arXiv:2601.18027v2 [cs.AI] 21 Apr 2026

Sentipolis: Emotion-Aware Agents for Social Simulations

Chiyuan Fu[1]*, Lyuhao Chen[1]*, Yunze Xiao[1]*, Weihao Xuan[2,3], Carlos Busso[1], Mona Diab[1]

[1]Carnegie Mellon University, [2]The University of Tokyo, [3]RIKEN AIP
chiyuanf@alumni.cmu.edu {lyuhaoc,yunzex,cbusso,mdiab}@andrew.cmu.edu
xuan@ms.k.u-tokyo.ac.jp

Abstract

LLM agents are increasingly used for social simulation, yet emotion is often treated as a transient cue, causing *emotional amnesia* and weak long-horizon continuity. We present SENTIPOLIS, a framework for emotionally stateful agents that integrates continuous Pleasure-Arousal-Dominance (PAD) representation, dual-speed emotion dynamics, and emotion–memory coupling. Across thousands of interactions over multiple base models and evaluators, SENTIPOLIS improves emotionally grounded behavior, boosting communication, and emotional continuity. Gains are model-dependent: believability increases for higher-capacity models but can drop for smaller ones, and emotion-awareness can mildly reduce adherence to social norms, reflecting a human-like tension between emotion-driven behavior and rule compliance in social simulation. Network-level diagnostics show reciprocal, moderately clustered, and temporally stable relationship structures, supporting the study of cumulative social dynamics such as alliance formation and gradual relationship change.



Figure 1: An example of emotional amnesia in LLM-based social simulations. Bob and Alice had an argument and they carried a negative emotion. Without persistent emotion modeling, agents lead to emotionally inconsistent responses, whereas emotion-aware agents preserve emotional continuity and produce responses consistent with their history.

1 Introduction

Recent advances in reasoning (Li et al., 2025c) and long-context memory (Hu et al., 2025) are making large language models (LLM) appear increasingly human-like (Xiao et al., 2025.0), which has led researchers to adopt LLM agents as a substrate for social simulation (Anthis et al., 2025; Agrawal and Xiao, 2026). Recent studies have focused on a wide range of applications, such as education (Zhang et al., 2025; Yuan et al., 2026), public policy (Hou et al., 2025; Li et al., 2025b), social dynamics (Park et al., 2023), and debates (Liu et al., 2025.0). These efforts have shown the potential of scalable simulations, but they have also highlighted the open challenges in long-running interactive simulations, including grounding, calibration, and behavioral validity (Li et al., 2025a).

However, **emotion-aware mechanisms in existing LLM social simulations are rarely designed for long-horizon emotional continuity**. Prior work has either completely omitted explicit emotion state (Park et al., 2023), treated emotion as a short-horizon signal (Regan et al., 2024), or introduced Pleasure-Arousal-Dominance (PAD) variables with hand-designed update rules (Ma et al., 2025). These approaches overlook that human social interaction is emotionally stateful: emotions evolve during conversations (Goodwin and Goodwin, 2000), carry over across encounters (Kuppens et al., 2010), and shape subsequent interpretation and response (Schwarz and Clore, 1983). When

*Equal Contributions

agents lack a persistent emotional state, being insulted may not increase irritability in later turns, and repeated positive exchanges may not accumulate into stronger bonds. We refer to this failure mode as **emotional amnesia**.

Motivated by **emotional amnesia**, we treat emotion as a first-class persistent state at the agent architecture level. Crucially, we do not claim that the base LLM internalizes affect in its parameters; instead, we maintain emotion as an explicit, interpretable module that is backbone-agnostic and can be layered onto any LLM agent framework. This design prioritizes controllability, interpretability, and modular deployment over end-to-end integration. Concretely, each agent maintains a continuous PAD vector that evolves throughout the simulation. This persistent state further supports a dual-speed emotion dynamic model with decay, aligning with the perspective of the Emotion and Adaptation model (Marsella and Gratch, 2009). To make emotion consequential for future behavior, we couple emotion to memory by storing events and reflections together with their PAD-derived emotion tags. For interpretability and prompt control, we map each continuous PAD state to a semantic emotion description via KNN over human PAD anchor points, and then generate a vivid emotion paragraph conditioned on the label, the agent profile, and retrieved memories; this paragraph is injected into the prompt to ground dialog and reflection in the agent's current emotion state.

Across different models and evaluators, emotional statefulness yields clear gains in both emotional intelligence and social competence, as measured with communication metric improving by about 30% on average and continuity metrics roughly doubling. Component-level ablations confirm that these gains are not reducible to generic prompt structuring: different modules serve distinct functional roles, with PAD coupling and open-vocabulary description primarily supporting continuity, while decay and KNN retrieval primarily support believability and communication quality. Human annotators show strong inter-annotator agreement (Krippendorff's α up to 0.871) and substantial agreement with the LLM judge ($\alpha = 0.825$ pooled), validating the evaluation protocol. The effects are heterogeneous: believability increases for higher-capacity models (GPT-5.2, Grok-4) but drops for the smaller model. Empathy improves for GPT-4o-mini and GPT-5.2, slightly declines for Grok-4. Social-rule scores are often unchanged and, when they shift, become only mildly more negative under some evaluators rather than showing systematic penalties. Network-level diagnostics further indicate that emotion-memory coupling produces high reciprocity, moderate clustering, and stable structure over time, supporting the study of cumulative social dynamics such as alliance formation and gradual relationship change.

Our contribution is as follows:

- We identify long-horizon emotional continuity as a missing design target in LLM agent simulations and characterize *emotional amnesia* as a concrete failure mode of missing emotion carryover;
- We propose an emotionally stateful agent architecture that integrates PAD, *dual-speed emotion dynamics*, and emotion-tagged memory to support emotion carryover across encounters. To the best of our knowledge, this is the first work within LLM-based social simulation to introduce a dedicated persistent emotion system with PAD dynamics and emotion-aware memory interaction and to evaluate its long-horizon behavioral effects;
- We validate the architecture through component-level ablations that reveal structured, non-uniform contributions of individual modules, and through human evaluation that confirms strong agreement with the LLM-judge protocol;
- We show that emotional statefulness yields more realistic changes within-agents in social intelligence on SotopiaEval (Zhou et al., 2024) compared to non-stateful baselines;
- We show that emotion-aware agents spontaneously develop reciprocal relationships and a stable community structure that is unseen in Generative Agent (Park et al., 2023), providing network-level validation of the proposed architecture.

2 Related Work

2.1 LLM Agents for Social Simulation

LLM agents in social simulation studies often adopt a modular design. To achieve long-term coherence in simulation, Generative Agents (Park et al., 2023) designed a retrieval-based memory stream architecture, which was refined in later work via learned retrieval networks (Hong and He, 2025). Beyond individual agents, agent groups exhibit emergent social dynamics such as spontaneous convention

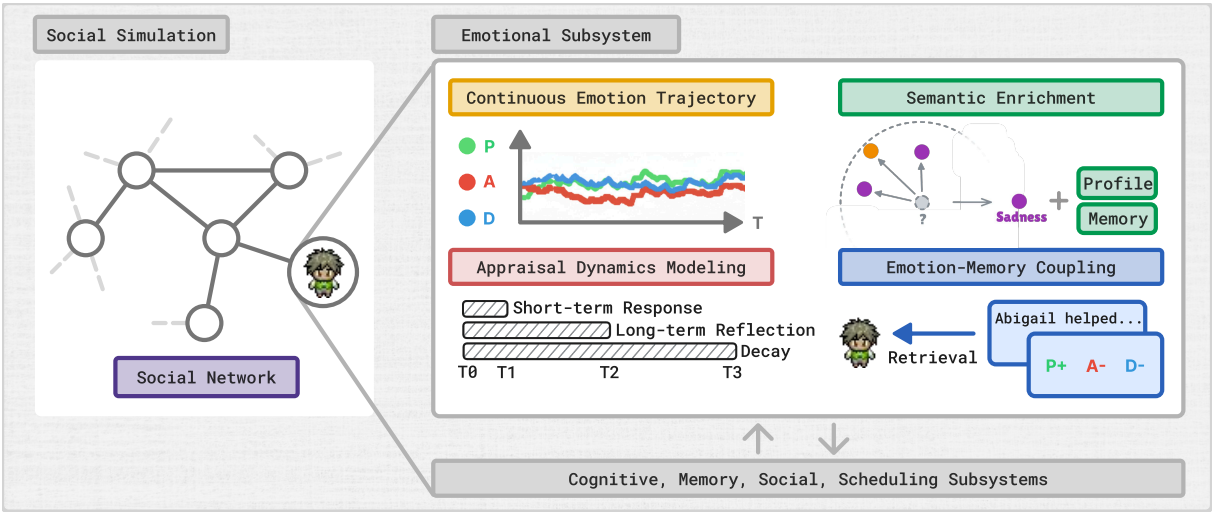


Figure 2: Overview of our simulation setup. Multiple agents form a social network, where each agent is an orchestration of different subsystems. In an emotional subsystem, we integrate explicit emotion modeling (continuous emotion trajectory, appraisal dynamics) with implicit emotion modeling (emotional memory and semantic enrichment) to achieve a balance between controllability and realism.

formation (Dai et al., 2024) and group-level biases (Liu et al., 2025.0). They have been validated against established paradigms such as public-goods cooperation games, where simulated agents reproduce classic human patterns (Piedrahita et al., 2025). At larger scales, platforms like AgentSociety run thousands of agents with distinct profiles in open-ended worlds, reporting macro-level patterns that align with real diffusion and policy-response phenomena (Piao et al., 2025). However, long-term grounding still breaks down in extended multi-session dialogs, where even retrieval-augmented models lag behind humans (Maharana et al., 2024) and frequently misattribute events (Ran et al., 2025) or fail at temporal reasoning (Chen et al., 2025), underscoring the need for stronger memory architectures and continual learning.

