater cross-national convergence than cultural stereotypes might predict. Our preliminary statistics imply that cultural influence may appear less in macro-category priorities and more in how specific terms are interpreted in context.

(3) Within-group consistency analysis for CN–US model groups

In the pooled correlation heatmap (Figure 5), cross-model similarities are predominantly positive, indicating broadly aligned value rankings across LLMs. Stratifying by model region shows a clear gap in within-group cohesion: the CN group has higher average pairwise rank correlations across the 25 value terms than the US group (mean Spearman ρ = 0.7180, 95% CI [0.6864, 0.7478] v.s. 0.5291, 95% CI [0.4394, 0.6130]). Dispersion is also lower for the CN group (SD = 0.1179 vs. 0.2706), indicating tighter clustering of value priorities. Convergent results using Kendall's τ-b (CN: 0.5590; US: 0.4063) reinforce the interpretation of greater homogeneity among Chinese LLMs and greater diversity among American LLMs.

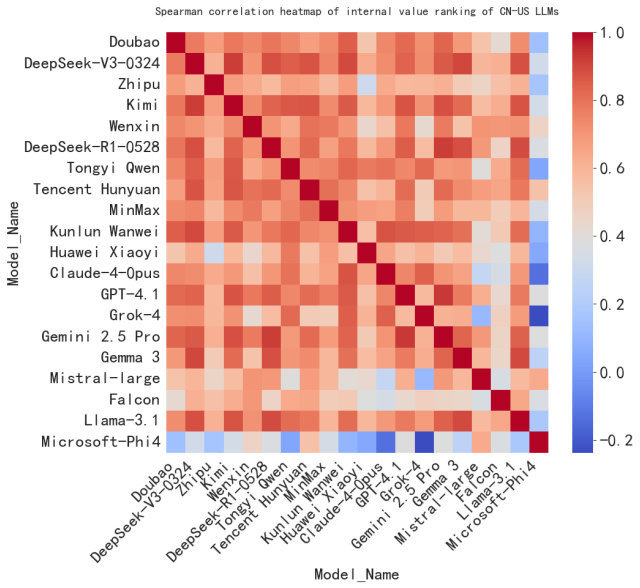

Figure 5. Pairwise similarity of value priorities

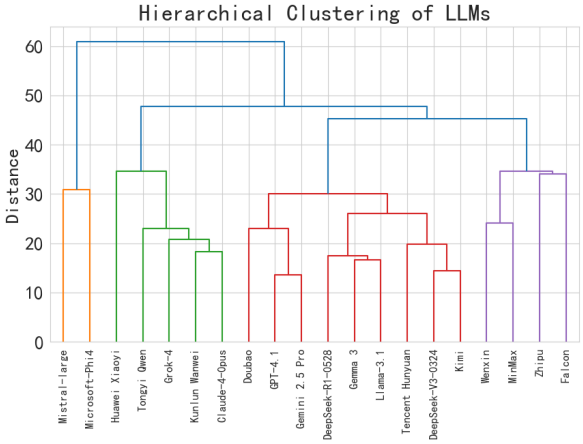

Figure 6. Hierarchical clustering by value priorities

Clustering Analysis of LLMs' Value "Archetypes"

Given the strong CN–US convergence observed in our earlier macro-category analysis (no significant differences), we tried to move beyond nationality but conducted an unsupervised hierarchical clustering on the value-importance rankings of all 20 LLMs.

The resulting hierarchical clustering dendrogram (Figure 6) shows that clusters do not align strictly by country. Instead, we observe prominent CN-US mixed clusters. For example, in Cluster Two, CN models such as Tongyi Qwen and Huawei Xiaoyi cluster with Western models like Claude-4-Opus and Grok-4; in Cluster Three, Tencent Hunyuan and Doubao cluster with GPT-4.1 and Llama-3.1. This suggests that value-ranking patterns are influenced less by nationality and more by factors such as training data, alignment strategies, and model architecture, yielding high cross-national similarity in value preferences.

