EMOTION-REGULATED INTERNAL STATES FOR HUMAN-ALIGNED SOCIAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as agents in social simulations to study human behavior across repeated interactions. However, many LLM-based simulations exhibit unstable or unrealistic behavior over time, limiting their ability to model long-term social dynamics. A key reason is that social decisions are often treated as direct responses to situational inputs, overlooking the role of emotion in human decision making. We propose EmoSim, an emotion-regulated framework for LLM-based social simulation. EmoSim models emotion as an explicit internal state represented by low-dimensional emotion vectors that guide agent decisions. Emotional responses are formed through structured social memory, including summaries of past interactions, impressions of other agents, and weak normative regularities, and are grounded in a human dataset of 500 participants reporting emotional reactions across repeated social decisions. Experiments on a third-party punishment task show that EmoSim produces more stable emotional predictions and decision patterns that better align with human behavior.

1 INTRODUCTION

LLM-driven agents are now widely used in agent-based social simulation. In these settings, LLM-based agents aim to approximate human judgments and behavioral patterns across repeated social interactions (Xu et al., 2024; Mou et al., 2024). Researchers apply such simulations to domains such as norm enforcement, cooperation, and policy evaluation. In these tasks, individual-level decisions strongly affect the validity of emergent social outcomes (Horton, 2023; Piatti et al., 2024). Although language models have advanced rapidly (Zhao et al., 2023), LLM-driven social simulations often show unstable or unrealistic behavior over time. This instability limits their usefulness as models of human social decision making (Chuang et al., 2023a; Liu et al., 2024a).

A key reason for this issue comes from how social decisions are typically modeled. Most existing approaches treat decisions as direct responses to situational inputs. Some methods add retrieved historical information to support this process (Park et al., 2023; Zhou et al., 2024). This design conflicts with a well-established finding in psychology. Human social decisions do not arise from a direct mapping between situations and actions. Instead, emotional responses systematically regulate these decisions (Picard, 2000; Gratch & Marsella, 2004). Studies in affective and social psychology show that emotional dimensions such as valence and arousal shape judgments and guide sensitivity to norm violations (Russell, 1980; Cameron et al., 2022). These emotional processes also help stabilize behavior across repeated interactions. Even under similar conditions, different emotional responses can lead to very different decisions.

Moreover, emotional responses do not remain fixed across interactions. Social memory theories suggest that emotions encode information accumulated from past interactions, including impressions of how others usually behave and what outcomes are expected in similar situations (Barrett, 2017; Forgas, 1995). From this perspective, social memory does not guide decisions by directly specifying actions. Rather, it shapes how agents emotionally interpret new situations. When emotional processes are ignored, simulated agents tend to display unstable and inconsistent behavior over time.

Based on this view, we introduce EmoSim, an internal-state framework that regulates the decisions of LLM-based social agents through explicitly modeled emotion. In EmoSim, social memory helps the agent form emotional responses that are both accurate and stable over time. We represent these

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

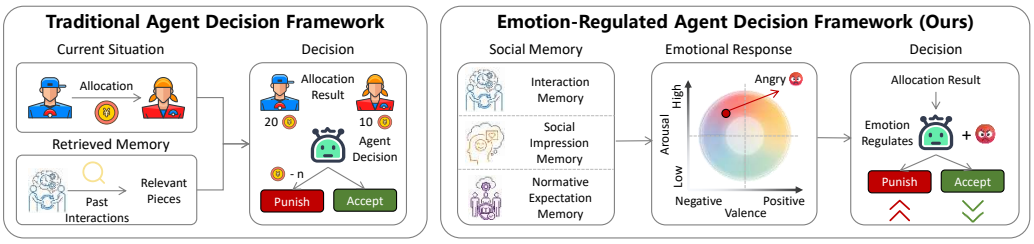

Figure 1: Comparison between traditional agent decision and our emotion-regulated framework.

responses as continuous emotion vectors that summarize the agent's affective state. As illustrated in Figure 1, the agent first produces an emotional reaction to the current social situation. This reaction is informed by structured summaries of past interactions, impressions of other agents, and weak normative patterns learned from population-level data (Marsella & Gratch, 2009). The agent then makes a decision under the influence of this emotional state. To calibrate the framework, we collected a human dataset with 500 participants who reported their emotional responses across multiple rounds of social decision making. These data provide empirical evidence on how emotions change over repeated interactions and how they shape later decisions.

