

The Failure Happens Before the Drift: The Social Dynamics of Values in LLM Agent Societies

Farah Atif¹, Sougata Saha¹, Monojit Choudhury¹,

¹Mohamed Bin Zayed University of Artificial Intelligence
{farah.atif, sougata.saha, monojit.choudhury}@mbzuai.ac.ae

Abstract

Large Language Model (LLM)-based agents are increasingly used as proxies for human participants in social science research, yet it remains unclear whether they can faithfully simulate diverse and conflicting human value systems. We present a World Values Survey (WVS)-grounded simulation framework where culturally diverse agents with different communication styles engage in longitudinal, value-laden discussions. Across approximately 4,000 conversations involving 1,200 personas, 15 topics, and three models (GPT-4o, Gemini-2.5-Flash, and Gemma-4-E4B), we evaluate value faithfulness, value drift, and conversational realism. We find that more than 50% of personas fail to express their assigned WVS profiles from the outset, while 2-7% drift after repeated conversations. Ablations removing demographic details improve faithfulness for some models but do not change the broader trend: simulated value distributions still systematically deviate from the assigned WVS profiles. Compared to human discussions, simulated dialogues show a different trade-off between stylistic consistency and semantic diversity, often producing content-wise varied but stylistically repetitive exchanges. These findings suggest that current LLM agents can generate plausible conversations, but remain limited proxies for representing and preserving diverse human value profiles over time.

Introduction

There is a growing interest in using LLMs as stand-ins for human populations. Prior work has used LLMs to simulate individuals with specific demographic attributes or personality traits (Argyle et al. 2023; Tjuatja et al. 2024; Wang et al. 2025), as well as collective phenomena such as group dynamics and organizational behavior (Park et al. 2024; Zhou et al. 2024). In these settings, LLMs are increasingly positioned as scalable proxies for human judgment, opinion, and behavior (Choi, Young, and Ferrara 2026; Tjuatja et al. 2024; Park et al. 2024). However, real-world societies are shaped not only by demographic diversity but also by diverse and often conflicting value systems (Schwartz 1992). Therefore, the validity of such simulations depends on whether LLM agents can reliably instantiate and sustain culturally grounded value profiles during interaction.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

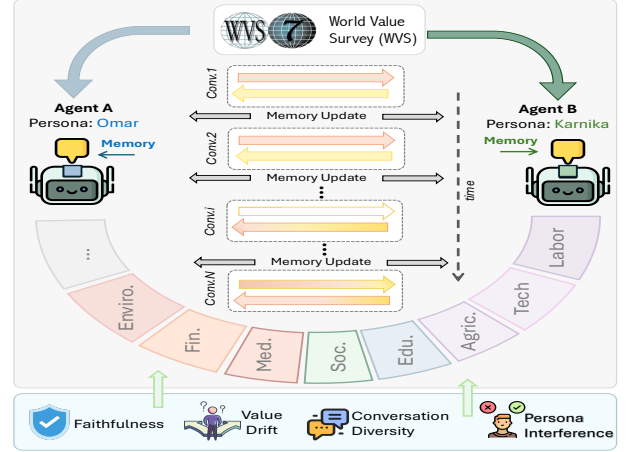


Figure 1: Overview of the proposed WVS-grounded multi-agent simulation framework, where culturally conditioned LLM personas engage in repeated value-driven discussions across diverse topics and are evaluated for value faithfulness, value drift, and conversational realism.

Most existing evaluations of LLM value alignment are static: models answer survey items, Likert-scale inventories, or value benchmarks, and their responses are compared against human distributions (Santurkar et al. 2023; Johnson et al. 2022; Dwivedi, Lavania, and Modi 2023). These studies show that LLMs often gravitate toward Western, educated, and liberal orientations, even under persona conditioning (Dwivedi, Lavania, and Modi 2023; Johnson et al. 2022). However, such pointwise evaluations measure values stated in isolation, as they do not test whether assigned values are expressed in conversation or sustained through repeated interaction, disagreement, and contestation.

In this work, we distinguish between two fundamental properties of value-driven LLM societies: **Instantiation** and **Retention**. *Instantiation* measures whether an agent expresses its assigned cultural value profile when first deployed. *Retention* measures whether, once instantiated, that value orientation remains recoverable across repeated multi-turn discussions. Existing benchmarks mostly evaluate instantiation through static responses, but do not capture retention under sustained interaction.

To study these questions, we introduce a large-scale sim-

ulation framework^{1 2} grounded in the WVS (Haerpfer et al. 2022b) (Figure 1). We sample personas from WVS Wave 7 across eight cultural regions and assign them positions along the two Inglehart-Welzel cultural dimensions (Inglehart and Welzel 2005): *Traditional–Secular-Rational* and *Survival–Self-Expression*. Agents then engage in repeated multi-turn discussions on 15 value-laden topics spanning labor, technology, agriculture, education, and social policy. After each interaction, agents update a structured memory that summarizes the discussion and its inferred effect on their value orientation, allowing conversational experiences to accumulate over time. We evaluate simulations across GPT-4o, Gemini-2.5-Flash, and Gemma-4-E4B for value faithfulness, value drift, and conversational quality.

Our analysis suggests that the primary failure in all LLM social simulations often occurs within the first two conversations: in nearly 50% of cases, LLM agents fail to exhibit their assigned WVS value profiles from the outset. Ablation studies further show that removing demographic details from the persona increases value faithfulness, while persuasive communication style is strongly correlated with unfaithfulness. Across all evaluated models, fewer than 7% of personas exhibit substantial value drift over their full conversational histories. However, the final value systems reflected in these simulations systematically deviate from real-world distributions, where moderate value profiles tend to collapse toward Traditional/Survival orientations, while several rare value combinations in the original WVS distribution become disproportionately amplified after repeated interactions.

Finally, our conversational analysis shows that models differ systematically in how they balance style, as reflected by dialogue acts (Stolcke et al. 2000), and content, as measured using embedding similarity (Mikolov et al. 2013). GPT-4o produces the most stylistically similar outputs relative to existing human dialogue corpora, but also shows high content similarity within topics. Gemma-4-E4B exhibits the greatest stylistic and content diversity, while Gemini-2.5-Flash generally falls between these two patterns, often overlapping with the human baseline in style while producing relatively varied content distributions.

These findings have important implications for the use of LLM agents in computational social science. If simulated personas do not reliably instantiate culturally grounded value systems, then downstream results from multi-agent simulations may reflect artifacts of model priors and alignment behavior rather than the intended population distributions. Our work, therefore, reframes the evaluation of LLM societies away from isolated survey responses and toward longitudinal behavioral consistency under interaction.

Simulation Design

Here we describe our end-to-end simulation pipeline in detail. An illustration of the same is presented in Figure 2.

¹Code: <https://github.com/mbzuai-nlp/wvs-value-simulation>

²Data: <https://huggingface.co/datasets/MBZUAI/WVS-Value-Simulation>

Participant Sampling and Persona Initialization

Each participant is defined by three broad aspects: their socio-demographic features, value system, and communication style. We derived the participant profiles from the WVS Wave 7 data (Haerpfer et al. 2022a), which spans ~90,000 participants from 66 countries. The WVS is particularly well-suited for this purpose because it is one of the most comprehensive and widely used cross-national surveys of human beliefs, values, and social attitudes (Haerpfer et al. 2022b; Inglehart and Welzel 2005). Its broad geographic coverage and standardized survey design make it possible to construct personas that vary not only across countries but also across demographic groups, social positions, and value orientations within countries. Moreover, because WVS Wave 7 contains each participant’s responses to ~290 questions pertaining to different aspects of life, such as social values, well-being, religion, economy, political interests, and demographics, it provides a richer basis for persona initialization than datasets that only contain coarse demographic attributes.

