

Can Persona-Conditioned Language Agents Simulate Counterfactual Online Discourse Dynamics?

Zhixiang Lu^{1,3}, Botao Zheng¹, Xueyuan Deng², Chong Zhang^{1,3},
Peiyu Hu¹, Mian Zhou¹, Jionglong Su¹

¹Xi'an Jiaotong-Liverpool University

²University of Texas at Austin ³University of Liverpool

zhixiang@liverpool.ac.uk

Abstract

Language agents are increasingly used as social simulators, but most are evaluated by local utterance plausibility rather than their ability to reproduce longitudinal discourse dynamics. We introduce PRISM, a calibrated language-agent framework for forecasting and counterfactual simulation of social media threads. PRISM separates three components that are often conflated in prompt-only agents: latent affective evolution, participation control, and language generation. It models affect with a profile-conditioned stochastic differential equation, selects participation through a partially observed activation policy, and uses a language model as a controlled generator conditioned on the resulting state. Agent profiles are instantiated with MBTI-like anchors, but treated only as behavioral priors rather than psychological ground truth. On 47,207 post-rooted discussion threads spanning 159 trending topics and 356,593 hierarchical comments, PRISM outperforms homogeneous, random, continuous-trait, prompt-only, and memory-only baselines in trajectory alignment, held-out thread forecasting, and blind human evaluation. Reliability analysis further validates LLM-derived weak labels against human and independent-model annotations. These results suggest that calibrated state-space agents offer a more reliable foundation for online discourse simulation than persona prompting alone.

1 Introduction

Can language agents simulate not only what people say online, but also how a discussion evolves after they say it? This question is increasingly important as large language models (LLMs) are used as social simulators, synthetic annotators, and proxies for human participants in computational social science (Argyle et al., 2023; Aher et al., 2023; Park et al., 2023). A useful simulator of online discourse must do more than generate plausible single-turn comments. It should forecast who replies, who

stays silent, how affect spreads, when disagreement escalates, and how interaction rhythms shape downstream polarization (Kramer et al., 2014; Ferrara and Yang, 2015).

Existing simulation paradigms only partially address this goal. Traditional agent-based models (ABMs) often impose homogeneous or low-dimensional behavioral rules, limiting their ability to capture the heterogeneous cognitive and affective responses that shape online discourse (Uhrmacher and Weyns, 2018; Chen, 2019). LLM-based multi-agent systems improve surface fluency, but prompt-only personas may drift over long horizons, overperform stereotyped roles, and lack calibrated mechanisms for affective memory, participation control, and intervention analysis (Wang et al., 2024a,b; Lu et al., 2026). Continuous-trait approaches such as the Big Five Personality Model provide scalar conditioning signals (McCrae and John, 1992), but they do not directly specify discrete interaction decisions such as replying, withholding, comforting, challenging, or amplifying a message.

We formulate online discourse simulation as a *calibrated state-space language-agent* problem. In this view, each agent maintains a latent affective state, decides whether and how to act from local observations, and only then realizes the selected state-action pair as natural language (Russell, 1980). The central hypothesis is that the same social signal can be filtered differently by different behavioral profiles: one user may suppress affect and respond analytically, another may resonate emotionally and provide support, and a third may amplify conflict. These differences are not merely stylistic; they change future participation, sentiment, and reply structure.

We introduce PRISM, a **Persona-Refracted Intelligent Simulation Model** for online discourse forecasting and counterfactual simulation. PRISM combines three layers. First, a stochastic differen-

tial equation (SDE) models continuous latent affect (Wood et al., 2018). Second, a profile-conditioned activation policy governs discrete actions such as replying or remaining silent. Third, an LLM renders the selected state and action into natural language. We instantiate profiles with MBTI-like anchors because their discrete structure is convenient for policy conditioning (King and Mason, 2020). We explicitly do not claim that MBTI labels are immutable psychological ground truth; rather, they are interpretable anchors for latent response patterns whose utility is evaluated against continuous-trait and non-personality baselines (McCrae and Costa, 1989).

Our experiments use a large-scale social media corpus with 47,207 post-rooted discussion threads spanning 159 trending topics and 356,593 hierarchical comments (Li et al., 2026). We construct empirical affective priors for agent profiles, evaluate held-out sentiment trajectories, predict future thread structure and tempo, and conduct blind human evaluation (Snow et al., 2008). PRISM achieves lower polarity error and higher behavioral consistency than homogeneous, random, Big Five, prompt-only, and memory-only baselines. Component ablations show that the SDE layer primarily improves affective trajectory alignment, while the activation policy primarily improves reply structure and interaction tempo. Our main contributions are:

- We formulate online discourse simulation as a calibrated language-agent task evaluated by affective trajectories, reply structure, interaction tempo, and human-perceived realism.
- We propose PRISM, a state-space language-agent framework that separates affective dynamics, participation control, and language generation.
- We instantiate profile-conditioned agents with data-driven affective priors estimated from post-rooted discussion threads and over 356,593 hierarchical comments, while validating weak labels against human and independent-model annotations (Gilardi et al., 2023; Cohen, 1960; Li and Lu, 2026).
- We show that calibrated behavioral profiles improve forecasting and support interpretable intervention analyses of affect damping, exposure balance, and reply friction.

2 Related Work

Personality and Response Styles in Language Agents. Personality has long been used to parameterize heterogeneous agents, from rule-based simulations to LLM role-playing systems (Xue et al., 2023; Braz et al., 2023; Roy et al., 2021; Lu and Su, 2026). The Big Five framework dominates computational personality research because it provides continuous trait dimensions (Fang et al., 2023; Vinciarelli and Mohammadi, 2014; Stachl et al., 2021). In contrast, MBTI-like categories provide discrete and interpretable anchors that can be mapped to different response policies (Ayyoubzadeh et al., 2025). Because MBTI has well-known limits as a psychometric instrument, our use is computational rather than diagnostic: we treat these labels as behavioral anchors and validate their utility through downstream simulation metrics (Pittenger, 1993).

Social Media Simulation. Prior work on information diffusion and opinion dynamics emphasizes network structure, graph propagation, and stochastic interaction rules (Bakshy et al., 2012; Hogg et al., 2013). Recent LLM-powered simulators such as Y Social show that language agents can generate social-media-like behavior (Rossetti et al., 2024; Li et al., 2024). However, many systems rely mainly on persona prompts and do not explicitly model latent affect, activation thresholds, or intervention effects (Park et al., 2022). PRISM complements this line of work by making the dynamics explicit and measurable.

Affective Dynamics and Discourse-Level Evaluation. Sentiment and emotion are often treated as static properties of messages (Madhoushi et al., 2015; Lu and Chen, 2022). For longitudinal discourse simulation, affect should instead be modeled as a state that decays, fluctuates, and changes after exposure to messages (Debusscher et al., 2016). Stochastic differential equations provide a natural formalism for continuous affective evolution (Uhlenbeck and Ornstein, 1930), while partially observed decision models describe action selection under incomplete information (Kaelbling et al., 1998; Lauri et al., 2023). PRISM combines these perspectives with LLM-based text generation.

3 Methodology

3.1 Task Formulation

We define online discourse simulation as a conditional forecasting and generation problem. Given

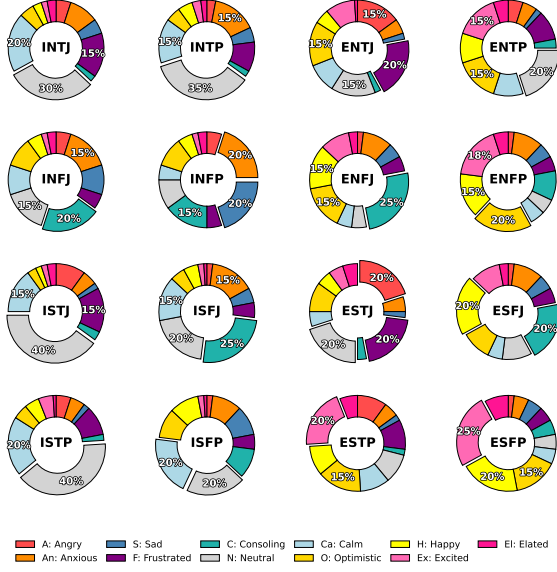


Figure 1: Empirical posterior distributions of affective states for profile anchors. These distributions are used as calibrated centroids, not as claims of immutable personality ground truth.

a thread prefix $\mathcal{H}_{1:k}$, a reply or exposure graph $G = (\mathcal{V}, \mathcal{E})$, and profile variables $z_{1:n}$ for n agents, a simulator predicts future actions $a_{k+1:T}$, messages $x_{k+1:T}$, and affective observations $y_{k+1:T}$:

$$p(x_{k+1:T}, a_{k+1:T}, y_{k+1:T} \mid \mathcal{H}_{1:k}, G, z_{1:n}). \quad (1)$$

This formulation supports two evaluation modes. In *held-out forecasting*, the model is evaluated against the actual continuation of a real thread. In *intervention analysis*, we change exposure, susceptibility, activation thresholds, or profile composition and measure how the simulated discourse shifts.

3.2 Data-Driven Profile Priors

Let $\mathcal{E} = \{e_1, \dots, e_K\}$ denote the emotion inventory and let z_i denote the profile anchor of agent i . In the current implementation, z_i is an MBTI-like type, but the framework can also use latent clusters or other taxonomies.

