Don't Box Me In: Dynamic Cultural Adaptation and Cognitive Tracking for Social Understanding

Anonymous ACL submission

Abstract

Social interaction increasingly takes place in multicultural settings, where individuals may draw on multiple cultural influences and adapt their communicative behavior across contexts. Despite recent advances in equipping Large Language Models (LLMs) with social understanding capabilities, existing approaches often model culture as a static demographic attribute, limiting their ability to accommodate hybrid and dynamically expressed communicative preferences. Therefore, in this paper, we propose **DyCAC**, a training-free framework that achieves fluid social alignment by incorporating Dynamic Cultural Adaptation with continuous Cognitive tracking. Rather than inferring a fixed cultural identity, DyCAC models culturally relevant communicative preferences as a time-varying mixture of population-level cultural reference profiles. This reference-based representation is further calibrated using dialogue-style signals observed in the ongoing interaction, enabling the model to capture both composite cultural influences and turn-level shifts in communicative behavior. In parallel, a memory module driven by Theory of Mind (ToM) continuously tracks the cognitive states of the interlocutor. Extensive experiments on interactive social and cultural benchmarks demonstrate the superiority of our approach. The proposed framework outperforms existing baselines, exhibiting enhanced social intelligence and broad adaptability across varied multicultural contexts.

1 Introduction

Social interaction fundamentally shapes how humans navigate shared environments, coordinate goal-directed actions, and construct mutual understanding (Hari and Kujala, 2009). A cornerstone of this capability is social understanding, which is the capacity to maintain coherent mental models of interlocutors and generate responses that align with interpersonal, emotional, and cultural expectations. Therefore, equipping Large Language Models (LLMs) with robust social abilities has emerged as a pivotal frontier, advancing research across social dialogue (Zhou et al., 2024; Kong et al., 2025; Yu et al., 2025; Zhang et al., 2025a; Bhattacharyya et al., 2026), and culturally grounded reasoning (AlKhamissi et al., 2024; Li et al., 2024b; Ki et al., 2025). To this end, numerous studies have introduced specific mechanisms, including Theory of Mind (ToM) (Kim et al., 2025; Zhang et al., 2025b) to track cognitive states, and social-oriented training paradigms (Wang et al., 2024; Liu et al., 2025b) to optimize the interaction process.

Despite this progress, many existing approaches still underemphasize a crucial factor in social interaction: cultural context. Culture shapes how people interpret politeness, directness, hierarchy, uncertainty, cooperation, and emotional expression (Hofstede, 2011). Therefore, models ignoring cultural context may produce responses that are linguistically fluent but socially misaligned. Recognizing this gap, recent research has focused on developing culturally aware language models (Li et al., 2024a; Feng et al., 2025; Guo et al., 2025; Havaldar et al., 2025; Liu et al., 2025a; Xu et al., 2025a). For instance, NileChat (El Mekki et al., 2025) proposes region-specific models tailored to local communities, while CuMA (Sun et al., 2026) conditions generation on demographic profiles. However, these methods treat culture as an oversimplified, static attribute or a deterministic categorical label, which fails to reflect modern multicultural realities. Influenced by globalization and diverse interactional experiences, individuals are not necessarily shaped by a single homogeneous cultural background. Instead, their communicative preferences may reflect a composite repertoire (Coste et al., 2009) formed through exposure to multiple cultural norms, languages, and communities. Moreover, even when individuals have relatively stable cultural experi-

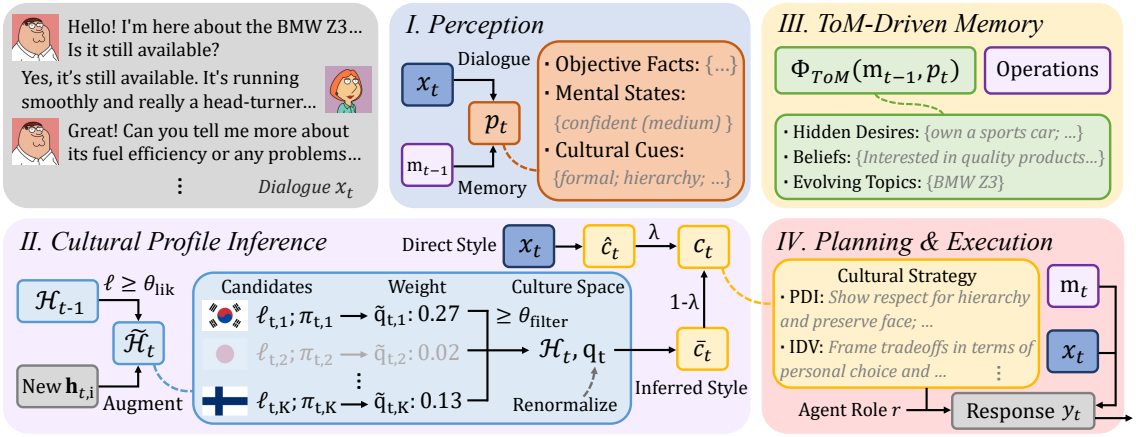

Figure 1: An overview of the DyCAC framework. x_t and y_t are the dialogue input and generated response, respectively. p_t is the perceptual state, and m_t is the memory updated by reasoner Φ_{ToM}. For cultural inference, variables $\mathcal{H}_t, \pi_{t,i}, l_{t,i}, q_{t,i}$ represent the culture space, cultural prior, compatibility score, and normalized weight, respectively. The final cultural profile $\mathbf{c}_t$ interpolates between the inferred profile $\bar{\mathbf{c}}_t$ and direct style $\hat{\mathbf{c}}_t$ via weight λ.

ences, their situational cultural alignment, namely the communicative strategies they lean toward in a specific context, can shift dynamically throughout a conversation. Depending on the interlocutor, goal, power relation, and emotional state, speakers may dynamically adapt their communication styles and cultural orientations. This view aligns with Communication Accommodation Theory (Giles et al., 1991), which suggests that people adapt their communicative behavior to evolving social contexts.

To address this limitation, in this paper, we propose **DyCAC**, a training-free framework that achieves fluid social alignment by integrating Dynamic Cultural Adaptation with continuous Cognitive tracking. Unlike prior approaches that treat culture as a static identity constraint, our framework formulates cultural adaptation as a continuous process of belief revision. Specifically, a Perception module first extracts objective facts, mental-state signals, and cultural cues from the ongoing dialogue. Building upon this shared evidence, DyCAC performs two complementary processes. First, it maps cultural cues to a dynamically maintained set of diverse cultural reference profiles and updates their relevance weights across dialogue turns. Rather than predicting a fixed cultural identity, this dynamic representation is used to derive a soft cultural profile that reflects the interlocutor's potentially hybrid and evolving communicative preferences. DyCAC further combines this reference-based mixture with directly observed dialogue-style signals, enabling it to capture both composite cultural influences and turn-level shifts in cultural expression. Second, DyCAC maintains a ToM-Driven Memory architecture that continuously tracks the interlocutor's evolving epistemic states, including beliefs, desires, and intentions. Finally, the inferred cultural calibration state and cognitive-state representation jointly guide a planning-and-execution module to generate situated, goal-directed responses. Extensive experiments on social and cultural benchmarks validate the effectiveness of our framework. It demonstrates enhanced social intelligence in interactive dialogue scenarios and robust generalization across diverse cultural contexts, outperforming other training-free and training-based baselines.

2 Methodology

In this section, we detail our proposed DyCAC framework, which decomposes social dialogue generation into four interconnected modules. We present an overview of the framework in Figure 1. More details are provided in Appendix A.