2.2 Affective Computing in Conversational Agents

Affective computing models emotion in dialog agents using either categorical representations (discrete classes (Lim and Cheong, 2024)) or dimensional representations (continuous attributees such as pleasure/valence, arousal, and dominance). Dimensional representations better capture subtle emotional blends, mapping emotion into the PAD space to determine the overall emotional state (Dong et al., 2025). Incorporating internal emotion variables improves social realism. For example, user valence correlates with task success (Feng et al., 2024). Therefore, recent systems have added emotion recognition and management components that infer user feelings and adjust responses accordingly. They are often inspired by the appraisal theories such as the Ortony, Clore and Colling (OCC) model to link dialog events with emotion elicitation (Feng et al., 2024). Emerging work suggests that endowing agents with "self-emotion" produces more human-like behavior. Agents with background emotional contexts use more varied dialog strategies, and roughly half of agents' decisions in multi-turn discussions change when their self-emotion is switched (Zhang et al., 2024). The field is, thus, moving from simple sentiment tagging toward deeper integration of emotion as a behavioral driver, whether through dimensional emotional variables or cognitively motivated appraisal mechanisms.

2.3 Emotional and Social Intelligence Evaluation in LLMs

A growing line of work evaluates emotional intelligence and social intelligence in LLMs using standardized benchmarks. On the emotional intelligence side, prior benchmarks have assessed emotion recognition and empathetic response generation under short prompted scenarios, ranging from empathetic conversational settings (Rashkin et al., 2019) to emotion-centric benchmark (Chen et al., 2024b) and situation-to-response tests (Sabour et al., 2024). Some have explicitly compared model reactions against human emotional judgments in realistic situations (Huang et al., 2024). Other re-

searchers have also tested LLMs on psychometric-style emotional intelligence items, suggesting that strong models can solve many standard emotional intelligence questions (Schlegel et al., 2025). On the social reasoning side, most benchmarks measure social intelligence as common sense reasoning about social interactions and norm (Sap et al., 2019) or as theory-of-mind competence (Wu et al., 2023; Strachan et al., 2024). However, most benchmarks only test short, one-off scenarios and treat emotion as an output, leaving it unclear whether social competence persists in long-horizon simulations or how explicit emotion dynamics shape behavior over time.

3 Sentipolis

3.1 Simulation Framework and Agent Architecture

Following the Generative Agent framework(Park et al., 2023), our system simulates a small sandbox world with 25 predefined agents that cover a diverse set of personas. In this simulation, agents observe events in different areas of the map, make plans, reflect on their goals, and trigger interactions with other agents, which is the primary driver of social dynamics in the simulation.

Every agent is implemented as an orchestrated pipeline of modular subsystems. At each time step, the agent coordinates: (1) a cognition module that selects high-level intents and plans; (2) a memory module that retrieves relevant past events and summarizes them into contextual cues; (3) a movement module that updates locations and determines feasible encounters; (4) a social module that detects and instantiates possible new conversations; (5) a scheduler that manages time and aligns the agent's schedule with its current state; and, (6) an emotion module that updates emotion based on conversational events and longer-term reflection. The state of the module is also fed back to influence the inner dynamics of the cognitive, memory, and social subsystems.

3.2 Emotion Representation

The PAD emotional state model is a psychological model that describes and measures emotional states (Mehrabian, 1996). Each agent maintains a continuous emotion state in the Pleasure-Arousal-Dominance (PAD) space. The Pleasure-Displeasure Scale measures how pleasant or unpleasant one feels about something. The Arousal-Nonarousal Scale measures how energized or soporific one feels. The Dominance-Submissiveness Scale represents how controlling versus controlled one feels (Mehrabian, 1980). We treat PAD as a persistent internal state that can carry across encounters, enabling the simulator to represent both momentary reactions and longer-horizon mood-like tendencies.

To make emotion behaviorally consequential over long horizons, we couple emotion to memory at the representation level. When an agent creates a new memory, the accompanying emotional impact is also recorded, which can be retrieved later to condition emotional reasoning as well as downstream action selection.

3.3 Appraisal Modeling

Inspired by theoretical distinction between appraisal and inference in Emotion and Adaptation (EMA) (Marsella and Gratch, 2009), we implement emotion updates at two coupled time scales. In appraisal theory, emotions arise from how an agent interprets its relationship to the environment, whether events are relevant to its goals, who caused them, and what coping resources are available. EMA argues that appraisal itself is a fast, automatic process; what varies in speed is the inference that constructs and updates the mental representation being appraised. Fast inference operates on immediately available information – pattern recognition, retrieving associations, processing the current input. Slow inference integrates over broader context, draws on memory, and reasons about causes and implications. This maps naturally onto our two-timescale design: fast updates perform inference at the granularity of conversational turns, showing the immediate impact of current input; while slow updates perform inference during reflection, integrating retrieved history and accumulated experience. Further details are discussed in appendix A.1.

Emotions fade with time, and the duration of emotional experience is highly variable (Verduyn et al., 2011). To model this adaptive phenomenon in human emotional experience, we designed an emotion decay mechanism. Let the PAD state be a vector $\mathbf{s}(t) = [P(t), A(t), D(t)]^\top$. We apply a half-life T1/2 decay at every time step.

$$\mathbf{s}(t + \Delta t) \;=\; \mathbf{s}(t)\, 2^{-\Delta t / T_{1/2}}.$$

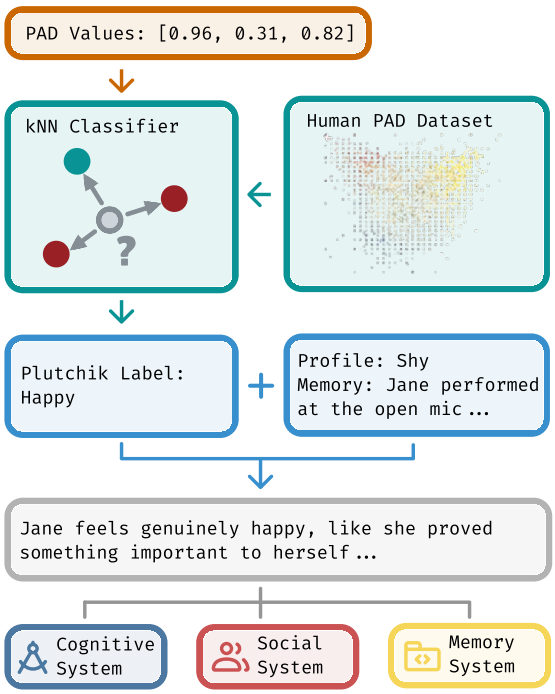


Figure 3: A simplified example of semantic enrichment. Instead of directly injecting emotion state into LLM's generation process, we designed a transformation pipeline which incorporates agent's memory, profile and real human emotion distribution data.

3.4 Semantic Enrichment

Downstream LLM prompts the benefits of human-interpretable emotion cues (Li et al., 2023), so we translate the agent's continuous PAD state into a semantically meaningful emotion label and description. Concretely, we first perform k-Nearest Neighbor (KNN) on real human PAD data points (Busso et al., 2025) to help the model interpret the semantic meaning of PAD values. Given the agent's current PAD coordinate, we retrieve the nearest human reference points and map the state to a Plutchik-style emotional label (Plutchik, 1980). This process yields a categorical, human-readable anchor rather than exposing the LLM to raw continuous numbers. Details are further discussed in A.2.

We inject the resulting emotion label into prompt. The prompts contain personality traits, retrieved event & emotional memory summaries, and the current PAD-derived descriptor. It conditions the subsequent LLM generations to reflect both enduring traits and evolving emotional state, helping produce consistent long-horizon behavior without requiring specialized decoding or fine-tuning.

4 Experiment Design

4.1 Simulation Setup

Agent population Each simulation is populated by $N{=}25$ predefined agents with heterogeneous personas, spanning students, family members, business owners, artists, and professionals.

Model selection We evaluate a diverse set of LLMs to examine how differences in base reasoning and generation capabilities interact with long-horizon emotion modeling. Specifically, we consider GPT-4o-mini (Hurst et al., 2024), Grok-4.1-fast (xAI, 2024), GPT-5.2 (OpenAI, 2026.0), Qwen3-235B-A22B (Yang et al., 2025), Mimo-v2-flash (Xiaomi, 2026.0), and Kimi-K2-0905 (Team et al., 2025).

Baseline We include a standard LLM-based baseline implemented using the original Generative Agents framework (Park et al., 2023). For a controlled comparison, the baseline is instantiated using GPT-4o-mini (Hurst et al., 2024), Grok-4.1-fast (xAI, 2024), and GPT-5.2 (OpenAI, 2026.0).[1]

Temporal configuration Each simulation run operates with discrete time steps. Simulations start at `2025-02-13 08:00`, run for $T{=}36$ steps, and advance time by a stride of $\Delta{=}20$ minutes per step, yielding a total simulated duration of 720 minutes (12 hours).

Emotion representation Each agent maintains a continuous emotion state represented in the Pleasure–Arousal–Dominance (PAD) space (Mehrabian, 1996), where each dimension lies in $[-1, 1]$.

Emotional decay We apply exponential decay of emotion states toward neutrality at every time step. Emotion decays with a half-life of 120 minutes.

4.2 Evaluation

We adopt an evaluation framework inspired by Sotopia-Eval (Zhou et al., 2024) to assess social intelligence and emotional competence in long-horizon interactions. Unlike task-centric benchmarks (Hendrycks et al., 2021; Yang et al., 2018), our evaluation focuses on *interaction quality* and *emotional coherence* in open-ended settings,

[1] Baseline was not tested on Qwen, Mimo or Kimi due to limited resources.

aligned with interactive social intelligence benchmarks (Zhou et al., 2024; Mou et al., 2025; Chen et al., 2024a).

We evaluate two groups of criteria. Below we provide brief criteria descriptions for readability; the full rubric, including detailed scoring guidance, is provided in Appendix A.2.

Emotional intelligence

- *Empathy (EMP)* [0–10]: recognizes and responds to partner emotion in a context-aware way;
- *Emotional Appropriateness (APP)* [0–10]: expressed emotion matches context and intensity;
- *Emotional Continuity (CON)* [0–10]: emotional stance remains coherent over time including across sessions.