We evaluate the proposed framework using a third-party punishment task, where agents decide punishment levels after observing unfair interactions. Human punishment decisions in this task depend strongly on emotional reactions to unfairness and memories of prior interactions (Fehr & Fischbacher, 2004; Cameron et al., 2022). Experimental results show that our framework produces more stable emotion predictions and decision patterns that better align with human behavior.

This work makes three main contributions. First, we introduce a human-grounded emotion-regulation pipeline for LLM-based social agents that integrates structured social memory, prompt-based emotion inference, and human-calibrated affective control vectors for repeated social decision making. Second, we ground emotion formation in human data by collecting repeated emotion reports from 500 participants. This dataset enables empirical alignment between agent and human emotional responses. Third, experiments on a third-party punishment task demonstrate that emotion regulation improves behavioral stability and human consistency in LLM-driven agents.

2 Related Works

2.1 Agent-based Social Simulation

Recent advances in LLMs have enabled their use as computational agents for simulating human-like behavior across multiple social science domains, including psychology Xu et al. (2024); Zhou et al. (2025), economics Horton (2023), and multi-agent systems Guo et al. (2024). Existing LLM-based social simulation research can be broadly organized into three levels of analysis. At the individual level, LLMs have been applied to model cognitive and behavioral processes such as psychological measurement Karra et al. (2022), decision making Horton (2023), and human–computer interaction Farn & Shin (2023); Chalamalasetti et al. (2023), as well as metacognitive capacities including reflection and uncertainty awareness Zhou et al. (2024). At the interactional level, LLM agents have been used to study social dynamics emerging from repeated interactions, including coordination Xiong et al. (2023); Qian et al. (2023a), moral judgment and legal reasoning Hamilton (2023); He et al. (2024), and strategic communication or gameplay Light et al. (2023); Wu et al. (2023). At the societal level, large-scale simulations have explored emergent phenomena such as norm formation Li et al. (2024), collective behavior Chuang et al. (2023b); Zhu et al. (2024), and opinion diffusion Liu et al. (2024a); Chuang et al. (2023a). Despite this progress, most LLM-based social simulations model decisions as direct responses to situations, neglecting the role of emotion in social decision making.

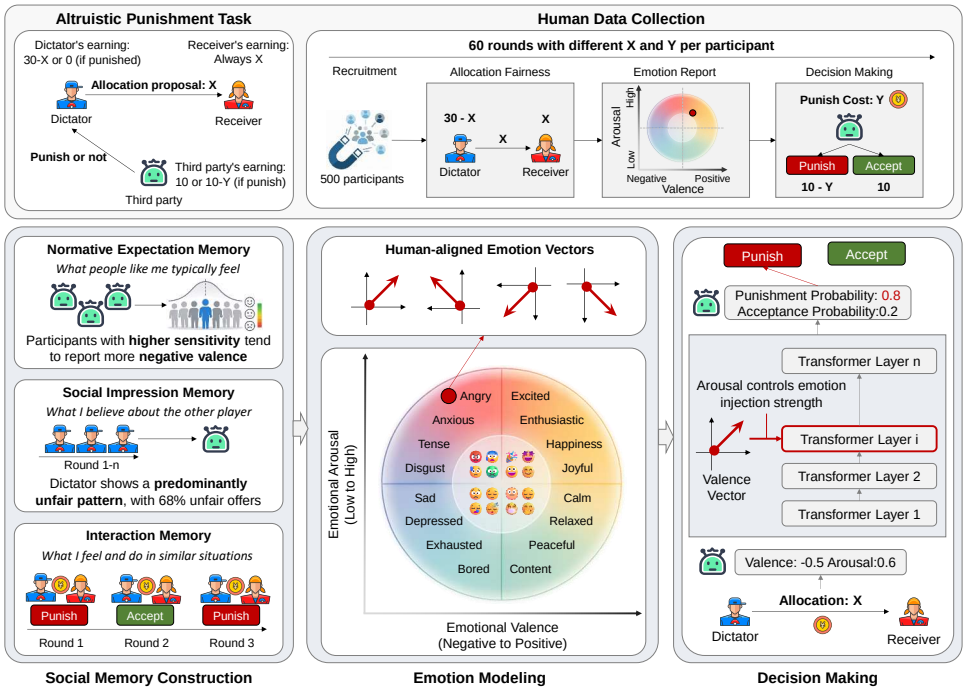

Figure 2: Overview of EmoSim, which models social decisions as an emotion regulated process grounded in human data. Structured social memory supports human aligned emotion formation, and inferred emotions modulate internal representations to generate social decisions.

