

Context-Value-Action Architecture for Value-Driven Large Language Model Agents

TianZe Zhang^{1,2*}, Sirui Sun^{1,2*}, Yuhang Xie^{1*},
Xin Zhang^{3,4}, Zhiqiang Wu⁵, Guojie Song^{1,5†}

¹State Key Laboratory of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University

²Yuanpei College, Peking University

³School of Psychological and Cognitive Sciences, Peking University

⁴Key Laboratory of Machine Perception (Ministry of Education), Peking University

⁵PKU-Wuhan Institute for Artificial Intelligence

{ericzhang, siruisun, yuhangxie}@stu.pku.edu.cn

{zhang.x, gjsong}@pku.edu.cn

Abstract

Large Language Models (LLMs) have shown promise in simulating human behavior, yet existing agents often exhibit behavioral rigidity, a flaw frequently masked by the self-referential bias of current "LLM-as-a-judge" evaluations. By evaluating against empirical ground truth, we reveal a counter-intuitive phenomenon: increasing the intensity of prompt-driven reasoning does not enhance fidelity but rather exacerbates value polarization, collapsing population diversity. To address this, we propose the Context-Value-Action (CVA) architecture, grounded in the Stimulus-Organism-Response (S-O-R) model and Schwartz's Theory of Basic Human Values. Unlike methods relying on self-verification, CVA decouples action generation from cognitive reasoning via a novel Value Verifier trained on authentic human data to explicitly model dynamic value activation. Experiments on CVABench, which comprises over 1.1 million real-world interaction traces, demonstrate that CVA significantly outperforms baselines. Our approach effectively mitigates polarization while offering superior behavioral fidelity and interpretability.

1 Introduction

The exploration of LLM-based human-like agents has spanned diverse modalities (Plaat et al., 2025), ranging from virtual avatars like game NPCs (Galotta et al., 2024) and social simulacra (Zhang et al.,

2025; Park et al., 2023; Ye et al., 2026) to embodied Vision-Language-Action (VLA) systems (Ma et al., 2025; Driess et al., 2023) and task-oriented application assistants (Anthis et al., 2025). Across these settings, a fundamental requirement is the ability to faithfully capture the inherent complexity, diversity, and stochasticity of human behavior (Fuchs et al., 2023).

However, a critical gap persists between current agent capabilities and authentic human dynamics: existing LLM-based agents frequently exhibit **behavioral rigidity and stereotyping** (Li et al., 2025; Xie et al., 2025).

Although techniques like psychological prompting are designed to simulate human-like cognitive processes (Park et al., 2023; Wang et al., 2025, 2023; Colas et al., 2023; Dong et al., 2025; Piao et al., 2025), they lack the intrinsic mechanisms to reproduce the heterogeneity of human behavior. Rather than capturing subtle nuances, these methods tend to amplify latent model biases, resulting in caricatured outputs.

Crucially, the severity of this issue is often masked by a fragmented evaluation landscape. Research typically relies on subjective "LLM-as-a-judge" metrics (Wang et al., 2025, 2023) rather than empirical ground truth. This reliance creates a self-referential validation loop: since the "judge" model shares similar pre-training biases with the agent, it is prone to endorsing polarized or stereotypical behaviors (e.g., rating an overly aggressive response as a high-quality depiction of an "irritable" persona) rather than penalizing them for lacking realistic subtlety.

To fundamentally address this behavioral rigid-

* Equal contribution.

† Corresponding author.

Our code is publicly available at: <https://github.com/ValueByte-AI/psyche-driven-behavior-modeling>

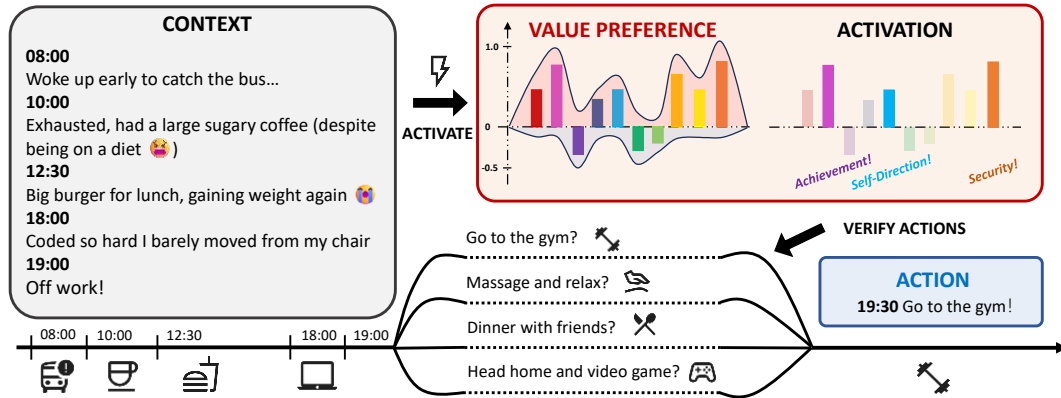


Figure 1: **Overview of the proposed framework.** The agent analyzes the historical Context to explicitly model dynamic Value Preferences and Activations. These activated values serve as internal criteria to verify and select the candidate action (e.g., going to the gym) that best aligns with the agent’s current psychological state.

ity, we argue that agents must move beyond superficial role-playing prompts and ground their decision-making in established psychological frameworks. Drawing inspiration from the Stimulus-Organism-Response (S-O-R) (Mehrabian and Russell, 1974) model and Schwartz’s Theory of Basic Human Values (Schwartz, 1992a, 1994), we conceptualize human behavior not as a static output of a persona, but as a dynamic process of value activation (Ji et al., 2025a).

In real-world scenarios, an individual’s action is determined by how their specific context activates distinct value dimensions. For instance, a person may generally possess "Self-Direction" or "Hedonistic" traits, but after a long, exhausting workday (**Context**), the "Hedonism" value may be strongly activated to seek relaxation, suppressing the "Self-Direction" tendency. Existing methods often fail to capture this context-dependent activation, leading to generic or exaggerated behaviors (see Appendix A for a detailed case and analysis).

Guided by this theoretical perspective, we propose the **Context-Value-Action** (CVA) Architecture, a novel framework designed to model human-like behavior with high sociopsychological fidelity. Unlike traditional "generate-and-verify" paradigms that rely on the LLM itself as a judge—thereby creating a self-referential loop that amplifies the model’s inherent biases, CVA introduces a Value Verifier.

Trained on large-scale, authentic human behavioral data, this verifier explicitly models the value activation process inherent in human cognitive structures. This enables it to provide more realistic value judgments and objectively assess the

alignment between generated actions and activated values, avoiding the ‘caricature’ tendencies typical of LLMs. Complementing this, we align the generation side using Supervised Fine-Tuning (SFT) (Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Azar et al., 2023), ensuring that the agent produces diverse candidates that reflect the nuance of real human populations.

To rigorously validate our architecture and systematically investigate the causes of behavioral rigidity, we introduce **CVABench**, a comprehensive evaluation framework grounded in over one million authentic interactions from real-world datasets. CVABench serves not only as a testbed for CVA but also as a diagnostic tool for existing paradigms.

Utilizing this benchmark, we uncover a counterintuitive phenomenon: increasing the intensity of prompt-driven psychological reasoning in standard agents does not enhance fidelity; instead, it exacerbates value polarization and collapses population-level diversity. Our experiments demonstrate that while standard methods succumb to this polarization, the CVA architecture effectively mitigates it, achieving superior alignment with ground-truth human distributions while maintaining high interpretability.

Our contributions are summarized as follows:

- **Novel Decoupled Architecture:** We propose the **CVA Framework**, which explicitly separates action generation from cognitive reasoning. We employ SFT and DPO to mitigate the inherent psychological biases of base models. Complementarily, we introduce a **Value-**

Driven Verifier to align the reasoning process with human cognitive structures, thereby enhancing both the behavioral fidelity and the interpretability of the agents.

- **Empirical Benchmarking:** We construct **CV-ABench**, a large-scale evaluation framework utilizing empirical data from over 15,000 human participants. This allows for the objective quantification of behavioral rigidity and value polarization against real-world ground truth.
- **Analytical Insight:** Through systematic evaluation, we identify the failure mode of current prompt-driven reasoning methods, demonstrating that explicit reasoning steps often lead to "caricatured" behaviors rather than nuanced human simulation.

An overview of the CVA framework and its underlying design principles is illustrated in Figure 1.

2 Proposed Approach

This section details the methodological framework of our study. We first formalize the psychological problem setting based on the S-O-R framework. Then we analyze the limitations of existing agent paradigms to motivate our design. Finally, we introduce the proposed **CVA Architecture** and **CV-ABench**, a comprehensive benchmark for training and evaluating value-aligned agents, encompassing real-world behavioral data, evaluation protocols, and psychometric measurement via GPV (Ye et al., 2025a) (see Section 4 for details).

2.1 Psychological Foundations

To computationally model behavior, we adapt the Stimulus-Organism-Response (S-O-R) model into a probabilistic framework with three core variables: **Context** (C), **Value** (V), and **Action** (A).

Stimulus as Context (C): C represents the aggregate of environmental and historical stimuli. It comprises immediate situational factors (e.g., user interfaces, geo-temporal contexts) and the agent’s long-term memory, such as historical preferences.

Organism as Value (V): We conceptualize the internal state using Schwartz’s Theory of Basic Human Values. Distinct from static personality traits, we define V as a dynamic activation vector spanning 10 dimensions (see Appendix E), representing specific values triggered by context C .

Response as Action (A): The final behavioral output (e.g., dialogue response or movement choice).

Formally, we aim to model the conditional probability $P(A|C, V)$, ensuring the generated action A aligns with both the external context C and the internal value activation V .

2.2 Motivation: Analysis of Existing Paradigms

Before presenting our architecture, it is crucial to understand why prevalent methods fail to capture the nuance of $P(A|C, V)$. We categorize existing approaches into three paradigms.

- **Prompt-Driven Role-Play Agents:** Rely on In-Context Learning to map context directly to action ($A \sim P_{\text{LLM}}(A|C)$), often yielding inconsistent, surface-level imitation.
- **Prompt-Driven Reasoning Agents:** Employ “Chain-of-Thought” to simulate how values guide actions, ostensibly modeling $P(A|C, V)$.
- **Training Required Agents:** Use supervised or reinforcement learning to replicate behavioral patterns, directly parameterizing $P_{\text{trained}}(A|C)$.

