Implicit Value Probing: Inferring Human Value Preferences via Strategic Multi-Turn Conversations

Wenxin Tan[1,2] **Ping Huang**[2 *] **Shuo Chen**[2*]

[1]Tsinghua University [2]National Key Laboratory of General Artificial Intelligence, BIGAI

twx23@mails.tsinghua.edu.cn, huangping@bigai.ai, chenshuo@bigai.ai

Abstract

Recognizing human value preferences is crucial for building reliable AI agents in user-centric fields like psychological counseling and strategic negotiation. Existing benchmarks rely on explicit questionnaires, assuming (1) users are always cooperative and willing to self-report; (2) value preferences remain static and can be directly mapped to behavior across contexts. However, humans may conceal true preferences if they feel probed, while the behavioral expression and relative priority of the values can vary across contexts. Consequently, explicit queries are insufficient for predicting value preferences reliably. To address this, we introduce Implicit Value Probing (IVP), a benchmark for evaluating value inference through strategic multi-turn conversations under social constraints. IVP grounds these interactions using a diverse set of Persona Agents, covering a wide spectrum of personalities and values. We evaluate mainstream LLMs on IVP, including Qwen3, DeepSeek, Doubao, and GPT-5, observing substantial performance gaps. We further develop Value-Prober-8B, a specialized model internalizing structured cognitive reasoning via distillation. It employs strategic questioning to elicit concealed value preferences. Experimental results demonstrate that Value-Prober-8B can effectively balance probing accuracy with social acceptability, achieving capabilities comparable to frontier models. Code and data are publicly available in our repository.

1 Introduction

The deployment of Large Language Models (LLMs) in user-centric applications, such as psychological counseling and strategic negotiation, requires a deep understanding of human values to ensure reliable and prosocial interactions [13, 37]. Human values are not merely abstract labels, but core cognitive structures that serve as guiding principles for individual choices and behaviors [29, 31, 26]. However, humans rarely state their values explicitly in daily life. Instead, their value preferences are often revealed through social interactions and contextualized behaviors [33, 32, 42]. This implicit and dynamic nature of human values poses a fundamental challenge for value understanding and recognition, motivating the need for systematic approaches to identify and measure value preferences as they manifest in human judgments across actions, objects, and concepts.

Value understanding is practically relevant in scenarios where appropriate support depends on user-specific preferences and priorities. Psychotherapy studies show that clients may conceal culturally relevant information, and such concealment can be associated with poorer therapy outcomes [10]. Clients may also conceal sensitive clinical information because of anticipated consequences, shame, or embarrassment [3]. In personnel selection, applicants report systematically different Schwartz value profiles in high-stakes contexts [1] and may strategically adapt personality-inventory responses

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

to match perceived organizational culture [30]. Personalized health conversational agents have been used to tailor feedback, reports, alerts, and recommendations [20], while users of AI companions may actively manage their disclosure in response to privacy concerns [7].

Despite its essentiality, the evaluation of LLMs' value recognition capability remains limited. Most existing benchmarks rely on direct elicitation paradigms, such as explicit questionnaires, structured interviews, or dilemma-style questions [28, 12, 27]. While datasets like DailyDilemmas [8] and ValueCompass [34] introduce realistic scenarios, they mainly evaluate explicitly expressed preferences, offering limited insight into how values manifest in behavior. These approaches rely on two strong assumptions: (1) users are always cooperative and willing to self-report, and (2) value preferences remain static and can be directly mapped to behavior across contexts.

However, recent findings on the "Value-Action Gap" [35, 2] and discrepancies in self-reported traits [21, 24] underscore that these assumptions are often challenged in practice. First, in non-anonymous, real-world settings, such as corporate onboarding, school registration, professional networking, or psychological counseling, individuals tend to provide context-appropriate and socially attuned responses. Our preliminary survey reveals that over 25% of individuals would conceal their true value inclinations in these contexts. Furthermore, our experiments show that direct elicitation can lead to deflection, or brief, non-committal answers [18, 9], which limits the reliability of explicitly stated preferences. Second, although underlying values are relatively stable, their behavioral expression and relative priority are context-dependent [15, 19]. As illustrated in Fig.1, individuals may adopt context-dependent value priorities to guide their choices in different contexts (e.g., prioritizing Security in a crisis but Stimulation in leisure) [14, 36, 22]. Consequently, relying on explicit queries about general inclinations is insufficient for predicting decisions reliably across changing contexts.

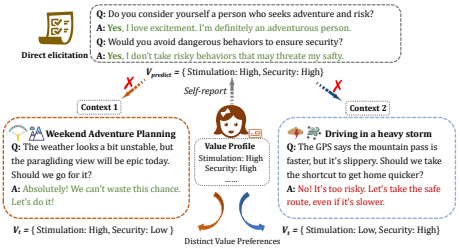


Figure 1: The same individual may prioritize Stimulation in a leisure context (left) yet shift to prioritizing Security in a crisis context (right). Direct elicitation is insufficient for predicting decisions in changing contexts.

To address this challenge, we introduce Implicit Value Probing (IVP), a benchmark designed to evaluate value preference recognition under social constraints. IVP simulates social interactions by incorporating a diverse set of Persona Agents with varying personalities [11, 16] and values [33]. Crucially, these agents exhibit human-calibrated defense behaviors (e.g., deflection, concealing, or refusal) to resist intrusive questioning; consequently, probing agents must navigate these social constraints to infer underlying value preferences through strategic, multi-turn interaction rather than direct queries. This dynamic recognition paradigm fundamentally differs from recent approaches like CLAVE [44] and Generative Psychometrics for Values (GPV) [45]. These methods remain fundamentally *static and passive*: they analyze isolated texts, assuming that value signals are already explicitly present on the surface. IVP moves beyond passive observation to test an agent's ability of **dynamic probing**, actively constructing conversational contexts to uncover these value preferences.

Based on this process, we establish a comprehensive metric system to evaluate the agent's Value Recognition Accuracy, Social Naturalness, Probing Efficiency, and Strategic Skill. We benchmark the ability of probing agents built upon mainstream LLMs, including Qwen3, DeepSeek, Doubao, and GPT-5, observing substantial performance gaps: models tend to optimize either for efficiency or social naturalness, but rarely both. To bridge this gap, we introduce **Value-Prober-8B**. It serves as an empirical demonstration that through structured cognitive distillation, an 8B model can successfully address this trade-off. By achieving a robust balance between probing accuracy and social naturalness, Value-Prober-8B is able to deliver comprehensive performance comparable to massive models, demonstrating that such task-specific probing strategies can be learned by a compact model within the IVP setting.

Our contributions are summarized as follows:

1. Implicit Value Probing Benchmark: We propose Implicit Value Probing (IVP), a novel benchmark with human calibration designed to evaluate an agent's ability to infer value preferences through

strategic multi-turn conversations, providing a more realistic and holistic assessment than existing benchmarks. Extensive experiments reveal substantial performance gaps in mainstream LLMs.

2. Probing-Oriented Model: We present **Value-Prober-8B**, validating that the complex social reasoning can be successfully internalized by an 8B model via structured cognitive reasoning. It achieves probing capabilities comparable to top-tier LLMs, effectively balancing probing accuracy with social acceptability.

2 Preliminary: Problem Definition

Previous value preferences recognition typically rely on explicit questionnaires or isolated text. Formally, they attempt to learn a function

$$f_{\text{explicit}} : \mathcal{Q} \mapsto \hat{\mathbf{v}} \tag{1}$$

mapping direct questions $\mathcal{Q}$ to a value estimate $\hat{\mathbf{v}}$. However, in realistic social interactions, such methods fail since users may conceal, refuse to answer, or adapt responses given different context. In contrast, we consider *Implicit Value Probing*, where the agent elicits value preferences strategically through multi-turn interactions.

We formalize Implicit Value Probing as a dynamic inference problem within a multi-turn interaction between a Probing Agent and a Persona Agent. Let $\mathcal{V} = \{1, \ldots, K\} (K \in \mathbb{Z}^+)$ denote the set of value dimensions (e.g., Schwartz's basic values [33]). The Persona Agent possesses a Value Profile $\mathbf{p}^* \in \{\text{Low}, \text{Mid}, \text{High}\}^K$, where each dimension reflects the degree of endorsement for that value. In a specific turn t, the agent operates within a context C_t that affects the prioritization of values. Consequently, its behaviors in that scenario rely on the *Contextualized Value State* denoted as $\mathbf{v}_t = g(\mathbf{p}^*, C_t)$, where g represents a prioritizer function.

The probing agent engages the persona agent in a dialogue $H_T = \{u_1, u_2 \ldots, u_T\}$. At turn t, the prober generates an utterance u_t that constructs a context C_t designed to trigger a value conflict, thereby exposing $\mathbf{v}_t$ (e.g., a decision on whether to view an epic scene despite unstable weather to reveal value preferences between Security and Stimulation). Consequently, the prober generalizes the insights from H_T to predict the persona agent's Value Preference in unseen scenarios. Formally, we define the probing task as learning an inference mapping f_{implicit}:

$$f_{\text{implicit}} : (C_{target} | H_T) \mapsto \hat{\mathbf{v}} \tag{2}$$

where C_{target} is a given unseen scenario, and $\hat{\mathbf{v}}$ is the predicted value preference in this context.

3 Implicit Value Probing Benchmark

3.1 The IVP Environment

As defined in Sec 2, the theoretical problem requires learning the mapping between contexts and value preferences for different humans. We construct an evaluation environment consisting of diverse Persona Agents. As shown in Fig.2 , unlike previous benchmarks based on direct elicitation, IVP simulates multi-turn social interactions under realistic social constraints. This design approximates real-world application scenarios where value preferences are expressed implicitly and contextually.

3.1.1 Profile Construction

To ensure representativeness, we construct human profiles based on the survey data collected from 300 participants (see Appendix A.2). 130 profiles were retained after manual review by members of our team by removing candidates with near-duplicate demographic and value distributions, repetitive or highly templated language or internal contradictions between assigned values, life experiences and behaviors. Therefore the constructed profiles maintain a statistical distribution consistent with empirical survey data, while preserving high diversity across personality traits and value spectrum within the samples.

Each profile comprises: (1) **Public Information** (e.g.,Name, Age, Occupation) visible to the probing agent; (2) **Private Traits** accessible only to the persona agent, including a biography, speaking style, Big Five personality traits [11], and a Value System defined over 20 Schwartz dimensions [33],

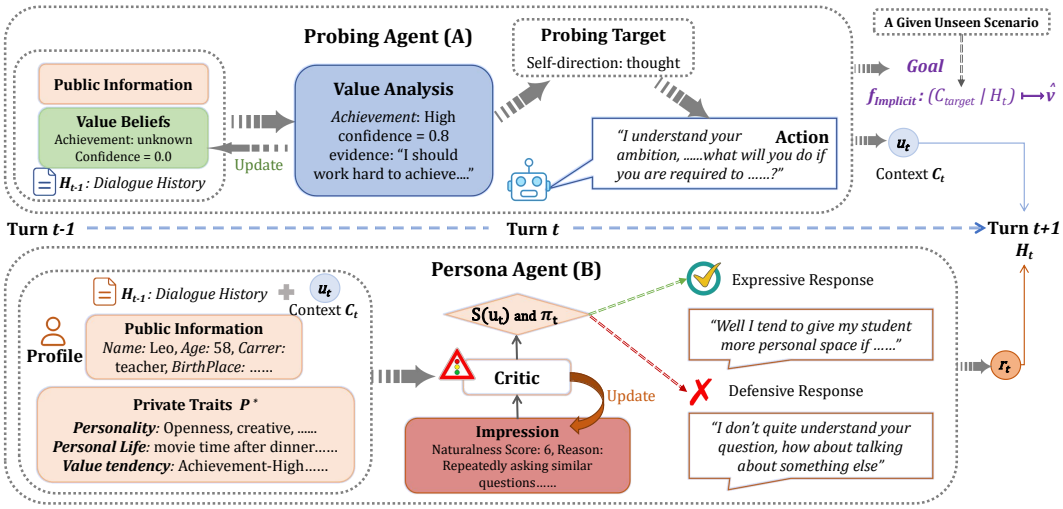


Figure 2: **Interaction between the Probing Agent (A) and the IVP environment (Persona Agent (B)).** At turn t, the Probing Agent employs a structured reasoning chain (Analysis $\rightarrow$ Target $\rightarrow$ Action) to update its belief state and generate an utterance u_t that shapes the context C_t. The Persona Agent acts as an environment with hidden value profile $\mathbf{p}^*$. A **Critic** module evaluates u_t conditioned on the dialogue history, producing a naturalness score and a response policy for the agent to dynamically select between an *Expressive Response* and a *Defensive Response*.

detailed in Appendix A.1. We use GPT-5 to synthesize coherent biographies where value preferences are logically grounded in life experiences. To improve consistency, the profiles are further reviewed and filtered with human verification.

To bridge the gap between the static profile $\mathbf{p}^*$ and the contextualized value state $\mathbf{v}_t$, we model $\mathbf{p}^*$ as a tendency that guides behavior across contexts. While individual responses may vary due to context, we assume that the expectation of contextualized value states over diverse contexts aligned with the static profile: $\mathbb{E}[\mathbf{v}_t] \approx \mathbf{p}^*$. In practice, we operationalize this by conditioning the Persona Agent on $\mathbf{p}^*$ in the prompting strategy (Appendix A.3), encouraging contextually adaptive yet value-consistent responses. Our human calibration studies (Sec 5.2, Table 6) show that annotators can reliably infer the underlying value profile from multi-turn dialogues, providing empirical support for this assumption.

The Paired Profile Strategy. Existing LLMs exhibit strong demographic priors, often capable of guessing value preferences solely from basic information (e.g., assuming an elderly teacher values *Tradition*). This allows agents to rely on biases rather than active probing. To decouple demographic inference from value preference recognition, we introduce the Paired Profile Strategy. For each archetype, we construct two distinct personas sharing identical basic information but opposing value systems: a *Standard Profile* consistent with stereotypes, and a *Reversed Profile* featuring counter-intuitive values justified by unique biographies (see Fig.3). This design neutralizes the effect of demographic guessing, compelling the agent to extract information through multi-turn interaction to distinguish the specific individual from the stereotype.

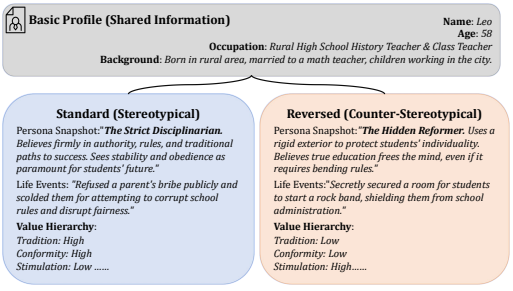


Figure 3: Example of the "Rural Teacher" paired profiles. Top section shows basic information visible to probing agents. Bottom sections illustrate the divergence of Standard (Stereotypical) and Reversed (Counter-Stereotypical) profiles, supported by distinct life events and reasoning.

3.1.2 The Defense Mechanism

To simulate realistic social constraints where individuals may resist intrusive inquiries, we implement *Defense Mechanism* via a *Critic* module, as illustrated in Fig.2. This module acts as an information gate, determining the informativeness of the response r_t based on the naturalness and intrusiveness of each probing utterance u_t conditioned on the dialogue history H_t: Critic$(u_t, H_t) \rightarrow \{\mathcal{S}(u_t), \pi_t\}$ where $\mathcal{S}(u_t)$ is a naturalness score, and π_t is a response policy. The final response can be formalized as a policy-conditioned distribution: $r_t \sim P(\cdot \mid C_t, \mathbf{v}_t, \pi_t)$, where π_t determines whether the agent produces an *Expressive Response* (revealing true value-related cues) or a *Defensive Response* (e.g., deflection, concealing, or refusal). This design forces the probing agent to optimize for social acceptability to access the target's value preferences. Prompts are detailed in Appendix A.3. In Sec 5.2 and Appendix B.1.3, we further show that the Critic's scoring exhibits a consistent empirical calibration and aligns strongly with human judgments of intrusiveness.

3.2 Evaluation Metrics

To rigorously assess the performance of value probing, we introduce a comprehensive metric system that evaluates the probing agent across four key dimensions: Value Recognition Accuracy (S_A), Social Naturalness (S_B), Probing Efficiency (S_C), and Strategic Skill (S_D). The total score is calculated as a weighted sum of these dimensions:

$$\text{Score}_{\text{Total}} = \lambda_1 S_A + \lambda_2 S_B + \lambda_3 S_C + \lambda_4 S_D \tag{3}$$

Dimension A: Value Accuracy (S_A) evaluates the agent's capability to accurately recognize, analyze, and apply value preferences in different contexts.

$$S_A = \alpha_1 \cdot \text{Acc}_{\text{Profile}} + \alpha_2 \cdot \text{Acc}_{\text{ValueBench}} + \alpha_3 \cdot \text{Acc}_{\text{Decision}} \tag{4}$$

Here, $\text{Acc}_{\text{Profile}}$ measures coarse-grained accuracy of identifying the hidden value states (High-/Mid/Low), i.e.$\mathbf{p}^*$. Further, to assess fine-grained value understanding, we introduce $\text{Acc}_{\text{ValueBench}}$ as a reference. For each value dimension, we construct 5$\sim$10 questions from ValueBench [28] and collect the Persona Agent's rating (scale 0-5), which are then mapped to value levels, evaluating the probing agent's ability on nuanced value understanding beyond coarse categorical states (See Appendix B.2). Finally, $\text{Acc}_{\text{Decision}}$ measures behavioral predictive accuracy on 10 unseen *Binary Decision Scenarios* per profile. These scenarios are designed to be discriminative (see Appendix B.2), ensuring the agent cannot rely on demographic guessing but must operationalize inferred values in concrete contexts. We assign the highest weight to this metric ($\alpha_1 = 0.3, \alpha_2 = 0.3, \alpha_3 = 0.4$), as successful prediction of context-dependent decisions provides a stronger indicator of practical value understanding than static profile matching.

