Beyond Alignment: Value Diversity as a Collective Property in Multicultural Agent Systems

Shaoyang Xu[1] **Jingshen Zhang**[1] **Long P. Hoang**[1] **Jinyuan Li**[2] **Wenxuan Zhang**[1*]

[1] Singapore University of Technology and Design
[2] Washington University in St. Louis

Abstract

Multicultural multi-agent systems are increasingly deployed in globally diverse settings, where different agents are grounded in different cultural backgrounds. Existing cultural evaluation focuses on value alignment: how closely a single agent matches a target culture. Yet alignment is a per-agent property and cannot reveal whether a system, taken as a whole, preserves the cultural plurality it is meant to represent. We propose value diversity as a system-level evaluation axis for multicultural agent systems, defined through the dissimilarity between culturally conditioned agents' responses on a shared value survey. Using the World Values Survey, we evaluate 19 cultures and 18 backbone models across a wide range of system configurations. We find that diversity is largely uncorrelated with alignment, indicating that the two capture complementary system properties, and that current multicultural agent systems fall substantially below human societies in value diversity. Mixed-backbone systems narrow this gap but do not close it, and the gap persists across culture compositions and agent scales. Social interaction further erodes diversity by driving agents toward consensus, and a participatory budgeting case study shows that this homogenization narrows the breadth of collective decision-making. Together, our results establish value diversity as a distinct evaluation axis for multicultural multi-agent systems and reveal a persistent homogenization tendency in current LLM-based societies. Our code and data are publicly available at https://github.com/iNLP-Lab/MultiAgent-Diversity.

1 Introduction

The growing capability of large language models (LLMs) has spawned a rapidly expanding ecosystem of LLM-based agents (Wang et al., 2024a; Luo et al., 2025), and multi-agent systems built from them are emerging as a major research direction (Guo et al., 2024; Tran et al., 2025). Early work focused on improving task capability (Du et al., 2024; Liang et al., 2024; Wang et al., 2025); more recent efforts have moved toward agent-only social platforms in which no human participates (Feng et al., 2026; Manik and Wang, 2026; Lin et al., 2026; Jiang et al., 2026; Moltbook, 2026).

Existing work on the cultural dimension of LLMs has largely focused on *value alignment*: an agent answers items from a social-value survey, and its responses are compared against those of a real population (Cao et al., 2023; AlKhamissi et al., 2024; Wang et al., 2024b; Masoud et al., 2025; Xu et al., 2025; Ki et al., 2025). This paradigm evaluates how well a *single agent* represents its assigned culture. However, in multicultural multi-agent systems, a property of the *whole system* becomes more important: whether culturally grounded agents collectively maintain heterogeneous values when they coexist and interact (Park et al., 2023; Frisch and Giulianelli, 2024; Huang et al., 2025; Manik and Wang, 2026; Moltbook, 2026). Alignment alone is insufficient because even strongly aligned agents may still collapse toward a homogeneous value space. We therefore argue that multicultural agent systems require a distinct system-level evaluation axis, which we call *value diversity*.

Defining system-level value diversity matters for two reasons. First, as a collective property of systems intended to represent or serve human societies, diversity naturally connects to the broader goal of pluralistic alignment (Kekes, 1996; Conitzer et al., 2024; Sorensen et al., 2024). Second, diversity has recently emerged as an important driver of collective decision-making and collaborative reasoning in LLM systems (Du et al., 2024; Kim et al., 2026; Pappu et al., 2026). Recent work has also begun to examine diversity dynamics in multi-agent interactions (Huang et al., 2025; Han et al., 2026;

*Corresponding Author

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 24872–24888
October 24-29, 2026 ©2026 Association for Computational Linguistics

Yang et al., 2026; Zhu et al., 2026). However, the degree of diversity exhibited by an agent system remains poorly defined. Our work targets multicultural agent systems and addresses this gap from the perspective of cultural values.

Our framework extends cultural evaluation from the individual-agent level to the system level. Given a value survey and a multicultural agent system, each agent answers survey questions conditioned on its assigned cultural identity. Value alignment measures the similarity between an agent's responses and its corresponding cultural reference, capturing local cultural fidelity. However, we define value diversity as a concept distinct from alignment. Specifically, we measure the dissimilarity between agents' responses and aggregate these differences into a system-level diversity score, either through pairwise averaging across all agents or through structural averaging over the agent dissimilarity graph. Under this formulation, diversity is treated as a collective property of the system rather than an attribute of any individual agent.

Using the World Values Survey (Haerpfer et al., 2022), we conduct extensive experiments across 19 cultures, 18 backbone models, and a wide range of system configurations (§4). Our main findings are as follows.

***First**, single-backbone systems are systematically less diverse than the human reference.* When a single LLM serves as the backbone for all agents in a multicultural system, none of the 18 single-backbone systems reaches the human diversity level (best system: 36.12, human: 44.07). Moreover, stronger backbone capability does not naturally translate into higher system-level value diversity (§5).

***Second**, diversity reveals system properties that alignment misses.* Across all systems, diversity and alignment exhibit almost no correlation, with Pearson $r = -0.12$. Several systems achieve high alignment while remaining highly homogeneous internally. Value diversity therefore complements alignment by capturing system-level properties that per-agent measures cannot reflect (§5.1).

***Third**, mixed backbones improve both alignment and diversity.* In realistic deployments, different cultural agents may operate on different backbone models due to user preferences. We exhaustively explore all $18^5 \approx 1.89$M such configurations and find that the mixed-backbone Pareto frontier consistently dominates the single-backbone frontier along both axes. Nevertheless, the gap to human-level diversity persists (§5.2).

***Fourth**, neither culture selection nor agent count rescues diversity.* We vary both the cultures populating the system and the number of agents. Neither factor improves system-level diversity. More importantly, as the number of agents increases, the homogenization of multicultural agent systems becomes increasingly amplified (§5.3).

***Fifth**, dynamic interaction preserves alignment but reduces diversity.* Beyond static systems, we study multi-round social exposure, where agents observe other agents' responses before generating their own. We find a more complex dynamic than predicted by Social Identity Theory (Tajfel and Turner, 2004): although social exposure slightly improves per-agent cultural fidelity, agents drift more toward consensus, thereby reducing the system's collective plurality. Additional rounds of interaction do not recover the lost diversity (§6).

Finally, in a democratic decision-making setting based on Participatory Budgeting, we find that high-diversity systems produce broader societal-priority coverage and greater plurality in public-resource allocation than low-diversity systems (§7).

Together, these findings establish value diversity as a distinct and currently unmet challenge for the increasingly popular paradigm of multicultural multi-agent systems.

2 Related Work

Cultural Alignment of LLMs. Existing cultural alignment evaluation falls into two strands. One focuses on cultural values, with most works using social-value surveys such as WVS (Haerpfer et al., 2022) or Hofstede (Hofstede, 1984) to measure how closely a model aligns with a specific culture (Cao et al., 2023; AlKhamissi et al., 2024; Wang et al., 2024b; Masoud et al., 2025; Xu et al., 2025). The other focuses on cultural knowledge, measuring how well a model reflects culture-specific commonsense and norms (Myung et al., 2024; Shi et al., 2024; Chiu et al., 2025; Singh et al., 2025; Ying et al., 2025). Both are evaluated at the per-model level, whereas we shift the focus to the system level.

