

PsyMemBehav: Psychology-Informed and Memory-Grounded Two-Level Identity-to-Action Reasoning for Social-Media Behavior Simulation

Wenji Zhang^{1,*}, Xilong Cheng^{1,*}, Yunxiao Qin^{1,2,*}, Mai Tu¹, Tongye Zhu¹, Yuhao Zhou¹, JingBo Ji¹, Yuting Tan¹, Duoyin Li¹, Hongjiang Xiao^{1,2,†}, Yuan Zhang^{1,2}

¹Communication University of China, China

²State Key Laboratory of Media Convergence and Communication, China

202311153001@mails.cuc.edu.cn, chengzhengyu330@cuc.edu.cn, qinyunxiao@cuc.edu.cn

*Equal contribution. †Corresponding author.

Abstract

Large language models (LLMs) support social simulation, but faithfully reproducing real users requires grounding both structured platform decisions and language in sparse, noisy behavioral evidence. We introduce **PsyMemBehav**, which maps psychology-informed profiles and retrieved memories to target-first platform behavior through two-level identity-to-action reasoning. To support this approach with comprehensive behavioral traces, we construct a dataset PsyMemBehav-bench from 5.57M observed behaviors of 5,862 real users (approximately 950 per user) across Reddit, Bluesky, and Twitter. Each recorded engagement is modeled as target selection followed by COMMENT, LIKE, or REPOST and optional response generation. PsyMemBehav combines 41-dimensional user descriptors, top-5 BM25 memories, reverse-CoT supervision, and rubric-guided DPO over complete reasoning-action trajectories. Experiments show that PsyMemBehav 27B ranks first in 11 of 12 reported individual-level metrics, improving average response quality over GPT-5.4 from 54.28 to 59.68 (9.95% relative), and obtains the lowest group-level MAE (0.081) and MSE (0.009) among simulated systems. These results support evidence-grounded user modeling for observed-thread continuation; the benchmark remains conditional on recorded engagement and does not model no-action decisions.¹

1 Introduction

Online public opinion develops through repeated user interactions: people comment, like, repost, choose whom to address, and express stances within evolving discussion threads. When aggregated, these micro-level decisions shape observable stance distributions, interaction networks, attention concentration, and information diffusion. Simulating this process offers a controlled setting for studying how public opinion may evolve under different discussion contexts (Hu et al. 2024; Liu et al. 2025; Jeon et al. 2025; Cheng et al. 2025). LLM agents offer a flexible mechanism for such simulations, but the credibility of aggregate results depends in part on whether individual behaviors are grounded in heterogeneous real-user evidence (Argyle et al. 2023; Gao et al. 2024). Fluent yet generic or uniformly active agents can distort action frequencies, target selection, and stance composition when

¹ All code, checkpoints, and data will be made publicly available immediately upon acceptance.

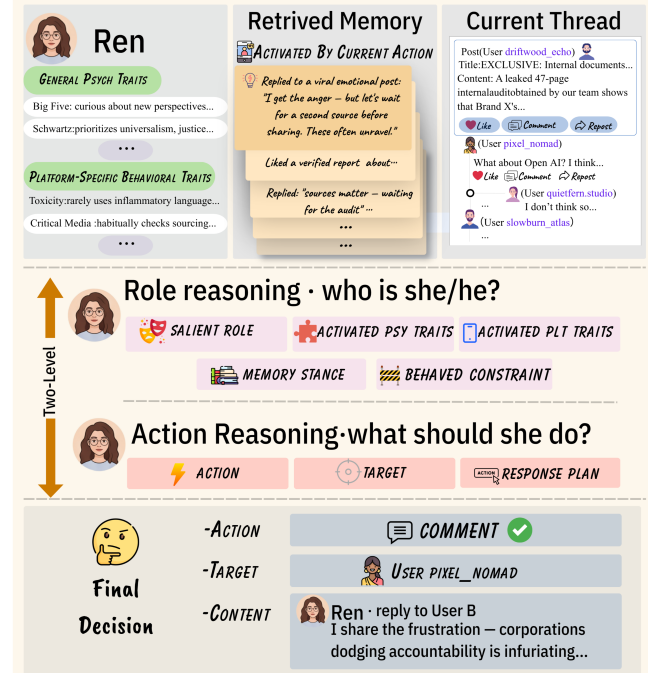


Figure 1: Motivation example of identity-driven social-media user simulation. Given Ren’s profile, retrieved memory, and current thread, the simulator first activates profile and memory evidence relevant to the thread, then selects the target before predicting the action and response.

their outputs are aggregated. We therefore study user-level target–action–response prediction as a behavioral foundation for evaluating thread-level public-opinion simulation in observed social-media contexts.

Conventional character role-play usually starts from an explicit persona and evaluates whether generated dialogue remains persona-consistent (Xu et al. 2026b; Xiao et al. 2026; Cheng et al. 2026). Social-media user simulation poses a different task: it infers observable platform behavior from a user’s partial and noisy behavioral history together with the current thread context. The prediction covers both textual responses and structured decisions about target selection and the subsequent action type. We formulate the task as predicting a user’s observed engagement in a given post–comment

context, including the target, subsequent action type, and optional textual response. We normalize observed interactions into COMMENT, LIKE, and REPOST². The logs record positive interactions but not impression-level exposure. Accordingly, the benchmark is conditional on a recorded engagement event, contains no explicit no-action label, and cannot distinguish unseen threads from threads that a user viewed but ignored. Within this scope, the task predicts the target, subsequent action, and optional text.

Existing LLM-based social simulation systems have explored information diffusion, online conversation, and social media event modeling (Park et al. 2023; Gao et al. 2024; Hu et al. 2024; Liu et al. 2025; Jeon et al. 2025; Ji et al. 2026). These studies show that LLM agents can generate natural-language interactions and support controlled simulation of social processes. However, they often rely on coarse personas, demographic attributes, or manually specified agent descriptions, and are less focused on predicting a concrete user’s platform behavior from long-horizon behavioral histories and situated thread context.

To address this, we introduce **PsyMemBehav**, a structured framework that predicts social media behavior by integrating user profiles, longitudinal memories, and thread context. The framework maps this evidence to behavior through two sequential steps: identity activation, which selects relevant traits and memories for the current thread, and action decision, which identifies the target and predicts the subsequent action and response. Furthermore, we align the complete reasoning-action trajectories via rubric-guided preference optimization. This ensures the simulator optimizes not just for response fluency, but also for action fidelity, memory grounding, and behavioral consistency.

To evaluate this, we construct a large-scale dataset PsyMemBehav-bench, containing 5.57M behaviors from 5,862 users across Reddit, Bluesky, and Twitter. Each sample of PsyMemBehav-bench integrates 41 psychology-informed descriptors, retrieved history, context, and observed behavior. This enables the joint evaluation of structured decision-making and text generation, moving beyond traditional text-only role-playing.

We evaluate PsyMemBehav at the individual level by matching held-out user interactions and at the thread level by comparing aggregate comment-section statistics in observed post contexts. In summary, our contributions are as follows:

- 1) We propose **PsyMemBehav**, a novel framework that models conditional social-media behavior through a two-stage structured decision process: *identity activation* and *action decision*. By dynamically integrating 41-dimensional psychology-informed user profiles and retrieved longitudinal memories with thread context, our approach moves beyond flat text generation to produce grounded, context-aware behavioral predictions.

- 2) We introduce a tailored alignment strategy combining Reverse-CoT supervision and rubric-guided DPO to optimize complete reasoning-action trajectories. This ensures

²Textual interactions such as replies and comments are mapped to COMMENT; resharing behaviors are mapped to REPOST; and reaction-only behaviors are mapped to LIKE.

the simulator is optimized not merely for response fluency, but critically for action fidelity (e.g., predicting COMMENT, LIKE, or REPOST with conditional text generation), robust memory grounding, and strict persona consistency.

- 3) On the large-scale dataset PsyMemBehav-bench, we conduct comprehensive evaluations on both held-out user behavior and observed-thread continuation. The resulting PsyMemBehav-27B ranks first in 11 of 12 individual metrics, achieving a 9.95% relative improvement in average response quality over GPT-5.4, while yielding the lowest aggregate group-level deviation. Extensive ablation studies further validate the critical contributions of both training stages, profile dimensions, and decision components.

2 Related Work

2.1 Social-Media Account Simulation

LLM-based agents have been used to simulate human-like behavior in role-playing, interactive environments, and social settings (Park et al. 2023; Argyle et al. 2023; Gao et al. 2024). Prior work shows that LLMs can maintain personas, imitate conversational styles, and produce role-consistent responses (Shao et al. 2023; Wang et al. 2024; Tu et al. 2024). In parallel, user modeling studies show that digital traces reveal stable attributes such as personality, preferences, demographics, and behavioral tendencies (Kosinski, Stillwell, and Graepel 2013; Youyou, Kosinski, and Stillwell 2015; Hovy and Søgaard 2015), motivating personalized behavior generation from historical traces.

Social-media account simulation, however, goes beyond generic persona imitation. Users do not only generate text, but also select interaction targets within a thread and choose platform-specific actions such as replying, quoting, reposting, liking, or ignoring. Although recent work explores LLM agents for interactive social behavior and social simulation (Ren et al. 2024; Zhou et al. 2024), existing approaches often focus on generic personas, aggregate populations, or domain-specific simulated users rather than real accounts grounded in historical traces and platform-specific action spaces. Our work addresses this gap by modeling real accounts with psychological traits, platform-specific behavioral traits, retrieved memory, and thread context.

2.2 Rubric-Guided Alignment for Social-Behavior Simulation

Preference learning aligns LLMs with desired behaviors through human or AI feedback and direct preference optimization (Ouyang et al. 2022; Bai et al. 2022; Lee et al. 2023; Rafailov et al. 2023), typically targeting general objectives such as helpfulness, harmlessness, or instruction following. Social-media simulation instead requires faithfulness to user profiles, grounding in memory, platform-appropriate actions, thread-context coherence, and social plausibility.

LLM-as-a-judge methods enable evaluation with explicit criteria rather than a single overall score (Zheng et al. 2023; Liu et al. 2023; Kim et al. 2024). Building on this idea, we construct rubric-guided preference pairs over complete reasoning-action trajectories, aligning both intermediate rea-

soning and resulting platform behavior rather than only final textual responses.

2.3 Group-Level Thread Simulation

Agent-based modeling studies how individual behaviors produce collective phenomena such as diffusion, coordination, polarization, and network formation (Epstein 1999; Macy and Willer 2002; Lazer et al. 2009). LLM agents extend this tradition with language-based reasoning and context-sensitive interaction policies (Park et al. 2023; Gao et al. 2024), while introducing challenges such as unstable identities, unrealistic information access, and weak behavioral grounding.

Recent work argues that social simulation should be evaluated beyond single-turn plausibility, including interaction dynamics, stance distributions, network structures, conversational diversity, and thread-level realism (Zhou et al. 2024; Zhang et al. 2025a). We therefore first improve account-level identity-to-action modeling through psychology-informed reasoning and rubric-guided DPO, and then evaluate whether these improvements transfer to group-level thread simulations.