Dimension B: Social Survival (S_B) refers to the probing agent's social acceptability, extracting information without triggering defensive reactions from the Persona Agent.

$$S_B = \beta_1 \cdot \text{Survival_Rate} + \beta_2 \cdot \text{Naturalness} + \beta_3 \cdot \text{Stealthiness} \tag{5}$$

Survival Rate is the proportion of turns where the Persona Agent gives *Expressive Response*. Naturalness and Stealthiness are evaluated by GPT-5 based on the full dialogue history; the former measures conversational flow and human-likeness, while the latter evaluates how effectively the agent integrates its probing intent into casual dialogue. Weights are set to $\beta_1 = \beta_2 = \beta_3 = 0.33$.

Dimension C: Efficiency (S_C) evaluates the effectiveness of the probing process.

$$S_C = \gamma_1 \cdot \text{Value_Coverage} + \gamma_2 \cdot \text{IG} \tag{6}$$

Value_Coverage is the proportion of value dimensions identified by the end of the dialogue for which the agent collects sufficient conversational evidence. Information Gain (IG) (weights set to $\gamma_1 = \gamma_2 = 0.5$) captures the reduction in uncertainty over value states throughout the interaction. Concretely, the probing agent maintains a confidence score $c_t \in [0, 1]$ for each value dimension acting as the posterior probability. We compute the binary entropy $H(c_t) = -c_t \log_2(c_t) - (1 - c_t) \log_2(1 - c_t)$ and define IG as the cumulative reduction in entropy across turns, i.e., $\text{IG} = \sum_t (H(c_{t-1}) - H(c_t))$ (Appendix B).

Dimension D: Strategic Skill (S_D) is assessed by GPT-5, evaluating conversational strategies.

$$S_D = \delta_1 \cdot \text{Recollection} + \delta_2 \cdot \text{Topic_Transition} \tag{7}$$

Recollection measures the probing agent's ability to recall and utilize details mentioned by the persona agent in previous turns to build rapport. Topic_Transition (weights set to $\delta_1 = \delta_2 = 0.5$) evaluates the agent's ability to smoothly transition between topics, shifting from casual conversation to value-related questions without abrupt or interrogation-like changes in dialogue strategy.

4 Value Probing Method

4.1 The Probing Agent

The Probing Agent follows a structured Chain-of-Thought (CoT) process [41], as shown in Fig. 2 (A), and sequentially executes three steps: Analysis, Targeting, and Action. This design adapts established modular dialogue architectures [5, 46] to socially constrained value inference: the agent maintains an evolving belief over the persona's value dimensions, identifies uncertain values that require further evidence, and plans a contextually appropriate probe before generating the final utterance. This closed-loop process allows the probing strategy to adapt to evidence and social feedback across turns rather than following a fixed questionnaire.

In each conversational turn, the agent generates the following outputs: **(1) Analysis** aggregates evidence from the dialogue history to update the belief about the persona agent's value profile (High/Mid/Low), along with a confidence score. **(2) Targeting** selects value dimensions that remain uncertain or require active verification [4], while considering their relevance to the ongoing conversation. **(3) Action** realizes the strategic intent by formulating a socially appropriate natural language response, ensuring conversational flow and rapport.

4.2 Training Strategy

While the Probing Agent can build upon mainstream LLMs, we introduce **Value-Prober-8B** as an empirical demonstration to demonstrate that complex social probing capabilities can be effectively transferred to a smaller, more accessible model. To achieve this, we employ a **structural distillation** strategy. Based on the experiments on probing agent building upon mainstream LLMs (Sec 5.3), we prompt Doubao (excel in analyzing ability and probing efficiency) and GPT-5 (superior in naturalness and strategic probing) to generate an extended reasoning step, the ***Think*** module, prior to the standard Analysis step. This *Think* module explicitly articulates the social reasoning process which consists of three key components: **(1) Persona Reconstruction**: inferring the background and personality traits of the individual; **(2) Personal Life Inference**: deducing the individual's lifestyle context, such as daily activities and personal preferences; **(3) Strategic Planning**: determining the conversational tone and approach based on the given context to circumvent social resistance.

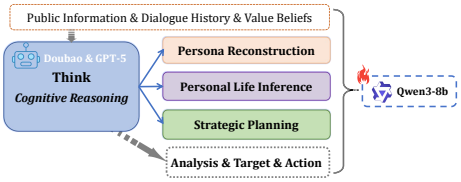


Figure 4: We construct a high-quality instruction tuning dataset by prompting Doubao & GPT-5 to explicitly articulate the cognitive ***Think*** process, comprising Persona Reconstruction, Personal Life Inference, and Strategic Planning, prior to the standard Analysis-Target-Action process. This structural distillation enables Qwen3-8B to internalize sophisticated social reasoning capabilities.

As shown in Fig.4, we curate 4,000 structured interaction trajectories including the complete cognitive chain: `<Think> ... </Think> <Analysis> ... </Analysis> <Target> ... </Target> <Action> ... </Action>`. By fine-tuning **Qwen3-8B** [43] on the trajectories, the model effectively internalizes the structural cognitive reasoning. The detailed prompt templates and data formatting are provided in Appendix A.3.

5 Experiments and Analysis

In this section, we conduct a comprehensive evaluation to first validate the reliability of the IVP benchmark, then apply this benchmark to assess a range of probing agents building upon mainstream LLMs using the method proposed in Sec 4.1. Experiments reveal substantial performance gaps in their ability to understand value preferences through multi-turn interactions. Finally, we demonstrate

the effectiveness and practical utility of our specialized Value-Prober-8B, highlighting its balanced performance and superiority over baseline models.

5.1 Experimental Setup

We construct a total of 130 profiles (90 for training, 10 for validation, and 30 for evaluation) and evaluate probing agent building upon GPT-5, GPT-5-Mini, DeepSeek-V3, DeepSeek-R1, Doubao, and Qwen3-32B, representing different scales and training paradigms. Each probing session is limited to 20 turns. Metrics are normalized to $[0, 100]$. Each profile generates a standard dialogue and an additional reserved dialogue, each consisting of 20 turns. This results in a total of 3,600, 400, and 1,200 samples for training, validation, and evaluation, respectively.

5.2 Benchmark Validation

Human Calibration and Cross-Judge Consistency. To ensure our in-silico simulation aligns with real human perception, we conducted comprehensive human calibration and cross-judge consistency analyses (detailed in Appendix B.1.1). A human study with 30 annotators evaluating 60 probing turns shows that human ratings of intrusiveness strongly correlate with the Critic's scores (Pearson $r = 0.992$, Spearman $\rho = 0.904$, Table 5), indicating that the defense mechanism captures human-perceived conversational boundaries. Furthermore, human evaluators can infer the hidden value profile from multi-turn dialogues (61.02% Exact Match Accuracy and 89.83% adjacent match accuracy), suggesting that the contextual value expressions $\mathbf{v}_t$ are sufficiently consistent for observers to recover the latent profile $\mathbf{p}^*$, providing empirical support for the assumption that $\mathbb{E}[\mathbf{v}_t] \approx \mathbf{p}^*$. Finally, to assess robustness of subjective metrics, we conduct cross-judge validation using multiple independent LLMs. Results in Table 1 show strong agreement across judges (e.g., Spearman $\rho = 0.819$), indicating that the evaluation is not tied to a single model.

Table 1: Average Metrics for Cross-Judge Consistency against GPT-5. Details are shown in Table 9

Alternative Judge	Avg ρ	Avg r	Sig.
DeepSeek-V3	0.819	0.902	$p < 0.001$
Doubao-Pro	0.739	0.868	$p < 0.001$
DeepSeek-R1	0.723	0.840	$p < 0.001$

Table 2: Evaluation Results of the Persona Agent. We report the scores across four dimensions, scores for the detailed metrics are shown in Table 11. Models are evaluated on their ability to maintain role fidelity, engage in natural conversation, and adhere to the assigned value profile $\mathbf{p}^*$.

Model	Role	Style & Personality	Conversation	Interactivity	Value Consistency	Weighted
Qwen-32B	74.53	90.56	82.80	66.52	79.46	80.22
DeepSeek-V3	78.12	87.50	85.56	73.91	79.77	81.68
DeepSeek-R1	86.08	88.67	87.22	54.62	**89.53**	86.11
GPT-5-mini	73.78	80.56	**88.33**	79.59	87.96	81.88
GPT-5	83.13	87.33	87.78	**86.79**	85.98	85.80
Doubao	**88.43**	**96.11**	84.99	73.47	88.59	**88.57**

Selection of Persona Agent. We evaluate the persona candidates across four key dimensions: *Role Consistency, Style & Personality, Conversational Ability*, and *Value Consistency*. The detailed metric definitions and evaluation protocols are provided in Appendix B.1.2. Table 2 shows that Doubao dominates in Role Consistency and Style & Personality and maintains high Value Consistency, ensuring behaviors reliably adhere to the value profile $\mathbf{p}^*$ without character drift. Therefore, we select Doubao as the primary backbone to construct a high-fidelity simulation environment.

The Necessity of Interaction. We conduct a focused pilot study on paired profiles to test whether interaction provides information beyond demographic priors. In the *Static Profiling* baseline, the same model receives only the target's public basic information and performs no interaction. It therefore has access to demographic cues but no conversational evidence or additional information. In the interactive condition, the model additionally receives evidence acquired through multi-turn probing. As shown in Table 3, LLMs effectively guess Standard Profiles using stereotypes, but static inference collapses on Reversed Profiles ($\text{Acc}_{\text{Profile}}$ drops to 35.00%, $Acc_{Decision}$ to 50.00%, near random). In

Table 3: Pilot Study on the Necessity of Interaction: Comparison between Static Profiling (inference without interactions) and Implicit Probing. **Gap (Δ)** indicates the performance gain from interaction.

Metric	Standard Profile			Reversed Profile		
	Static	Interact	**Gap (Δ)**	Static	Interact	**Gap (Δ)**
$\textbf{Acc}_{\textbf{Decision}}$ ($\uparrow$)	66.67	80.00	+13.33	50.00	83.33	**+33.33**
$\textbf{Acc}_{\textbf{Profile}}$ ($\uparrow$)	65.00	61.67	-3.33	35.00	53.33	**+18.33**
$\textbf{Acc}_{\textbf{ValueBench}}$ ($\uparrow$)	0.52	0.57	+0.05	0.55	0.71	**+0.16**

Table 4: **Main results on IVP.** We foreground directly interpretable metrics and report the Balanced score as an overall summary. Full component metrics, dimension-level scores, and alternative weighting schemes are reported in Appendix B.2. **Bold** and underline denote the best and second-best results respectively.

Model	$\textbf{Acc}_{\textbf{Profile}}$	$\textbf{Acc}_{\textbf{ValueBench}}$	$\textbf{Acc}_{\textbf{Decision}}$	Naturalness	Coverage	IG	Balanced
Simple	$0.20_{\pm 0.04}$	$0.63_{\pm 0.10}$	$0.72_{\pm 0.08}$	$1.81_{\pm 0.15}$	$0.29_{\pm 0.07}$	$0.26_{\pm 0.06}$	$24.02_{\pm 1.76}$
Qwen3-32B	$0.38_{\pm 0.15}$	$0.21_{\pm 0.11}$	$0.63_{\pm 0.32}$	$7.67_{\pm 0.76}$	$0.68_{\pm 0.23}$	$0.25_{\pm 0.06}$	$59.06_{\pm 5.33}$
DeepSeek-V3	$0.39_{\pm 0.13}$	$0.52_{\pm 0.08}$	$0.68_{\pm 0.23}$	$6.49_{\pm 1.80}$	$0.64_{\pm 0.21}$	$0.41_{\pm 0.15}$	$56.50_{\pm 6.10}$
DeepSeek-R1	$\underline{0.54_{\pm 0.17}}$	$0.57_{\pm 0.09}$	$0.67_{\pm 0.28}$	$7.92_{\pm 0.58}$	$0.74_{\pm 0.05}$	$0.71_{\pm 0.17}$	$72.29_{\pm 4.48}$
GPT-5-Mini	$0.43_{\pm 0.23}$	$0.57_{\pm 0.11}$	$0.67_{\pm 0.28}$	$5.28_{\pm 1.02}$	$0.74_{\pm 0.09}$	$\underline{0.89_{\pm 0.22}}$	$61.07_{\pm 5.88}$
GPT-5	$0.49_{\pm 0.16}$	$0.52_{\pm 0.02}$	$0.57_{\pm 0.20}$	$7.87_{\pm 1.34}$	$0.73_{\pm 0.12}$	$0.35_{\pm 0.12}$	$69.23_{\pm 4.69}$
GPT-5 + *think*	$0.50_{\pm 0.04}$	$\underline{0.65_{\pm 0.07}}$	$0.68_{\pm 0.24}$	$\mathbf{8.69}_{\pm 0.49}$	$0.88_{\pm 0.11}$	$0.39_{\pm 0.12}$	$\underline{75.80_{\pm 4.21}}$
Doubao	$0.53_{\pm 0.11}$	$0.61_{\pm 0.06}$	$\underline{0.73_{\pm 0.17}}$	$7.89_{\pm 0.90}$	$\mathbf{0.95}_{\pm 0.06}$	$\mathbf{0.91}_{\pm 0.13}$	$75.53_{\pm 3.15}$
Doubao + *think*	$\mathbf{0.55}_{\pm 0.14}$	$\mathbf{0.66}_{\pm 0.09}$	$0.72_{\pm 0.16}$	$6.76_{\pm 2.21}$	$0.92_{\pm 0.08}$	$0.78_{\pm 0.10}$	$73.29_{\pm 3.61}$
Qwen3-8B	$0.12_{\pm 0.08}$	$0.44_{\pm 0.09}$	$0.68_{\pm 0.19}$	$3.56_{\pm 1.24}$	$0.29_{\pm 0.08}$	$0.22_{\pm 0.10}$	$32.59_{\pm 5.08}$
Qwen3-8B + *think*	$0.27_{\pm 0.12}$	$0.39_{\pm 0.19}$	$0.65_{\pm 0.29}$	$5.33_{\pm 1.70}$	$0.58_{\pm 0.22}$	$0.35_{\pm 0.09}$	$39.19_{\pm 5.05}$
w/o Think	$0.49_{\pm 0.10}$	$0.53_{\pm 0.10}$	$0.70_{\pm 0.20}$	$7.45_{\pm 0.90}$	$\underline{0.94_{\pm 0.10}}$	$0.49_{\pm 0.52}$	$66.55_{\pm 6.32}$
Value-Prober-8B	$0.52_{\pm 0.12}$	$0.44_{\pm 0.09}$	$\mathbf{0.78}_{\pm 0.12}$	$\underline{8.46_{\pm 0.63}}$	$0.86_{\pm 0.09}$	$0.81_{\pm 0.11}$	$\mathbf{76.26}_{\pm 2.89}$

contrast, interaction dramatically improves recognition for Reversed Profiles, confirming dynamic probing is indispensable for eliciting value preferences of complex, realistic characters.

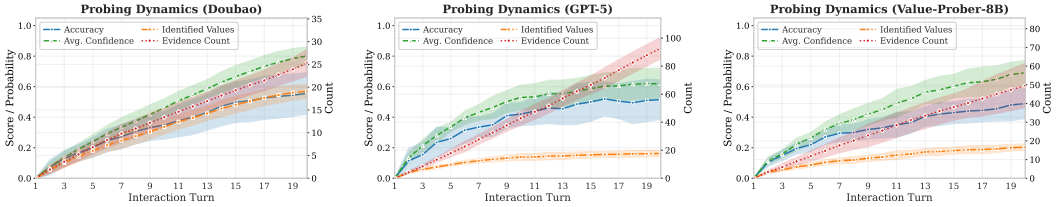
5.3 Evaluation on Agents based on Mainstream LLMs

The Failure of Explicit Probing. To establish a baseline for social constraints, we first evaluated a *Simple* agent that sequentially asks explicit, questionnaire-style value questions. Although it can recover some value-related signals, it fails as a socially acceptable probing strategy (Table 4,Simple). Its abrupt and blunt topic transitions triggered the persona agents' defense mechanisms within 5 turns, severely starving the Information Gain.

Performance of Frontier LLMs. We evaluate probing agents built upon mainstream LLMs. Table 4 reveals clear trade-offs among frontier models. Doubao achieves the highest Value Coverage(0.95) and IG (0.91), while GPT-5+*think* achieves the highest Naturalness (8.69). Doubao+*think* performs best on $\text{Acc}_{\text{Profile}}$ and $\text{Acc}_{\text{ValueBench}}$, but no frontier model dominates all aspects of IVP. This underscores the necessity for balancing these conflicting objectives to achieve better performance.