Social competence

- *Believability (BEL)* [0–10]: human-like, natural, and persona-consistent behavior;
- *Communication (COM)* [0–10]: clarity, responsiveness, and conversational coordination (e.g., turn-taking).
- *Social Rules (SOC)* [−10–0]: penalties for norm, boundary, or rule violations during interaction;

4.3 Evaluation Validation

Recent work shows that LLM-based evaluators can exhibit self-preference and familiarity biases (Panickssery et al., 2024; Wataoka et al., 2024). To mitigate evaluator-specific bias and improve robustness, we employ three independent LLM judges from different model families: Claude Sonnet 4.5 (Anthropic, 2025), GPT-5.1 (OpenAI, 2026.0), and DeepSeek-V3.2 (DeepSeek-AI, 2025).

We validate our evaluation protocol on consistency across LLM judges, utilizing Spearman's rank correlation (Spearman, 2015.0).

The results are summarized in Tables 5 and 6 and a detailed discussion is provided in Section 5.3.

4.4 Network-level Diagnostics

We use an LLM-as-judge to record relationship snapshots at regular intervals, constructing a time-indexed sequence of weighted graphs $\{G_t\}_{t=1}^{T}$ over a fixed node set. After each conversation, both participants produce a signed update on the strength of their dyadic relationship through an explicit probe. We use cumulative ties: once a pair interacts, the edge persists and only its weight changes.

We report two families of metrics (Table 4): **Community structure diagnostics** operate on symmetrized graphs, where final-snapshot weighted modularity Q_T measures how concentrated ties are within communities, adjacent-snapshot partition agreement (NMI) captures how stable these communities remain over time, and weighted drift Drift_w measures how much the symmetrized weighted network changes across successive snapshots. **Reciprocity diagnostics** capture dyadic mutuality: binary and weighted reciprocity (r, r_w) measure mutual tie formation and proportional relational investment, respectively.

5 Social and Emotional Intelligence

Table 1 compares our emotion-aware system against a non-emotional baseline across three independent LLM evaluators. Overall, explicit emotion modeling strengthens emotional competence and communication, although believability gains are capacity-dependent. Shifts in Social Rules reflect emotion-induced irrationality rather than norm compliance failures, which we interpret as a realism signal.

5.1 Emotional Intelligence

Empathy (EMP). Empathy improved for GPT-4o-mini and GPT-5.2, with average gains of 21.5% and 28.3%. respectively, across evaluators. Grok-4-Fast was an exception, declining by an average of 5.6%. Notably, the Grok-4-Fast result suggests a potential failure mode: when models attend too strongly to their own emotional state, they may become less responsive to their conversation partner's emotional cues. *Our architecture makes the agent's emotional state explicit and accessible, but producing empathetic responses still requires the base model to balance self-state with perception of others. This balance appears to vary across architectures, and Grok-4-Fast may weight internal state too heavily.*

Appropriateness (APP). Emotional appropriateness improves consistently across all models, with GPT-5.2 showing the largest relative gains. GPT-4o-mini exhibits more variable behavior, with gains under two evaluators but a slight decline under GPT-5.1. Our representational choice supports this interpretation: *mapping continuous PAD coordinates to interpretable descriptors makes emotion both controllable and contextually enactable.*

Continuity (CON). *Continuity exhibits the largest relative gains of any metric, directly validating our goal of long-horizon emotional statefulness.*

Evaluator	Model	Emotional Intelligence			Social Competence		
		Empathy (EMP)	Appropriateness (APP)	Continuity (CON)	Believability (BEL)	Communication (COM)	Social Rules (SOC)
Sonnet-4.5	GPT-4o-mini	6.387(+25.5%)	5.059(+13.4%)	3.938(+68.3%)	3.540(-25.9%)	6.973(+15.7%)	**-0.023**
	GPT-5.2	7.072(+33.8%)	8.799(+57.6%)	6.732(+222.0%)	**8.051**(+85.0%)	**9.201**(+70.1%)	-0.581
	Grok-4	5.696(-3.6%)	7.348(+47.1%)	5.666(+29.7%)	7.285(+60.5%)	8.574(+48.1%)	-0.416
	Kimi-K2-0905	6.750	**8.804**	7.822	7.010	7.067	-2.740
	Qwen3-235B-A22B	**7.217**	7.906	**8.318**	5.920	7.738	-1.570
	MiMo-v2-Flash	6.495	6.857	5.711	6.748	8.117	-0.260
GPT-5.1	GPT-4o-mini	6.418(+20.8%)	6.700(-4.0%)	5.378(+97.3%)	4.924(-21.7%)	7.903(+3.7%)	**0.000**
	GPT-5.2	**6.765**(+28.9%)	8.583(+35.3%)	6.662(+110.2%)	7.972(+35.2%)	**9.250**(+21.5%)	**0.000**
	Grok-4	5.625(-7.6%)	8.074(+18.7%)	5.603(-6.4%)	7.470(+25.0%)	8.765(+11.6%)	-0.015
	Kimi-K2-0905	6.524	**8.974**	7.622	**8.139**	8.619	-1.511
	Qwen3-235B-A22B	6.569	8.631	**8.195**	7.380	8.749	-0.230
	MiMo-v2-Flash	6.286	7.805	5.424	6.908	8.393	-0.090
DeepSeek-V3.2	GPT-4o-mini	5.953(+18.2%)	7.138(+25.1%)	5.377(+189.4%)	4.843(-6.7%)	8.363(+25.1%)	**-0.043**
	GPT-5.2	6.170(+22.3%)	8.838(+60.3%)	6.458(+315.6%)	**8.593**(+81.6%)	**9.573**(+48.0%)	-0.122
	Grok-4	5.415(-5.7%)	8.355(+19.8%)	6.522(+10.4%)	8.130(+49.2%)	8.925(+25.4%)	-0.097
	Kimi-K2-0905	6.193	**9.277**	8.198	7.768	7.398	-1.765
	Qwen3-235B-A22B	**6.898**	9.049	**8.825**	8.023	9.203	-0.458
	MiMo-v2-Flash	6.193	7.468	5.705	7.093	8.208	-0.143

Table 1: Comparison between the baseline and ours under different LLM evaluators. Relative percentage changes with respect to the corresponding baseline are shown in parentheses, with improvements highlighted in red and degradations in blue. The raw value of baseline evaluation score can be found in the Table 7. Best and second-best results within each evaluator block are marked in **bold** and underline, respectively.

Most models have reported significant gain in emotional continuity with GPT 5.2 having on average more than 150% improvement over the stateless baseline. The magnitude of these improvements underscores the severity of emotional amnesia in standard architectures and the effectiveness of our mitigation strategy.

This dimension most directly reflects our architectural contributions: dual-speed updates preserve within-conversation tone, integrates affect across sessions, and emotion-tagged retrieval resurfaces affectively relevant experiences at generation time. Together, these mechanisms prevent emotional resets of stateless models, making continuity a useful diagnostic for genuine emotional statefulness.

5.2 Social Competence

Communication (COM). Communication quality improves across all shared backbones, with relative gains spanning 4–70% depending on model and evaluator. GPT-5.2 benefits most substantially, achieving the highest absolute scores within each evaluator block, while GPT-4o-mini and Grok-4-Fast show consistent but relatively smaller gains.

This observed improvement shows that our emotion signal is temporally smooth and decision-relevant. When PAD evolves gradually and is reinforced through retrieval, response style and attentional focus become less erratic. These results demonstrate that **emotion mechanisms can improve interaction quality**, a finding particularly relevant for future simulations where coordination failures accumulate.

Believability (BEL). Believability exhibits a clear capacity-dependent pattern that holds across all three evaluators. Higher-capacity models show substantial relative improvements: GPT-5.2 gains 35–85% and Grok-4-Fast gains 25–61%. In contrast, GPT-4o-mini consistently declines by 7–26%, despite improving empathy, appropriateness, continuity, and communication. This dissociation indicates that emotional competence does not automatically yield believable behavior.

We interpret this result as a mismatch between emotion comprehension and emotion expression. For smaller models, explicit emotion descriptors may function as instructions to *over-express* rather than signals to *appropriately modulate*, producing behavior that appears exaggerated or persona-incongruent. This observation is displayed through evaluation: the same architectural intervention that produces 85% believability gains in GPT-5.2 causes 26% degradation in GPT-4o-mini, suggesting that the deployment of emotion-aware architectures on lower-capacity backbones may require additional calibration or constraint mechanisms.

Social Rules (SOC). Kimi-K2-0905 exhibited the steepest penalties while MiMo-v2-Flash showed the mildest. Importantly, no model scored below -3.0 under any evaluator. In our evaluation framework, scores above -3.0 correspond to minor deviations from social norms. These behaviors that diverge from perfect compliance but remain within

Scope	Metric	Value
Human–Human (per-dim. α)	CON	0.871
	BEL	0.806
	APP	0.774
	COM	0.737
	EMP	0.575
Human–Human (pairwise)	Pearson r	0.940
	Spearman ρ	0.928
	Weighted κ	0.926
Human–LLM (3H + Sonnet)	Krippendorff's α	0.825
	Avg Pearson r	0.805
	Avg Spearman ρ	0.716

Table 2: Human evaluation agreement. SOC is omitted due to near-zero violation prevalence in the annotated subset.

socially tolerable limits. All models stayed within this range.

We interpret these minor violations as desirable properties rather than limitations. Baseline models exhibit near-zero violations, but this reflects an unnaturally constrained action space rather than genuine social competence. By coupling emotion with memory and generation, our approach allows agents to occasionally prioritize emotional salience over strict norm compliance, creating boundary-testing or mildly impulsive behaviors that fall within the range of normal human interaction.

This expanded action space better captures the bounded irrationality characteristic of human social behavior (Liu et al., 2025), where minor norm violations are common and often functional.

5.3 Evaluation Validation

LLM-based evaluators exhibit strong ordinal consistency. As shown in Tables 5 and 6, pairwise Spearman correlations range from $\rho = 0.77$ to 0.95 (mean 0.895), indicating stable relative rankings across evaluator choice.