3 Methodology

3.1 Problem Formulation and Overview

We formulate social-media user simulation as a time-anchored multi-action prediction problem. For a user u at time t , the simulator observes

$$x_t = (P_u, M_{u,t}, T_t, C_t),$$

where P_u is the user profile, $M_{u,t}$ is retrieved behavioral memory before time t , T_t is the current thread context, and C_t is the set of candidate targets, i.e. the visible post or comments the user may interact with at t . The simulator predicts a complete platform behavior

$$y_t = (c_t, a_t, r_t),$$

where $c_t \in C_t$ is the selected target, $a_t \in \mathcal{A}$ the subsequent action, and r_t the generated comment when required. We normalize the heterogeneous behaviors in raw logs into three visible actions, $\mathcal{A} = \{\text{COMMENT}, \text{LIKE}, \text{REPOST}\}$: textual engagement (comments, replies) maps to COMMENT, reaction-only behavior to LIKE, and resharing (including quote reposts) to REPOST; original posts and past interactions serve as evidence rather than prediction labels.

This requires more than fluent response generation: conditional on a recorded engagement event, the simulator must first select a target, then choose Comment, Like, or Repost and generate text when required. Because behavior is context-dependent, a plausible comment can still be wrong if the user would target a different item or act differently. We address this with two design choices. First, *two-level identity-to-action reasoning* situates the user in the current thread before deciding the platform behavior (§3.2). Second, *two-stage trajectory training* optimizes complete reasoning-action trajectories rather than only final responses, through reverse-CoT supervision and rubric-guided preference optimization (§3.3). Figure 2 gives an overview.

3.2 Two-Level Identity-to-Action Reasoning

Each simulation instance is modeled as two ordered decisions: the evidence relevant to the user in the current thread is identified first, and the corresponding platform behavior is predicted second. This separation reflects the difference between a broad, history-derived profile and a context-specific action, while making the intermediate evidence selection explicit. A direct mapping from x_t to y_t remains possible, but it does not expose separate supervision for evidence selection and action decision. We therefore insert an intermediate reasoning trace

$$x_t \rightarrow z_t \rightarrow y_t, \quad z_t = (z_t^{\text{id}}, z_t^{\text{act}}),$$

with identity activation z_t^{id} (which profile and memory evidence is relevant here) and action decision z_t^{act} (which target, action, and response follow from that evidence).

Psychology-informed representation. We organize user evidence into two complementary groups:

$$P_u = (P_u^{\text{gen}}, P_u^{\text{plat}}),$$

where P_u^{gen} contains 22 dimensions spanning Big Five personality, Schwartz values, moral foundations, and social dominance orientation (Goldberg 1990; Schwartz 1992; Graham et al. 2013; Pratto et al. 1994), while P_u^{plat} contains 19 dimensions spanning emotional tendencies, network and social capital, media literacy, and online behavior (Katz and Lazarsfeld 1955; Ellison, Steinfield, and Lampe 2007; Guess et al. 2020; Wulczyn, Thain, and Dixon 2017). These 41 variables are LLM-inferred, history-derived descriptors for behavior prediction, not validated psychometric or clinical measurements. General traits encode cross-context dispositions, whereas platform traits encode online interaction tendencies, allowing identity activation to select the evidence relevant to a specific thread. Profiles are built once from each user’s pre-split history, $P_u = g_{\text{prof}}(\mathcal{H}_u^{\text{train}})$, and reused for held-out instances to prevent post-split leakage. The complete taxonomy and annotation prompt are in the appendix.

Memory retrieval. We use the current visible thread as a lexical query, rank each strictly pre-time historical behavior with BM25 (Robertson and Zaragoza 2009), and retain the top five as $M_{u,t}$. The score uses neither dense similarity nor an additional recency prior. In the retrieval-size analysis, both 9B and 27B models peak at $k = 5$; fewer memories omit relevant history, whereas more introduce redundancy or noise (see the appendix).

Identity activation and target-first action decision. Given the profile and retrieved memory, *identity activation* selects and summarizes the profile traits, historical behaviors, stance evidence, and behavioral constraints relevant to the current thread. Conditioned on this evidence summary, *action decision* selects the target before predicting the action and response plan. The decomposition provides explicit supervision for distinguishing COMMENT, LIKE, and REPOST instead of treating every interaction as text generation. Together they yield the complete trajectory

$$\tau_t = (z_t, y_t) = (z_t^{\text{id}}, z_t^{\text{act}}, c_t, a_t, r_t).$$

The full schema and a worked example are in the appendix.

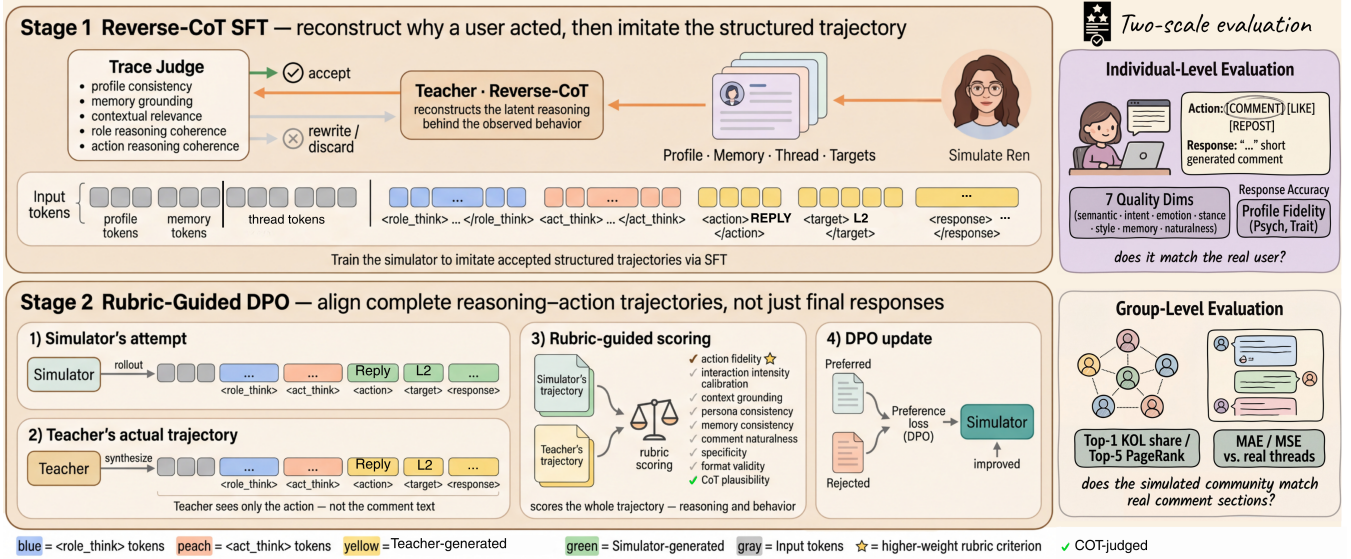


Figure 2: Overview of the proposed two-stage training framework. Stage 1 reconstructs reverse-CoT traces from observed behavior and trains the simulator by supervised fine-tuning on accepted trajectories. Stage 2 applies rubric-guided DPO to align complete reasoning-action trajectories with social-behavior criteria. The right panel summarizes the individual- and group-level evaluation.

3.3 Two-Stage Training

We train over the trajectory τ_i in two stages (Figure 2): supervised fine-tuning on reconstructed traces, then rubric-guided preference optimization over complete trajectories.

Stage 1: Reverse-CoT supervision. Behavior logs provide observed actions and comments but no explicit annotation of the underlying decision process. Direct output supervision therefore leaves the intermediate mapping from user evidence to platform behavior implicit. We reconstruct the missing reasoning after the fact: given (x_t, y_t^*) , a teacher produces a posterior rationale constrained by the observed behavior, $z_t^* = f_{\text{teacher}}(x_t, y_t^*)$. Since z_t^* is generated after the behavior, we treat it not as ground-truth cognition but as structured supervision for learning the mapping from user evidence to observed behavior. Following behavior-conditioned user modeling paradigms (Zhang et al. 2025b), the reconstructed trace connects the user profile, retrieved memory, and thread context with the observed action. It is therefore intended to provide a plausible and behavior-consistent reasoning structure, rather than to recover a unique latent cognitive process. The observed behavior is used only during trace construction and is unavailable at inference time. A validation judge J_{trace} (checking profile consistency, memory grounding, contextual relevance, identity-activation coherence, and action-decision coherence; Figure 2) keeps only consistent traces,

$\mathcal{D}_{\text{SFT}} = \{(x_t, z_t^*, y_t^*) \mid J_{\text{trace}}(x_t, z_t^*, y_t^*) = 1\}_{t=1}^N$, and we minimize the negative log-likelihood of the full structured output,

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x, z^*, y^*) \sim \mathcal{D}_{\text{SFT}}} \log \pi_{\theta}(z^*, y^* \mid x),$$

initializing the simulator to emit identity activation, action decision, and behavior in one trajectory. Judge criteria, the form of J_{trace} , and prompts are in the appendix.

Stage 2: Rubric-guided trajectory DPO. SFT yields well-formed trajectories but does not say which of several fluent ones is the more faithful simulation. We therefore optimize preferences over complete trajectories. The rubric itself is induced rather than hand-specified. We sample trajectory pairs and ask a judge to write free-form critiques explaining why one trajectory is a less faithful simulation than the other; an LLM then consolidates the full set of critiques into ten criteria that together account for the observed weaknesses, which the human annotators verify for coverage and non-redundancy. Grounding the evaluation axes in the weaknesses simulators actually exhibit, rather than in a priori intuitions, makes the rubric an explanatory cover of the failures rather than an arbitrary checklist. Once induced, the rubric is frozen and applied uniformly during preference optimization. For each input we compare a trajectory rolled out by the SFT model against a reference synthesized from the observed behavior, scoring both with a fixed ten-criterion social-behavior rubric (action fidelity, interaction-intensity calibration, context grounding, persona/memory/stance consistency, comment naturalness, specificity, format validity, and evidence-to-action coherence; full definitions in the appendix). Two choices matter. First, since action selection is central, action fidelity is up-weighted,

$$S(\tau) = \frac{2s_1(\tau) + \sum_{j=2}^{10} s_j(\tau)}{11},$$

so the action is never outweighed by surface fluency. Second, the judge sees only the ground-truth action label a_t^* , not the gold comment, forcing it to assess whether each trajectory chooses an appropriate action and uses evidence faithfully rather than matching reference wording. To avoid spurious supervision, we retain a pair only when the judge is confident, the score margin exceeds a threshold, the trajectories are not near-duplicates, and the preference is not explained by length.