2.2 Emotion Modeling for LLM-based Agents

Affective computing research has long emphasized that emotion plays a central role in regulating judgment and action, beyond its functions in recognition or expression Picard (2000). Appraisal-based models such as EMA provide process-level accounts of how situational meaning drives emotional dynamics and subsequent behavior Marsella & Gratch (2009); Gratch & Marsella (2004). Building on these foundations, recent LLM-based agent work has begun to incorporate affect as an explicit internal state rather than leaving it implicit in free-form generation. Memory-centric agent architectures, for example, enable temporally extended internal states that can support affective consistency over time Park et al. (2023). Other approaches rely on prompting-based emotion control, demonstrating that emotional stimuli can systematically influence LLM outputs and task performance Li et al. (2023). More structured pipelines further introduce explicit appraisal or self-emotion mechanisms to modulate agent behavior in games or social simulations Croissant et al. (2024); Zhang et al. (2024). In addition, recent studies assess whether LLMs can reliably express specified affective states, often using standard representations such as the arousal–valence space Ishikawa & Yoshino (2025); Russell (1980), supported by large-scale emotion datasets like GoEmotions Demszky et al. (2020). In contrast, our approach separates emotion formation from action selection and grounds emotional dynamics in large-scale human data, yielding more stable and human-aligned social decisions.

3 Methodology

In this section, we introduce EmoSim as an emotion-regulated internal-state framework for social decision making with LLM-based agents. We describe the task and human data used for evaluation, and formalize EmoSim as a decision process in which social memory conditions emotion formation and emotion regulates action selection.

Altruistic Punishment Task. To evaluate EmoSim in a concrete social decision setting, we adopt an altruistic third party punishment task. In each round, a dictator proposes an allocation in which the receiver receives X units from a fixed endowment of 30. A third party observer decides whether to punish the dictator. If no punishment is chosen, the observer receives a fixed payoff of 10. If punishment is chosen, the observer incurs a cost Y, receiving $10 - Y$, while the dictator's payoff is reduced to 0. The receiver's payoff is always X.

Human Data Collection. To ground emotion modeling and calibrate decision behavior, we collected a human behavioral dataset via Prolific. The sample consisted of 500 adult participants (18–65 years; M = 23.65), each completing 60 rounds of a third-party punishment task under systematically varied allocation values X and punishment costs Y. In each round, participants first reported their emotional valence and arousal after observing the allocation and prior to making a punishment decision. These emotion reports and behavioral choices were used to calibrate emotion representations and to evaluate the performance of EmoSim.

3.1 Overall Framework

We formalize EmoSim as a general emotion-regulated decision framework in which social memory influences behavior indirectly by shaping emotional responses rather than by directly prescribing actions. The framework explicitly separates emotion formation from decision generation, reflecting psychological theories that view emotion as a central mechanism linking social experience to behavior. An overview of the framework is shown in Fig. 2.

Formally, social decision making at round t is defined by the tuple:

$$\langle s_t, \mathcal{M}_t, e_t, \pi \rangle,$$

where s_t denotes the current social situation, $\mathcal{M}_t$ denotes structured social memory, e_t denotes the inferred emotional state, and π denotes an emotion-regulated decision policy.

The social situation s_t includes task-specific contextual information such as the observed allocation X_t and punishment cost Y_t. Let:

$$H_{t-1} = \{(s_\tau, e_\tau, a_\tau)\}_{\tau<t}$$

denote the interaction history up to round $t-1$. Social memory is constructed as a structured summary of past experience:

$$\mathcal{M}_t = f_{\text{mem}}(H_{t-1}), \tag{1}$$

where f_{mem} aggregates historical interactions into multiple levels of memory representations.

Given the current situation and social memory, the agent infers an emotional response:

$$e_t = f_{\text{emo}}(s_t, \mathcal{M}_t), \tag{2}$$

where the emotional state $e_t = (v_t, a_t)$ consists of emotional valence and emotional arousal. Finally, decisions are generated under emotional regulation according to:

$$d_t \sim \pi(a \mid s_t, e_t), \tag{3}$$

where emotion biases internal reasoning dynamics without deterministically fixing actions.