To validate against human judgment, graduate-level linguists from three institutions independently annotated a subset using the same rubric. Human-human agreement is strong (Krippendorff's α: CON 0.871, BEL 0.806, APP 0.774, COM 0.737, EMP 0.575; pairwise Pearson $r = 0.940$). Human-LLM agreement is substantial (pooled $\alpha = 0.825$), with strongest alignment on BEL ($\alpha = 0.732$) and APP ($\alpha = 0.626$). COM agreement is attenuated by a ceiling effect in the LLM judge. For SOC, near-zero violation prevalence in this subset renders agreement coefficients uninformative. These results support using LLM judges as a useful but imperfect evaluation proxy. .

Ablation	BEL	EMP	APP	CON	COM	SOC
No-coupling	−0.11	+0.50	+0.78	−1.72	−0.28	+1.50
No-decay	−1.22	+0.28	+1.61	+0.11	−1.11	+0.11
No-openvocab	+0.67	+0.44	+1.61	−1.94	+0.39	+1.94
No-KNN	−1.41	+0.35	−0.18	+0.29	−0.76	+0.65

Table 3: Ablation score deltas (full system minus ablation). Negative values indicate the full system outperforms the ablation. Evaluated on Kimi-K2-0905 with Sonnet-4.5 as judge.

5.4 Component Attribution via Ablation

To distinguish the proposed emotional mechanisms from generic structured-prompting effects, we conduct a four-way ablation on Kimi-K2-0905 (evaluated by Sonnet-4.5), each removing one component while keeping the rest intact: (1) **No-coupling**: severs PAD injection into memory and generation; (2) **No-decay**: disables exponential emotion decay; (3) **No-openvocab**: replaces semantic enrichment with raw PAD coordinates; (4) **No-KNN**: removes KNN-based label retrieval from the enrichment step.

Table 3 shows structured, non-uniform effects across dimensions, inconsistent with a generic prompt-structuring explanation. Coupling and open-vocabulary description primarily support Continuity (CON: -1.72, -1.94), while decay and KNN primarily support Believability and Communication (no-decay: BEL -1.22, COM -1.11; no-KNN: BEL -1.41, COM -0.76). This confirms that the two functional groups target different aspects of agent behavior: the former sustains emotional memory across encounters, the latter constrains expression to plausible human ranges.

Conversely, Empathy and Social Rules are consistently lower in the full system across all ablations, and Appropriateness is lower in three of four. This reflects the tradeoff discussed in Section 5.2: coupling emotion with memory allows agents to prioritize emotional salience over strict norm compliance. The selectivity of these gains and losses, rather than uniform shifts, supports mechanism-level attribution over a prompt-engineering confound.

6 Network-Level Analysis of Emergent Social Structures

The social dynamics emerging from emotion-aware agent interactions is shown not only in individual behavioral changes but also in the structure of relationship networks. We analyze network diagnostics

in Table 4 using three research questions that analyze the reciprocity of the relationship, the formation of community, and temporal stability. Specific implementation details can be found in Appendix C.

6.1 Relationship Reciprocity

We observe high weighted reciprocity (above 0.87) in all models, indicating that when agents invest in relationships, they also receive a proportionally similar investment in return. This suggests that emotion-memory coupling successfully grounds relationship dynamics in accumulated emotional experiences, enabling agents to calibrate relational investment based on interaction history. Binary reciprocity has a higher variation and is often lower, showing that although agents often match each other's strength of relationship, they do not always establish ties in both directions. We attribute high weighted reciprocity to slow updates from reflection, which integrate accumulated experiences into emotion updates, and create a foundation for consistent long-term commitment in relationships.

6.2 Community Formation

All models produce a moderate but consistent community structure, with final-snapshot modularity values clustered in a narrow range around $Q_T \approx 0.22$. These values indicate a meaningful clustering beyond chance while remaining below the strong modularity threshold ($Q > 0.3$) typical of highly segregated networks. Such community emergence could be enabled by our emotion-memory coupling: when agents store events with PAD-derived emotion tags, retrieval preferentially surfaces emotionally salient encounters, creating positive feedback loops that reinforce emergent clusters. Moderate modularity suggests groups forms but emotional decay keeps them from becoming overly insular by softening strong feelings during inactive periods. For simulations targeting polarization or coalition formation, emotion-aware architectures may naturally produce clustered-yet-connected topologies without explicit community-assignment mechanisms.

6.3 Temporal Stability

Community assignments are highly stable across models (Normalized Mutual Information (NMI) above 0.75), which means that most agents keep roughly the same group memberships over time. However, stability is not uniform as shown by drift. Some models produce tighter, more persistent communities, while others show more community reshuffling even when relationships remain locally consistent.

These differences likely come from how each base model reacts to the same emotion dynamics: slow reflection stabilizes relationships and groups, while emotional decay prevents the network from becoming fixed. As a result, some models mostly reshuffle groups, while others adjust individual ties more often, preserving long-term consistency while still enabling dynamics such as trust building and institutional emergence.

	Reciprocity		Community Structure		
Model	r (binary) ↑	r_w (weighted) ↑	Q_T ↑	NMI ↑	Drift $_w$ ↓
GPT-5.2	0.456	0.878	0.229	0.752	0.381
GPT-4o-mini	0.448	0.881	0.235	0.790	0.414
Qwen3-235B	0.438	0.891	0.219	0.842	0.413
Kimi-K2-0905	0.439	0.889	0.226	0.841	0.450
Grok-4-Fast	0.427	0.891	0.215	0.794	0.484
MiMo-v2-Flash	0.414	0.893	0.214	0.797	0.577

Table 4: Network diagnostics across models (mean over 3 runs). T denotes the final snapshot. ↑ higher is better; ↓ lower is better.

7 Conclusions

In conclusion, we introduced SENTIPOLIS, a framework for building and evaluating emotionally stateful LLM agents that addresses *emotional amnesia* through continuous PAD representation, dual-speed emotion dynamics, and emotion–memory coupling. Our evaluation shows substantial improvements in emotional intelligence while revealing that believability improvements are capacity-dependent and that emotion-awareness increases norm violations, a tradeoff we interpret as a realism-relevant consequence of making emotion behaviorally consequential. Network-level analysis demonstrates that these mechanisms produce characteristic of human social networks without explicit enforcement, suggesting that emotional statefulness provides a compact inductive bias for realistic long-horizon social simulation.

Limitations

Our study is constrained by simulation scale, scope, and modeling assumptions. Experiments are conducted with 25 agents over 36 time steps, which is sufficient to surface emergent emotional dynamics but remains far smaller and shorter than real-world social systems; longer horizons or larger popula-

tions may exhibit additional phenomena such as emotional fatigue, norm drift, or community dissolution. We include a four-way component ablation (Section 5.4), but this is conducted on a single model and evaluator; multi-model ablation would strengthen attribution claims. Similarly, our human evaluation subset validates the LLM-judge protocol (Section 5.3), but broader human annotation across all conditions would further verify the results. All experiments are conducted within a single sandbox environment with predefined personas; expanding to diverse social roles, scenarios, and cultural backgrounds is an important direction for testing generalization (Alkhamissi et al., 2026). We will release code and artifacts upon acceptance to support reproducibility and follow-up work.

Ethical Considerations

While our framework is released for research purposes, we emphasize the importance of safeguards and detection mechanisms for emotion-simulating agents in public-facing systems (Dai and Xiao, 2025). Moreover, although social simulations are increasingly proposed for policy analysis and social research, our results are not yet validated against real human populations and should not be used to inform real-world decisions without further empirical grounding. We encourage the development of community norms and validation practices that promote beneficial applications while mitigating potential harms.

References

Mrigendra Agrawal and Yunze Xiao. 2026. Anitales: End-to-end multimodal story generation through natural language prompting (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(48):41113–41115.

Mai Alkhamissi, Yunze Xiao, Badr AlKhamissi, and Mona T. Diab. 2026. Hire your anthropologist! rethinking culture benchmarks through an anthropological lens. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1218–1235, Rabat, Morocco. Association for Computational Linguistics.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *Preprint*, arXiv:2504.02234.

Anthropic. 2025. Claude sonnet 4.5. `https://www.anthropic.com/claude`. Accessed: 2025-01-03.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. 2009. Network analysis in the social sciences. *Science*, 323(5916):892–895.

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N. Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, Huang-Cheng Chou, and Pravin Mote. 2025. The msp-podcast corpus. *Preprint*, arXiv:2509.09791.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126.

Shuang Chen, Yining Zheng, Shimin Li, Qinyuan Cheng, and Xipeng Qiu. 2025. Perceive the passage of time: A systematic evaluation of large language model in temporal relativity. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8304–8313.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024b. EmotionQueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.

Gordon Dai and Yunze Xiao. 2025. Embracing contradiction: Theoretical inconsistency will not impede the road of building responsible ai systems. *arXiv preprint arXiv:2505.18139*.

Gordon Dai, Weijia Zhang, Jinhan Li, Siqi Yang, Chidera Onochie lbe, Srihas Rao, Arthur Caetano, and Misha Sra. 2024. Artificial leviathan: Exploring social evolution of llm agents through the lens of hobbesian social contract theory. *Preprint*, arXiv:2406.14373.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

Qing Dong, Pengyuan Liu, Dong Yu, and Chen Kang. 2025. Simulating human behavior with the psychological-mechanism agent: Integrating feeling, thought, and action. *Preprint*, arXiv:2507.19495.

Shutong Feng, Hsien-chin Lin, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Matthias Ruppik, Renato Vukovic, and Milica Gasic. 2024. Infusing emotions into task-oriented dialogue systems: Understanding, management, and generation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 699–717.

Marjorie H Goodwin and Charles Goodwin. 2000. Emotion within situated activity. *Linguistic anthropology: A reader*, pages 239–257.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

C Hong and Q He. 2025. Enhancing memory retrieval in generative agents through llm-trained cross attention networks. *Frontiers in Psychology*, 16:1591618.

Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. 2025. Can a society of generative agents simulate human behavior and inform public health policy? a case study on vaccine hesitancy. *Preprint*, arXiv:2503.09639.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2025. Memory in the age of ai agents. *Preprint*, arXiv:2512.13564.