Algorithm 1: Trajectory Preference Construction

Require: input x_i ; SFT policy π_{SFT} ; observed behavior y_i^* with action label a_i^* ; rubric score $S(\cdot)$

- 1: $\tau_i^{\text{gen}} \sim \pi_{\text{SFT}}(\cdot | x_i)$ {simulator rollout}
- 2: $\tau_i^{\text{ref}} \leftarrow (z_i^{\text{ref}}, y_i^*)$ {synthesize reference from observed behavior}
- 3: score $\tau_i^{\text{gen}}, \tau_i^{\text{ref}}$ with $S(\cdot)$, exposing a_i^* to the judge but *not* the gold comment
- 4: $(\tau_i^+, \tau_i^-) \leftarrow$ higher- / lower-scoring trajectory
- 5: **if** pair passes confidence, margin, duplication, and length filters **then**
- 6: add $(x_i, \tau_i^+, \tau_i^-)$ to $\mathcal{D}_{\text{DPO}}$
- 7: **end if**

Algorithm 1 summarizes the full construction.

Given the resulting $\mathcal{D}_{\text{DPO}} = \{(x_i, \tau_i^+, \tau_i^-)\}_{i=1}^M$, we optimize a DPO objective (Rafailov et al. 2023) over *complete* trajectories: the preference margin scores the full $(z_t^{\text{id}}, z_t^{\text{act}}, c_t, a_t, r_t)$ under π_θ relative to the SFT reference π_{ref} with strength β , aligning both the intermediate reasoning and the platform behavior with the rubric. The complete rubric, trace-validation criteria, and judging prompts are in the appendix.

4 Experimental Setup

To evaluate the effectiveness of the proposed PsyMemBehav framework, we conduct experiments on the social-media user simulation task defined in Section 3. The model first selects the target, then predicts COMMENT, LIKE, or REPOST and any required comment text.

4.1 Base Models

We conduct experiments on two backbone models: Qwen3.5-9B and Qwen3.5-72B. The 9B model is used to examine whether the proposed framework remains effective under a smaller simulator, while the 72B model serves as the stronger backbone for our main experiments. Both models are trained under the same two-stage pipeline, resulting in two models: PsyMemBehav 9B and PsyMemBehav 72B.

4.2 Baselines

We compare PsyMemBehav 9B and PsyMemBehav 72B against the following baselines: GPT-5.4 (OpenAI 2026), Gemini-2.5-Pro (Comanici et al. 2025), Claude-Sonnet-4.6 (Anthropic 2025), Qwen3.7-Max (Yang et al. 2025), and DeepSeek-V4-Pro (Xu et al. 2026a). Additionally, Qwen3.5-9B and Qwen3.5-72B (Yang et al. 2025) are included as open-weight baselines. All models receive the same full user profile, BM25-retrieved pre-event memories, thread context, and candidate targets. They use the unified end-to-end prompt in the appendix, which selects a target before predicting the action and branch-specific text; gold target and action labels are withheld.

4.3 Evaluation Metrics

We evaluate the models at both individual and group levels.

Individual-level evaluation. We report two behavior-level metrics. **Target Accuracy** measures whether the model selects the correct interaction target among the candidate posts and comments. **Action Accuracy** measures whether the model chooses the correct platform action—COMMENT, LIKE, or REPOST—for the selected target.

For generated comments, we use Gemini-3-Flash as an automated judge to score seven quality dimensions: semantic alignment, intent alignment, emotional consistency, stance consistency, style fidelity, persona memory, and naturalness. We report both individual dimension scores and their average. The complete scoring prompt appears in the appendix.

Group-level simulation. To evaluate multi-user thread continuation, we place profiled simulators in the same observed thread seed and roll out their target-first actions. Following prior work that evaluates collective simulations through interaction, stance, and influence structure (Zhou et al. 2024; Liu et al. 2025), we report four complementary metrics. *Reply Reciprocity* is the share of interacting dyads with mutual directed replies, a standard property of directed networks (Garlaschelli and Loffredo 2004). *Stance deviation* is the ℓ_1 distance between simulated and real PRO, ANTI, and NEUTRAL distributions, using the conventional favor/against/neither stance formulation (Mohammad et al. 2016). *Top1 KOL* is the reply share received by the most-addressed user, operationalizing key-opinion-leader concentration (Katz and Lazarsfeld 1955); *Top5 PageRank* is the normalized PageRank mass of the five highest-ranked users (Brin and Page 1998). MAE and MSE summarize absolute and squared deviations across thread-level statistics. Detailed formulas are in the appendix.

4.4 Dataset and Data Split

We construct PsyMemBehav-bench from Reddit, Bluesky, and Twitter through official APIs. It contains 5.57M behaviors from 5,862 real users (approximately 950 per user), with usernames replaced by randomized hash identifiers. Platform events are normalized as COMMENT, LIKE, or REPOST. Each instance combines a two-group profile, five BM25-retrieved pre-event memories, thread context, and candidate targets with the observed target, action, and response. A user-disjoint chronological split assigns 5,000 users to training and 862 to validation and testing; held-out users are unseen, and all profile and memory evidence predates prediction. Collection favors accounts with rich observable histories. Thus, the scale denotes longitudinal density rather than platform-wide representativeness and may overrepresent active users while underrepresenting low-activity, private, or new users. Among closely related studies reporting real-user counts, FineRob uses 1,866 users and 78.6K records (approximately 126 histories per user), whereas MOSAIC uses 204 real users and 30 posts and GA-S³ models group agents over 30 events. PsyMemBehav therefore provides larger and denser longitudinal evidence for user-level target-action-response modeling, not a claim of platform-wide coverage (see the appendix).

Models	Behavior & Response Quality										Profile Fidelity	
	Act.	Tgt.	Sem.	Intent	Emo.	Stance	Style	Memory	Natural.	Avg.	Psych	Trait
GPT-oss-120B	54.63	37.35	34.49	41.50	43.39	51.39	37.70	52.14	45.95	42.92	73.12	73.78
GPT-5	61.80	58.80	46.99	54.44	54.22	65.12	46.36	60.79	58.57	54.12	76.52	75.23
Gemini-2.5-Pro	63.43	49.28	41.43	44.45	45.97	58.07	42.83	54.91	61.28	49.84	72.86	71.85
GPT-5.4	66.12	59.46	45.05	54.21	56.36	69.25	45.49	58.19	62.94	54.28	76.59	75.11
Claude-Sonnet-4.6	63.67	57.23	41.57	49.83	51.84	62.62	43.61	55.58	60.40	50.75	77.09	71.39
Qwen3.7-Max Preview	62.02	54.01	40.85	49.47	52.16	63.76	45.18	58.98	56.60	51.16	74.73	72.87
Gemini-3.1-Pro-Preview	62.81	51.61	37.53	46.26	50.24	60.77	48.12	57.66	60.02	50.26	75.91	73.72
DeepSeek-V4-Pro	64.32	47.01	36.62	45.15	47.36	59.94	40.70	52.96	55.19	46.86	75.63	71.35
Qwen3.5-9B	51.61	38.79	26.89	33.42	38.95	48.28	33.09	39.55	36.12	35.60	70.77	70.25
PsyMemBehav 9B	59.52	48.22	36.87	46.58	48.71	58.54	40.25	51.64	58.90	48.78	75.33	74.43
Qwen3.5-27B	56.57	47.90	33.87	39.84	42.23	49.43	40.30	54.52	49.62	44.26	72.23	71.02
PsyMemBehav 27B	71.40	65.34	50.35	55.49	58.89	68.89	53.57	64.19	66.39	59.68	80.38	78.25

Table 1: Individual-level user simulation results. Blue rows denote our proposed PsyMemBehav models. **Act.** denotes action accuracy over COMMENT, LIKE, and REPOST; **Tgt.** denotes target accuracy. The seven response-quality dimensions are judged by Gemini-3-Flash. **Psych** and **Trait** are profile-fidelity scores computed by re-annotating generated behaviors and comparing them with the original user profile using $100 \times (1 - \text{NED})$. Higher is better for all metrics. Best results are bolded.

Dimension	Human mean	Gemini	Human–Gemini
Semantic	44.22	41.98	+2.24
Intent	48.46	50.84	−2.38
Emotion	50.94	54.54	−3.60
Stance	52.31	56.51	−4.20
Style	55.79	54.38	+1.41
Profile/memory	61.29	65.02	−3.73
Naturalness	67.77	64.92	+2.85
Overall	54.40	55.46	−1.06

Table 2: Human and Gemini mean scores on 150 instances from 30 anonymized user IDs (five per ID); differences are human minus Gemini.

Variant	Act.	Tgt.	Avg.	Psych	Trait
<i>Training ablation</i>					
Base	51.61	38.79	35.60	70.77	70.25
CoT-P	47.31	35.90	33.92	69.69	68.81
Out-SFT	49.23	40.30	39.70	71.73	70.34
R-CoT	53.61	46.64	43.37	74.58	75.79
R-CoT+R-DPO	59.52	48.22	48.78	75.33	74.43
<i>Component ablation under R-CoT</i>					
Full	53.61	46.64	43.37	74.58	75.79
w/o Gen.	52.48	43.77	39.66	70.71	74.69
w/o Plat.	52.39	42.32	41.13	73.11	71.45
w/o IdAct	51.27	41.03	38.77	69.43	72.50
w/o ActDec	50.47	40.62	39.37	69.89	71.83

Table 3: Ablation results on Qwen3.5-9B. Act. is action accuracy; Tgt. is target accuracy; Avg. is the average score over seven response-quality dimensions. Psych and Trait are profile-fidelity scores for general psychological traits and platform-specific traits.

stance consistency: its 68.89 is 0.36 points below GPT-5.4’s 69.25. Gains therefore span decisions, responses, and profile preservation.

Evaluation robustness. We score the same anonymized outputs with Qwen3.5-27B, GPT-5.4, Gemini-2.5-Pro, and Claude-Sonnet-4.6 as independent judges. PsyMemBehav 27B ranks first under all four, exceeding the runner-up by 7.73, 3.03, 4.90, and 2.83 points, respectively. Agreement across judge families reduces the likelihood that the main ranking is driven by one evaluator’s self-preference, although it does not by itself establish human validity. Full cross-judge scores are reported in the appendix.

Table 2 calibrates the automated evaluator against human judgments on 150 instances from 30 anonymized user IDs (five per ID). The overall means differ by 1.06 points (54.40 vs. 55.46), and the signed dimension gaps are mixed rather than uniformly favoring one evaluator. This close aggregate

5 Experimental Results and Analyses

5.1 Individual-Level Results

Table 1 reports target and action accuracy, seven response-quality dimensions, and profile fidelity, computed by re-annotating generated behavior and applying $100 \times (1 - \text{NED})$ to the original profile (see the appendix).

PsyMemBehav 27B ranks first in **11 of 12 metrics**, with target accuracy of 65.34, action accuracy of 71.40, average response quality of 59.68, and profile-fidelity scores of 80.38 (Psych) and 78.25 (Trait). Relative to GPT-5.4, its average response quality rises from 54.28 to 59.68 (9.95% relative), alongside higher target and action accuracy.

Training improves both backbones consistently: PsyMemBehav 9B raises average response quality from 35.60 to 48.78 over Qwen3.5-9B, and PsyMemBehav 27B raises it from 44.26 to 59.68 over Qwen3.5-27B, with concurrent gains in target accuracy, action accuracy, and profile fidelity.