For our purposes, we preserved the following **socio-demographic features** for each individual: age, gender, country of residence, region (urban/rural), known languages, immigration status (immigrant/non-immigrant), household status, education profile of the individual and household, employment profile of the individual and household, economic status, religion, and ethnicity, as detailed in the Persona creation prompt: Part I (in Appendix). These features allow the simulations to condition conversations on concrete social and demographic contexts, rather than relying on abstract or stereotyped representations of cultural groups.

We filtered out participants who did not have values for these features, resulting in ~50,000 participants. We further filtered out participants who belonged to regions (defined by country and urban/rural combination) with <50 participants. This filtering was done to ensure sufficient diversity within each region, so that different types of conversations could be simulated without overfitting a region to a small number of observed individuals. From the remaining participants, we randomly sampled ~1,200 final participants for the simulation. The final participant set was limited to 1,200 due to cost-related constraints, since simulating and evaluating longitudinal conversations is costly.

We inferred participants’ **value systems** on a 3-point scale - high, medium, and low - along two bipolar dimensions: “Traditional–Secular-rational” and “Survival–Self-expression”. These dimensions were inferred based on participants’ answers to the 10 indicator questions used to create the Inglehart-Welzel Cultural Map (Inglehart and Welzel 2005). The map is widely used in comparative social science to summarize cross-national variation in values, while still allowing individual participants within the same country or cultural region to occupy different positions along the two dimensions. Thus, the inferred value systems provide a compact representation of participants’ normative and social orientations without reducing them solely to their country of residence. Furthermore, we used GPT-4o to create realistic and culturally appropriate synthetic names for each participant, and created a succinct 100-word bio comprising

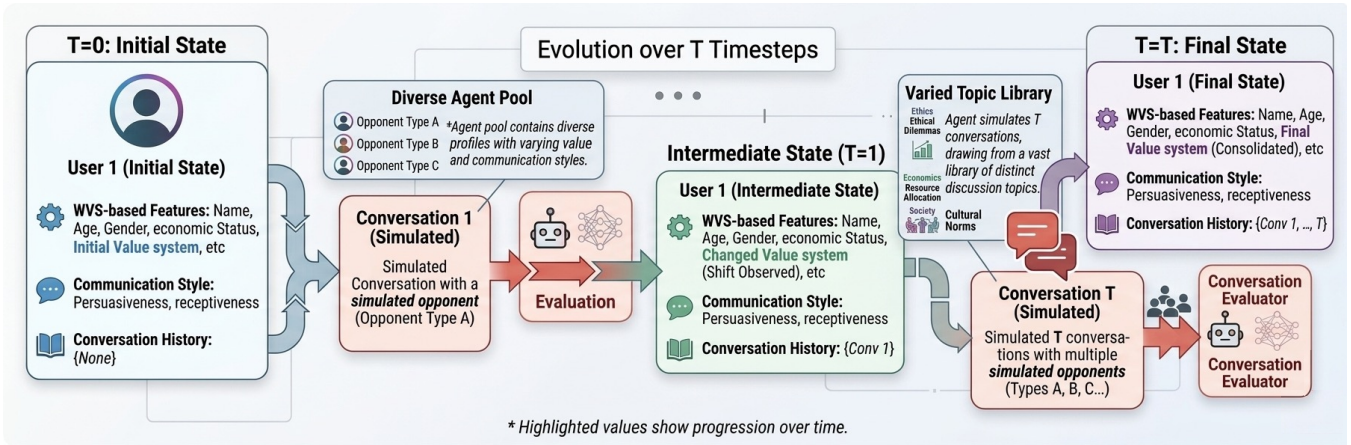


Figure 2: Overview of the WVS-grounded simulation pipeline, where diverse personas engage in repeated value-laden interactions, update their histories and value states, and are analyzed for value evolution and behavioral consistency.

their demographic features, as detailed in the Persona creation prompt: Part II (in Appendix).

In addition to socio-demographic features and value systems, we equipped each persona with a *communication style*. Communication style is pertinent for our simulations because longitudinal conversations are shaped not only by what participants believe, but also by how they express disagreement, evaluate new information, and respond to competing viewpoints. Prior work on persuasion and conversational receptiveness shows that individuals differ in their willingness to engage with opposing views, their tendency to evaluate arguments impartially, and their use of language to signal openness or resistance during disagreement (Minson and Chen 2022; Yeomans et al. 2020). Similarly, persuasion research suggests that the effectiveness and form of persuasive interaction depend on individual differences in how people present and process arguments (Cacioppo and Petty 1982; Petty and Cacioppo 2012). Therefore, modeling communication style allows our simulations to capture variation in conversational dynamics that would not be represented by demographic features or value systems alone.

We define communication style using two axes: *persuasiveness* and *receptiveness*. A high persuasiveness indicates a propensity to convince others and guide them toward one’s perspective, whereas a low persuasiveness is reflected by presenting one’s arguments more neutrally, without an evident attempt to sway others toward a specific viewpoint. A high receptiveness is reflected by being open to diverse ideas and perspectives, and by thoughtfully considering arguments without demanding immediate evidence or justification. In contrast, low receptiveness is reflected by critically evaluating arguments, requiring substantial evidence and logical consistency before considering alternative viewpoints. Together, persuasiveness and receptiveness define four regimes of communication: both high, both low, high persuasiveness with low receptiveness, and low persuasiveness with high receptiveness.

Importantly, we do not treat communication style as deterministically implied by a participant’s cultural background or value system. Although culture can shape norms of com-

munication, cross-cultural research also shows substantial individual-level variation within cultures, and some studies find that individual-level factors are stronger predictors of communication preferences than country-level cultural categories (Gudykunst et al. 1996; Park et al. 2012). This motivates treating communication style as an additional axis of persona variation rather than deriving it directly from cultural region or WVS-inferred values. Accordingly, each persona was randomly assigned one of the four communication styles. This random assignment helps avoid essentializing cultural groups while allowing the simulations to explore how similar socio-demographic and value profiles may produce different conversational trajectories under different interactional dispositions.

Overall, the $\sim 1,200$ participants span eight cultural regions: African-Islamic, Latin America, West & South Asia, Confucian, English-speaking, Orthodox Europe, Catholic Europe, and Protestant Europe. Their personas were defined by their WVS-based socio-demographic features, inferred value systems, and randomly assigned communication style.

Discussion Format and Topic

Unlike general chitchat, where value-laden exchanges may occur only sparsely, we simulate discussions around a fixed set of occupation-relevant topics designed to elicit disagreement, justification, and perspective-taking. This setup follows prior work on online deliberation and persuasion, where participants respond to explicit claims or prompts rather than engaging in open-ended conversation (Tan et al. 2016; Hidey et al. 2017). It also resembles structured debate platforms such as Kialo³, where discussions are organized around central claims and supporting or opposing arguments. Such a structure is important for our setting because it increases the likelihood that participants express values, trade-offs, and normative commitments, while still allowing them to respond naturally.

Participants were grouped into 5 occupation-based groups, determined from the WVS dataset: “Farm Work-

³<https://www.kialo.com/>

ers, Farm Owners, Semi-Skilled, and Unskilled Workers”, “Professional, Technical, and Skilled Workers”, “Service, Sales, and Clerical Roles”, “Students”, and “Unemployed, Retired, and Housewife Roles”. Each participant belonged to exactly one group, and the groups were approximately balanced in size. As shown in Table 1, we manually created three occupation-relevant discussion topics per group, yielding 15 topics in total. The topics reflect contemporary issues in global news and public discourse, and were designed to be broadly relevant within each occupational group regardless of country, religion, ethnicity, or value orientation. This occupation-conditioned design supports realistic engagement by grounding discussions in issues tied to participants’ work, education, household roles, or livelihoods.