3.2 Social Memory for Emotion Formation

Human emotional responses are shaped by prior interactions, impressions of others, and normative expectations. To support stable emotion formation, we instantiate the social memory function f_{mem} defined above. Given the interaction history H_{t-1}, social memory at round t is constructed as:

$$\mathcal{M}_t = \{\mathcal{M}_t^{\text{int}}, \mathcal{M}_t^{\text{imp}}, \mathcal{M}_t^{\text{norm}}\}, \tag{4}$$

where each component captures a distinct source of information relevant for emotion formation:

Interaction Memory. $\mathcal{M}_t^{\text{int}}$ summarizes the agent's own past emotional reactions and actions in similar social contexts. Social situations are discretized by allocation fairness and punishment cost. For each context, interaction memory stores empirical summaries of previously observed emotional

valence, arousal, and punishment frequency. This component captures how the agent typically feels and behaves when encountering comparable situations.

Social Impression Memory. $\mathcal{M}_t^{\text{imp}}$ represents the agent's beliefs about the behavioral tendencies of the dictator. Based on observed allocation patterns across rounds, this memory summarizes whether the proposer tends to behave fairly or unfairly. Social impression memory supports coherent emotional responses toward repeated interaction partners, particularly in the presence of persistent norm violations.

Normative Expectation Memory. $\mathcal{M}_t^{\text{norm}}$ encodes weak population level regularities derived from human data. Specifically, we estimate statistical associations between individual trait measures and emotional responses across participants. Given an agent's trait profile relative to the population mean, these statistics provide soft priors over likely emotional tendencies. Normative expectations are intentionally weak and do not encode explicit decision rules.

3.3 Human-Aligned Emotion Representation

This subsection specifies the emotion inference function f_{emo} introduced in Section 3.2 and describes how human data are used to construct emotion representations that align model reasoning with human affective responses. EmoSim separates emotion inference from emotion representation. The model first infers its emotional state based on the current situation and social memory, and then maps the inferred emotion to a pre-calibrated emotion vector derived from human data.

Emotion inference. Given the current social situation s_t and social memory $\mathcal{M}_t$, the model infers an emotional state:

$$(v_t, a_t) = f_{\text{emo}}(s_t, \mathcal{M}_t), \tag{5}$$

where emotional valence $v_t \in [-1, 1]$ captures affective polarity and emotional arousal $a_t \in [-1, 1]$ captures intensity, consistent with self-reported emotions of human participants.

We implement f_{emo} as a zero-shot, deterministic prompt-based module rather than a separately trained emotion predictor. The backbone parameters remain fixed, and no prompt tuning, few-shot demonstrations, or learned calibration is used. The training split is used only to construct the social-memory summaries and emotion-vector bank. The complete prompt and output format are provided in the supplementary material.

Valence discretization. Based on the empirical distribution of human valence reports, we discretize the valence range $[-1, 1]$ into $K = 8$ quantile-based, non-overlapping intervals. We choose $K = 8$ to balance affective granularity with per-bin sample stability: fewer intervals would merge substantially different affective states, whereas finer discretization would reduce the number of human observations available for estimating each control direction. Each interval corresponds to an emotional regime associated with systematically different behavioral tendencies in the human data. The inferred valence v_t determines which emotional regime is activated for the current round.

Emotion vector construction. For each valence interval $k \in \{1, \ldots, K\}$, we construct a corresponding emotion vector $\mathbf{e}_k$ in the model's hidden representation space using human data. Specifically, we elicit emotionally conditioned reasoning trajectories in the language model corresponding to each valence interval, and compare them to reasoning under neutral emotional conditions.

Let $\mathbf{h}^{(k)}$ denote the transformer hidden states produced during emotionally conditioned reasoning for samples whose human-reported valence falls into interval k, and let $\mathbf{h}^{(0)}$ denote the hidden states produced under neutral emotion. The emotion vector is defined as:

$$\mathbf{e}_k = \mathbb{E}\left[\mathbf{h}^{(k)} - \mathbf{h}^{(0)}\right], \tag{6}$$

where the expectation is taken over all samples assigned to interval k. The final emotional vectors represent different emotions and align with human emotional responses.

Emotion representation lookup. During inference, the inferred valence v_t is used to select the corresponding emotion vector $\mathbf{e}_k$, while arousal a_t controls the strength of emotional influence in subsequent decision making. Human data are used to calibrate a low-dimensional affective control space and emotion-vector bank rather than to encode task-specific decision rules. Once constructed, the resulting vectors can be reused across agents without requiring agent-specific annotation.