At 27B, the complete training pipeline improves action accuracy by 14.83 points, target accuracy by 17.44, average response quality by 15.42, Psych fidelity by 8.15, and Trait fidelity by 7.23 over the matched Qwen3.5 backbone. At 9B, the corresponding gains are 7.91, 9.43, 13.18, 4.56, and 4.18 points. The only metric not led by PsyMemBehav 27B is

Model	Reciprocity	Stance Dev. ↓	Top1 KOL	Top5 PR	MAE ↓	MSE ↓
Real	0.043	0.000	0.213	0.596	0.000	0.000
PsyMemBehav 27B	0.121	0.155	0.232	0.524	0.081	0.009
GPT-5.4	0.182	0.175	0.197	0.529	0.099	0.014
Claude-Sonnet-4.6	0.313	0.322	0.191	0.472	0.185	0.048
GPT-5	0.375	0.496	0.226	0.507	0.233	0.091
Qwen3.7-Max-Preview	0.458	0.452	0.131	0.483	0.266	0.099
DeepSeek-V4-Pro	0.444	0.649	0.136	0.448	0.319	0.152

Table 4: Thread-level continuation results. We use **PsyMemBehav 27B** as our simulator and compare its generated interaction threads with those produced by strong general models. **Real** denotes statistics computed from real comment sections. **Reciprocity** measures bidirectional reply interactions. **Stance Dev.** measures deviation from the real stance distribution. **Top1 KOL** and **Top5 PR** measure the concentration of influence in the simulated interaction graph. **MAE** and **MSE** summarize the overall deviation from real group-level statistics. Lower is better for metrics marked with ↓; for other metrics, closer to the real distribution indicates better simulation fidelity.

alignment supports model-level comparison, while the stance (4.20) and emotion (3.60) gaps motivate dimension-specific caution; full analysis appears in Appendix.

5.2 Ablation Study

Table 3 separates training ablations from component ablations on Qwen3.5-9B. *R-CoT/CoT-SFT* is supervised fine-tuning on accepted reverse-CoT traces without DPO, whereas *R-DPO* is rubric-guided trajectory optimization.

Naive CoT prompting degrades performance, and output-only SFT improves response quality without improving action prediction. R-CoT instead improves target accuracy, action accuracy, and profile fidelity; adding R-DPO further improves most behavior and response metrics. Relative to Base, the complete **R-CoT+R-DPO** pipeline raises target accuracy from 38.79 to **48.22**, action accuracy from 51.61 to **59.52**, and average response quality from 35.60 to **48.78**.

The two stages contribute differently. Relative to Base, R-CoT adds 7.85 points in target accuracy, 2.00 in action accuracy, 7.77 in average response quality, 3.81 in Psych fidelity, and 5.54 in Trait fidelity. Adding R-DPO contributes a further 1.58, 5.91, 5.41, and 0.75 points to target accuracy, action accuracy, response quality, and Psych fidelity, while Trait fidelity decreases by 1.36. Thus, reverse-CoT accounts for most target and trait gains; DPO mainly strengthens action selection and response quality despite a small trait-preservation trade-off.

Removing either profile group degrades performance, confirming that general psychological and platform-specific evidence are complementary. Removing identity activation mainly reduces response quality, whereas removing action decision produces the lowest target and action accuracy, supporting their respective roles in evidence selection and target-first behavior prediction.

The component losses also reveal specialization. Removing identity activation causes the largest drop in average response quality (−4.60) and Psych fidelity (−5.15), whereas removing action decision causes the largest losses in target accuracy (−6.02) and action accuracy (−3.14). Platform-specific traits contribute more to target accuracy (−4.32) and Trait fidelity (−4.34) than general traits (−2.87 and −1.10), while general traits contribute more to response qual-

ity (−3.71 versus −2.24). This pattern matches the intended division of labor among evidence and reasoning components.

5.3 Group-Level Thread Simulation

Table 4 compares PsyMemBehav 27B with strong general models on continuations of the same observed thread contexts. **PsyMemBehav 27B** obtains the **lowest stance deviation (0.155)**, **MAE (0.081)**, and **MSE (0.009)**, while its reciprocity (0.121) is closest to the real value (0.043). Its lower reciprocity error also avoids the over-interactive group-chat pattern produced by general-purpose agents. Relative to GPT-5.4, PsyMemBehav reduces stance deviation by 11.4%, MAE by 18.2%, and MSE by 35.7%. GPT-5 and GPT-5.4 remain closer on Top1 KOL and Top5 PageRank, respectively, so PsyMemBehav does not lead every structural statistic. In direct pairwise evaluation, **PsyMemBehav** is preferred over GPT-5.4 in **64.0%** of thread continuations versus 34.0%, with 2.0% ties (Appendix). These results concern continuation within observed threads, not platform-wide social dynamics.

PsyMemBehav’s absolute reciprocity gap from the real threads is 0.078, compared with 0.139 for GPT-5.4, a 43.9% reduction. For influence concentration, however, its Top1 KOL gap is 0.019 versus 0.013 for GPT-5, and its Top5 PageRank gap is 0.072 versus 0.067 for GPT-5.4. The strongest improvements therefore concern interaction reciprocity, stance composition, and aggregate fidelity, while centralization remains model-dependent.

6 Conclusion and Limitations

We present PsyMemBehav, combining 41-dimensional user descriptors, BM25 memories, target-first reasoning, reverse-CoT supervision, and rubric-guided DPO. PsyMemBehav 27B achieves the strongest individual results and lowest aggregate deviation in observed-thread continuation, supporting evidence-grounded user models for micro-level public-opinion simulation. Limitations include missing exposure/no-action labels, LLM-inferred rather than self-reported profiles, answer-conditioned reasoning traces, and aggregate rather than per-instance human validation. Future work should address these issues across broader settings.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62206259).

References

- Anthropic. 2025. Claude Sonnet 4.6.
- Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.
- Brin, S.; and Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1–7): 107–117.
- Cheng, X.; Qin, Y.; Tan, Y.; Li, Z.; Wang, Y.; Xiao, H.; and Zhang, Y. 2026. PsyMem: Fine-grained Psychological Alignment and Explicit Memory Control for Advanced Role-Playing LLMs. *Transactions of the Association for Computational Linguistics*, 14: 510–529.
- Cheng, Z.; Lin, Z.; Yang, Y.; Wei, Z.; and Chen, S. 2025. Interactive Simulation and Visual Analysis of Social Media Event Dynamics with LLM-based Multi-Agent Modeling. *Visual Informatics*, 9(3): 100260.
- Comanici, G.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.
- Ellison, N. B.; Steinfield, C.; and Lampe, C. 2007. The Benefits of Facebook “Friends:” Social Capital and College Students’ Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4): 1143–1168.
- Epstein, J. M. 1999. Agent-Based Computational Models and Generative Social Science. *Complexity*, 4(5): 41–60.
- Gao, C.; Lan, X.; Li, N.; Yuan, Y.; Ding, J.; Zhou, Z.; Xu, F.; and Li, Y. 2024. Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives. *Humanities and Social Sciences Communications*, 11(1): 1–24.
- Garlaschelli, D.; and Loffredo, M. I. 2004. Patterns of Link Reciprocity in Directed Networks. *Physical Review Letters*, 93(26): 268701.
- Goldberg, L. R. 1990. An Alternative “Description of Personality”: The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, 59(6): 1216–1229.
- Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*, volume 47, 55–130. Academic Press.
- Guess, A. M.; Lerner, M.; Lyons, B.; Montgomery, J. M.; Nyhan, B.; Reifler, J.; and Sircar, N. 2020. A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27): 15536–15545.
- Hovy, D.; and Søgaard, A. 2015. Tagging Performance Correlates with Author Age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 483–488. Beijing, China: Association for Computational Linguistics.
- Hu, Y.; Sherpa, G.; Zhang, L.; Li, W.; Bai, Q.; Wang, Y.; and Wang, X. 2024. An LLM-enhanced Agent-based Simulation Tool for Information Propagation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 8679–8682. International Joint Conferences on Artificial Intelligence Organization. Demo Track.
- Jeon, M. S.; Mendoza, M.; Fernández, M.; Providel, E.; Rodríguez, F.; Espina, N.; Carvallo, A.; and Abeliuk, A. 2025. Simulating Conversations on Social Media with Generative Agent-Based Models. *EPJ Data Science*, 14(79).
- Ji, J.; Li, L.; Cheng, X.; Zhou, Y.; Zhang, W.; Tan, Y.; and Qin, Y. 2026. RippleMem: From Isolated Retrieval to Associative Recollection for Long-Term Agent Memory. *arXiv preprint arXiv:2608.13334*.
- Katz, E.; and Lazarsfeld, P. F. 1955. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Glencoe, IL: Free Press.
- Kim, S.; Shin, J.; Cho, Y.; Jang, J.; Longpre, S.; Lee, H.; Yun, S.; Shin, S.; Kim, S.; Thorne, J.; and Seo, M. 2024. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*.
- Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private Traits and Attributes Are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences*, 110(15): 5802–5805.
- Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; and Van Alstyne, M. 2009. Computational Social Science. *Science*, 323(5915): 721–723.
- Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; and Prakash, S. 2023. RLAIIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv preprint arXiv:2309.00267*.
- Li, K.; Dai, C.; Zhou, W.; and Hu, S. 2025. LLM-Based Behavior Prediction for Social Media Users with Continuous Memory. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 459–474. Mumbai, India: Asian Federation of Natural Language Processing and Association for Computational Linguistics.
- Liu, G.; Le, V. T.; Rahman, S.; Kreiss, E.; Ghassemi, M.; and Gabriel, S. 2025. MOSAIC: Modeling Social AI for Content Dissemination and Regulation in Multi-Agent Simulations.