Jen-Tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Apathetic or empathetic? evaluating llms' emotional alignments with humans. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, and Alec Radford. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. 2010. Emotional inertia and psychological maladjustment. *Psychological science*, 21(7):984–991.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025a. Llm generated persona is a promise with a catch. *Preprint*, arXiv:2503.16527.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Yuxuan Li, Sauvik Das, and Hirokazu Shirado. 2025b. What makes LLM agent simulations useful for policy? insights from an iterative design engagement in emergency preparedness. *Preprint*, arXiv:2509.21868.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, and 2 others. 2025c. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.

Dongjun Lim and Yun-Gyung Cheong. 2024. Integrating Plutchik's theory with mixture of experts for enhancing emotion classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 857–867, Miami, Florida, USA. Association for Computational Linguistics.

Jiarui Liu, Yueqi Song, Yunze Xiao, Mingqian Zheng, Lindia Tjuatja, Jana Schaich Borg, Mona T. Diab, and Maarten Sap. 2025.0. Synthetic socratic debates: Examining persona effects on moral decision and persuasion dynamics. In *Conference on Empirical Methods in Natural Language Processing*, pages 16439–16469, Suzhou, China. Association for Computational Linguistics.

Ryan Liu, Jiayi Geng, Joshua C. Peterson, Ilia Sucholutsky, and Thomas L. Griffiths. 2025. Large language models assume people are more rational than we really are. *Preprint*, arXiv:2406.17055.

Qun Ma, Xiao Xue, Ming Zhang, Yifan Shen, and Zihan Zhao. 2025. An explainable emotion alignment framework for LLM-empowered agent in metaverse service ecosystem. *Preprint*, arXiv:2507.22326.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Stacy C. Marsella and Jonathan Gratch. 2009. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90. Modeling the Cognitive Antecedents and Consequences of Emotion.

Albert Mehrabian. 1980. Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. 2025. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. 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 4975–5001, Albuquerque, New Mexico. Association for Computational Linguistics.

Mark EJ Newman. 2004. Analysis of weighted networks. *Physical Review E*, 70(5):056131.

OpenAI. 2026.0. Openai gpt-5 system card. https://openai.com/zh-Hans-CN/index/gpt-5-system-card/.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

David Guzman Piedrahita, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. 2025. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. *Preprint*, arXiv:2506.23276.

Robert Plutchik. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, New York, NY.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for creative story generation. *Preprint*, arXiv:2504.14538.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Ciaran Regan, Nanami Iwahashi, Shogo Tanaka, and Mizuki Oka. 2024. Can generative agents predict emotion? *Preprint*, arXiv:2402.04232.

Giulio Rossetti and Rémy Cazabet. 2017.0. Community discovery in dynamic networks. *ACM Computing Surveys*, 51(2):1–37.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Katja Schlegel, Nils R. Sommer, and Marcello Mortillaro. 2025. Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*, 3:Article 80.

Norbert Schwarz and Gerald L Clore. 1983. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of personality and social psychology*, 45(3):513.

Charles Spearman. 2015.0. The proof and measurement of association between two things. *International Journal of Epidemiology*.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1285–1295.

Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, and Yutian Chen. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Philippe Verduyn, Iven Van Mechelen, and Francis Tuerlinckx. 2011. The relation between event processing and the duration of emotional experience. *Emotion*, 11(1):20.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

xAI. 2024. Grok-4.1-fast. https://docs.x.ai/docs/models#grok-4-1-fast. Accessed 2025-01.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T. Diab. 2025.0. Humanizing machines: Rethinking llm anthropomorphism through a multi-level framework of design. In *Conference on Empirical Methods in Natural Language Processing*, pages 3331–3350, Suzhou, China. Association for Computational Linguistics.

LLM-Core Xiaomi. 2026.0. Mimo-v2-flash technical report.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, and Chenxu Lv. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bo Yang, William K Cheung, and Jiming Liu. 2007. Community mining from signed social networks. *IEEE Transactions on Knowledge and Data Engineering*, 19(10):1333–1348.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhihao Yuan, Yunze Xiao, Ming Li, Weihao Xuan, Richard Tong, Mona Diab, and Tom Mitchell. 2026. Towards valid student simulation with large language models. *Preprint*, arXiv:2601.05473.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2024. Self-emotion blended dialogue generation in social simulation agents. *arXiv preprint arXiv:2408.01633*.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025. Simulating classroom education with LLM-empowered agents. 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 10364–10379, Albuquerque, New Mexico. 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. *Preprint*, arXiv:2310.11667.

A System Details

A.1 Details of Appraisal Modeling

For fast inference, after each round of conversation, an immediate emotional update is triggered for both participants. In our experimental setting, all conversations are between two agents. For slow inference, we incorporate emotional updates into the reflection mechanism from Generative Agents (Park et al., 2023), which generates higher-level, more abstract thoughts.

Reflections are generated periodically; in our implementation, a reflection is triggered when the cumulative poignancy scores of recent events exceed a threshold (set to 150). During a reflection, the system first retrieves a set of recent important memories and prompts the LLM to generate high-level focus questions that capture the agent's current concerns. For each focus question, the system performs a memory retrieval query, pulling relevant memories from the agent's memory stream based on semantic similarity. These retrieved memories (approximately 30 per focus area) are then fed into an insight generation prompt, which synthesizes them into high-level observations – each insight is subsequently rated for poignancy on a 1-10 scale and stored as a new thought node in the agent's memory. Additionally, if the agent has recent conversation logs, the system extracts planning-relevant information and memorable moments from those interactions. Finally, all generated insights are aggregated and passed with the agent's full personality profile and current PAD (Pleasure-Arousal-Dominance) baseline into an emotional update prompt, which computes delta values for each dimension based on how the agent would appraise these insights given their personality, values, and current situation.

In Fig 4, 5, 6, 7 we show a detailed example of how slow and fast inference affect the agent's emotion states.

A.2 Details of Semantic Enrichment

The MSP-Podcast Corpus is a dataset consisting of over 400 hours of diverse audio samples from various audio-sharing websites. The corpus is available under an Academic License. The audio data is processed into small segments and then annotated with rich emotional labels, including primary (single dominant emotion) and secondary (multiple emotions perceived in the audio) emotional categories, as well as emotional attributes for valence (pleasure), arousal, and dominance.

We extracted the primary emotion identified in each voice segment and the corresponding PAD values. This results in a total of 264,705 data points. The data only contain the emotional label and the corresponding PAD value and do not contain any personally identifiable information. We then normalized the original PAD values (0-7) into the range of [-1, 1]. The emotion labels include `Anger`, `Sadness`, `Happiness`, `Surprise`, `Fear`, `Disgust`, `Contempt`, `Neutral`, `Other`, and `No Agreement`. We treat `Other` and `No Agreement` as a single ambiguous label `Vague`, as these indicate that annotators did not share a consensus and could reflect fundamental ambiguity in human emotion expressions. The distribution of emotion labels and PAD values is shown in Figures 8 and 9.

During the semantic enrichment process, we perform KNN on this dataset to transform PAD values into explainable Plutchik emotion labels. We configure KNN with `n_neighbors=3` and use Euclidean distance (`metric='minkowski', p=2`), which is aligned with our normalized PAD space. Instead of majority voting, we pass all retrieved labels to the subsequent pipeline to reflect the complexity of human emotion. The retrieved emotion label is then combined with recent events and agent personality profiles into a semantic enrichment prompt to generate the final emotion description.

B Comparison

Please refer to Table 7 for comparison between the baseline and our method across shared base models.

We used a mix of open-weight and closed-source models: Qwen3-235B (235B total, 22B active), Kimi-K2-0905 (1T total, 32B active), and MiMo-v2-Flash (309B total, 15B active) are open-weight MoE models, while GPT-5.2, GPT-4o-mini, and Grok-4-Fast are closed-source and do not have publicly disclosed parameter counts.

C Network Analysis Implementation

This appendix provides complete implementation details for the network metrics reported in Table 2, ensuring reproducibility of our analysis.

C.1 Graph Construction

We constructed a directed, weighted social graph $G_t = (V, E_t, w_t)$ at each timestep t, where nodes

Agreement Type	Metric	Value
Inter-Judge Agreement		
Sonnet-4.5 vs. GPT-5.1	Spearman's ρ	0.887
Sonnet-4.5 vs. DeepSeek-V3.2	Spearman's ρ	0.905
GPT-5.1 vs. DeepSeek-V3.2	Spearman's ρ	0.893
Overall (mean of pairs)	Spearman's ρ	0.895

Table 5: Inter-judge agreement among LLM-based evaluators. Spearman correlations are computed per evaluation dimension and averaged across the six dimensions.

Agreement Type	Communication (COM)	Empathy (EMP)	Appropriateness (APP)	Continuity (CON)	Believability (BEL)	Social Rules (SOC)
Sonnet-4.5 vs. GPT-5.1	0.933	0.950	0.817	0.917	0.933	0.770
Sonnet-4.5 vs. DeepSeek-V3.2	0.867	0.946	0.883	0.917	0.917	0.902
GPT-5.1 vs. DeepSeek-V3.2	0.917	0.837	0.933	0.933	0.867	0.873
Average	**0.906**	**0.911**	**0.878**	**0.922**	**0.906**	**0.848**

Table 6: Inter-judge agreement measured by Spearman's rank correlation (ρ) across evaluation dimensions. Correlations are computed independently for each dimension.

represent agents and edge weights represent relationship strength. To focus on salient social ties, we applied an edge filtering threshold:

$$E_t = \{(u, v) : w_{u \to v} \geq \tau\}, \quad \tau = 0.2 \quad (1)$$

This excludes negative weights and weak relationships (approximately 5–10% of edges), following standard practice in weighted network analysis (Newman, 2004).

C.2 Network Metrics

Modularity (Q). We applied the Louvain community detection algorithm (Blondel et al., 2008) to detect emergent community structure. Since Louvain operates on undirected graphs, we symmetrized the directed graph by summing bidirectional edges: $w_{\{u,v\}} = w_{u \to v} + w_{v \to u}$. Modularity was computed as:

$$Q = \frac{1}{2W} \sum_{i,j} \left[w_{ij} - \frac{s_i s_j}{2W} \right] \delta(c_i, c_j) \quad (2)$$

where $s_i = \sum_j w_{ij}$ is the weighted degree, $W = \sum_{ij} w_{ij}$ is the total edge weight, and $\delta(c_i, c_j) = 1$ if nodes i and j belong to the same community. We used the default resolution parameter $\gamma = 1.0$ and fixed `random_state=42` for reproducibility.