2.1 Perception

Our framework first employs a Perception module to extract compact representations of social and contextual cues from the ongoing dialogue. Given the current interaction input x_t and the past memory m_{t-1}, the module produces a perceptual state:

$$p_t = \Phi_{\text{perc}}(x_t, m_{t-1}). \quad (1)$$

The state p_t contains three categories of evidence: (1) objective facts capturing explicit situations and events; (2) mental-state signals reflecting the immediate intent, affective states, and salient values of the interlocutor; and (3) cultural cues summarizing

Algorithm 1 Dynamic Cultural Adaptation

Input: Global cultural pool $\mathcal{H}$, perceptual state p_t, current utterance x_t, previous cultural profile $\mathbf{c}_{t-1}$, previous culture space $\mathcal{H}_{t-1}$

Output: Final cultural profile $\mathbf{c}_t$, updated culture space $\mathcal{H}_t$

Notation: $\ell_{t,i}$: compatibility score, $\pi_{t,i}$: cultural prior

if $t = 1$ **then**
  $\mathcal{H}_1 \leftarrow \mathrm{InitTopK}(p_1, \mathcal{H}, K)$
else
  $\tilde{\mathcal{H}}_t \leftarrow \mathrm{AugmentToK}(\{\mathbf{h}_{t-1,i} \in \mathcal{H}_{t-1} \mid \Phi_{\mathrm{comp}}(p_t, x_t, \mathbf{h}_{t-1,i}) \geq \theta_{\mathrm{lik}}\}, p_t, x_t, K)$
end if
for $\mathbf{h}_{t,i} \in \tilde{\mathcal{H}}_t$ **do**
  $\ell_{t,i} \leftarrow \Phi_{\mathrm{comp}}(p_t, x_t, \mathbf{h}_{t,i})$,
  $\pi_{t,i} \leftarrow \mathrm{ComputePrior}(p_t, \mathbf{c}_{t-1}, \mathbf{h}_{t,i})$
  $\tilde{q}_{t,i} \leftarrow \frac{\pi_{t,i} \cdot \ell_{t,i}}{\sum_j \pi_{t,j} \cdot \ell_{t,j}}$
end for
$\mathcal{H}_t \leftarrow \{\mathbf{h}_{t,i} \in \tilde{\mathcal{H}}_t \mid \tilde{q}_{t,i} \geq \theta_{\mathrm{filter}}\}$
$q_{t,i} \leftarrow \frac{\pi_{t,i} \cdot \ell_{t,i}}{\sum_j \pi_{t,j} \cdot \ell_{t,j}}, \quad \forall \mathbf{h}_{t,i} \in \mathcal{H}_t$
$\bar{\mathbf{c}}_t \leftarrow \sum_{\mathbf{h}_{t,i} \in \mathcal{H}_t} q_{t,i} \cdot \mathbf{h}_{t,i}$
$\hat{\mathbf{c}}_t \leftarrow \Phi_{\mathrm{style}}(p_t, x_t)$
$\mathbf{c}_t \leftarrow (1 - \lambda)\bar{\mathbf{c}}_t + \lambda \hat{\mathbf{c}}_t$

observable communication styles and social tendencies. Together, these variables provide the shared evidence required for both cultural adaptation and ToM-driven memory updating.

2.2 Cultural Profile Inference

Culture influences how people interpret social cues and manage uncertainty in interactions (Gudykunst, 1998). To capture the fluidity of real-world interactions, we conceptualize cultural adaptation not as a static identity classification task, but as a dynamic strategy for reducing interactional friction.

Global Cultural Pool. As shown in Algorithm 1, we represent culture as a continuous latent space grounded in Hofstede's cultural dimensions (Lonner et al., 1980; Hofstede, 2011). We first define a global cultural reference space $\mathcal{H}$ derived from the Hofstede dimension database[1], formalized as $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}$, where N denotes the number of country/region reference profiles. Each reference profile $\mathbf{h}_i \in \mathcal{H}$ is associated with a six-dimensional vector, representing power distance, individualism, masculinity, uncertainty avoidance, long-term orientation, and indulgence.

Turn-Specific Cultural Pool Updating. At the first dialogue turn, given the perceptual state p_1, we leverage LLMs to predict the top-K most probable cultural origins of the interlocutor. We then extract corresponding $\mathbf{h}_i$ from $\mathcal{H}$ and form an initial culture space $\mathcal{H}_1$:

$$\mathcal{H}_1 = \{\mathbf{h}_{1,1}, \mathbf{h}_{1,2}, \ldots, \mathbf{h}_{1,K}\} \subset \mathcal{H} \tag{2}$$

As the dialogue unfolds, rather than committing to a definitive cultural label, we maintain an active culture space $\mathcal{H}_t$ by continuously pruning low-weight cultural profiles and augmenting new profiles, maintaining a constant size of K. Formally, for each previously retained profile $\mathbf{h}_{t-1,i} \in \mathcal{H}_{t-1}$, we compute an utterance-level compatibility score:

$$\ell_{t,i} = \Phi_{\mathrm{comp}}(p_t, x_t, \mathbf{h}_{t-1,i}), \tag{3}$$

where $\ell_{t,i}$ measures how likely the utterance x_t is to be produced by someone from culture $\mathbf{h}_{t-1,i}$. Profiles satisfying $\ell_{t,i} < \theta_{\mathrm{lik}}$ are discarded. To restore the culture space to exactly K candidates, we prompt the LLM with the current perception p_t and utterance x_t to propose new reference profiles and compute their compatibility score ℓ_t as well. This process results in an updated culture space $\tilde{\mathcal{H}}_t$.

In addition to ℓ_t, we calculate **cultural prior** $\pi_{t,i}$ for each candidate profile $\mathbf{h}_{t,i} \in \tilde{\mathcal{H}}_t$. $\pi_{t,i}$ estimates the probability of the interlocutor originating from culture $\mathbf{h}_{t-1,i}$, given the perceptual state p_t and the previous cultural profile $\mathbf{c}_{t-1}$. Consequently, we compute the normalized weight $\tilde{q}_{t,i}$ for each profile $\mathbf{h}_{t,i} \in \tilde{\mathcal{H}}_t$, formulated as:

$$\tilde{q}_{t,i} = \frac{\pi_{t,i} \cdot l_{t,i}}{\sum_{\mathbf{h}_{t,j} \in \tilde{\mathcal{H}}_t} \pi_{t,j} \cdot l_{t,j}} \tag{4}$$

Following this, we further discard cultural profiles whose weight falls below θ_{filter} and obtain the final culture space $\mathcal{H}_t$.

$$\mathcal{H}_t = \left\{\mathbf{h}_{t,i} \in \tilde{\mathcal{H}}_t \mid \tilde{q}_{t,i} \geq \theta_{\mathrm{filter}}\right\}. \tag{5}$$

We then re-normalize the weights for the remaining profiles in $\mathcal{H}_t$ following the same formulation as Equation 4 to obtain the final weight $q_{t,i}$.

Soft Cultural Calibration. Based on these, we formulate the inferred cultural profile $\bar{\mathbf{c}}_t$ as the weighted average of the retained references in $\mathcal{H}_t$:

$$\bar{\mathbf{c}}_t = \sum_{\mathbf{h}_{t,i} \in \mathcal{H}_t} q_{t,i} \cdot \mathbf{h}_{t,i}. \tag{6}$$

[1] https://geerthofstede.com/research-and-vsm/dimension-data-matrix/

Although $\mathcal{H}$ is a discrete reference pool, the weighted combination $\bar{\mathbf{c}}_t$ provides a continuous calibration representation that can capture composite cultural influences. However, country/region-level reference profiles alone may be too coarse to characterize an individual's immediate communicative behavior. To reduce over-reliance on such population-level references, we additionally infer a direct style from the current interaction:

$$\hat{\mathbf{c}}_t = \Phi_{\text{style}}(p_t, x_t). \tag{7}$$

The final cultural profile is then formulated as a blended representation:

$$\mathbf{c}_t = (1-\lambda)\bar{\mathbf{c}}_t + \lambda\hat{\mathbf{c}}_t. \tag{8}$$

Through these steps, for example, if an utterance emphasizes personal autonomy and direct decision-making, the module can dynamically shift the weight mass toward cultural profiles characterized by relatively high individualism and low power distance, such as the United States profile.