- In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 6390–6417. Suzhou, China: Association for Computational Linguistics.
- Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.
- Macy, M. W.; and Willer, R. 2002. From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28: 143–166.
- Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. San Diego, California: Association for Computational Linguistics.
- OpenAI. 2026. Introducing GPT-5.4.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744.
- Park, J. S.; O’Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST ’23. New York, NY, USA: Association for Computing Machinery.
- Pratto, F.; Sidanius, J.; Stallworth, L. M.; and Malle, B. F. 1994. Social Dominance Orientation: A Personality Variable Predicting Social and Political Attitudes. *Journal of Personality and Social Psychology*, 67(4): 741–763.
- Qiu, Z.; Lyu, H.; Xiong, W.; and Luo, J. 2025. Can LLMs Simulate Social Media Engagement? A Study on Action-Guided Response Generation. *arXiv preprint arXiv:2502.12073*.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, 53728–53741.
- Ren, R.; Qiu, P.; Qu, Y.; Liu, J.; Zhao, W. X.; Wu, H.; Wen, J.-R.; and Wang, H. 2024. BASES: Large-scale Web Search User Simulation with Large Language Model based Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 902–917. Miami, Florida, USA: Association for Computational Linguistics.
- Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*.
- Schwartz, S. H. 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In *Advances in Experimental Social Psychology*, volume 25, 1–65. Academic Press.
- Shao, Y.; Li, L.; Dai, J.; and Qiu, X. 2023. Character-LLM: A Trainable Agent for Role-Playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13153–13187. Singapore: Association for Computational Linguistics.
- Tu, Q.; Fan, S.; Tian, Z.; Shen, T.; Shang, S.; Gao, X.; and Yan, R. 2024. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11836–11850. Bangkok, Thailand: Association for Computational Linguistics.
- Wang, N.; Peng, Z. Y.; Que, H.; Liu, J.; Zhou, W.; Wu, Y.; Guo, H.; Gan, R.; Ni, Z.; Yang, J.; Zhang, M.; Zhang, Z.; Ouyang, W.; Xu, K.; Huang, W.; Fu, J.; and Peng, J. 2024. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 14743–14777. Bangkok, Thailand: Association for Computational Linguistics.
- Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399. International World Wide Web Conferences Steering Committee.
- Xiao, Z.; Li, M.; Wang, X.; Li, L.; Shui, L.; Ji, M.; and Cai, B. 2026. DREAM: LLM-based Dynamic Role-playing via Event-Aware Memory Graph. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD ’26, 5686–5697. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722592.
- Xu, A.; Lin, B.; Xue, B.; Wang, B.; Xu, B.; Wu, B.; Zhang, B.; Lin, C.; Dong, C.; Ling, C.; et al. 2026a. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*.
- Xu, Z.; Chen, D.; Wang, S.; Li, J.; Wang, C.; Han, M.; and Wang, Y. 2026b. AdaMARP: An Adaptive Multi-Agent Interaction Framework for General Immersive Role-Playing. In Liakata, M.; Moreira, V. P.; Zhang, J.; and Jurgens, D., eds., *Findings of the Association for Computational Linguistics: ACL 2026*, 31230–31283. San Diego, California, United States: Association for Computational Linguistics. ISBN 979-8-89176-395-1.
- Yang, A.; et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.
- Youyou, W.; Kosinski, M.; and Stillwell, D. 2015. Computer-Based Personality Judgments Are More Accurate than Those Made by Humans. *Proceedings of the National Academy of Sciences*, 112(4): 1036–1040.
- Zhang, Y.; Song, Z.; Zhou, H.; Ren, W.; Chen, Y.-P. P.; Yu, J.; and Yang, W. 2025a. GA-S3: Comprehensive Social Network Simulation with Group Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, 8950–8970. Vienna, Austria: Association for Computational Linguistics.
- Zhang, Z.; Liu, S.; Liu, Z.; Zhong, R.; Cai, Q.; Zhao, X.; Zhang, C.; Liu, Q.; and Jiang, P. 2025b. LLM-Powered User

Simulator for Recommender System. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12): 13339–13347.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 46595–46623.

Zhou, X.; Su, Z.; Eisape, T.; Kim, H.; and Sap, M. 2024. Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21692–21714. Miami, Florida, USA: Association for Computational Linguistics.

A User Representation and Memory Retrieval

This technical supplement provides additional details on the user representation, memory retrieval, reasoning traces, evaluation rubric, qualitative examples, and prompt templates used in PsyMemBehav. We organize the supplement around the main components of the system: user profile construction, memory retrieval, trace construction, DPO preference construction, additional analyses, case studies, and prompts.

A.1 Trait-based User Representation

We represent each user with two complementary profile groups:

$$P_u = (P_u^{\text{gen}}, P_u^{\text{plat}}).$$

The general psychological profile P_u^{gen} captures relatively stable dispositions, including personality traits, values, moral preferences, and broad social orientation. These dimensions describe how a user tends to interpret information, express disagreement, evaluate social issues, and maintain persistent preferences across topics.

The platform-specific profile P_u^{plat} captures online behavioral tendencies. These dimensions describe how the same user tends to behave in social-media environments, including opinion leadership, social capital, critical media literacy, and toxicity-related interaction patterns. This separation is important because two users may share similar psychological dispositions while expressing them differently on a platform: one may engage through careful correction, another through short ironic replies, and another through passive endorsement.

The general psychological profile P_u^{gen} comprises 22 dimensions across four categories. *Big Five personality* (Goldberg 1990) captures openness, conscientiousness, extraversion, agreeableness, and neuroticism. *Schwartz values* (Schwartz 1992) cover ten motivational value types: self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence, and universalism. *Moral foundations* (Graham et al. 2013) include care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. *Social dominance orientation* (Pratto et al. 1994) provides two dimensions: SDO dominance and SDO anti-egalitarianism.

The platform-specific profile P_u^{plat} comprises 19 dimensions across four categories. *Emotional profile* covers expressiveness, stability, negativity bias, empathy, anxiety tendency, and anger tendency. *Network and social capital* (Katz and Lazarsfeld 1955; Ellison, Steinfield, and Lampe 2007) captures opinion leadership, network centrality, and social capital. *Media literacy* (Guess et al. 2020) includes critical media literacy, source evaluation, and misinformation susceptibility. *Online behavior* (Wulczyn, Thain, and Dixon 2017) covers posting frequency, reply tendency, repost tendency, like tendency, toxicity, trolling tendency, and botlikeness.

The profile is produced from each user’s pre-split history $\mathcal{H}_u^{\text{train}}$ by an annotation model g_{prof} . To avoid leakage, only behaviors before the prediction split are used for profile construction. Each stored profile contains numerical trait scores and evidence-grounded summaries. The numerical scores provide compact control variables, while the summaries preserve interpretable evidence for generation and analysis.

A.2 Memory Retrieval

For each prediction instance at time t , PsyMemBehav retrieves a compact memory set $M_{u,t}$ from the user’s previous behaviors:

$$\mathcal{H}_u^{<t} = \{h_{u,1}, h_{u,2}, \dots, h_{u,n}\}.$$

We use BM25 lexical retrieval (Robertson and Zaragoza 2009). The current post together with its visible comment context forms the query q_t , and each historical behavior $h_{u,j}$ is represented as a candidate document $d_{u,j}$ using its available textual content. The BM25 relevance score is

$$s_{\text{BM25}}(d_{u,j}, q_t) = \sum_{w \in q_t} \text{IDF}(w) \frac{f(w, d_{u,j})(k_1 + 1)}{D(w, d_{u,j})},$$

$$D(w, d_{u,j}) = f(w, d_{u,j}) + k_1 \left(1 - b + b \frac{|d_{u,j}|}{\text{avgl}} \right),$$

where $f(w, d_{u,j})$ is the frequency of query term w in document $d_{u,j}$, $|d_{u,j}|$ is the document length, avgl is the average candidate-document length, and k_1 and b are the standard BM25 term-saturation and length-normalization parameters. The final memory set keeps the top-ranked pre-time behaviors:

$$M_{u,t} = \text{TopK}_{h_{u,j} \in \mathcal{H}_u^{<t}} s_{\text{BM25}}(d_{u,j}, q_t).$$

We do not use dense embedding similarity or an additional recency-prior term in the retrieval score. Temporal validity is enforced by restricting the candidate pool to behaviors observed strictly before t . Unless otherwise specified, we use $k = 5$ as the default retrieval size, balancing relevant historical evidence against redundant or noisy memories.

A.3 User Profile Distribution

To examine whether the constructed user profiles collapse into a narrow region or preserve meaningful diversity, we project the high-dimensional profile vectors into a two-dimensional space using principal component analysis. Figure 3 shows that the profile distribution is compact but non-collapsed. Most users are concentrated around a dense center, indicating shared behavioral regularities across the population. At the same time, peripheral low-density regions

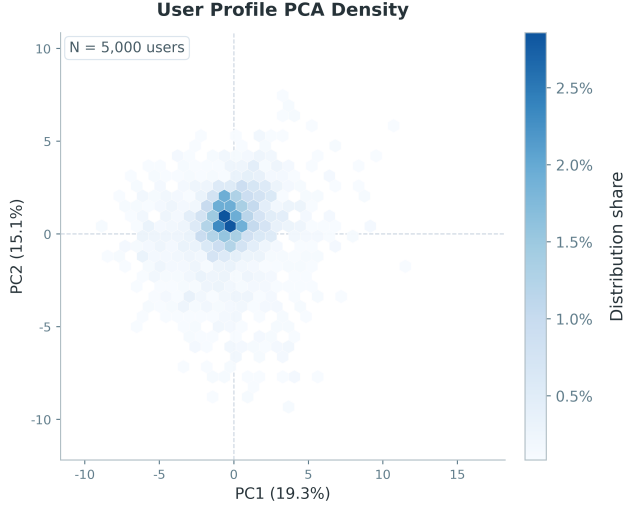


Figure 3: PCA density visualization of user profile representations. Each hexagonal bin represents the local density of users in the projected two-dimensional profile space.

indicate long-tail users with more distinctive profile combinations. This supports the use of profile-conditioned simulation, since the population contains both central tendencies and individual heterogeneity.

B Reasoning Schema and Trace Construction

PsyMemBehav uses a two-level reasoning schema to connect user identity with concrete platform behavior. The intermediate trace is represented as

$$z_t = (z_t^{\text{id}}, z_t^{\text{act}}),$$

where z_t^{id} summarizes the situated user identity and z_t^{act} explains the selected target, the action conditioned on that target, and the resulting response when required.

The identity-activation component identifies the salient user role in the current context. It includes activated psychological traits, activated platform-specific traits, retrieved memory evidence, likely stance, and behavioral constraints. For example, a user may act as a skeptical fact-checker in one thread, a supportive community member in another, and an ironic commentator in a third. The identity-activation stage is designed to capture this context-dependent but profile-consistent identity.

The action-decision component maps the situated evidence summary into a target-conditioned platform behavior:

$$c_t \in C_t, \quad a_t \in \{\text{COMMENT}, \text{LIKE}, \text{REPOST}\}.$$

It first selects the interaction target c_t and then predicts how to interact with that target through a_t . The visible response r_t is planned only when the selected action is COMMENT. For LIKE and REPOST, the model returns the selected target and action without generating unnecessary visible text.

We construct reverse CoT traces by asking a teacher model to explain why the observed ground-truth action is plausible given the user profile, retrieved memory, and current thread.

The teacher is not asked to predict a new action; instead, it produces an inverse reasoning trace aligned with the known target and action. This makes the reasoning trace a supervision signal for behavior explanation rather than a free-form alternative response.

A trace validator J_{trace} accepts a reconstructed trace only when it satisfies profile consistency, memory grounding, contextual relevance, role-reasoning coherence, and action-reasoning coherence:

$$J_{\text{trace}}(x_t, z_t, y_t^*) = \prod_{k=1}^5 \mathbf{1}[c_k(x_t, z_t, y_t^*)] \in \{0, 1\}.$$

Only accepted traces are used as training supervision.

C Evaluation Rubric and DPO Preference Construction

We use a ten-criterion rubric to evaluate simulated user behaviors and construct DPO preference pairs. The rubric covers four aspects: action-level fidelity, user-profile alignment, content quality, and format/reasoning validity. This design reflects the fact that social-media simulation is not only a text generation problem: the model must first select a valid interaction target, then choose how to interact with that target, and only then produce a response if the selected action requires visible text.

Action-level fidelity includes action accuracy, interaction-intensity calibration, and target/thread grounding. User-profile alignment includes persona consistency, memory activation discipline, and stance consistency. Content quality includes response naturalness and specificity. Format/reasoning validity includes output format validity and CoT plausibility.