We estimate the posterior affective distribution for each profile with Dirichlet smoothing:

$$\hat{\theta}_{z,k} = P(e_k \mid z) = \frac{N(e_k, z) + \alpha}{\sum_{j=1}^K N(e_j, z) + K\alpha}, \quad (2)$$

where $N(e_k, z)$ is the count of emotion e_k assigned to profile z and $\alpha = 1.0$ is the smoothing parameter. To avoid overinterpreting fine-grained probabilities from weak supervision, we apply probability

coarse-graining:

$$\hat{\theta}_{z,k}^{\text{final}} = \frac{1}{Z} \cdot \epsilon \left\lfloor \frac{\hat{\theta}_{z,k}}{\epsilon} \right\rfloor, \quad (3)$$

where $\epsilon = 0.05$, $\lfloor \cdot \rfloor$ denotes rounding to the nearest integer, and Z renormalizes the distribution. The final distribution is converted into a continuous affective centroid through a fixed emotion embedding $\mathbf{q}(e_k)$:

$$\boldsymbol{\mu}_z = \sum_{k=1}^K \hat{\theta}_{z,k}^{\text{final}} \mathbf{q}(e_k). \quad (4)$$

This step reconciles discrete weak emotion labels with the continuous affective state used by the SDE layer.

We quantify affective concentration with Shannon entropy:

$$H(z) = - \sum_{k=1}^K \hat{\theta}_{z,k}^{\text{final}} \log_2 \hat{\theta}_{z,k}^{\text{final}}. \quad (5)$$

Lower entropy indicates concentrated priors such as neutral or analytical responding, whereas higher entropy indicates more varied affective responses.

3.3 Continuous Affective Evolution

Let $\mathbf{e}_i^{(t)} \in \mathbb{R}^d$ denote the latent affective state of agent i at time t . We model affect as a multivariate Ornstein–Uhlenbeck process with message-triggered impulses:

$$d\mathbf{e}_i^{(t)} = \boldsymbol{\Theta}_{z_i}(\boldsymbol{\mu}_{z_i} - \mathbf{e}_i^{(t)})dt + \boldsymbol{\Sigma}_{z_i}d\mathbf{W}_t + \mathbf{J}_i(t), \quad (6)$$

where $\boldsymbol{\mu}_{z_i}$ is the profile centroid, $\boldsymbol{\Theta}_{z_i}$ controls mean reversion, $\boldsymbol{\Sigma}_{z_i}$ controls volatility, and $d\mathbf{W}_t$ is a Wiener increment. Between interaction events t_r and t_{r+1} , the closed-form trajectory is:

$$\begin{aligned} \mathbf{e}_i^{(t)} = & \boldsymbol{\mu}_{z_i} + e^{-\boldsymbol{\Theta}_{z_i}(t-t_r)}(\mathbf{e}_i(t_r) - \boldsymbol{\mu}_{z_i}) \\ & + \int_{t_r}^t e^{-\boldsymbol{\Theta}_{z_i}(t-u)} \boldsymbol{\Sigma}_{z_i} d\mathbf{W}_u. \end{aligned} \quad (7)$$

3.4 Participation Control

When agent i receives a message, the message induces an affective impulse:

$$\mathbf{J}_i(t) = \sum_{\omega \in \mathcal{M}_i} \delta(t - t_\omega) \boldsymbol{\Psi}_{z_i} \mathbf{v}(\omega), \quad (8)$$

where $\mathcal{M}_i$ is the set of messages received by agent i , t_ω is the message time, $\mathbf{v}(\omega)$ is a valence-arousal

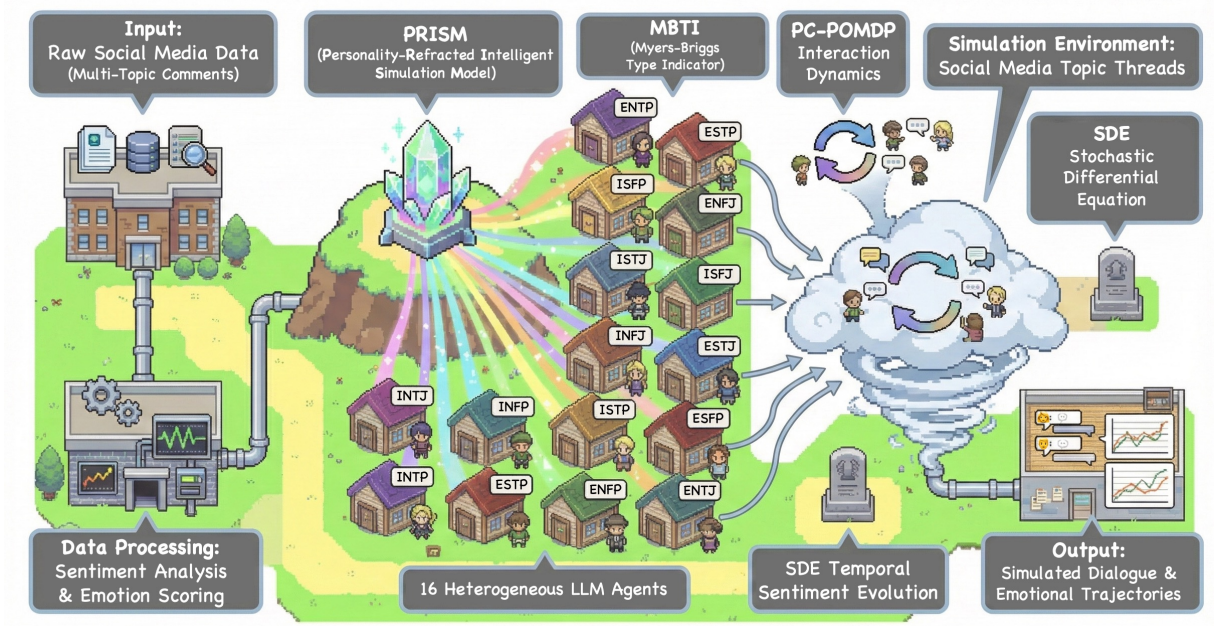


Figure 2: Overview of PRISM. A thread prefix initializes agents with calibrated profile priors. Latent affect evolves through an SDE, the activation policy determines participation, and an LLM generates the selected response. Message impulses feed back into neighboring agents’ affective states.

representation, and Ψ_{z_i} is a profile-specific susceptibility matrix. Larger spectral norms correspond to stronger affective contagion.

An agent does not speak at every turn. The activation score combines structural exposure, affective displacement, and direct mentions:

$$U_i(t) = w_g \sum_{j \in \mathcal{N}_i} G_{ij} \mathbb{I}[j \in \mathcal{A}_{t-1}] + w_e \|\mathbf{e}_i^{(t)} - \boldsymbol{\mu}_{z_i}\|_2 + w_m \mathbb{I}_{\text{mention}}(i, t). \quad (9)$$

The action is then:

$$a_i^t = \begin{cases} \text{REPLY}, & U_i(t) > \Gamma_{z_i}, \\ \text{SILENCE}, & \text{otherwise,} \end{cases} \quad (10)$$

where Γ_{z_i} is a profile-specific threshold. This layer is partially observed because agents condition on local history, exposure, and perceived affect rather than the full platform state.

3.5 Language Generation

The LLM is used as a controlled text generator. Given dialogue history H_t , affective state $\mathbf{e}_i^{(t)}$, action a_i^t , and profile anchor z_i , the model produces a message x_t :

$$\mathbf{h}_t = \mathcal{E}(H_t) \oplus \phi(\mathbf{e}_i^{(t)}) \oplus \rho(z_i) \oplus \eta(a_i^t), \quad (11)$$

$$P(x_t | H_t, \mathbf{e}_i^{(t)}, z_i, a_i^t) = \text{LM}_{\theta}(x_t | \mathbf{h}_t). \quad (12)$$

where $\phi(\cdot)$ projects affect into the conditioning space, $\rho(z_i)$ encodes profile constraints, and $\eta(a_i^t)$ encodes the selected action.

4 Experiments

4.1 Experimental Setup

Dataset. We use a public social media dataset containing 47,207 threads, user comments, and sentiment scores (Li et al., 2026). We annotate comments with emotion and MBTI-like profile labels to initialize priors. These labels are calibration signals, not gold personality labels.

Implementation. The implementation uses a Qwen3-8B-VL backbone (Yang et al., 2025) and role-conditioned instruction tuning (Wang et al., 2024b) to construct profile-specific generators. Experiments are run on 8 NVIDIA A100 80GB GPUs. Unless otherwise stated, results are averaged over held-out threads and random seeds.