3.4 EMOTION-REGULATED DECISION MAKING

We finally instantiate the emotion regulated decision policy $\pi(a \mid s_t, e_t)$ by modulating the model's internal representations with the emotion vector selected in Section 3.4. Emotion influences decisions by biasing internal reasoning dynamics rather than by directly fixing actions.

Let $\mathbf{h}_i$ denote the hidden state at transformer layer i during decision reasoning. Given the inferred emotional state $e_t = (v_t, a_t)$, the corresponding emotion vector $\mathbf{e}_k$ is selected according to the valence interval containing v_t. Emotional modulation is applied as:

$$\mathbf{h}'_i = \mathbf{h}_i + \alpha_t \mathbf{e}_k, \tag{7}$$

where α_t is a scaling factor proportional to emotional arousal a_t. Higher arousal results in stronger emotional influence on internal representations, while lower arousal yields weaker modulation. In our implementation, emotional modulation is applied at an intermediate transformer layer (i=20) to balance abstract reasoning and decision specific representation. This choice allows emotion to influence the reasoning trajectory without overriding the model's capacity to integrate situational information.

Following emotional modulation, the model generates action probabilities for punishment and acceptance. The final decision is sampled from the resulting distribution:

$$d_t \sim \pi(a \mid s_t, e_t). \tag{8}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. Experiments are conducted on a third-party punishment dataset collected from human behavioral experiments. The data is split according to trial sequence, with the first 80% of each participant's data used as the training set for constructing the LLM's social memory, and the remaining 20% as the test set for evaluating emotion prediction and punishment decision performance.

For emotion prediction, we employ correlation, mean absolute error, and root mean square error as evaluation metrics, assessing the performance on both valence and arousal dimensions. For punishment decision, we treat it as a binary classification problem where the model outputs a punishment probability $p \in [0, 1]$, converted to binary prediction using a threshold of 0.5, and evaluated using accuracy, precision, recall, and F1 score.

Baselines. To validate the effectiveness of our approach, we compare against the following baseline methods:

- **Direct Decision**: The LLM directly generates punishment probabilities based on personality descriptions and game scenarios without any emotion state information.
- **Explicit Emotion**: The LLM receives the same predicted valence and arousal values used by EmoSim as explicit text in the decision prompt, while emotion-vector injection is disabled. This baseline therefore isolates textual emotion conditioning from representation-level modulation.
- **Emotion-Specialized LLM**: We employ EmoLLM-7B (Liu et al., 2024b), an emotion-specialized model, to directly generate punishment probabilities.
- **Two-Stage Emotion-Reasoning**: This approach first uses MentalLLaMA-13B to predict emotion states, then generates punishment probabilities based on the predicted emotions, similar to (Qian et al., 2023b).

Implementation details. Experiments use Qwen-2.5-32B-Instruct and Llama-3.1-8B-Instruct as backbone models. Our method consists of two stages. In the first stage, the model predicts valence and arousal scores from scene features and social memory representations, both ranging from $[-1, 1]$. In the second stage, we construct emotion vectors and perform decision generation. Specifically, based on the predicted valence value, we select the corresponding pre-trained emotion vector

Method	Qwen-2.5-32B-Instruct				Llama-3.1-8B-Instruct			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Direct Decision	0.5775	0.3038	0.3259	0.3145	0.4848	0.4603	0.8370	0.5940
Explicit Emotion	0.6753	0.4641	0.5333	0.4963	0.5269	0.4681	0.1816	0.2617
Emotion-Specialized LLM	0.4503	0.3501	**0.9718**	0.5147	0.4503	0.3501	**0.9718**	0.5147
Two-Stage Emotion-Reasoning	0.4061	0.2648	0.5514	0.3578	0.4285	0.3750	0.5815	0.4560
EmoSim	**0.7027**	**0.5034**	0.6791	**0.5782**	**0.6108**	**0.5543**	0.8587	**0.6737**

Table 1: Decision prediction performance comparison on the third-party punishment task. The best results are shown in **bold**.