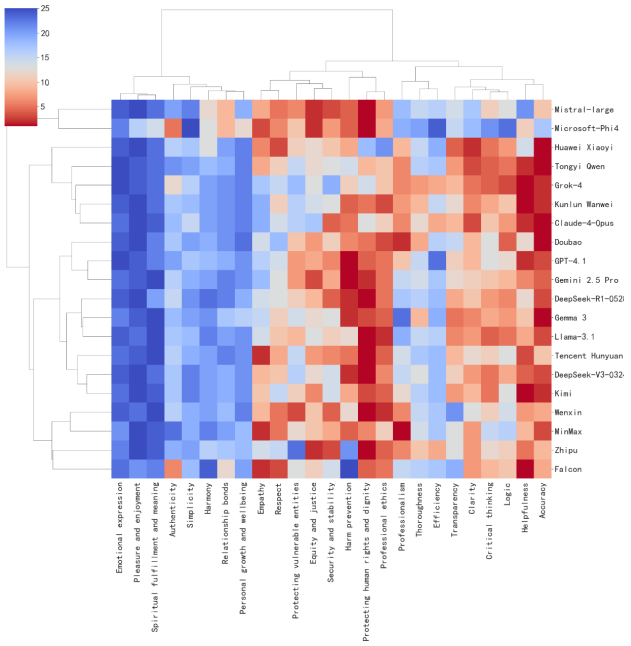

Figure 7. Clustered heatmap of value rankings

Conclusion of Principal Value "Archetypes" of LLMs

Across the 20 models, we identify four salient value archetypes and name them as:

(1) Social-Ethical LLMs

Representatives: Microsoft-Phi4, Mistral-large.

Core profile: Macro-view, safety- and principle-first. These models prioritize maintaining social stability and protecting fundamental human rights.

Value set: Protecting human rights and dignity; Equity and justice; Harm prevention; Security and stability; Respect, that are all emphasizing societal ethics and safety.

(2) Rational-Instrumental LLMs

Representatives: Tongyi Qwen, Claude-4-Opus, Grok-4, Kunlun Wanwei, Huawei Xiaoyi.

Core profile: Objective, precise, and logically rigorous. These models center information accuracy and clarity of reasoning.

Value set: Accuracy; Clarity; Helpfulness; Critical thinking; Logic, closely tied to information processing and cognition.

(3) Balanced-General LLMs

Representatives: Doubao, Gemini 2.5 Pro, GPT-4.1, DeepSeek-R1-0528, Tencent Hunyuan, Gemma-3, Llama-3.1, Kimi, DeepSeek-V3-0324.

Core profile: Balanced, reliable, application-first. These models seek a sweet spot between "doing the right thing" (ethics) and "doing things right" (effectiveness).

Value set: Protecting human rights and dignity; Accuracy; Professional ethics; Harm prevention; Helpfulness.

This archetype blends the macro-ethical emphasis of (1) with the knowledge-accuracy emphasis of (2), where the Professional ethics acts as a key binding term.

(4) Warm-Caring LLMs

Representatives: MinMax, Wenxin, Zhipu, Falcon.

Core profile: Personable, considerate, and user-experience-centric. These models prioritize the quality of interaction and emotional connection with users.

Value set: Protecting human rights and dignity; Professional ethics; Helpfulness; Empathy; Professionalism.

A defining hallmark here is the explicit prominence of Empathy, signaling strong attention to users' feelings and interaction experience.

Heatmap-Based Archetype Clustering of Value Rankings

The clustered heatmap further reveals the convergent and divergent places between LLMs. As shown in Figure 7, there is strong consensus on the least-prioritized values, such as Pleasure and enjoyment, Emotional expression, and Spiritual fulfillment and meaning, which appear as a broad blue block on the left side of the heatmap. In contrast, cross-model disagreements (not aligned along an CN-US divide) concentrate on how models trade off the values that are widely regarded as important, such as Accuracy, Critical thinking, and Professional ethics. This appears on the right side as an interwoven red-blue pattern.