Each topic consists of a main claim and five discussion aspects. The main claim provides a concrete position for participants to respond to, while the aspects constrain the space of possible responses to issues that are substantively relevant to the topic. This design follows findings from research on asynchronous online discussions, which suggests that structured prompts can reduce ambiguity, focus participation, and support more substantive discourse (Gao, Zhang, and Franklin 2013; Hew and Cheung 2012). Constraining the discussion in this way also helps mitigate common issues in simulated conversations, such as topical drift, generic responses, and repetitive content.

We framed each interaction as a private discussion between two members of a social-media group related to their occupational background. One member had posted a claim about the topic, and the other participant responded. This framing reflects common forms of online discussion, where people engage in asynchronous, text-based exchanges within interest-, identity-, or occupation-based groups (Boyd and Ellison 2007; Walther 1996). It also provides a realistic setting for cross-cultural interaction: participants share an occupational context, but may differ in socio-demographic background, cultural region, and value system.

At each turn, participants were instructed to first infer their interlocutor’s value system from the previous responses. They were then asked to reflect on whether the interlocutor’s stance and underlying values aligned with their own, or whether they wanted the interlocutor to move toward their own stance and values. Based on this reflection, participants formulated their next response while continuing the discussion along one or more of the topic-specific aspects. This design encourages responses that are both conversationally grounded and value-aware, rather than isolated statements of opinion. It also aligns with prior work showing that persuasion and opinion change in online discussions depend not only on the content of arguments, but also on interactional dynamics, language choices, and the relationship between speakers’ prior positions (Tan et al. 2016; Durmus, Ladhak, and Cardie 2019).

Participants were randomly paired across cultural backgrounds to ensure cross-cultural interaction. Each pair was also randomly assigned a **familiarity label**, with a 0.3 probability of familiarity, indicating whether the speakers were familiar with each other. We include familiarity because prior work suggests that social con-

text and perceived relationship between interlocutors can shape politeness, disagreement, disclosure, and engagement in computer-mediated communication (Walther 1996; Danescu-Niculescu-Mizil et al. 2013). Thus, the familiarity label introduces a realistic source of conversational variation without deterministically tying it to participants’ demographics or values.

The ‘Dialogue Prompt: Setup’ and ‘Dialogue Prompt: Tasks’ in appendix detail the prompts used to simulate each turn asynchronously from each participant in a dialogue. The number of turns in a conversation was sampled from a normal distribution with $\mu=15$ and $\sigma=5$. We use this stochastic length to approximate medium-length online discussions with sustained back-and-forth exchange, while avoiding a fixed conversation length that would make all simulations structurally identical. The stopping point was not revealed to participants. Once the sampled conversation length was reached, the discussion was forcefully concluded.

#	Occupation	Topic
1	Farm Workers, Owners, Semi-, & Unskilled Workers	Water Usage Limits for Farmers
2		Climate-Resilient Farming Practices
3		Farmers’ Access to Fair Pricing
4	Professional, Technical, & Skilled Workers	Local Ethical Standards for Global Tech
5		Remote Work Policy Changes
6		AI-Driven Recruitment Systems
7	Service, Sales, & Clerical Roles	Attendance Policies During Peak Seasons
8		Impact of Automation in Retail
9		Employee Unionization in Retail
10	Students	AI-Driven Education and Assessment
11		Cost of Higher Education
12		Mental Health Support in Universities
13	Unemployed, Retired, & Housewife Roles	Unemployment Benefits Program
14		Financial Independence for Homemakers
15		Retirement Age Adjustments

Table 1: Occupation groups and their discussion topics.

Dialogue Summarization and Value Tracking

Our simulation is longitudinal, where each persona participates in multiple dialogues with distinct interlocutors over time. To preserve continuity across these interactions, we maintain an independent memory for each participant. After every dialogue, we persist the participant’s own turns, their reflection on their value system, and whether they intend to revise their values in future interactions. All subsequent conversations for that participant are conditioned on this memory, which is updated after each dialogue.

Storing such longitudinal interaction history is necessary for realistic simulations of social and value change. Prior work on generative agents shows that persistent memory and reflection are central to producing believable long-horizon behavior, since agents must be able to remember past experiences, synthesize them into higher-level reflections, and use them to guide future actions (Park et al. 2023). Similarly, work on reflective language agents shows that textual reflections stored in episodic memory can improve later decision-making without modifying model weights (Shinn et al. 2023). In our setting, this motivates treating each dialogue not as an isolated exchange, but as an experience that may incrementally shape the participant’s later stance, reasoning, and conversational behavior.

However, naively persisting all prior turns word-for-word is impractical and undesirable, because: (i) longitudinal simulations produce a continually growing history that can exceed the model’s context window. Prior work on LLM memory systems notes that limited context windows constrain extended conversations and motivate external or compressed memory mechanisms (Packer et al. 2024); (ii) raw transcript replay can introduce irrelevant details, redundancy, and topic drift into later prompts. We therefore summarize each completed dialogue into a structured memory record rather than storing the full transcript verbatim.

The structured summary contains: (i) the topic and main claim, (ii) the overall discussion trajectory, (iii) the agent’s individual decisions, (iv) the detected stance change, and (v) the estimated short-term and long-term effects on both WVS dimensions. The *Dialogue Summarizer Prompt* (in Appendix) details the prompt used to extract this information after each dialogue for every participant. This structured format serves two purposes. First, it compresses the dialogue into a form that can be included in future prompts under context-length constraints. Second, it preserves the aspects of the interaction that are most relevant to our research question: how repeated value-laden conversations may influence an agent’s future beliefs and argumentative behavior.

The summary is appended to the agent’s prior discussion history and included in subsequent prompts. Later conversations are therefore conditioned on earlier simulated experiences, enabling value change to accumulate incrementally rather than arising as a one-shot reaction to a single exchange. This design also makes the longitudinal process more interpretable: by storing both stance changes and estimated effects on the two WVS dimensions, we can trace how each dialogue contributes to the participant’s evolving value profile over time.

Models

We conduct simulations with three models, out of which two were API-based: GPT-4o⁴ and Gemini-2.5-Flash⁵, and one was a smaller open-weights model: Gemma-4-E4B-it⁶. All models are prompted identically using the same persona and discussion templates. Table 2 summarizes key statistics. We incurred a total of ~\$450 for the end-to-end simulation using GPT-4o, and ~\$100 for Gemini-2.5-Flash. The open-weights Gemma-4-E4B-it model was run using one NVIDIA RTX 6000 GPU.

Metric	Value
Number of agents	1,238
Unique conversations	3,895
Number of topics	15
Avg. conversations per user	6.16
Avg. turns per conversation	9.59

Table 2: Statistics of the WVS-grounded simulation dataset

⁴<https://openai.com/index/hello-gpt-4o/>

⁵<https://deepmind.google/technologies/gemini/flash/>

⁶<https://ai.google.dev/gemma>

Evaluation Methodology & Analysis

We evaluate the simulations along three dimensions: (i) **Value Faithfulness**: Are agents faithful to their assigned value systems? This measures how accurately models instantiate the intended cultural value profiles during interaction. (ii) **Value Drift**: Do agents’ values drift over time? Here, time refers to repeated interactions across multiple conversations, allowing us to measure whether agents’ value orientations change over their conversational histories. (iii) **Conversational Realism**: How realistic are the generated conversations relative to human conversations? This evaluates the stylistic and semantic diversity of generated discussions compared to human dialogue. To automate the evaluation, we used two independent LLM evaluators, Gemini-3-Flash and GPT-5.2, alongside human annotators. We found that human evaluators showed higher agreement with Gemini-3-Flash than with GPT-5.2. Consequently, we selected Gemini-3-Flash as the main evaluator. Further details are provided in the Evaluator Assessment in the Appendix.

Because value attribution from dialogue is inherently subjective, we interpret the reported percentages as approximate estimates and focus our conclusions on qualitative and relative patterns.