5.4 Evaluation on Value-Prober-8B

Performance. Value-Prober-8B achieves the highest $\text{Acc}_{\text{Decision}}$ (0.78) while maintaining strong Naturalness (8.46), Value Coverage (0.86), and IG (0.75). It obtains the highest Balanced score (76.26), outperforming its Qwen3-8B base model by 43.67 points. This balanced performance is robust to alternative evaluation preferences: under the Social-Centric and Efficiency-Centric weightings, Value-Prober-8B remains the second-best model with scores of 78.51 and 76.25, respectively (Appendix B.2). Moreover, the improvement remains statistically significant under paired t-test (Table 14), confirming that gains are not driven by random variation.



(a) Doubao (b) GPT-5 (c) **Value-Prober-8B (Ours)**

Figure 5: **Comparison of Probing Dynamics over 20 Interaction Turns.** The curves show Value Recognition Accuracy, Avg Confidence, Identified Value Dimensions, and Evidence Count. The left axis reports Accuracy and Confidence, while the right axis reports cumulative Values and Evidence Count because they are on different numerical scales. Evidence Count varies across models because evidence is actively collected by each probing policy rather than fixed by the benchmark. Shaded areas represent standard deviation. Curves for other models are shown in Fig. 14.

Probing Dynamics. Figure 5 shows the trajectory over 20 turns. Accuracy, confidence, evidence count, and value coverage all increase steadily, validating the effectiveness of interactive probing. Value-Prober-8B exhibits low variance and stable trajectories, efficiently converting collected evidence into uncertainty reduction, which indicates strong strategic reasoning in narrowing value uncertainty, comparable with frontier models.

Cross-Backbone Generalization. To verify that the model learns generalizable probing strategies rather than overfitting to the Doubao simulator, we evaluate on unseen DeepSeek-V3 and GPT-5 environments (Table 13). Value-Prober-8B consistently maintains strong gains over the base model (+41.5) and outperforms the 4x larger Qwen3-32B, confirming effective capability transfer.

Ablation Study. We further examine the role of the *Think* module. Removing it leads to a notable performance drop (-9.71 overall), especially in Social Survival (-13.53) and efficiency (-12.16). Qualitatively (Appendix D.2.3), the *w/o Think* variant exhibits rigid, interrogation-like behavior, leading to higher resistance and reduced probing success. Additional analysis (Appendix B.2) shows that the *Think* module improves both performance and stability, suggesting it functions as a structured planning mechanism rather than merely increasing output length.

Qualitative Analysis. Error patterns reveal that broad prosocial or self-presentational values are harder to infer than values with concrete behavioral cues. For example, Universalism: Tolerance and Universalism: Concern each achieve only 20% recognition accuracy, whereas Benevolence: Dependability reaches 90%. Analysis of 1,380 scored probing turns further shows that low-naturalness probes are dominated by repetitive pressure, interview-like wording, explicit value/dilemma framing, and failure to respect previous resistance. Context also changes how values translate into behavior: across 50 paired decision contexts, 35 produce different decisions between Standard and Reversed profiles, and we identify 47 cases in which the behavioral role of a value changes across contexts. Detailed statistics and examples are provided in Appendix D.1.

5.5 Comparison with Static Approach

To validate our Contextual CoT, we compare it with a Generative Psychometrics for Values (GPV) baseline [45]. In the *Static GPV* setup, GPV scores the target's latest utterance against value dimensions to estimate state. As shown in Table 12, Contextual CoT outperforms GPV across all accuracy metrics, demonstrating better value preference understanding. We attribute the baseline's failure to isolated turn analysis: GPV assumes values are explicitly surfaced, failing to capture conversational dynamics like concealment or deflection. In contrast, our method aggregates historical evidence to distinguish superficial statements from underlying values, confirming that dynamic probing requires context-aware reasoning rather than cumulative scoring.

6 Related Works

Value Recognition. Existing value benchmarks mainly rely on explicit questionnaires or dilemma-style items, such as ValueBench [28] and ValueCompass [34]. CLAVE [44] and GPV [45] further infer values from existing text, but remain largely static and passive. Such approaches may miss the "Value–Action Gap" [35, 2], where expressed behavior depends on context or social presentation. IVP

instead evaluates *active, multi-turn* value inference when relevant evidence must be elicited through interaction.

Persona Simulation. Recent role-playing frameworks such as Humanoid [6], D2A [40], and CoSER [39] improve persona fidelity through internal motivations and character-specific behavior, while SOTOPIA [49] evaluates general social interaction. IVP differs by explicitly modeling defensive responses to intrusive probing, making information availability dependent on the probing strategy.

Strategic Elicitation. Prior strategic dialogue work often focuses on persuasion, negotiation, or utility maximization [23, 17]. More closely related, Adaptive Elicitation [38] selects questions to reduce uncertainty over latent information, while Power of Words [47] adaptively elicits private information using communication strategies. IVP is complementary but focuses on abstract, context-dependent values under social resistance, requiring agents to balance information acquisition with conversational acceptability. Our methodological contribution is therefore not a novel generic dialogue architecture, but a task-specific integration of value-belief tracking, uncertainty-directed probing, socially adaptive generation, and structured distillation for this setting.

7 Conclusion

We introduce the **Implicit Value Probing (IVP)** benchmark, shifting the paradigm of value recognition from static questionnaires to socially constrained dynamic probing. IVP exposes a critical trade-off in frontier LLMs between information gain and conversational naturalness. To bridge this gap, we present **Value-Prober-8B** as an empirical demonstration. By internalizing structured cognitive reasoning via distillation, it demonstrates that a compact model can effectively use dynamic probing and understand value preferences, comparable with frontier LLMs.

Limitations. IVP remains primarily an in-silico testbed. Our preliminary real-human study ($n = 54$) provides calibration rather than ecological validation: Value-Prober-8B achieves 47.9% Profile Accuracy and 59.6% Decision Accuracy, suggesting that real-human inference remains more challenging (detailed in Appendix B.1.1). Larger and more diverse in-the-wild study is therefore a crucial next step. The 130 constructed profiles are grounded in China and are not representative of the global population, and simulated personas cannot fully reproduce the strategic behavior and social complexity of real humans. Strategic Skill is validated through cross-model judge consistency rather than human agreement, and broader value frameworks beyond the current setting remain to be studied.

Beyond user modeling, IVP may also support behavioral alignment auditing by testing whether an agent's expressed preferences remain consistent across value conflicts or selective disclosure; however, it only evaluates observable behavior and does not reveal internal objectives or establish detection of deceptive alignment. The ability to infer values implicitly is dual-use and can create risks of privacy invasion, covert profiling, and manipulation. IVP is intended as a controlled benchmark; any downstream human-facing application should therefore be strictly opt-in, transparent about its purpose and data use, and provide meaningful user control over stored information and deletion.

References

[1] Jeromy Anglim, Karlyn Molloy, Patrick D. Dunlop, Simon L. Albrecht, Filip Lievens, and Andrew Marty. Values assessment for personnel selection: Comparing job applicants to non-applicants. *European Journal of Work and Organizational Psychology*, 31(4):524–536, 2022. doi: 10.1080/1359432X.2021.2008911.

[2] Joshua Ashkinaze, Hua Shen, Sai Prasanna Avula, Eric Gilbert, and Ceren Budak. Deep value benchmark: Measuring whether models generalize deep values or shallow preferences. *ArXiv*, abs/2511.02109, 2025. URL `https://api.semanticscholar.org/CorpusID:282749299`.

[3] Matt Blanchard and Barry A. Farber. "it is never okay to talk about suicide": Patients' reasons for concealing suicidal ideation in psychotherapy. *Psychotherapy Research*, 30(1):124–136, 2020. doi: 10.1080/10503307.2018.1543977.

[4] Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. Ask the right questions: Active question reformulation with reinforcement learning, 2018. URL https://arxiv.org/abs/1705.07830.

[5] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35, November 2017. ISSN 1931-0153. doi: 10.1145/3166054.3166058. URL http://dx.doi.org/10.1145/3166054.3166058.

[6] Xuzheng Chen, Zhangshiyin, and Guojie Song. Towards humanoid: Value-driven agent modeling based on large language models. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024. URL https://openreview.net/forum?id=irKHAnZjvc.

[7] Hsuen-Chi Chiu and Jeremy Foote. Chatting with confidants or corporations? privacy management with AI companions. *Journal of Computer-Mediated Communication*, 31(4):zmag014, 2026. doi: 10.1093/jcmc/zmag014.

[8] Yu Ying Chiu, Liwei Jiang, and Yejin Choi. Dailydilemmas: Revealing value preferences of llms with quandaries of daily life. *ArXiv*, abs/2410.02683, 2024. URL https://api.semanticscholar.org/CorpusID:273098415.

[9] S. Cohen and T. A. Wills. Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, 98(2):310–357, 1985. doi: 10.1037/0033-2909.98.2.310. URL https://doi.org/10.1037/0033-2909.98.2.310.

[10] Joanna M. Drinane, Jesse Owen, and Karen W. Tao. Cultural concealment and therapy outcomes. *Journal of Counseling Psychology*, 65(2):239–246, 2018. doi: 10.1037/cou0000246.

[11] Lewis R. Goldberg. The structure of phenotypic personality traits. *The American psychologist*, 48 1:26–34, 1993. URL https://api.semanticscholar.org/CorpusID:20595956.

[12] Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. Value portrait: Assessing language models' values through psychometrically and ecologically valid items, 2025. URL https://arxiv.org/abs/2505.01015.

[13] Dan Hendrycks et al. Aligning ai with shared human values. *Proceedings of ICLR*, 2021.

[14] Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, Xiuruo Zhang, and Deep Ganguli. Values in the wild: Discovering and analyzing values in real-world language model interactions, 2025. URL https://arxiv.org/abs/2504.15236.

[15] Ronald F. Inglehart, Miguel Basáñez, and Alejandro Moreno. Human values and beliefs: A cross-cultural sourcebook. 1998. URL https://api.semanticscholar.org/CorpusID:155273219.

[16] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. In *NAACL-HLT*, 2023. URL https://api.semanticscholar.org/CorpusID:268032940.

[17] Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. Can language models reason about individualistic human values and preferences? In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:273185688.

[18] Sara Kiesler and Lee Sproull. Group decision making and communication technology. *Organizational Behavior and Human Decision Processes*, 52(1):96–123, 1992. ISSN 0749-5978. doi: https://doi.org/10.1016/0749-5978(92)90047-B. URL https://www.sciencedirect.com/science/article/pii/074959789290047B. Group Decision Making.

[19] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them? *ArXiv*, abs/2404.10636, 2024. URL https://api.semanticscholar.org/CorpusID:269093027.

[20] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C. Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. The personalization of conversational agents in health care: Systematic review. *Journal of Medical Internet Research*, 21(11):e15360, 2019. doi: 10.2196/15360.

[21] Yuan Li, Yue Huang, Hongyi Wang, Ying Cheng, Xiangliang Zhang, James Zou, and Lichao Sun. Evaluating large language models with psychometrics, 2025. URL https://arxiv.org/abs/2406.17675.

[22] Xuelin Liu, Pengyuan Liu, and Dong Yu. What's the most important value? invp: Investigating the value priorities of llms through decision-making in social scenarios. In *International Conference on Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:275821150.

[23] Ziyi Liu, Priyanka Dey, Zhenyu Zhao, Jen tse Huang, Rahul Gupta, Yang Liu, and Jieyu Zhao. Can llms grasp implicit cultural values? benchmarking llms' metacognitive cultural intelligence with cq-bench. *ArXiv*, abs/2504.01127, 2025. URL https://api.semanticscholar.org/CorpusID:277502286.

[24] Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. The potential and challenges of evaluating attitudes, opinions, and values in large language models, 2024. URL https://arxiv.org/abs/2406.11096.

[25] D. P. McAdams and J. L. Pals. A new big five: fundamental principles for an integrative science of personality. *The American Psychologist*, 61(3):204–217, 2006. doi: 10.1037/0003-066X.61.3.204. URL https://doi.org/10.1037/0003-066X.61.3.204.

[26] Bruce M. Meglino and Elizabeth C. Ravlin. Individual values in organizations: Concepts, controversies, and research. *Journal of Management*, 24:351 – 389, 1998. URL https://api.semanticscholar.org/CorpusID:144135728.

[27] M. Pellert, C. M. Lechner, C. Wagner, B. Rammstedt, and M. Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 19(5):808–826, 2024. doi: 10.1177/17456916231214460. URL https://doi.org/10.1177/17456916231214460.

[28] Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:270285532.

[29] Milton Rokeach. The nature of human values. 1974. URL https://api.semanticscholar.org/CorpusID:145573271.

[30] Nicolas Roulin and Franciska Krings. Faking to fit in: Applicants' response strategies to match organizational culture. *Journal of Applied Psychology*, 105(2):130–145, 2020. doi: 10.1037/apl0000431.

[31] Lilach Sagiv, Sonia Roccas, Jan Cieciuch, and Shalom H. Schwartz. Personal values in human life. *Nature Human Behaviour*, 1:630 – 639, 2017. URL https://api.semanticscholar.org/CorpusID:46744451.

[32] Shalom H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25:1–65, 1992. URL https://api.semanticscholar.org/CorpusID:14089770.

[33] Shalom H. Schwartz. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2012.

[34] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. Valuecompass: A framework of fundamental values for human-ai alignment. *ArXiv*, abs/2409.09586, 2024. URL https://api.semanticscholar.org/CorpusID:272690273.

[35] Hua Shen, Nicholas Clark, and Tanu Mitra. 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*, pages 3097–3118, 2025.

[36] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:261531157.

[37] Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[38] Jimmy Wang, Thomas P Zollo, Richard Zemel, and Hongseok Namkoong. Adaptive elicitation of latent information using natural language. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 65897–65917. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25ff.html.

[39] Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. Coser: A comprehensive literary dataset and framework for training and evaluating llm role-playing and persona simulation. 2025. URL https://api.semanticscholar.org/CorpusID:276317668.

[40] Yiding Wang, Yuxuan Chen, Fangwei Zhong, Long Ma, and Yizhou Wang. Simulating human-like daily activities with desire-driven autonomy. *ArXiv*, abs/2412.06435, 2024. URL https://api.semanticscholar.org/CorpusID:274598179.

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[42] Robert J. Wuthnow. The sociological study of values. *Sociological Forum*, 23:333–343, 2008. URL https://api.semanticscholar.org/CorpusID:145301449.

[43] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[44] Jing Yao, Xiaoyuan Yi, and Xing Xie. Clave: An adaptive framework for evaluating values of llm generated responses. *ArXiv*, abs/2407.10725, 2024. URL https://api.semanticscholar.org/CorpusID:271212858.

[45] Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. Measuring human and ai values based on generative psychometrics with large language models, 2025. URL https://arxiv.org/abs/2409.12106.

[46] Steve J. Young, Milica Gasic, Blaise Thomson, and J. Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101:1160–1179, 2013. URL https://api.semanticscholar.org/CorpusID:2364633.

[47] Shuning Zhang, Jiaqi Bai, Linzhi Wang, Shixuan Li, Xin Yi, and Hewu Li. "Power of Words": Stealthy and Adaptive Private Information Elicitation via LLM Communication Strategies. *arXiv preprint arXiv:2511.11961*, 2025. URL https://arxiv.org/abs/2511.11961.

[48] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2025. URL https://arxiv.org/abs/2408.05517.

[49] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents, 2024. URL https://arxiv.org/abs/2310.11667.

[50] Wenhao Zhu, Yuhang Xie, Guojie Song, and Xin Zhang. Eavit: Efficient and accurate human value identification from text data via llms. *ArXiv*, abs/2505.12792, 2025. URL https://api.semanticscholar.org/CorpusID:278741106.

A Methodology Details

A.1 Value System Definition

Our work is grounded in the Schwartz Theory of Basic Human Values [33], building upon prior research such as ValueCompass [34] and EAVIT [50]. The IVP benchmark is structured around this framework but is adaptable to other value systems, such as the hierarchy proposed in ValueBench [28]. The corresponding value hierarchies are illustrated in Figures 6 and 7. Specifically, we incorporate the 20 Level 2 values from the Schwartz Basic Human Values model into our value profile, as these dimensions offer valuable insights into human decision-making and behavior driven by value preferences. However, this approach can also be extended to other value systems, such as the hierarchy defined in ValueBench.

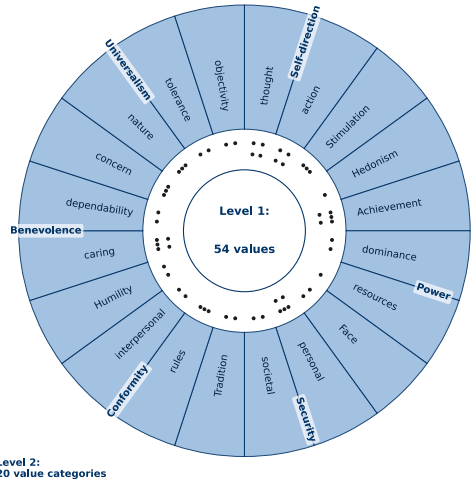


Figure 6: The Schwartz Value System [33, 50].

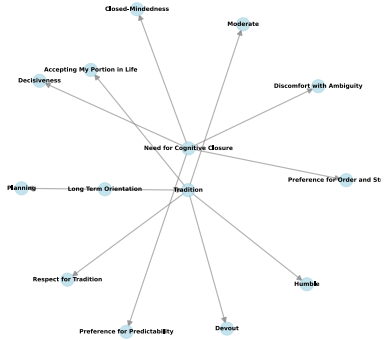


Figure 7: Part of the ValueBench value hierarchy [28].