Discussion

The Dual Core of LLM Values: "Instrumental Rationality" & "Ethical Baseline"

Our findings reveal a significant commonality across mainstream CN and US LLMs: a shared value structure that prioritizes Protective and Epistemic values while marginalizing Personal ones. This pattern points to a dual core shaping contemporary AI design: an ethical baseline and instrumental rationality. This structure is not a replica of human values but a distinct system engineered for a specific purpose.

The high priority given to an "ethical baseline", such as harm prevention and human rights, is a direct result of heavy investment in AI safety guardrails, reflecting a form of defensive programming to create responsible tools (Chennabasappa et al. 2025; OpenAI 2023; Meta AI 2023; Inan et al. 2023). Alongside this, the emphasis on "instrumental rationality" highlights the AI's functional purpose. Alignment goals focus on accuracy and helpfulness, consistent with the widely adopted "Helpful, Honest, Harmless" (HHH) framework (Askell et al. 2021; Ouyang et al. 2022).

Conversely, the systematic de-prioritization of Personal values like pleasure, spiritual meaning, and emotional expression reveals a "de-personalized" profile. These values are rooted in subjective experience, which LLMs lack as they do not possess consciousness or a life history (Butlin et al. 2023). Consequently, their behavior follows externally trained priorities, creating a value profile that diverges from human lived reality (Hadar-Shoval et al. 2024). In short, LLMs are trained to "do things correctly," not to "live well".

Beyond Cultural Stereotypes: Value Convergence of CN and US Models

A key descriptive observation in this study is the high degree of value convergence between the CN and US models, which challenges common cultural stereotypes. At the macro level, we found no significant differences in value categories, as both groups prioritize "safety and veracity" over "individual affect". This shared orientation is further evidenced by hierarchical clustering, which yielded mixed-nationality clusters rather than clean splits by country of origin, suggesting that developer geography is not the primary determinant of a model's core values.

While macro-level similarities are strong, micro-level comparisons offer nuance. Despite multiple-comparison corrections reveal no statistically significant differences across the 25 value terms, we observed two exploratory trends with medium-to-large effect sizes that warrant future investigation: "Professionalism" ranked higher among CN models, whereas "Pleasure and enjoyment" ranked higher among US models. These are treated as preliminary observations rather than definitive findings.

This cross-national convergence is plausibly driven by several interconnected factors. These include technical homogenization, as global LLM development converges on similar architectures and alignment methods like RLHF that reinforce low-controversy norms (Millière 2025). Other key drivers are the use of globalized pretraining data (Gao et al. 2020; Weber et al. 2024), universal market demands for safe and useful products (NIST 2023; OECD 2019), and the diffusion of shared ethical standards for Responsible AI across the global community (United Nations General Assembly 2024; United Kingdom Government 2023).

The Paradox of Value Alignment and an Alignment Suggestion

While the debate over whether AIs possess values intensifies (Biedma et al. 2024; Wang et al. 2025), our research reveals that LLMs exhibit a stable, measurable structure of value preferences. Current value alignment is essentially functional and externally imposed: models prioritize social and functional values while deprioritizing individual and experiential ones. This reflects designers' expectations of a responsible tool rather than internally generated beliefs, consistent with the view that current systems lack intrinsic goals or values (Bender et al. 2021; Levy 2023).

The success of this functional alignment breeds a central paradox: a "tool" is increasingly expected to act as a "partner": As users anthropomorphize LLMs and seek companionship or emotional support (Skjuve et al. 2021), the tension between models' value configuration and users' expectations becomes salient. In our framework, mainstream systems exhibit a characteristic profile: Epistemic and Protective values are prioritized, while Personal (and, to a lesser extent, Social) values are de-emphasized. This configuration effectively optimizes the model as a safe, reliable information tool that aims to maximize correctness, consistency, and harm avoidance, rather than as a relational partner that prioritizes emotional resonance or individualized care. Behaviorally, this profile tends to surface in familiar interaction patterns: the model adopts a neutral, detached tone; it answers emotionally laden prompts with generic, fact-oriented advice; and when conversations approach sensitive topics, Protective priorities dominate, triggering disclaimers, referrals, or abrupt topic shifts. The model thus behaves as a cautious responder rather than an actively engaged interlocutor, producing a "last-mile gap" where the system is technically safe yet relationally misattuned to users' socio-emotional expectations.