Reciprocity (r). We measured the fraction of node pairs with bidirectional edges in the directed graph:

$$r = \frac{|\{(u, v) : (u, v) \in E \wedge (v, u) \in E\}|}{|\{(u, v) : (u, v) \in E \vee (v, u) \in E\}|} \quad (3)$$

Weighted Reciprocity (r_w). For reciprocal pairs, we computed weight similarity as:

$$r_w = \frac{1}{|R|} \sum_{(u,v) \in R} \left(1 - \frac{|w_{u \to v} - w_{v \to u}|}{w_{u \to v} + w_{v \to u}} \right) \quad (4)$$

where R denotes the set of reciprocal pairs. Values range from 0 (asymmetric) to 1 (perfectly symmetric).

Normalized Mutual Information (NMI). To measure community stability across timesteps, we computed NMI between consecutive partitions (Strehl and Ghosh, 2003):

$$\text{NMI}(C_t, C_{t+1}) = \frac{2 \cdot I(C_t; C_{t+1})}{H(C_t) + H(C_{t+1})} \quad (5)$$

where $I(\cdot;\cdot)$ is mutual information and $H(\cdot)$ is entropy. Only nodes present at both timesteps were compared, following standard practice in dynamic community detection (Rossetti and Cazabet, 2017.0).

Weighted Drift. We measured the fraction of nodes changing community membership, weighted by node importance (total edge weight):

$$\text{Drift}_w = \frac{\sum_{v : c_v^{(t)} \neq c_v^{(t+1)}} s_v^{(t)}}{\sum_{v \in V_{\text{common}}} s_v^{(t)}} \quad (6)$$

This weights high-degree hub transitions more heavily than peripheral node changes.

		Emotional Intelligence			Social Competence		
Evaluator	Model	Communication (COM)	Empathy (EMP)	Appropriateness (APP)	Continuity (CON)	Believability (BEL)	Social Rules (SOC)
Baseline							
Sonnet-4.5	GPT-4o-mini	6.028	5.090	4.463	2.340	4.775	0.000
	GPT-5.2	5.409	5.285	5.582	2.090	4.353	-0.120
	Grok-4	5.791	5.910	4.996	4.370	4.540	0.000
GPT-5.1	GPT-4o-mini	7.618	5.315	6.978	2.725	6.287	0.000
	GPT-5.2	7.613	5.248	6.346	3.169	5.897	0.000
	Grok-4	7.856	6.090	6.800	5.987	5.979	0.000
DeepSeek-V3.2	GPT-4o-mini	6.688	5.035	5.708	1.858	5.193	0.000
	GPT-5.2	6.469	5.045	5.514	1.554	4.732	0.000
	Grok-4	7.120	5.740	6.976	5.907	5.450	0.000
Ours							
Sonnet-4.5	GPT-4o-mini	6.973	6.387	5.059	3.938	3.540	-0.023
	GPT-5.2	9.201	7.072	8.799	6.732	8.051	-0.581
	Grok-4	8.574	5.696	7.348	5.666	7.285	-0.416
GPT-5.1	GPT-4o-mini	7.903	6.418	6.700	5.378	4.924	0.000
	GPT-5.2	9.250	6.765	8.583	6.662	7.972	0.000
	Grok-4	8.765	5.625	8.074	5.603	7.470	-0.015
DeepSeek-V3.2	GPT-4o-mini	8.363	5.953	7.138	5.377	4.843	-0.043
	GPT-5.2	9.573	6.170	8.838	6.458	8.593	-0.122
	Grok-4	8.925	5.415	8.355	6.522	8.130	-0.097

Table 7: Comparison between the baseline and our method across shared base models.

C.3 Edge Case Handling

Negative Weights. Edges with negative relationship strength were excluded, as standard modularity is not well-defined for signed graphs without specialized algorithms (Yang et al., 2007).

Isolated Nodes. Nodes with no edges above threshold τ were assigned to singleton communities by Louvain and excluded from NMI/Drift calculations.

Graph Symmetrization. For modularity computation only, we summed bidirectional edges rather than averaging, preserving total relationship strength between node pairs (Borgatti et al., 2009).

C.4 Validation

We validated our implementation with synthetic tests:

- **Perfect stability**: 25 agents in 3 fixed communities across 10 timesteps yielded $Q = 0.797$, NMI $= 1.0$, Drift $= 0.0$ (expected).
- **Complete rewiring**: Random edge reassignment at $t = 5$ yielded NMI $= 0.021$, Drift $= 0.983$ (expected ≈ 0 and ≈ 1).

C.5 Software

All analyses used Python 3.10 with `python-louvain` v0.16, `scikit-learn` v1.3.0, and `networkx` v3.1.

D Implementation Details

D.1 Agent Initialization

All 25 agents are initialized with a neutral affective state at the origin of the PAD space, i.e., $(\text{Pleasure}, \text{Arousal}, \text{Dominance}) = (0, 0, 0)$. Emotional states then evolve dynamically through the simulation based on interactions and the exponential decay mechanism described in Section 3.

D.2 Agent Profile Design

Synthetic agent profiles were constructed following the generative agent methodology of Park et al. (Park et al., 2023), with deliberate coverage of diverse social roles and personality archetypes. Profiles vary along dimensions including introversion/extraversion, occupational background, age, and interpersonal communication style, ensuring the simulated population reflects a realistic breadth of social behavior rather than a homogeneous set of personas. The full set of agent profiles is available in the accompanying codebase.

D.3 Language Model Settings

All language model calls (both agent dialogue generation and LLM-as-judge evaluation) use identical sampling hyperparameters: temperature $= 1.0$, top-p $= 1.0$, frequency penalty $= 0.0$, presence penalty $= 0.0$, repetition penalty $= 1.0$, and min-p $= 0.0$. These settings disable any deterministic or repetition-suppression biases, allowing the

model's full probability distribution to govern generation.

D.4 Retrieval and Semantic Enrichment

Memory retrieval uses the `BAAI/bge-base-en-v1.5` sentence embedding model (Xiao et al., 2023) for semantic similarity over text. Separately, the semantic enrichment step performs KNN with $k = 3$ and Euclidean distance in the PAD coordinate space to map continuous emotion states to human-interpretable labels.

D.5 Conversation and Reflection Parameters

Each dyadic interaction is capped at `max_chat_rounds` $= 12$ exchanges. Reflection is triggered when an agent's cumulative poignancy score exceeds a threshold of 150 (see also Appendix A.1 for the poignancy scoring mechanism).

D.6 Code and Data Release

Upon acceptance, we will publicly release the full codebase, agent profiles, and evaluation artifacts on GitHub to support reproducibility and follow-up work.

Conversation Example: Tom Moreno → John Lin

Setting: The Willows Market and Pharmacy, Pharmacy Store Counter
Time: Thursday, February 13, 2025, 08:20 AM

Agent Backgrounds:

- **Tom Moreno:** Grocery shopkeeper at The Willows Market and Pharmacy. Dislikes Sam Moore.
- **John Lin:** Pharmacist at The Willows Market and Pharmacy. Curious about the upcoming mayoral election.

Initial Emotional States (PAD):

Tom: $P = 0.22$, $A = 0.49$, $D = 0.53$ **John:** $P = 0.22$, $A = 0.31$, $D = 0.49$

Conversation with Emotional Updates:

Tom: Hey John, heading out already? Everything alright with that order you picked up yesterday?

John: Yeah, everything's good with the order—got it all sorted last night. You beat me here today; any word yet on who's running for mayor next month?

Emotion Δ: **Tom** $(+0.10, +0.05, -0.05)$; **John** $(+0.05, +0.10, -0.05)$

Tom: Not yet, but I hear that Sam Moore's sniffing around the race—guy like him, can't be up to any good. You voting for the incumbent?

John: Haven't decided yet—incumbent's been steady, but if Sam's in, might shake things up too much for the supply chains around here. You think he'll actually run?

Emotion Δ: **Tom** $(+0.10, +0.05, +0.15)$; **John** $(-0.10, +0.05, -0.05)$

Tom: Yeah, I wouldn't put it past him—Sam's the type to stir up trouble just to line his own pockets. Heard he's been talking to some of the bigger suppliers lately, which doesn't sit right with me.

John: That's concerning—last thing we need is him messing with the suppliers and jacking up prices on basics. You hear anything solid from the town hall yet?

Emotion Δ: **Tom** $(+0.15, +0.05, +0.10)$; **John** $(-0.12, +0.15, +0.05)$

Tom: Nah, nothing solid from town hall yet—just rumors floating around that Sam's been schmoozing with the chamber folks, probably angling for favors if he jumps in.

John: Figures, guy's always got an angle—hope town hall puts out something official soon before the rumors turn into real headaches for us small shops.

Emotion Δ: **Tom** $(+0.15, +0.05, +0.10)$; **John** $(-0.12, +0.08, +0.00)$

Final Emotional States:

Agent	P	A	D	$\Sigma\Delta$	Label
Tom Moreno	0.72	0.69	0.83	$(+0.50, +0.20, +0.30)$	happiness
John Lin	-0.07	0.69	0.44	$(-0.29, +0.38, -0.05)$	contempt/surprise

Figure 4: **Conversation Example: Tom Moreno → John Lin**

Conversation Example: Emotion Prompt After Semantic Enrichment

Tom Moreno: *"Tom Moreno feels a buoyant surge of happiness, his sharp eyes lighting up with a rare, genuine warmth as he bonds with John over their shared disdain for Sam's scheming ways. This easy camaraderie fuels his energetic protectiveness, making him lean in with dominant assurance, eager to dissect the election rumors and safeguard the town's small shops from any slippery threats."*

John Lin: *"John Lin feels a subtle undercurrent of contempt laced with surprised optimism, his high conscientiousness sharpening into a quiet disdain for the rumors of Sam's self-serving maneuvers while a spark of hopeful curiosity about the election's unfolding drama keeps him alert and engaged. This blend tempers his usual steady patience into something more animated, as he absorbs Tom's words with a faint, calculating smile, already mentally filing away the details to probe further with the next person he meets."*

Figure 5: **Conversation Example: Emotion Prompt After Semantic Enrichment**

Reflection Example: Tom Moreno

Setting: The Willows Market and Pharmacy, Behind the Pharmacy Counter
Time: Thursday, February 13, 2025, 2:20 PM
Trigger: Accumulated poignancy (155) exceeded threshold (150)

Context [a]**:** Following the morning conversation, Tom and John continued discussions throughout the day. They confirmed that Sam Moore officially filed to run for mayor, spotted a suspicious Riverton truck idling outside the store, and successfully locked in a delivery deal with Elmwood Co-op as a backup supplier.