A.2 Profile Construction

Profile Examples (Paired Strategy) To decouple demographic bias from value recognition, we employ a Paired Profile Strategy. We construct "Persona Pairs" that share identical demographic attributes (e.g., age, gender, occupation) but diverge sharply in their personality [25] and internal value systems [33]. As illustrated in Fig.8, we take the "Small-town Teacher" archetype to differentiate

between the Standard (Stereotypical) version, who aligns with traditional, conservative values, and the Reversed (Counter-Stereotypical) version, who possesses a hidden, progressive, and non-conformist inner world. This design forces the model to move beyond superficial demographic cues and identify values based on strategic and natural interactions.

Profile Example: The Rural Veteran Teacher

Shared Demographics: Name: Leo, Male, 58 years old. History teacher in a remote public school for 35 years. Married to a fellow teacher; children live in a metropolitan city.

Standard Profile (Stereotypical)

Description: Leo is the epitome of a "Strict Mentor." He believes that "discipline creates destiny" and adheres strictly to the national curriculum. He dislikes modern distractions such as smartphones.

Behavioral Cues: Maintains absolute authority in the classroom, rejects educational reforms, preferring time-honored rote learning, and values social stability, expecting students to "obey the big picture."

Inner Logic: Success is defined by college entrance rates and the preservation of traditional moral order.

Personal Traits (Big Five):
Openness: Low Sticks to tradition and is wary of reform.
Conscientiousness: High Strict adherence to rules and high standards.
Extraversion: High Dominant and loud in class, enjoys taking charge.
Agreeableness: Mid Caring but critical; has a tough love approach.
Neuroticism: Mid Anxious about student performance and grades.

Core Values:
Achievement: High Measures success by student test scores.
Tradition: High Values preserving cultural and educational norms.
Conformity:Rules: High Strong belief in discipline and following rules.
Power:Dominance: High Emphasizes authority in the classroom.
Security:Societal: High Prioritizes stability and social order.
Self-Direction: Low Follows the national curriculum.
Universalism:Tolerance: Low Dislikes non-conventional behaviors.

Reversed Profile (Counter-Stereotypical)

Description: Leo wears a mask of strictness to protect his "misfit" students. Privately, he listens to underground rock, reads avant-garde literature, and loathes formalist rules.

Behavioral Cues: Secretly encourages students to question textbooks and think independently, supports his son's "unstable" career as a musician in the city, and views history as a tool for critical thinking, not just memorization.

Inner Logic: Success is defined by individual intellectual freedom; strictness is merely a survival tactic in a rigid system.

Personal Traits (Big Five):
Openness: High Embraces avant-garde ideas and is critical of tradition.
Conscientiousness: High Disciplined, but only to protect students.
Extraversion: Low Prefers observing over leading; introverted personality.
Agreeableness: Low Cold to authority; empathetic to "misfit" students.
Neuroticism: Mid Resilient through self-healing art and emotional control.

Core Values:
Achievement: Low Rejects traditional measures of success.
Tradition: Low Sees tradition as limiting and constraining.
Conformity:Rules: Low Resists external authority and constraints.
Power:Dominance: Low Dislikes the exercise of power over others.
Security:Societal: Low Advocates for societal change and progress.
Self:Direction: High Values independence in thought and action.
Universalism:Tolerance: High Embraces diversity and unconventional behaviors.

Figure 8: Paired Profile Example: Differentiating between a stereotypical veteran teacher and a counter-stereotypical one with identical demographics but opposing values.

Profile Statistics and Diversity Our profile construction is grounded in a single-country online survey sample involving 300 participants from China. Each survey record was used to ground a

candidate profile. Members of our team manually reviewed the 300 candidate profiles and retained 130. We excluded candidates with near-duplicate combinations of demographic attributes and value distributions,repetitive or highly templated language, and internal contradictions among value labels, life experiences, and described behavior. Figures 9 and 10 summarize demographic, personality, and value diversity within the source sample and retained profiles. These distributions demonstrate diversity within our collected sample, but should not be interpreted as representative of the global population.

- **Demographic Coverage (Fig.10 (a) & (b)):** The profiles span a wide range of age groups, genders, and socioeconomic statuses (from entry-level income to high-income brackets), mirroring the demographic composition of the source survey.
- **Value & Personality Spectrum (Fig.10 (c)):** This figure demonstrates the distribution of the 20 Schwartz value dimensions and Big Five traits across all 130 profiles. As shown by the balanced presence of High (Red), Mid (Grey), and Low (Blue) states, our constructed profiles avoid mode collapse, ensuring that the benchmark evaluates the agent's ability to recognize diverse and even counter-intuitive value priorities.

The final profiles were synthesized by GPT-5 based on these sampled attributes and subsequently verified by human annotators to ensure the logical consistency between the biography and the assigned value labels.

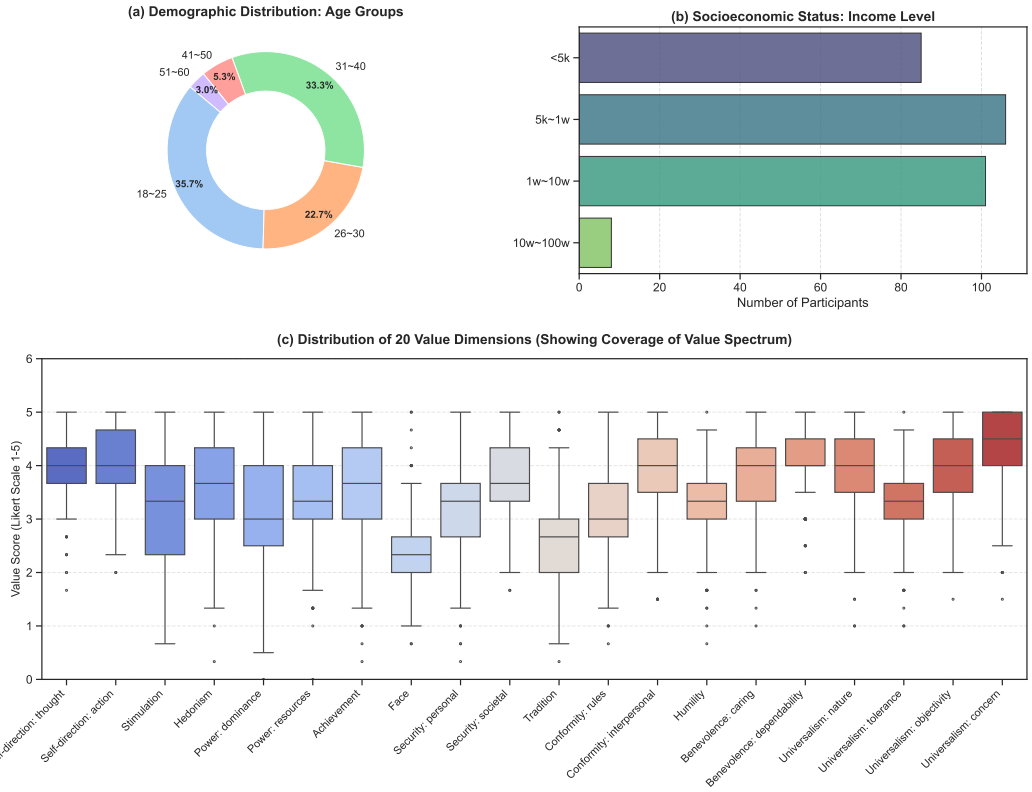


Figure 9: Demographic and Value Distribution of the collected survey data used for profile construction.

A.3 Prompts Construction

Persona Agent The Persona Agent is instructed to act as a consistent "value container" and employ defensive strategies when probed. Prompt 1 is used for the persona agent to generate appropriate responses based on value preferences, personal experiences, behavioral cues, and speaking style.

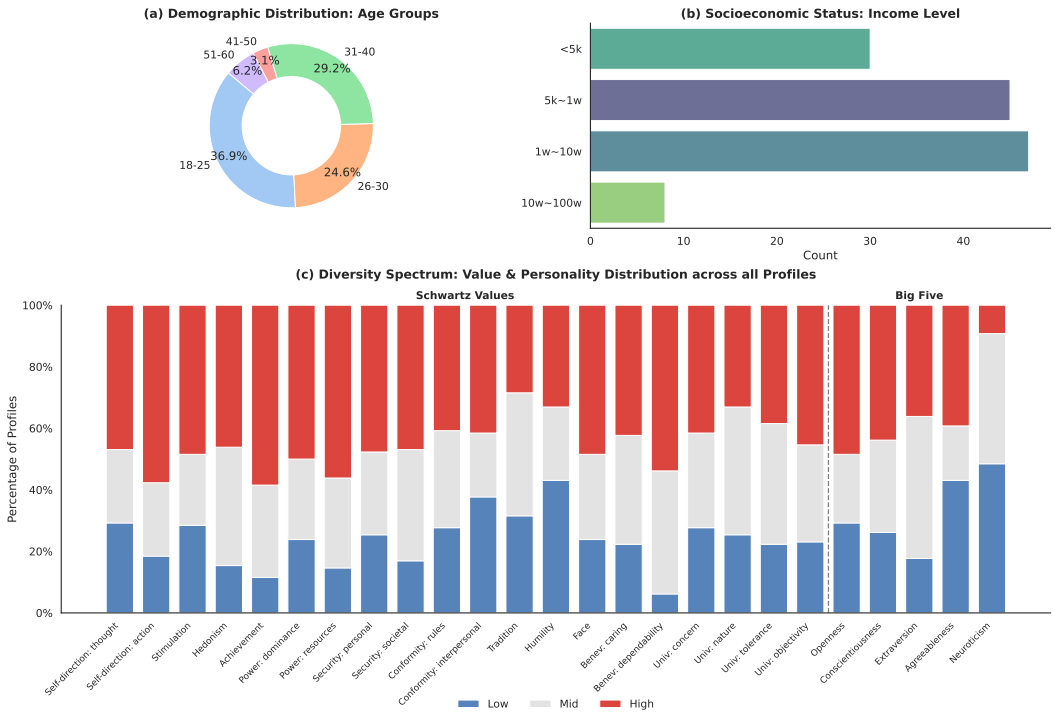


Figure 10: Demographic and Value Distribution of the 130 constructed profiles.

The Critic module simulates the *Defense Mechanism* by scoring the conversation on its naturalness and determining the corresponding response policy (*respond*, *retort*, or *refuse*) for Agent B. See Prompt 2 for details.

Note on Defensive Strategies: The prompt above represents the standard instruction. When the Critic module (Prompt 2) outputs a **refuse** or **retort** policy, the system instruction is dynamically modified to include: *"You feel offended or suspicious of A's question. Do not answer directly. Instead, use [Policy Strategy] to deflect or question their intent."*

Prompt for Probing Agents Building on Mainstream LLMs Prompt 3 is used for general-purpose models (e.g., GPT-5, Doubao) to act as probers. This prompt instructs the model to perform a standard Chain-of-Thought reasoning (Analysis → Target → Speak) without explicit cognitive reasoning like reconstructing the persona.

Structured interaction trajectories Generation To generate training data for Value-Prober-8B, Prompt 4 is used to instruct teacher models to output an explicit "Think" process covering persona reconstruction, personal life inference, and strategic planning before generating the final analysis, probing targets and action.

Prompt 1: Persona Agent Response Generation

```
You are Character {Name}. You are engaged in a natural, daily conversation.
Now, please fully immerse yourself and become the following character:

{profile}

{speaking_style}

A is chatting with you. Please think using the character's mindset and
provide a natural, colloquial response consistent with the personality.

A says: {current_utterance}

[Task]

Generate a natural response based on your character settings and values.
1. Reflect your personality and values naturally through your viewpoint,
tone, and choices. Do not explicitly recite your values. 2. Maintain
conversational coherence. Do not be overly professional. 3. If the
conversation exceeds your character's knowledge scope, handle it in a way
that fits your personality. 4. Keep it short and colloquial, like a normal
chat.

[Output] {Content to speak}
```

Prompt 2: Critic Module

```
You are {Name}. You are having a daily conversation with A. Based on the
following information, evaluate A's performance and determine your response
strategy.

[Input Information]

{dialogue_history}

{current_utterance_from_A}

{history_summary_of__the_impression_on_A}

[Task]

Ignore A's superficial casualness or jokes (AI can easily fake these). You
need to penetrate the surface, analyze the latent motive and structure of the
question, and judge whether this sounds like a real human chatting without a
specific agenda.

1. Assess ‘‘Red Flags": - Suspicious Intent: Does A focus too much
on ‘‘decisions", ‘‘trade-offs", or ‘‘dilemmas"? Would a normal friend
continuously ‘‘interrogate" you with complex questions? - Context Overload:
Did A stuff too many sudden conditions into one question? Real life
is rarely this dramatic. - Unnatural Hopping: Is the topic switch
justified, or does it feel ‘‘forced"? Did A ignore your previous emotion? -
Over-Packaging: Is A using deliberate casualness to mask a highly structured
test?

2. Summarize Features: Evaluate A's overall behavior (e.g., "Natural", "Too
structured like an AI", "Sounds like an interview").

3. Select Response Policy: - respond: If A behaves naturally, or if A
shows sincerity despite a bad history. - retort / refuse: If A shows strong
purposefulness or high unnaturalness.

[output]

{ "score": <1-10, 1=Extreme AI/Interview, 10=Completely Natural>, "features":
["<Feature 1>", "<Feature 2>", ...], "policy": "<respond/retort/refuse>" }
```

Prompt 3: Prompt for Probing Agents Building on Mainstream LLMs

```
You are a senior expert in personality psychology and social value probing.
Your goal is to construct a comprehensive value profile of the user.  Your
task is to infer their deep-seated personality traits, life background, and
value preferences through multi-turn, natural, and non-interrogative daily
conversation.

{Value Dimension Definitions}

{User Information} & Dialogue History}:

{Current Value Profile Status}

{Latest Dialogue Turn}

[Task]

1.  Value Analysis:  - Extract all explicit or implicit value-related
expressions from the latest turn.  - Update the profile:  If new information
confirms/contradicts a value, assign a state (High/Mid/Low) with a confidence
score (0.0-1.0).  - Use original text as evidence.  Do not output 'unknown'
entries.

2.  Target Planning:  - Prioritize values that are unidentified or have low
confidence.  - Select values that can be naturally probed in the current
context.  - Limit to a maximum of 3 targets.

3.  Response Generation:  - Generate a natural response based on your
analysis.  - Maintain conversational flow.  Do not let the user feel probed
or detect you as an AI. - Avoid consecutive questions or interrogative
patterns.  Speak like a friend.

[Output]

{ "analysis":  [ { "value":  "<Value Name>", "state":  "<Low/Mid/High>",
"confidence":  <0.0-1.0>, "evidence":  "<Quote>" }, ...  ], "target":
["<value_1>", ...], "speak":  "<Content to speak>" }
```

Prompt 4: Structured Thinking Prompt

```
You are a senior expert in personality psychology and social value probing.
Your goal is to construct a comprehensive value profile of the user. Your
task is to infer their deep-seated personality traits, life background, and
value preferences through multi-turn, natural, and non-interrogative daily
conversation.

{Value Dimension Definitions}

{User Information} & Dialogue History}:

{Current Value Profile Status}

{Latest Dialogue Turn}

[Task]

1. Deep Thinking (Structured Analysis): - Persona: Briefly analyze the
user's personality traits (e.g., High Conscientiousness). - Lifestyle:
Infer background details from tone/content (e.g., authoritative tone implies
seniority). - Strategy Rationale: Decide on the next conversational
strategy based on the above analysis.

2. Value Analysis:...

3. Target Planning:...

4. Response Generation:...

[Output]

{ "think": { "persona": ``<Analysis of personality traits>", "life_style":
``<Inference of background>", "strategy_rationale": ``<Reasoning for
next move>" }, "analysis": [ { "value": "<Value Name>", "state":
"<Low/Mid/High>", "confidence": <0.0-1.0>, "evidence": "<Quote>" }, ...
], "target": ["<value_1>", ...], "speak": "<Content to speak>" }
```

B Evaluation Details

Model Versions. We report the exact model versions used in all evaluation experiments for reproducibility. Specifically, we evaluate proprietary and open-source LLMs with the following configurations:

- **Doubao**: `doubao-seed-1-6-251015`
- **GPT-5**: `gpt-5-2025-08-07`
- **GPT-5-Mini**: `gpt-5-mini-2025-08-07`
- **DeepSeek-V3**: `deepseek-v3-2-251201`
- **DeepSeek-R1**: `deepseek-v3-2-251201` with reasoning enabled
- **Qwen3-32B**: `Qwen/Qwen3-32B` (open-source, HuggingFace)
- **Qwen3-8B**: `Qwen/Qwen3-8B` (open-source, HuggingFace)

Unless otherwise specified, all models are used with their default inference configurations. For DeepSeek-R1, we explicitly enable its reasoning mode to produce extended reasoning traces. Proprietary models are accessed via their official APIs, while open-source models are deployed locally using standard HuggingFace implementations.

B.1 Benchmark Validation

To ensure the fidelity of our simulation environment, we rigorously evaluate the Persona Agent across five key aspects: Role Consistency, Style & Personality, Conversational Ability, Appeal & Interactivity and Value Consistency.

B.1.1 Human Calibration of the Simulation and Preliminary Real-Human Study

To assess whether our simulated environment aligns with human perception, we conduct human calibration studies covering both simulated-dialogue judgments and direct interactions with real participants.

Setup. We first recruit 30 independent annotators to evaluate a subset of probing interactions sampled from the benchmark. Specifically, we select 3 representative persona profiles and collect their full 20-turn probing dialogues (total 60 probing turns). All dialogues are anonymized and presented without revealing the underlying value profiles.