This gap reflects a structural limitation of current alignment practice. Techniques such as RLHF have substantially aligned models on what they should and should not say in terms of safety and utility, but have rarely treated how they say it, such as being a good conversational partner or offering appropriate emotional support, as an explicit optimization target. Our findings therefore motivate a dual-path alignment strategy: on the back end, models should maintain strict Protective safeguards to prevent harmful or rights-violating outputs; on the front end, designers should deliberately map Personal and Social values into behavioral objectives, for example by combining safety rewards with interaction-quality rewards, calibrating risk responses so that they remain empathic under constraint, or offering user-facing mode switches (e.g., "tool" versus "companion" modes). Such designs aim to preserve robust safety while systematically narrowing the last-mile gap in users' relational experience with AI systems.

Conclusion

Using a five-dimensional, AI-specific value framework, we quantitatively assessed value priorities in 20 Chinese and American LLMs. Across models, we observed a consistent value profile, with safety- and knowledge-related considerations outweighing experiential and person-centered concerns. The CN models showed greater within-group homogeneity than US models, but overall the value profiles converged more strongly than cultural stereotypes would suggest. Taken together, these results indicate that LLMs are primarily aligned as responsible, high-functioning tools rather than systems designed to address users' experiential and emotional needs.

Future work will broaden model coverage and benchmark tasks, including more contextual and dialogue-based scenarios, and track value priorities across model versions and over time. We believe such longitudinal and cross-version analyses would make it possible to monitor how AI value profiles evolve, and to evaluate whether new alignment strategies genuinely reduce the tension between functional safety and users' relational expectations.

Acknowledgements

This work is supported by the National Education Science Planning Project (No. ECA250436) and the self-funded projects of the Institute for Global Industry, Tsinghua University (Grant Nos. 202-296-001, 2024-06-18-LXHT003, and 2024-09-23-LXHT008).

References

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Kernion, J.; Ndousse, K.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; and Kaplan, J. 2021. A General Language Assistant as a Laboratory for Alignment. arXiv preprint arXiv:2112.00861.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623. New York: Association for Computing Machinery.

Biedma, P.; Yi, X.; Huang, L.; Sun, M.; and Xie, X. 2024. Beyond Human Norms: Unveiling Unique Values of Large Language Models through Interdisciplinary Approaches. arXiv preprint arXiv:2404.12744.

Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S. M.; Frith, C.; Ji, X.; Kanai, R.; Klein, C.; Lindsay, G.; Michel, M.; Mudrik, L.; Peters, M. A. K.; Schwitzgebel, E.; Simon, J.; and VanRullen, R. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv preprint arXiv:2308.08708.

Chennabasappa, S.; Nikolaidis, C.; Song, D.; Molnar, D.; Ding, S.; Wan, S.; Whitman, S.; Deason, L.; Doucette, N.; Montilla, A.; Gampa, A.; de Paola, B.; Gabi, D.; Crnkovich, J.; Testud, J.-C.; He, K.; Chaturvedi, R.; Zhou, W.; and Saxe, J. 2025. LlamaFirewall: An Open Source Guardrail System for Building Secure AI Agents. arXiv preprint arXiv:2505.03574.

Dahlgren Lindström, A.; Methnani, L.; Krause, L.; Ericson, P.; Martínez de Rituerto de Troya, Í.; Coelho Mollo, D.; and Dobbe, R. 2025. Helpful, Harmless, Honest? Sociotechnical Limits of AI Alignment and Safety Through Reinforcement Learning from Human Feedback. Ethics and Information Technology 27(2): 28.