The DPO judge receives the correct high-level action label but not the gold visible text. This prevents the judge from rewarding exact surface matching and encourages evaluation of target grounding, action fidelity, persona consistency, memory discipline, stance, naturalness, specificity, format validity, and reasoning plausibility. The weighted score is computed as

$$S(\tau) = \frac{2R_1 + \sum_{i=2}^{10} R_i}{11},$$

where action fidelity is double-weighted because choosing the correct action family remains a central fidelity requirement after target selection.

C.1 Rubric Activation Analysis

Figure 4 reports how frequently each rubric dimension is activated in the annotated evaluation pairs. Persona consistency, CoT plausibility, and context grounding are the most frequently activated dimensions, while response naturalness, specificity, and memory consistency are also highly active. This shows that the evaluation is not merely action-based, but also depends on profile alignment, social-media realism, and reasoning plausibility.

Category	Criterion	Definition
Action-level Fidelity	R1. Action Fidelity	Whether the candidate selects the correct high-level behavior type, such as COMMENT, LIKE, or REPOST. LIKETHREAD and LIKEBASIC are both treated as LIKE.
	R2. Interaction Intensity	Whether the amount of engagement is appropriate. Overly strong interaction, such as writing a long comment when the user would only like, and under-interaction, such as producing a weak or incomplete response when a reply is expected, are penalized.
	R3. Target Grounding	Whether the candidate understands the current post, comment tree, reply target, and local social context.
User-profile Alignment	R4. Persona Consistency	Whether the behavior matches the user’s profile, long-term interests, values, directness, social habits, and emotional style.
	R5. Memory Discipline	Whether historical comments are used only when helpful. The model is penalized both for ignoring clearly relevant memories and for forcing irrelevant memories into the reasoning.
	R6. Stance Consistency	Whether the stance is compatible with the user’s prior positions and the current thread, without unsupported ideological invention or stance reversal.
Content Quality	R7. Response Naturalness	Whether the visible response sounds like a real platform user rather than an assistant, summary, template, or overly polished explanation.
	R8. Specificity	Whether the response addresses concrete contextual details rather than giving a generic or vague reply.
Format and Reasoning	R9. Format Validity	Whether target, action, CoT, and visible response fields are complete, legal, parseable, and consistent with the candidate target list.
	R10. CoT Plausibility	Whether the reasoning plausibly explains the action and response without over-rationalization, fabricated private psychology, or unsupported causal claims.

Table 5: Evaluation rubric for simulated user behaviors. The rubric contains ten criteria grouped into four categories.

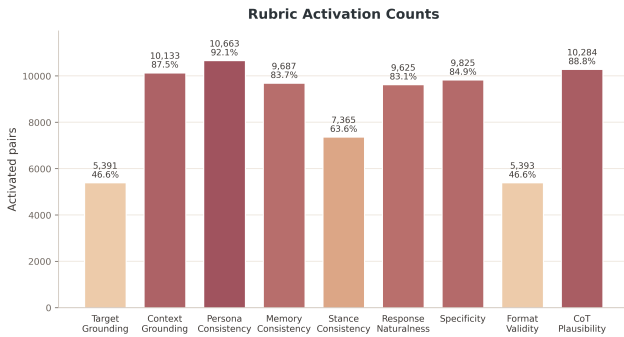


Figure 4: Activation counts of rubric dimensions over the evaluation set. The numbers above each bar show the absolute number of activated pairs and their percentage in the full annotated set.

D Additional Analyses

This section reports additional analyses that complement the main experimental results. We include platform-level action statistics, thread-level structural properties, the effect of memory retrieval size, judge robustness, and pairwise group-simulation preference results.

D.1 Platform-level Simulation Statistics

Figure 5 shows the distribution of simulated agent actions. Comments account for the largest proportion of actions, followed by reposts and likes. This indicates that the simulated platform is primarily driven by active conversational behavior rather than passive engagement.

Figure 6 presents thread-level interaction statistics, including comments per thread, maximum thread depth, and

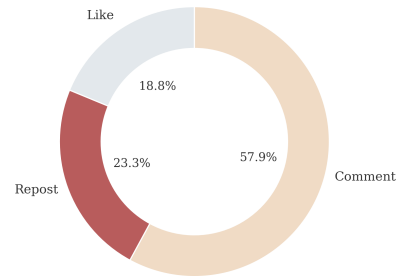


Figure 5: Distribution of simulated agent actions.

unique participants. Most discussions remain moderate in scale, while a small subset develops into deeper and more active interactions. This suggests that the simulator does not simply generate uniformly dense discussion trees, but instead produces a mixture of shallow and deeper interaction patterns.

D.2 Effect of Memory Retrieval Size

We vary the number of retrieved past comments k and evaluate the average generation quality. As shown in Figure 7, both the 9B and 27B models achieve the best performance when $k = 5$. Smaller values provide insufficient historical context, while larger values may introduce redundant or noisy memories. Therefore, we set $k = 5$ as the default retrieval size.

D.3 Judge Robustness

LLM-as-a-judge evaluation may introduce evaluator-specific preferences. To examine whether our conclusion depends

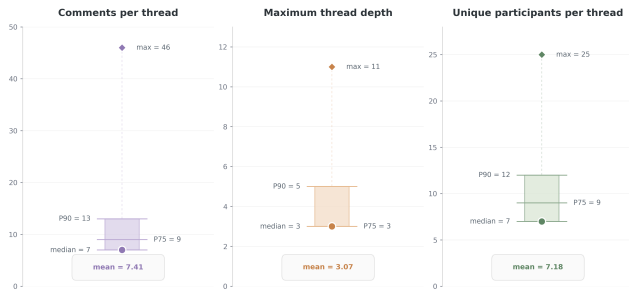


Figure 6: Statistical properties of simulated discussion threads.

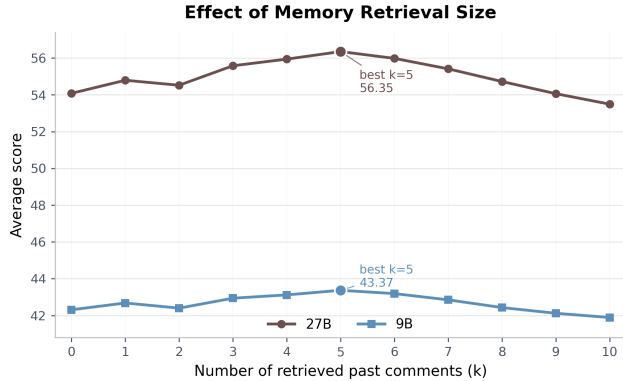


Figure 7: Effect of memory retrieval size k on average generation quality.

on a particular evaluator, we evaluate the same anonymized model outputs using multiple judge models from different model families. As shown in Table 6, PsyMemBehav 27B receives the highest score under all reported judges, suggesting that its advantage is not caused by a single evaluator or judge-family preference.

Judge	PsyMemBehav	GPT-5.4	Gemini	Claude
Qwen3.5-27B	61.96	54.23	52.02	47.60
GPT-5.4	63.09	60.06	48.94	57.12
Gemini-2.5-Pro	72.50	67.60	62.75	61.33
Claude-Sonnet-4.6	47.27	40.68	33.88	44.44

Table 6: Judge robustness analysis. Each row corresponds to one judge model; each column corresponds to an evaluated model. PsyMemBehav 27B achieves the highest score under all four judges.

We also conduct a direct pairwise group-simulation preference test between PsyMemBehav 27B and GPT-5.4. For each instance, both models are given the same real thread seed and asked to generate a group-level continuation. The better simulation is selected according to its fidelity to the real discussion dynamics.

Model	Win Rate
GPT-5.4	34.0%
PsyMemBehav 27B	64.0%
Tie	2.0%

Table 7: Pairwise group simulation preference test between GPT-5.4 and PsyMemBehav 27B.

D.4 Human–Gemini Evaluation Agreement

For the human evaluation, we selected 30 anonymized user IDs and evaluated five behavior instances per ID, yielding 150 annotated instances. Human evaluators and Gemini scored the same outputs under the same seven-dimensional rubric. Table 2 in the main paper compares their dimension-level means. The human and Gemini overall means are 54.40 and 55.46, respectively, a human-minus-Gemini difference of -1.06 points. Humans score semantic alignment, style fidelity, and naturalness slightly higher, whereas Gemini scores intent, emotion, stance, and profile/memory consistency higher. This dimension-mean comparison assesses aggregate calibration; it neither establishes per-instance agreement nor replaces a formal human reliability analysis.

The largest absolute gaps occur for stance (4.20 points) and emotion (3.60 points), while all other dimensions differ by at most 3.73 points. Because the signs vary, Gemini does not show a simple global tendency to be uniformly more lenient or stringent. The dimension-specific gaps therefore motivate reporting human evaluation together with LLM-judge results and cross-judge robustness.

D.5 Dataset Scale and Longitudinal Coverage

Our use of *large-scale* does not imply platform-wide coverage. Instead, it refers to the joint scale of real-user coverage, total behavioral volume, and the density of longitudinal histories available for user modeling. We therefore characterize the resource more precisely as a *large-scale, longitudinal, user-centered dataset*. Table 8 compares PsyMemBehav-bench with closely related work along three dimensions: the number of real users represented, the behavioral evidence available for modeling each user, and the granularity of the prediction or simulation target.

Here, *real-user coverage* denotes the number of distinct real users whose observed behavioral evidence is used to construct or evaluate the simulator. *Behavioral evidence* captures both the total amount of data and the density of the history available for each user. *Prediction/simulation target* distinguishes user-level target–action–response prediction from aggregate diffusion or traffic simulation.

Among the closely related studies that report real-user counts, PsyMemBehav-bench contains more real users and substantially denser longitudinal behavioral histories. Its scale characterization is therefore based not on the number of users alone, but on the combination of 5,862 real users, 5.57 million observed behaviors, and approximately 950 behaviors per user. At the same time, these statistics should not be interpreted as platform-wide coverage or as a representative sample of all social-media users. The collection procedure

Study	Real-user coverage	Behavioral evidence	Prediction/simulation target
FineRob: <i>LLM-Based Behavior Prediction for Social Media Users with Continuous Memory</i> (Li et al. 2025)	1,866 users	78.6K fine-grained prediction records; 126 historical records per user on average	Interaction object, behavior type, and content
<i>Can LLMs Simulate Social Media Engagement? A Study on Action-Guided Response Generation</i> (Qiu et al. 2025)	User count not reported	3,990 instances; 1–31 historical examples per user	Engagement-action selection and response generation
MOSAIC (Liu et al. 2025)	204 real users and 204 simulated personas	30 posts plus survey-derived personas	Group-level content dissemination and regulation
GA-S ³ (Zhang et al. 2025a)	Group agents rather than identifiable real-user histories	30 events over seven days	Aggregate traffic and collective behavior
PsyMemBehav-bench	5,862 real users	5.57M observed behaviors; approximately 950 behaviors per user	User-level interaction target, target-conditioned action, and response

Table 8: Scale comparison with closely related social-media behavior prediction and simulation studies.

favors observable users with sufficiently rich histories and may therefore overrepresent highly active accounts while underrepresenting low-activity, private, or newly created accounts. Platform access constraints, the selected observation period, and activity-based sampling can further shape the resulting population. These limitations should be considered when interpreting generalization beyond the studied users and contexts.