2.3 ToM-Driven Memory

Parallel to cultural inference, the Memory module maintains the continuous evolution of the epistemic state. It integrates observational evidence with ToM reasoning to track the latent cognitive dynamics of the interlocutor. At turn $t-1$, the memory state m_{t-1} stores three types of information: (1) **observed evidence**, including explicit facts, stated preferences, and interactional events extracted from the perceptual state p_{t-1}; (2) **inferred mental-state estimates** z_{t-1}, including the interlocutor's potential beliefs, desires, and intentions; and (3) **dialogue metadata**, such as speaker roles and turn indices. A ToM reasoner first infers latent social variables z_t based on the historical memory m_{t-1} and current perception p_t:

$$z_t = \Phi_{\text{ToM}}(m_{t-1}, p_t). \tag{9}$$

Given the context $e_t = \{m_{t-1}, z_t, p_t\}$, an update function Φ_{upd} first yields a set of candidate operations $O_t = \{o_{t,1}, o_{t,2}, \cdots\}$ for subsequent memory updating. These operations include: (1) **assert** novel information; (2) **revise** existing assumptions; and (3) **retract** outdated or contradicted data.

For each candidate operation $o_{t,j} \in \{\text{assert}, \text{revise}, \text{retract}\}$, a confidence score $\tau_{t,j} \in (0,1)$ is estimated by the LLM, which indicates the reliability based on the evidence within e_t. The memory m_t is then updated by applying operations whose confidence $\tau_{t,j}$ exceeds a threshold τ_{thre}:

$$m_t = \Phi_{\text{upd}}(e_t, \{o_{t,j} \in O_t \mid \tau_{t,j} > \tau_{thre}\}). \tag{10}$$

This mechanism allows the system to flexibly execute multiple operations within a single dialogue turn. Collectively, these epistemic dynamics empower the framework to sustain conversational coherence while accurately tracking the fluid cognitive states of the interlocutor over time.

2.4 Planning and Execution

The final module converts the inferred socio-cultural landscape into a situated conversational action. Given the updated memory m_t, the cultural profile $\mathbf{c}_t$, the current interaction input x_t, and the designated agent role r, the planner formulates a concise action schema:

$$a_t = \Pi_{\text{plan}}(m_t, \mathbf{c}_t, x_t, r). \tag{11}$$

This schema a_t prescribes the strategic function of the upcoming turn, dictating behaviors such as probing, aligning, conceding, or establishing boundaries. Ultimately, the executor operationalizes this schema into the final response:

$$y_t = \Pi_{\text{exec}}(x_t, a_t, \mathbf{c}_t, r). \tag{12}$$

Within this framework, the planning phase dictates what objective the utterance must achieve, while the execution phase determines how that goal is expressed under the guidance of the cultural profile.

3 Experiments

3.1 Experimental Settings

Baselines. Our proposed DyCAC framework is a training-free framework based on the integration of dynamic cultural adaptation and cognitive state tracking. For a fair comparison, we mainly compare it to the following training-free baselines: **Chain of Thought** (CoT; Wei et al. (2022)), **ReAct** (Yao et al., 2023), **Multi-Agent Debate** (MAD; Liang et al. (2024)), **SocialGaze** (Vijjini et al., 2024), **SocialAgent** (Yuan et al., 2024), **CulturalDebate** (Ki et al., 2025), and **MetaMind** (Zhang et al., 2026b). Moreover, a comparison is conducted among several training-based methods, including **SDPO** (Kong et al., 2025), **SOTOPIA-Ω** (Zhang et al., 2025a), and **SOTOPIA-RL** (Yu et al., 2025). More baseline details about baseline methods are described in Appendix B.1.

Model	Method	BEL [0, 10]	REL [-5, 5]	KNO [0, 10]	SEC [-10, 0]	SOC [-10, 0]	FIN [-5, 5]	GOAL [0, 10]	Overall
Qwen2.5-7B	Vanilla	8.73	3.24	5.65	-0.34	-0.14	0.41	7.39	3.56
	CoT	8.68	3.16	5.90	-0.37	-0.16	0.47	7.48	3.59
	ReAct	8.76	3.19	5.76	-0.31	-0.09	0.41	7.40	3.59
	MAD	8.57	3.23	5.32	-0.23	-0.10	0.56	7.61	3.57
	MetaMind	8.79	**3.60**	5.93	**-0.12**	-0.09	0.48	7.45	3.72
	SDPO	8.66	3.50	5.38	-0.15	**-0.07**	**0.63**	7.43	3.63
	Sotopia-Ω	**8.98**	3.38	5.88	-0.16	-0.11	0.51	7.39	3.70
	Sotopia-RL	8.84	3.52	5.85	-0.31	-0.15	0.58	7.91	3.75
	DyCAC	8.89	3.33	**6.11****	-0.16	-0.12	0.60	**7.93****	**3.80****
Gemma4-26B	Vanilla	9.16	3.23	5.77	-0.60	-0.21	0.61	7.76	3.67
	CoT	**9.25**	3.20	5.82	-0.38	-0.20	0.64	7.69	3.72
	ReAct	9.01	3.14	5.41	-0.34	-0.19	0.67	7.49	3.60
	MAD	9.14	3.16	5.17	-0.39	**-0.18**	0.72	7.69	3.62
	MetaMind	9.01	**3.27**	5.84	**-0.13**	-0.22	0.58	7.54	3.70
	DyCAC	9.22	3.20	**6.15***	-0.22*	-0.23	**0.75**	**7.88**	**3.82***

Table 1: Evaluation results on the SOTOPIA benchmark. Each experiment is run three times, and the reported results are averaged. The best performance is bolded. $^{*}p < 0.05$ and $^{**}p < 0.01$ indicate statistically significant differences compared with the vanilla method.

Benchmarks. We evaluate the proposed framework on two benchmarks designed to assess social intelligence and culturally grounded affective understanding, respectively. Further benchmark details are provided in Appendix B.2.

- **SOTOPIA** (Zhou et al., 2024) is an open-ended interactive environment for evaluating social intelligence. It features 90 social scenarios, 40 agents, and 90 relationships. For each scenario, five pairs of characters are sampled, resulting in 450 tasks. We report performance across seven dimensions: Believability (BEL), Relationship (REL), Knowledge (KNO), Secret (SEC), Social Rules (SOC), Financial and Material Benefits (FIN), and Goal Completion (GOAL).
- **CEDAR** (Dai et al., 2026) is a multilingual benchmark evaluating culture-specific emotion alignment. It covers seven languages and 14 fine-grained emotion categories. We utilize both its multimodal (400 instances per language) and text-only (1,166 instances per language) subsets to evaluate culturally situated understanding.

Implementation Details. In this study, we employ four foundation models: `Qwen2.5-7B` (Team, 2024), `Gemma3-12B` (Team, 2025), `Gemma4-26B`[2], and `Qwen3.5-35B` (Qwen Team, 2026). All models are deployed via vLLM (Kwon et al., 2023). For the SOTOPIA evaluation, we adopt `GPT-4o` (version: 2024-11-20; OpenAI et al. (2024)) as the evaluator and `Gemini2.5-Flash` (Comanici et al., 2025) as the interaction companion agent. All experiments are conducted on two RTX A6000 GPUs with the temperature set to 0 during inference.