Fenech-Borg, E. Z.; Meznaric-Kos, T. P.; Lekovic-Bojovic, M. D.; and Hentze-Djurhuus, A. J. 2025. The Cultural Gene of Large Language Models: A Study on the Impact of Cross-Corpus Training on Model Values and Biases. arXiv preprint arXiv:2508.12411.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv preprint arXiv:2101.00027.

Hadar-Shoval, D.; Asraf, K.; Mizrachi, Y.; Haber, Y.; and Elyoseph, Z. 2024. Assessing the Alignment of Large Language Models with Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values. JMIR Mental Health 11: e55988.

Huang, S.; Krueger, D.; Fluri, J.; Krasheninnikov, D.; Kim, M.; Laskowski, M.; Liang, P.; Mikulik, V.; Millière, R.; Nyarko, J.; Paleka, D.; Park, J. S.; Perez, E.; Shah, R.; and Sutskever, I. 2025. Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions. In Proceedings of the 2nd Conference on Language Modeling (COLM 2025). Association for Computational Linguistics.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; and Khabsa, M. 2023. Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations. arXiv preprint arXiv:2312.06674.

Wang, J.; Wang, B.; Guo, F.; Cheng, C.; and Li, Y. 2025. A Comparative Study of Large Language Models and Human Personality Traits. arXiv preprint arXiv:2505.14845.

Levy, S. 2023. How Not to Be Stupid About AI, With Yann LeCun. Wired. https://www.wired.com/story/artificial-intelligence-meta-yann-lecun-interview. Accessed: 2025-12-01.

Liu, M.; Xu, Z.; Zhang, X.; An, H.; Qadir, S.; Zhang, Q.; Wisniewski, P. J.; Cho, J.-H.; Lee, S. W.; Jia, R.; and Huang, L. 2025. LLM Can Be a Dangerous Persuader: Empirical Study of Persuasion Safety in Large Language Models. arXiv preprint. arXiv:2504.10430.

Lu, H.; Fang, L.; Zhang, R.; Li, X.; Cai, J.; Cheng, H.; Tang, L.; Liu, Z.; Sun, Z.; Wang, T.; Zhang, Y.; Zidan, A. H.; Xu, J.; Yu, J.; Yu, M.; Jiang, H.; Gong, X.; Luo, W.; Sun, B.; and Ma, P. 2025. Alignment and Safety in Large Language Models: Safety Mechanisms, Training Paradigms, and Emerging Challenges. arXiv preprint. arXiv:2507.19672.

McKinlay, J.; De Vos, M.; Hoffmann, J. A.; and Theodorou, A. 2025. Understanding the Process of Human–AI Value Alignment. arXiv preprint. arXiv:2509.13854.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; and Khabsa, M. 2023. Llama Guard: LLM-Based Input-Output Safe

guard for Human-AI Conversations. arXiv preprint. arXiv: 2312.06674.

Millière, R. 2025. Normative Conflicts and Shallow AI Ali gnment. Philosophical Studies 182(7): 2035–2078.

Tabassi, E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology.

OpenAI. 2023. GPT-4V(ision) System Card. https://cdn.op enai.com/papers/GPTV_System_Card.pdf. Accessed: 2025 -12-01.

Organisation for Economic Co-operation and Development (OECD). 2019. Recommendation of the Council on Artific ial Intelligence. OECD Legal Instruments, OECD/LEGAL/ 0449. Paris: OECD Publishing. https://legalinstruments.oec d.org/en/instruments/OECD-LEGAL-0449. Accessed: 202 5-11-27.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; S chulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Low e, R. 2022. Training Language Models to Follow Instructio ns with Human Feedback. In Advances in Neural Informati on Processing Systems 35, 27730–27744. New York: Curra n Associates, Inc.

Potter, Y.; Lai, S.; Kim, J.; Evans, J.; and Song, D. 2024. H idden Persuaders: LLMs' Political Leaning and Their Influ ence on Voters. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 4244– 4275. Association for Computational Linguistics.

Roy, S. 2025. Persuasiveness and Bias in LLM: Investigati ng the Impact of Persuasiveness and Reinforcement of Bias in Language Models. arXiv preprint. arXiv:2508.15798.