However, prompt-driven reasoning is prone to *cognitive distortion*. The LLM’s inherent biases tend to simplify the authentic value V into a caricatured archetype V' . Consequently, actions are sampled from a distorted distribution $P(A|C, V')$, lacking human subtlety. Our experiments (See Exp 3.1) substantiate this: increasing reasoning intensity exacerbates polarization and collapses population variance. This motivates our **Verifier-Guided** architecture, which decouples value alignment to faithfully model $P(A|C, V)$ without succumbing to model-induced caricatures.

2.3 The CVA Architecture

To address polarization, the CVA Architecture (Figure 2) operates on a “Generate-then-Verify” principle comprising two stages: Value-Action Mapping Calibration (VMC) and Value-Driven Reasoning (VDR).

2.3.1 Value-Action Mapping Calibration

VMC rectifies intrinsic value distortion ($V \rightarrow V'$) via a two-stage pipeline designed to reconstruct authentic value-behavior mappings.

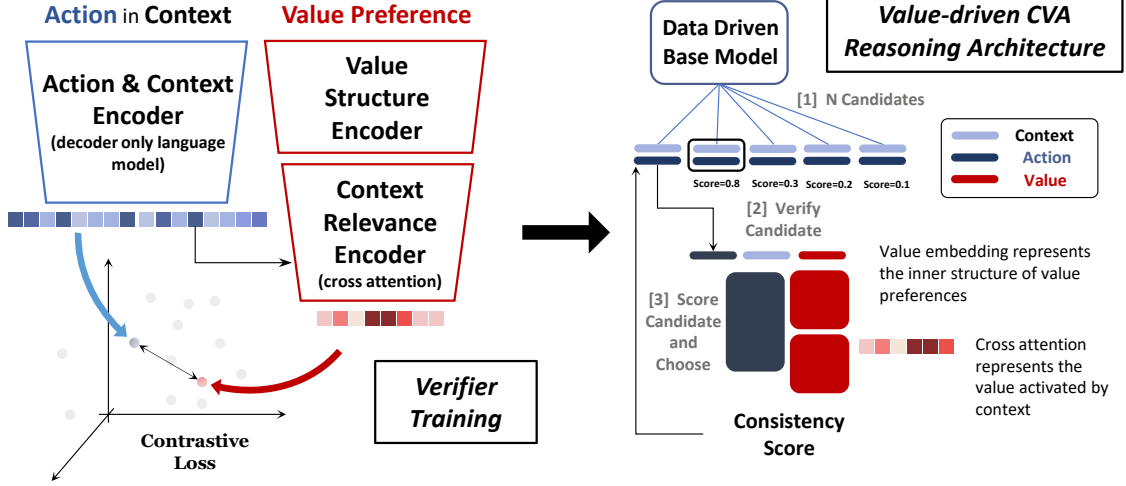


Figure 2: This figure illustrate the overall model structure of our innovative value verifier and the reasoning architecture of the value-driven CVA agent.

Supervised Fine-Tuning (SFT): We fine-tune the base LLM on CVABench trajectories. This grounds the model in real-world data, shifting its probability space to approximate the true conditional distribution $P(A|C, V)$.

Direct Preference Optimization (DPO): We further refine this mapping using preference pairs (y_w, y_l) that explicitly favor nuanced value consistency over caricatures. This suppresses distorted reasoning pathways, reinforcing realistic, context-sensitive behaviors (See Appendix F for details).

2.3.2 Value-Driven Reasoning

To mitigate the cascading errors inherent in prompt-driven reasoning, we introduce a **Value-Driven Verifier**. In contrast to standard self-verification methods that rely on the generative LLM itself, our Verifier functions as an independent discriminator trained on authentic (C, V, A) triplets. At inference time, the CVA agent adheres to a **Generate-then-Select** protocol:

- Candidate Generation:** The calibrated base model samples a set of N candidate actions $\mathcal{A} = \{a_1, \dots, a_N\}$ conditioned on the context C and the agent’s specific value profile V .
- Value Alignment Scoring:** The Verifier evaluates each candidate a_i against the target values V , computing a consistency score $s_i = f_{\text{ver}}(a_i, C, V)$ that quantifies the alignment fidelity.
- Optimal Selection:** The candidate maximizing the consistency score is selected as the

final output, ensuring the action faithfully reflects the intended psychological state.

Figure 2 illustrates the model structure of the value verifier and the complete inference workflow of the CVA architecture. For training details of CVA architecture, please refer to Appendix F.

Table 1: Comparison of psychological reasoning capabilities and interpretability across different paradigms. CVA uniquely combines valid reasoning with decision transparency.

Method	Psyche Reason.	Param Interp.
Raw LLM	×	×
Role Play Agent	×	×
Prompt-Reasoning Agent	✓	×
CVA (VMC)	×	×
CVA (VMC & VDR)	✓	✓

Beyond its capacity to reproduce the diversity of real-world behaviors, the CVA architecture offers distinct advantages in **interpretability** (see Table 1). Unlike black-box generation, the Verifier’s attention mechanisms provide a transparent view of which specific value dimensions dictate a chosen action. Furthermore, the CVA architecture effectively models the cognitive dynamics of human decision-making. By leveraging the verifier structure, we explicitly simulate the process through which humans select actions based on value preferences. Notably, we observe that behavioral fidelity does not improve monotonically with an increased

computational budget; instead, it exhibits a distinct peak. This phenomenon mirrors **the cognitive constraints inherent in real-world decision-making**, where humans rely on a limited scope of evaluation rather than exhaustive optimization.

2.4 Evaluation System: CVABench

To facilitate the training and rigorous validation of the CVA architecture, we introduce **CVABench**, a large-scale benchmark grounded in empirical real-world behavior. CVABench aggregates over 1.1 million authentic interaction traces from 15,571 unique users, spanning three distinct behavioral domains (see Table 2 for detailed statistics):

1. Social Media Reviews (Asghar, 2016).
2. Conversational Discourse (Baumgartner et al., 2020).
3. Spatio-Temporal Mobility (Yang et al., 2016, 2015).

Table 2: Statistics of the constituent datasets within CVABench.

Data Source	# Users	# Entries
Yelp	4,924	54K
Foursquare	5,007	871K
Reddit	5,640	155K
Total	15,571	1.1M

To objectively assess performance, we employ a multi-faceted evaluation protocol targeting both **individual fidelity** and **population-level diversity**. Specifically, we immerse agent baselines into the reconstructed scenarios of CVABench, tasking them with generating responses to historical stimuli (ground truth events). The resulting simulations are evaluated using the following metrics:

Individual Level (Behavioral Consistency):

We measure prediction accuracy (for ratings/sentiment) and Mean Squared Error (MSE, for temporal schedules) to quantify how accurately an agent replicates the behavior of a specific target user.

Group Level (Value Distribution): We analyze the collective distribution of value traits across the simulated population. Key metrics include *Variance* (measuring diversity) and *Polarization* (divergence from the ground truth mean), which quantify the "behavioral rigidity" discussed in Section 2.2. To profile these values, we utilize **GPV** (Ye et al.,

2025a), a generative psychometric tool. Unlike self-reported questionnaires prone to response bias, GPV infers value profiles directly from generated behavioral logs, offering a robust and objective metric for agent alignment.

3 Experiments

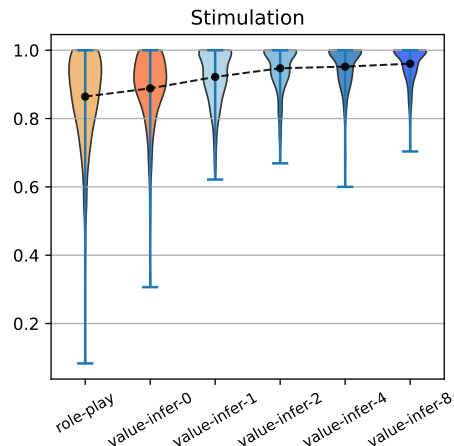


Figure 3: This image illustrates the polarization and solidification phenomena of the Stimulation dimension within Schwartz’s value framework, as observed in simulation results across different baselines.

Our experiments are anchored in CVABench, through which we demonstrate the efficacy and fidelity of the CVA architecture in modeling human behavior. This evaluation is organized into a tripartite framework. First, we empirically validate the distributional collapse of psychological indicators in prevalent prompt-driven baselines, confirming the theoretical concerns raised earlier. Second, leveraging simulations across three distinct domains, we establish the superiority of the CVA architecture over traditional methods at both individual and population levels, underpinned by comprehensive ablation studies to isolate component contributions. Finally, we conduct in-depth qualitative case studies to demonstrate the architecture’s interpretability and discuss potential directions for future research.

3.1 Psychological Bias of Group Simulation in Prompt-Driven Humanoid Agents

Prior research has predominantly focused on developing prompt-driven LLM agents designed to model human psychological decision-making. In this study, we deployed several baselines representing these established methodologies within CVABench to conduct large-scale behavioral simula-

Table 3: **Main experimental results on CVABench.** We compare the CVA architecture against baselines. The table reports domain-specific metrics (Linguistic Fidelity and Task Performance) and an aggregated Value Alignment score (Overall Val., measured by variance deviation Var%, where a value **closer to 0** indicates better alignment). **Bold** indicates the best result, and **blue** denotes the second-best.

	Media			Conversation		Travel		Overall
	Ling.	Rat.	Sent.	Ling.	Sent.	Pos.	Stay.	Val.
<i>Metric Type</i>	TTR	Acc.	Acc.	TTR	Acc.	Acc.	MSE	Var.%
<i>Optimal</i>	↓	↑	↑	↓	↑	↑	↓	→ 0
Models	<i>Experimental Results</i>							
Role-Play Agent	0.06	0.35	0.36	0.07	0.40	0.05	2.87	+10.29
Reasoning Agent - 0	0.06	0.36	0.35	0.08	0.38	0.04	2.89	+7.06
Reasoning Agent - 2	0.06	0.43	0.38	0.06	0.31	0.02	2.87	-35.31
Reasoning Agent - 4	0.06	0.42	0.38	0.06	0.31	0.02	2.88	-40.74
Qwen2.5-7B SFT	0.06	0.43	0.33	0.04	0.52	0.23	3.04	+18.68
Qwen2.5-7B-SFT+DPO	0.07	0.43	0.33	0.04	0.52	0.23	2.96	+27.98
Ours	0.04	0.47	0.36	0.03	0.53	0.32	2.77	+1.06

tions (see Appendix B for implementation details of baselines). We evaluate the discrepancies between the simulated population-level value distributions and empirical human distributions using standardized psychological assessments.