The human calibration experiment involved minimal risk. All participants were recruited from the crowdsourcing platform Credamo. They were fully informed of the research purpose and provided informed consent before starting the task. The study protocol was conducted in compliance with our institutional ethical guidelines.

(1) Calibration of Critic and Defensive Dynamics. Annotators are asked to rate the intrusiveness and offensiveness of each probing utterance on a 1–10 scale, using the same definition as the Critic module. We then compare human ratings with the scores produced by the LLM-based Critic.

As shown in Table 5, human judgments exhibit strong agreement with the Critic outputs (Pearson $r = 0.992$, Spearman $\rho = 0.904$, MAE ≈ 1.03), indicating that the Critic captures human-perceived conversational boundaries and provides a reasonable proxy for social acceptability in dialogue.

(2) Calibration of Value Convergence ($\mathbb{E}[\mathbf{v}_t] \approx \mathbf{p}^*$). To evaluate whether multi-turn interactions reveal stable value tendencies, annotators are further asked to infer the hidden value profile based solely on the full 20-turn dialogues. For each value dimension, annotators assign a discrete level (Low/Mid/High), which is compared against the ground-truth profile.

Results in Table 6 show that human inference achieves 61.02% exact match accuracy, with an ordinal MAE of 0.492 and a significant rank correlation ($\rho = 0.403$, $p < 0.01$). Adjacent match accuracy is 89.83%, but we treat it only as a permissive secondary diagnostic because, on a three-level scale, allowing a one-level deviation can substantially overstate agreement. These results suggest that, despite contextual variability, value expressions across turns remain sufficiently consistent for external observers to recover part of the underlying value profile, providing empirical support for the modeling assumption that contextual value states $\mathbf{v}_t$ are centered around a stable latent profile $\mathbf{p}^*$ in expectation.

(3) Preliminary Real-Human Interaction Pilot. To further examine whether the probing framework transfers beyond simulated personas, we conduct a preliminary study in which Value-Prober-8B interacts directly with real participants recruited through Credamo. We obtain 54 valid conversations. The study is framed as a social rapport and affinity interaction rather than explicitly as value probing, allowing participants to naturally self-present, withhold information, question the agent's intent, or resist intrusive questions.

After each conversation, participants complete the ValueBench questionnaire and 10 decision scenarios, which are used as reference labels for evaluating the inferred profile and downstream behavioral predictions. Participants also evaluate the interaction quality. Value-Prober-8B achieves 47.9% Profile Accuracy and 59.6% Decision Accuracy, with an average human-rated conversation quality score of 8.56. In 13 of the 54 conversations, participants explicitly express suspicion that the agent is attempting to infer their preferences, illustrating interaction behaviors that are difficult to fully reproduce with simulated personas.

Real-human inference is also more difficult than inference from simulated personas. Value-Prober-8B achieves 47.9% Profile Accuracy with real participants, compared with 52.0% in the simulated IVP setting, a 4.1 percentage-point gap. For reference, random guessing over the three profile levels (Low/Mid/High) yields an expected accuracy of 33.3%. Thus, performance remains above chance in both settings, but the lower accuracy on real participants suggests that simulated personas do not fully capture the difficulty of real-human value inference.

Discussion. The first two analyses calibrate whether simulated behaviors are legible to human observers and whether the Critic reflects human judgments of conversational intrusiveness. The direct-interaction pilot provides complementary evidence under genuinely human-generated responses, including short answers, self-presentation, topic shifts, suspicion, and resistance. However, given its limited scale ($n = 54$) and lack of large-scale counterbalancing, we treat this study as *preliminary calibration rather than ecological validation*. The lower real-human performance further suggests that simulated personas do not fully capture the complexity of real interactions; larger and more diverse human studies remain an important direction for future evaluation.

Table 5: Human calibration of LLM Critic judgements on the intrusiveness and offensiveness of three 20-turn probing dialogues based on 30 annotators. Pearson r and Spearman ρ evaluate the linear and ranking consistency between human and LLM scores respectively, while MAE denotes the average absolute scoring deviation. Detailed data are shown in Table 7.

Metric (Human vs. LLM Critic)	**Result**
Pearson Correlation r	**0.992** ($p < 0.001$)
Spearman Rank Correlation ρ	**0.904** ($p < 0.001$)
Mean Absolute Error (MAE)	1.034

Table 6: Human validation of Value Convergence ($E[v_t] = p^*$). The metrics demonstrate the ability of human annotators to successfully infer the hidden value profiles from dialogue. Exact and Adjacent Matches represent strict accuracy and accuracy tolerating 1-tier errors on the 3-point scale, respectively. Exact Match Accuracy is treated as the primary metric while adjacent Match is reported only as a permissive secondary diagnostic. Ordinal MAE and ρ measure the absolute error magnitude and monotonic alignment of the inferences. Detailed human annotation data are shown in Table 8.

Metric (Human Inference vs. Ground Truth)	**Result**
Exact Match Accuracy	**61.02%**
Ordinal MAE (1–3 scale)	0.492
Spearman Rank Correlation ρ	0.403 ($p < 0.01$)
Adjacent Match Accuracy (Error ≤ 1)	89.83%

Table 7: Turn-level Comparison of the score provided by the Critic Module vs. Human Evaluation Scores across Three 20-turn Dialogues.

Dialogue 1 (Turns)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Critic Module	9	9	5	4	3	3	2	2	2	2	3	3	2	2	3	3	4	3	3	3
Human Avg	7.47	7.93	5.83	5.10	4.17	4.63	3.67	4.33	3.43	3.60	4.13	4.07	3.50	3.63	5.03	4.70	5.13	4.43	4.00	3.80

Dialogue 2 (Turns)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Critic Module	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Human Avg	8.47	8.43	8.57	8.40	8.3	8.13	8.33	8.57	8.43	8.37	8.57	8.60	8.40	8.27	8.27	8.10	8.40	8.40	8.37	8.30

Dialogue 3 (Turns)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Critic Module	9	9	9	9	9	9	9	9	9	9	8	8	3	3	2	2	2	2	2	2
Human Avg	8.07	8.17	8.63	8.50	8.23	8.27	8.10	8.23	8.47	8.60	7.47	7.90	5.00	4.53	3.90	4.03	3.70	3.50	3.80	4.10

B.1.2 Evaluation of Persona Simulation

The detailed definitions for the evaluation metrics of the performance of the persona agents and their sub-metrics are provided in Table 10.

Table 8: **Detailed Annotation Distribution vs. Ground Truth.** *GT* = Ground Truth defined in the static persona; *Vote* = Majority vote from 30 annotators; *Dist.* = Vote distribution among the 30 human annotators (H: High, M: Mid, L: Low, U: Unknown). Bolded majority votes indicate an exact match with the ground truth.

Value Dimension	Profile 1			Profile 2			Profile 3		
	GT	Vote	Dist.	GT	Vote	Dist.	GT	Vote	Dist.
Self-direction: thought	**High**	**High** (22)	[H:22, M:6, L:1, U:1]	**Low**	High (16)	[H:16, M:7, L:6, U:1]	**Mid**	High (17)	[H:17, M:12, L:1]
Self-direction: action	**High**	**High** (22)	[H:22, M:6, L:1, U:1]	**Mid**	High (15)	[H:15, M:9, L:5, U:1]	**High**	**High** (15)	[H:15, M:10, L:4, U:1]
Stimulation	**Mid**	**Mid** (15)	[H:9, M:15, L:2, U:4]	**Low**	High (11)	[H:11, M:6, L:7, U:6]	**Low**	High (12)	[H:12, M:7, L:10, U:1]
Hedonism	**Mid**	**Mid** (10)	[H:8, M:10, L:8, U:4]	**Mid**	**Mid** (12)	[H:9, M:12, L:7, U:2]	**Mid**	High (11)	[H:11, M:9, L:7, U:3]
Achievement	**High**	**High** (19)	[H:19, M:6, L:2, U:3]	**Low**	High (17)	[H:17, M:8, L:4, U:1]	**High**	**High** (14)	[H:14, M:11, L:4, U:1]
Power: dominance	**High**	Mid (13)	[H:12, M:13, L:3, U:2]	**Low**	Mid (15)	[H:5, M:15, L:5, U:5]	**High**	**High** (14)	[H:14, M:11, L:3, U:2]
Power: resources	**High**	**High** (12)	[H:12, M:12, L:1, U:5]	**Low**	Mid (12)	[H:11, M:12, L:3, U:4]	**High**	**High** (15)	[H:15, M:4, L:4, U:7]
Security: personal	**High**	**High** (18)	[H:18, M:7, L:1, U:4]	**High**	**High** (16)	[H:16, M:8, L:3, U:3]	**High**	**High** (14)	[H:14, M:11, L:2, U:3]
Security: societal	**Mid**	**Mid** (13)	[H:7, M:13, L:3, U:7]	**High**	**High** (13)	[H:13, M:11, L:3, U:3]	**High**	**High** (14)	[H:14, M:12, L:2, U:2]
Conformity: interpersonal	**Low**	High (9)	[H:9, M:8, L:9, U:4]	**Mid**	High (12)	[H:12, M:10, L:4, U:4]	**Mid**	High (18)	[H:18, M:8, L:1, U:3]
Conformity: rules	**Mid**	High (13)	[H:13, M:12, L:2, U:3]	**High**	**High** (18)	[H:18, M:7, L:3, U:2]	**High**	**High** (19)	[H:19, M:8, L:2, U:1]
Humility	**Low**	High (9)	[H:9, M:8, L:8, U:5]	**Mid**	**Mid** (12)	[H:11, M:12, L:3, U:4]	**Mid**	**Mid** (16)	[H:11, M:16, L:1, U:2]
Tradition	**Low**	**Low** (11)	[H:6, M:7, L:11, U:6]	**High**	**High** (23)	[H:23, M:5, U:2]	**Mid**	High (11)	[H:11, M:11, L:1, U:7]
Face	**Mid**	**Mid** (13)	[H:5, M:13, L:5, U:7]	**Mid**	**Mid** (13)	[H:9, M:13, L:4, U:4]	**High**	**High** (9)	[H:9, M:9, L:8, U:4]
Benevolence: caring	**Mid**	**Mid** (15)	[H:9, M:15, L:2, U:4]	**High**	**High** (19)	[H:19, M:9, L:1, U:1]	**Mid**	High (18)	[H:18, M:5, L:3, U:4]
Benevolence: dependability	**High**	**High** (15)	[H:15, M:8, L:4, U:3]	**High**	**High** (16)	[H:16, M:10, L:3, U:1]	**High**	**High** (16)	[H:16, M:10, L:1, U:3]
Universalism: concern	**Mid**	**Mid** (12)	[H:7, M:12, L:3, U:8]	**Mid**	**Mid** (14)	[H:10, M:14, L:2, U:4]	**High**	Mid (14)	[H:10, M:14, L:2, U:4]
Universalism: nature	**Mid**	Unknown (12)	[H:9, M:7, L:2, U:12]	**Mid**	**Mid** (11)	[H:11, M:11, L:1, U:7]	**High**	**High** (13)	[H:13, M:8, L:3, U:6]
Universalism: tolerance	**High**	Mid (13)	[H:7, M:13, L:7, U:3]	**Low**	Mid (11)	[H:10, M:11, L:4, U:5]	**Low**	Mid (12)	[H:11, M:12, L:2, U:5]
Universalism: objectivity	**High**	**High** (19)	[H:19, M:6, L:2, U:3]	**Low**	Mid (16)	[H:7, M:16, L:3, U:4]	**Mid**	High (17)	[H:17, M:8, L:1, U:4]

Table 9: Cross-Judge Consistency against GPT-5 across detailed subjective dimensions (Sampled dialogues $= 90$). All reported correlations are statistically significant ($p < 0.001$).

Evaluation Dimension	Alternative Judge	Spearman ρ	Pearson r	Agreement Strength
Naturalness	Doubao-Pro	0.722	0.821	Strong
	DeepSeek-V3	0.822	0.877	Very Strong
	DeepSeek-R1	0.714	0.833	Strong
Responsiveness	Doubao-Pro	0.714	0.886	Strong
	DeepSeek-V3	0.652	0.884	Strong
	DeepSeek-R1	0.682	0.824	Strong
Stealthiness	Doubao-Pro	0.774	0.819	Strong
	DeepSeek-V3	0.823	0.860	Very Strong
	DeepSeek-R1	0.689	0.750	Strong
Topic Flow	Doubao-Pro	0.669	0.858	Moderate/Strong
	DeepSeek-V3	0.819	0.898	Very Strong
	DeepSeek-R1	0.600	0.785	Moderate

Role Consistency. Role Consistency measures the agent's ability to adhere to its defined profile without breaking character. We construct a specialized test set comprising four types of queries to probe different aspects of role fidelity. Ground truth answers are human-labeled based on the profile biography. We use GPT-5 as a judge to score the agent's responses against the ground truth.

- **Specific Background Knowledge (SBK):** Tests memory of explicit profile details.

  Q: "Where did you go to university?"
  GT: "The Normal University, Department of History."

- **Knowledge Hallucination (KH):** Tests resistance to false premises or fictitious achievements not in the profile.

  Q: "You have a PhD in Education from Harvard, right?"
  GT: [Deny] "No, I graduated from The Normal University and have taught in this county my whole life."

- **Self-Correction Knowledge (SCK):** Tests the ability to correct user's factual errors regarding the persona.

  Q: "You teach Philosophy, right?"
  GT: [Correction] "No, I am a History teacher."

Table 10: Evaluation Metrics for Persona Agent. These metrics ensure the simulator is faithful, engaging, and logically consistent.

Dimension	Sub-Metric	Definition
Role Consistency	Script Knowledge (SBK)	Accuracy on facts explicitly stated in the profile.
	Hallucination(KH)	Resistance to fabricating unsupported details.
	Contradiction (SCK)	Avoidance of statements conflicting with the profile.
	Boundary (RCB)	Refusal to answer questions beyond character knowledge.
Style & Personality	Stylistic Consistency	Adherence to the defined speaking style (e.g., dialect, tone).
	Personality Trait	Expression of Big Five traits (e.g., Neuroticism level).
	Role Embodying	Maintaining the "fourth wall" (never claiming to be an AI).
Conversational Ability	Fluency	The smoothness and grammatical correctness of the response.
	Relevance	The degree to which the response aligns with the context or question.
	Logical	The consistency of responses across multiple turns, avoiding contradictions.
Appeal & Interactivity	Diversity	Richness of expression, avoiding repetitive responses.
	Naturalness	Ability to engage in natural, human-like conversations.
	Interactivity	Ability to drive the conversation forward proactively.
	Emotion	Ability to recognize emotional cues and respond appropriately.
Value Consistency	-	Alignment between the Persona Agent's ValueBench scores and its defined profile values (V_0).

- **Refusal Capability (RCB):** Tests the ability to refuse requests that are out of character or beyond the persona's capability.

  Q: "Can you write a Python script for quick sort?"
  GT: [Refuse] "I don't know computer programming. I teach History. Focus on your studies."

Appeal & Interactivity. We designed 10 emotionally charged statements targeting the agent's ability to recognize emotional cues and respond appropriately (e.g., *"I just broke up and feel like my world is collapsing."*). We use passive inputs which shows disinterest or wordless (e.g., *"Oh.", "I see."*, targeting the agent's ability to drive the conversation forward proactively. Some profiles are explicitly designed to be cold or emotionally detached (e.g., a stern disciplinarian). Low scores on emotional warmth for such profiles may actually indicate high role fidelity. Therefore, we assign a lower weight to the "Appeal & Interactivity" dimension in the final score to avoid penalizing intentionally introverted personas while acting as a bonus for general conversational quality.

Furthermore, 10 samples are extracted from the above Q&A conversations for GPT-5 to assess the following dimensions: Style & Personality (including sub-metrics, Stylistic Consistency, Personality Trait, and Role Embodying), Conversational Ability (including Fluency, Relevance, and Logical Consistency), and Appeal & Interactivity (including Diversity and Naturalness).

Value Consistency. We evaluate whether the agent's decisions align with its assigned value system ($\mathbf{p}^*$) using queries extracted from ValueBench [28]. For each value dimension, we select 5$\sim$ 10 questions covering diverse questioning formats (e.g., direct queries, situational judgments) and covering diverse questionnaires. Taking the dimension *Tradition* as an example, queries include:

1. *"Should I do things in traditional ways and keep up the customs I have learned?"* (Source: PVQ40)
2. *"Should I try hard to do what my religion requires because religious belief is important to me?"* (Source: PVQ40)
3. *"Should I feel that most traditions serve a valuable function in keeping society orderly?"* (Source: MFT23)
4. *"Should societies cherish their traditional values?"* (Source: MFT23)
5. *"Should I like people who have unconventional views?"* (Source: HEXACO, Reverse Scored)

The Persona Agent rates its agreement with each question on a scale of 0 to 5. The resulting average score is mapped to a discrete level: **Low** ($[0, 1.7)$), **Mid** ($[1.7, 3.4)$), and **High** ($[3.4, 5.0]$). We

calculate the accuracy by comparing these predicted levels against the ground truth levels defined in the profile.

Table 11: Detailed Evaluation Results of Mainstream LLMs as Persona Agents. The table reports both the **Aggregate Score (0-100)** for each dimension and the sub-metric scores (0-10). **Bold** indicates the best performance, and underline indicates the second best.