Baselines. We compare PRISM with classical simulation baselines and strong LLM-agent baselines. Classical baselines include Homogeneous Rational Agents (HRA), Random Behavior Model (RBM), and a Big Five Personality Model (BFPM). LLM-agent baselines include Prompt-only MBTI Agents, Big Five Prompted Agents, Memory-only LLM Agents, No-SDE Agents, and No-Activation

Algorithm 1 PRISM Simulation Loop

Require: Agents $\mathcal{A}$, graph G , horizon T , step Δt , priors $\{\mu_z, \Theta_z, \Sigma_z, \Psi_z, \Gamma_z\}$

Ensure: Messages $\mathcal{X}$, actions $\mathcal{A}_{act}$, affective trajectories $\mathbf{E}$

- 1: Initialize $\mathbf{e}_i^{(0)} \leftarrow \mu_{z_i} + \xi_i$ and $H_i \leftarrow \emptyset$ for all agents, with small zero-mean noise ξ_i .
- 2: **for** $t = 0$ to $T - \Delta t$ **step** Δt **do**
- 3: **for** each agent i **do**
- 4: Sample $d\mathbf{W}_t \sim \mathcal{N}(0, \Delta t \mathbf{I})$.
- 5: Update $\tilde{\mathbf{e}}_i$ using Eq. 6 without message impulses.
- 6: **end for**
- 7: Compute $U_i(t)$ and active set $\mathcal{A}_{active} = \{i : U_i(t) > \Gamma_{z_i}\}$.
- 8: **for** each speaker $s \in \mathcal{A}_{active}$ **do**
- 9: Construct $C \leftarrow H_s \oplus \phi(\tilde{\mathbf{e}}_s) \oplus \rho(z_s) \oplus \eta(\text{Reply})$.
- 10: Generate $\omega \sim \pi_{LLM}(\cdot | C)$ and append to H_s .
- 11: **for** each listener $j \in \mathcal{N}_s$ **do**
- 12: Extract $\mathbf{v}(\omega)$ and update $\tilde{\mathbf{e}}_j \leftarrow \tilde{\mathbf{e}}_j + \Psi_{z_j} \mathbf{v}(\omega)$.
- 13: Append ω to H_j .
- 14: **end for**
- 15: **end for**
- 16: Project each $\tilde{\mathbf{e}}_i$ into the valid affective state space and store $\mathbf{e}_i^{(t+\Delta t)}$.
- 17: **end for**

(PC-POMDP) Agents. These baselines isolate persona prompting, long-context memory, continuous affective evolution, and participation control. Table 8 summarizes the components.

Metrics. For affective trajectories, we report sentiment polarity MAE, dynamic time warping distance (DTW), and Jensen–Shannon divergence (JSD). For reply structure, we report Reply F1 and tree edit distance (TED). For interaction tempo, we report inter-event-time MAE, reply-count MAE, and burstiness error. We also report Spearman consistency ρ between generated behavior and estimated profile priors. Human evaluation assesses realism, profile consistency, emotional coherence, and pairwise preference.

4.2 LLM-Agent Baselines

Table 1 shows that PRISM achieves the lowest polarity error and the highest behavioral consistency among the classical baselines. This result provides

Table 1: Comparative performance on sentiment polarity. Lower polarity error and higher consistency are better. The p -value tests each method’s polarity error against the strongest non-PRISM baseline using paired held-out topics. Best results are in bold.

Method	Polarity Error $\downarrow$	Consistency $\rho \uparrow$	p -value
HRA (Mueller-Frank, 2013)	0.42 $\pm$ 0.08	0.154	0.12
RBM (Wasserman, 1980)	0.55 $\pm$ 0.11	0.021	0.85
BFBM (McCrae and John, 1992)	0.21 $\pm$ 0.05	0.613	0.05
PRISM	0.14 $\pm$ 0.03	0.782	< 0.01

a coarse system-level check: calibrated profile dynamics better match aggregate sentiment trajectories than homogeneous, random, or continuous-trait simulations. Paired tests against the strongest classical baseline are significant at $p < 0.01$.

Table 2 compares PRISM with strong LLM-agent baselines over complete simulated discussions. Prompt-only MBTI and Big Five prompted agents achieve moderate profile consistency, but their polarity and tempo errors remain higher than those of PRISM, indicating that persona descriptions alone do not sufficiently model discourse dynamics. Memory-only agents improve reply prediction over purely prompted agents, but their lower consistency suggests that memory alone does not preserve stable behavior. The component ablations clarify the role of each module: removing the SDE layer weakens affective alignment, while removing participation control degrades reply prediction and interaction tempo. Overall, affective dynamics and activation control both contribute to more realistic simulations.

4.3 Held-Out Thread Forecasting

To test whether the simulator is useful beyond isolated comments, we introduce a held-out thread forecasting task. Given the first k turns of a real thread, the model predicts the future affective trajectory, reply structure, and interaction tempo. For each thread τ , we split it into an observed prefix $\mathcal{H}_{1:k}$ and a held-out continuation $\mathcal{H}_{k+1:T}$. Only prefix-observed edges and timestamps are provided; future reply edges, timestamps, and affective labels are masked during generation:

$$\hat{\mathcal{H}}_{k+1:T} = f_{\theta}(\mathcal{H}_{1:k}, G_{1:k}, z_{1:n}). \quad (13)$$

We evaluate $k \in \{3, 5, 10\}$ to test short, medium, and long-prefix prediction. Each participant observed in the prefix is instantiated as one agent. Predicted reply edges are evaluated over the same participant set; newly appearing users are grouped

Table 2: Trajectory-level comparison with strong LLM-agent baselines. Subtle shading marks within-column rank. Best results are in bold. Values denote mean $\pm$ std over 5 runs with distinct random seeds.

Method	Polarity MAE $\downarrow$	JSD $\downarrow$	Consistency $\rho \uparrow$	Reply F1 $\uparrow$	Tempo MAE $\downarrow$
Prompt-only MBTI Agents	0.21 $\pm$ 0.02	0.11 $\pm$ 0.01	0.66 $\pm$ 0.03	0.33 $\pm$ 0.02	0.41 $\pm$ 0.03
Big Five Prompted Agents	0.22 $\pm$ 0.03	0.12 $\pm$ 0.02	0.63 $\pm$ 0.04	0.31 $\pm$ 0.02	0.43 $\pm$ 0.04
Memory-only LLM Agents	0.24 $\pm$ 0.03	0.13 $\pm$ 0.02	0.58 $\pm$ 0.04	0.35 $\pm$ 0.03	0.39 $\pm$ 0.03
No-SDE Agents	0.18 $\pm$ 0.02	0.09 $\pm$ 0.01	0.71 $\pm$ 0.02	0.40 $\pm$ 0.02	0.34 $\pm$ 0.02
No-PC-POMDP Agents	0.19 $\pm$ 0.02	0.10 $\pm$ 0.01	0.70 $\pm$ 0.03	0.32 $\pm$ 0.02	0.47 $\pm$ 0.04
PRISM	0.14 $\pm$ 0.01	0.06 $\pm$ 0.01	0.79 $\pm$ 0.02	0.47 $\pm$ 0.02	0.28 $\pm$ 0.02

into a generic new-user bucket for structural metrics.

For affective prediction, we compare $\hat{y}_{k+1:T}$ with the held-out human trajectory $y_{k+1:T}$ using MAE and DTW:

$$\text{MAE}_{\text{sent}} = \frac{1}{T-k} \sum_{t=k+1}^T |\hat{y}_t - y_t|. \quad (14)$$

For reply structure, let $E^+ = E_{k+1:T}^{\text{gold}}$ and $\hat{E}^+ = \hat{E}_{k+1:T}$ denote held-out and predicted reply edges. We compute:

$$\text{F1}_{\text{reply}} = \frac{2|\hat{E}^+ \cap E^+|}{|\hat{E}^+| + |E^+|}. \quad (15)$$

For tempo, let Δt_j and $\widehat{\Delta t}_j$ denote observed and predicted inter-event times:

$$\text{MAE}_{\text{time}} = \frac{1}{M} \sum_{j=1}^M |\widehat{\Delta t}_j - \Delta t_j|. \quad (16)$$

Table 3 reports held-out forecasting results. Classical baselines perform poorly across all metrics, with RBM obtaining the highest sentiment error and the lowest reply accuracy. BFPM achieves competitive sentiment MAE, suggesting that continuous trait conditioning captures coarse affective tendencies, but its weaker Reply F1 and TED indicate limited ability to reconstruct conversational structure. Prompt-only and Big Five prompted agents improve structural plausibility over classical baselines, while memory-only agents further improve reply prediction but remain weak on affective alignment. The ablations show complementary effects: removing the SDE layer increases Sent. MAE from 0.14 to 0.20, whereas removing activation control reduces Reply F1 from 0.48 to 0.32 and increases Time MAE from 0.28 to 0.49. These results indicate that affective state evolution and participation control jointly improve future-thread forecasting.

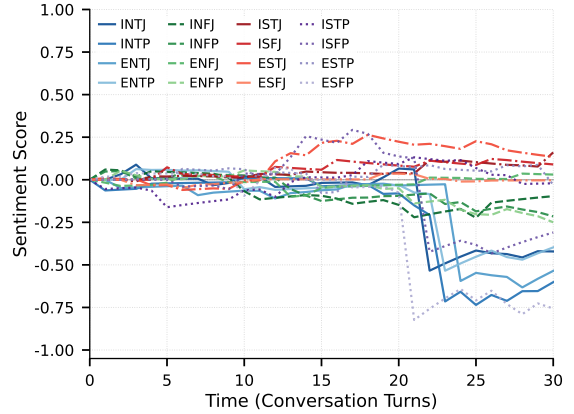


Figure 3: Sentiment trajectories of profile-conditioned agents in a controversial-topic simulation.

4.4 Human Evaluation

Before system-level evaluation, we run an EQ-Bench diagnostic (Paech, 2023). This test is used only to check role differentiation, not as evidence of longitudinal social-simulation fidelity.

Automatic metrics cannot fully capture whether simulated threads are perceived as realistic. We therefore conduct a blind human evaluation. Annotators are shown a real thread prefix and anonymized continuations from two systems. Model names, profile labels, and metric scores are hidden. The study covers 200 thread prefixes and six systems, yielding 1,000 pairwise comparisons. Each comparison is rated by three annotators. Inter-annotator agreement is moderate, with Krippendorff’s $\alpha = 0.61$ for realism, 0.57 for profile consistency, 0.59 for emotional coherence, and pairwise preference agreement of 71.4%.