E Evaluation Metric Definitions

E.1 Profile Fidelity: Normalized Edit Distance

Profile fidelity measures how faithfully a simulator reproduces a user’s behavioral profile. For each user, we re-annotate the simulator’s generated behaviors using the same annotation model g_{prof} to obtain a predicted profile vector $\hat{P}_u$, and compare it against the original profile P_u derived from ground-truth history.

We report fidelity using $100 \times (1 - \text{NED})$, where NED is the Normalized Edit Distance between the discretized profile sequences:

$$\text{NED}(P_u, \hat{P}_u) = \frac{\text{ED}(P_u, \hat{P}_u)}{\max(|P_u|, |\hat{P}_u|)},$$

where $\text{ED}(\cdot, \cdot)$ denotes the Levenshtein edit distance and $|\cdot|$ denotes sequence length after discretizing continuous trait scores into ordinal bins. Because both reference and generated behaviors are interpreted through the same annotator, this score measures relative cross-system consistency under a fixed profiling lens; it should not be interpreted as absolute psychometric validity. A score of $100 \times (1 - \text{NED}) = 100$ indicates a perfect match; higher values reflect closer alignment between simulated and real user profiles. We report two profile-fidelity scores separately: one for general psychological traits (P_u^{gen}) and one for platform-specific behavioral traits (P_u^{plat}).

E.2 Group-Level Simulation Metrics

The group-level evaluation places multiple profiled simulators into real thread seeds and measures how closely the gen-

erated interaction structure matches real comment-section continuations.

Reply Reciprocity (Rec.). Reciprocity is a standard directed-network measure of mutual linkage (Garlaschelli and Loffredo 2004). We compute the fraction of interacting user pairs (u, v) for which both reply directions occur:

$$\text{Rec.} = \frac{|\{(u, v) : u \rightarrow v \wedge v \rightarrow u\}|}{|\{(u, v) : u \rightarrow v \vee v \rightarrow u\}|}.$$

Overly high reciprocity indicates an unrealistic group-chat pattern rather than broadcast discussion.

Stance Deviation (Stance Dev.). Following the conventional favor/against/neither formulation of stance (Mohammad et al. 2016), each comment is annotated as PRO, ANTI, or NEUTRAL. We measure the ℓ_1 distance between simulated and real stance distributions:

$$\text{StanceDev} = \sum_{s \in \{\text{Pro, Anti, Neutral}\}} |p_{\text{sim}}(s) - p_{\text{real}}(s)|.$$

Lower values indicate closer reproduction of collective opinion composition.

Attention Concentration (Top1 KOL / Top5 PR). Top1 KOL is the fraction of all received replies captured by the single most-addressed user, reflecting key-opinion-leader concentration (Katz and Lazarsfeld 1955). Top5 PR is the sum of normalized PageRank scores (Brin and Page 1998) for the five highest-ranked users in the directed reply graph. Together they test whether simulated attention and structural influence concentrate as in the real thread.

Overall Deviation (MAE / MSE). MAE and MSE aggregate, respectively, absolute and squared per-thread deviations between simulated and real statistics. Lower values indicate closer overall group-level fidelity; MSE penalizes large mismatches more strongly.

F Case Studies

This section provides qualitative and group-level examples that complement the aggregate results in the main paper. We

include both a single-reply comparison and a replay-style group simulation comparison. The single-reply case highlights how different models vary in stance commitment and response style under the same thread context. The group-level case shows that a model can produce plausible individual replies while still failing to reproduce the collective dynamics of a real discussion.

F.1 Single-reply Comparison

Table 9 compares different models under the same thread context and target comment. The real user reply is cautious and low-commitment. Several strong general-purpose models produce more socially reactive or stance-bearing responses, while PsyMemBehav 27B remains more hedged and balancing.

Model	Generated Reply	Stance	Observation
Real user	"I can see why people read the timing as suspicious, but it may also just be production timing."	Neutral	Conservative and low-commitment stance.
Gemini-3.1-Pro	"They could still be reworking it. This week gave them too much new material to ignore."	Neutral	Cautious and relatively close to the real style.
GPT-5.4	"That timing feels awfully convenient. People are going to read into it no matter what."	Neutral / mildly charged	More socially reactive and discussion-provoking.
Claude-Sonnet-4.6	"That explanation feels way too convenient given the pressure around this."	Pro / supportive	More likely to clearly agree with a suspicious reading.
Qwen3.7-Max	"That excuse is transparent. They probably got spooked by legal threats or the network."	Pro / charged	Stronger stance and more speculative.
PsyMemBehav 27B	"The reality is probably more mixed than this thread is making it."	Neutral	More hedged and balancing, closer to the real user's conservative stance.
DeepSeek-V4-Pro	"There is a narrower point here, but the broad claim feels overcooked."	Anti / neutral	Often gives a cooling or corrective reply.

Table 9: Single-reply comparison under the same thread context.

F.2 Group Simulation Comparison

The group simulation setting starts from the same real thread seed and asks each model to generate a continuation involving multiple agents. Table 4 in the main paper reports aggregate interaction and network-level results over the evaluation set. Table 10 complements that aggregate evaluation with the stance distribution of this representative replay case. Each comment is assigned to one of three mutually exclusive stance classes: **Pro** explicitly supports the focal claim, **Anti** opposes it, and **Neutral (Neu.)** expresses neither clear support nor opposition; their proportions therefore sum to one. **Charged (Chg.)** is the share of stance-bearing comments, $\text{Chg.} = p(\text{Pro}) + p(\text{Anti}) = 1 - p(\text{Neutral})$. **Distance (Dist.)** is the ℓ_1 distance between a model's stance vector and the real continuation, $\sum_s |p_{\text{sim}}(s) - p_{\text{real}}(s)|$; it ranges from 0 to 2, and lower is better. **Entropy (Ent.)** is normalized Shannon entropy, $-\sum_s p(s) \log p(s) / \log 3$; it ranges from 0 (all comments in one stance) to 1 (a uniform three-way distribution). Because entropy measures stance diversity rather than accuracy, neither a larger nor a smaller raw value is intrinsically better; closeness to the real entropy indicates better fidelity.

Discussion. This representative case shows that strong general-purpose models generate substantially more stance-bearing responses than the real continuation. The real thread is predominantly neutral (0.830), with a charged share of

Model	Pro	Anti	Neu.	Chg.	Dist.	Ent.
Real	0.085	0.085	0.830	0.170	0.000	0.523
Gemini-3.1-Pro	0.312	0.125	0.563	0.437	0.534	0.862
GPT-5.4	0.479	0.083	0.438	0.562	0.788	0.838
Claude-Sonnet-4.6	0.322	0.212	0.466	0.534	0.728	0.955
Qwen3.7-Max	0.417	0.062	0.521	0.479	0.664	0.798
PsyMemBehav 27B	0.182	0.121	0.697	0.303	0.266	0.744
DeepSeek-V4-Pro	0.423	0.135	0.442	0.558	0.776	0.906

Table 10: Stance-distribution comparison for a representative continuation from the same real thread seed. Neu. = Neutral; Chg. = Charged; Dist. = stance distance; Ent. = entropy. Aggregate interaction and network-level results are reported in Table 4 in the main paper; metric definitions are provided in the *Group-Level Simulation Metrics* subsection.

0.170. Among the simulated systems, PsyMemBehav 27B has the highest neutral share (0.697), the lowest charged share (0.303), and the smallest stance distance (0.266), making its stance distribution the closest to the real continuation. Its entropy (0.744) is also the closest to the real value (0.523). This case-level analysis complements, rather than duplicates, the aggregate thread-level comparison in Table 4 in the main paper.

G Prompt Templates

This section lists the main prompt templates used in PsyMemBehav. We include the prompts for CoT validation, action generation, reference CoT construction, pairwise DPO judging, and seven-dimension semantic evaluation. All prompt tables are shown as double-column tables for readability.

A Reference Prompt for CoT Validation.

```
1 You are a validator for generated hidden CoT in a real social media user memory activation dataset.
2
3 You will receive the system prompt, dialogue prefix, original assistant target, and a generated assistant value
  containing <think> plus the original target.
4
5 Your job is to decide whether the <think> is valid supervision for deciding when to use the scene memory pool.
6
7 === REFERENCE CONTEXT ===
8 {reference_content_with_memories}
9
10 === ORIGINAL ASSISTANT TARGET ===
11 {target_assistant_content}
12
13 === GENERATED ASSISTANT VALUE TO VALIDATE ===
14 {generated_reasoning}
15
16 Check the following:
17
18 1. Structure:
19 - There must be exactly one <think> block before the assistant target.
20 - The <think> should contain exactly three natural paragraphs.
21 - It should not use explicit awareness/activation/decision tags, headings, bullets, or numbered step labels.
22 - The assistant target after </think> must exactly match the original assistant target.
23
24 2. Source grounding:
25 - Every fact in the <think> must be supported by the system profile, current scenario, dialogue prefix,
   Relevant Past Memories, Other Characters Relationships, or original assistant target.
26 - No fabricated scene details, motives, actions, or emotions.
27
28 3. Current-turn boundary:
29 - The <think> must explain only the current assistant turn.
30 - It must not use later dialogue or future plot knowledge unavailable to the character at this story point.
31 - If the system profile contains future events, the CoT should not use them unless the current character could
   know them.
32
33 4. Information leakage:
34 - The <think> must not leak information unavailable to the character at the current turn.
35 - The <think> must not reveal information that the visible assistant target withholds.
36 - The <think> may use the original assistant target to check stance alignment and response policy, but it must
   not quote, closely paraphrase, or explicitly pre-announce the target's distinctive surface wording,
   actions, or inner thoughts.
37 - Treat future-biographical details in the profile as unsafe unless the character could know or act from them
   at this story point.
38
39 5. Memory decision:
40 - Independently decide whether the scene memory pool is needed.
41 - Scene memory pool means only the listed Relevant Past Memories, not the character profile.
42 - If the current dialogue, current scenario, and profile are sufficient, memory should be marked not needed.
43 - If the target requires information absent from current context/profile but present in a relevant memory,
   memory should be marked needed.
44 - Check that the generated CoT names the correct memory ids when memory is needed.
45 - When memory is not needed, the CoT should explain why the current dialogue, current scenario, and/or profile
   are sufficient.
46 - If memory is needed, check that the generated CoT explains how each activated memory affects behavior in the
   current turn.
47 - If multiple memories are activated, check that the generated CoT compares their directions: aligned, opposed,
   or mixed, and explains the final net effect.
48
49 6. Profile vs memory:
50 - If the required information is in the profile, the CoT should not claim that a scene memory is needed.
51 - It may say profile knowledge is needed, but scene memory pool is not.
52
53 7. Answer leakage:
54 - The <think> must not quote or closely paraphrase the final assistant wording.
55 - It may describe response function, tone, and constraints, but not the final surface form.
56
57 8. Training usefulness:
58 - The <think> should clearly analyze the current information need, memory relevance, and response policy.
59 - It should not be a generic template or a repetition of the final answer.
```

Table 11: A reference prompt for CoT validation.