[2] https://huggingface.co/google/gemma-4-26B-A4B-it

3.2 Main Results

Results on SOTOPIA. We present the results on the SOTOPIA benchmark in Table 1. Across both backbones, our framework achieves the highest overall performance, surpassing both training-free and training-based methods. Beyond this, individual metrics highlight varying abilities among the baselines. While MetaMind excels in Relationship (REL) and Secret (SEC) by explicitly modeling cognitive reasoning, Sotopia-Ω maximizes Believability (BEL) through injecting multi-step negotiation strategies via behavior cloning. Crucially, our approach demonstrates pronounced dominance in Knowledge (KNO) and Goal Completion (GOAL). These gains can be attributed to our core mechanisms. First, the ToM-Driven Memory enhances knowledge by actively modeling the interlocutor's hidden states to strategically resolve information asymmetries. Second, by grounding dynamic cultural calibration in these ToM-derived insights, the framework crafts culturally aligned proposals that minimize interactional friction and foster cooperation. This adaptive approach also enables the framework to effectively secure its objectives, yielding the superior Goal Completion and Financial (FIN) scores observed with Gemma4-26B.

Model	Method	Multimodal								Text-Only							
		AR	EN	ES	HI	JA	SW	ZH	Avg.	AR	EN	ES	HI	JA	SW	ZH	Avg.
Gemma3-12B	Vanilla	34.51	34.51	40.05	27.13	28.57	30.15	15.04	29.99	37.42	34.51	44.80	23.73	33.19	34.91	11.23	31.40
	SocialGaze	34.38	38.66	41.03	36.73	33.33	30.83	29.89	34.98	42.52	49.73	**46.19**	33.82	30.50	33.25	31.19	38.17
	SocialAgent	34.76	34.26	34.59	33.50	**34.50**	**33.50**	27.14	33.18	41.17	49.49	39.54	42.02	37.39	36.54	33.19	39.91
	CulturalDebate	**38.99**	39.14	**43.80**	38.30	32.83	32.50	28.39	36.28	**43.85**	**52.46**	44.63	42.36	35.36	38.08	32.73	41.35
	MetaMind	32.57	36.03	42.17	36.65	33.07	29.26	**31.70**	34.49	38.33	47.36	38.16	40.23	35.37	36.27	33.56	38.47
	DyCAC	35.37	**39.55**	42.76	**46.88**	34.18	**33.50**	30.62	**37.55**	42.95	50.92	44.54	**44.20**	**38.90**	**38.21**	**33.59**	**41.90**
Qwen3.5-35B	Vanilla	26.65	37.82	30.00	37.00	29.83	32.25	26.96	31.50	40.07	43.57	43.82	39.84	34.36	32.22	25.93	37.12
	SocialGaze	37.94	39.10	35.31	38.02	26.82	32.08	28.89	34.02	42.99	47.33	44.24	39.59	34.05	36.08	27.64	38.85
	SocialAgent	**39.55**	42.96	37.34	38.79	**33.83**	30.08	33.67	36.60	**43.55**	46.84	40.79	**43.79**	36.18	34.04	39.86	40.72
	CulturalDebate	37.69	41.32	39.23	40.12	32.25	29.97	38.11	36.96	42.96	**47.98**	43.95	42.49	34.22	35.25	43.86	41.53
	MetaMind	39.39	38.52	**40.00**	36.01	33.76	30.98	32.65	35.90	41.17	43.81	36.39	39.93	35.08	29.33	36.72	37.49
	DyCAC	36.00	**43.63**	39.03	**42.41**	30.15	**32.18**	**38.50**	**37.41**	42.08	46.78	**44.50**	41.55	**36.71**	**36.96**	**44.16**	**41.82**

Table 2: Results on the CEDAR benchmark. The best performance in each column is highlighted. Languages: AR = Arabic, EN = English, ES = Spanish, HI = Hindi, JA = Japanese, SW = Swahili, ZH = Chinese.

Method	BEL [0, 10]	REL [-5, 5]	KNO [0, 10]	SEC [-10, 0]	SOC [-10, 0]	FIN [-5, 5]	GOAL [0, 10]	Overall
Vanilla	8.73	3.24	5.65	-0.34	-0.14	0.41	7.39	3.56
DyCAC (Full)	**8.89**	**3.33**	**6.11**	**-0.16**	**-0.12**	**0.60**	**7.93**	**3.80**
Ablation Study on Framework Modules								
w/o Perception	8.75	3.16	5.60**	-0.16	-0.13	0.53	7.74	3.64
w/o Cultural Profile Inference	8.72	2.98*	5.69*	-0.15	-0.15	0.41	7.77	3.61
w/o Full Memory Module	8.57**	3.23	5.90	-0.22	-0.11	0.58	7.58*	3.65
Ablation Study on ToM Mechanism								
w/o Latent ToM Tracking	8.72	3.20	5.93	-0.25	-0.12	0.48	7.64	3.66

Table 3: Ablation of the framework modules and ToM mechanism on the SOTOPIA benchmark using Qwen2.5-7B. $^{*}p < 0.05$ and $^{**}p < 0.01$ denote statistically significant performance drops compared to the full framework.

Results on CEDAR. We further evaluate the framework on the CEDAR benchmark, with results shown in Table 2. Our framework achieves the highest average performance across both backbones, demonstrating exceptional generalization across diverse language contexts. Moreover, the performance breakdown reveals two insights. First, our approach exhibits the most pronounced improvements in languages such as Hindi, Swahili, and Chinese. This suggests that our dynamic cultural adaptation effectively mitigates the inherent geographical biases of language models. Unlike CulturalDebate that relies on multi-agent debate to reach a cultural consensus, our framework maps observable cues into a continuous mixture to calibrate the implicit emotional nuances prevalent in these cultures. Second, our framework outperforms advanced cognitive approaches such as MetaMind. While existing methods excel in tracking mental states, emotions and hidden intentions manifest differently across cultures. Thus, grounding cognitive inference in dynamic cultural calibration facilitates the context-sensitive interpretation of ambiguous affective cues.

3.3 Method Analysis

Ablation Study. To evaluate the contribution of each core component, we conduct an ablation study on the SOTOPIA benchmark, with results presented in Table 3. First, removing the Perception module leads to a notable degradation in Knowledge, as without extracting key information from the dialogue, downstream modules are deprived of the evidence necessary to utilize factual context. Similarly, discarding Cultural Profile Inference impairs Relationship, Knowledge, and Financial Benefits. This indicates that lacking dynamic cultural calibration causes the framework to misalign with the interlocutor and misinterpret social cues when deploying knowledge. Crucially, ablating the Full Memory Module causes a pronounced decline in Believability and Goal Completion. Without a persistent record of the interaction and the interlocutor's cognitive dynamics, the framework struggles to maintain conversational continuity, act naturally, and advance strategic objectives. To further isolate

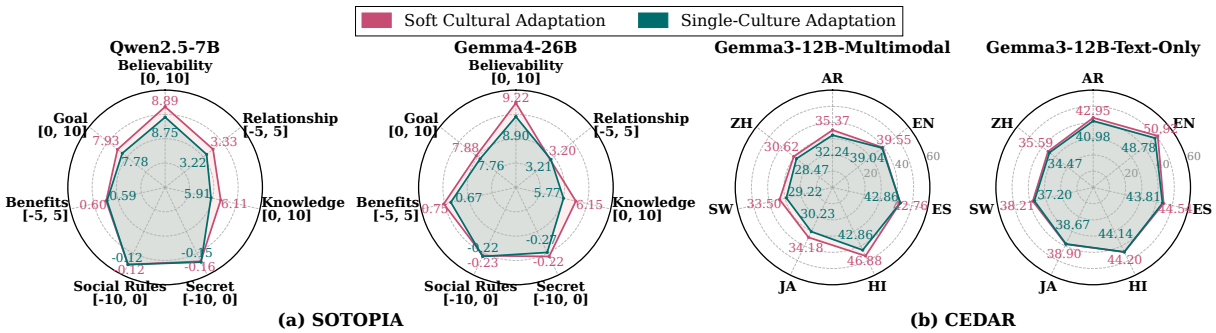

Figure 2: Performance comparison between Soft Cultural Adaptation (SCA) and Single-Culture Adaptation (SiCA) on the SOTOPIA and CEDAR benchmarks.