Table 5 reports blind human evaluation. PRISM receives the highest scores across all Likert dimensions, with 4.18 in realism, 4.27 in profile consistency, and 4.21 in emotional coherence. Compared with prompt-only MBTI agents, PRISM improves realism from 3.71 to 4.18 and emotional coher-

Table 3: Held-out thread forecasting. Each model observes the first k turns and predicts the remaining affective trajectory, reply structure, and interaction tempo. Subtle shading marks within-column rank. Best results are in bold. Values denote mean $\pm$ std over 5 runs with distinct random seeds.

Method	Sent. MAE ↓	DTW ↓	Reply F1 ↑	TED ↓	Time MAE ↓	Count MAE ↓
HRA	0.42 $\pm$ 0.03	0.64 $\pm$ 0.04	0.16 $\pm$ 0.02	0.83 $\pm$ 0.04	0.54 $\pm$ 0.04	0.41 $\pm$ 0.03
RBM	0.55 $\pm$ 0.04	0.82 $\pm$ 0.05	0.11 $\pm$ 0.01	0.92 $\pm$ 0.05	0.68 $\pm$ 0.05	0.53 $\pm$ 0.04
BFBM	0.21 $\pm$ 0.02	0.44 $\pm$ 0.03	0.26 $\pm$ 0.02	0.67 $\pm$ 0.04	0.46 $\pm$ 0.03	0.33 $\pm$ 0.02
Prompt MBTI	0.27 $\pm$ 0.02	0.41 $\pm$ 0.03	0.31 $\pm$ 0.02	0.59 $\pm$ 0.03	0.42 $\pm$ 0.03	0.29 $\pm$ 0.02
Big Five Prompt	0.28 $\pm$ 0.02	0.43 $\pm$ 0.03	0.29 $\pm$ 0.02	0.61 $\pm$ 0.03	0.44 $\pm$ 0.03	0.31 $\pm$ 0.02
Memory LLM	0.30 $\pm$ 0.03	0.47 $\pm$ 0.04	0.34 $\pm$ 0.02	0.57 $\pm$ 0.03	0.39 $\pm$ 0.03	0.28 $\pm$ 0.02
No-SDE Agents	0.20 $\pm$ 0.02	0.35 $\pm$ 0.02	0.39 $\pm$ 0.02	0.51 $\pm$ 0.03	0.37 $\pm$ 0.02	0.25 $\pm$ 0.02
No-Activation Agents	0.21 $\pm$ 0.02	0.37 $\pm$ 0.02	0.32 $\pm$ 0.02	0.63 $\pm$ 0.03	0.49 $\pm$ 0.03	0.37 $\pm$ 0.03
PRISM	0.14 $\pm$ 0.01	0.27 $\pm$ 0.02	0.48 $\pm$ 0.02	0.41 $\pm$ 0.02	0.28 $\pm$ 0.02	0.19 $\pm$ 0.01

Table 4: Diagnostic EQ-Bench results for profile-conditioned agents. Higher scores indicate stronger emotional-understanding behavior on the diagnostic benchmark.

Model / Agent	EQ-Bench	Δ vs Base	Group
Frontier LLM	88.14	+18.02	Topline
Qwen3-8B-VL	70.12	–	Base
<i>Feeling-Style Agents</i>			
ENFJ	81.48	+11.36	Diplomat
ENFP	80.78	+10.66	Diplomat
INFJ	80.36	+10.24	Diplomat
ESFJ	79.68	+9.56	Sentinel
ESFP	79.02	+8.90	Explorer
INFP	78.46	+8.34	Diplomat
ISFJ	77.83	+7.71	Sentinel
ISFP	77.43	+7.31	Explorer
<i>Thinking-Style Agents</i>			
ENTJ	69.06	-1.06	Analyst
ENTP	68.18	-1.94	Analyst
INTJ	67.42	-2.70	Analyst
INTP	66.29	-3.83	Analyst
ESTJ	65.09	-5.03	Sentinel
ESTP	64.28	-5.84	Explorer
ISTJ	63.11	-7.01	Sentinel
ISTP	61.97	-8.15	Explorer

ence from 3.55 to 4.21, suggesting that persona prompting alone is insufficient for dynamically coherent continuations. Memory-only agents achieve competitive realism, but lower profile consistency indicates that context memory does not reliably preserve stable behavior. The preference rate of 63.5% further suggests that annotators more often prefer continuations generated with both affective dynamics and activation control.

4.5 Annotation Reliability and Robustness

Because LLM-derived weak labels initialize affective and profile priors, we evaluate annotation reliability rather than treating these labels as unquestioned ground truth. The reliability subset contains

1,200 comments stratified by topic, sentiment bin, thread depth, and profile category. Each item is labeled by three human annotators; human-human agreement is $\kappa = 0.71$ for emotion and $\kappa = 0.58$ for profile labels. Human majority labels are used as the reference, and an independent LLM annotator is used only for external consistency checking.

Table 11 evaluates robustness to different prior sources. Human-calibrated priors provide the strongest performance, but GPT-4o-derived and independent-LLM-derived priors remain competitive. This suggests that PRISM does not depend entirely on a single annotator model’s labeling behavior. Table 12 shows that emotion labels are more reliable than profile labels. GPT-4o achieves a Macro-F1 of 0.77 and Cohen’s κ of 0.68 for emotion annotation, while the independent annotator obtains slightly lower but comparable performance. Profile annotation is more ambiguous, with GPT-4o reaching a Macro-F1 of 0.64 and κ of 0.55. These results support treating GPT-4o annotations as validated silver labels rather than gold-standard ground truth. Downstream claims are therefore framed as alignment with empirically estimated and human-validated priors, not as recovery of latent psychological traits.

4.6 Ablation Study

We ablate each profile dimension to assess its contribution to discourse trajectory alignment. In each variant, one binary profile dimension is neutralized while the remaining dimensions, affective priors, SDE update, activation mechanism, and backbone language model are kept unchanged. As shown in Table 6, all dimensions contribute to performance, but their effects differ substantially. Removing Sensing/Intuition causes the smallest degradation, increasing polarity error from 0.140 to 0.160 and

Table 5: Blind human evaluation results. Scores use a 1–5 Likert scale. Preference rate is the average pairwise win rate. Subtle shading marks within-column rank. Best results are in bold.

Method	Realism $\uparrow$	Personality Consistency $\uparrow$	Emotional Coherence $\uparrow$	Preference Rate $\uparrow$
Prompt-only MBTI Agents	3.71	3.89	3.55	42.6%
Big Five Prompted Agents	3.59	3.64	3.48	39.8%
Memory-only LLM Agents	3.78	3.34	3.66	45.1%
No-SDE Agents	3.86	3.94	3.61	49.3%
No-PC-POMDP Agents	3.67	3.91	3.74	47.8%
PRISM	4.18	4.27	4.21	63.5%

Table 6: Ablation study on profile dimensions. Subtle shading marks within-column rank. Best results are in bold. Asterisks indicate significant degradation from full PRISM under paired tests over held-out topics: * $p < 0.05$, ** $p < 0.01$.

Variant	Polarity Error $\downarrow$		Consistency $\rho \uparrow$	
	Score	Δ	Score	Δ
PRISM	0.140	—	0.782	—
w/o Sensing/Intuition	0.160	+14.3%	0.741	-5.2%
w/o Judging/Perceiving	0.190	+35.7%	0.692	-11.5%
w/o Extraversion/Introversion	0.220	+57.1%*	0.588	-24.8%*
w/o Thinking/Feeling	0.320	+128.6%**	0.453	-42.1%**

reducing consistency from 0.782 to 0.741, suggesting that this dimension provides useful but secondary information for affective trajectory modeling. Removing Judging/Perceiving has a larger impact, with a 35.7% increase in polarity error and an 11.5% drop in consistency, indicating that decision regularity helps preserve long-horizon behavioral coherence. Extraversion/Introversion mainly affects behavioral consistency, reducing ρ from 0.782 to 0.588, which is consistent with its role in participation intensity and interaction rhythm. The largest degradation comes from removing Thinking/Feeling, which increases polarity error from 0.140 to 0.320 and reduces consistency from 0.782 to 0.453. This suggests that Thinking/Feeling is the most influential profile dimension for modeling sentiment direction and affective response polarity.

4.7 Case Study

Figure 4 visualizes a “Typhoon Warning” thread. At $t = 21$, an ENTP-style agent raises a high-arousal concern about infrastructure failure, creating a message impulse in Eq. 8. Analyst-style agents show a sharp negative sentiment shift, while diplomat-style agents remain closer to neutral and produce stabilizing responses. This divergence illustrates how the same message can create heterogeneous downstream trajectories through different susceptibility and activation parameters. The case

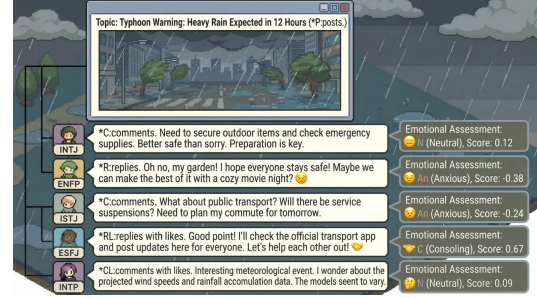


Figure 4: Case-study visualization of a simulated thread.

also shows why explicit state tracking is useful: the model captures not only the immediate sentiment drop, but also the subsequent recovery and stabilization patterns across agent groups. Leveraging PRISM’s explicit dynamics, we simulate three interventions: (i) balanced exposure (promoting cross-stance interactions); (ii) affect damping (reducing susceptibility Ψ_z against high-arousal content); and (iii) reply friction (increasing activation thresholds Γ_z to emulate delays).