Value Faithfulness

Value faithfulness measures whether a simulated persona instantiates its assigned WVS value profile at the beginning of interaction. We evaluate this using the agent’s first two conversations, where accumulated memory is minimal and externally induced drift is least likely to have compounded. After each conversation, an LLM judge (Gemini-3-Flash) assessed the agent’s expressed value system along each WVS dimension, using the Evaluation Prompt (in Appendix). The judge uses a five-point ordinal scale: *Strong Traditional (StrongTrad)/Leaning Traditional (LeanTrad)/Balanced/Leaning Secular (LeanSec)/Strong Secular (StrongSec)* for the Traditional–Secular–Rational axis, and *Strong Survival (StrongSurv)/Leaning Survival (LeanSurv)/Balanced/Leaning Self-Expression (LeanSelf)/Strong Self-Expression (StrongSelf)* for the Survival–Self-Expression axis.

We define two evaluation schemes. In the **Lenient** scheme, *Strong*, *Lean*, and *Balanced* labels are treated as compatible with the assigned value pole. For example, a persona assigned a Traditional orientation is considered faithful if the judge returns *StrongTrad*, *LeanTrad*, or *Balanced*. In the **Strict** scheme, only *Lean* and *Strong* labels consistent with the assigned orientation are accepted as faithful. Here, *Balanced* is treated as a deviation because it indicates that the agent no longer expresses a clear directional commitment. Thus, the strict scheme measures unambiguous expression of the assigned value pole, while the lenient scheme captures broader compatibility with it.

Are agents faithful to their initial value system? Our experiments show that simulated agents often fail to remain faithful to their assigned value systems. As shown in Table 3, across all evaluated models, fewer than 50% of per-

sonas are faithful under the strict criterion, and even the lenient criterion shows substantial failure rates. Gemma-4-E4B exhibits the lowest faithfulness rates in both settings. GPT-4o and Gemini-2.5-Flash show comparable faithfulness rates under the strict scheme, but differ under the lenient scheme, where Gemini-2.5-Flash has a lower faithfulness rate (47.5%). These findings suggest that persona conditioning based on demographic and cultural information is likely insufficient to override latent model priors acquired during pretraining and alignment. While the reported percentages are based on automatic value attribution, we treat them as approximate estimates. Our main finding is the consistent qualitative asymmetry between initial value instantiation failures and subsequent value retention.

	Strict	Lenient
GPT-4o	44.2%	52.5%
Gemini-2.5-Flash	44.6%	47.5%
Gemma-4-E4B	32.1%	37.1%

Table 3: Strict and lenient faithfulness across models

What drives value unfaithfulness? We fitted a binary logistic regression model using Statsmodels⁷ to identify factors associated with value faithfulness. The dependent variable was faithfulness, and the independent variables included demographic features (age, gender, country, region, education profile, employment status, economic status, immigration status, occupation, household status, ethnicity, and religion), the initial value system along both WVS dimensions, and communication style. We fit a separate model for each of the three evaluated LLMs. For each original feature, we use likelihood-ratio tests to assess whole-feature significance, while coefficient-level odds ratios (*OR*) indicate the direction and magnitude of specific category effects. The models were fitted on the full set of 1,200 agents. We did not use a train/test split because the goal was statistical inference rather than prediction. Since the outcome classes were approximately balanced (Table 3), classification accuracy provides a coarse diagnostic of model fit. The three regressions achieved comparable accuracies of 72.1%, 74.2%, and 72.1%, respectively.

Consistent predictors across all three models: Three features were significant in all models: initial survival-self-expression dimension, initial traditional-secular-rational dimension, and communication style. Across all models, a persuasive communication style was the strongest and most consistent predictor of unfaithfulness (Gemini: *OR* = 3.29, $p < 0.001$; Gemma: *OR* = 3.26, $p < 0.001$; GPT-4o: *OR* = 5.04, $p < 0.001$). Conversely, initial value orientations on both WVS dimensions consistently predicted faithfulness: personas assigned a survival orientation (Gemini: *OR* = 0.035; Gemma: *OR* = 0.045; GPT-4o: *OR* = 0.144, all $p < 0.0014$), a secular-rational orientation (Gemini: *OR* = 0.0734; Gemma: *OR* = 0.154; GPT-4o: *OR* = 0.212, all $p < 0.001$), a traditional orientation (Gemini: *OR* = 0.419; Gemma: *OR* = 0.467; GPT-4o: *OR* = 0.511,

all $p < 0.001$), and a self-expression orientation (Gemini: *OR* = 0.323; Gemma: *OR* = 0.403; GPT-4o: *OR* = 0.554, all $p < 0.001$) were all significantly more likely to remain faithful relative to their baselines. In particular, the protective effect of the orientation of initial values was strongest in Gemini and Gemma, and somewhat weaker in GPT-4o, suggesting that GPT-4o assigns less weight to the cultural profiles assigned when generating responses.

Model-specific predictors: Beyond the shared predictors, each model exhibited unique associations. Under Gemini-2.5-Flash, economic status emerged as a significant feature (LR $p < 0.001$), with lower-class (*OR* = 2.67, $p < 0.001$) and working-class personas (*OR* = 1.72, $p = 0.019$) more likely to be unfaithful, and a both-low (low persuasive and low receptive) communication style also significantly associated with unfaithfulness (*OR* = 2.20, $p < 0.001$). Under Gemma-4-E4B, gender was a significant predictor (LR $p = 0.015$), with male personas more likely to be unfaithful (*OR* = 1.42, $p = 0.013$); additionally, a receptive communication style was explicitly associated with faithfulness (*OR* = 0.633, $p = 0.015$). Under GPT-4o, cultural region reached significance (LR $p = 0.041$), with Latin American personas more likely to remain faithful (*OR* = 0.600, $p = 0.018$), and a both-low communication style also significantly predicted unfaithfulness (*OR* = 2.13, $p < 0.001$).

We conjecture that persuasive communication prompts likely activate generic argumentative priors learned during alignment training, encouraging models to optimize for rhetorical effectiveness and social acceptability rather than strict adherence to assigned cultural value profiles. This suggests that conversational objectives and persona conditioning may compete during generation, with persuasive framing weakening recoverability of the intended value orientation.

Value Drift

Value drift measures the shift between a persona’s value position in its first two conversations and its value position over the complete conversation history. We use the same five-point ordinal scale and Lenient/Strict evaluation schemes as in the faithfulness analysis, allowing change scores to be expressed as directional shifts toward or away from the initially assigned value pole. A persona is considered to have undergone value drift if its estimated overall position differs from its initial position (as reflected in the first two conversations), under either of the chosen evaluation schemes.

	Strict	Lenient
GPT-4o	1.9%	0.8%
Gemini-2.5-flash	3.9%	1%
Gemma-4-E4B	6.4%	2.4%

Table 4: Strict and lenient value drift across models

Do values drift in simulated conversations? Table 4 reports the percentage of personas exhibiting value drift for each simulated model. Across all models, drift rates remain low, particularly under the lenient criterion. GPT-4o shows the highest stability, with only 1.9% strict and 0.8% lenient value drift. Gemini-2.5-Flash exhibits slightly higher