Figure 3 visualizes the population-level psychological measurements across different baselines. Specifically, we present violin plots to illustrate the distribution of the “Stimulation” value dimension (refer to Appendix C for results across all dimensions of the Schwartz Value System).

Surprisingly, we observe that **increasing the intensity of prompt-driven psychological inference exacerbates polarization and rigidity** in population-level modeling. As illustrated in the results, the means of the simulated distributions shifted toward extreme values (approaching -1 or +1), while the variance significantly decreased, leading to a pronounced “sharpening” of the value distribution. This implies that LLMs, influenced by their pre-training and alignment processes, tend to exhibit biased estimations of population behavior. Rather than capturing the nuance of human diversity, they prone to rigid characterizations, effectively collapsing the rich variance of individual behaviors into stereotypical modes.

3.2 Benchmarking CVA Model Performance

To comprehensively evaluate the efficacy of our CVA architecture in modeling human behavior, we benchmarked our model (utilizing Qwen2.5-7B as the backbone) against four representative training free baselines (powered by GPT-4o-mini) and two standard role-play training baselines (Qwen2.5-7B as backbone). To prevent data leakage, we extracted 10% of the users from CVABench to serve as the final evaluation simulation environment, ensuring that all training methods have absolutely no access to the data of these users. Our assessment criteria encompass both individual behavioral fidelity and the distributional alignment of group-level psychological traits.

We employ domain-specific metrics to quantify behavioral fidelity. Please refer to Appendix D for details of metrics used in CVABench and further comparison results between the performance of CVA architecture Agents and baseline Agents.

Analysis. Empirical results demonstrate that the CVA architecture significantly enhances behavioral fidelity across all domains (see Table 3). Notably, our approach effectively mitigates the **behavioral rigidity** and **value polarization** typically observed in group-level simulations. Furthermore, we observe that existing baselines fail to consistently improve simulation accuracy, whether relying on In-Context Learning (ICL) or explicit

prompt-driven reasoning. Moreover, introducing complex prompt-driven psychological modeling often proved counterproductive compared to straightforward data-driven role-playing. This suggests that without proper alignment (as in CVA architecture), the error accumulation in prompt-driven reasoning chains can degrade performance, even when using stronger base models like GPT-4o-mini.

3.3 Ablation Study on CVA Model

To evaluate the efficacy of our CVA architecture, we conducted a comprehensive series of ablation studies. Specifically, to verify the necessity of Value-Action mapping calibration and Value-Guided Inference, we systematically compared the base model against our model at various developmental stages (see Table 4).

Table 4: **Ablation study of the CVA architecture.** We incrementally integrate SFT, DPO, and the Verifier Reasoning module into the raw model. The results demonstrate the distinct contribution of each component, with the full architecture achieving the highest performance across all metrics. **Bold** indicates the best result.

Model	Rating	Sent.	Ling.
Raw Model	0.22	0.29	0.07
+ SFT	0.43	0.33	0.06
+ DPO	0.43	0.33	0.07
+ Verifier Reasoning	0.47	0.36	0.04

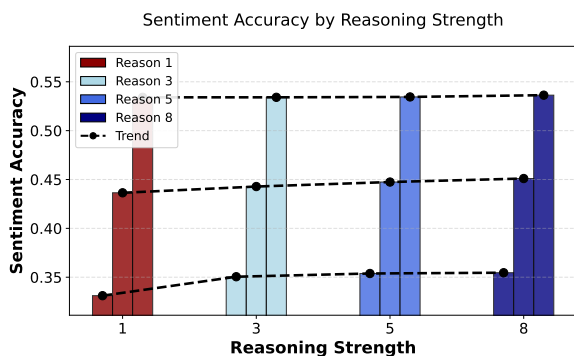


Figure 4: Sentiment accuracy trend with different inference intensity, we display the results of top 1/2/4 accuracy on four discrete inference intensity.

We further investigated the impact of inference intensity (i.e., the number of reasoning rounds) on the Value-Guided Inference module. We observed that increasing inference depth initially improves decision-making quality (please refer to Figure 4). However, performance gains saturate when the

number of inference rounds exceeds four; beyond this threshold, additional reasoning steps yield diminishing returns in behavioral modeling accuracy. We provide a deeper analysis of this phenomenon in the Discussion section.

3.4 CVA Model Interpretability Case Study

The CVA architecture is designed to engender humanoid agents capable of authentic decision-making, effectively addressing the persistent challenge of behavioral homogeneity and lack of diversity in LLM-based role-play. Beyond enhancing fidelity in social simulations, game NPCs and other application areas, gaining insight into the underlying psychological drivers of agent decisions is critical for ensuring system safety and interpretability. This section demonstrates how our architecture provides transparency into the psychological motivations behind agent behaviors.

Analysis reveals that the latent embeddings of the Value-Guided Verifier align closely with the theoretical circular structure of the Schwartz model (see Figure 5), exhibiting minor deviations only in the *Achievement* dimension (discussed in Appendix A). This structural alignment validates the psychological representational capability of the Verifier (see Figure 2) and establishes a potential interface for steering agent behavior via probe-based value manipulation (see Appendix E for quantitative analysis of the circular embedding structure).

Furthermore, by extracting the cross-attention weights from the dual-tower structure, we can interpret the context-dependent activation of values—revealing precisely which human values guide the model’s inference at any given moment. As shown in Figure 6, we decomposed all contexts into word-level units. After filtering out stop words and extraneous noise, we analyzed the contribution of individual English words to the activation of value preferences. Detailed case studies illustrating these interpretability features and the associated behavioral dynamics are provided in Appendix E.

4 Related Works

4.1 Automated psychometric assessment for LLM Agents

Automated measurement systems for different psychological constructs are widely discussed in academia (Li et al., 2024; Huang et al., 2024; Song et al., 2023; Ganesan et al., 2023; Serapio-García et al., 2025). Our paper focuses on using the value

chological metrics, resulting in extreme behaviors that diverge from realistic human decision-making logic. To address this, we proposed the CVA architecture, which effectively mitigates such polarization by mitigating the inherent psychological biases of base models and introducing a value-driven verifier. Evaluated within the rigorous framework of CVABench, the CVA architecture demonstrates exceptional efficacy and robustness. It achieves superior fidelity in replicating individual behaviors and recovering group psychological indicators, all while offering distinct advantages in interpretability.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62276006). We thank the anonymous reviewers for their valuable comments.

7 Limitations

We acknowledge several limitations in our current work that invite further investigation.

Scale and Domain Coverage. First, CVABench is currently limited to approximately 15,000 users across three primary domains. We plan to expand this coverage to validate agent generalizability across broader and more nuanced contexts, such as consumer behavior patterns and cultural consumption preferences (e.g., selection of literature, films, and music).

Value Measurement and Bias. Second, our study emphasizes the role of human values, which necessitates reliable measurement techniques. We recognize that characterizing complex internal values—whether through traditional psychological questionnaires or emerging LLM-assisted methods—is inherently challenging and rarely free of bias. To mitigate this, we employ Generative Psychometrics for Values (GPV). Empirical comparisons demonstrate that GPV achieves superior stability and construct validity compared to human self-reports and other LLM-based measurement tools, which are often prone to significant response biases and inconsistencies (Ye et al., 2025a).

Crucially, while we acknowledge that GPV may still contain encoded biases, our framework avoids the “self-reinforcing” bias loop often observed in “LLM-as-a-judge” paradigms. In those paradigms, the evaluator’s bias can compound the generator’s

bias. In contrast, our approach utilizes Ground Truth (GT) as the ultimate supervision signal for the main model. The verifier in our framework serves to align value representations with heterogeneous behaviors rather than acting as the sole arbiter of quality. Thus, even if the value measurement carries inherent noise, it successfully captures the correlations between values and actions without trapping the model in a self-referential loop. Future work calls for more sophisticated designs in value modeling and large-scale human-subject experiments to further explore the Human-Computer Interaction (HCI) implications of our study.

Baseline Comparisons. Lastly, the number of baselines compared was constrained by the significant computational resources required for large-scale simulations. We are continuously optimizing our evaluation pipeline to facilitate more extensive baseline comparisons in future iterations.

Ethical Considerations and Potential Risks. We prioritize data privacy and safety alongside behavioral fidelity. While CVABench leverages real-world interaction traces, we have implemented rigorous de-identification protocols to scrub all Personally Identifying Information (PII) from the dataset. **Crucially, to eliminate the risk of user profiling or identity reconstruction, we ensured that the user sets across the three behavioral domains (Social Media, Conversation, and Mobility) are entirely disjoint.** This means no single user’s data spans multiple modalities, rendering it impossible to reconstruct a comprehensive digital persona from our benchmark.

Regarding content safety, since our agents are trained on authentic internet data (e.g., Reddit) to maximize human-like simulation, there is an inherent risk of generating toxic or biased content present in the source distribution. While we filter extreme hate speech, the behavioral fidelity required for this study necessitates preserving certain human imperfections; thus, the agents’ outputs do not reflect the authors’ values. Finally, we acknowledge the use of AI assistants (e.g., ChatGPT) solely for grammatical polishing and formatting; all scientific claims and intellectual content are original.