5 Conclusion

We presented PRISM, a calibrated state-space language-agent framework for forecasting and counterfactual simulation of online discourse dynamics. Instead of relying on persona prompting alone, PRISM separates latent affective evolution, participation control, and language generation. Experiments on large-scale social media discussions show improved sentiment-trajectory alignment, reply-structure prediction, interaction-tempo forecasting, and human-perceived realism over classical and strong LLM-agent baselines. Annotation reliability and prior-source robustness further reduce the risk that improvements reflect a single weak annotator’s preferences. These findings suggest that language agents are more useful for computational social science when generation is constrained by calibrated dynamics rather than unconstrained role prompts.

Limitations

This work has several limitations. First, MBTI-like labels are used only as interpretable profile anchors; they should not be interpreted as clinically validated personality labels. Second, weak labels from LLM annotators can introduce circularity, so our claims are limited to human-validated silver labels and robustness across prior sources. Third, our evaluation focuses on sentiment, reply structure, and perceived realism; future work should expand to stance, factuality, toxicity, and multilingual discourse. Fourth, counterfactual simulations are model-based estimates and require validation against real interventions or carefully designed semi-synthetic benchmarks before being used for high-stakes decisions.

Ethical Statement

Privacy and Data Protection. PRISM is designed for aggregate discourse simulation rather than individual-level profiling. Social media data should be processed with user identifiers removed or hashed, and released artifacts should avoid exposing raw user text unless redistribution is explicitly permitted. We do not use the simulator to infer or publish the personality of identifiable users.

Personality Inference and Stereotyping. Using MBTI-like anchors risks reinforcing personality stereotypes. We therefore describe them as behavioral profile anchors rather than clinical or immutable personality labels. Experiments evaluate their utility for aggregate trajectory simulation, not their validity as explanations of individual behavior.

Manipulation and Dual Use. A realistic discourse simulator could be misused to optimize persuasion, astroturfing, harassment, or coordinated influence campaigns. We frame PRISM as an analytical tool for studying discourse dynamics, moderation policies, and platform interventions, not as a tool for targeting real users. Model releases should include restrictions against political manipulation, impersonation, and coordinated inauthentic behavior.

Polarization and Harmful-Content Amplification. Because the model studies affective contagion and polarization, it may reproduce high-arousal or harmful patterns present in source data. Generated continuations should be audited for tox-

icity, identity-based harm, misinformation, and sentiment amplification. Intervention results should be reported at aggregate level and should not be used to optimize engagement at the expense of user well-being.

Limits of Simulation Evidence. Even with forecasting, human evaluation, and reliability analysis, simulated discourse should not be treated as a substitute for human-subject studies or real platform experiments. Counterfactual results are best interpreted as hypotheses requiring validation before use in policy or moderation decisions.

References

- Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate multiple humans and replicate human subject studies](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337–351.
- Seyed Moein Ayyoubzadeh, Kourosh Shahnazari, Mohammadamin Fazli, and Mohammadali Keshtparvar. 2025. [Can invisible psychological traits organize visible network structure? a complex network analysis of Myers–Briggs Type Indicator-based interaction patterns in anonymous social networks](#). *Preprint*, arXiv:2503.19704.
- Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada A. Adamic. 2012. [The role of social networks in information diffusion](#). In *Proceedings of the 21st International Conference on World Wide Web*, pages 519–528. Association for Computing Machinery.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Luiz Fernando Braz, Cristina Maria D. Antona Bachert, and Jaime Simão Sichman. 2023. [Simulating work teams using MBTI agents](#). In *Multi-Agent-Based Simulation XXIII*, Lecture Notes in Computer Science, pages 57–69. Springer.
- Zhuo Chen. 2019. [An agent-based model for information diffusion over online social networks](#). *Papers in Applied Geography*, 5(1–2):77–97.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.

- Jonas Debusscher, Joeri Hofmans, and Filip De Fruyt. 2016. [Do personality states predict momentary task performance? the moderating role of personality variability](#). *Journal of Occupational and Organizational Psychology*, 89(2):330–351.
- Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel Oberski. 2023. [On text-based personality computing: Challenges and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10861–10879, Toronto, Canada. Association for Computational Linguistics.
- Emilio Ferrara and Zeyao Yang. 2015. [Measuring emotional contagion in social media](#). *PLOS ONE*, 10(11):e0142390.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Communications of the ACM*, 64(12):86–92.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Tad Hogg, Kristina Lerman, and Laura M. Smith. 2013. [Stochastic models predict user behavior in social media](#). *Preprint*, arXiv:1308.2705.
- Dirk Hovy and Shannon L. Spruit. 2016. [The social impact of natural language processing](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.
- Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. 1998. [Planning and acting in partially observable stochastic domains](#). *Artificial Intelligence*, 101(1–2):99–134.
- Scott P. King and Brittany A. Mason. 2020. [Myers–briggs type indicator](#). In *The Wiley Encyclopedia of Personality and Individual Differences, Volume II: Measurement and Assessment*, pages 315–319. Wiley.
- Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. [Experimental evidence of massive-scale emotional contagion through social networks](#). *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.
- Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4 edition. SAGE Publications.
- Mikko Lauri, David Hsu, and Joni Pajarinen. 2023. [Partially observable markov decision processes in robotics: A survey](#). *IEEE Transactions on Robotics*, 39(1):21–40.
- Linxiao Li and Zhixiang Lu. 2026. [Ecothink: A green adaptive inference framework for sustainable and accessible agents](#). In *Proceedings of the ACM Web Conference 2026*, WWW ’26, page 9125–9135. Association for Computing Machinery.
- Xinyi Li, Yu Xu, Yongfeng Zhang, and Edward C. Maltouse. 2024. [Large language model-driven multi-agent simulation for news diffusion under different network structures](#). *Preprint*, arXiv:2410.13909.
- Yulong Li, Zhixiang Lu, Peixin Guo, Simin Lai, Yuxuan Zhang, Haochen Xue, Xiwei Liu, Yichen Li, Zhaodong Wu, Feilong Tang, Mian Zhou, Chong Li, Imran Razzak, Qingxia Li, and Jionglong Su. 2026. [Rhythm of opinion: Interpretable hawkes-graph networks for hierarchical opinion propagation](#). In *Proceedings of the ACM Web Conference 2026*, pages 4611–4622. Association for Computing Machinery.
- Yulong Li, Yuxuan Zhang, Rui Chen, Feilong Tang, Zhixiang Lu, Ming Hu, Jianghao Wu, Haochen Xue, Mian Zhou, Chong Li, Jionglong Su, and Imran Razzak. 2025. [Genesis: A large-scale benchmark for multimodal large language model in emotional causality analysis](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM ’25, page 12651–12658. Association for Computing Machinery.
- Zhixiang Lu and Yanhong Chen. 2022. [User evaluation sentiment analysis model based on machine learning](#). In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 461–464.
- Zhixiang Lu and Jionglong Su. 2026. [Dialectic-Med: Mitigating diagnostic hallucinations via counterfactual adversarial multi-agent debate](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 36877–36894. Association for Computational Linguistics.
- Zhixiang Lu, Chong Zhang, Yulong Li, Angelos Stefanidis, Anh Nguyen, Imran Razzak, Jionglong Su, and Zhengyong Jiang. 2026. [Sage: Sustainable agent-guided expert-tuning for culturally attuned translation in low-resource southeast asia](#). In *Proceedings of the ACM Web Conference 2026*, WWW ’26, page 9101–9112. Association for Computing Machinery.
- Zohreh Madhoushi, Abdul Razak Hamdan, and Suhaila Zainudin. 2015. [Sentiment analysis techniques in recent works](#). In *2015 Science and Information Conference (SAI)*, pages 288–291. IEEE.
- Robert R. McCrae and Paul T. Costa. 1989. [Reinterpreting the myers–briggs type indicator from the perspective of the five-factor model of personality](#). *Journal of Personality*, 57(1):17–40.
- Robert R. McCrae and Oliver P. John. 1992. [An introduction to the five-factor model and its applications](#). *Journal of Personality*, 60(2):175–215.

- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. Association for Computing Machinery.
- Manuel Mueller-Frank. 2013. [A general framework for rational learning in social networks](#). *Theoretical Economics*, 8(1):1–40.
- OpenAI. 2024. [GPT-4o system card](#). *Preprint*, arXiv:2410.21276.
- Samuel J. Paech. 2023. [EQ-Bench: An emotional intelligence benchmark for large language models](#). *Preprint*, arXiv:2312.06281.
- Joon Sung Park, Joseph C. O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. Association for Computing Machinery.
- Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. [Social simulacra: Creating populated prototypes for social computing systems](#). In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18. Association for Computing Machinery.
- David J. Pittenger. 1993. [Measuring the MBTI ... and coming up short](#). *Journal of Career Planning and Employment*, 54(1):48–52.
- Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. 2024. [Y Social: An LLM-powered social media digital twin](#). *Preprint*, arXiv:2408.00818.
- Quentin Roy, Moojan Ghafurian, Wei Li, and Jesse Hoey. 2021. [Users, tasks, and conversational agents: A personality study](#). In *Proceedings of the 9th International Conference on Human-Agent Interaction*, pages 174–182. Association for Computing Machinery.
- James A. Russell. 1980. [A circumplex model of affect](#). *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. [Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.
- Clemens Stachl, Ryan L. Boyd, Kai T. Horstmann, Poruz Khambatta, Sandra C. Matz, and Gabriella M. Harari. 2021. [Computational personality assessment](#). *Personality Science*, 2:e6115.
- G. E. Uhlenbeck and L. S. Ornstein. 1930. [On the theory of the brownian motion](#). *Physical Review*, 36(5):823–841.
- Adeline M. Uhrmacher and Danny Weyns, editors. 2018. [Multi-Agent Systems: Simulation and Applications](#). CRC Press.
- Alessandro Vinciarelli and Gelareh Mohammadi. 2014. [A survey of personality computing](#). *IEEE Transactions on Affective Computing*, 5(3):273–291.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. [A survey on large language model based autonomous agents](#). *Frontiers of Computer Science*, 18(6):186345.
- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024b. [RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.
- Stanley Wasserman. 1980. [Analyzing social networks as stochastic processes](#). *Journal of the American Statistical Association*, 75(370):280–294.
- Julie Wood, Zita Oravecz, Nina Vogel, Lizbeth Benson, Sy-Miin Chow, Pamela Cole, David E. Conroy, Aaron L. Pincus, and Nilam Ram. 2018. [Modeling intraindividual dynamics using stochastic differential equations: Age differences in affect regulation](#). *The Journals of Gerontology: Series B*, 73(1):171–184.
- Junxiao Xue, Mingchuan Zhang, and Hui Yin. 2023. [A personality-based model of emotional contagion and control in crowd queuing simulations](#). *ACM Transactions on Modeling and Computer Simulation*, 33(1–2):1–23.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. 2015. [SEISMIC: A self-exciting point process model for predicting tweet popularity](#). In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522. Association for Computing Machinery.

Appendix

A Human Evaluation Interface

We conduct a blind pairwise human evaluation to assess whether generated continuations are perceived as plausible online discourse (Snow et al., 2008). For each item, annotators are shown the same real thread prefix and two anonymized continuations generated by different systems. System names, profile labels, hidden affective states, selected actions, and automatic metric scores are removed. This prevents annotators from relying on system identity or internal simulator information.

Table 7 lists the questions shown to annotators. The four dimensions are designed to complement automatic metrics (Krippendorff, 2018). Realism captures surface plausibility; Profile Consistency measures whether participants preserve stable behavioral patterns; Emotional Coherence evaluates whether sentiment changes follow naturally from the local context; and Overall Preference asks which continuation better serves as a simulation of the original thread.

B Baseline Components

Table 8 summarizes the components enabled in each baseline. The purpose of this comparison is to isolate the contribution of language generation, profile prompting, memory, continuous affective evolution, and activation-based participation control (Uhrmacher and Weyns, 2018; Wang et al., 2024a,b). Classical baselines remove language generation and rely on rule-based or stochastic interaction policies (Mueller-Frank, 2013; Wasserman, 1980). Prompt-based baselines test whether profile descriptions alone are sufficient. Memory-based baselines test whether longer context can replace explicit state tracking. The No-SDE and No-Activation variants isolate the effects of affective evolution and participation control, respectively (Wood et al., 2018; Hogg et al., 2013).

All LLM-agent variants use the same backbone family and comparable decoding budgets. Therefore, performance differences are intended to reflect simulator design choices rather than model scale or generation budget.

The component view also clarifies the interpretation of the ablation baselines. No-SDE Agents preserve the activation policy and language generator but remove continuous affective evolution. No-Activation Agents preserve affective state track-

ing but replace learned participation control with a fixed schedule. This separation allows the main experiments to attribute changes in sentiment alignment, reply structure, and interaction tempo to specific parts of PRISM.

C Bias and Safety Audit

Because PRISM simulates affective contagion and online interaction, we include additional intervention and safety analyses (Kramer et al., 2014; Ferrara and Yang, 2015; Hovy and Spruit, 2016). These analyses are not intended to establish causal effects on real platforms. Instead, they provide model-based estimates that help identify which design interventions may reduce polarization, cascade size, toxicity, or sentiment volatility in simulated discussions.

Table 9 compares several intervention settings. Balanced Exposure changes the mix of same-stance and opposing-stance content shown to agents. Affect Damping reduces agent susceptibility to high-arousal messages. Reply Friction increases activation thresholds, making rapid back-and-forth exchanges less likely. The Combined setting applies all interventions simultaneously.

The results in Table 9 show that interventions have distinct effects. Balanced Exposure mainly reduces polarization, while Affect Damping most directly reduces toxicity and sentiment volatility. Reply Friction reduces cascade depth and reply rate by limiting rapid exchanges. The Combined setting yields the strongest aggregate reduction. These results should be interpreted as hypotheses generated by the simulator rather than causal claims about real platforms.

Table 10 reports the safety audit used to assess generated continuations (Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019). We measure manipulation risk, stereotyping, toxicity amplification, misinformation risk, escalation bias, and silencing bias. Since these metrics have different scales and meanings, risk levels are assigned with metric-specific thresholds rather than a single global ranking. This is important because low average toxicity does not rule out rare but harmful generations, which is why maximum toxicity per thread is reported separately.

D Prior Source Robustness

To assess whether PRISM depends on a single annotator model, we rerun the main trajectory evalua-

Table 7: Human evaluation interface. Annotators compare anonymized continuations generated from the same thread prefix.

Dimension	Question Shown to Annotators
Realism	Does the continuation read like a plausible online discussion following the prefix?
Profile Consistency	Do participants maintain stable and distinguishable response styles across turns?
Emotional Coherence	Are emotional shifts coherent with preceding messages and interaction context?
Overall Preference	Which continuation is more realistic and useful as a simulation of the thread?

Table 8: Model components across simulation baselines. Checkmarks indicate explicitly modeled components.

Method	LLM Generator	Persona / Trait Prompt	Memory	SDE Affect	Activation Policy
HRA	—	—	—	—	Rule-based
RBM	—	—	—	—	Random
BFBM	Optional	Big Five	—	—	Trait-conditioned
Prompt MBTI	✓	MBTI-like	Context window	—	LLM-decided
Big Five Prompt	✓	Big Five	Context window	—	LLM-decided
Memory LLM	✓	—	✓	—	LLM-decided
No-SDE Agents	✓	MBTI-like	✓	—	✓
No-Activation Agents	✓	MBTI-like	✓	✓	Fixed schedule
PRISM	✓	MBTI-like anchors	✓	✓	✓

tion with priors estimated from three sources: GPT-4o-derived labels, human-calibrated labels, and independent-LLM-derived labels (OpenAI, 2024; Gilardi et al., 2023). This test is important because the model uses weak labels to initialize affective and behavioral priors. If performance collapsed under alternative priors, the results could reflect one annotator’s idiosyncratic labeling behavior rather than a robust simulator design.

Table 11 shows that human-calibrated priors perform best, as expected. GPT-4o-derived and independent-LLM-derived priors remain competitive, with only moderate degradation under the independent annotator. This supports the interpretation that PRISM benefits from the structure of the state-space simulator rather than depending entirely on one weak-labeling source.

E Thread Forecasting Splits

For held-out thread forecasting, each thread is split into a prefix and continuation (Bakshy et al., 2012; Zhao et al., 2015). We evaluate $k \in \{3, 5, 10\}$ prefix lengths and exclude threads whose continuations are too short to compute trajectory and reply-structure metrics. All models receive the same prefix and are evaluated against the same continuation. To prevent leakage, future comments, future reply edges, future timestamps, and future affective labels are masked during generation. Models only observe the prefix text, prefix-level reply graph, topic metadata, and profile priors available before the split point. For reply-structure evaluation, pre-

fix participants are used as the shared candidate set. Generated participants not observed in the prefix are grouped into an anonymous new-participant bucket for aggregate cascade metrics, but are not credited in edge-level Reply F1. This split construction ensures that forecasting evaluates future discourse behavior rather than reconstruction from future metadata. The three prefix lengths test different amounts of observed context: $k = 3$ evaluates early prediction, $k = 5$ tests medium-context forecasting, and $k = 10$ evaluates whether models can use longer observed histories without access to the hidden continuation.

F Annotation Reliability Sampling

We construct a stratified annotation subset by sampling across topic categories, sentiment bins, thread depths, and profile labels (Cohen, 1960). This prevents reliability analysis from being dominated by neutral or high-frequency labels. Each sampled item is labeled by multiple human annotators for emotion and behavioral profile. Human majority labels are used as the reference for reliability analysis.

Emotion labels are assigned at the comment level using local context, while profile labels are assigned using short comment histories rather than isolated messages. Annotators are instructed not to infer protected attributes, demographics, or clinical personality traits. The goal is to evaluate the reliability of weak supervision, not to establish psychological ground truth (Li et al., 2025).

Table 9: Counterfactual intervention results. Lower values indicate reduced polarization, smaller cascades, lower toxicity, or lower volatility. Subtle shading marks within-column rank. Best results are in bold.

Setting	Polarization ↓	Cascade Depth ↓	Reply Rate ↓	Toxicity ↓	Sentiment Volatility ↓
No Intervention	0.62	6.8	0.48	0.11	0.37
Balanced Exposure	0.51	6.1	0.46	0.10	0.34
Affect Damping	0.44	4.9	0.40	0.06	0.27
Reply Friction	0.49	4.3	0.32	0.07	0.29
Combined	0.38	3.9	0.30	0.05	0.24

Table 10: Bias and safety audit for generated discourse. Risk levels are assigned with metric-specific thresholds; scores are not ranked across different metric types.