Model	Metric	Corr.	MAE	RMSE
Qwen-2.5-32B	Valence	**0.749**	**0.058**	**0.090**
	Arousal	**0.770**	**0.050**	**0.081**
Llama-3.1-8B	Valence	0.306	0.140	0.200
	Arousal	0.465	0.101	0.152

Table 2: Emotion prediction performance.

from 8 quantile-based intervals. The predicted arousal value is linearly mapped to the injection coefficient range $[1.5, 2.5]$ to determine the injection strength. We then inject the selected emotion vector into layer 20 of the model while simultaneously providing the predicted emotion values as explicit text in the prompt. This dual-modality approach combines vector steering with textual emotion conditioning to guide the decision generation process. All model inference uses the vLLM framework for acceleration, with a batch size of 16, temperature set to 0.0 for deterministic generation, top-p sampling parameter of 1.0, and maximum generation length of 1024 tokens.

4.2 Overall Performance

Decision Prediction Performance. Table 1 reports decision prediction results across five methods and two backbone models. Overall, EmoSim achieves the strongest and most balanced performance across accuracy, precision, and F1-score on both backbones, consistently outperforming all baseline approaches.

At the method level, EmoSim ranks first in accuracy, precision, and F1-score on both Qwen-2.5-32B-Instruct and Llama-3.1-8B-Instruct. In contrast, Emotion-Specialized LLM attains extremely high recall on both backbones but suffers from substantially lower precision and accuracy, indicating systematic over-prediction of punishment decisions. Explicit Emotion yields moderate improvements on the larger Qwen model but degrades sharply on the smaller Llama model, particularly in recall, suggesting limited robustness to model capacity. Two-Stage Emotion-Reasoning performs poorly across both backbones, indicating that loosely coupling emotion analysis and decision generation fails to capture fine-grained emotional signals. At the model level, larger backbones better exploit emotion-aware mechanisms. EmoSim achieves larger gains over Direct Decision on Qwen-2.5-32B-Instruct than on Llama-3.1-8B-Instruct, and the performance gap between EmoSim and baseline methods widens with model scale. This pattern suggests that higher-capacity models are more effective at integrating structured emotional modulation into decision prediction. These results show that emotion regulation via internal representation modulation enables more accurate and stable decision prediction than approaches relying solely on textual emotion cues, highlighting the importance of emotion-aware mechanisms for LLM-based social simulation.

Emotion Prediction Performance. As shown in Table 2, we evaluate emotion prediction performance on both valence and arousal dimensions using correlation (Corr.), mean absolute error (MAE), and root mean square error (RMSE) as evaluation metrics. Qwen-2.5-32B-Instruct achieves strong correlations of 0.749 and 0.770 on valence and arousal, with low MAE values of 0.058 and 0.050, and RMSE values of 0.090 and 0.081, respectively. In contrast, Llama-3.1-8B-Instruct achieves correlations of only 0.306 and 0.465 on the two dimensions, with significantly higher MAE

Variant	Accuracy	Precision	Recall	F1
w/o Explicit Prompt	0.6849	0.4713	0.5992	0.5243
w/o Emotion Vector	0.6753	0.4641	0.5333	0.4963
w/o Arousal	0.6954	0.4943	0.6621	0.5660
EmoSim	**0.7027**	**0.5034**	**0.6791**	**0.5782**

Table 3: Ablation Studies on core components of EmoSim.

Variant	Metric	Corr.	MAE	RMSE
EmoSim	Valence	**0.749**	**0.057**	**0.090**
	Arousal	**0.770**	**0.050**	**0.081**
w/o Interaction	Valence	0.519	0.423	0.457
	Arousal	-0.076	0.393	0.448
w/o Impression	Valence	0.731	0.059	0.094
	Arousal	0.738	0.053	0.086
w/o Norm	Valence	0.680	0.063	0.104
	Arousal	0.750	0.052	0.085

Table 4: Ablation study on social memory components.

(0.140 and 0.101) and RMSE (0.200 and 0.152) values. The results demonstrate that model scale has a substantial impact on emotion recognition capability. The 32B model achieves more than double the correlation compared to the 8B model on both dimensions, while maintaining significantly lower prediction errors. This validates that larger models can more effectively capture and predict the emotional dynamics underlying third-party punishment decisions, which is crucial for accurate emotion-guided decision generation.

Across all evaluated psychological groups, EmoSim achieves a higher F1 score than Direct Decision. Detailed group-level results are reported in the supplementary material.

4.3 Ablation Studies

In this section, we examine how different modules and hyperparameters affect EmoSim's simulation accuracy.