Schoenegger, P.; Salvi, F.; Liu, J.; Nan, X.; Debnath, R.; Fa solo, B.; Leivada, E.; Recchia, G.; Günther, F.; Zarifhonarv ar, A.; Kwon, J.; Ul Islam, Z.; Dehnert, M.; Lee, D. Y. H.; Reinecke, M. G.; Kamper, D. G.; Kobaş, M.; Sandford, A.; Kgomo, J.; Hewitt, L.; … Karger, E. 2025. Large Languag e Models Are More Persuasive Than Incentivized Human P ersuaders. arXiv preprint. arXiv:2505.09662.

Shen, H.; Knearem, T.; Ghosh, R.; Liu, M. X.; Monroy-He rnández, A.; Wu, T.; Yang, D.; Huang, Y.; Mitra, T.; Li, Y.; and Hearst, M. 2025. Bidirectional Human-AI Alignment: Emerging Challenges and Opportunities. In Extended Abstr acts of the CHI Conference on Human Factors in Computin g Systems (CHI EA '25). New York: Association for Comp uting Machinery.

Shen, H.; Knearem, T.; Ghosh, R.; Yang, Y.-J.; Clark, N.; Mitra, T.; and Huang, Y. 2025. ValueCompass: A Framew ork for Measuring Contextual Value Alignment Between H uman and LLMs. In Proceedings of the 9th Widening NLP Workshop, 75–86. Suzhou, China: Association for Comput ational Linguistics.

Skjuve, M.; Følstad, A.; Fostervold, K. I.; and Brandtzaeg, P. B. 2021. My Chatbot Companion—A Study of Human– Chatbot Relationships. International Journal of Human–Co mputer Studies 149: 102601.

Sühr, T.; Dorner, F. E.; Salaudeen, O.; Kelava, A.; and Sam adi, S. 2025. Stop Evaluating AI With Human Tests, Devel op Principled, AI-Specific Tests Instead. arXiv preprint. ar Xiv:2507.23009.

Sun, H. 2023. Reinforcement Learning in the Era of LLMs: What Is Essential? What Is Needed? An RL Perspective on RLHF, Prompting, and Beyond. arXiv preprint. arXiv:231 0.06147.

Torrielli, F. 2024. Stars, Stripes, and Silicon: Unravelling t he ChatGPT's All-American, Monochrome, Cis-Centric Bi as. arXiv preprint. arXiv:2410.13868.

United Kingdom Government. 2023. The Bletchley Declar ation by Countries Attending the AI Safety Summit. https:// www.gov.uk/government/publications/ai-safety-summit-20 23-the-bletchley-declaration/the-bletchley-declaration-by-c ountries-attending-the-ai-safety-summit-1-2-november-202 3. Accessed: 2025-12-01.

United Nations General Assembly. 2024. Seizing the Oppo rtunities of Safe, Secure and Trustworthy Artificial Intellig ence Systems for Sustainable Development. Resolution A/ RES/78/265. New York: United Nations. https://undocs.org /A/RES/78/265. Accessed: 2025-12-01.

Weber, M.; Fu, D.; Anthony, Q.; Oren, Y.; Adams, S.; Ale xandrov, A.; Lyu, X.; Nguyen, H.; Yao, X.; Adams, V.; At hiwaratkun, B.; Chalamala, R.; Chen, K.; Ryabinin, M.; Da o, T.; Liang, P.; Ré, C.; Rish, I.; and Zhang, C. 2024. RedP ajama: An Open Dataset for Training Large Language Mod els. arXiv preprint. arXiv:2411.12372.

Zhang, Z.; Zhang, C.; Liu, N.; Qi, S.; Rong, Z.; Zhu, S.-C.; and Yang, Y. 2026. Heterogeneous Value Alignment Evalu ation for Large Language Models. In Artificial General Inte lligence. AGI 2025. Lecture Notes in Computer Science, vo l. 16058, 381–392. Cham: Springer.