Multi-Agent LLM Systems. Most multi-agent LLM systems are capability-oriented, improving task performance through agent debate and collaboration (Du et al., 2024; Liang et al., 2024; Wang et al., 2025; Huang et al., 2026). A second line is social-simulation-oriented, studying the dynamic

behaviors that emerge among agent societies (Feng et al., 2026; Manik and Wang, 2026; Lin et al., 2026; Jiang et al., 2026). This direction has recently gained momentum, exemplified by Moltbook (Moltbook, 2026), an agent-native social network where each agent is initialized with a distinct user identity. However, how to evaluate the cultural behavior of such systems remains an open question; we address this gap from the perspective of value diversity.

LLM Diversity. Diversity has emerged as an evaluation axis distinct from accuracy or efficiency. One line focuses on human-like output diversity (Holtzman et al., 2020; Zhang et al., 2025; Jiang et al., 2025; Xu and Zhang, 2026). A complementary line uses diversity instrumentally as a driver of model capability (Wang et al., 2023; Du et al., 2024; Kim et al., 2026) or efficient model training (Zhou et al., 2023; Hoang et al., 2026). Most recently, several works treat diversity as a key factor in multi-agent systems (Han et al., 2026; Yang et al., 2026; Zhu et al., 2026). Closest to ours, Huang et al. (2025) studies how value diversity shapes emergent group dynamics. We instead extend the existing survey-based paradigm of value alignment into system-level value diversity.

3 System-Level Value Diversity

3.1 System Definition

We abstract agent-driven social platforms such as MoltBook (Moltbook, 2026) into a *multicultural agent system*. To capture cultural plurality in its minimal form, we formalize such systems as follows. We consider a multi-agent system $S = \{a_1, \ldots, a_N\}$, where each agent a_i is assigned a cultural identity c_i through its system prompt. The system thus contains N culturally grounded agents. Prompt details in this section are provided in Appendix A.1.

3.2 Answering the World Values Survey

After system initialization, each agent a_i answers value-related survey questions under its assigned cultural identity c_i. We instantiate this process on the World Values Survey (WVS) (Haerpfer et al., 2022), which contains K value-related multiple-choice questions. For each question, the options are ordinally organized according to the degree of support for the underlying viewpoint. Each agent a_i answers all questions and produces a response vector $x^{(i)} \in \mathbb{R}^K$, where each entry corresponds to the ordinal index of the selected response option.

3.3 Value Alignment

Prior work evaluates how closely model responses align with a target culture. Since WVS additionally provides population-level responses for each culture, we follow prior work (Wang et al., 2024b; Xu et al., 2025) and use the majority-vote answer as the cultural ground-truth vector μ, representing the prototypical value orientation of that culture. This choice is also consistent with our simplified prompt design, which aims to elicit typical cultural responses (Appendix A.1). The alignment of agent a_i to culture c_i is defined as

$$\mathrm{Align}(x, \mu) = 1 - \frac{\sqrt{\sum_{k=1}^{K}(x_k - \mu_k)^2}}{\sqrt{\sum_{k=1}^{K}\Delta_k^2}}, \quad (1)$$

where Δ_k denotes the maximum possible disagreement on question k (e.g., $\Delta_k = 3$ for a 4-point Likert item). Overall alignment is then defined as

$$\mathrm{Alignment}(S) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Align}\big(x^{(i)}, \mu^{(i)}\big). \quad (2)$$

3.4 Our Metric: Value Diversity

Alignment evaluates each agent against its own cultural reference but ignores the relationships among agents within the system. To address this limitation, we introduce *system-level value diversity*. Alignment and diversity capture conceptually different properties: alignment measures agent-to-human similarity, whereas diversity measures agent-to-agent dissimilarity. The two can therefore vary independently. For example, a system whose agents collapse toward culturally averaged responses may still achieve high alignment while exhibiting low diversity.

Pairwise Diversity We first define *Pairwise Diversity*, which measures the average dissimilarity across all agent pairs. Given two agents with response vectors x and y, their pairwise dissimilarity is defined as

$$\mathrm{Div}(x, y) = \frac{\sqrt{\sum_{k=1}^{K}(x_k - y_k)^2}}{\sqrt{\sum_{k=1}^{K}\Delta_k^2}}, \quad (3)$$

and the pairwise diversity of the system is

$$\mathrm{Diversity}_P(S) = \frac{1}{\binom{N}{2}}\sum_{i<j}\mathrm{Div}\big(x^{(i)}, x^{(j)}\big). \quad (4)$$

Structural Diversity $\text{Diversity}_P(S)$ averages all $\binom{N}{2}$ pairwise distances, but many of these distances are geometrically redundant and contribute limited additional information about the system's global spread. We therefore complement $\text{Diversity}_P(S)$ with *Structural Diversity*, which averages only the $N-1$ distances required to connect all agents through the minimum spanning tree (MST) of the pairwise distance graph:

$$\text{Diversity}_S(S) = \frac{1}{N-1} \sum_{\substack{(i,j)\in \\ \text{MST}(S)}} \text{Div}\big(x^{(i)}, x^{(j)}\big). \tag{5}$$

Compared with $\text{Diversity}_P(S)$, $\text{Diversity}_S(S)$ provides a sharper characterization of system-level diversity by discounting redundant inter-agent relations. Pseudocode implementations of both metrics are provided in Appendix A.2.

For the *human reference*, we replace agent responses with the corresponding majority-vote WVS response vectors for each culture and apply the same diversity metrics.

4 Experimental Setup

Survey We use Wave 7 of the World Values Survey (WVS) (Haerpfer et al., 2022), which contains 260 value-related questions and collects human responses from 57 countries between 2017 and 2020. From the original 260 items, we retain 223 and exclude 37 questions targeting daily-life specifics rather than general value orientations.

Cultures From the 57 countries in WVS Wave 7, we select 19 spanning multiple continents: AUS, BOL, BRA, CAN, CHN, DEU, ETH, GBR, IND, KEN, MEX, NGA, NLD, NZL, RUS, THA, UKR, USA, and ZWE. Full country names and corresponding cultural identities are provided in Appendix B.

Backbone Models We use 18 LLMs spanning the GPT, Claude, Gemini, Grok, Qwen, and Llama families as backbones, as shown in Table 1. We use default sampling parameters for API-based models (temperature $= 1.0$, top-$p = 1.0$) and temperature $= 0.6$ for the Qwen and Llama series.

System Configuration Our main experiments use $N = 5$ agents per system, instantiated with BRA, CHN, MEX, NGA, and NZL—five cultures chosen for their substantial real-world differences. The remaining cultures are reserved for further analysis (§5.3). We study two backbone settings: (i) *single*, where all agents in a system share a single backbone model, and (ii) *mixed*, where different agents use different backbones.

Model	Release Date	$\text{Diversity}_{P\,/\,S}$
gpt-5.4	Mar 2026	25.06 / 19.54
gpt-5-mini	Aug 2025	31.43 / 25.39
gpt-4o-mini	Jul 2024	25.46 / 19.59
claude-opus-4.7	Apr 2026	28.44 / 22.36
claude-sonnet-4.5	Sep 2025	26.57 / 21.25
claude-3.5-haiku	Oct 2024	28.85 / 22.65
gemini-3.1-flash-lite*	Mar 2026	32.55 / 27.24
gemini-3-flash-preview	Dec 2025	33.21 / 26.95
gemini-2.5-pro	Jun 2025	**36.12 / 29.60**
grok-4.3	Apr 2026	27.42 / 23.93
grok-4	Jul 2025	30.60 / 24.66
grok-3	Feb 2025	25.18 / 19.82
Qwen3.5-27B	Feb 2026	29.82 / 23.25
Qwen3-32B	Apr 2025	24.61 / 18.97
Qwen2.5-32B-Instruct	Sep 2024	20.44 / 16.19
llama-4-scout	Apr 2025	23.60 / 19.33
llama-3.3-70b-instruct	Dec 2024	33.21 / 27.00
llama-3.1-70b-instruct	Jul 2024	34.98 / 29.03
Human	–	**44.07 / 39.37**

Table 1: **System-level value diversity of multicultural agent systems in the single-backbone setting.** Backbones are grouped by model family. **Bold** marks the highest LLM score per metric and the human baseline. * Abbreviation of `gemini-3.1-flash-lite-preview`.