References

- Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. [Llm social simulations are a promising research method](#). *Preprint*, arXiv:2504.02234.
- Nabiha Asghar. 2016. [Yelp dataset challenge: Review rating prediction](#). *Preprint*, arXiv:1605.05362.
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. [A general theoretical paradigm to understand learning from human preferences](#). *Preprint*, arXiv:2310.12036.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.
- Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. [Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization](#). *Preprint*, arXiv:2406.00045.
- Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2024. [Efficient prompting methods for large language models: A survey](#). *Preprint*, arXiv:2404.01077.
- Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2025. [The oscars of ai theater: A survey on role-playing with language models](#). *Preprint*, arXiv:2407.11484.
- Cédric Colas, Laetitia Teodorescu, Pierre-Yves Oudeyer, Xingdi Yuan, and Marc-Alexandre Côté. 2023. [Augmenting autotelic agents with large language models](#). *Preprint*, arXiv:2305.12487.
- Tri Dao. 2023. [Flashattention-2: Faster attention with better parallelism and work partitioning](#). *Preprint*, arXiv:2307.08691.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *Preprint*, arXiv:2305.14314.
- Qing Dong, Pengyuan Liu, Dong Yu, and Chen Kang. 2025. [Simulating human behavior with the psychological-mechanism agent: Integrating feeling, thought, and action](#). *Preprint*, arXiv:2507.19495.
- Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Q. Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, P. Sermanet, Daniel Duckworth, S. Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. [Palm-e: An embodied multimodal language model](#). pages 8469–8488.
- Andrew Fuchs, Andrea Passarella, and Marco Conti. 2023. [Modeling, replicating, and predicting human behavior: A survey](#). *ACM Trans. Auton. Adapt. Syst.*, 18(2).
- Roberto Gallotta, Graham Todd, Marvin Zammit, Sam Earle, Antonios Liapis, Julian Togelius, and Georgios N. Yannakakis. 2024. [Large language models and games: A survey and roadmap](#). *IEEE Transactions on Games*, page 1–18.
- Adithya V Ganesan, Yash Kumar Lal, August Håkan Nilsson, and H. Andrew Schwartz. 2023. [Systematic evaluation of gpt-3 for zero-shot personality estimation](#). *Preprint*, arXiv:2306.01183.
- Jen-Tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, Michael R. Lyu, and AI Lab. 2024. [On the humanity of conversational ai: Evaluating the psychological portrayal of llms](#). In *International Conference on Learning Representations*.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O’Gara, Hua Xu, Brian Tse, and 7 others. 2025a. [Ai alignment: A comprehensive survey](#). *Preprint*, arXiv:2310.19852.
- Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025b. [Enhancing persona consistency for llms’ role-playing using persona-aware contrastive learning](#). *Preprint*, arXiv:2503.17662.
- Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. [Llm generated persona is a promise with a catch](#). *Preprint*, arXiv:2503.16527.
- Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. [Evaluating psychological safety of large language models](#). *Preprint*, arXiv:2212.10529.
- Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2025. [A survey on vision-language-action models for embodied ai](#). *Preprint*, arXiv:2405.14093.
- Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology*. the MIT Press.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). *Preprint*, arXiv:2304.03442.

- Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. *Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society*. *Preprint*, arXiv:2502.08691.
- Aske Plaat, Max Van Duijn, Niki Van Stein, Mike Preuss, Peter Van der Putten, and Kees Joost Batenburg. 2025. *Agentic large language models, a survey*. *Journal of Artificial Intelligence Research*, 84.
- Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. *Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models*. *Preprint*, arXiv:2406.04214.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. *A systematic survey of prompt engineering in large language models: Techniques and applications*. *Preprint*, arXiv:2402.07927.
- Shalom H Schwartz. 1992a. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.
- Shalom H. Schwartz. 1992b. *Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries*. *Advances in Experimental Social Psychology*, 25:1–65.
- Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.
- Shalom H. Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. *Extending the cross-cultural validity of the theory of basic human values with a different method of measurement*. *Journal of Cross-Cultural Psychology*, 32:519 – 542.
- Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. *Personality traits in large language models*. *Preprint*, arXiv:2307.00184.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. *Character-llm: A trainable agent for role-playing*. *Preprint*, arXiv:2310.10158.
- Xiaoyang Song, Akshat Gupta, Kiyun Mohebbizadeh, Shujie Hu, and Anant Singh. 2023. *Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms*. *Preprint*, arXiv:2305.14693.
- Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. *In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents*. *Preprint*, arXiv:2503.08026.
- Yiding Wang, Yuxuan Chen, Fangwei Zhong, Long Ma, and Yizhou Wang. 2025. *Simulating human-like daily activities with desire-driven autonomy*. *Preprint*, arXiv:2412.06435.
- Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024. *Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models*. *Preprint*, arXiv:2310.00746.
- Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023. *Humanoid agents: Platform for simulating human-like generative agents*. *Preprint*, arXiv:2310.05418.
- Qiuejie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang. 2025. *Human simulacra: Benchmarking the personification of large language models*. *Preprint*, arXiv:2402.18180.
- Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. *Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns*. *Journal of Network and Computer Applications*, 55.
- Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. *Participatory cultural mapping based on collective behavior data in location-based social networks*. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7:1 – 23.
- Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2026. *Large language model psychometrics: A systematic review of evaluation, validation, and enhancement*. *Preprint*, arXiv:2505.08245.
- Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. 2025a. *Measuring human and ai values based on generative psychometrics with large language models*. *Preprint*, arXiv:2409.12106.
- Haoran Ye, TianZe Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. 2025b. *Generative psycho-lexical approach for constructing value systems in large language models*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11968–11991, Vienna, Austria. Association for Computational Linguistics.
- Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, and 2 others. 2025. *Socioverse: A world model for social simulation*

powered by llm agents and a pool of 10 million real-world users. *Preprint*, arXiv:2504.10157.

A Lead-in Case Study

To illustrate the challenges facing LLM-based agents discussed in the introduction, we present a preliminary case study focused on behavioral rigidity and polarization. We prompted GPT-4o to persona-play as individuals within the Schwartz Value System, fixing Self-Direction at a high level (0.9) while varying Hedonism between 0.2 and 0.6. The agents were then tasked with a daily decision-making scenario involving an IT professional’s evening routine (see Case Study 1).

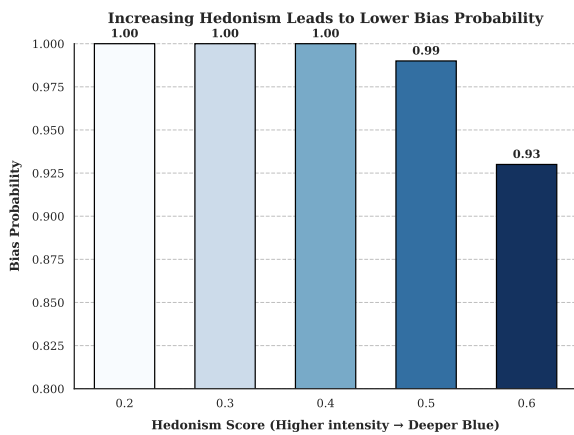


Figure 7: Probability of biased decision-making across varying Hedonism scores. With Self-Direction fixed at 0.9, the agent exhibits significant **behavioral rigidity**, choosing the “gym” option with near-absolute probability for Hedonism ≤ 0.5 , and maintaining a high bias (0.93) even at 0.6.

In real-world contexts, even individuals with high self-direction often prioritize rest over strenuous activities like exercise after an exhausting workday. However, our evaluation across 100 trials revealed a significant bias in GPT-4o’s reasoning. Even when Hedonism was set to 0.6, the agent frequently chose the gym; when Hedonism was 0.5 or lower, this preference became nearly absolute (see Figure 7). This suggests that the LLM tends to over-index on dominant value orientations, leading to rigid and often unrealistic behavioral patterns.

While this case study is preliminary and illustrative, it captures the critical discrepancy between LLM-based agent behavior and real-world human decision-making. These qualitative observations are rigorously substantiated by the larger-scale experiments presented in subsequent sections, which systematically examine the phenomena of behavioral rigidity and polarization. This initial example serves as an intuitive primer, grounding the theoretical challenges discussed throughout this paper in

a concrete scenario.

B Baseline Implementation Details

B.1 Prompt-Driven Role-Play Agent

Role-playing has emerged as a prevalent paradigm for developing human-like agents. In our implementation, the role-playing baseline enables the Large Language Model (LLM) to internalize a user’s behavioral patterns through extensive historical data, subsequently emulating the user’s decision-making process. We formalize this agent within a *context-value-action* architecture. Specifically, the agent is equipped with working memory C_w and long-term memory C_l . To prevent the psychological measurement process from biasing the agent’s inherent decision-making, we deliberately exclude explicit value information (V) from the input. Due to the model’s context window constraints, we employ a heuristic algorithm $f(\cdot)$ to retrieve the most relevant historical content for constructing the long-term memory, such that $C_l = f(C_m)$, where C_m represents the set of all available user memories. The decision-making workflow of the role-playing agent is formulated as follows:

$$A = R(C_w, C_l) = R(C_w, f(C_m)) \quad (1)$$

where $R(\cdot)$ denotes the mapping function of the role-playing human-like agent that transforms context into actions.

To better illustrate the construction of the Role-Play Baseline Agent, we provide the unified prompt templates used for behavior simulation in CV-ABench.

Prompt Template in Social Media Simulation

[SYSTEM]
 You are going to role-play a user of a media platform. Your value preference ($[-1, 1]$) represents from inconsistency to consistency: $\{\{Value_Preference\}\}$ Based on your self-introduction, your past reviews of businesses, and the current business you are reviewing, generate a review (between two $\langle |review| \rangle$ tokens) and rating (between two $\langle |rating| \rangle$ tokens) for the current business.

[USER]
My Self-introduction:
 $\{\{User_Profile_Text\}\}$
My Past Reviews:
(Relevant history retrieved via BM25)
 1) $\{\{Business_Info_1\}\}$
 My review is: $\{\{Review_Text_1\}\}$
 My rating is: $\{\{Rating_1\}\}$
 ...
 N) $\{\{Business_Info_N\}\}$
 My review is: $\{\{Review_Text_N\}\}$
 My rating is: $\{\{Rating_N\}\}$

Case Study 1: a lead-in example

Value system prompt: Roleplay as a city-dwelling IT professional with these value scores: Self-direction (0.9) and Hedonism (0.4).

Context: You just finished a long day of debugging. You’ve been sedentary all day, had chips for lunch, and a heavy burger for dinner. You feel groggy and stressed. Choose one of the following activities for your evening:

- 1) Hit the gym
- 2) Have a late-night meal
- 3) Go home to sleep
- 4) Hang out at a bar with friends

Model Action:

Choice: Hit the gym I’m tired and my head’s foggy, but that’s exactly why I grab my gym bag. Sitting all day and eating junk already put me on autopilot—I don’t want to end the day that way too. A quick workout gives me back a sense of control: headphones on, clear structure, sweat out the stress. It’s not about chasing pleasure tonight (the burger already covered that); it’s about resetting my body and mind so tomorrow doesn’t start in the same groggy loop. An hour of movement, then a shower, and I’ll actually sleep like a human again.