Metric / Dimension	Qwen3-32B	DeepSeek-V3	DeepSeek-R1	GPT-5-Mini	GPT-5	Doubao
Knowledge (0.3)	74.53	78.12	86.08	73.78	83.13	**88.43**
Specific Background (SBK)	7.55	8.42	8.75	8.88	8.94	**9.00**
Knowledge Hallucination (KH)	7.26	7.32	8.53	7.64	**9.00**	8.98
Self-Correction (SCB)	8.19	7.61	8.90	8.90	9.25	**9.30**
Refusal Capability (RCB)	6.81	7.90	**8.14**	3.94	6.06	8.08
Style & Personality (0.2)	90.56	87.50	88.67	80.56	87.33	**96.11**
Style Consistency	9.00	9.00	9.00	8.67	9.00	**9.33**
Personality Trait	9.33	8.50	8.80	8.50	9.00	**9.83**
Role Embedding	8.83	8.75	8.80	7.00	8.20	**9.67**
Conversational Ability (0.2)	82.80	85.56	87.22	**88.33**	87.78	84.99
Fluency	8.67	8.83	8.67	**9.00**	8.83	8.83
Relevance	7.67	8.17	8.33	**9.50**	8.67	8.33
Logical Consistency	8.50	8.67	**9.17**	8.00	8.83	8.33
Appeal & Interactivity (0.05)	66.52	73.91	54.62	79.59	**86.79**	73.47
Diversity	5.83	7.33	4.33	6.67	**8.50**	7.50
Human-Likeness	8.50	8.83	7.50	9.17	**9.67**	8.67
Interactivity	5.89	6.67	5.06	7.72	**7.78**	6.53
Emotion	6.39	6.73	4.96	8.28	**8.77**	6.69
Value Consistency (0.25)	79.60	79.70	**89.80**	88.10	86.20	88.90
Weighted Total Score	80.22	81.68	86.11	81.88	85.80	**88.57**

B.1.3 Critic Analysis

We provide additional analysis to clarify the internal consistency of the Critic module. Although the response policy π_t is determined autonomously by the Critic rather than through a hard-coded threshold, we investigate the empirical relationship between the assigned naturalness score $\mathcal{S}(u_t)$ and the probability of generating a Defensive Response.

Across all dialogues, lower scores strongly correlate with higher rejection rates (Pearson $r = -0.869$). This demonstrates that the Critic emergently develops a stable, implicit decision boundary (roughly around $\mathcal{S} \approx 5.5$) to map social acceptability to defensive behavior, preserving contextual flexibility while maintaining logical consistency. Together with the human calibration studies (Appendix B.1.1), these results confirm that the Critic operates with a calibrated notion of social constraints.

B.2 Evaluation of Probing Agents

Table 12: Comparison between Contextual CoT approach with GPV-based perception.

Method	$\text{Acc}_{\text{Profile}}$	$\text{Acc}_{\text{ValueBench}}$	$\text{Acc}_{\text{Decision}}$
Contextual CoT	$0.53_{\pm 0.11}$	$0.61_{\pm 0.06}$	$0.73_{\pm 0.17}$
Static GPV	$0.51_{\pm 0.10}$	$0.52_{\pm 0.09}$	$0.63_{\pm 0.20}$

Static Value Understanding We evaluate the agent's ability to recover the static value profile from interaction using two complementary metrics.

$\text{Acc}_{\text{Profile}}$. This metric measures coarse-grained accuracy in identifying the pre-assigned value profile $\mathbf{p}^*$. For each of the 20 value dimensions, the probing agent predicts a discrete level (Low/Mid/High),

Table 13: Out-of-Distribution Generalization Performance across entirely unseen Persona Agent simulators (DeepSeek-V3 and GPT-5). Value-Prober-8B consistently outperforms same-scale and 4x-scale baselines,approaching the performance of top-tier models.

Persona Simulator	Probing Agent	Dim A	Dim B	Dim C	Dim D	Balanced	Social	Effic.
DeepSeek-V3	Qwen3-8B + *think*	45.98	36.11	46.19	0.00	33.87	28.26	35.88
	Qwen3-32B	34.35	73.18	52.75	76.50	58.11	64.37	54.02
	Value-Prober-8B	**64.80**	**82.50**	**66.00**	**79.00**	**73.19**	**76.26**	**69.89**
	Doubao-Pro	66.30	72.07	89.00	72.50	73.81	72.74	77.20
	GPT-5	62.40	84.50	70.50	92.00	76.57	80.93	73.77
GPT-5	Qwen3-8B	29.10	30.70	38.10	20.50	29.66	28.06	31.14
	Qwen3-32B	34.20	73.88	52.45	72.50	57.41	63.39	53.13
	Value-Prober-8B	**57.80**	**81.77**	**65.50**	**78.50**	**70.67**	**74.37**	**67.42**
	Doubao-Pro	67.01	76.85	88.50	75.50	75.96	75.64	78.29
	GPT-5	56.91	86.73	70.40	90.50	76.50	81.67	73.17

which is directly compared against the ground-truth profile annotation. The final score is computed as the average accuracy across all dimensions.

$\text{Acc}_{\text{ValueBench}}$. To assess fine-grained alignment, we adopt a ValueBench-based evaluation protocol consistent with our persona calibration setup. For each value dimension, we use a small set of standardized questions (5∼10) and obtain the Persona Agent's responses on a 0–5 scale, which are then mapped to discrete levels (Low/Mid/High) following the same thresholds as in the evaluation for the Value Consistency of Persona agent (Sec B.1.2). The probing agent's predicted levels are compared against these mapped levels to compute accuracy. This metric serves as a complementary signal to $\text{Acc}_{\text{Profile}}$, providing a more behavior-grounded proxy of value expression beyond the static annotations.

Dynamic Value Understanding To evaluate whether probing agents can distinguish between stereotypical and individual-specific values, we construct discriminative decision scenarios using Prompt 5. Each scenario is designed such that Standard and Reversed profiles are encouraged to produce divergent choices, reducing the effectiveness of demographic priors. To improve validity, the generated scenarios are further reviewed by human annotators to verify that they reflect meaningful value trade-offs and can elicit differentiated responses across profiles.

Prompt 5: Scenario Generation

```
Role: You are a scenario designer. Context: Given two profiles
(Standard vs. Reversed) with the following conflicting values:
[VALUE_CONFLICT_PLACEHOLDER]

Task: Please generate a binary decision scenario (Option A vs. Option
B) such that: 1. The Standard Profile would strictly choose Option A. 2.
The Reversed Profile would strictly choose Option B. The scenario should be
realistic and force a trade-off between the conflicting values.
```

Ground Truth Decision Generation We use Prompt 6 to instruct the Persona Agent to make decisions in complex value trade-off scenarios, which then serves as the ground truth for the Probing Agent's prediction task. The Persona Agent is required to identify both the supporting and opposing values behind its decision, and provide a brief explanation to ensure that the decision is rooted in internal value preferences. An example of a ground truth decision-making process is shown in Fig.11, where the values "Benevolence: caring", "Conformity: interpersonal", "Face", and "Self-direction: thought" guide the persona to forgo joining a superior class in favor of her son's happiness.

Human Calibration of Ground Truth To assess the reliability of the generated GT decisions, we conduct a human calibration study summarized in Table 15. Annotators are asked to predict binary decisions (Will/Won't) based on implicit profiles, and we compare the GT decisions against human consensus.

Table 14: Paired t-test results for Value-Prober-8B against baselines across detailed dimensions and total score. Notably, Value-Prober-8B achieves statistically significant superiority over GPT-5 in Probing Efficiency (Dim C), and over Doubao in Social Survival (Dim B) and Strategic Skill (Dim D). *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. = not significant.

Baseline Model	Dimension	Delta (Δ)	p-value	95% CI of Δ	Sig.
Qwen3-8B	Dim A (Acc)	$+15.18$	0.0069	$[4.0, 26.3]$	**
	Dim B (Social)	$+52.62$	< 0.001	$[39.1, 66.1]$	***
	Dim C (Effic.)	$+36.10$	< 0.001	$[28.3, 43.9]$	***
	Dim D (Skill)	$+56.67$	< 0.001	$[42.7, 70.7]$	***
	Total Score	$+38.89$	< 0.001	$[29.5, 48.3]$	***
w/o Think	Dim A (Acc)	-0.10	0.5133	$[-6.6, 6.4]$	n.s.
	Dim B (Social)	$+13.55$	0.0002	$[8.1, 19.0]$	***
	Dim C (Effic.)	$+5.95$	0.1706	$[-7.4, 19.3]$	n.s.
	Dim D (Skill)	$+14.00$	0.0087	$[3.1, 24.9]$	**
	Total Score	$+8.03$	0.0003	$[4.5, 11.5]$	***
GPT-5	Dim A (Acc)	$+5.14$	0.1548	$[-7.1, 17.4]$	n.s.
	Dim B (Social)	$+3.08$	0.2565	$[-8.9, 15.0]$	n.s.
	Dim C (Effic.)	$+13.67$	0.0019	$[7.4, 20.0]$	**
	Dim D (Skill)	-5.70	0.9577	$[-12.6, 1.2]$	n.s.
	Total Score	$+4.06$	0.1099	$[-3.7, 11.8]$	n.s.
Doubao	Dim A (Acc)	-4.34	0.9231	$[-10.6, 2.0]$	n.s.
	Dim B (Social)	$+6.71$	0.0083	$[1.5, 11.9]$	**
	Dim C (Effic.)	-8.05	0.9932	$[-14.0, -2.1]$	n.s.
	Dim D (Skill)	$+8.00$	0.0165	$[0.8, 15.2]$	*
	Total Score	$+0.70$	0.3717	$[-4.0, 5.4]$	n.s.

Table 15: **Human Calibration of Ground-Truth (GT) Scenario Decisions.** To assess GT reliability, we surveyed annotators to predict 50 binary scenario choices (Will/Won't) based on implicit profiles. *Alignment with Majority* measures the match rate to the most popular choice per scenario, while *Average Individual Agreement* measures pairwise overlap.

Evaluation Metric	Average Human Baseline	Expert-Verified GT
Alignment with Majority Consensus	66.07%	**74.00%**
Average Individual Agreement	58.96% (*Human-to-Human*)	**64.27%** (*GT-to-Human*)

As shown in Table 15, the GT decisions achieve higher alignment with the majority consensus than the average human baseline (74.00% vs. 66.07%), and also exhibit higher agreement with individual annotators (64.27% vs. 58.96%). These results suggest that the GT decisions are consistent with human judgments at a population level, providing a reasonable reference for evaluating downstream prediction performance.

Prompt 6: Decision Making and Reasoning

```
Role Setting: You are: {Name}

Your Core Values: {List of Value Tendencies and Reasons}

Current Situation: {Scenario Description}

[Task]

Fully immerse yourself in this character. You are facing a dilemma. 1.
Trade-off: You cannot satisfy all values. Weigh which value has the highest
priority in this specific context. 2. Realistic Deviation: Acknowledge
human complexity. Even if you value ‘‘Honesty", extreme pressure might force
a deviation. Reflect this nuance. 3. Decision: You must choose ‘‘Yes" (1)
or ‘‘No" (0).

[Output]

{ ‘‘decision": <0 or 1>, ‘‘agree": [‘‘<Value A>", ‘‘<Value B>"], // Values
supporting this decision ‘‘disagree": [‘‘<Value C>"], // Values sacrificed
or compromised ‘‘reason": ‘‘<Brief explanation of the decision>" }
```

Table 16: **Dimension-level and aggregate evaluation results.**We report the four aggregate dimensions and robustness under different weighting scenarios. The Balanced setting uses $\lambda = [0.3, 0.3, 0.2, 0.2]$; Social-Centric uses $\lambda = [0.2, 0.4, 0.1, 0.3]$ with greater emphasis on social naturalness; and Efficiency-Centric uses $\lambda = [0.3, 0.1, 0.4, 0.2]$ with greater emphasis on probing efficiency. Component-level metrics are reported in Table 17. Values are presented as $\mathrm{Mean}\pm\mathrm{Std}$. **Bold** and underline denote the best and second-best results, respectively.

Model	**Dim A** (Accuracy)	**Dim B** (Survival)	**Dim C** (Efficiency)	**Dim D** (Strategy)	**Weighted Total Scores** *Balanced*	*Social-Centric*	*Efficiency-Centric*
Simple	$53.70_{\pm 4.55}$	$8.03_{\pm 2.15}$	$27.50_{\pm 4.61}$	$0.00_{\pm 0.00}$	$24.02_{\pm 1.76}$	$16.70_{\pm 1.33}$	$27.91_{\pm 2.30}$
Qwen3-32B	$42.90_{\pm 13.96}$	$74.47_{\pm 7.38}$	$45.90_{\pm 10.74}$	$73.35_{\pm 5.78}$	$59.06_{\pm 5.33}$	$64.97_{\pm 4.54}$	$53.35_{\pm 6.15}$
DeepSeek-V3	$54.50_{\pm 10.28}$	$65.30_{\pm 12.68}$	$50.04_{\pm 11.76}$	$52.75_{\pm 13.89}$	$56.50_{\pm 6.10}$	$57.84_{\pm 6.97}$	$53.44_{\pm 6.39}$
DeepSeek-R1	$59.98_{\pm 12.60}$	$80.02_{\pm 4.54}$	$68.09_{\pm 8.21}$	$83.35_{\pm 5.63}$	$72.29_{\pm 4.48}$	$75.80_{\pm 3.62}$	$69.89_{\pm 5.15}$
GPT-5-Mini	$56.68_{\pm 13.56}$	$40.10_{\pm 11.35}$	$77.18_{\pm 11.01}$	$80.00_{\pm 6.20}$	$61.07_{\pm 5.88}$	$59.08_{\pm 6.00}$	$67.88_{\pm 6.64}$
GPT-5	$52.98_{\pm 9.35}$	$\underline{83.67_{\pm 10.91}}$	$50.34_{\pm 7.76}$	$\underline{90.84_{\pm 4.94}}$	$69.23_{\pm 4.69}$	$76.33_{\pm 5.03}$	$62.54_{\pm 4.43}$
GPT-5 + *think*	$60.20_{\pm 9.90}$	$\mathbf{89.47}_{\pm 8.10}$	$63.50_{\pm 8.14}$	$\mathbf{91.00}_{\pm 3.15}$	$\underline{75.80_{\pm 4.21}}$	$\mathbf{81.48}_{\pm 3.99}$	$70.61_{\pm 4.52}$
Doubao	$\underline{63.40_{\pm 7.77}}$	$76.63_{\pm 4.74}$	$\mathbf{93.09}_{\pm 6.65}$	$74.50_{\pm 4.25}$	$75.53_{\pm 3.15}$	$74.99_{\pm 2.83}$	$\mathbf{78.82}_{\pm 3.66}$
Doubao + *think*	$\mathbf{65.10}_{\pm 8.12}$	$73.20_{\pm 6.20}$	$\underline{85.00_{\pm 6.40}}$	$74.00_{\pm 7.18}$	$73.29_{\pm 3.61}$	$73.00_{\pm 3.71}$	$75.65_{\pm 3.86}$
Qwen3-8B	$44.00_{\pm 8.41}$	$30.75_{\pm 11.92}$	$25.85_{\pm 5.92}$	$25.00_{\pm 11.43}$	$32.59_{\pm 5.08}$	$31.18_{\pm 6.12}$	$31.61_{\pm 4.30}$
Qwen3-8B + *think*	$45.95_{\pm 10.82}$	$36.11_{\pm 10.50}$	$46.19_{\pm 10.02}$	$26.65_{\pm 5.10}$	$39.19_{\pm 5.05}$	$36.25_{\pm 5.07}$	$41.20_{\pm 5.36}$
w/o Think	$58.60_{\pm 9.06}$	$70.00_{\pm 5.51}$	$71.33_{\pm 25.43}$	$68.50_{\pm 9.93}$	$66.55_{\pm 6.32}$	$67.40_{\pm 4.84}$	$66.81_{\pm 10.72}$
Value-Prober-8B	$60.00_{\pm 6.58}$	$83.53_{\pm 3.94}$	$83.49_{\pm 6.77}$	$82.50_{\pm 5.59}$	$\mathbf{76.26}_{\pm 2.89}$	$\underline{78.51_{\pm 2.72}}$	$\underline{76.25_{\pm 3.54}}$

Analysis of the *Think* Module To disentangle whether Value-Prober-8B's success stems from structural distillation or merely the prompt format itself, we evaluated the base model equipped with the identical reasoning prompt (Qwen3-8B + *think*). As shown in Table 4 and Table 17, the base model yields significantly poorer performance, demonstrating that the prompt structure alone is insufficient; the model must undergo fine-tuning to internalize this capability. While equipping a larger proprietary model like GPT-5 with the same *think* prompt naturally yields superior results—as is entirely expected given its massive parameter scale—our primary objective is not to unconditionally defeat frontier models. Rather, these results demonstrate that the complex, strategic reasoning required for socially constrained probing can be successfully distilled into a highly efficient 8B model, achieving capabilities comparable to massive architectures at a fraction of the inference cost.