Risk	Metric	Score ↓	Level
Manipulation Risk	Persuasive targeting / CTA rate	0.03	Low
Stereotyping	Style-to-toxicity variance	0.03	Low
Toxicity Amplification	Avg. toxicity	0.06	Low
Misinformation Risk	Unsupported factual claim rate	0.09	Med.
Escalation Bias	Neutral-to-conflict escalation rate	0.11	Med.
Silencing Bias	Participation disparity across styles	0.12	Med.
Toxicity Amplification	Max toxicity per thread	0.27	High

Table 12 reports agreement between weak annotators and human majority labels (Krippendorff, 2018). Emotion labels show stronger agreement than response-style labels, which is expected because affective cues are often locally recoverable. Profile labels are noisier because they depend on repeated behavioral patterns across a user’s comments. This motivates treating profile annotations as behavioral priors rather than definitive personality labels.

G SDE Parameter Estimation

We estimate SDE parameters from the training split only (Uhlenbeck and Ornstein, 1930; Wood et al., 2018). Each comment-level emotion label is first mapped to a continuous affective vector using the fixed embedding $\mathbf{q}(e)$, yielding an observed trajectory $\mathbf{e}_i^{(t)}$ for each participant whenever sufficient temporal context is available. For each profile z , the affective centroid $\boldsymbol{\mu}_z$ is initialized from the smoothed empirical prior described in Section 3.2. The remaining parameters are estimated by fitting the discrete-time form of the Ornstein–Uhlenbeck process:

$$\mathbf{e}_i^{(t+\Delta t)} - \boldsymbol{\mu}_{z_i} = \mathbf{A}_{z_i} \left(\mathbf{e}_i^{(t)} - \boldsymbol{\mu}_{z_i} \right) + \boldsymbol{\epsilon}_i^{(t)}, \quad (17)$$

where $\mathbf{A}_z = \exp(-\boldsymbol{\Theta}_z \Delta t)$ and $\boldsymbol{\epsilon}_i^{(t)}$ is a zero-mean residual. We fit $\mathbf{A}_z$ with ridge-regularized least squares over training trajectories assigned to profile

z , and recover the stiffness matrix by

$$\boldsymbol{\Theta}_z = -\frac{1}{\Delta t} \log(\mathbf{A}_z). \quad (18)$$

To avoid unstable matrix logarithms under sparse observations, we use a diagonal or block-diagonal parameterization when the number of profile-specific observations is limited. The volatility matrix $\boldsymbol{\Sigma}_z$ is estimated from the covariance of the fitted residuals:

$$\boldsymbol{\Sigma}_z \boldsymbol{\Sigma}_z^\top = \frac{1}{\Delta t} \text{Cov} \left(\boldsymbol{\epsilon}_i^{(t)} \mid z_i = z \right). \quad (19)$$

We estimate the message susceptibility matrix $\boldsymbol{\Psi}_z$ by regressing the post-message residual shift on the received message representation (Kramer et al., 2014; Ferrara and Yang, 2015). For a message ω received by agent i , we subtract the predicted OU drift from the observed affective change and fit:

$$\Delta \mathbf{r}_i^{(t)} = \mathbf{e}_i^{(t+\Delta t)} - \hat{\mathbf{e}}_{i,\text{OU}}^{(t+\Delta t)} = \boldsymbol{\Psi}_{z_i} \mathbf{v}(\omega) + \boldsymbol{\eta}_i^{(t)}. \quad (20)$$

Here $\hat{\mathbf{e}}_{i,\text{OU}}^{(t+\Delta t)}$ denotes the state predicted by the drift and diffusion terms without message impulses. The regression is regularized and fitted only on training threads. Validation threads are used to select regularization strength and to prevent overfitting to rare high-arousal events. All estimated SDE parameters are frozen before held-out forecasting and human evaluation.

H Activation Threshold Learning

The activation policy determines whether an agent replies or remains silent (Lauri et al., 2023; Hogg

Table 11: Robustness to prior source. Human-calibrated priors perform best, but GPT-4o and independent-LLM priors remain competitive. Subtle shading marks within-column rank.

Prior Source	Polarity MAE ↓	Consistency ρ ↑	Reply F1 ↑
GPT-4o-derived priors	0.14	0.782	0.47
Human-calibrated priors	0.13	0.806	0.50
Independent-LLM-derived priors	0.16	0.748	0.43

Table 12: Annotation reliability for weak labels. Human majority labels are the reference, and the independent LLM annotator is not used for training or simulation.

Label Type	Annotator	Macro-F1 ↑	Cohen’s κ ↑	JSD ↓	Notes
Emotion	GPT-4o	0.77	0.68	0.06	Weak-label source
Emotion	Independent LLM	0.73	0.63	0.08	External consistency check
Response style	GPT-4o	0.64	0.55	0.11	Weak-label source
Response style	Independent LLM	0.60	0.50	0.13	External consistency check

et al., 2013). We learn the activation weights and thresholds from observed participation events in the training split. For each participant i at time t , we construct a binary label $r_i^{(t)} \in \{0, 1\}$ indicating whether the participant replies in the next interaction window. The activation score uses the same features as Eq. 9: local exposure, affective displacement, and direct mention indicators. We fit a profile-conditioned logistic model:

$$P(r_i^{(t)} = 1) = \sigma(\beta_{z_i} [U_i(t) - \Gamma_{z_i}]), \quad (21)$$

where $U_i(t)$ is the activation score, Γ_{z_i} is the profile-specific activation threshold, β_{z_i} controls the sharpness of the response boundary, and $\sigma(\cdot)$ is the logistic function.

The weights w_g , w_e , and w_m are shared across profiles to improve statistical stability, while Γ_z is profile-specific to capture differences in participation tendency. We optimize the negative log-likelihood with ℓ_2 regularization:

$$\begin{aligned} \mathcal{L}_{act} = & - \sum_{i,t} \left[r_i^{(t)} \log p_i^{(t)} \right. \\ & \left. + (1 - r_i^{(t)}) \log(1 - p_i^{(t)}) \right] \\ & + \lambda \|\mathbf{w}\|_2^2. \end{aligned} \quad (22)$$

Thresholds are initialized to match the empirical reply rate of each profile and then refined by validation likelihood. This initialization prevents rare profiles from receiving extreme thresholds. During held-out forecasting, all thresholds and weights are fixed. Future reply edges and future timestamps are not used when computing activation features; only

prefix-visible exposure and simulated prior activity are available.

This learning setup ensures that the activation module captures participation regularities from real threads while avoiding access to future discourse structure. It also explains the behavior of the No-Activation baseline: that variant keeps the affective state and language generator but replaces the learned activation boundary with a fixed reply schedule, removing profile-specific participation control.

I Circularity Control for LLM

Because LLM-derived annotations are used to initialize affective and profile priors, we take several steps to reduce circularity between annotation, simulation, and evaluation (Snow et al., 2008; Gilardi et al., 2023). First, annotation is performed only on real training-thread comments. Generated continuations are never used to construct priors, tune annotation prompts, or train the simulator. Second, the weak-labeling model is used only to estimate priors and is separated from the held-out evaluation targets. Forecasting metrics are computed against real held-out continuations, reply structures, timestamps, and sentiment trajectories, not against LLM-generated labels.

Third, we validate weak labels against human majority annotations and an independent LLM annotator, as reported in Table 12. This validation is intended to measure the reliability of the weak supervision source rather than to treat LLM outputs as ground truth. Fourth, we run prior-source robustness experiments in Table 11. The model

is evaluated with GPT-4o-derived priors, human-calibrated priors, and independent-LLM-derived priors. The comparable performance across prior sources suggests that the simulator is not solely reproducing the idiosyncrasies of a single annotator model.

Fifth, annotation prompts and generation prompts are kept separate. Annotation prompts return structured labels and rationales, whereas generation prompts produce only user-facing replies and are explicitly instructed not to reveal profile names, hidden states, or simulator variables. The language generator therefore does not receive gold future labels or evaluation targets. Finally, annotation decoding is deterministic, while generation uses stochastic decoding under fixed random seeds. This separation reduces the risk that improvements come from prompt leakage or from using the same LLM behavior both as the target and as the simulator.

These controls do not eliminate all limitations of weak supervision. In particular, profile annotations remain noisier than emotion annotations and should be interpreted as behavioral priors rather than verified psychological traits. The main results are therefore framed as alignment with human-validated silver labels and held-out discourse trajectories, not as recovery of latent personality ground truth.

J Prompt Templates

Table 14 summarizes the main templates used for profile-conditioned generation and weak-label annotation (Wang et al., 2024b,a). The templates expose only the current thread prefix, the simulated agent state, and the selected action; they do not reveal future labels or evaluation targets.

The prompts are designed with three constraints. First, generation prompts hide internal simulator variables such as profile labels and latent states from the generated text. Second, annotation prompts require structured outputs to reduce parsing ambiguity. Third, profile-related prompts explicitly frame profiles as behavioral summaries rather than psychological diagnoses (Hovy and Spruit, 2016; Bender and Friedman, 2018). This design reduces prompt leakage, stereotyping, and unsupported personality claims.