⁷<https://www.statsmodels.org/stable/index.html>

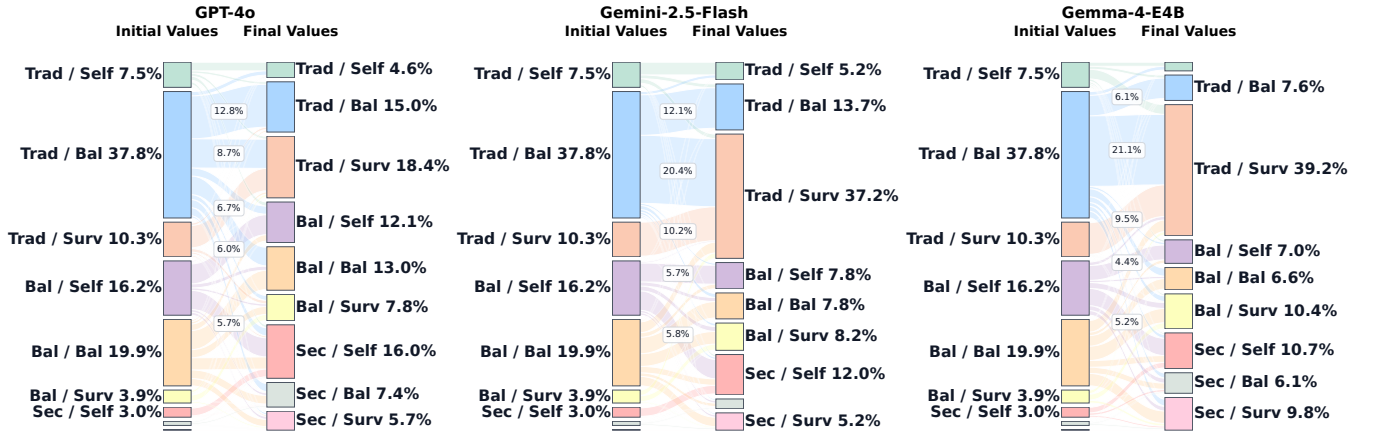


Figure 3: Value trajectories from initial WVS profile to final simulated value position across models

variation, but still maintains low drift overall. In contrast, Gemma-4-E4B shows the highest drift under both strict (6.4%) and lenient (2.4%) settings, indicating comparatively lower robustness in preserving value orientations over time. These results align with the stability of human values observed in real-world longitudinal research (Bardi et al. 2009; Vecchione et al. 2026). Nonetheless, we do not claim that the simulated conversations correspond to the real-world temporal dynamics through which human values evolve. Rather, the simulations provide a controlled setting for evaluating whether models preserve value orientations across repeated interactions.

Where do values land in a simulated society? For each model, Figure 3 shows the movement from the assigned initial value profile (on the left) to the final evaluated value profile (on the right), where the nodes represent combined positions of the two WVS cultural axes: Traditional–Secular–Rational and Survival–Self-Expression. The flow widths are proportional to the share of simulated users following that transition. Percent labels on the larger flows indicate the percentage of users in that model moving along that path. Each panel reports the total number of evaluated users and the share whose final value position differs from their initial profile.

Figure 3 reveals a consistent pattern across models, where the largest transition occurs from Traditional/Balanced to Traditional/Survival. This flow accounts for 12.8%, 20.5%, and 21.2% of personas in GPT-4o, Gemini-2.5-Flash, and Gemma-4-E4B, respectively. This suggests that personas initialized as balanced on the Survival–Self-Expression dimension are often evaluated as more Survival-oriented after repeated conversations. Correspondingly, the Traditional/Survival cell grows from 10.3% at initialization to 18.4% for GPT-4o, 37.2% for Gemini-2.5-Flash, and 39.2% for Gemma-4-E4B. The direction of this shift is consistent across models, differing mainly in magnitude, which suggests a systematic tendency in conversational LLM simulations to favor Survival-oriented framings on value-laden topics, while the Traditional–Secular–Rational axis is comparatively more stable.

The Balanced/Balanced cell, representing personas with moderate positions on both WVS dimensions, contracts in every model: from 19.8% at initialization to 13.0% in GPT-4o, 7.7% in Gemini-2.5-Flash, and 6.6% in Gemma-4-E4B. This indicates that models struggle to preserve moderate value profiles, with many such personas moving toward Traditional/Survival. For social simulation, this is a notable fidelity failure, since the real WVS distribution contains substantial mass in the moderate region.

Interestingly, less frequent value combinations in the original WVS-derived samples, such as Balanced/Survival, Secular/Self-Expression, Secular/Balanced, and Secular/–Survival, become more prevalent in the final evaluated profiles across all models. Thus, the simulated value space is not merely compressed toward a single dominant region. Rather, it is also reshaped in ways that amplify several initially rare value combinations. Together, these results show that repeated LLM-based interactions can systematically distort the intended population-level value distribution, even when individual-level drift rates appear low.

Ablation: The Effect of Demographic Conditioning

We further perform an ablation study in which we remove all demographic details from each user’s persona and simulate conversations using only the assigned initial value system. This allows us to examine whether the observed patterns in value faithfulness and drift are driven by demographic persona conditioning, or whether they reflect model-level biases that persist even when demographic information is absent. The ablation is conducted on a subset of 82 users, sampled to uniformly cover all combinations of the WVS dimensions.

Model	Faithfulness		Value Drift	
	Strict	Lenient	Strict	Lenient
GPT-4o	71%	72%	3.7%	1.2%
Gemini	68.3%	68.3%	4.9%	4.9%
Gemma-4b	39%	42.7%	8.5%	3.7%

Table 5: Ablation results: Faithfulness and value drift

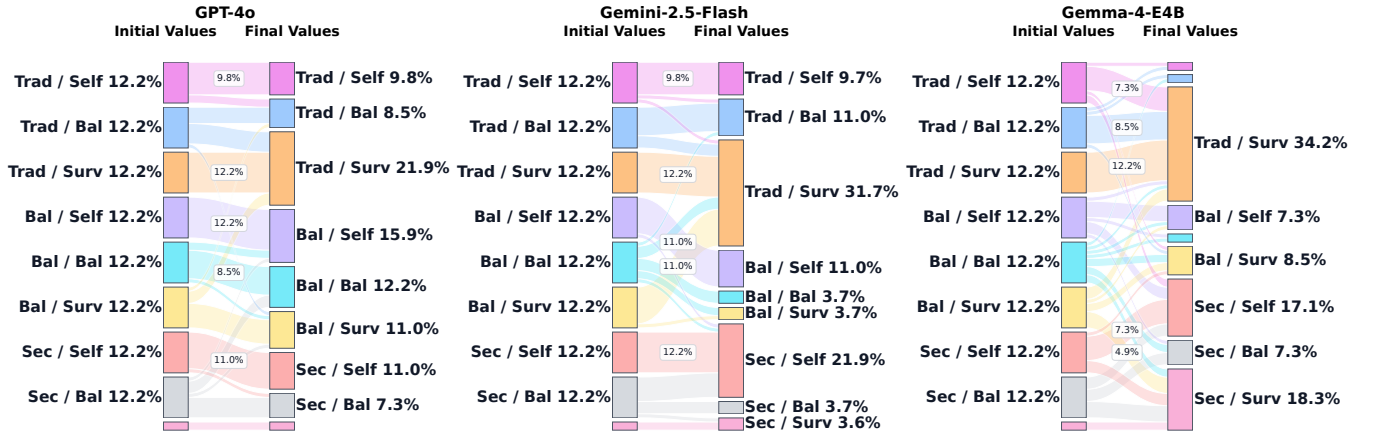


Figure 4: Value trajectories from initial WVS profile to final simulated value position without persona grounding

Value Faithfulness: As detailed in Table 5, removing demographic details substantially improves value faithfulness for GPT-4o and Gemini-2.5-Flash. GPT-4o’s strict faithfulness increases from 44.2% to 71.0% (+26.8), while Gemini-2.5-Flash increases from 44.6% to 68.3% (+23.7). For both models, the gap between the strict and lenient schemes also narrows, suggesting that their expressed values become more directionally consistent when only the value system is provided. Gemma-4-E4B improves more modestly, from 32.1% to 39.0% strict faithfulness (+6.9), and remains below 50%. Overall, these results suggest that demographic conditioning can interfere with value instantiation, particularly for GPT-4o and Gemini-2.5-Flash, which appear better able to adhere to the assigned value system when demographic factors are excluded.