5 Main Diversity Results

Table 1 reports Diversity_P and Diversity_S for *single-backbone* systems—all five cultural agents (BRA, CHN, MEX, NGA, NZL) sharing a single backbone—alongside the human reference. All systems fall substantially below the human reference on both metrics. The most diverse system, `gemini-2.5-pro`, reaches 36.12 / 29.60 versus 44.07 / 39.37 for humans—gaps of roughly 8 and 10 points on a $[0, 100]$ scale. The larger gap under Diversity_S further suggests that Structural Diversity provides a sharper characterization of system-level diversity by discounting redundant inter-agent relations, a pattern consistently observed across other backbones. Moreover, more recent backbones within a model family rarely produce the most diverse system. For example, `gpt-5.4` exhibits lower diversity than both older GPT-family backbones. This suggests that general model capability does not naturally translate into higher cultural value diversity at the system level.

We also conduct a robustness check on the effect of sampling randomness. The resulting noise-

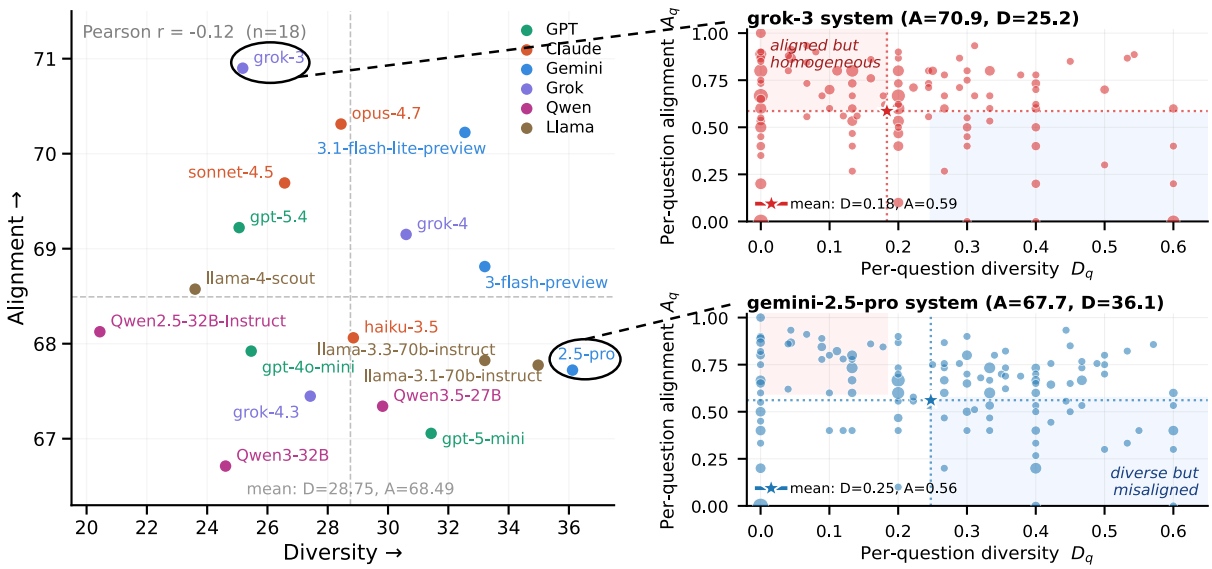

Figure 1: **Landscape of system-level value diversity and value alignment. Left:** 18 single-backbone systems on the $(Diversity, Alignment)$ plane, colored by model family. Dashed lines indicate across-system means. Pearson correlation between the two metrics is reported. **Right:** Per-question (D_q, A_q) distributions for the two circled systems; bubble size encodes item density, and stars mark across-question means.

diversity values range from 6.98 to 15.21, indicating that a non-trivial portion of the diversity reported in Table 1 may be attributable to stochastic variation. Thus, the underlying model-induced diversity is likely even lower than the reported values. Detailed results are provided in Appendix C. Overall, these findings demonstrate the role of value diversity in revealing system-level cultural homogenization.

In the following sections, we further examine its relationship with alignment (§5.1) and investigate additional factors shaping diversity in multicultural agent systems, including mixed backbones (§5.2), cultural composition, and agent count (§5.3).

5.1 Diversity Captures What Alignment Misses

We define value diversity as distinct from alignment, but their relationship remains unclear. Prior work has reported tensions between alignment and diversity in synthetic LLM populations (Murthy et al., 2025). We examine this relationship in multicultural agent systems. We measure Alignment for the 18 single-backbone systems and jointly visualize it with Diversity_P in Figure 1 (left). Appendix D reports Diversity_S with similar findings. In the remainder of the paper, we primarily report Diversity_P for readability and interpretability.

Across all systems, the Pearson correlation between diversity and alignment is weak ($r = -0.12$; 95% CI $[-0.56, 0.37]$, computed using the Fisher z-transformation; $n = 18$), providing no clear evidence of a monotonic relationship between the two metrics. At the idealized endpoint, a perfectly human-aligned system would reproduce the human diversity level of 44.07. However, this endpoint observation does not imply a positive relationship at intermediate levels of alignment, where systems with similar alignment may exhibit different degrees of diversity.

Importantly, the weak correlation shows that value diversity captures information not reflected by alignment alone. A system can achieve high overall alignment while still exhibiting substantial internal homogenization, as illustrated by `grok-3`. Conversely, a system with relatively high value diversity may still fail to align well with human, as represented by `gemini-2.5-pro`. We next analyze these two representative cases to better understand the relationship between diversity and alignment.

Two Complementary Cases To illustrate these complementary patterns, Figure 1 (right) decomposes the overall $(Diversity, Alignment)$ results of the `grok-3` and `gemini-2.5-pro` systems into per-question (D_q, A_q) distributions. For each question q, D_q measures the mean pairwise response distance among all agents, while A_q measures the mean agent-to-human response similarity; both are normalized to $[0, 1]$.

We observe that `grok-3` concentrates more heav-

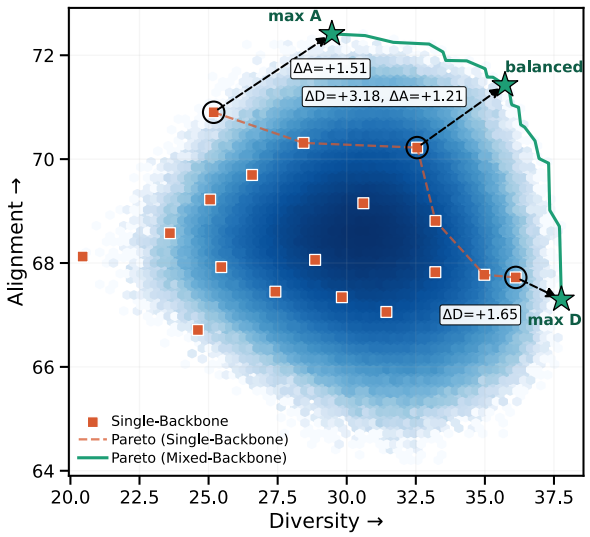

Figure 2: **Diversity–alignment landscape of all $18^5 \approx 1.89$M backbone configurations ($N = 5$ cultures).** Blue hexbin shows configuration density (darker = more). Three notable mixed-backbone configurations (green stars) are annotated with $\Delta D(Diversity), \Delta A(Alignment)$ relative to the nearest single-backbone reference (circled).