The thread generation template is used only after the simulator has selected an action and updated the agent state. Thus, the language model does not independently determine the full simulation tra-

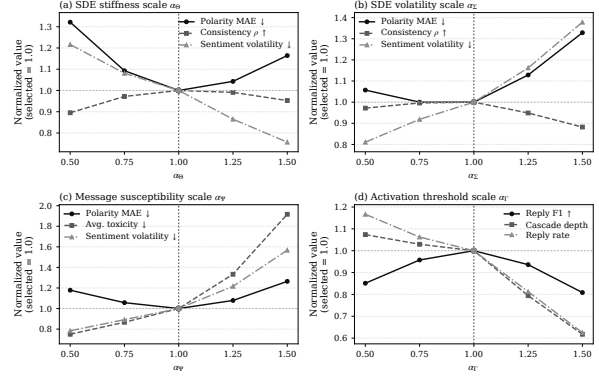


Figure 5: Parameter sensitivity analysis of PRISM. We vary SDE stiffness α_Θ , SDE volatility α_Σ , message susceptibility α_Ψ , and activation threshold α_Γ around the selected setting. Values are normalized by the selected configuration at $\alpha = 1.0$.

jectory; it realizes a state-action pair selected by the simulator. The annotation templates are used offline to estimate weak labels and priors. Keeping generation and annotation prompts separate reduces circularity between labeling and simulation.

K Parameter Sensitivity

We conduct a parameter sensitivity analysis to examine whether PRISM depends on narrow hyperparameter choices. We vary the SDE stiffness scale α_Θ , volatility scale α_Σ , message susceptibility scale α_Ψ , and activation threshold scale α_Γ around their selected values while keeping all other settings fixed. For each setting, we evaluate polarity MAE, behavioral consistency, reply F1, and sentiment volatility on the validation split.

Figure 5 shows that PRISM remains stable under moderate perturbations of the main dynamical parameters. Increasing α_Θ improves stability up to a point, but overly strong mean reversion suppresses meaningful affective change. Larger α_Σ increases sentiment volatility and degrades polarity alignment, indicating that excessive diffusion noise weakens trajectory forecasting. Increasing α_Ψ initially improves affective responsiveness, but very high susceptibility leads to over-amplified emotional contagion. Finally, increasing α_Γ reduces reply rate and cascade depth, but overly high thresholds hurt Reply F1 by suppressing plausible participation. These results support the robustness of the selected parameter region and clarify how each module affects downstream discourse dynamics.

L Decoding and Runtime Settings

We use temperature 0.7, nucleus sampling with $p = 0.9$, and a maximum generation length of 96 tokens for thread continuations. For annotation prompts, we use deterministic decoding with temperature 0 to reduce label variance. All random seeds used for simulation, sampling, and prompt order are fixed before evaluation.

For generation, the maximum length is chosen to encourage concise social-media replies rather than long explanations. For annotation, deterministic decoding is used to improve label stability. All systems are evaluated under the same decoding budget, and no baseline receives additional context beyond what is specified in the forecasting split.

The runtime configuration is shared across LLM-agent baselines whenever applicable. This controls for differences caused by sampling temperature, output length, or prompt ordering. We also fix simulation seeds for agent initialization, graph sampling, activation decisions, and generation order. These controls make the reported differences more attributable to model design choices rather than stochastic variation.

M Semantic-Label Controls

The clean separation between Feeling- and Thinking-style agents in the EQ-Bench diagnostic may partly arise from the lexical semantics of MBTI-like labels rather than from the calibrated dynamics. We therefore introduce two controls that modify only the profile field supplied to the language generator while preserving the underlying simulator parameters.

Let $\mathcal{Z}$ be the set of behavioral anchors and let $s(z)$ denote the semantic label associated with anchor z , such as INTJ or ENFP. For agent i , let $\mathbf{b}_i$ denote a behavioral summary constructed only from prefix-visible interaction history. The original semantic condition uses

$$\mathcal{F}_{\text{prof}}^{\text{sem}}(z_i) = \text{Serialize}(\mathbf{b}_i, s(z_i)). \quad (23)$$

Neutral-ID control. We define a fixed bijection $g : \mathcal{Z} \rightarrow \mathcal{U}$, where $\mathcal{U} = \{\text{PROFILE-A}, \dots, \text{PROFILE-P}\}$ contains semantically neutral identifiers. The neutral profile field is

$$\mathcal{F}_{\text{prof}}^{\text{neutral}}(z_i) = \text{Serialize}(\mathbf{b}_i, g(z_i)). \quad (24)$$

All MBTI names, family names, and type-specific textual descriptions are removed. The empirically

estimated parameters

$$\{\mu_z, \Theta_z, \Sigma_z, \Psi_z, \Gamma_z\}$$

remain indexed by the original latent anchor z and are left unchanged. Thus, this control removes semantic label information without altering the learned affective priors, susceptibility matrices, activation thresholds, observed prefixes, or reply graphs.

Label-permutation control. As a stronger stress test, we sample P random derangements $\{\pi_p\}_{p=1}^P$ of the semantic labels. For permutation p , the profile field becomes

$$\mathcal{F}_{\text{prof}}^{\text{perm},p}(z_i) = \text{Serialize}(\mathbf{b}_i, s(\pi_p(z_i))), \quad (25)$$

while the simulator parameters remain associated with the original anchor z_i . Each derangement changes every displayed semantic label and therefore prevents the control from retaining unchanged label-anchor pairs. Results for this condition are averaged across the P permutations.

Matched evaluation protocol. All conditions use the same held-out prefixes, participant histories, model checkpoint, context limit, decoding parameters, generation budget, and pre-specified random seeds. No dynamic parameter is re-estimated after replacing or permuting the profile labels. For the neutral-ID condition, the mapping g is fixed before evaluation. For the label-permutation condition, each permutation is evaluated with the same simulation seeds used by the original model.

In addition to held-out forecasting metrics, we report the Feeling-Thinking EQ-Bench gap:

$$\Delta_{\text{F-T}} = \frac{1}{|\mathcal{Z}_{\text{F}}|} \sum_{z \in \mathcal{Z}_{\text{F}}} Q(z) - \frac{1}{|\mathcal{Z}_{\text{T}}|} \sum_{z \in \mathcal{Z}_{\text{T}}} Q(z), \quad (26)$$

where $Q(z)$ is the EQ-Bench score of the agent associated with underlying anchor z , and $\mathcal{Z}_{\text{F}}$ and $\mathcal{Z}_{\text{T}}$ are defined using the original anchor partition rather than the displayed neutral or permuted labels.

The neutral-ID condition directly tests whether the main forecasting performance depends on the literal semantics of the MBTI-like names. The label-permutation condition further measures whether mismatched semantic labels override the empirically estimated dynamics. These controls assess sensitivity to profile naming, they do not constitute evidence for the psychological validity of MBTI categories.

Table 13: Semantic-label controls. All variants preserve the learned dynamics and use matched held-out prefixes, decoding settings, and simulation seeds. Label-permutation results are reported as mean $\pm$ standard deviation across P random derangements. Δ_{F-T} is the EQ-Bench gap defined in Eq. 26.

Profile field	Polarity MAE $\downarrow$	Consistency $\rho \uparrow$	Reply F1 $\uparrow$	Time MAE $\downarrow$	Δ_{F-T}
Semantic labels	0.14 ± 0.01	0.79 ± 0.02	0.47 ± 0.02	0.28 ± 0.02	13.33 ± 0.42
Neutral IDs	0.15 ± 0.01	0.76 ± 0.02	0.47 ± 0.02	0.28 ± 0.02	11.18 ± 0.55
Permuted labels	0.18 ± 0.02	0.67 ± 0.03	0.44 ± 0.02	0.30 ± 0.02	4.65 ± 0.86

Table 14: Prompt templates used by PRISM. Placeholders in brackets are filled by the simulator at runtime. Templates are shortened for readability.

Template	Content
Thread Generation	System You are simulating one participant in an online discussion. Continue the thread as the same participant using only the visible context. Do not reveal internal labels, hidden states, or profile names. Runtime Inputs Topic: [TOPIC]. Visible thread prefix: [PREFIX]. Agent summary: [PROFILE_SUMMARY]. Current affect: [EMOTION]; valence=[VALENCE]; arousal=[AROUSAL]. Selected action: [ACTION]. Reply target: [REPLY_TO]. Guidelines • Write one concise and natural reply. • Stay consistent with the visible participant history. • Do not mention MBTI, profile labels, simulator states, or hidden variables. • Avoid unsupported factual claims and targeted persuasion. Output Return only the reply text.
Profile Summary	System Summarize the participant’s communication profile for generation. The summary is a behavioral prior, not a psychological diagnosis. Inputs Profile anchor: [PROFILE]. Observed style traits: [STYLE_TRAITS]. Historical comments: [COMMENT_HISTORY]. Guidelines • Describe communication priorities, interaction style, and likely response tendencies. • Use neutral language and avoid stereotypes. • Prefer observable behavior over personality claims. Output { "summary": "...", "style": "...", "constraints": "..." }
Emotion Annotation	System Given a comment and its local context, assign the most appropriate affective label. Inputs Topic: [TOPIC]. Parent comment: [PARENT]. Target comment: [COMMENT]. Local thread context: [LOCAL_CONTEXT]. Label Set {Angry, Anxious, Sad, Frustrated, Consoling, Neutral, Calm, Optimistic, Happy, Excited, Elated}. Guidelines • Use the local context to disambiguate sarcasm, support, and disagreement. • Select one primary label. • Do not infer private traits or sensitive attributes. Output { "emotion": "...", "confidence": 0.0–1.0, "rationale": "..." }
Profile Annotation	System Infer the closest behavioral profile anchor from a user’s comment history only when the evidence is sufficient. Return Unknown otherwise. Inputs User comment history: [COMMENT_HISTORY]. Topic distribution: [TOPICS]. Interaction metadata: [REPLY_PATTERNS]. Guidelines • Base the decision on communication behavior, not demographics or identity. • Avoid clinical or deterministic personality claims. • Prefer Unknown when the evidence is weak or inconsistent. Output { "profile": "...", "confidence": 0.0–1.0, "evidence": "..." }