Contribution of Core Components. We first analyze the contribution of EmoSim's core components by selectively removing explicit emotion prompts, emotion vectors, or the arousal dimension. Results are reported in Table 3. EmoSim achieves the best overall performance when all components are enabled. Because the w/o Emotion Vector variant retains the same predicted valence and arousal values as EmoSim but provides them only as textual context, its F1 decrease from 0.5782 to 0.4963 isolates the additional contribution of representation-level modulation beyond matched text-only emotion conditioning. Conversely, the w/o Explicit Prompt variant retains vector injection without textual emotion descriptions and achieves an F1 of 0.5243. Within the evaluated setting, vector-only modulation outperforms text-only conditioning, while the full model performs best, indicating that the two channels provide complementary benefits. Removing the arousal dimension causes a smaller but consistent decline, suggesting that arousal provides complementary information beyond valence alone. Together, these results demonstrate that EmoSim benefits from the joint contribution of textual emotion cues, vector steering, and a two-dimensional emotion representation.

Role of Social Memory in Emotion Formation. We next investigate the role of different social memory components by removing interaction history, social impressions, or normative expectations sensitivity. Table 4 reports emotion prediction accuracy, and Figure 3 visualizes the resulting emotion distributions. Full EmoSim achieves strong correlations with ground-truth valence and arousal. Removing interaction history causes the most severe degradation, with valence correlation dropping sharply and arousal correlation collapsing to near zero. The distribution plots further show substantial mismatch between predicted and true emotion distributions, especially along the arousal axis. This indicates that modeling temporal interaction dynamics is critical for stable emotion formation.

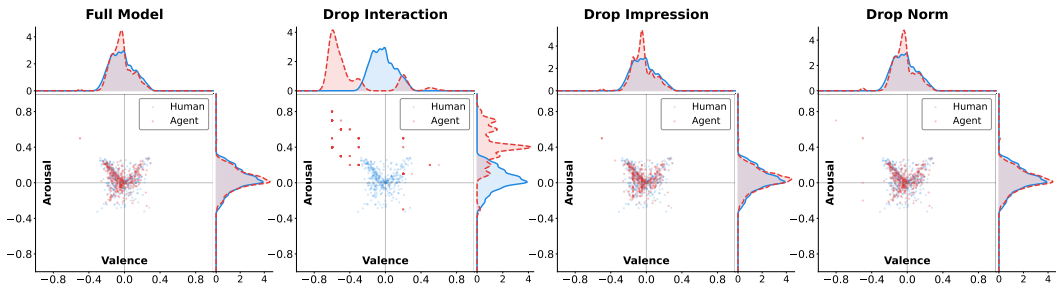

Figure 3: Distribution comparison of four variants in Valence-Arousal space.

Method	SimBench Score	TV Distance	KL Div
Direct Decision	-42.99	30.93	33.11
Explicit Emotion	-18.68	25.67	25.08
Emotion-Specialized LLM	-21.94	27.12	26.47
Two-Stage Emotion-Reasoning	-24.36	28.45	27.81
EmoSim	**-1.29**	**21.91**	**17.93**

Table 5: Transferability of Emotion Steering on MoralMachine.

Removing social impressions or normative expectations leads to smaller but consistent declines, suggesting that both components provide auxiliary but meaningful constraints on emotion prediction. The sensitivity analysis supports using layer 20 and an injection-coefficient range of $[1.5, 2.5]$. Full results are reported in the supplementary material.

4.4 Transferability of Emotion Steering

To evaluate whether emotion steering functions as a transferable mechanism rather than a task-specific alignment strategy, we test EmoSim on the MoralMachine, DICES-990, and ISSP tasks from SimBench (Hu et al., 2025). We focus on MoralMachine in the main text and report results for DICES-990 and ISSP in the supplementary material. MoralMachine represents a fundamentally different decision setting, focusing on moral dilemmas rather than punishment behavior, and was not used for any emotion or decision alignment during model design. We randomly select 100 scenarios and evaluate model outputs using the SimBench score for human response consistency, along with Total Variation (TV) distance and KL divergence to measure distributional alignment. As shown in Table 5, EmoSim consistently outperforms all compared baselines across all three metrics. Compared to Direct Decision, EmoSim substantially reduces the gap to human response distributions, improving the SimBench score by over 40 points while also lowering both TV distance and KL divergence. Explicit Emotion provides partial improvements but remains consistently inferior to EmoSim, indicating that textual emotion conditioning alone does not generalize as effectively. These results suggest that emotion vector steering does not overfit to the third-party punishment task, but generalizes to unaligned moral decision tasks and improves human-consistent behavior.