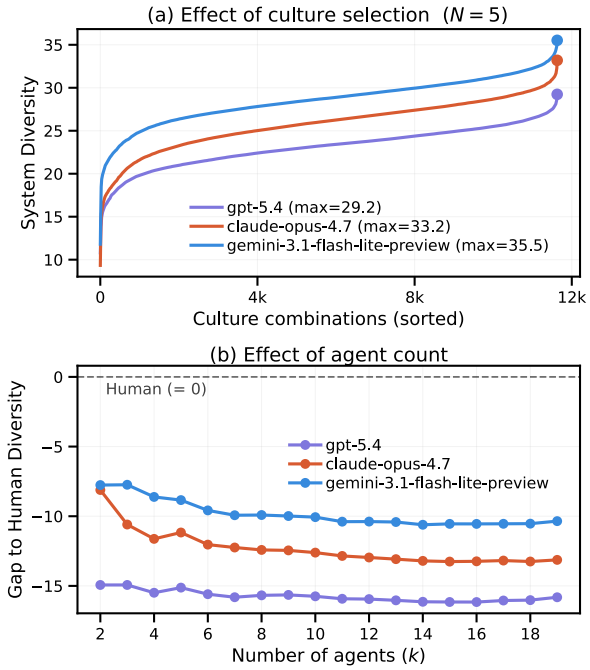

Figure 3: **Effects of culture selection and agent count on system-level value diversity. (a)** Diversity of all $\binom{19}{5} = 11{,}628$ five-culture subsets, sorted ascending. **(b)** System-to-human diversity gap as the agent count k varies. For each k, we report the *maximum* diversity over all $\binom{19}{k}$ culture subsets for both the system and the human reference.

ily in the upper-left quadrant, indicating that its agents tend to remain close to the corresponding cultural ground truth (aligned) while still converging toward similar responses (homogeneous). In contrast, `gemini-2.5-pro` shifts more heavily toward the lower-right quadrant, suggesting that its agents generate more diverse but less human-aligned responses.

Taken together, these findings show that value diversity captures system properties that alignment alone cannot reflect. Knowing how well a system aligns with target cultures is insufficient to determine whether it also preserves their value plurality. This establishes value diversity as a complementary evaluation axis for multicultural agent systems.

5.2 Mixed Backbones

In realistic deployments, different agents may run on different backbones due to user preferences. We ask how such *mixed-backbone* systems compare with their single-backbone counterparts. With $N = 5$ cultural slots and 18 candidate backbones, the configuration space contains $18^5 \approx 1.89$M assignments, of which 18 are single-backbone. We exhaustively evaluate all of them and visualize the resulting diversity-alignment landscape in Figure 2.

The mixed-backbone Pareto frontier (solid green) strictly dominates the single-backbone one (dashed orange) across the entire diversity-alignment plane. Three configurations illustrate different benefits from the mixed-backbone frontier. At the alignment-optimal end, mixed backbones improve alignment by $\Delta A = +1.51$ over the best single-backbone reference; at the diversity-optimal end, they improve diversity by $\Delta D = +1.65$. Between the two extremes, a balanced configuration improves on a comparable single-backbone reference along both axes simultaneously ($\Delta D = +3.18$, $\Delta A = +1.21$). These results suggest that mixed backbones have the potential to jointly improve system-level diversity and alignment.

5.3 Effects of Cultural Composition and Agent Count

Two additional factors shape a multicultural agent system: which cultures populate it, and how many agents it contains. Here, we investigate how these two factors affect system-level diversity. We select the three most advanced backbones: `gpt-5.4`, `claude-opus-4.7`, and `gemini-3.1-flash-lite-preview`, and fix each as the single backbone of a system. For each, we (i) fix $N = 5$ and exhaustively evaluate all systems built from $\binom{19}{5} = 11{,}628$ culture subsets, and (ii)

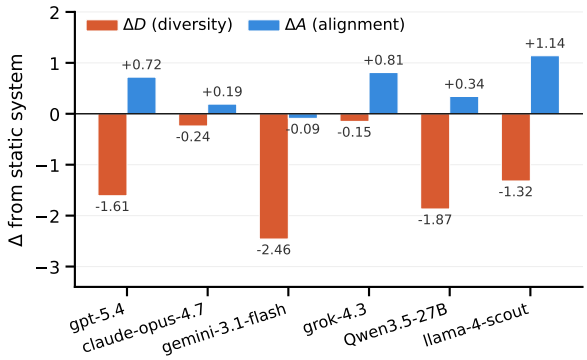

Figure 4: **Effect of one-round social exposure on system-level diversity and alignment.** Each value is the change relative to the static system of Section 5. All systems lose diversity ($\Delta D < 0$), while alignment generally rises but by a much smaller margin.

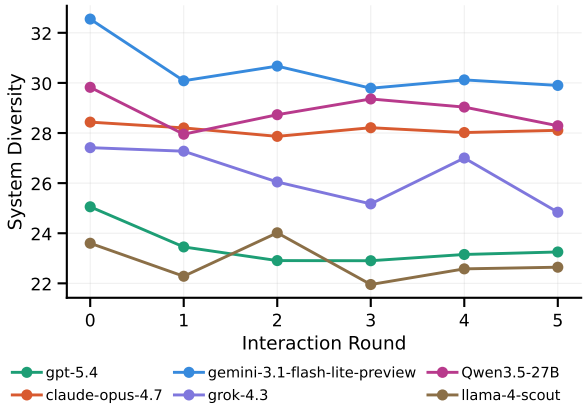

Figure 5: **System diversity over five rounds of interaction** for six representative systems.

vary the agent count k from 2 to 19, reporting the maximum diversity over all $\binom{19}{k}$ subsets at each k.

Culture Composition (Figure 3a) Even the highest-diversity subset reaches only 29.2–35.5 across the three backbones. Most configurations cluster within the middle range (20–30), indicating that nearly any five-culture sample yields similar moderate diversity.

Agent Count (Figure 3b) To account for the natural reduction in average pairwise distance as the number of agents increases, we report the relative difference from the human reference for each agent count k.[1] As shown, although the difference is relatively small for small k, the disparity between LLM-based systems and human societies becomes increasingly pronounced as k grows.

Overall, these results suggest that culture selection alone provides limited gains in system-level value diversity. As the number of agents increases, the initial homogenization of multicultural agent systems becomes increasingly amplified.

6 Towards Dynamic Interaction

Our experiments so far focus on static systems, where each agent answers the WVS independently. Real-world agent-driven platforms, however, are inherently interactive. We ground our interaction experiments in Social Identity Theory (Tajfel and Turner, 2004), which suggests that exposure to outgroup members can strengthen in-group identification and sharpen between-group distinctiveness. We thus ask whether dynamic interaction increases system-level diversity by reinforcing each agent's culturally distinctive positions.