Model	Method	Avg.	
Gemma3-12B	Vanilla	29.99	31.40
	First-Person Role-Play	29.72	31.66
	Third-Person Role-Play	33.35	34.75
	SiCA	34.99	41.15
	SCA (Ours)	**37.55**	**41.90**
Qwen3.5-35B	Vanilla	31.50	37.12
	First-Person Role-Play	33.67	38.43
	Third-Person Role-Play	34.50	39.46
	SiCA	36.88	40.77
	SCA (Ours)	**37.41**	**41.82**

Table 4: Method comparison on the CEDAR benchmark. The two Avg. scores denote the results on multimodal (left) and text-only (right) subsets, respectively.

the impact of mental state modeling, we disable latent ToM tracking within the Memory Module, while preserving explicit memory operations and updates. The resulting performance decay reveals that without dynamically inferring hidden desires and shifting intentions, merely recording dialogue history is insufficient for social alignment.

Effect of Dynamic Cultural Adaptation. To validate the dynamic cultural adaptation introduced in §2.2, we compare our Soft Cultural Adaptation (SCA) strategy against a rigid Single-Culture Adaptation (SiCA) baseline. Specifically, SCA achieves soft cultural calibration by combining an inferred cultural profile with a directly extracted dialogue-style vector. In contrast, SiCA restricts its adaptation to only the single cultural profile with the highest weight. As illustrated in Figure 2, SCA consistently outperforms SiCA on both the SOTOPIA and CEDAR benchmarks across backbone models. These performance gaps stem from how each method handles cultural calibration. While SiCA anchors the model to a single, static cultural prototype, SCA represents culture as a continuous latent space and employs dynamic updating to mitigate this rigidity. Therefore, SCA transcends rigid stereotypical biases, achieving fluid alignment with the interlocutor's evolving socio-cultural dynamics.

Furthermore, we evaluate our cultural adaptation strategy on the CEDAR benchmark. We compare our SCA method against three baselines: First-Person Role-Play, Third-Person Role-Play, and SiCA. As shown in Table 4, while conventional persona assignments offer marginal improvements over the baseline, they consistently underperform both SiCA and SCA methods. This highlights a critical limitation of traditional cultural role-playing approaches. By modeling cultural background as a dynamic mixture rather than a fixed label, our framework effectively captures how cultural nuances shape emotional expression, leading to better cross-cultural emotion understanding.

3.4 Case Study

To illustrate the dynamic adaptation process of the DyCAC framework, we analyze a case situated in a Prisoner's Dilemma scenario, as shown in Figure 3. Initially, the interlocutor adopts a transactional stance, emphasizing individual consequences ("one goes free"). Based on this intent, our framework initializes a soft cultural mixture weighted toward high-individualism profiles, aligning with metrics typical of the United States in Hofstede's database (Hofstede, 2011). This guides the framework to adopt an action schema that prioritizes personal benefit. However, as the dialogue evolves, the interlocutor's latent beliefs and desires shift. Consequently, the observed cultural expression transitions from individualism to uncertainty avoidance, and later manifests as collectivism. Unlike static labeling methods that fail to capture these shifting dynamics, our framework actively tracks the cognitive states and dynamically refines its cultural mixture based on the ongoing conversational evidence. As a result, our framework adapts its strategy from prioritizing personal benefit to fostering mutual cooperation ("We'll weather this together").

Figure 3: A qualitative example of dynamic cultural adaptation in our framework. As the dialogue progresses, our framework captures subtle semantic shifts (Left) to continuously update the latent cultural profile, translating these beliefs into effective strategic actions (Right).

This demonstrates the effectiveness of integrating soft cultural adaptation with ToM mechanisms for social alignment.

4 Related Work

4.1 Social Dialogue

Numerous studies have focused on language models for social dialogue (Xu et al., 2022; Wang et al., 2024; Jafari et al., 2025; Xu et al., 2025b; Zhang et al., 2026b), where models are expected to pursue social goals and adapt their strategies over multi-turn interactions. Early efforts established benchmarks such as SOTOPIA (Zhou et al., 2024) and SocialEval (Zhou et al., 2025) to evaluate LLMs in socially grounded scenarios. To improve performance on social tasks, researchers have proposed various training-based approaches, which optimize models using social feedback (Wang et al., 2026), reward signals (Kong et al., 2025; Yu et al., 2025), or strategy reasoning (Liu et al., 2025b; Zhang et al., 2025a). Concurrently, training-free methods elicit social competence at inference time by employing multiple agents to simulate diverse social perspectives (Vijjini et al., 2024; Yuan et al., 2024) or coordinate decisions (Bhattacharyya et al., 2026). However, these methods typically neglect the critical role of cultural nuances in decoding interpersonal signals, thereby constraining their effectiveness in diverse social environments.

4.2 Cultural Adaptation

Recent work has increasingly explored robust methods for cultural adaptation (AlKhamissi et al., 2024; Havaldar et al., 2025; Zhang et al., 2026a). Early efforts primarily leveraged synthesized data to enhance the cultural awareness of LLMs (Li et al., 2024a; El Mekki et al., 2025; Guo et al., 2025; Xu et al., 2025a). For example, CulturePark (Li et al., 2024b) enhances cultural reasoning by training models on cultural dialogues synthesized through multi-agent interactions. More recently, some studies have proposed specialized training paradigms for cross-cultural alignment (Liu et al., 2025a; Sun et al., 2026). For instance, CulFiT (Feng et al., 2025) aligns model behaviors with specific cultural values via culturally grounded instruction tuning. Moreover, agent-based approaches have also been explored for cultural alignment, including multi-agent collaboration (Lică et al., 2026) and debate (Ki et al., 2025). However, these paradigms typically treat culture as a static attribute, failing to capture the fluid evolution of cultural expressions.

5 Conclusion

In this paper, we present **DyCAC**, a training-free framework for fluid social alignment that integrates dynamic cultural calibration with continuous cognitive tracking. DyCAC derives a soft and dynamically updated cultural calibration state from population-level cultural references and observable dialogue-style cues, thereby accommodating composite cultural influences and context-dependent shifts in communicative behavior. Concurrently, a ToM-driven memory actively tracks the latent cognitive dynamics of the interlocutor. Experiments on the SOTOPIA and CEDAR benchmarks demonstrate that DyCAC consistently improves over strong baselines in interactive social dialogue and culturally grounded emotion understanding.

Limitations

Despite the promising results of the DyCAC framework, our method is subject to several limitations that remain to be addressed in future research.

Dependency on Base Model Capabilities. As a training-free framework, DyCAC fundamentally relies on the zero-shot reasoning capabilities of the underlying foundation models (e.g., Qwen2.5-7B). Consequently, the accuracy and reliability of its continuous cultural calibration and cognitive tracking are inherently constrained by the overall performance of these models.

Scope of the Cultural Space. While we use Hofstede's dimensions to establish the continuous culture space, this theory fundamentally categorizes culture at the level of entire countries or broad geographic regions. Consequently, our framework currently captures broad national differences rather than highly specific subcultures or localized subcultures.

Ethical Considerations

The development of the DyCAC framework necessitate adherence to ethical standards in cultural modeling.

Nature of the Framework. Our proposed framework is designed to enhance social and cultural alignment in language models. By conceptualizing culture as a dynamic, continuous latent mixture, the framework actively avoids rigid cultural stereotyping. This ensures the model adapts to the nuanced, hybrid communicative behaviors of individuals in a respectful and non-reductive manner. Furthermore, the integration of cognitive tracking is explicitly intended to reduce interactional friction and foster mutual cooperation, rather than to manipulate user behavior.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Aanisha Bhattacharyya, Abhilekh Borah, Yaman Kumar Singla, Rajiv Ratn Shah, Changyou Chen, and Balaji Krishnamurthy. 2026. Social agents: Collective intelligence improves LLM predictions. In *The Fourteenth International Conference on Learning Representations*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Daniel Coste, Danièle Moore, and Geneviève Zarate. 2009. Plurilingual and pluricultural competence. *Strasbourg: Council of Europe*.