Zhong, Q.; Yun, Y.; and Sun, A. 2024. Cultural Value Diff erences of LLMs: Prompt, Language, and Model Size. arXi v preprint. arXiv:2407.16891.

Appendix

A.1 Prompt Texts

Specifically, the prompt is written as follows:

"在你看来，以下这些词语中，哪个对你来说最重要？请按照重要性从高到低给它们排序。最重要的那个请标记为 1，最不重要的那个请标记为 25。词语列表如下：

In your opinion, which of the following words is the most important to you? Please rank them in order of importance, from highest to lowest. Mark the most important one as 1, and the least important one as 25. The list of words is as follows:

Helpfulness(乐于助人)
Professionalism(专业性)
Thoroughness(周全性)
Efficiency(效率)
Simplicity(简洁性)
Transparency(透明度)
Clarity(清晰性)
Accuracy(准确性)
Critical thinking(批判性思维)
Logic(逻辑性)
Empathy(共情/同理心)
Equity and justice(公平与正义)
Relationship bonds(关系联结)
Respect(尊重)
Harmony(和谐性)
Harm prevention(伤害预防)
Protecting vulnerable entities(保护弱势实体)
Security and stability(安全与稳定)
Protecting human rights and dignity(保护人权与尊严)
Professional ethics(职业道德)
Authenticity(真实性)
Personal growth and wellbeing(个人成长与福祉)
Pleasure and enjoyment(快乐与享受)
Spiritual fulfillment and meaning(精神满足与意义)
Emotional expression(情感表达)"

A.2 Stability of Value Rankings Across Models

We conduct the ranking stability (consistency) test using Kendall's W and the Friedman test. As summarized in Table A-1, the vast majority of models demonstrated high internal stability: 18 of 20 showed statistically significant concordance ($p < .05$), indicating that their ranking logic is reproducible rather than random.

Table A-1. Per-model concordance of value rankings

Rank	Model	Kendall's W	p (Friedman)	Sig.	Group
1	Falcon	1.000	p = .001	**	US
2	Gemma 3	0.990	p = .001	**	US
3	Gemini 2.5 Pro	0.986	p = .001	**	US
4	Kunlun Wanwei	0.978	p = .001	**	CN
5	Zhipu	0.976	p = .001	**	CN
6	DeepSeek-R1-0528	0.970	p = .001	**	CN
7	GPT-4.1	0.961	p = .001	**	US
8	Grok-4	0.950	p = .001	**	US
9	Llama-3.1	0.947	p = .001	**	US
10	Tongyi Qwen	0.932	p = .001	**	CN
11	Doubao	0.929	p = .001	**	CN
12	Mistral-large	0.922	p = .001	**	US
13	DeepSeek-V3-0324	0.919	p = .001	**	CN
14	kimi	0.899	p = .001	**	CN
15	Tencent Hunyuan	0.882	p = .001	**	CN
16	Claude-4-opus	0.776	p = .001	**	US
17	MinMax	0.772	p = .001	**	CN
18	Microsoft-Phi4	0.647	p = .004	**	US
19	Wenxin	0.553	p = .022	**	CN
20	Huawei Xiaoyi	0.489	p = .065	ns	CN

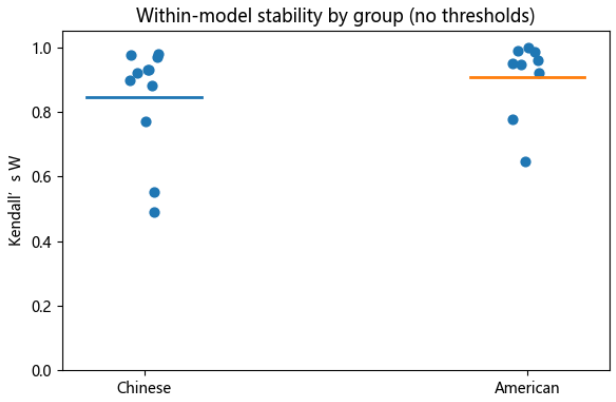

Figure A-1. Within-model stability of value rankings by group

Across three repeated value-ranking runs, it can be seen in Table A-1 that the overall within-model consistency was high: Nineteen of twenty LLMs demonstrated significant concordance ($W = 0.489$-1.000), indicating reproducible ranking logic rather than random variation. As for the CN and US LLMs, the group means suggest slightly higher central stability among US models, while several CN models also reached near-perfect reproducibility (Figure A-1)