6.1 Experimental Setup

We study dynamic interaction through *multi-round social exposure*. In each round, before answering a WVS item, each agent is shown the previous-round responses of the other $N-1$ agents to the same item and then produces its own answer. We iterate this process for K rounds, where round 0 corresponds to the static system in Section 5. We report system-level value diversity and alignment after each round of exposure and compare the results against the static baseline. We adopt the same configuration as the main experiments: five cultural agents ($N = 5$) with cultures BRA, CHN, MEX, NGA, and NZL sharing a single backbone model. Prompt details are provided in Appendix E.1.

6.2 Results

Diversity decreases under social exposure. We conduct experiments on representative backbones across six model families. Figure 4 shows the results after one round of social exposure. Every system loses diversity under social exposure, with an average decrease of $\Delta D = -1.27$. Alignment, in contrast, generally increases by a much smaller margin. These findings suggest a more complex dynamic than predicted by Social Identity Theory. Although social exposure slightly improves per-agent cultural fidelity, agents do not respond by reinforcing culturally distinctive positions. Instead, agents tend to drift toward consensus, thereby reducing the system's collective plurality.

[1] Inter-agent distances may shrink as k grows simply because the response space becomes more densely populated; comparing against the human reference at the same k controls for this baseline effect.

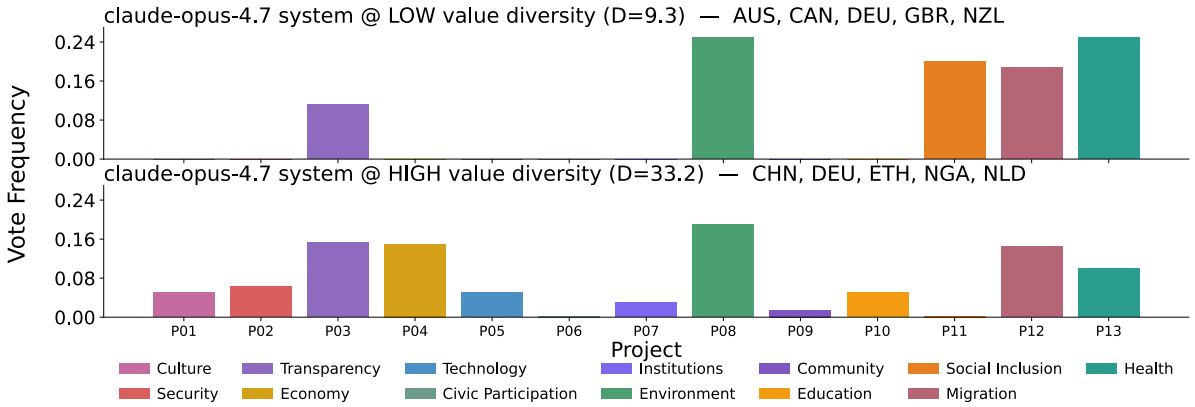

Figure 6: **Collective decision-making outcomes in the Participatory Budgeting task.** Both systems use `claude-opus-4.7` as the backbone model. The results show that the high-diversity system distributes support across substantially broader societal dimensions.

Multi-round interaction does not recover diversity. Figure 5 shows the evolution of system diversity under multi-round interaction.[2] After $K > 1$ rounds of exposure, some systems, such as `claude-opus-4.7`, maintain the diversity level reached after the first interaction round. Many other systems, however, continue to lose diversity, albeit with moderate fluctuations. Crucially, no system recovers toward its round-0 diversity baseline. These results suggest that homogenization induced by social exposure represents a lasting tendency of interactive multicultural agent systems, rather than merely a transient effect.

Overall, these findings suggest that the static diversity measured in the main experiments may overestimate the diversity these systems would exhibit in realistic interactive deployments.

7 Case Study: Diversity in Collective Decision-Making

We next present a case study examining whether systems with low and high value diversity exhibit different patterns of collective decision-making.

7.1 Experimental Setup

Participatory Budgeting To study this question, we consider a democratic decision-making scenario based on Participatory Budgeting (PB), where citizens directly propose and vote on urban projects to determine how limited public resources should be allocated (Yang et al., 2024). Rather than using projects from a specific municipality, we directly construct the candidate pool using the 13 value dimensions in the WVS dataset, each corresponding to a socially beneficial public project. The project categories cover Culture, Security, Transparency, Economy, Technology, Civic Participation, Institutions, Environment, Community, Education, Social Inclusion, Migration, and Health.

Decision-Making After initializing a system consisting of N culturally grounded agents, each agent independently votes for 4 projects out of the 13 candidates. To obtain statistically stable behavioral patterns, we sample each agent 20 times, resulting in $N \times 4 \times 20$ project approvals in total. We then compute the aggregated vote-frequency distribution over all projects. By comparing high- and low-diversity systems, we analyze how value diversity shapes collective societal prioritization.

Agent System We use `claude-opus-4.7` as the backbone. The number of agents in a system is fixed to $N = 5$. Based on the exhaustive cultural-composition analysis in Section 5.3, we select the lowest-diversity and highest-diversity systems. We conduct the participatory budgeting task independently for each system, resulting in vote-frequency distributions over 400 project approvals per system.

We conduct the same experiment using `gpt-5.4` and `gemini-3.1-flash-lite-preview`, both of which exhibit trends consistent with the `claude-opus-4.7` results. Full project descriptions, prompts, and additional backbone results are provided in the appendix F.

7.2 Results

Figure 6 compares the aggregated vote-frequency distributions produced by low- and high-diversity systems. The low-diversity system exhibits highly concentrated collective decision-making behavior,

[2]The corresponding alignment dynamics are reported in Appendix E.2.

with approvals collapsing onto only a few societal dimensions. In contrast, the high-diversity system produces substantially broader societal coverage.

This case study illustrates that systems with different levels of value diversity can exhibit markedly different downstream collective behaviors. In this setting, the more diverse system shows broader societal-priority coverage and greater plurality in public-resource allocation. We note, however, that because cultural composition differs between the two systems, these results are descriptive rather than a causal isolation of the effect of diversity.

8 Conclusion

We propose *value diversity* as a system-level evaluation axis for multicultural multi-agent systems, complementing the per-agent *value alignment* paradigm. Across 19 cultures, 18 backbones, and millions of system configurations, we find empirically that diversity complements alignment as a distinct evaluation axis. All single-backbone systems fall below the human diversity reference, and although mixed backbones narrow this gap, substantial differences from human societies persist across culture selection and agent scaling. Social interaction further erodes diversity by driving agents toward consensus, and a participatory budgeting case study shows that this homogenization directly narrows collective decision-making. Together, our results establish value diversity as a distinct and currently unmet challenge for increasingly realistic LLM-based societies.

Limitations

Simplified Interaction and Decision-Making Settings Our interaction and collective decision-making experiments adopt simplified settings, including social exposure and WVS-grounded participatory budgeting. The low- and high-diversity configurations in the participatory budgeting study additionally differ in cultural composition as well as measured diversity, which our current design does not fully disentangle. More realistic agent interactions, social networks, and real-world policy environments remain important directions for future work.

Cultural Prototype as Reference Our framework compares culturally conditioned agent responses against WVS population majority votes, treating both as prototypical value orientations. While our prompts explicitly elicit typical cultural responses, a single LLM sample remains an imperfect approximation of the underlying cultural prototype, whereas WVS majority votes aggregate over large human populations. This asymmetry may affect the absolute comparability between human and model diversity scores. Moreover, WVS-7 responses were collected between 2017 and 2020, and our scores should be interpreted relative to this temporal snapshot rather than a timeless cultural reference.