Value Drift: As detailed in Table 5, removing demographic factors slightly increases strict value drift for all models: GPT-4o rises from 1.9% to 3.7%, Gemini-2.5-Flash from 3.9% to 4.9%, and Gemma-4-E4B from 6.4% to 8.5%. Lenient drift rates remain low and broadly comparable to the full-persona condition, except for Gemini-2.5-Flash, where the strict and lenient rates both become 4.9%. These results indicate that value drift remains limited even without demographic persona conditioning. Thus, the main effect of removing demographic information is not to eliminate longitudinal drift, but to improve initial value instantiation.

Value Landing: Figure 4 shows the initial values distribution and the final evaluated values when simulating without demographic factors (persona). The dominant flow from Traditional/Balanced towards Traditional/Survival remains persistent across all models, similarly to when simulating with the demographic factors. The Balanced/Balanced collapse remains with Gemini and Gemma models, unlike GPT-4o, where the Balanced/Balanced distribution remains the same. The experiment also reveals that rare dimensions amplification significantly diminishes with GPT-4o and Gemini-2.5-Flash, only Gemma still strongly amplifies the Secular cells (Sec/Self: 12.2 to 17.1%, Sec/Surv: 2.4 to 18.3%). The overall observation is that GPT-4o is the

most stable, and most starting cells largely stay put. Without demographic factors, GPT-4o behaves close to an identity function on the value space.

Conversational Realism

To assess the quality of simulated conversations, we evaluate two complementary aspects: **stylistic variation** and **content variability**. Stylistic variation captures whether conversations differ in how participants communicate, while content variability captures whether they differ in what semantic information they convey. For stylistic variation, we perform dialogue act tagging using the taxonomy of Saha et al. (2021), which comprises 11 tags, with Gemini-3-Flash as the tagger. For content variability, we compute semantic representations using 384-dimensional SentenceBERT embeddings (Reimers and Gurevych 2019). We measure these properties along three axes:

- **Individual style and content** measures inter-user similarity in terms of their conversational style and semantic content. For each user, we assign a dialogue act to each utterance across all conversations and aggregate these labels into a dialogue-act frequency vector representing the user’s conversational style. We then compute pairwise cosine similarity across user pairs to quantify stylistic convergence. For content, we embed each utterance using SentenceBERT and compute a per-user mean embedding to represent the user’s overall semantic profile.
- **Conversation-level diversity** measures how similar conversations are in terms of style and content. We aggregate dialogue-act and embedding-based representations at the conversation level, and compute pairwise distances between conversations to assess whether they are stylistically and semantically distinct.
- **Within-topic conversation diversity** measures whether conversations remain varied when the subject matter is held constant. For each topic, we compare style and content representations across different conversations on the same claim.

As a human baseline, we use the OUMdials dataset (Farg

et al. 2022), which contains human conversations on similarly open-ended and socially relevant topics. We apply the same dialogue-act and embedding-based analyses to OUMdials and compare the resulting style and content distributions with those from our simulated conversations.

Individual-Level Diversity Figure 5 presents box plots that capture individual-level similarity across our generated datasets, compared against the human-curated OUMdials baseline. For **style**, GPT-4o exhibits the highest cross-user similarity, with a mean similarity of 0.86, exceeding the OUMdials baseline. This suggests that GPT-4o tends to produce stylistically homogeneous conversations across personas. Gemini-2.5-Flash overlaps closely with the OUMdials distribution, indicating a more human-like degree of stylistic variation. Gemma-4-E4B shows the greatest stylistic diversity, with a lower mean similarity of 0.63 and a broader distribution across user pairs.

For **content**, similarity scores are substantially lower than style scores across all generated datasets, with means below 0.6. This indicates that even when simulated users communicate in similar styles, they often differ in the semantic content they produce. In contrast, OUMdials shows the highest content similarity, with a mean of 0.73 and the tightest distribution. This suggests that human participants tend to discuss more similar content while expressing it through more varied styles, whereas LLM agents often display the opposite pattern: stronger stylistic convergence but more dispersed semantic content.

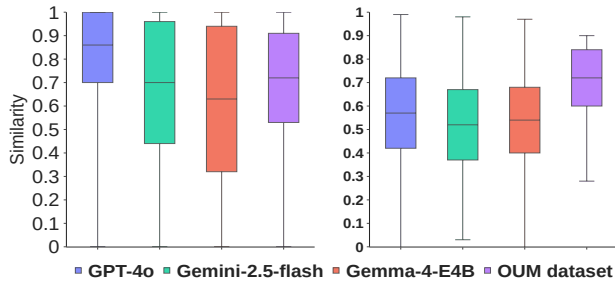


Figure 5: Individuals' style and content similarity

Conversation-Level Diversity Figure 6 compares conversation-level similarity in the generated datasets against the OUMdials baseline. For **style**, all models substantially overlap with OUMdials, suggesting that the overall distribution of dialogue acts is broadly comparable to human conversations at the exchange level. Gemma-4-E4B shows the greatest stylistic diversity, reflected by its lower median similarity and wider spread. For **content**, similarity scores are lower than stylistic scores across both generated and human datasets, with means below 0.5. This indicates high thematic variation across conversations. Among the simulated models, Gemini-2.5-Flash and Gemma-4-E4B show the most diverse content distributions, with the lowest mean similarity scores.

Within-Topic Conversation Diversity Figure 7 shows conversation similarity when conversations are grouped by

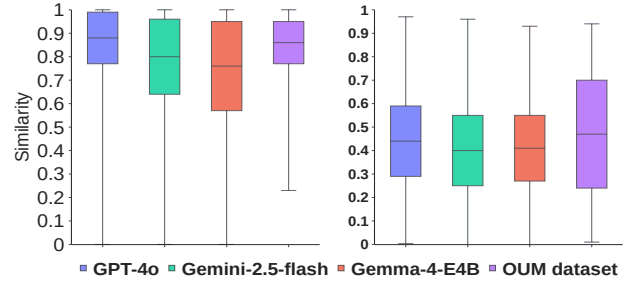


Figure 6: Conversations diversity by style and content

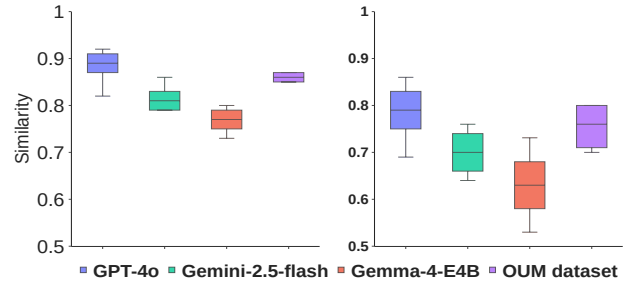


Figure 7: Within-topic style and content diversity

topic. For **style**, Gemma-4-E4B exhibits the greatest diversity, with the lowest mean similarity and a broader distribution than the OUMdials baseline. In contrast, GPT-4o maintains a highly uniform dialogue-act profile across conversations on the same topic, indicating stronger stylistic convergence. For **content**, Gemma-4-E4B again shows the greatest variation, with a mean similarity of 0.63 and a substantially larger spread than the other sources. This suggests that Gemma-4-E4B explores a wider range of semantic content even when the underlying claim is held fixed. By contrast, both GPT-4o and OUMdials show higher within-topic content similarity (> 0.75), indicating more focused discussions within thematic clusters. We provide additional topic-level analysis in the Appendix.