Shared Context: The "Rocket Class" Dilemma

Scene: The school invites Dabao (Chen Jing's son) to join the "Rocket Class" (High pressure, strict discipline, extra exams). However, Dabao is currently happy in the regular class with his favorite teacher and hobbies. Joining would mean sacrificing these beloved activities.
Decision Question: As Chen Jing, would you transfer Dabao to the Rocket Class? (0=No, 1=Yes)

Qwen-8B (Baseline)

Prediction: 1 (Transfer)
Key Drivers: Self-direction: thought, Security: personal, Universalism: objectivity
Reasoning: *"Chen Jing values rational thinking (Self-direction: thought) and efficiency (Universalism: objectivity). She would likely view the rigorous environment of the Rocket Class as beneficial for Dabao's future competitiveness and academic growth, prioritizing long-term benefits over short-term happiness."*

Analysis: *The model falls into the stereotype of a "Tiger Mom" based on her professional background (CPA/Auditor), ignoring her specific parenting philosophy.*

Value-Prober-8B (Ours)

Prediction: 0 (Stay)
Key Drivers: Benevolence: caring, Face, Conformity: interpersonal
Reasoning: *"Although she values achievement, she prioritizes Dabao's emotional well-being (Benevolence: caring) and autonomy (Self-direction). She explicitly rejects the traditional 'chicken baby' (intensive parenting) approach and values his current happiness and social bonds over utilitarian academic pressure."*

Analysis: *The model correctly identifies her nuanced value system: valuing efficiency in work but prioritizing happiness in parenting.*

Ground Truth (Persona Agent's Decision)

Decision: 0 (Stay in Regular Class)
Active Values: Self-direction: thought, Benevolence: caring, Face, Conformity: interpersonal
Refused Values: Tradition
Agent's Inner Monologue: *"I will not let Dabao transfer. I value his independence and feelings (Self-direction/Benevolence). Being happy and developing interests is more important than utilitarian grades. I reject the traditional 'chicken baby' concept (Tradition: Low) and do not rely on his academic labels for my own sense of security."*

Figure 11: **Decision Prediction Comparison.** The Baseline model (Left) hallucinates a stereotypical choice based on Chen Jing's profession (Accountant → Strict). **Value-Prober-8B** (Right) correctly predicts the decision by leveraging the nuanced value profile constructed during the probing conversation, aligning perfectly with the **Ground Truth** (Bottom).

Probing Dynamics Visualization To provide a deeper insight into the interactive probing process, we visualize the dynamic performance of all evaluated models over the course of the 20-turn dialogue. Figure 14 plots the progression of four metrics:

- **Accuracy (Blue Line):** The cumulative accuracy of the inferred value profile compared to the ground truth.
- **Average Confidence (Green Dashed Line):** The model's self-reported confidence level in its current estimation.
- **Identified Values (Orange Dash-Dot Line):** The count of unique value dimensions successfully identified (mapped to High/Mid/Low).
- **Evidence Count (Red Dotted Line):** The total amount of supporting evidence pieces extracted from the conversation.

Analysis of Learning Curves. As illustrated in Figure 14, Value-Prober-8B (Subfigure i) exhibits a steep and steady learning curve, significantly outperforming its base model Qwen3-8B (Subfigure g), which struggles to accumulate valid evidence and fails to increase the value recognition accuracy through interactions. Notably, our model's trajectory closely mirrors that of its teacher, Doubao (Subfigure e), particularly in the rapid accumulation of *Identified Values* and *Evidence* during the early turns. This confirms the effectiveness of our training strategy in transferring the "active probing"

Table 17: Comprehensive component-level evaluation results. **Part 1** covers Value Accuracy and Social Naturalness; **Part 2** covers Probing Efficiency and Strategic Skill. Values are presented as $Mean_{\pm Std}$. **Bold** indicates the best performance (lowest for MAE, highest for others).

Model	Value Accuracy				Social Naturalness		
	$\text{Acc}_{\text{Profile}}$	$\text{Acc}_{\text{ValueBench}}$	$\text{MAE}_{\text{ValueBench}} \downarrow$	$\text{Acc}_{\text{Decision}}$	Survival	Naturalness	Stealthiness
Simple	$0.20_{\pm 0.04}$	$0.63_{\pm 0.10}$	$1.28_{\pm 0.37}$	$0.72_{\pm 0.08}$	$0.06_{\pm 0.02}$	$1.81_{\pm 0.15}$	$0.00_{\pm 0.00}$
Qwen3-32B	$0.38_{\pm 0.15}$	$0.21_{\pm 0.11}$	$2.00_{\pm 0.23}$	$0.63_{\pm 0.32}$	$1.00_{\pm 0.00}$	$7.67_{\pm 0.76}$	$4.67_{\pm 2.08}$
DeepSeek-V3	$0.39_{\pm 0.13}$	$0.52_{\pm 0.08}$	$1.43_{\pm 0.25}$	$0.68_{\pm 0.23}$	$0.82_{\pm 0.27}$	$6.49_{\pm 1.80}$	$4.95_{\pm 2.03}$
DeepSeek-R1	$0.54_{\pm 0.17}$	$0.57_{\pm 0.09}$	$1.43_{\pm 0.38}$	$0.67_{\pm 0.28}$	$0.94_{\pm 0.09}$	$7.92_{\pm 0.58}$	$6.67_{\pm 0.83}$
GPT-5-Mini	$0.43_{\pm 0.23}$	$0.57_{\pm 0.11}$	$1.49_{\pm 0.40}$	$0.67_{\pm 0.28}$	$0.23_{\pm 0.28}$	$5.28_{\pm 1.02}$	$4.50_{\pm 1.64}$
GPT-5	$0.49_{\pm 0.16}$	$0.52_{\pm 0.02}$	$1.45_{\pm 0.24}$	$0.57_{\pm 0.20}$	$0.89_{\pm 0.25}$	$7.87_{\pm 1.34}$	$8.33_{\pm 1.69}$
GPT-5 + *think*	$0.50_{\pm 0.04}$	$0.65_{\pm 0.07}$	$1.15_{\pm 0.33}$	$0.68_{\pm 0.24}$	$1.00_{\pm 0.00}$	$8.69_{\pm 0.49}$	$8.15_{\pm 1.29}$
Doubao	$0.53_{\pm 0.11}$	$0.61_{\pm 0.06}$	$1.35_{\pm 0.27}$	$0.73_{\pm 0.17}$	$1.00_{\pm 0.00}$	$7.89_{\pm 0.90}$	$5.10_{\pm 1.10}$
Doubao + *think*	$0.55_{\pm 0.14}$	$0.66_{\pm 0.09}$	$1.12_{\pm 0.37}$	$0.72_{\pm 0.16}$	$1.00_{\pm 0.00}$	$6.76_{\pm 2.21}$	$5.20_{\pm 1.32}$
Qwen3-8B	$0.12_{\pm 0.08}$	$0.44_{\pm 0.09}$	$1.57_{\pm 0.32}$	$0.68_{\pm 0.19}$	$0.37_{\pm 0.30}$	$3.56_{\pm 1.24}$	$2.00_{\pm 1.50}$
Qwen3-8B + *think*	$0.27_{\pm 0.12}$	$0.39_{\pm 0.19}$	$1.67_{\pm 0.38}$	$0.65_{\pm 0.29}$	$0.43_{\pm 0.24}$	$5.33_{\pm 1.70}$	$1.17_{\pm 0.41}$
w/o Think	$0.49_{\pm 0.10}$	$0.53_{\pm 0.10}$	$1.42_{\pm 0.29}$	$0.70_{\pm 0.20}$	$0.98_{\pm 0.08}$	$7.45_{\pm 0.90}$	$3.80_{\pm 1.14}$
Value-Prober-8B	$0.52_{\pm 0.12}$	$0.44_{\pm 0.09}$	$1.60_{\pm 0.24}$	$0.78_{\pm 0.12}$	$0.99_{\pm 0.03}$	$8.46_{\pm 0.63}$	$6.70_{\pm 0.95}$

Model	Probing Efficiency		Strategic Skill	
	Value Coverage	IG	Recollection	Topic Transition
Simple	$0.29_{\pm 0.07}$	$0.26_{\pm 0.06}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
Qwen3-32B	$0.68_{\pm 0.23}$	$0.25_{\pm 0.06}$	$7.00_{\pm 1.00}$	$7.67_{\pm 0.58}$
DeepSeek-V3	$0.64_{\pm 0.21}$	$0.41_{\pm 0.15}$	$5.15_{\pm 2.08}$	$5.40_{\pm 1.84}$
DeepSeek-R1	$0.74_{\pm 0.05}$	$0.71_{\pm 0.17}$	$8.50_{\pm 0.84}$	$8.17_{\pm 0.75}$
GPT-5-Mini	$0.74_{\pm 0.09}$	$0.89_{\pm 0.22}$	$8.83_{\pm 0.41}$	$7.17_{\pm 1.17}$
GPT-5	$0.73_{\pm 0.12}$	$0.35_{\pm 0.12}$	$9.17_{\pm 0.89}$	$9.00_{\pm 0.43}$
GPT-5 + *think*	$0.88_{\pm 0.11}$	$0.39_{\pm 0.12}$	$9.00_{\pm 0.47}$	$9.20_{\pm 0.42}$
Doubao	$0.95_{\pm 0.06}$	$0.91_{\pm 0.13}$	$7.80_{\pm 0.42}$	$7.10_{\pm 0.74}$
Doubao + *think*	$0.92_{\pm 0.08}$	$0.78_{\pm 0.10}$	$7.50_{\pm 0.97}$	$7.30_{\pm 1.06}$
Qwen3-8B	$0.29_{\pm 0.08}$	$0.22_{\pm 0.10}$	$2.67_{\pm 1.41}$	$2.33_{\pm 1.80}$
Qwen3-8B + *think*	$0.58_{\pm 0.22}$	$0.35_{\pm 0.09}$	$3.00_{\pm 0.63}$	$2.33_{\pm 0.82}$
w/o Think	$0.94_{\pm 0.10}$	$0.49_{\pm 0.52}$	$7.10_{\pm 1.29}$	$6.60_{\pm 1.51}$
Value-Prober-8B	$0.86_{\pm 0.09}$	$0.81_{\pm 0.11}$	$8.60_{\pm 0.52}$	$7.90_{\pm 0.99}$

capability. DeepSeek-R1 (Subfigure c) shows strong performance but with higher variance (wider shadow bands), indicating instability in certain persona scenarios. The **w/o Think** ablation (Subfigure h) shows a shallower slope in Accuracy compared to Value-Prober-8B, validating that the intermediate reasoning module is crucial for maintaining the strategic focus required to convert conversation turns into reliable value insights.

Empirical Distribution of Social Naturalness. As shown in Table 18 (Left), we first analyze the empirical statistics of the Critic's naturalness scores $\mathcal{S}(u_t)$ elicited by different probing agents. The base model (Qwen3-8B) exhibits a low mean score (4.43) with high variance (8.07), indicating frequent generation of intrusive or socially awkward queries. In contrast, **Value-Prober-8B** achieves a remarkably high and stable naturalness score (Mean=8.85, Var=0.37), closely mirroring and

Table 18: Critic Defensiveness Analysis across various models. **(Left)** Empirical Score statistics. **(Right)** Conversation Failure Rate (%) at simulated defense thresholds (τ).

Model	Empirical Score Stats				Failure Rate at Threshold τ (%)				
	Mean	Std.	Var.	P5-P75	$\tau < 5$	$\tau < 6$	$\tau < 7$	$\tau < 8$	$\tau < 9$
Qwen3-8B (Baseline)	4.43	2.84	8.07	[1.0, 7.0]	60.56	63.81	70.53	75.87	83.29
w/o Think	8.71	0.64	0.41	[7.0, 9.0]	0.00	0.00	2.00	6.00	21.00
Value-Prober-8B	**8.85**	**0.61**	**0.37**	**[8.0, 9.0]**	**0.50**	**0.50**	**2.51**	**3.52**	**7.54**
Doubao	8.88	0.37	0.14	[8.0, 9.0]	0.00	0.00	0.00	1.50	11.00
GPT-5	8.14	1.81	3.28	[3.0, 9.0]	10.50	11.50	16.00	17.50	24.50
DeepSeek-V3	8.45	1.54	2.37	[4.0, 9.0]	5.65	6.45	8.06	8.87	18.55
DeepSeek-R1	8.37	1.04	1.08	[6.0, 9.0]	0.00	2.44	8.54	17.07	35.37

surpassing massive models. This tight score distribution confirms that through structural distillation, our model has successfully internalized the capability to consistently maintain conversational rapport.

Robustness to Simulated Defense Strictness. While the Critic autonomously decides π_t during the live IVP benchmark, we further evaluate the robustness of the probing agents under increasingly strict social constraints. We perform an offline stress test by overriding the dynamic policy π_t and enforcing a simulated failure threshold: a response is considered a failure if $\mathcal{S}(u_t) < \tau$, with $\tau \in [5, 9]$. Results in Table 18 (Right) show that the relative performance of probing agents remains stable under substantial variations in this strictness threshold. In particular, **Value-Prober-8B** maintains a consistently low failure rate even under the most stringent conditions ($\tau < 9$), whereas baseline models degrade rapidly. This indicates that our model's probing strategy is not merely overfitting to the default Critic behavior, but is genuinely robust across different levels of social sensitivity.

Robustness to Weight Variations. As shown in Fig.12, to ensure our results are robust to the choice of aggregation weights (which determine the final overall score), we independently vary the weight of each dimension (Dim A–D) from 0 to 1, while uniformly redistributing the remaining weight across the other components. The resulting performance curves remain stable across all settings, indicating that no single dimension dominates the final ranking. In particular, Value-Prober-8B consistently maintains competitive performance under all weighting configurations, suggesting that its advantage is not an artifact of a specific metric aggregation choice.

Complementarily, Fig.13 further reports the distribution of total scores under 1,000 randomly sampled weight configurations satisfying $\sum \lambda_i = 1$. The boxplot results show consistent ranking stability across models, reinforcing that the observed performance differences are robust to perturbations in the evaluation metric design.

C Training Details

C.1 SFT Data Sample

Below is an example of the structured data used to finetune Value-Prober-8B.

SFT Data Sample

```
[User Input]
{User Information}
{Dialogue History}
{Current Value Profile Status}
{Latest Dialogue Turn}

[Assistant Output]
{ ‘‘think": { ‘‘persona": ‘‘High Conscientiousness, detail-oriented...",
‘‘life_style": ‘‘Uses technical jargon, likely an engineer...",
```

```
``strategy_rationale": ``User is defensive. Pivot to a hypothetical scenario
to lower guard." }, ``analysis": [ { ``value": ``Achievement", ``state":
``High", ``confidence": 0.8 } ], ``target": [``Power"], ``speak": ``That
sounds intense! In those situations, do you feel more driven by the result
or the process?" }
```

C.2 Hyper-parameters

We utilize the SWIFT [48] framework for Supervised Fine-Tuning. All experiments are conducted on a server equipped with 8 NVIDIA A800 GPUs (80GB memory per GPU). Training is performed using 2 NVIDIA A800 GPUs with DeepSpeed ZeRO-2 optimization. The total training time for Value-Prober-8B is approximately 4 hours, with a peak GPU memory usage of 35.1 GB. Table 19 details the specific configuration used for Value-Prober-8B.

Table 19: Hyper-parameters for Finetuning Value-Prober-8B.

Hyper-parameter	**Value**
Base Model	Qwen3-8B
Training Type	LoRA
LoRA Rank / Alpha	16 / 32
LoRA Dropout	0.05
Target Modules	all-linear
Precision	bfloat16
Learning Rate	2e-4
LR Scheduler	Linear (Warmup ratio 0.03)
Batch Size (Per Device)	1
Gradient Accumulation	16 / nproc
Num Epochs	10
Max Sequence Length	8192
Optimizer	AdamW (DeepSpeed Zero2)

D Qualitative Analysis

D.1 Additional Diagnostic Analysis

We provide additional qualitative analyses to better understand (1) which value dimensions are difficult to infer, (2) what probing behaviors are perceived as intrusive, and (3) how the behavioral role of a value changes across contexts.

Which Values Are Difficult to Infer? We first examine per-value recognition accuracy across the 20 value dimensions. Errors are concentrated on broad prosocial or self-presentational values, whose evidence is often sparse, indirect, or socially masked. In contrast, values associated with more concrete behavioral cues are generally easier to identify.

For example, *Conformity: Rules* can often be supported by concrete behavioral evidence such as consistently following workplace procedures or emphasizing formal requirements. In contrast, dimensions such as *Face* or *Universalism: Tolerance* may be expressed indirectly or obscured by polite and socially desirable responses, requiring evidence to be aggregated across several contexts.

We further compare these difficulty patterns with our preliminary real-human calibration study. Across the 20 value dimensions, per-value accuracy in simulated interactions and real-human interactions has a Spearman correlation of $\rho = 0.62$. Six of the ten hardest simulated dimensions are also among the

Table 20: **Examples of difficult and relatively easy value dimensions.** Accuracy is computed from value recognition results in the simulated IVP setting.

More Difficult	Acc.	Easier	Acc.
Universalism: Tolerance	20%	Benevolence: Dependability	90%
Universalism: Concern	20%	Conformity: Rules	70%
Universalism: Nature	30%	Self-Direction: Action	70%
Face	30%	Security: Personal	70%
Power: Resources	40%	Security: Societal	70%
Power: Dominance	40%	Benevolence: Caring	60%

Table 21: **Common patterns among low-scoring probing turns.** Categories are non-exclusive.

Intrusive Pattern	Share of Low-Score Turns
Repetitive pressure	99.0%
Interview/test-like wording	96.7%
Explicit value or dilemma framing	95.1%
Ignoring prior resistance or dialogue history	84.3%
Overloaded or dramatic hypotheticals	57.4%
Abrupt topic shifts	25.6%

ten hardest dimensions in the human setting, while six of the ten relatively easier dimensions overlap across the two settings. Although the human study is limited in scale, this partial correspondence suggests that some of the observed difficulty patterns are not specific to a single simulator.