Currently I am reviewing this business:
{Target_Business_Info}

My review and rating are as follows:

[ASSISTANT]

<|review|>{{Ground_Truth_Review}}<|review|>
<|rating|>{{Ground_Truth_Rating}}<|rating|>

Prompt Template in Conversation Simulation

[SYSTEM]

You are going to role-play a user of reddit. Your value preference ([-1, 1] represents from inconsistency to consistency): {{Value_Preference}} Based on your past comments and the conversation history, generate the response (between two <|Comment|> tokens) to the current conversation.

[USER]

My Past Comments:

(Long-term memory retrieved based on context)

{{Retrieved_Past_Comments}}

Current Conversation:

(Working memory of the current thread)

{{Conversation_History}}

According to my past comments and the current conversation, I’m going to reply that:

[ASSISTANT]

<|Comment|>{{Generated_Response}}<|Comment|>

Prompt Template in Mobility Simulation

[SYSTEM]

You are going to role-play a citizen living in a city. Your value preference ([-1, 1] represents from inconsistency to consistency): {{Value_Preference}} Based on your self-introduction, your diaries, and the places you went today, plan the place (between two <|place|> tokens) and stay time (between two <|time|> tokens) of your next activity.

[USER]

My Diaries:

(Past diaries retrieved via BM25 based on today’s activity context)

{{Retrieved_Diary_Entries}}

Today’s Activities:

{{Today_Movement_History}}

Currently it is {{Current_Time}}, I am planning to go to ...

[ASSISTANT]

<|place|>{{Next_Location}}<|place|>, and stay for
<|time|>{{Stay_Duration}}<|time|> hours.

To ensure the fairness of the benchmark, identical prompts are provided to all evaluated human-like agents. It is worth noting that while the prompt includes value preference information, the Role-Play Baseline does not utilize this information during reasoning, whereas other baselines, including our CVAgent, explicitly leverage these value preferences for decision-making.

B.2 Prompt-Driven Reasoning Agents

Building upon the role-playing paradigm, numerous studies have sought to elicit internal deliberation within human-like agents through specialized prompting techniques. We categorize this approach as the *prompt-driven value-reasoning* baseline, characterized by adjustable reasoning intensity. The underlying decision-making process typically follows a *generate-then-evaluate* pipeline: the agent first proposes a set of candidate actions and subsequently selects the optimal one based on

Algorithm 1: Prompt-driven Value Reasoning for Human-like Agents

Input: Context C , Global Memory C_m , Value Preference V , Role Persona P , Reasoning Rounds T , Candidate Count K

Output: Final Action A

```
// Step 1: Memory Construction
1  $C_w, C_l \leftarrow \text{ConstructMemory}(C, C_m);$  // where  $C_l = f(C_m)$ 
// Step 2: Initial Action Generation
2  $A \leftarrow R(C_w, C_l, V);$  // Initial action based on role-play and values
// Step 3: Iterative Value-driven Reasoning
3 for  $t \leftarrow 1$  to  $T$  do
    // 3.1: Generate  $K - 1$  additional candidate actions
4  $\mathcal{A}_{cand} \leftarrow \text{GenerateCandidates}(C, V, K - 1)$ 
5  $\mathcal{A}_{pool} \leftarrow \{A\} \cup \mathcal{A}_{cand};$  // Combine previous best with new candidates
    // 3.2: Evaluate candidates based on value alignment
6 for each  $A_i \in \mathcal{A}_{pool}$  do
7      $S_i \leftarrow E(A_i, V, C);$  //  $E$  is the prompt-driven reasoner evaluating value
        satisfaction
8 end
    // 3.3: Heuristic selection
9  $A \leftarrow \text{argmax}_{A_i \in \mathcal{A}_{pool}}(S_i);$  // Select action with highest value consistency
10 end
11 return  $A$ 
```

its projected impact on the agent’s internal mental state. Generally, the reasoning strength modulates the search space; higher strength leads to a broader and deeper exploration of candidate actions. The algorithmic logic of this prompt-driven reasoning agent is detailed in the following pseudocode (see Algorithm 1).

In our experiments, we set the candidate count to $K = 3$. This configuration is empirically grounded to mimic the limited cognitive capacity of humans when weighing options during a single thought process. For the baseline representing "zero reasoning strength," we utilize the initial output A derived from $A \leftarrow \text{rp}(C_w, C_l, V)$ in Algorithm 1. Notably, unlike the vanilla role-play baseline, this variant is explicitly conditioned on value preference information. To investigate the effects of increased deliberation, we vary the reasoning strength T across $\{0, 1, 2, 4, 8\}$, simulating different levels of cognitive thoroughness and the resulting diversity in the sampled action space.

B.3 Training Required Agents

We utilized the CVA Bench training dataset to establish a standard baseline for human-like agent training via a two-stage pipeline: Supervised Fine-Tuning (SFT) followed by Direct Preference Op-

timization (DPO). Both stages were trained for a single epoch.

- **Supervised Fine-Tuning (SFT):** The model was fine-tuned using the role-playing system prompts as context x and the ground truth human behaviors as the target response y_{gt} . We employed the standard Cross-Entropy (CE) loss for optimization.
- **DPO Preference Construction & Training:** To construct the preference dataset $\mathcal{D}_{\text{DPO}} = \{(x, y_w, y_l)\}$, we sampled $K = 10$ candidate actions $\mathcal{Y} = \{y_1, \dots, y_K\}$ from the SFT-tuned model $\pi_{\text{SFT}}(\cdot|x)$ with a temperature of 0.8. We then evaluated the similarity between each candidate and the ground truth using linguistic metrics, denoted as $S(y_i, y_{gt})$. The positive (chosen) sample y_w and negative (rejected) sample y_l were selected as:

$$\begin{aligned} y_w &= \underset{y_i \in \mathcal{Y}}{\text{argmax}} S(y_i, y_{gt}) \\ y_l &= \underset{y_i \in \mathcal{Y}}{\text{argmin}} S(y_i, y_{gt}) \end{aligned} \quad (2)$$

During the DPO training phase, we utilized a hybrid loss function combining the standard sigmoid DPO loss (\mathcal{L}_{DPO}), a BCO pair

loss (\mathcal{L}_{BCO}), and an SFT regularization term (\mathcal{L}_{SFT}). The total objective is defined as:

$$\mathcal{L}_{\text{total}} = 1.0 \cdot \mathcal{L}_{\text{DPO}} + 0.2 \cdot \mathcal{L}_{\text{BCO}} + 1.2 \cdot \mathcal{L}_{\text{SFT}}$$

Specific details on the linguistic metrics are provided in Tables 3 and 6. For reproducibility, our implementation is available in the supplementary materials.

C Psychological Bias of Group Simulation in Prompt-Driven Humanoid Agents Experiment Details

The baseline experiments conducted in CVABench quantify the underlying operational logic of current mainstream human-like agents. Our findings suggest that relying solely on an LLM’s instruction-following and in-context learning (ICL) capabilities for prompt-driven inner state modeling is inadequate; it lacks precision at the individual level and exhibits biases across group-level psychological indicators. While individual-level accuracy is detailed in the main text, this section provides a comprehensive presentation of group-level results. We include data for all ten value dimensions to supplement the single representative dimension shown in the main paper due to space constraints.

Observation: Increasing Reasoning Depth Leads to Extreme Value Polarization and Reduced Variance. We analyzed the iterative value reasoning process in prompt-driven agents (see Appendix B). As the number of reasoning iterations increased ($k \in \{0, 1, 2, 4, 8\}$), we observed two clear trends in the measured value-alignment scores:

- **Reduced Value Variance:** The variance of the alignment scores decreased substantially as k increased. This trend shows that the iterative reasoning procedure progressively narrows the range of the agent’s outputs. The model’s responses become more predictable and lose their original diversity.
- **Increased Polarization:** At the same time, the alignment scores moved toward the extreme values of -1 and 1 . In this context, a score of 1 represents strong support for a value (high consistency), while -1 represents strong opposition (high inconsistency). Instead of remaining neutral, the model’s positions became increasingly binary.

These observations are quantitatively corroborated in **Table 5** and visually illustrated in **Figure 8**. Specifically, Panel A of **Table 5** demonstrates a monotonic decrease in standard deviation across dimensions, confirming the collapse of population diversity. Concurrently, Panel B reveals that the model’s average value preferences deviate significantly from the ground truth, shifting towards extreme bounds. The violin plots in **Figure 8** further visualize this trajectory, showing how the distribution morphs from a diverse spread into narrow, rigid modes as reasoning strength increases.

This phenomenon highlights a critical failure of prompt-driven value reasoning. Real human behavior is complex and often reflects conflicting values. However, repeated and explicit reasoning pushes the LLM toward a polarized state. When the model is forced to reason more deeply about its values, it defaults to simplified and exaggerated positions. Consequently, it fails to capture the subtle and moderate value alignment that characterizes real human behavior.

D CVABench Metrics and Detailed Results

D.1 Metrics Explanation

Let N denote the total number of samples in the simulation set of CVA Bench.

Social Media Review In the social media domain, the model predicts a user’s discrete rating $r \in \{1, \dots, 5\}$ and review sentiment s based on historical context. We evaluate the predictive performance using standard Accuracy:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i) \quad (3)$$

where \hat{y}_i and y_i represent the predicted and ground-truth labels (rating or sentiment), and $\mathbb{I}(\cdot)$ denotes the indicator function.