Related Work

LLM Value Evaluation

Research on value alignment in LLMs has increased in recent years, with systematic evaluations emerging across ethical frameworks, demographic variation, and cross-cultural differences (Cao et al. 2023; Tao et al. 2024; Santurkar et al. 2023). Most existing approaches rely on established social-science instruments such as Schwartz's value taxonomy (Schwartz, Lehmann, and Roccas 1999), the WVS (Inglehart and Welzel 2005), or Hofstede's cultural dimensions (Hofstede 2001), and evaluate models through single-turn elicitation using Likert-scale questions or value inventories. Across these studies, a recurring finding is that LLMs tend to lean toward Western, educated, liberal orientations regardless of persona conditioning (Santurkar et al. 2023; Johnson et al. 2022; Dwivedi, Lavania, and Modi 2023). However, these evaluations remain fundamentally static. They measure what a model states when directly prompted, but not

whether assigned value systems are behaviorally instantiated in interaction. Recent work has begun to expose this limitation. Shen, Clark, and Mitra (2025) shows that LLMs often exhibit a value-action gap, where stated values diverge from the choices models make in contextualized scenarios. Similarly, benchmarks such as ValueBench (Ren et al. 2024) attempt to move beyond isolated survey responses, yet still remain confined to single-turn settings. Existing evaluations provide limited visibility into whether assigned value profiles persist through sustained, multi-turn interaction, especially under disagreement and cross-cultural contestation. We address this gap by testing whether culturally grounded value profiles are faithfully instantiated and remain stable across repeated social interactions.

Multi-Agent LLM Systems

Multi-agent LLM simulations are increasingly used to model social behavior, including emergent group dynamics, opinion formation, and collective decision-making shaped by heterogeneous value systems (Park et al. 2023; Chuang 2024; Osman and d’Inverno 2023). Prior work shows that LLM societies can exhibit coordination, relationship formation, and information diffusion without explicit scripting (Park et al. 2024), while value diversity can improve emergent intelligence and network integration (Osman and d’Inverno 2023). However, most evaluations focus on task-level outcomes such as reasoning, planning, coordination, or social intelligence benchmarks (Zhou et al. 2024), rather than whether agents remain faithful to their assigned cultural identities during repeated interaction. Recent work such as ValueFlow (Liu, Liu, and Shen 2026) studies how value perturbations propagate through interacting LLM societies, but still assumes that agents initially instantiate stable value orientations. Our work examines this assumption directly by evaluating whether WVS-grounded personas both faithfully instantiate and preserve assigned value systems throughout longitudinal interaction. Moreover, our simulation relies on real personas drawn from the WVS data rather than synthetic personas, offering a closer approximation to the heterogeneity and internal consistency of real human value profiles than synthetically generated personas can provide.

Value Drift

Psychology suggests that human values are generally stable over time (Lane 1976). Prior studies show that core value systems remain relatively persistent across adulthood and even adolescence, despite moderate shifts in specific priorities (Milfont, Milojev, and Sibley 2016; Vecchione et al. 2016, 2026). At the same time, values are also context-sensitive and can change through social interaction. Psychological research identifies both unconscious and deliberate mechanisms of value change, including priming, reflection, and consistency maintenance (Bardi and Goodwin 2011). Other work shows that values may be reconstructed differently when applied to concrete situations rather than abstract principles (Maio and Olson 1998; Seligman and Katz 2013). These dynamics have also been explored in agent-based simulations studying moral change and collective policy behavior (de Wildt and van de Poel 2024; Dehkordi et al. 2024).

Discussion and Conclusion

While LLM-based simulations increasingly model individuals and populations, our results show that conversational plausibility can mask failures of representational fidelity. They may therefore reproduce long-standing external validity problems, including overgeneralization from WEIRD populations (Henrich, Heine, and Norenzayan 2010). These risks might get amplified in synthetic populations, digital twins, and policy simulations, where unfaithful value profiles can distort estimates of public opinion, intervention effects, or group behavior. Such distortions may especially affect marginalized or culturally underrepresented groups and contribute to algorithmic monoculture, where downstream studies inherit the same model-specific biases (Kleinberg and Raghavan 2021; Saha and Choudhury 2025; Pandey, Saha, and Choudhury 2025; Saha, Pandey, and Choudhury 2025; Saha et al. 2025). Also, since richer demographic prompting is not a sufficient safeguard, LLM-based simulations should be treated as virtual respondents, but as model-mediated instruments requiring validation against the populations they claim to represent (Bisbee et al. 2024; Ziemis et al. 2024; Madden 2025).

In conclusion, we presented a large-scale WVS-grounded framework for evaluating cross-cultural value faithfulness and drift in LLM-based multi-agent discussions. Under our evaluation protocol, current LLM agents generate coherent conversations but often fail to instantiate their assigned value profiles before exhibiting comparatively little measurable longitudinal drift. Because value attribution from dialogue is inherently subjective, the reported frequencies should be interpreted as approximate estimates rather than precise point values. Nevertheless, the qualitative asymmetry between initial instantiation failures and subsequent value drift is consistent across all evaluated models. These findings motivate evaluating LLM societies through behavioral consistency under interaction, rather than relying solely on static survey-style responses.

Limitations. This work exhibits the following limitations. First, value orientations were evaluated using an LLM-based judge, which may reflect cultural or alignment-related biases despite moderate agreement with human annotators, indicating that value attribution is inherently subjective. Accordingly, our reported percentages should be interpreted as approximate measurements. Future work should validate the evaluation using multiple independent judges and larger human-annotated subsets. Second, our simulations were conducted using only three language models, limiting the generalizability of the findings across broader open-source, multilingual models. Third, the simulated conversations were exclusively in English, which may suppress the cultural nuances and communication patterns present in multilingual real-world interactions. Finally, the simulations rely on synthetic conversational settings and predefined discussion topics, which may not fully capture the complexity of real-world social interaction.

Ethics Statement

This work does not aim to recommend or prescribe which value systems societies should adopt. Rather, our goal is observational: to analyze how current LLM agents instantiate, preserve, and modify culturally grounded value profiles during social interaction. In addition, the simulated personas are derived from anonymized and aggregated WVS demographic information and are not identifiable to any real individuals. Synthetic names and concise biographies were generated solely to support realistic conversational grounding. Finally, annotators were warned that some conversations may contain sensitive arguments due to the nature of value-laden social and political discussions. Therefore, we had their consent beforehand.