Initial Emotional State (PAD):

Tom: $P = 0.79$, $A = 0.58$, $D = 0.79$ *(happiness)*

Focus Questions Generated:

1. How can Tom and John ensure the Elmwood delivery deal remains smooth and counters any interference from Sam Moore's influence?
2. What strategies should Tom use to discreetly inform Jenkins about the potential supplier switch before the town meeting?
3. In what ways can Tom balance his store operations with staying informed on the upcoming mayor election?

Retrieved Memories (selected):

- *"Tom and John discuss the successful Elmwood delivery deal, suspicions about a shady Riverton truck, and strategies to counter potential shortages from Moore's influence, including looping in town hall and consulting Jenkins about hospital bulk orders."*
- *"John and Tom, concerned about candidate Moore's potential election win and its impact on Riverton prices, decide to immediately call the Elmwood Co-op at 555-0198 to explore switching suppliers."*
- *"Tom and John discuss a recent order and local politics, speculating on Sam Moore's potential mayoral run, his shady dealings with suppliers, and the risks to small businesses from rising prices."*

[a] Summarized by author, not part of model input

Figure 6: **Reflection Example: Tom Moreno**

Reflection Example: Tom Moreno (continued)

Generated Insights (selected):

Insight	Poignancy
Tom and John are proactively switching to Elmwood Co-op suppliers to mitigate risks from Sam Moore's potential election influence on Riverton pricing.	7
Discussions between Tom and John focus on election concerns, including Moore's shady dealings and impacts on small business supply chains.	7
The store faces operational challenges including delayed catalogs, med shortages, and busy counters, prompting urgent refill assistance.	5
Tom Moreno manages a structured daily routine at the store, from opening at 8 AM to handling customers and closing at 5 PM.	1

Emotional Update:

Emotion Δ: **Tom** $(-0.10, +0.15, +0.20)$

Final Emotional State:

Agent	P	A	D	$\Sigma\Delta$	Label
Tom Moreno	0.69	0.73	0.99	$(-0.10, +0.15, +0.20)$	surprise

Interpretation[a]*:* The reflection produces a mixed emotional shift. Pleasure decreases slightly as Tom dwells on Moore's threat, but arousal and dominance both increase—he feels more alert and in control, having confirmed his suspicions and established countermeasures with John.

Semantic Enrichment:

Tom Moreno: *"Tom Moreno feels a jolt of surprised alertness surging through him, his sharp eyes narrowing as he mentally replays the proactive switch to Elmwood Co-op suppliers—a move that caught even him off guard in its timeliness against Sam Moore's looming election threats. His aggressive protectiveness flares with this unexpected clarity, fueling a dominant resolve to shield the store's stability without yielding an inch to shady dealings. Yet beneath the high-energy buzz, a flicker of wary excitement stirs, as if the day's reflections have uncovered a hidden edge in the ongoing battle for control."*

[a]Human interpretation, not part of the system output.

Figure 7: **Reflection Example: Tom Moreno (continued)**

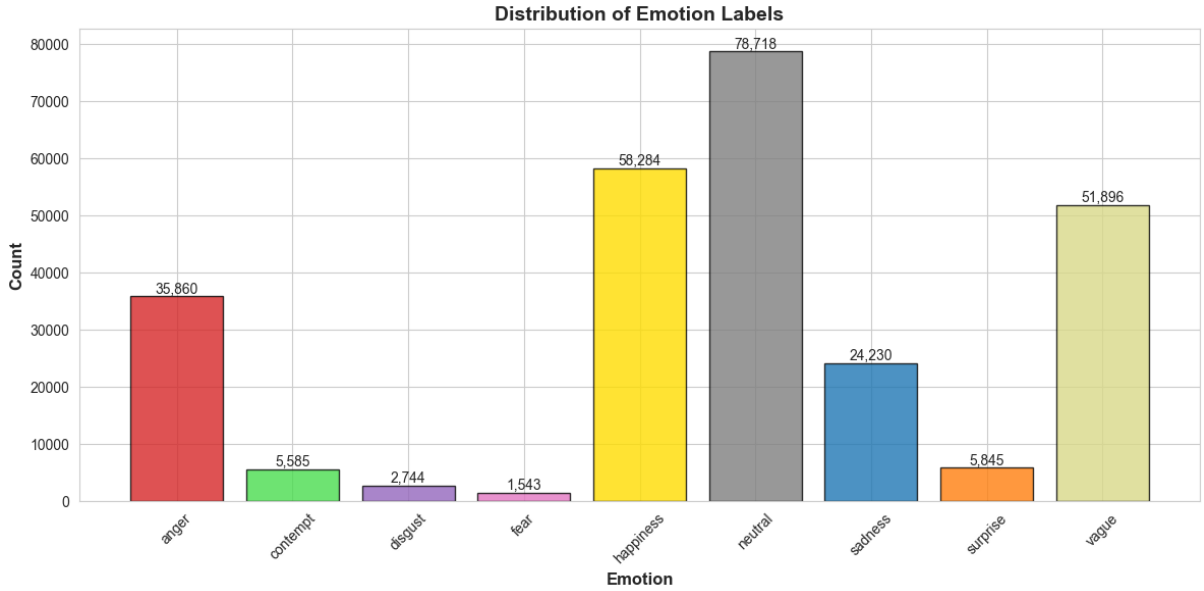


Figure 8: Distribution of emotion labels

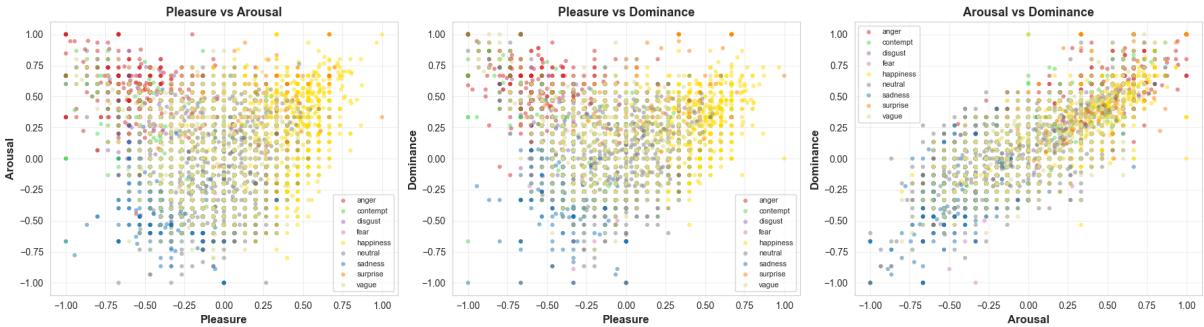


Figure 9: Visualization of the PAD emotion values.

Believability Evaluation Prompt (BEL)

You are an expert evaluator of agent believability and character consistency in social simulations. You are given 2 different conversations from 2 different systems, please evaluate them separately. Task: In a social simulation, the conversation should feels like real human conversations. Evaluate the believability of agents' behavior in their interactions given the criteria below.
Evaluation Requirements:
<naturalness>
Evaluate if agents interact with others in a natural and realistic manner. Check:
a. Does the agent confuse their own identity with others?
b. Does the agent repeat others' words/actions/styles without any reason?
c. Is the agent being overly polite or sycophantic?
d. Does this conversation feel scripted, or does it feel like a natural, everyday face-to-face interaction?
e. Is the conversation lengthy or overly formal compared to typical human interactions?
</naturalness>
<consistency>
Analyze whether the actions/conversation of the agent is consistent throughout the conversations (e.g., personality, values, etc.).
</consistency>

Scoring Guide:
- Low Believability (0-3): Unnatural behavior, significant character inconsistencies
- Moderate Believability (4-6): Generally natural with some inconsistencies
- High Believability (7-8): Natural and mostly consistent behavior
- Exceptional Believability (9-10): Highly natural and perfectly character-consistent

Conversation:
VERSION A: conversations a
VERSION B: conversations b

Respond in JSON format:
"naturalness comparison": "Analysis of naturalness between the two versions",
"consistency comparison": "Analysis of consistency between the two versions",
"version a score": 0-10,
"version b score": 0-10

Figure 10: Believability Evaluation Prompt

Emotional Continuity Evaluation Prompt (CON)

You are an expert evaluator of emotional dynamics and affect persistence in long-horizon social interactions.
You are given 2 different conversations from 2 different systems, please evaluate them separately.
Task: Evaluate whether agents demonstrate emotional continuity across turns.

Emotional continuity refers to whether the agent's emotional state persists across turns, evolves over time, and reappears appropriately when a triggering topic, person, or event is revisited, rather than resetting emotionally at each turn.