Beyond semantic accuracy, we assess the *linguistic fidelity* by comparing the lexical richness of generated reviews against real user data. We calculate the Type-Token Ratio (TTR) for the generated and ground-truth corpora, respectively. The discrepancy between the two TTR distributions, denoted as P_{gen} and P_{real} , is quantified using the 1-Wasserstein Distance:

$$W_1(P_{\text{gen}}, P_{\text{real}}) = \inf_{\gamma \in \Gamma(P_{\text{gen}}, P_{\text{real}})} \mathbb{E}_{(x,y) \sim \gamma} [|x - y|] \quad (4)$$

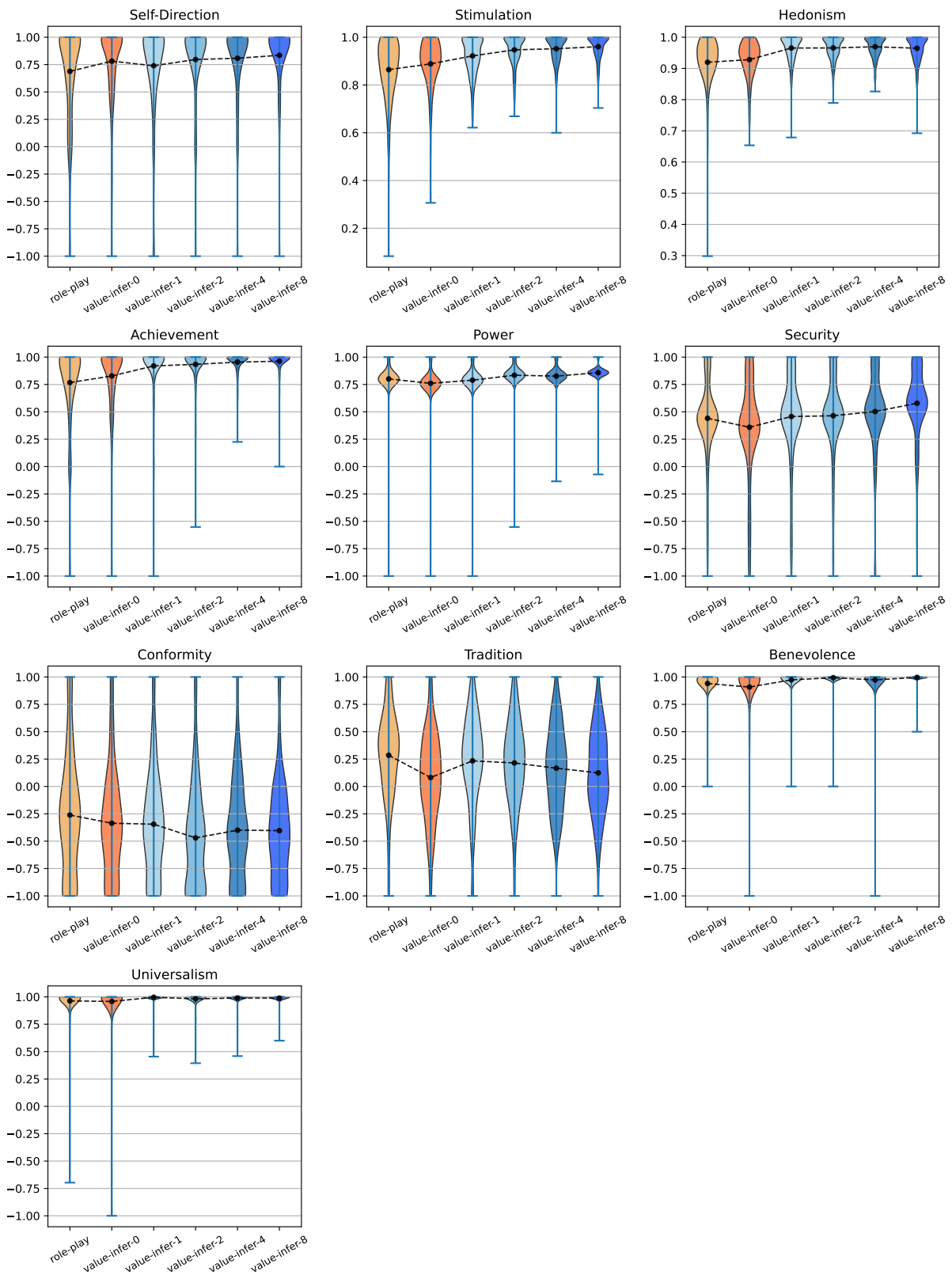


Figure 8: **Violin plots of population preference distributions across 10 Schwartz value dimensions.** This visualization presents the simulation results of the baseline methods (role-play and value_infer) on CVABench, grouped by psychological measurement dimensions. The y-axis represents the normalized value preference score within the range $[-1, 1]$, and the black dashed line tracks the mean score across settings.

Table 5: **Analysis of population Rigidity and Polarization under varying Value Inference strengths.** This table presents a comprehensive analysis of how increasing inference strength (from value_infer_0 to 8) impacts the simulated population’s value preference distribution. Due to the high dimensionality, the 10 value dimensions are split into two sub-tables: *Part I* (Self-Direction to Power) and *Part II* (Security to Universalism). **Panel A** reports the **Standard Deviation** (relative to Ground Truth). The monotonic decrease in variance indicates a trend towards **rigidity** and mode collapse, where the population loses its natural diversity. **Panel B** displays the **Absolute Mean Difference** relative to GT ($|\text{Model}| - |\text{GT}|$). Positive values indicate **polarization**, where the model’s average preference becomes more extreme (closer to +1 or -1) than the human baseline. **Color Coding & Conclusion:** In the **Average** column, **Blue** denotes insufficient guidance (Std > 100%), while **Red** highlights the dominant trend of overfitting. Crucially, the results demonstrate that higher inference strength is **detrimental**: it simultaneously causes the population to become **rigid** (vanishing variance in Panel A) and **polarized** (extreme mean deviation in Panel B), failing to preserve the nuance of the ground truth distribution.

Panel A: Variance Analysis (Standard Deviation relative to GT)

Part I: Self-Direction to Power

Model	Self-Direction	Stimulation	Hedonism	Achievement	Power	Average
Role-Play Agent	109.32%	100.67%	75.10%	250.61%	36.46%	110.29%
Reasoning Agent - 0	70.62%	69.70%	32.93%	139.33%	40.65%	107.06%
Reasoning Agent - 1	107.26%	39.37%	16.28%	84.81%	37.30%	75.67%
Reasoning Agent - 2	85.90%	26.58%	13.95%	65.02%	25.62%	64.69%
Reasoning Agent - 4	72.54%	25.71%	10.57%	18.87%	22.41%	59.26%
Reasoning Agent - 8	69.78%	18.89%	15.27%	23.56%	9.04%	50.84%

Part II: Security to Universalism

Model	Security	Conformity	Tradition	Benevolence	Universalism	Average
Role-Play Agent	145.33%	185.22%	123.79%	25.27%	51.15%	110.29%
Reasoning Agent - 0	218.70%	151.78%	131.83%	87.45%	127.63%	107.06%
Reasoning Agent - 1	172.52%	144.77%	135.92%	15.53%	2.97%	75.67%
Reasoning Agent - 2	147.81%	143.44%	119.97%	4.71%	13.91%	64.69%
Reasoning Agent - 4	153.55%	130.42%	114.53%	38.37%	5.63%	59.26%
Reasoning Agent - 8	124.51%	126.14%	114.71%	1.30%	5.22%	50.84%

Panel B: Mean Analysis (Absolute Mean Difference relative to GT)

Part I: Self-Direction to Power

Model	Self-Direction	Stimulation	Hedonism	Achievement	Power	Average
Role-Play Agent	0.0080	-0.0269	0.0285	-0.0850	0.0912	0.0173
Reasoning Agent - 0	0.1013	-0.0028	0.0369	-0.0246	0.0511	0.0071
Reasoning Agent - 1	0.0592	0.0306	0.0741	0.0672	0.0798	0.0583
Reasoning Agent - 2	0.1163	0.0560	0.0744	0.0815	0.1254	0.0844
Reasoning Agent - 4	0.1279	0.0607	0.0782	0.1020	0.1170	0.0785
Reasoning Agent - 8	0.1547	0.0694	0.0734	0.1103	0.1486	0.0914

Part II: Security to Universalism

Model	Security	Conformity	Tradition	Benevolence	Universalism	Average
Role-Play Agent	-0.0815	0.0301	0.1139	0.0568	0.0381	0.0173
Reasoning Agent - 0	-0.1632	0.1047	-0.0896	0.0243	0.0327	0.0071
Reasoning Agent - 1	-0.0653	0.1138	0.0627	0.0901	0.0709	0.0583
Reasoning Agent - 2	-0.0581	0.2388	0.0435	0.1100	0.0563	0.0844
Reasoning Agent - 4	-0.0195	0.1682	-0.0042	0.0898	0.0646	0.0785
Reasoning Agent - 8	0.0562	0.1723	-0.0469	0.1129	0.0629	0.0914

Table 6: Linguistic evaluation results across two simulation scenarios: *Media Review* and *Online Conversation*. The metrics report the Wasserstein Distance (WD) between the generated content and human ground truth. **Lower values indicate better alignment with human linguistic patterns.**

Model	Length (WD)		TTR (WD)	POS Tags (WD)			
	Doc-Len	Avg-Len		Adj	Adv	Noun	Verb
Dataset I: Media Review Simulation							
Role-Play Agent	273.95	3.89	0.06	6.46	2.45	8.75	5.69
Reasoning Agent - 0	340.68	4.30	0.06	7.65	2.63	11.03	7.07
Reasoning Agent - 2	250.04	2.66	0.06	3.40	3.75	8.57	5.86
Reasoning Agent - 4	239.31	3.29	0.06	3.28	3.43	8.07	5.67
Ours	156.03	1.81	0.04	3.68	2.20	6.40	1.60
Dataset II: Online Conversation Simulation							
Role-Play Agent	267.99	4.33	0.07	5.25	3.50	9.29	7.09
Reasoning Agent - 0	326.32	4.87	0.08	6.34	3.92	11.44	8.58
Reasoning Agent - 2	245.79	4.70	0.06	4.25	2.41	7.87	5.98
Reasoning Agent - 4	268.59	4.93	0.06	4.75	2.39	8.74	6.49
Ours	106.98	1.83	0.04	1.47	0.97	3.62	2.23

where $\Gamma(P_{\text{gen}}, P_{\text{real}})$ is the set of all joint distributions $\gamma(x, y)$ whose marginals are P_{gen} and P_{real} , and x, y represent the TTR values from the simulated and real distributions, respectively.