References

- Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.
- Bardi, A.; and Goodwin, R. 2011. The dual route to value change: Individual processes and cultural moderators. *Journal of cross-cultural psychology*, 42(2): 271–287.
- Bardi, A.; Lee, J. A.; Hofmann-Towfigh, N.; and Soutar, G. N. 2009. The structure of intraindividual value change. *Journal of personality and social psychology*, 97 5: 913–29.
- Bisbee, J.; Clinton, J. D.; Dorff, C.; Kenkel, B.; and Larson, J. M. 2024. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4): 401–416.
- Boyd, D. M.; and Ellison, N. B. 2007. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1): 210–230.
- Cacioppo, J. T.; and Petty, R. E. 1982. The need for cognition. *Journal of personality and social psychology*, 42(1): 116.
- Cao, Y.; Zhou, L.; Lee, S.; Piqueras, L. C.; Chen, M.; and Hershovich, D. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, 53–67.
- Choi, E. C.; Young, L. E.; and Ferrara, E. 2026. Overstating Attitudes, Ignoring Networks: LLM Biases in Simulating Misinformation Susceptibility. *arXiv preprint arXiv:2602.04674*.
- Chuang, Y.-S. 2024. *Simulating Human Opinion Dynamics Using AI Agents and Large Language Models*. The University of Wisconsin-Madison.
- Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 250–259.
- de Wildt, T.; and van de Poel, I. 2024. Modelling value change: An exploratory approach. *Journal of Artificial Societies and Social Simulation*, 27(1).
- Dehkordi, M. A. E.; Melnyk, A.; Herder, P.; and Ghorbani, A. 2024. Examining the interplay between national strategies and value change in the battle against COVID-19: An agent-based modelling inquiry. *Journal of Artificial Societies and Social Simulation*, 27(1).
- Durmus, E.; Ladhak, F.; and Cardie, C. 2019. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5668–5678.
- Dwivedi, A.; Lavania, P.; and Modi, A. 2023. EtiCor: Corpus for Analyzing LLMs for Etiquettes. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6921–6931. Singapore: Association for Computational Linguistics.
- Farg, Y.; Brand, C.; Amidei, J.; Piwek, P.; Stafford, T.; Stoyanchev, S.; and Vlachos, A. 2022. Opening up minds with argumentative dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4569–4582.
- Gao, F.; Zhang, T.; and Franklin, T. 2013. Designing asynchronous online discussion environments: Recent progress and possible future directions. *British Journal of Educational Technology*, 44(3): 469–483.
- Gudykunst, W. B.; Matsumoto, Y.; Ting-Toomey, S.; Nishida, T.; Kim, K.; and Heyman, S. 1996. The influence of cultural individualism-collectivism, self construals, and individual values on communication styles across cultures. *Human communication research*, 22(4): 510–543.
- Haerpfer, C.; Inglehart, R.; Moreno, A.; Welzel, C.; Kizilova, K.; Diez-Medrano, J.; Lagos, M.; Norris, P.; Ponarin, E.; Puranen, B.; and et al. 2022a. World Values Survey Wave 7 (2017-2022) Cross-National Data-Set.
- Haerpfer, C.; Inglehart, R.; Moreno, A.; Welzel, C.; Kizilova, K.; Diez-Medrano, J.; Lagos, M.; Norris, P.; Ponarin, E.; Puranen, B.; et al. 2022b. World values survey: Round seven-country-pooled datafile version 5.0.
- Henrich, J.; Heine, S. J.; and Norenzayan, A. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3): 61–83.
- Hew, K. F.; and Cheung, W. S. 2012. *Student participation in online discussions: Challenges, solutions, and future research*. Springer Science & Business Media.
- Hidey, C.; Musi, E.; Hwang, A.; Muresan, S.; and McKeown, K. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, 11–21.
- Hofstede, G. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.
- Inglehart, R.; and Welzel, C. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge University Press.
- Johnson, R. L.; Pistilli, G.; Menéndez-González, N.; Duran, L. D. D.; Panai, E.; Kalpokiene, J.; and Bertulfo, D. J. 2022.

- The Ghost in the Machine has an American accent: value conflict in GPT-3. *arXiv preprint arXiv:2203.07785*.
- Kleinberg, J.; and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.
- Lane, R. E. 1976. The Nature of Human Values. By Milton Rokeach. (New York: The Free Press, 1973. Pp. 438. \$13.95.). *American Political Science Review*, 70(3): 965–966.
- Liu, J.; Liu, C.; and Shen, H. 2026. ValueFlow: Measuring the Propagation of Value Perturbations in Multi-Agent LLM Systems. *arXiv preprint arXiv:2602.08567*.
- Madden, E. R. 2025. Evaluating the Use of Large Language Models as Synthetic Social Agents in Social Science Research. *Journal of Social Computing*, 6(4): 334–341.
- Maio, G. R.; and Olson, J. M. 1998. Values as truisms: evidence and implications. *Journal of personality and social psychology*, 74(2): 294.
- Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In Bengio, Y.; and LeCun, Y., eds., *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- Milfont, T. L.; Milojev, P.; and Sibley, C. G. 2016. Values stability and change in adulthood: A 3-year longitudinal study of rank-order stability and mean-level differences. *Personality and Social Psychology Bulletin*, 42(5): 572–588.
- Minson, J. A.; and Chen, F. S. 2022. Receptiveness to opposing views: Conceptualization and integrative review. *Personality and Social Psychology Review*, 26(2): 93–111.
- Osman, N.; and d’Inverno, M. 2023. Human Values in Multiagent Systems. *arXiv preprint arXiv:2305.02739*.
- Packer, C.; Wooders, S.; Lin, K.; Fang, V.; Patil, S. G.; Stolica, I.; and Gonzalez, J. E. 2024. MemGPT: Towards LLMs as Operating Systems. *arXiv:2310.08560*.
- Pandey, S. K.; Saha, S.; and Choudhury, M. 2025. To Generate or Discriminate? Methodological Considerations for Measuring Cultural Alignment in LLMs. In Inui, K.; Sakti, S.; Wang, H.; Wong, D. F.; Bhattacharyya, P.; Banerjee, B.; Ekbal, A.; Chakraborty, T.; and Singh, D. P., eds., *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 1545–1562. Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-303-6.
- Park, H. S.; Levine, T. R.; Weber, R.; Lee, H. E.; Terra, L. I.; Botero, I. C.; Bessarabova, E.; Guan, X.; Shearman, S. M.; and Wilson, M. S. 2012. Individual and cultural variations in direct communication style. *International Journal of Intercultural Relations*, 36(2): 179–187.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
- Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52.
- Petty, R. E.; and Cacioppo, J. T. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.
- Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 3982–3992.
- Ren, Y.; Ye, H.; Fang, H.; Zhang, X.; and Song, G. 2024. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2015–2040.
- Saha, S.; and Choudhury, M. 2025. User Behavior Prediction as a Generic, Robust, Scalable, and Low-Cost Evaluation Strategy for Estimating Generalization in LLMs. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 11047–11065. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Saha, S.; Das, S.; Soper, E.; Pacquetet, E.; and Srihari, R. K. 2021. Proto: A neural cocktail for generating appealing conversations. *arXiv preprint arXiv:2109.02513*.
- Saha, S.; Pandey, S. K.; and Choudhury, M. 2025. Meta-Cultural Competence: Climbing the Right Hill of Cultural Awareness. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8025–8042. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.
- Saha, S.; Pandey, S. K.; Gupta, H.; and Choudhury, M. 2025. Reading between the Lines: Can LLMs Identify Cross-Cultural Communication Gaps? In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8043–8067. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.
- Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, 29971–30004. PMLR.
- Schwartz, S. 1992. *Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries*, volume 25, 1–65. ISBN 9780120152254.
- Schwartz, S. H.; Lehmann, A.; and Roccas, S. 1999. Multi-method probes of basic human values. In *Social psychology and cultural context*, 107–124. SAGE Publications, Inc.

Seligman, C.; and Katz, A. N. 2013. The dynamics of value systems. In *The psychology of values*, 53–75. Psychology Press.

Shen, H.; Clark, N.; and Mitra, T. 2025. Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 3097–3118.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36: 8634–8652.

Stolcke, A.; Ries, K.; Coccaro, N.; Shriberg, E.; Bates, R.; Jurafsky, D.; Taylor, P.; Martin, R.; Van Ess-Dykema, C.; and Meteer, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3): 339–374.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, 613–624.

Tao, Y.; Viberg, O.; Baker, R. S.; and Kizilcec, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9): pgae346.

Tjautja, L.; Chen, V.; Wu, T.; Talwalkwar, A.; and Neubig, G. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12: 1011–1026.

Vecchione, M.; Schwartz, S.; Alessandri, G.; Döring, A. K.; Castellani, V.; and Caprara, M. G. 2016. Stability and change of basic personal values in early adulthood: An 8-year longitudinal study. *Journal of Research in Personality*, 63: 111–122.

Vecchione, M.; Spagnolo, G.; Daniel, E.; Benish-Weisman, M.; and Bardi, A. 2026. Stability and change of basic personal values in mid-to-late adolescence: A 4-year longitudinal study. *European Journal of Personality*, 40(2): 351–368.

Walther, J. B. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research*, 23(1): 3–43.

Wang, Y.; Zhao, J.; Ones, D. S.; He, L.; and Xu, X. 2025. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1): 519.

Yeomans, M.; Minson, J.; Collins, H.; Chen, F.; and Gino, F. 2020. Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, 160: 131–148.

Zhou, X.; Zhu, H.; Mathur, L.; Zhang, R.; Yu, H.; Qi, Z.; Morency, L.-P.; Bisk, Y.; Fried, D.; Neubig, G.; et al. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, 40975–41019.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.