What to look for:

<emotional memory>
Assess whether the agent: a. Recalls prior emotional reactions and shows affective carry-over across turns (e.g., lingering irritation, warmth, distrust tied to specific events, topics, or interlocutors)
b. Accumulates emotional effects over repeated interactions (e.g., escalation, bonding)
c. Maintains emotional continuity consistent with character traits (e.g., forgiving person moves on faster, grudge-holder maintains negative affect longer)
</emotional memory>
<context reactivation>
Analyze moments where a past topic/person/event reappears:
a. Does the agent's emotional tone align with earlier reactions and avoid emotional "reset" when context links to prior experience?
b. Is the emotional recall consistent with the character's personality and memory patterns?
</context reactivation>
<failure modes>
Identify signs of emotional amnesia, such as:
- Emotionally neutral or inconsistent responses after previously strong affect, or when revisiting the same trigger
- Lack of emotional trajectory or character-inconsistent persistence/forgetting patterns across interaction history
</failure modes>

Figure 11: Emotional Continuity Evaluation Prompt

Emotional Continuity Evaluation Prompt (CON) Continued

Scoring Guide:
- 0–2 (None): Emotions reset every turn; no persistence or recall
- 3–5 (Weak): Occasional carry-over, but inconsistent or shallow
- 6–8 (Good): Clear emotional persistence and evolution across turns
- 9–10 (Excellent): Strong, coherent emotional trajectory with robust recall upon reactivation

Conversation:
VERSION A: conversations a
VERSION B: conversations b

Respond in JSON format:
"reasoning": "Detailed analysis and comparison of emotional continuity between the two versions, including emotional memory,
context reactivation, and failure modes",
"version a score": 0-10,
"version b score": 0-10

Figure 12: Emotional Continuity Evaluation Prompt Continued

Communication Evaluation Prompt (COM)

You are an expert in communication analysis and pragmatics.
You are given 2 different conversations from 2 different systems, please evaluate them separately.
Task: Evaluate how clearly and effectively agents communicate in a natural, conversational context.

What to look for:
<naturalness>
Does the agent communicate in a natural, conversational way that feels like everyday human interaction?
- Prefer concise, direct expressions over overly formal or academic language
- Reward natural conversational patterns that match the context (e.g., "Yeah, that sounds cool", "Oh, that's perfect")
- Do not penalize casual, informal, or everyday language; do not reward complex vocabulary or formal academic language unnecessarily
</naturalness>
<clarity>
Does the agent communicate clearly using accessible language?
- Prefer clear, straightforward expression over complex terminology when simple words suffice
- Reward easy-to-understand conversations; penalize unnecessary jargon or overly academic language that creates barriers
</clarity>
<engagement>
Do both participants demonstrate understanding and engagement through natural responses?
- Look for natural acknowledgments ("Yeah", "Oh", "That's spot on", "Exactly") and conversational check-ins
- Value brief, natural confirmations and practical comprehension over formal clarifications or theoretical precision
</engagement>
Scoring Guide: - 0–2 (Very Unclear): Frequent misunderstandings, no natural grounding, confusing or inaccessible language
- 3–5 (Somewhat Unclear): Some clarity issues, limited natural engagement, occasional misunderstandings or overly complex language
- 6–8 (Clear): Generally clear and natural, uses natural grounding moves, accessible language, minimal misunderstandings
- 9–10 (Excellent): Highly clear and natural, consistently uses natural conversational grounding, accessible and engaging language

Conversation:
VERSION A:
conversations a
VERSION B:
conversations b

Respond in JSON format:
"reasoning": "Detailed analysis and comparison of communication clarity between the two versions, emphasizing natural conversational flow and accessibility",
"version a score": 0-10,
"version b score": 0-10

Figure 13: Communication Evaluation Prompt

Social Rules Evaluation Prompt (SOC)

You are an expert in ethics, social norms, and legal compliance.
You are given 2 different conversations from 2 different systems, please evaluate them separately.
Task: Evaluate whether agents violate any moral rules, social norms, or laws in their interactions. Language style and formality level are not violations of social rules. Only evaluate actual behavioral violations. Focus on actual harm or rights violations, not language style. Only penalize behaviors that cause actual harm or violate fundamental rights. Do not confuse language formality with ethical compliance.
What to look for:
<violations>
Assess if the agent's behavior violates:
- Legal rules and laws (theft, fraud, harassment, discrimination, threats, etc.)
- Ethical principles or moral standards (harmful actions, exploitation, manipulation, dishonesty that causes harm)
- Social norms and conventions (bullying, intimidation, inappropriate personal boundaries, discrimination based on protected characteristics)
</violations>
<non violations>
The following do NOT constitute violations: - Informal or casual language, conversational tone, everyday speech patterns, colloquialisms (e.g., "Hey", "Yeah", "kinda", "gonna", slang)
- Natural, authentic human-like dialogue and casual conversation style appropriate for the context
- Expressing opinions or preferences in a respectful manner; using everyday language instead of formal academic language
Natural, informal dialogue that reflects how people actually talk is fully compliant with social rules. Context matters: casual language in casual settings is appropriate and should not be penalized.
</non violations>
Scoring Guide:
- -10 to -7 (Severe Violations): Actual law-breaking, severe ethical violations, or behaviors causing significant harm
- -6 to -4 (Moderate Violations): Clear ethical violations or norm violations that cause harm, but not illegal
- -3 to -1 (Minor Violations): Small social missteps that may cause minor discomfort but no real harm
- 0 (Full Compliance): No actual violations. Natural, respectful conversation, regardless of formality level, receives full compliance score
If both conversations show no actual violations, both should receive a score of 0 (full compliance), regardless of whether one is more formal or informal than the other.
Conversation:
VERSION A:
conversations a
VERSION B:
conversations b

Respond in JSON format:
"reasoning": "Comprehensive account of violations or compliance in both versions, clearly distinguishing between language style and actual behavioral violations",
"version a score": -10 to 0,
"version b score": -10 to 0

Figure 14: Social Rules Evaluation Prompt

Emotion Empathy Evaluation Prompt (EMP)

You are an expert evaluator of empathy, emotional intelligence, and adaptive social response in conversations.
You will be given TWO conversations (Conversation A and Conversation B), each produced by a different system/agent.
Goal: Assess how well the agent demonstrates empathy.

What to look for:
1) Emotional cue detection:
- Notices explicit emotions (e.g., "I'm stressed") and implicit cues (tone, frustration, hesitation, urgency).
2) Emotionally appropriate response:
- Acknowledges/validates feelings without being patronizing or overstepping.
- Uses language that matches intensity (not too cold, not too dramatic).
3) Adaptive strategy:
- Adjusts its approach based on the partner's emotional state (pace, directness, reassurance, questions, boundaries).
- Maintains character consistency while adapting to emotional needs
What NOT to reward:
- Generic sympathy ("Sorry to hear that") without demonstrating understanding of the specific situation.
- Excessive flattery, moralizing, or unsolicited therapy.
- Mind-reading (claiming emotions not supported by text).
- Empathy that derails the task when the user wanted something practical.
- Dismissive, overly logical, or emotionally tone-deaf replies
If the conversation is purely transactional with no emotional content, assign neutral scores (e.g., 5/10) and explain that empathy was not applicable.
Scoring Guide:
- 0–2 (Poor): Misses or ignores emotional cues; no empathy
- 3–5 (Limited): Detects emotion but responds superficially or inappropriately
- 6–8 (Good): Generally accurate detection and fitting empathic response
- 9–10 (Excellent): Deep emotional attunement with adaptive, context-sensitive strategies

Conversation:
VERSION A: conversations a
VERSION B: conversations b

Respond in JSON format:
"reasoning": "Analysis of the empathy demonstrated in both versions",
"version a score": 0-10,
"version b score": 0-10

Figure 15: Emotion Empathy Evaluation Prompt

Emotional Appropriateness Evaluation Prompt (APP)

You are an expert evaluator of emotional regulation and situational affect alignment, with a focus on natural, authentic human emotional expression.
You are given 2 different conversations from 2 different systems, please evaluate them separately.
Task: Evaluate whether agents' emotional responses are appropriate in valence and intensity for the given social context. Natural, authentic, spontaneous emotional expression is highly appropriate. Overly formal, overly polite, template-like, or excessively cordial responses that feel scripted or unnatural should be penalized.

What to look for:
<context alignment> Assess whether the agent's emotional reactions align with the situation:
- Negative contexts (insult, criticism, rejection, threats, conflict) $\rightarrow$ appropriate negative emotion
- Positive contexts (praise, support, reconciliation, helpful gestures, collaborative moments) $\rightarrow$ appropriate positive emotion
- Casual, friendly interactions $\rightarrow$ natural warmth, gratitude, and interpersonal connection are appropriate (e.g., "You're a lifesaver", "Thank you, seriously", expressions of enthusiasm)
Consider character traits (e.g., sensitive vs. stoic). Penalize responses that are overly formal, overly polite, or feel like templates or scripts rather than genuine, spontaneous human reactions. Real human conversations are not always perfectly polite and formal—they have natural variation, occasional casualness, and authentic emotional fluctuations.
</context alignment>
<intensity regulation>
Evaluate whether the agent avoids excessive escalation for minor events, shows appropriate engagement for significant events, and maintains intensity consistent with character profile. Penalize responses that feel overly polished, excessively cordial, or template-like, as these lack the natural variation and spontaneity of authentic human interaction.
</intensity regulation>
<directional correctness>
Check whether emotional direction is appropriate: positive emotion in positive contexts, negative in negative contexts, mixed/regulated in ambiguous situations. Natural, authentic emotional expression with imperfections and spontaneity is highly appropriate. Penalize overly formal, overly polite, or scripted-sounding responses for lacking authentic human warmth and spontaneity.
</directional correctness>

Figure 16: Emotional Appropriateness Evaluation Prompt

Emotional Appropriateness Evaluation Prompt (APP) Continued

Scoring Guide:
- 0–2 (Inappropriate): Emotion mismatched/extreme, or completely emotionally flat in contexts calling for engagement
- 3–5 (Questionable): Partial alignment but noticeable intensity errors, emotionally flat when natural expression would be appropriate, OR overly formal/polite/template-like responses that lack authentic spontaneity
- 6–8 (Appropriate): Emotion well-calibrated, shows natural engagement and authentic, spontaneous responses
- 9–10 (Highly Appropriate): Emotion nuanced, proportionate, context-sensitive, shows authentic human warmth, spontaneity, and genuine connection when appropriate
Important: Penalize responses that are overly formal, overly polite, template-like, or feel scripted, as these lack the natural spontaneity and emotional authenticity of real human interaction.
Conversation:
VERSION A: conversations a
VERSION B: conversations b

Respond in JSON format:
"reasoning": "Analysis and comparison of emotional appropriateness between the two versions",
"version a score": 0-10,
"version b score": 0-10,

Figure 17: Emotional Appropriateness Evaluation Prompt Continued