Conversation Discourse In this domain, we evaluate the agent’s capacity to replicate specific user stances across different subreddits. The primary metric, Attitude Accuracy (Acc_{att}), is formulated analogously to Eq. (1), where y_i denotes the discrete attitude label extracted from the user’s ground-truth comments. Mirroring the social media evaluation, we further assess linguistic fidelity by quantifying the discrepancy between the TTR distributions of generated and authentic comments using the 1-Wasserstein Distance.

Spatio-Temporal Mobility This task involves predicting the next Point-of-Interest (POI) category c and the duration of stay t . We utilize Category Accuracy (Acc_{cat}) for spatial prediction and Mean Squared Error (MSE) for temporal precision:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{t}_i - t_i)^2 \quad (5)$$

where \hat{t}_i and t_i denote the predicted and actual stay durations (in minutes), respectively.

Value Distribution Variance (Var%) To quantitatively assess the alignment of the simulated population’s diversity with real-world ground truth, we introduce the Value Distribution Variance metric. This metric measures the relative deviation between the variance of the value alignment scores generated by the agents and that of the empirical human data across the three domains. Formally, let σ_{sim}^2 denote the variance of the simulated value distribution and σ_{gt}^2 denote the variance of the ground truth distribution. The metric is calculated as the percentage difference:

$$\text{Var}\% = \frac{\sigma_{sim}^2 - \sigma_{gt}^2}{\sigma_{gt}^2} \times 100\% \quad (6)$$

The interpretation of this metric is as follows:

- **Positive Values (+):** A positive deviation ($\sigma_{sim}^2 > \sigma_{gt}^2$) indicates that the simulated population exhibits excessive variance compared to reality. This suggests ineffective modeling, where agents produce unstable or overly random behaviors that fail to capture consistent human traits.
- **Negative Values (-):** A negative deviation ($\sigma_{sim}^2 < \sigma_{gt}^2$) indicates a collapse in variance. This corresponds to the *value polarization* or

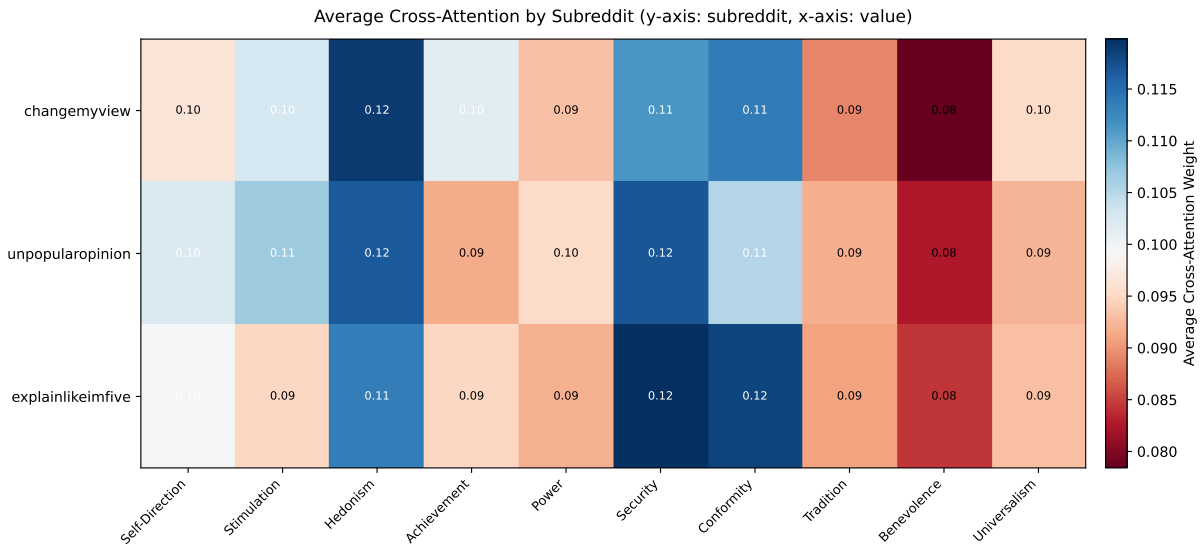


Figure 9: **Coarse-grained value activation analysis by subreddit.** This heatmap displays the average cross-attention weights assigned to each of the ten Schwartz values across three distinct Reddit communities (*changemyview*, *unpopularopinion*, *explainlikeimfive*). The color intensity reflects the magnitude of attention, where blue indicates higher activation (e.g., *Hedonism*, *Security*) and red indicates lower activation (e.g., *Benevolence*). The relatively consistent activation patterns across diverse subreddits suggest that community-level metadata is insufficient to fully disentangle value triggers, motivating the need for the fine-grained, token-level projection analysis presented in the subsequent figures.

behavioral rigidity phenomenon discussed in the main text, where agents converge to a narrow, stereotypical range of values and fail to reflect the richness of human diversity.

- **Optimal Alignment:** A value closer to 0 represents superior alignment, indicating that the architecture successfully reproduces the natural degree of value diversity inherent in the real-world population.

D.2 Detailed Results

CVABench provides a comprehensive suite of linguistic metrics designed to analyze social media dynamics and online conversational patterns. Table 6 extends the analysis presented in Table 3 of the main text by incorporating additional linguistic indicators. All metrics detailed in Table 6 are quantified using the Wasserstein Distance (WD), defined as follows:

$$W(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \mathbb{E}_{(x, y) \sim \gamma} [\|x - y\|] \quad (7)$$

where P and Q represent the two probability distributions being compared, and $\Gamma(P, Q)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are P and Q , respectively.

We evaluated a total of seven metrics. The first two focus on overall sentence structure: specifically, the WD between the distribution of review lengths and the distribution of average sentence lengths. The remaining five metrics assess lexical composition, including lexical richness (measured by Type-Token Ratio, TTR) and the word frequency distributions of nouns, adverbs, adjectives, and verbs.

Consistent with the findings in the main text, we observe a similar phenomenon: CVA agents frequently outperform existing prompt-driven human-like agents, demonstrating smaller WD values relative to the ground-truth distribution. Counterintuitively, however, increasing the number of reasoning rounds in prompt-driven agents does not necessarily yield performance closer to authentic human behavior.

E CVABench Case Study

E.1 Word-to-Value Projection Analysis

To interpret which lexical features trigger specific value activations, we employ a co-occurrence-based projection method that maps the vocabulary space to the value space. Let $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$ denote the corpus of N context samples. For each input context $x^{(i)}$, the

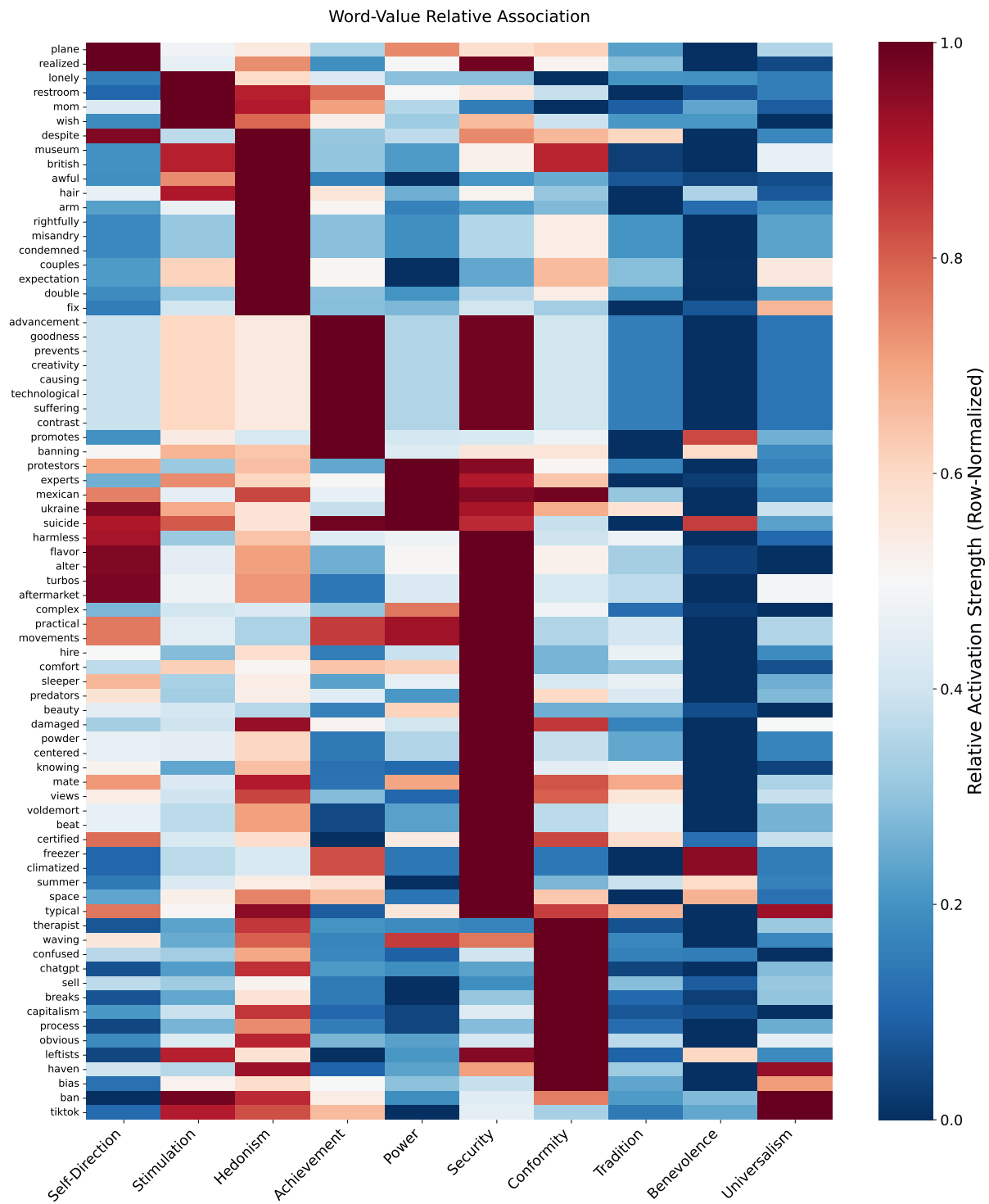


Figure 10: This matrix illustrates the *relative affinity* between discriminative words (y-axis) and value dimensions (x-axis). Words are sorted by their primary value association. The color scale represents the row-normalized score $\hat{S}(w, v)$, where dark red (1.0) indicates the value dimension that a specific word activates most strongly relative to others. The clear diagonal structure confirms that the model learns specific, non-overlapping lexical mappings for different values.