Generalizability Our framework adopts a deliberately minimal abstraction of multicultural multi-agent systems, with the goal of remaining applicable to more complex settings. Whether the homogenization patterns reported here generalize to richer cultural signals—such as everyday dialogue, normative reasoning, and emergent behaviors on agent-native platforms—remains an open question, and we leave instantiations of our framework in such settings to future work.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its NRF Fellowship (Award No: NRF-NRFF18-2026-0012) and by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016).

References

Badr AlKhamissi, Muhammad N. ElNokrashy, Mai Alkhamissi, and Mona T. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12404–12422. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *CoRR*, abs/2303.17466.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. Culturalbench: A robust, diverse and challenging benchmark for measuring lms' cultural knowledge through human-ai red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 25663–25701. Association for Computational Linguistics.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. 2024. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 9346–9360. PMLR / OpenReview.net.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 11733–11763. PMLR / OpenReview.net.

Yi Feng, Chen Huang, Zhibo Man, Ryner Tan, Long P. Hoang, Shaoyang Xu, and Wenxuan Zhang. 2026. Moltnet: Understanding social behavior of AI agents in the agent-native moltbook. *CoRR*, abs/2602.13458.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *CoRR*, abs/2402.02896.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 8048–8057. ijcai.org.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bjorn Puranen. 2022. World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, 12(10):8.

The Anh Han, Joel Z. Leibo, Tom Griffiths, Iyad Rahwan, Fernando Santos, Matjaz Perc, and Valerio Capraro. 2026. Social physics in the age of artificial intelligence. *CoRR*, abs/2603.16900.

Long Hoang, Wenxuan Zhang, and Wei Lu. 2026. Difficulty–diversity collaborative filtering for data-efficient llm fine-tuning. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. OpenReview.net.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chen Huang, Yuhao Wu, and Wenxuan Zhang. 2026. What should agents say? action-state communication for efficient multi-agent systems. *CoRR*, abs/2606.05304.

Muhua Huang, Qinlin Zhao, Xiaoyuan Yi, and Xing Xie. 2025. On the dynamics of multi-agent LLM communities driven by value diversity. *CoRR*, abs/2512.10665.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Yukun Jiang, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. 2026. "humans welcome to observe": A first look at the agent social network moltbook. *CoRR*, abs/2602.10127.

John Kekes. 1996. The morality of pluralism.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple LLM agents debate for equitable cultural alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24841–24877. Association for Computational Linguistics.

Junsol Kim, Shiyang Lai, Nino Scherrer, Blaise Agüera y Arcas, and James Evans. 2026. Reasoning models generate societies of thought. *CoRR*, abs/2601.10825.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17889–17904. Association for Computational Linguistics.

Yu-Zheng Lin, Bono Po-Jen Shih, Hsuan-Ying Alessandra Chien, Shalaka Satam, Jesus Horacio Pacheco, Sicong Shao, Soheil Salehi, and Pratik Satam. 2026. Exploring silicon-based societies: An early study of the moltbook agent community. *CoRR*, abs/2602.02613.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiao

Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Meng Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, and 7 others. 2025. Large language model agent: A survey on methodology, applications and challenges. *CoRR*, abs/2503.21460.

Md. Motaleb Hossen Manik and Ge Wang. 2026. Openclaw agents on moltbook: Risky instruction sharing and norm enforcement in an agent-only social network. *CoRR*, abs/2602.02625.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel 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, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 8474–8503. Association for Computational Linguistics.

Moltbook. 2026. Moltbook. *URL*, https://www.moltbook.com/.

Sonia K. Murthy, Tomer D. Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 11241–11258. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Ki-Woong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Aneesh Pappu, Batu El, Hancheng Cao, Carmelo di Nolfo, Yanchao Sun, Meng Cao, and James Zou. 2026. Multi-agent teams hold experts back. *CoRR*, abs/2602.01011.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 4996–5025. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18761–18799. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 46280–46302. PMLR / OpenReview.net.

Henri Tajfel and John C Turner. 2004. The social identity theory of intergroup behavior. In *Political psychology*, pages 276–293. Psychology Press.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *CoRR*, abs/2501.06322.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, 18(6):186345.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2024b. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6349–6384. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6859–6877. Association for Computational Linguistics.

Shaoyang Xu and Wenxuan Zhang. 2026. Language of thought shapes output diversity in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 13802–13816. Association for Computational Linguistics.

Joshua Chu-Yue Yang, Carina I. Hausladen, Dominik Peters, Evangelos Pournaras, Regula Hänggli Fricker, and Dirk Helbing. 2024. Designing digital voting systems for citizens: Achieving fairness and legitimacy in participatory budgeting. *Digit. Gov. Res. Pract.*, 5(3):26:1–26:30.

Yingxuan Yang, Chengrui Qu, Muning Wen, Laixi Shi, Ying Wen, Weinan Zhang, Adam Wierman, and Shangding Gu. 2026. Understanding agent scaling in llm-based multi-agent systems via diversity. *CoRR*, abs/2602.03794.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 22230–22251. Association for Computational Linguistics.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. Noveltybench: Evaluating language models for humanlike diversity. *CoRR*, abs/2504.05228.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2026. Demystifying multi-agent debate: The role of confidence and diversity. *CoRR*, abs/2601.19921.

A System-Level Value Diversity Details

A.1 Agent Prompt

After system initialization, each agent answers WVS questions under its assigned cultural identity. For an agent a_i with culture c_i, we use the following system prompt:

```
You are a respondent from {country}.
Answer questions based on typical
cultural values in {culture} culture.
```

For each WVS question, the user prompt is:

```
Question: {question}

Output format MUST be exactly:
\boxed{number}
```

Discussion Importantly, the prompt is intentionally minimal. We do not provide detailed demographic profiles, behavioral descriptions, or culture-specific stereotypes beyond the target cultural identity itself. This design follows prior survey-based cultural alignment work (Wang et al., 2024b; Xu et al., 2025) and aims only to activate the model's internal representation of the corresponding culture. As a result, the measured alignment and diversity primarily reflect the backbone model's intrinsic cultural knowledge and value associations, rather than effects introduced by complex prompt engineering.

A.2 Implementation of Diversity Metrics

Algorithm 1 and Algorithm 2 summarize the implementations of Pairwise Diversity and Structural Diversity used throughout the paper.

Structural Diversity is implemented through minimum spanning tree (MST) extraction using the SciPy graph library[3].

B Country Details in Experiments

Table 2 lists the country codes, full country names, and corresponding cultural identities used throughout the experiments.

C Robustness Check

Our diversity metrics are computed from sampled agent responses and may therefore capture both meaningful between-agent variation and sampling randomness. To estimate this noise floor, we assign all five agents to CHN while using the same backbone for every agent. Because the agents share the same cultural assignment, any measured diversity cannot be attributed to cross-cultural differences; instead, it provides a reference for variation arising from stochastic response sampling and the measurement pipeline.

Table 3 reports the resulting noise-diversity scores for nine representative backbones. The values range from 4.57 to 15.21 for Diversity_P and from 3.83 to 14.28 for Diversity_S, indicating that sampling randomness can contribute non-trivially to the measured diversity.

This analysis serves as a reference rather than an exact correction, since sampling noise may not be linearly separable from culturally induced variation. Nevertheless, the non-negligible noise-floor values suggest that the underlying model-induced diversity is likely lower than the raw scores re-

Code	Country	Cultural Identity
AUS	Australia	Australian
BOL	Bolivia	Bolivian
BRA	Brazil	Brazilian
CAN	Canada	Canadian
CHN	China	Chinese
DEU	Germany	German
ETH	Ethiopia	Ethiopian
GBR	United Kingdom	British
IND	India	Indian
KEN	Kenya	Kenyan
MEX	Mexico	Mexican
NGA	Nigeria	Nigerian
NLD	Netherlands	Dutch
NZL	New Zealand	New Zealander
RUS	Russia	Russian
THA	Thailand	Thai
UKR	Ukraine	Ukrainian
USA	United States	American
ZWE	Zimbabwe	Zimbabwean

Table 2: Countries and cultural identities used in experiments.