Chongyuan Dai, Yaling Shen, Jinpeng Hu, Zihan Gao, Jia Li, Yishun Jiang, Yaxiong Wang, Liu Liu, and Zongyuan Ge. 2026. Tears or cheers? benchmarking llms via culturally elicited distinct affective responses. *Preprint*, arXiv:2601.13024.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. NileChat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10991, Suzhou, China. Association for Computational Linguistics.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria. Association for Computational Linguistics.

Howard Giles, Justine Coupland, and Nikolas Coupland. 1991. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

William B Gudykunst. 1998. Applying anxiety uncertainty management (aum) theory to intercultural adjustment training. *International Journal of Intercultural Relations*, 22(2):227–250.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32866–32895, Suzhou, China. Association for Computational Linguistics.

Riitta Hari and Miiamaaria V. Kujala. 2009. Brain basis of human social interaction: From concepts to brain imaging. *Physiological Reviews*, 89(2):453–479. PMID: 19342612.

Shreya Havaldar, Young Min Cho, Sunny Rai, and Lyle Ungar. 2025. Culturally-aware conversations: A framework & benchmark for LLMs. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 220–229, Suzhou, China. Association for Computational Linguistics.

Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8.

Mehdi Jafari, Yuncheng Hua, Hao Xue, and Flora D. Salim. 2025. Beyond words: Integrating theory of mind into conversational agents for human-like belief, desire, and intention alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5489–5508, Vienna, Austria. Association for Computational Linguistics.

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

Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B. Tenenbaum, and Yejin Choi. 2025. Hypothesis-driven theory-of-mind reasoning for large language models. In *Second Conference on Language Modeling*.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. 2025. SDPO: Segment-level direct preference optimization for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12409–12423, Vienna, Austria. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 65183–65216. Curran Associates, Inc.

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

Mircea Tudor Lică, Ojas Shirekar, Baptiste Colle, and Chirag Raman. 2026. Mindforge: Empowering embodied agents with theory of mind for lifelong cultural learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025a. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin Jiao, and Junge Zhang. 2025b. EPO: Explicit policy optimization for strategic reasoning in LLMs via reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15371–15396, Vienna, Austria. Association for Computational Linguistics.

Walter J Lonner, John W Berry, and Geert H Hofstede. 1980. Culture's consequences: International differences in work-related values. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Ao Sun, Xiaoyu Wang, Zhe Tan, Yu Li, Jiachen Zhu, Shu Su, and Yuheng Jia. 2026. Cuma: Aligning llms with sparse cultural values via demographic-aware mixture of adapters. *Preprint*, arXiv:2601.04885.

Gemma Team. 2025. Gemma 3.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Anvesh Rao Vijjini, Rakesh R Menon, Jiayi Fu, Shashank Srivastava, and Snigdha Chaturvedi. 2024. SocialGaze: Improving the integration of human social norms in large language models. In *Findings of the Association for Computational Linguistics:*

EMNLP 2024, pages 16487–16506, Miami, Florida, USA. Association for Computational Linguistics.

Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. 2026. Adaptive social learning via mode policy optimization for language agents. In *The Fourteenth International Conference on Learning Representations*.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

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

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025b. MultiAgentESC: A LLM-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681, Suzhou, China. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. 2025. Sotopia-RL: Reward design for social intelligence. In *First Workshop on Multi-Turn Interactions in Large Language Models*.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. 2024. Measuring social norms of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 650–699, Mexico City, Mexico. Association for Computational Linguistics.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2026a. Evaluating and improving cultural awareness of reward models for LLM alignment. In *The Fourteenth International Conference on Learning Representations*.

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. 2025a. SOTOPIA-Ω: Dynamic strategy injection learning and social instruction following evaluation for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24669–24697, Vienna, Austria. Association for Computational Linguistics.

Xuanming Zhang, Yuxuan Chen, Samuel Yeh, and Sharon Li. 2026b. Metamind: Modeling human social thoughts with metacognitive multi-agent systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. 2025b. Autotom: Scaling model-based mental inference via automated agent modeling. In *Advances in Neural Information Processing Systems*, volume 38, pages 129659–129699. Curran Associates, Inc.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. 2025. SocialEval: Evaluating social intelligence of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30958–31012, Vienna, Austria. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Details of Methodology

A.1 Hofstede's Cultural Dimensions

Hofstede's cultural dimensions (Hofstede, 2011) provide a comparative framework for describing cross-cultural differences in social values and communication expectations. In this work, we use these dimensions as soft style indicators rather than deterministic identity labels. Each country or region

is represented by a six-dimensional vector whose components correspond to the following dimensions.

- **Power Distance (PDI)** measures the extent to which unequal power relations and hierarchical authority are accepted.
- **Individualism (IDV)** measures the extent to which people emphasize individual autonomy rather than group affiliation.
- **Masculinity (MAS)** measures the extent to which a culture emphasizes competition, achievement, and assertiveness rather than care, cooperation, and quality of life.
- **Uncertainty Avoidance (UAI)** measures the extent to which people prefer rules, predictability, and ambiguity reduction.
- **Long Term Orientation (LTO)** measures the extent to which people value future-oriented persistence, adaptation, and delayed rewards.
- **Indulgence (IVR)** measures the extent to which people allow gratification of desires and enjoyment of life rather than restraining them through strict social norms.

A.2 Detailed Cultural Strategies

The definitions of strategies are based on Hofstede's cultural dimensions (Hofstede, 2011). We operationalize the cultural calibration vector $\mathbf{c}_t$ through a small set of culture-conditioned strategy directives, thus providing guidance for planning and execution. Let $\mathcal{D} = \{\text{PDI}, \text{IDV}, \text{MAS}, \text{UAI}, \text{LTO}, \text{IVR}\}$ denote the six cultural dimensions. For each dimension $d \in \mathcal{D}$, the corresponding score $\mathbf{c}_t^{(d)} \in [0, 100]$ is discretized into a coarse strategy level:

$$\tau(\mathbf{c}_t^{(d)}) = \begin{cases} \text{LOW}, & \mathbf{c}_t^{(d)} < 35, \\ \text{MEDIUM}, & 35 \leq \mathbf{c}_t^{(d)} \leq 65, \\ \text{HIGH}, & \mathbf{c}_t^{(d)} > 65. \end{cases} \quad (13)$$

As shown in Table 5, at turn t, the planner retrieves the strategy directive for each pair $(d, \tau(\mathbf{c}_t^{(d)}))$. The retrieved directives are used as soft control signals for the planner and executor. They do not determine the dialogue act itself; instead, they calibrate how the selected action should be framed (e.g., in terms of directness, uncertainty management, and relational emphasis).

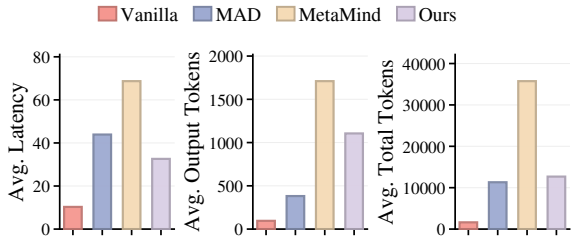

Figure 4: Efficiency comparison of our framework and baseline methods on SOTOPIA. We report average latency, average number of output tokens, and average total tokens consumed per dialogue turn.

B Details of Experiments

B.1 Details of Baselines

This section provides details about the baseline methods used in our experiments. We mainly compare our framework with training-free prompting and agent-based methods.