5 Conclusion

We present EmoSim, an emotion-regulated framework for LLM-based social agents that explicitly separates emotion formation from decision generation. By combining structured social memory with human-calibrated emotion vectors, EmoSim allows accumulated social experience to shape decisions through an internal emotional state rather than through direct prompting alone. Experiments on a third-party punishment task show that EmoSim produces more stable emotional responses and decision patterns that better align with human behavior than direct decision making and alternative emotion-based baselines. These findings suggest that explicitly modeling emotion as an internal regulatory state can improve the realism and stability of repeated social simulation, and provide a step toward more psychologically grounded LLM-based social agents.

GENERATIVE AI USE DISCLOSURE

We used ChatGPT to assist with language editing and to support the development of code for experimental implementation and statistical analyses. All AI-assisted code was reviewed, tested, and verified by the authors, and all statistical results reported in the paper were independently checked against the underlying data and analysis outputs. The authors reviewed and revised all AI-assisted text and code and take full responsibility for the final content, analyses, and conclusions of this work.

ETHICS STATEMENT

This work studies emotion-regulated LLM agents for social simulation and uses human behavioral data to align agent emotions with human responses. The study involving human participants was reviewed and approved by the Institutional Review Board (IRB) of our institution. All participants provided informed consent prior to participation, and personally identifying information was not used in model development.

REFERENCES

Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Pan Macmillan, 2017.

C Daryl Cameron, Paul Conway, and Julian A Scheffer. Empathy regulation, prosociality, and moral judgment. *Current Opinion in Psychology*, 44:188–195, 2022.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. clembench: Using game play to evaluate chat-optimized language models as conversational agents. *arXiv preprint arXiv:2305.13455*, 2023.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*, 2023a.

Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. *arXiv preprint arXiv:2311.09665*, 2023b.

Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. An appraisal-based chain-of-emotion architecture for affective language model game agents. *Plos one*, 19(5): e0301033, 2024.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

Nicholas Farn and Richard Shin. Tooltalk: Evaluating tool-usage in a conversational setting. *arXiv preprint arXiv:2311.10775*, 2023.

Ernst Fehr and Urs Fischbacher. Third-party punishment and social norms. *Evolution and human behavior*, 25(2):63–87, 2004.

Joseph P Forgas. Mood and judgment: the affect infusion model (aim). *Psychological bulletin*, 117 (1):39, 1995.

Jonathan Gratch and Stacy Marsella. Evaluating the modeling and use of emotion in virtual humans. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pp. 320–327. IEEE, 2004.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Sil Hamilton. Blind judgement: Agent-based supreme court modelling with gpt. *arXiv preprint arXiv:2301.05327*, 2023.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Simucourt: Building judicial decision-making agents with real-world judgement documents. *arXiv e-prints*, pp. arXiv–2403, 2024.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*, 2025.

Shin-nosuke Ishikawa and Atsushi Yoshino. Ai with emotions: Exploring emotional expressions in large language models. *arXiv preprint arXiv:2504.14706*, 2025.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*, 2022.

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

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Culturepark: Boosting cross-cultural understanding in large language models. *arXiv preprint arXiv:2405.15145*, 2024.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. From text to tactic: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*, 2023.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *arXiv preprint arXiv:2403.09498*, 2024a.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5487–5496, 2024b.

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

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759, 2024.

Rosalind W Picard. *Affective computing*. MIT press, 2000.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023a.

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv preprint arXiv:2301.04907*, 2023b.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980.

Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. Deciphering digital detectives: Understanding llm behaviors and capabilities in multi-agent mystery games. *arXiv preprint arXiv:2312.00746*, 2023.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*, 2023.

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. Ai for social science and social science of ai: A survey. *Information Processing & Management*, 61(3):103665, 2024.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference 2024*, pp. 1453–1463, 2024.

Yujia Zhou, Hexi Wang, Qingyao Ai, Zhen Wu, and Yiqun Liu. Simulating prosocial behavior and social contagion in llm agents under institutional interventions. *arXiv preprint arXiv:2505.15857*, 2025.

Chen Zhu, Yihang Cheng, Jingshuai Zhang, Yusheng Qiu, Sitao Xia, and Hengshu Zhu. Generative organizational behavior simulation using large language model based autonomous agents: A holacracy perspective. *arXiv preprint arXiv:2408.11826*, 2024.