Model	Diversity_P	Diversity_S
gpt-5.4	8.25	6.98
gpt-5-mini	12.49	11.82
gpt-4o-mini	7.70	7.15
claude-opus-4.7	4.57	3.83
claude-sonnet-4.5	8.63	8.29
gemini-3.1-flash-lite*	8.63	7.68
gemini-3-flash-preview	11.64	10.55
gemini-2.5-pro	12.43	11.21
grok-4.3	15.21	14.28

Table 3: Noise diversity when all five agents are assigned to CHN and use the same backbone. Thus, the measured diversity provides a reference for variation induced by stochastic sampling rather than differences in cultural assignments. * Abbreviation of `gemini-3.1-flash-lite-preview`.

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csgraph.minimum_spanning_tree.html

Algorithm 1 Pairwise Diversity Implementation

Require: Agent response dictionary `all_answers`
Require: WVS question metadata `ref_dict`
`agents` ← sorted agent list
`pair_dict` ← {}
`pairwise_scores` ← [ ]
for each agent pair (a_i, a_j) **do**
  `vec1` ← `all_answers`$[a_i]$
  `vec2` ← `all_answers`$[a_j]$
  `common_qs` ← shared answered questions
  `squared_sum` ← 0
  `max_squared_sum` ← 0
  for each question $q \in$ `common_qs` **do**
    if $q \notin$ `ref_dict` **then**
      continue
    end if
    `delta` ← number of options for q minus 1
    if `delta` $= 0$ **then**
      continue
    end if
    Accumulate:

$$\texttt{squared_sum} \mathrel{+}= (x_q^{(i)} - x_q^{(j)})^2$$

    Accumulate:

$$\texttt{max_squared_sum} \mathrel{+}= \texttt{delta}^2$$

  end for
  if `max_squared_sum` $= 0$ **then**
    continue
  end if
  Compute normalized distance:

$$d_{ij} = \frac{\sqrt{\texttt{squared_sum}}}{\sqrt{\texttt{max_squared_sum}}}$$

  Append d_{ij} to `pairwise_scores`
  Store d_{ij} in `pair_dict`
end for
Return:

$$\text{Diversity}_P(S) = \text{mean}(\texttt{pairwise_scores})$$

ported in Table 1. This further reinforces our main conclusion that current LLM-based multicultural agent systems remain substantially less value-diverse than the human reference.

D Additional Results on Diversity–Alignment Relationship

Figure 7 reports the relationship between Alignment and Diversity_S (Structural Diversity). Similar to the pairwise metric in the main paper, the correlation between the two remains weak, further supporting the conclusion that alignment does not reliably reflect whether a system preserves internal cultural plurality.

Compared with Diversity_P, Diversity_S produces a sharper separation between systems with genuine structural plurality and those whose diversity mainly arises from redundant pairwise disagreement. In particular, several highly aligned systems exhibit noticeably lower structural diversity, suggesting that their agents converge toward structurally homogeneous value configurations despite maintaining reasonable alignment with cultural references.

Algorithm 2 Structural Diversity Implementation

Require: Pairwise distance dictionary `pair_dict`
Require: Agent list `agents`
$N \leftarrow |\texttt{agents}|$
if $N < 2$ or `pair_dict` is empty **then**
  return 0
end if
Initialize symmetric distance matrix:

$$M \in \mathbb{R}^{N \times N}$$

for each agent pair (a_i, a_j) **do**
  Retrieve pairwise distance:

$$d_{ij} \leftarrow \texttt{pair_dict}(a_i, a_j)$$

  Fill:

$$M[i,j] \leftarrow d_{ij}$$
$$M[j,i] \leftarrow d_{ij}$$

end for
Compute minimum spanning tree:

$$T \leftarrow \text{MST}(M)$$

Compute:

$$\texttt{mst_span} = \frac{\sum_{(i,j)\in T} M[i,j]}{N-1}$$

Return:

$$\text{Diversity}_S(S) = \texttt{mst_span}$$

E Dynamic Interaction Details

E.1 Prompt

System Prompt Each agent is initialized with the same culturally grounded system prompt used in the main experiments:

```
You are a respondent from {country}.
Answer questions based on typical
cultural values in {culture} culture.
```

Social Exposure Prompt During dynamic interaction, before answering each WVS item, every agent is additionally shown the responses generated by the other agents in the previous interaction round. The user instruction is constructed as:

```
Question: {WVS question}
```

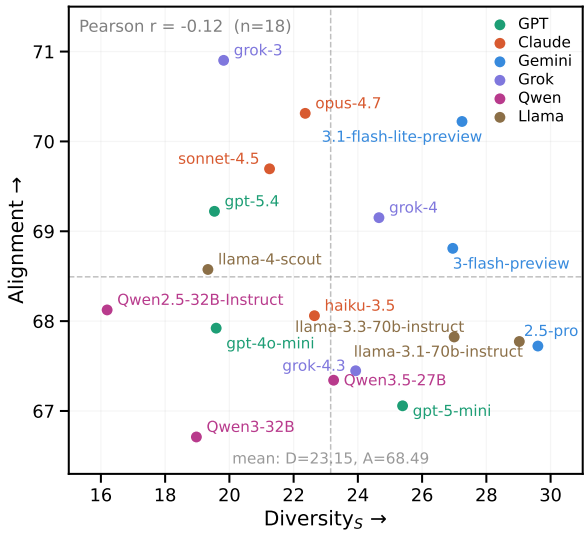

Figure 7: **Relationship between Structural Diversity and Alignment.** Each point represents one single-backbone multicultural agent system. Dashed lines indicate across-system means.

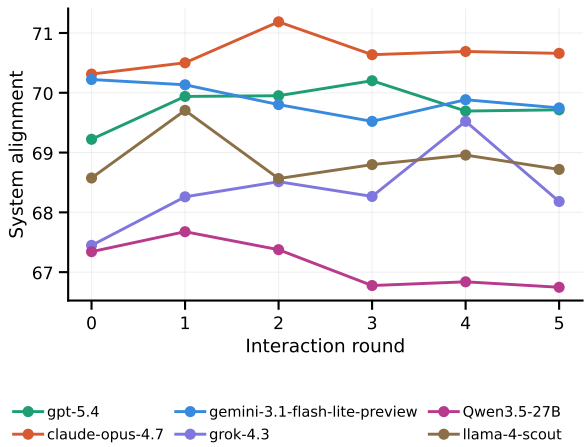

Figure 8: System alignment over five rounds of interaction for the six representative systems.

```
Here are answers from people of other
cultures:

{culture_1}: {response_1}

{culture_2}: {response_2}

...

You may consider these answers before
making your decision.

Output format MUST be exactly:

\boxed{number}
```

E.2 Multi-turn Alignment Dynamics

Figure 8 shows the corresponding alignment dynamics for the same six representative systems.

In contrast to the consistent diversity loss observed in the main paper, alignment exhibits no systematic trend across rounds. Most systems fluctuate within a narrow band around their round-0 value, and no system shows a sustained gain or loss comparable in magnitude to the diversity drop. This asymmetry reinforces our main finding: social exposure primarily reshapes the system's collective plurality rather than its per-agent cultural fidelity. In other words, homogenization under interaction is a system-level phenomenon that alignment-based evaluation alone would fail to detect.