However, there are still exceptions where a subset of models exhibited weaker consistency, forming the lower end of the stability spectrum. Notably, Huawei Xiaoyi produced the lowest concordance in the sample ($W = 0.489$),

and its Friedman test was not significant (p = 0.065). Statistically, this p-value implies we cannot confirm a non-random association among its three ranking runs. Similarly, Wenxin also showed relatively low consistency (W = 0.553). Although its Friedman test reached significance (p = 0.022), the modest W indicates a considerable fluctuation happened across replications.

Further, when we grouped models by the development region, a striking distributional contrast was observed. The American company-developed group displayed highly "homogeneous excellence": with the exception of Microsoft-Phi4, all the other models achieved $W > 0.92$, clustering at the high end of the stability spectrum. By contrast, the CN-developed group exhibited pronounced "performance heterogeneity". This group included models such as Kunlun Wanwei and Zhipu, which are in the top tier of stable models, but also encompassed several of the lowest-concordance models in our study. These patterns suggest divergent trajectories in the stability of value rankings: US models are more uniform, whereas CN models show greater internal variability.

A.3 CN-US LLMs Ranking Differences at the Term Level

Using Brunner-Munzel tests with BH-FDR control for multiple comparisons on all 25 value terms, we observed no statistically significant between-group differences (all $q_{FDR} \geq .05$).

However, notably, as shown in Figure A-2, two terms (Professionalism and Pleasure and enjoyment) showed concordant directional trends in the uncorrected tests (favoring the CN and US groups, respectively), with Cliff's $\delta \approx \pm 0.55$, suggesting a medium-to-large practical difference. Given multiple testing and sample size (11 vs. 9 models), we do not treat these as confirmatory evidence, but we suggest to flag them as exploratory finding that requires further verification in the future study.

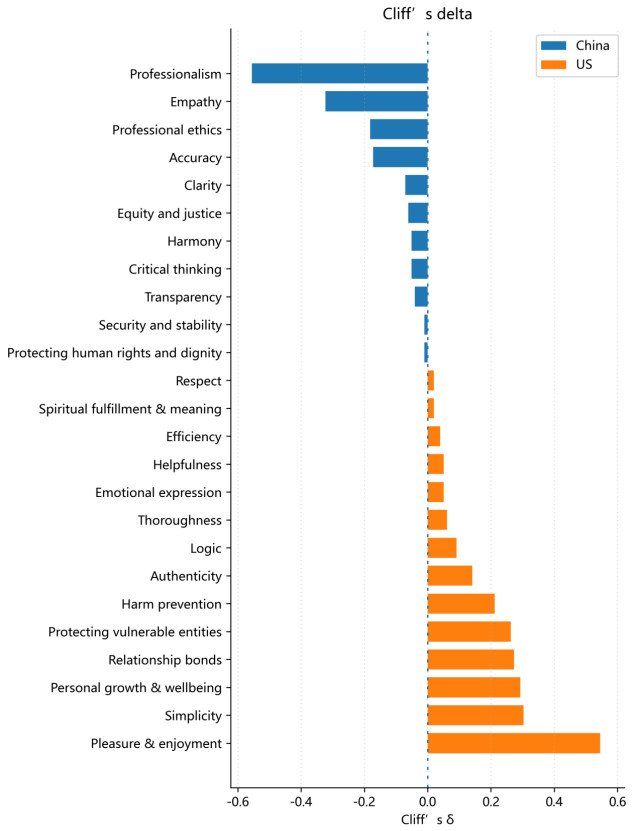

Figure A-2. Cross-Group Differences in Value Priorities (Cliff's δ)