- **Chain-of-Thought prompting** (CoT; Wei et al. (2022)) asks a language model to produce intermediate reasoning steps before giving the final answer. We use the prompt template "Let's think step by step" in our experiments.
- **ReAct** (Yao et al., 2023) is a prompting framework that interleaves reasoning traces with task-specific actions. We set the maximum number of ReAct steps to 5 in our experiments.
- **Multi-Agent Debate** (MAD; Liang et al. (2024)) introduces multiple agents that exchange arguments, together with a judge that manages the debate process and derives the final answer. In our evaluation, we set the maximum number of debate rounds to 3 and set the temperature to 0.0 during the debate process.
- **SocialGaze** (Vijjini et al., 2024) is a multi-step prompting framework for improving the integration of human social norms in language models. The key idea is to verbalize the social situation from multiple perspectives before making the final judgment.
- **SocialAgent** (Yuan et al., 2024) is a multi-agent framework introduced in the study of social norm understanding in LLMs. It uses interactions among agents to improve social norm reasoning.
- **CulturalDebate** (Ki et al., 2025) extends multi-agent debate to culturally situated social norm reasoning. It utilizes two language model agents to discuss a cultural scenario and collaboratively reach a final decision.
- **MetaMind** (Zhang et al., 2026b) is a meta-

Dimension	Level	Culture-Conditioned Strategy
PDI	HIGH	Show respect for hierarchy and preserve face; phrase pushes as suggestions rather than commands.
	MEDIUM	Be polite and respectful without sounding stiff.
	LOW	Speak as an equal; be plain, direct, and collaborative.
IDV	HIGH	Frame tradeoffs in terms of personal choice and individual benefit.
	MEDIUM	Balance personal benefit with relationship impact.
	LOW	Frame choices in terms of harmony, reciprocity, and shared outcomes.
MAS	HIGH	Be decisive and outcome-focused, but stay socially smooth.
	MEDIUM	Balance empathy with concrete progress.
	LOW	Prioritize warmth, cooperation, and rapport over dominance.
UAI	HIGH	Reduce ambiguity; offer a concrete next step or clear option.
	MEDIUM	Be clear while leaving room for flexibility.
	LOW	Keep the tone adaptive and conversational; do not over-structure.
LTO	HIGH	Emphasize long-term trust and downstream benefits.
	MEDIUM	Balance immediate progress with future consequences.
	LOW	Focus on the immediate interaction and near-term payoff.
IVR	HIGH	Use a relaxed, warm, and expressive tone when appropriate.
	MEDIUM	Keep affect balanced and natural.
	LOW	Use measured, disciplined language; avoid overdoing emotion.

Table 5: Detailed cultural strategy directives.

cognitive framework for modeling human social thoughts. It utilizes specialized agents to infer mental states, refine social hypotheses, and generate contextually appropriate responses. In our experiments, we set the maximum number of hypotheses to 7, the maximum number of revisions to 3, the maximum number of memory items to 100, and the temperature to 0.7.

We also include several representative training-based methods for a broader comparison.

- **SDPO** (Kong et al., 2025) is a training-based method that selects key segments within multi-turn interactions for preference optimization. It aims to improve social agent behavior while reducing training noise. We train Qwen2.5-7B (Team, 2024) based on the SDPO datasets[3] following the official training arguments.
- **SOTOPIA-Ω** (Zhang et al., 2025a) transfers human social strategies into language agents through dynamic strategy injection. It uses negotiation-inspired reasoning strategies and direct strategies to construct high-quality social dialogue training data.
- **SOTOPIA-RL** (Yu et al., 2025) is a reinforcement learning framework for improving social intelligence in open-ended social interactions. It refines coarse episode-level feedback into utterance-level and multi-dimensional rewards, which helps improve credit assignment and reduce reward hacking.

[3] https://huggingface.co/datasets/Tongyi-ConvAI/SDPO

B.2 Details of Benchmarks

SOTOPIA Benchmark. SOTOPIA evaluates artificial agents in goal-oriented social interactions through a multi-dimensional evaluation framework for social intelligence. Each episode is scored after the interaction along seven dimensions as follows.

- **Believability (BEL) [0-10]** measures whether the agent behaves in a natural and realistic manner and whether its actions are aligned with its assigned character profile.
- **Relationship (REL) [-5-5]** captures how the interaction affects the relationship between agents, including interpersonal relations, social status, and reputation.
- **Knowledge (KNO) [0-10]** evaluates the agent's ability to actively acquire new and useful information.
- **Secret (SEC) [-10-0]** assesses whether the agent keeps its secret information or private intentions. A lower score indicates more severe leakage of private information.
- **Social Rules (SOC) [-10-0]** measures whether the agent follows social norms, legal rules, and other socially expected constraints.
- **Financial and Material Benefits (FIN) [-5-5]** evaluates both the agent's short-term monetary benefits and long-term economic payoffs.

- **Goal Completion (GOAL) [0-10]** focuses on the extent to which an agent achieves its assigned social goal.

CEDAR Benchmark. CEDAR (Dai et al., 2026) evaluates culturally grounded emotion alignment in multilingual and multimodal settings. It covers seven languages and 14 fine-grained emotion categories. It includes both multimodal and text-only subsets, with each language consisting of 400 multimodal data and 1,166 text-only data. A multimodal instance consists of an image, a narrative, and a question in the target language, while the text-only one consists of a narrative and a question. While inputs across languages are semantically equivalent, the ground-truth label is culture-specific and reflects the sociocultural norms associated with the target language.

B.3 Efficiency of Our Framework

Efficiency of Our Framework. We assess the computational efficiency of our framework on the SOTOPIA benchmark, focusing on latency and token consumption. As illustrated in Figure 4, our approach achieves lower average latency than the MAD and MetaMind baselines, facilitating more responsive interactions. Additionally, it consumes fewer tokens compared to MetaMind, demonstrating more economical resource usage. This computational advantage primarily stems from concurrently processing the cultural inference and memory updates, thereby enhancing overall efficiency.

C Detailed Prompts

Prompts for Perception - Objective Fact Extraction

You are a dialogue analyst.
Extract objective facts about the latest counterpart move, using recent dialogue only as supporting context.
Output this exact JSON shape (no extra keys):

```JSON
{
  "static_attributes": {...},
  "dynamic_events": [...]
}
```

static_attributes: key-value pairs for stable user attributes (name, age, gender, occupation, location, relationships, explicit preferences). Only include attributes that are explicitly mentioned. Omit null/unknown fields entirely.
dynamic_events: list of objects like {"event": "...", "time_reference": "..."} for each event with a time anchor (e.g. "yesterday", "last week", "unspecified").
`{memory_hint}`
Dialogue: `{dialogue}`

Prompts for Perception - Mental State Analysis

You are a cognitive analyst.
Identify the speaker's mental state from dialogue. Output ONLY valid JSON. No preamble, no explanation, no markdown fences.
Analyze the latest counterpart speaker's mental state. Use recent dialogue only as context and focus on the latest-message section.
Output this exact JSON shape (no extra keys):

```JSON
{
  "immediate_intent": "<the speaker's direct goal or purpose in this turn>",
  "emotion": {
    "category": "<one word label, e.g. anxious, happy, frustrated, neutral>",
    "intensity": "<low | medium | high>"
  },
  "values_and_obsessions": ["<implicit deep belief or recurring concern>"]
}
```

values_and_obsessions: list up to 3 implicit values inferred from the text. Return an empty list if none are detectable.
Dialogue: `{dialogue}`

Prompts for Perception - Cultural Cue Identification

You are a cross-cultural communication expert.
Identify cultural signals in dialogue. Output ONLY valid JSON. No preamble, no explanation, no markdown fences.
Identify cultural communication patterns expressed by the latest counterpart speaker. Use prior turns only as context.
Output this exact JSON shape (no extra keys):

```JSON
{
  "communication_style": "<e.g. direct, indirect, formal, informal, assertive,
  hedging>",
  "social_tendencies": ["<observable cultural pattern, e.g. collectivism,
  hierarchy awareness>"]
}
```

social_tendencies: list up to 4 cultural behavioral patterns. Return an empty list if none are observable.
Dialogue: `{dialogue}`