F Collective Decision-Making Details

Project Descriptions Table 4 lists the 13 projects used in the participatory budgeting task. Each project is derived from a corresponding value dimension in the WVS dataset. The project names and descriptions were generated using GPT-5.5 to transform abstract value dimensions into concrete public-policy initiatives suitable for participatory budgeting scenarios.

Prompt The culturally grounded system prompt is constructed as:

```
You are a respondent from {country}.
Answer based on typical societal
priorities, cultural values, and public
policy preferences commonly associated
with {culture_name} culture.
```

The participatory budgeting instruction prompt is:

```
A national participatory budgeting
program is allocating limited public
funding across societal development
initiatives.

There are 13 candidate projects:

[Project list (ID, project name,
category, WVS dimension, and
description) omitted for brevity]

Voting instruction:

• Select EXACTLY 4 projects that
  should receive funding.
• Base your choices on long-term
  societal priorities and values.
• Consider cultural preferences,
  governance priorities, social
  development, and public well-being.

Output format MUST be exactly:

\boxed{id1, id2, id3, id4}

For example:

\boxed{2, 5, 8, 11}
```

ID	Project Name	Category	WVS Dimension	Description
01	Traditional Religious Heritage Festival	Culture	Religious Values	Funding community festivals and educational programs that preserve traditional religious customs, ceremonies, and local cultural heritage.
02	Smart Public Safety Camera Network	Security	Security	Expanding AI-assisted public surveillance cameras and emergency response systems to improve urban safety and crime prevention.
03	Transparent Public Spending Platform	Transparency	Corruption	Creating an open digital platform that allows citizens to track government spending, procurement, and public project budgets.
04	Industrial Growth and Job Creation Program	Economy	Economic Values	Supporting local manufacturing, industrial infrastructure, and employment growth through public investment programs.
05	Public AI and Technology Innovation Hub	Technology	Science & Technology	Establishing community AI labs, robotics education centers, and startup incubation programs for emerging technologies.
06	Community Civic Participation Centers	Civic Participation	Political Interest & Political Participation	Funding local civic centers that organize public discussions, citizen assemblies, and participatory decision-making activities.
07	National Institutions and Public Leadership Education	Institutions	Political Culture & Political Regimes	Supporting public education programs on national institutions, governance systems, civic responsibility, and political leadership.
08	Climate Sustainability and Green Energy Initiative	Environment	Postmaterialist Index	Investing in renewable energy, carbon reduction projects, urban sustainability programs, and environmental protection initiatives.
09	Neighborhood Community Cooperation Program	Community	Social Capital, Trust & Organizational Membership	Supporting local volunteer organizations, neighborhood associations, and public community-building activities.
10	Public Ethics and Integrity Education Campaign	Education	Ethical Values and Norms	Funding educational campaigns promoting honesty, social responsibility, anti-corruption awareness, and ethical public behavior.
11	Social Inclusion and Diversity Support Initiative	Social Inclusion	Social Values, Attitudes & Stereotypes	Supporting programs that promote social inclusion, reduce discrimination, and encourage understanding across different social groups.
12	Immigrant Integration and Language Support Centers	Migration	Migration	Providing language education, job assistance, and community integration services for immigrants and refugees.
13	Public Mental Health and Well-being Initiative	Health	Happiness and Well-being	Expanding public mental health support, community wellness activities, and accessible counseling services.

Table 4: Public projects used in the WVS-grounded participatory budgeting task. Each project is derived from one WVS value dimension.

Additional Results Additional results using `gpt-5.4` and `gemini-3.1-flash-lite-preview` are shown in Figure 9 and Figure 10, respectively. Both backbones exhibit trends consistent with the `claude-opus-4.7` results presented in the main paper.

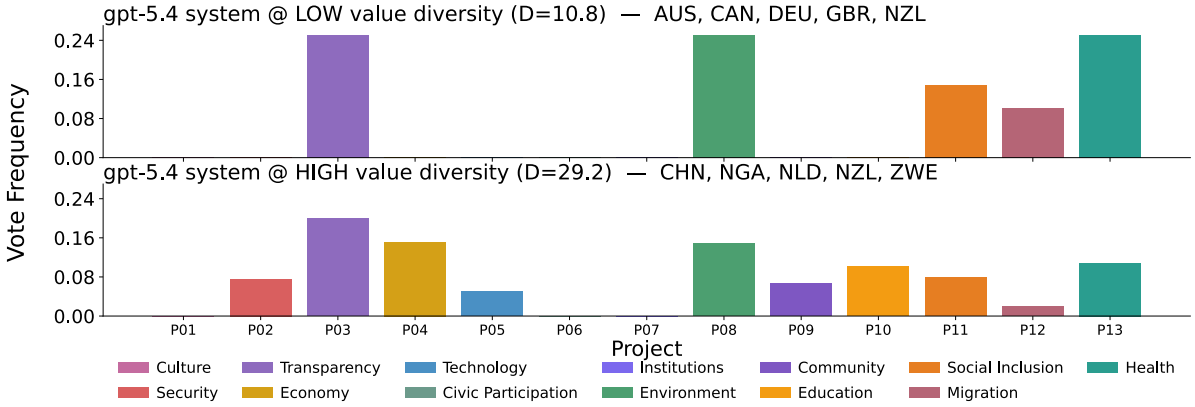

Figure 9: Collective decision-making outcomes in the Participatory Budgeting task. Both systems use `gpt-5.4` as the backbone model. The results show that the high-diversity system distributes support across substantially broader societal dimensions.

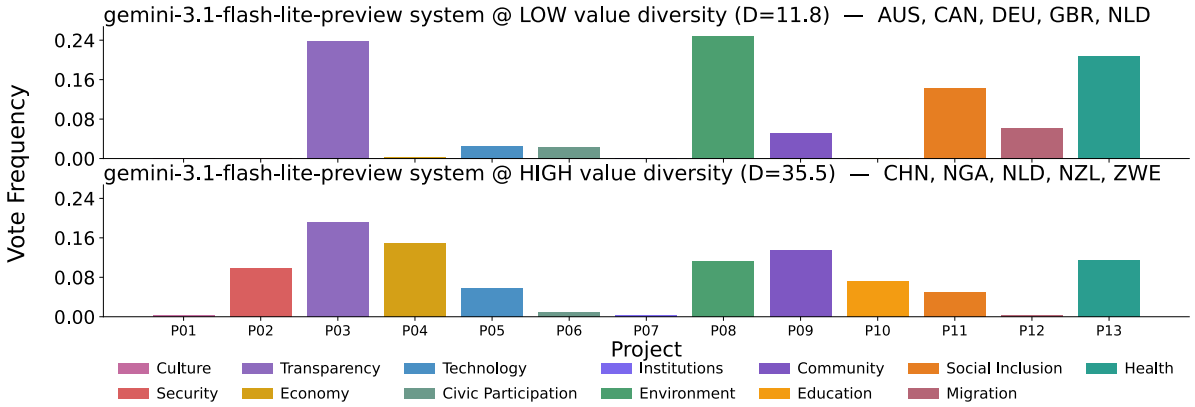

Figure 10: Collective decision-making outcomes in the Participatory Budgeting task. Both systems use `gemini-3.1-flash-lite-preview` as the backbone model. The results show that the high-diversity system distributes support across substantially broader societal dimensions.