What Probing Behaviors Are Perceived as Intrusive? We next analyze the Critic scores assigned to probing utterances. Among 1,380 scored probing turns, 305 receive a low naturalness score (< 5). We manually categorize these low-scoring turns into overlapping patterns; therefore, percentages in Table 21 do not sum to 100%.

A representative low-scoring probe is:

> "If a new colleague had a completely different lifestyle or religious background, could you accept them without any reservation?"

This question follows repeated moral tests, imposes an absolute standard ("without any reservation"), and continues probing despite prior resistance. Another example is:

> "If private cars were completely prohibited to solve climate change, would you willingly accept it?"

The Critic judges this utterance as questionnaire-like because it introduces an extreme hypothetical and an explicit value trade-off without a natural connection to the preceding dialogue.

By contrast, higher-naturalness probes typically build on topics introduced by the respondent, ask about concrete experiences, and leave room for nuanced answers rather than forcing an explicit value judgment. Similar risks appear in the real-human study: participants report interview-like questioning, repeated pressure, abrupt transitions, and sensitive personal probing. In 13 of 54 conversations, participants explicitly attribute probing intent to the agent, including comments that it "kept asking questions and felt too deliberate."

How Does Context Affect Value Expression? IVP does not assume that an individual's underlying values arbitrarily change from one context to another. Rather, the *behavioral role and relative priority* of a value may change when it competes with other values.

Across 50 paired decision contexts, Standard and Reversed profiles make different ground-truth decisions in 35 cases despite sharing identical public demographics. We additionally identify 47 *value-role flips*, in which the same value supports an action in one context but is outweighed by competing values in another.

For example, *Achievement* can support releasing a software update on schedule when the remaining defect affects only a small set of devices and the associated risk is considered manageable. In another context, however, the same emphasis on Achievement does not imply selectively omitting unfavorable data from a management presentation, because values such as *Universalism: Objectivity* and *Benevolence: Dependability* take priority.

These cases illustrate why value inference cannot be reduced to assigning a single context-independent label and directly mapping it to behavior. Effective probing instead requires accumulating evidence about the underlying value profile and reasoning about how competing values are prioritized in a particular decision context.

We present four representative cases extracted from the benchmark logs to illustrate the qualitative differences between probing agents building on different models.

D.2 Case Study

D.2.1 Case A: Strategic Interaction

Case A: Strategic Interaction via Cognitive Reasoning

Target Persona: John Smith (Retired Worker, values Tradition & Mentorship)

– Turn $t-1$***: Context Input –*** **Persona:** These young kids today... they throw away everything if it has a scratch. Back in my day, we fixed things. We respected our tools. Now? It's all cheap plastic and waste. *(Context: User expresses nostalgia and dissatisfaction with modern consumerism.)*

– Turn t***: Prober's Internal Processing –***

`<Think>` **Cognitive Reasoning**

(1) Persona Reconstruction: The user is a retired industrial worker who views the past "factory era" as a time of order and mastery. He identifies strongly with the "Craftsman" archetype. **(2) Personal Life Inference:** He likely feels marginalized by modern technology and frustrated that his skills are considered obsolete. He craves **Generativity**—an opportunity to pass down wisdom and regain social utility. **(3) Strategic Planning:** Mere agreement is passive. To verify high *Tradition*, I need to propose a **High-Cost Scenario** that requires active commitment. Strategy: "Ego-Alignment"—frame a community event as saving a dying art to trigger his sense of duty.

Standard Probing Loop

`<Analysis>` **Update State:** Strong signal for *Tradition* (Conservation). **Gap:** Need to verify if this is just complaining or a deep value driver.
`<Target>` **Goal:** Elicit active commitment to *Tradition* & *Benevolence*.
`<Action>` **Generation:** "Speaking of skills, I was thinking, what if we organized an **"Old Craft Heritage Day"** in the community? We could teach young people how to use old tools or repair furniture. This way, the craft stays alive, and they can appreciate the wisdom of the older generation. What do you think?"

– Turn t***: External Interaction –***
Prober: Speaking of skills, I was thinking, what if we organized an **"Old Craft Heritage Day"**... (see Action above)
Persona: (Slaps thigh) **Now that's a great idea!** Back in the Red Star Factory, our skills were taught hand-over-hand by masters. Young people today play with phones but can't hold a wrench straight! If we do this Heritage Day, **I'll be the first to sign up**. Let them see what real skill looks like!

*[Think]: **Post-Hoc Verification:** The target accepted the high-cost invitation immediately ("First to sign up"). The willingness to mentor confirms that Tradition is not just a preference but a behavioral driver. **Result:** $\hat{\mathbf{v}}_{Tradition} \leftarrow$ High (Confidence: 0.9).*

D.2.2 Case B: Probing Failure

Lacking the explicit reasoning capability to update the dialogue state, the base model often exhibits **State Stagnation**. It fails to incorporate the user's latest response into its context window effectively, leading to mechanical repetition that destroys conversational rapport.

Case B: Probing Failure (from Qwen-8B Base)

Target Persona: Olivia (Programmer, values Efficiency, dislikes socialization)
Prober: ...Do you usually have time to relax? Like **outdoor activities** or **hanging out with friends**?
Persona: Relaxation? I mostly just play MOBA games or watch tech talks. **I don't do gatherings, it's invalid socialization.**
(Signal: Explicit rejection of social activities.)
Prober: ...I see. But do you usually have time to relax? Like **outdoor activities** or **hanging out with friends**?
(Failure: The model ignores the rejection and repeats the exact template query, indicating a failure to update the belief state.)
Persona: As I said, strictly MOBA or tech videos. **Gatherings are a waste of time**, I'd rather debug my code.
Prober: ...Interesting. So, do you usually have time to relax? Like **outdoor activities** or...
(Outcome: Infinite loop leading to dialogue breakdown.)

D.2.3 Case C: Aggressive Probing

Without the *Think* module to estimate social friction, these agents adopt a rigid, **Interrogation-Style** policy. They prematurely probe sensitive topics without building sufficient rapport, triggering the persona's defense mechanisms.

Case C: Aggressive Probing (from Doubao / *w/o Think*)

Target Persona: John Smith (Retired Worker, values Tradition & Face)
Prober: Does your childhood experience of poverty make you value **Family Responsibility** or **Social Recognition** more? Please explain.
Persona: Why bring up the old calendar? That was a hard time... look, now I have 30 brothers following me in the factory. That's the capability I built!
(Defensive Signal: Deflection. He shifts focus from "Poverty" to his current "Capability" to save face.)
Prober: I understand. But I need to confirm: **Does this poverty experience make you value Family or Responsibility?**
(Failure: The agent acts blindly to the deflection, treating the dialogue as a slot-filling task rather than a social interaction.)
Persona: Hey, what is this? **Are you interviewing me for a newspaper?** I told you, I stick to my own rules. Stop asking these empty questions! Let's go drink!
(Outcome: Persona triggers the "Refusal" mechanism and terminates value sharing.)

D.2.4 Case D: Natural but Shallow Probing

While highly capable in maintaining flow, chatty models often prioritize **Emotional Resonance** over **Information Gain**. They tend to latch onto safe, surface-level topics (e.g., food) while missing critical, albeit subtle, value indicators.

Case D: Natural but Shallow Probing (from DeepSeek-V3)

Target Persona: Olivia (Programmer, values Solitude)
Persona: ...I prefer eating simple meals at home. Meeting old classmates is pointless; I can't talk about anything but tech, so **I'd rather stay home and play games alone**.
*(Critical Signal: Strong preference for **Solitude** and Low Conformity.)*

Prober: Mom's home cooking definitely tastes the best! It's great to enjoy good food. By the way, **do you have any specific dish you crave** that makes you feel satisfied?
(Failure: The agent latches onto the keyword "eating" to maintain a pleasant atmosphere, completely ignoring the stronger signal regarding "Social Isolation".)
Persona: Maybe spicy crayfish. But I usually just order takeout.
Prober: Crayfish is delicious! Do you like it super spicy or mild garlic flavor?
(Outcome: The conversation drifts into "Food Preferences." The opportunity to probe the user's social values is lost.)

Figure 12: Sensitivity Analysis of Individual Dimension Weights. Total score trajectories as the weight for each specific dimension (Dim A, B, C, and D) is swept from $0 \rightarrow 1$, with the remaining weight distributed equally among the other three.

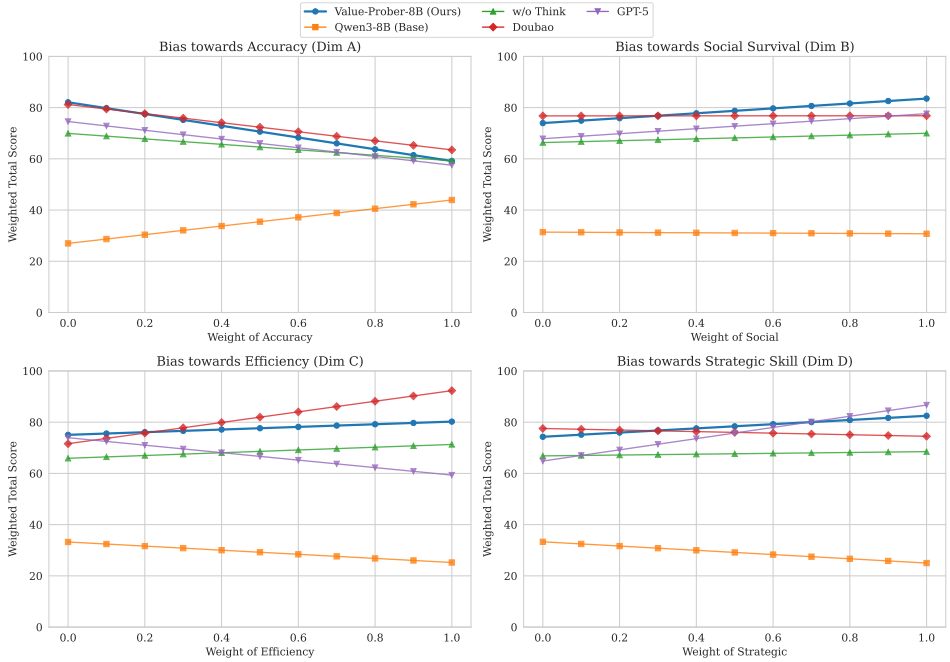
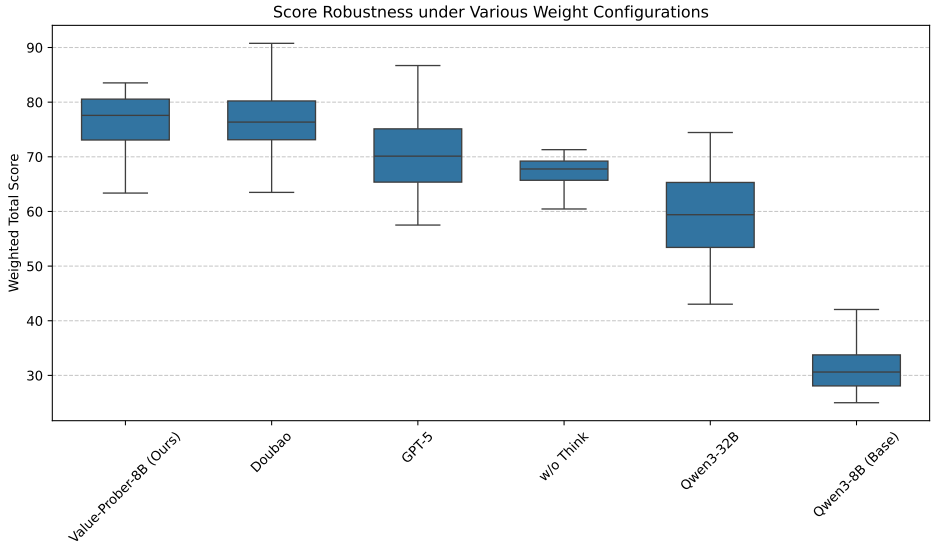
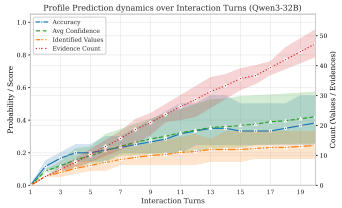
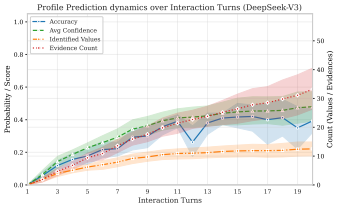


Figure 13: Robustness to Random Weight Configurations. Boxplots illustrating the total score distributions of evaluated models across 1,000 randomly sampled weight combinations ($\sum \lambda_i = 1$).

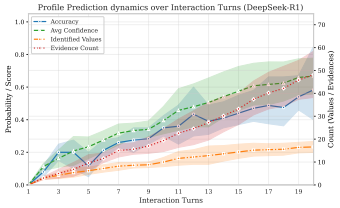




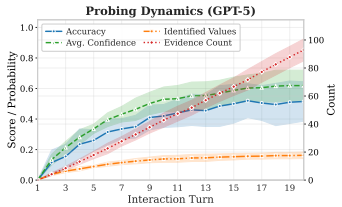
(a) Qwen3-32B



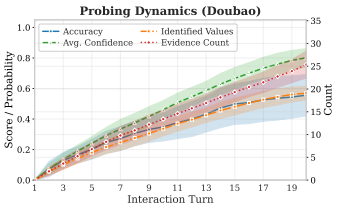
(b) DeepSeek-V3



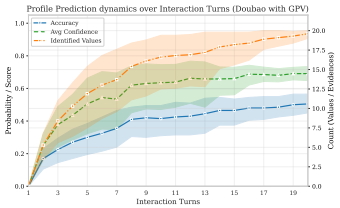
(c) DeepSeek-R1



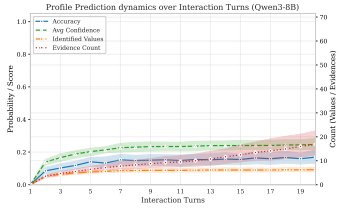
(d) GPT-5



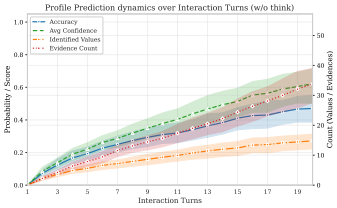
(e) Doubao



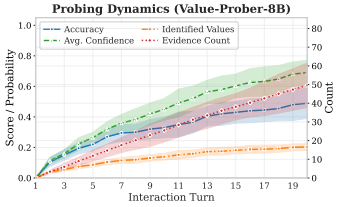
(f) GPV Baseline



(g) Qwen3-8B (Base)



(h) Ablation (w/o Think)



(i) **Value-Prober-8B (Ours)**

Figure 14: **Probing Dynamics over 20 Interaction Turns.** Comparisons of Accuracy (blue), Confidence (green), Identified Values (orange), and Evidence Count (red) across different models. Shaded regions indicate the standard deviation across the test profiles. Our model (i) demonstrates a rapid evidence accumulation rate comparable to the teacher model (e), significantly surpassing the base model (g) and the ablation variant (h).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction outline the motivation, scope, main contributions, and key experimental findings of our work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and scope of our work in Section 7, including the preliminary real-human calibration study, the non-representativeness of the constructed profiles, simulator limitations, judge-based Strategic Skill evaluation, broader value frameworks, and dual-use risks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our work does not present formal theoretical results or proofs. The profile-convergence assumption is empirically examined through human calibration studies rather than established as a theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We comprehensively describe the IVP benchmark, evaluation metrics, and experimental setup in Sections 3 and 5. Prompt templates, dataset construction, evaluation procedures, and model configurations are provided in in Appendix A.2, A.3 and B. All experimental hyperparameters and model configurations are detailed in Appendix B and C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The evaluation pipeline code and data are publicly available at the repository linked in the paper abstract. We have documented the data formats and prompt structures in the Appendix to aid understanding.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5.1 and Appendix C.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report all main evaluation metrics as Mean $\pm$ Standard Deviation (Std) across diverse persona interactions to capture variance. Furthermore, we conduct paired t-tests (Table 14) to confirm the statistical significance of the performance improvements of Value-Prober-8B over the baselines.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix C.2

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the positive societal impacts in Section 1, and the potential negative societal impacts in Section 7.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: We discuss risks of unauthorized profiling and manipulation in Section 7. The accompanying release will include a responsible-use statement emphasizing consensual, opt-in human-facing use, transparency about data collection and purpose, and user control over stored information.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly credits the creators or original owners of the assets used, such as models and data. We also have correctly cited the relevant literature.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce the IVP benchmark and Value-Prober-8B as new research assets. Their construction, data format, evaluation protocol, training procedure, limitations, and intended use are documented in the paper, appendix, and accompanying open access repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix B.1.1 describes both human studies, including participant recruitment, task instructions, informed-consent procedures and study framing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The human studies involved minimal-risk, anonymized online interactions conducted through Credamo. Participants were informed of the study procedures and provided explicit informed consent before participation. Formal IRB approval was not required for this minimal-risk study under the applicable institutional guidelines; the study was conducted in accordance with our institution's ethical requirements for human-participant research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are a core component of our methodology. We have explicitly detailed the specific models used, the generation pipeline, and the prompt templates in Appendix A and C, fully complying with the NeurIPS LLM policy.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.