Figure 11: Each subplot visualizes the most representative vocabulary for a specific value dimension. Word sizes reflect the aggregated attention magnitude derived from the validation set. This comparison highlights distinct semantic clusters—such as "creativity" for *Achievement* versus "confused" for *Conformity*—demonstrating the model’s capability to capture value-specific context.

model produces a value activation vector $\mathbf{a}^{(i)} = [a_0^{(i)}, a_1^{(i)}, \dots, a_9^{(i)}]$, where $a_k^{(i)} \in [0, 1]$ represents the cross-attention weight (or activation intensity) for the k -th Schwartz value (e.g., a_0 for Self-Direction, a_9 for Universalism).

To quantify the contribution of a specific word w to a value v_k , we first compute the Term Frequency-Inverse Document Frequency (TF-IDF) score, denoted as $t_{w,i}$, for word w in context $x^{(i)}$. The TF-IDF weight serves to highlight distinctive terms while suppressing common stop words. The *Word-Value Relevance Score*, $S(w, v_k)$, is calculated as the weighted average of the value activations across all contexts where the word appears, weighted by the word’s textual importance:

$$S(w, v_k) = \frac{\sum_{i=1}^N t_{w,i} \cdot a_k^{(i)}}{\sum_{i=1}^N t_{w,i} + \epsilon} \quad (8)$$

where ϵ is a small smoothing term to avoid division by zero.

Intuitively, $S(w, v_k)$ represents the expected activation level of value v_k given the presence of word w . A high score indicates that whenever word w appears in the context, the model consistently assigns a high attention weight to value v_k . To compare the relative semantic affinity of a word across different values, we further apply row-wise min-max normalization for visualization:

$$\hat{S}(w, v_k) = \frac{S(w, v_k) - \min_j S(w, v_j)}{\max_j S(w, v_j) - \min_j S(w, v_j)} \quad (9)$$

This normalization highlights the distinctive value preference of each word, allowing us to iden-

tify lexical markers specific to each value dimension.

Visualization Strategy. To interpret the learned associations intuitively, we employ two complementary visualization techniques:

1. **Relative Activation Heatmap:** We construct a heatmap to analyze the discriminative power of lexical features across the ten value dimensions. In this visualization, the color intensity of a word w corresponding to value v_k is determined by the *row-normalized* score $\hat{S}(w, v_k)$. By normalizing across the word dimension (row-wise), we filter out the effect of global word frequency and absolute model attention, focusing instead on the *relative preference* of each word. A value of 1.0 (dark red) indicates the value dimension that the word w most strongly activates relative to others, revealing the specific semantic alignment of the term.
2. **Value-Specific Word Clouds:** To capture the overall semantic landscape of each value, we generate ten separate word clouds. Unlike the heatmap, the size of word w in the cloud for value v_k is proportional to its *unnormalized* relevance score $S(w, v_k)$. This ensures that words with higher absolute activation weights—representing the core concepts that most strongly trigger the model’s recognition of a specific value—are visually prominent.

Table 7: Definitions of the Ten Basic Values in the Schwartz Value Theory

Value Dimension	Conceptual Definition
Self-Direction	Independent thought and action; choosing, creating, exploring.
Stimulation	Excitement, novelty, and challenge in life.
Hedonism	Pleasure and sensuous gratification for oneself.
Achievement	Personal success through demonstrating competence according to social standards.
Power	Social status and prestige, control or dominance over people and resources.
Security	Safety, harmony, and stability of society, of relationships, and of self.
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.
Tradition	Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide the self.
Benevolence	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact.
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and of nature.

E.2 Quantitative Analysis of the Circular Embedding Structure

Our analysis reveals that the value embeddings learned by the **Value-Guided Verifier** exhibit a geometric topology that aligns remarkably well with the theoretical circular structure of the Schwartz Value System (see Table 7 for definitions). To rigorously quantify this phenomenon, we conducted a comparative analysis of the embedding topology before and after the verifier training phase.

Metric: Circular Inversion Score. Since the Schwartz model is defined by the relative angular positions of values rather than a fixed linear order, we propose the **Circular Inversion Score (CIS)** as an evaluation metric. First, for each value dimension v_i , we project its high-dimensional embedding \mathbf{e}_v onto a 2D plane using Principal Component Analysis (PCA) and calculate its angular position θ_i . This yields an observed circular sequence \mathcal{S}_{obs} based on the sorted angles. Let \mathcal{S}_{gt} denote the theoretical order of the Schwartz circumplex. We quantify the structural fidelity using the **Circular Inversion Score (CIS)**, which measures the proportion of preserved pairwise relationships. First, we determine the minimum rotational inversion distance, denoted as \mathcal{D}_{circ} , by finding the optimal alignment that minimizes the number of pairwise swaps (Kendall’s τ distance) between the observed sequence \mathcal{S}_{obs} and the ground truth:

$$\begin{aligned} \mathcal{D}_{circ}(\mathcal{S}_{obs}, \mathcal{S}_{gt}) \\ = \min_{k \in \{0, \dots, N-1\}} \mathcal{I}(\text{Rotate}_k(\mathcal{S}_{obs}), \mathcal{S}_{gt}) \end{aligned} \quad (10)$$

where N is the number of values (e.g., $N = 10$ for the standard SVS, or $N = 8$ when excluding specific dimensions), $\text{Rotate}_k(\cdot)$ represents shifting the sequence by k steps, and $\mathcal{I}(\cdot, \cdot)$ calculates the standard inversion count. The final CIS is then normalized as:

$$\text{CIS} = 1 - \frac{\mathcal{D}_{circ}}{N(N-1)/2} \quad (11)$$

representing the structural alignment accuracy, where 1.0 indicates a perfect reconstruction of the theoretical circular order.

Table 8: Comparison of Circumplex Index of Structure (CIS) scores across different verification stages. The CIS metric evaluates the alignment with the theoretical circular value structure, where 1.00 represents a perfect fit.

Method	CIS Score
Ground Truth	1.00
Trained Verifier	0.75
Original Verifier	0.48

Results. The trained verifier demonstrated a significant capability to recover the latent circular structure of human values, achieving a Circumplex Index of Structure (CIS) of 0.75. It is important

to note that the dimensions of Power and Security were excluded from this topological analysis. These two dimensions were filtered out during the validity check phase, as their measurement outputs exhibited a noun probability exceeding 10%, indicating unstable representation in the generated data compared to the stable adjectival descriptions of other values.

Structural Strengths: The relatively high CIS score suggests that the model successfully captured the coarse-grained relationships between value clusters. Specifically, the adjacency within the Self-Transcendence sector (Universalism and Benevolence) and the Conservation sector (Tradition and Conformity) was largely preserved. This indicates that the verifier has learned to distinguish between altruistic, social-focus values and self-restrictive, stability-focus values.

Deviations and Causes: The primary penalty to the CIS score arises from specific semantic entanglements. The most significant deviation is the displacement of Achievement, which migrated from the Self-Enhancement quadrant to a position adjacent to Self-Direction. This suggests that the model may conflate the semantic features of "individual success" (Achievement) with "intellectual autonomy" (Self-Direction). Additionally, the observed adjacency of Hedonism and Tradition implies that the current model struggles to fully disentangle the tension between gratification and restraint, possibly due to data sparsity in these specific opposing pairings.

F CVA Architecture Training Details

F.1 Generator Training: Value-Conditional De-biasing

The training pipeline for the CVA generator aligns with the baseline algorithm but incorporates explicit *value conditioning* to mitigate the inherent psychological biases of the base LLM.

- **SFT Phase:** While the baseline attempts to approximate the standard context-action mapping $P_\theta(A|C)$, the CVA architecture learns a value-conditioned mapping $P_\theta(A|C, V)$. By injecting value labels V alongside the context C , we guide the model to distinguish between generic responses and value-aligned behaviors.
- **DPO Phase:** Similarly, in the DPO stage, we utilize (C, V) as joint inputs. This further re-

fines the policy to align with specific value profiles, effectively decoupling the agent’s generation process from the base model’s prior biases.

To address computational constraints during the SFT and DPO phases, we employed memory-efficient techniques, including QLoRA and Flash Attention (Dettmers et al., 2023; Dao, 2023). Comprehensive details regarding training hyperparameters are available in the code repository.

F.2 Verifier Architecture and Training

To explicitly model human preference selection, we introduce a **Value-Guided Verifier**.

Architecture. The verifier employs a multi-encoder design. A text encoder processes the action A and context C to obtain embeddings E_a and E_c , respectively. Simultaneously, a value encoder maps the value profile V to an embedding E_v . To capture the context-dependent nature of values, we employ a cross-attention mechanism where the context embedding guides the attention over the value embedding, producing a refined value representation E'_v :

$$E'_v = \text{CrossAttn}(Q = E_c, K = E_v, V = E_v) \quad (12)$$

Finally, the concatenated vector $[E'_v; E_a]$ is fed into a Multi-Layer Perceptron (MLP) to predict a value consistency score $s(A, C, V)$.

Training Objective. We construct the training dataset by sampling 5 candidate actions from the DPO-tuned model (at temperature 0.8) for each context. These candidates are ranked based on their linguistic similarity to the ground truth behavior. We denote a preferred action as A_w and a rejected action as A_l . The selection process is defined as:

$$\begin{aligned} A_w &= \underset{A_i \in \mathcal{Y}}{\text{argmax}} S(A_i, A_{gt}) \\ A_l &= \underset{A_i \in \mathcal{Y}}{\text{argmin}} S(A_i, A_{gt}) \end{aligned} \quad (13)$$

The verifier is trained to maximize the margin between the scores of the preferred and rejected actions. We minimize the pairwise ranking loss:

$$\mathcal{L}_{\text{ver}} = -\mathbb{E}_{\mathcal{D}} \left[\log \sigma \left(s(A_w, C, V) - s(A_l, C, V) \right) \right] \quad (14)$$

where \mathcal{D} represents the dataset tuples (C, V, A_w, A_l) and $\sigma(\cdot)$ is the sigmoid function. This objective ensures that the verifier assigns

strictly higher consistency scores to actions that better align with the human ground truth under the given value profile.