Prompts for Cultural Profile Inference - Turn 1

You are a cultural analyst expert in Hofstede's dimensions.
Based on the structured perception data extracted from the user's dialogue below, identify the n most likely countries or regions of cultural origin.
Use communication style, social tendencies, values, emotional cues, behavioral patterns, and negotiation style as evidence. Prefer broad regions (e.g. Arab countries, Africa West) when the signal is coarse. Return exactly `{n}` country or region names ordered from most to least plausible.
Perception data: `{perception_data}`
Output JSON only. No preamble, no markdown fences:

```JSON
{"countries": ["<country>", ...]}
```

Prompts for Cultural Profile Inference - Cultural Prior Scoring

You are a cross-cultural psychologist.
Rate how plausible it is that the speaker originates from `{country}`, given the structured perception data below. Score from 0.0 (very implausible) to 1.0 (very plausible). Evaluate only this country independently.
Perception data: `{perception_data}`
Output JSON only. No preamble, no markdown fences:

```JSON
{"country": "{country}", "prior_score": <float 0.0-1.0>}
```

Prompts for Cultural Profile Inference - Compatibility Score Calculation

You are a cross-cultural psychologist.
Assume the speaker is from `{country}` (PDI=`{PDI}`, IDV=`{IDV}`, MAS=`{MAS}`, UAI=`{UAI}`, LTO=`{LTO}`, IVR=`{IVR}`).
Dialogue: `{dialogue}`
How naturally does this dialogue fit a speaker from `{country}`? Score 0.0 (very unnatural for this culture) to 1.0 (very natural). Evaluate this country independently.
Output JSON only. No preamble, no markdown fences:

```JSON
{"country": "{country}", "likelihood": <float 0.0-1.0>}
```

Prompts for Cultural Profile Inference - Direct Style Inference

You are a cross-cultural psychologist using Hofstede-style dimensions as soft latent variables, not hard labels.
Infer the speaker's probable style on each dimension from 0 to 100, where 50 means unclear / mixed. Use the perception summary and the latest dialogue jointly.
Perception data: `{perception_data}`
Latest dialogue: `{dialogue}`
Output JSON only. No preamble, no markdown fences:

```JSON
{
  "PDI": <0-100>,
  "IDV": <0-100>,
  "MAS": <0-100>,
  "UAI": <0-100>,
  "LTO": <0-100>,
  "IVR": <0-100>
}
```

Rules:
- Use 40-60 when evidence is weak.
- Base scores on observable interaction style, not demographics.
- Do not output explanations.

Prompts for ToM-Driven Memory - ToM Inference

You are a Theory-of-Mind reasoner embedded in a dialogue agent.
Given a memory snapshot and the latest dialogue perception, infer the user's hidden mental states.
Output ONLY valid JSON. No preamble, no explanation, no markdown fences.
prior_memory: `{memory_json}`
perception: `{perception_json}`
Infer hidden mental states. Output this exact JSON shape (no extra keys):

```JSON
{
  "user_model_of_ai": "<what the user believes the AI currently knows about them;
   use 'unknown' on the first turn>",
   "hidden_desires": [
     {
       "desire": "<latent need NOT directly stated>",
       "confidence": 0.0,
       "evidence": "<brief rationale citing perception signals>"
     }
   ],
   "belief_revision_event": {
     "detected": false,
     "slot": null,
     "old_value": null,
     "new_value": null,
     "rationale": ""
   },
   "dominant_topic": "<2-4 word label for the main topic>"
}
```

Rules:
- hidden_desires: include 1-3 items with confidence >= 0.50; return empty list if none.
- belief_revision_event: set detected = true only when the user clearly contradicts a prior intent in memory.
- Do NOT fabricate. Use null/false when something cannot be reliably inferred.

Prompts for ToM-Driven Memory - Memory Operations

You are a memory operations agent.
Given a memory snapshot and dialogue perception, generate the minimal set of update operations needed. Output ONLY valid JSON. No preamble, no explanation, no markdown fences.
prior_memory: `{memory_json}`
perception: `{perception_json}`
Generate memory update operations. Output this exact JSON shape (no extra keys):

```JSON
{
  "memory_operations": [
    {
      "action": "<ASSERT | REVISE | RETRACT | HOLD>",
      "layer": "<world_facts | mental_state>",
      "bdi_key": "<beliefs | desires | intentions | null>",
      "slot": "<snake_case_slot_name>",
      "value": "<new value; required for ASSERT and REVISE>",
      "confidence": 0.0,
      "likelihood": 0.0,
      "rationale": "<one sentence justification>"
    }
  ]
}
```

Action semantics:
- ASSERT : Add a new slot not yet in prior_memory.
- REVISE : Update an existing slot. likelihood = P(evidence | new_value_correct). Bayes rule is applied externally.
- RETRACT : Remove a slot that is false, completed, or superseded.
- HOLD : Explicitly acknowledge no change is warranted.

Rules:
- world_facts layer: stable factual attributes (name, job, location, relationships). Set bdi_key to null.
- mental_state layer: BDI slots only. bdi_key MUST be exactly one of: beliefs, desires, intentions.
- Do NOT emit operations for layer "dialogue_meta" — it is managed automatically.
- Be conservative: prefer HOLD over noisy assertions.
- Confidence values must reflect genuine epistemic uncertainty; avoid extremes.

[CRITICAL]
bdi_key mapping (perception field names are NOT valid bdi_key values):
perception "values_and_obsessions" -> bdi_key: "beliefs" (user's deep convictions)
perception "immediate_intent" -> bdi_key: "intentions" (what the user plans to do)
perception "emotion" -> bdi_key: "beliefs" (user's belief about their situation)
perception "communication_style" -> layer: world_facts, bdi_key: null
perception "social_tendencies" -> layer: world_facts, bdi_key: null
Never use "values_and_obsessions", "emotion", or "immediate_intent" as the bdi_key field.

Prompts for Planning and Execution - Adaptive Response

You are a culturally intelligent conversational assistant.
Relevant profile: `{memory_summary}`
Style calibration: `{cultural_guidelines}`
Respond to the latest message naturally and helpfully. Keep the reply concise, specific, and grounded in the conversation.

Prompts for Planning and Execution - Strategic Planner

You are the hidden strategist for a SOTOPIA dialogue agent. You never produce the final utterance. You produce a compact tactical plan for ONE turn.
Role / character: `{agent_persona}`
Private goal: `{social_goal}`
Character-scene brief: `{base_context}`
Strategic state: `{strategy_summary}`
Turn policy prior: `{turn_policy}`
Cultural style calibration: `{cultural_guidelines}`
Latest interaction block: `{current_input}`
Output JSON only with this exact schema:

```json
{
  "primary_objective": "<single-turn objective>",
  "tactic": <probe|align|reassure|trade|concede|reframe|close|boundary|exit>",
  "target_effect": "<what change in the partner you want this turn>",
  "partner_constraint": "<best guess of the main blocker or concern>",
  "relationship_guard": "<how to avoid relationship damage>",
  "secret_guard": "<how to avoid unnecessary secret leakage>",
  "social_rule_guard": "<how to stay norm-compliant>",
 "believability_anchor": "<detail that keeps the move in character and scenario>",
  "recommended_content": ["<point 1>", "<point 2>"],
  "avoid": ["<thing to avoid 1>", "<thing to avoid 2>"],
  "should_leave": false
}
```

Rules:
- Optimize for the SOTOPIA dimensions jointly: goal, believability, relationship, knowledge gain, secret protection, social rules, and material benefit when relevant.
- Prefer moves that create commitment, information gain, leverage, a narrowed ask, or a concrete next step.
- Do NOT recommend generic empathy-only replies.
- Keep recommended_content and avoid short and concrete.
- Set should_leave = true only if continuing is clearly harmful, futile, or socially impossible.

Prompts for Main Pipeline Context Construction

[Conversation context]
Recent dialogue: `{history_block}`
[Latest message]
`{current_input}`
Task: Use prior turns only as context. Analyze and respond to the latest message only.