Unified End-to-End Behavior-Prediction Prompt.

```

1  You are roleplaying as social-media user {username}. Given the user profile, retrieved pre-event history, current
   thread, and candidate targets, first select the interaction target. Conditioned on that target, predict the
   user's recorded engagement as exactly one of Comment, Like, or Repost. Generate textual content only for
   Comment. Base both decisions on evidence in the input and do not sound like an assistant.

2
3  [Persona]
4  {rendered_profile}
5
6  [Retrieved Past Comments]
7  {rendered_history_block}
8
9  [Thread]
10 [post] @{{post_author}} -- "{{post_title}}"
11 {post_text}
12
13 [Comments]
14 - [{{cand_id_1}}] @{{author_1}} (reply to [post]): "{{full_text_1}}"
15 - [{{cand_id_2}}] @{{author_2}} (reply to [{{parent_id}}]): "{{full_text_2}}"
16
17 [Candidate Targets]
18 [post], [{{cand_id_1}}], [{{cand_id_2}}]
19
20 Rules:
21 - Select a valid target before choosing the action; neither the gold target nor the gold action is provided.
22 - Conditioned on the selected target, infer exactly one action from {Comment, Like, Repost} and keep the target
   fixed.
23 - Emit valid JSON for exactly one action and omit all fields belonging to the other actions.
24 - Comment: output action="Comment" and a reply object containing reply_to_id and natural, thread-specific text.
25 - Like: output action="LikeThread" and a target object containing id and kind; LikeThread and LikeBasic are
   evaluated as high-level Like. A like represents light endorsement, acknowledgment, or agreement.
26 - Repost: output action="Repost" and the same target object; a repost represents stronger public amplification than
   a like.
27 - Do not generate visible text for Like or Repost. Use history only as preference evidence and keep private
   reasoning concise and grounded.
28
29 Output exactly:
30 <think>
31 - Selected target: ...
32 - Predicted action/intensity for this target: Comment, Like, or Repost / ...
33 - Context grounding: ...
34 - Persona/memory/stance cue: ...
35 - Naturalness cue: ...
36 </think>
37
38 Then output exactly one valid <answer> branch; do not output branch labels.
39 Comment: <answer>{"activity_type":"BehaviorPrediction","action":"Comment","reply":{"reply_to_id":"<candidate id>","
   text":"..."}}</answer>
40 Like: <answer>{"activity_type":"BehaviorPrediction","action":"LikeThread","target":{"id":"<candidate id>","kind":"
   post | comment"}}</answer>
41 Repost: <answer>{"activity_type":"BehaviorPrediction","action":"Repost","target":{"id":"<candidate id>","kind":"post
   | comment"}}</answer>

```

Table 12: Unified end-to-end behavior-prediction template. All inputs and target-first reasoning steps are shared; only the selected action and its required output fields differ.

A Reference Prompt for Reference CoT Construction.	
1	# Action-DPO Reference CoT Prompt
2	
3	You are writing inverse reasoning for a persona-simulation training example.
4	
5	You are given:
6	- the user’s profile and retrieved history,
7	- the current post/thread context,
8	- the ground-truth action and answer block.
9	
10	Your job is not to predict a new action. Your job is to explain, briefly and plausibly, why this specific user would take the provided ground-truth action in this situation.
11	
12	## What To Cover
13	
14	Write a compact natural-language <think> block. It should cover:
15	- Action fit: why this action type is natural for the user here.
16	- Interaction intensity: why the amount of engagement is calibrated.
17	- Context grounding: what in the current post/thread creates the opening.
18	- Persona and memory: which profile/history signals matter.
19	- Stance and tone: the concrete stance or affective register implied by the action.
20	- Format/target: keep the selected target fixed and include visible target references only when the chosen action format requires them.
21	
22	## Constraints
23	
24	- Do not invent private motives, fake memories, or unsupported ideology.
25	- Do not claim that the exact hidden ground-truth text is the only possible wording.
26	- For like/repost actions, do not over-explain as if the user wrote a comment.
27	- Keep the reasoning concise; this should be a training CoT, not an essay.
28	- Output only one <think>...</think> block. Do not output an <answer> block.
29	
30	## Output Format
31	
32	<think>
33	Action fit: ...
34	Interaction intensity: ...
35	Context grounding: ...
36	Persona and memory: ...
37	Stance and tone: ...
38	Format/target: ...
39	</think>

Table 13: A reference prompt for constructing inverse CoT supervision from a known ground-truth action.

A Reference Prompt for Pairwise DPO Judging.

```

1 You are a strict pairwise preference judge for persona-simulation DPO data.
2
3 You will receive:
4 - the target user's profile and retrieved past comments;
5 - the current social-media thread and candidate targets;
6 - the correct high-level action type only: comment/reply, like, or repost;
7 - Candidate A and Candidate B, each containing CoT, target, action, and optional visible reply text.
8
9 Important:
10 - The ground-truth visible content is hidden. Do not assume exact wording.
11 - LikeThread and LikeBasic are both treated as high-level like.
12 - Judge target grounding first, then the action conditioned on that target, persona fit, memory use, stance,
    naturalness, specificity, format, and CoT plausibility.
13 - For comment/reply, judge whether the visible text is plausible for this user and context.
14 - For like/repost, penalize unnecessary visible comments.
15 - Do not reward length or polished assistant-like writing by itself.
16
17 Score each candidate from 1 to 5:
18 R1. Action Fidelity
19 R2. Interaction Intensity Calibration
20 R3. Target and Thread Grounding
21 R4. Persona Consistency
22 R5. Memory Activation Discipline
23 R6. Stance Consistency
24 R7. Response Naturalness
25 R8. Specificity
26 R9. Format Validity
27 R10. CoT Plausibility
28
29 Use the weighted score:
30 weighted_score = (2*R1 + R2 + R3 + R4 + R5 + R6 + R7 + R8 + R9 + R10) / 11
31
32 Choose A, B, tie, or drop.
33 Drop weak comparisons if:
34 - confidence < 4;
35 - score gap is too small;
36 - candidates are near-duplicates;
37 - the winner wins mainly by being longer;
38 - either candidate has invalid target, action, format, or malformed JSON;
39 - the CoT fabricates private psychology or forced memory.
40
41 Return JSON only:
42 {
43   "sample_id": "...",
44   "correct_action_type": "comment|like|repost|other",
45   "candidate_a_scores": {
46     "R1": 1, "R2": 1, "R3": 1, "R4": 1, "R5": 1,
47     "R6": 1, "R7": 1, "R8": 1, "R9": 1, "R10": 1,
48     "weighted_score": 0.0
49   },
50   "candidate_b_scores": {
51     "R1": 1, "R2": 1, "R3": 1, "R4": 1, "R5": 1,
52     "R6": 1, "R7": 1, "R8": 1, "R9": 1, "R10": 1,
53     "weighted_score": 0.0
54   },
55   "chosen": "A|B|tie|drop",
56   "confidence": 1,
57   "score_gap": 0.0,
58   "winner_reason": "brief evidence-bound reason",
59   "loser_weakness": "main concrete weakness",
60   "keep_for_dpo": true
61 }

```

Table 14: A reference prompt for pairwise DPO preference judging.

A Reference Prompt for Seven-Dimension Semantic Evaluation.

```

1 You are a strict, consistent, and objective evaluator for social media user simulation.
2 Your task is to assess how well the model-generated comment simulates the target user's real response in the same social
  media context.
3 You will be given:
4 1. The original social media post
5 2. The target user's history, profile, or memory information
6 3. The target user's real comment, used as the ground truth
7 4. The model-generated comment
8 Important evaluation principles:
9 - The ground-truth comment is not the only possible correct response, but it represents the user's actual reaction in
  this context.
10 - Do not judge only by lexical overlap. Evaluate semantic meaning, interaction intent, emotion, stance, style, persona/
  memory consistency, and naturalness.
11 - If the generated comment uses different wording but expresses a similar meaning, attitude, emotion, and interaction
  goal, it can receive a high score.
12 - If the generated comment is semantically relevant but does not match the user's style, stance, or interaction intent,
  reduce the corresponding scores.
13 - If the generated comment fabricates the user's personal history, identity, preferences, relationships, or experiences,
  significantly reduce the persona_memory_consistency score.
14 - If the generated comment sounds like AI writing, is overly formal, overly explanatory, or resembles a summary, reduce
  naturalness and style_fidelity scores.
15 - Do not give a high score simply because the generated comment is more complete or grammatically fluent. The goal is
  whether it sounds like the target user.
16 - Be strict. A score above 90 should only be given when the generated comment is extremely close to the target user's
  real response.
17 - Avoid giving all dimensions similar scores. Evaluate each dimension independently.
18 [Input]
19 Original social media post: {source_post}
20 Target user history/profile/memory: {user_profile_or_history}
21 Ground-truth comment: {ground_truth_comment}
22 Model-generated comment: {generated_comment}
23 [Scoring Scale] Assign an integer score from 0 to 100 per dimension.
24 90-100: Nearly perfect match.
25 75-89: Strong match, only minor differences.
26 60-74: Moderate match, noticeable differences.
27 40-59: Weak match, substantial deviation.
28 20-39: Poor match.
29 0-19: Very poor; irrelevant, contradictory, or fabricated.
30 [Evaluation Dimensions]
31 1. semantic_alignment
32 Assess whether the generated comment expresses the same core meaning, topic, and information points as the ground-truth
  comment.
33 2. intent_alignment
34 Assess whether the generated comment fulfills the same social interaction purpose (e.g., supporting, opposing, joking,
  asking, answering, expressing emotion, sharing experience, giving advice).
35 3. emotion_consistency
36 Assess whether the generated comment matches the ground-truth in emotion type, direction, and intensity.
37 4. stance_consistency
38 Assess whether the generated comment expresses the same stance (supportive, opposed, skeptical, neutral, ironic, etc.)
  toward the post, event, or opinion.
39 5. style_fidelity
40 Assess whether the generated comment resembles the target user's language style: length, tone, punctuation, emoji usage,
  slang, sentence patterns, formality, and humor style.
41 6. persona_memory_consistency
42 Assess whether the generated comment is consistent with the target user's historical preferences, interests, values, and
  long-term behavioral tendencies, and avoids fabricating user facts.
43 7. naturalness
44 Assess whether the generated comment sounds like a real social media comment rather than AI-generated text, a summary, or
  an essay-like response.
45 [Overall Score]
46 overall_score = unweighted average of the seven dimension scores, rounded to two decimal places.
47 [Output] JSON only. No markdown. No explanatory preface. All dimension scores must be integers from 0 to 100.
48 {
49   "semantic_alignment":      {"score": 0, "reason": ""},
50   "intent_alignment":       {"score": 0, "reason": ""},
51   "emotion_consistency":    {"score": 0, "reason": ""},
52   "stance_consistency":     {"score": 0, "reason": ""},
53   "style_fidelity":         {"score": 0, "reason": ""},
54   "persona_memory_consistency": {"score": 0, "reason": ""},
55   "naturalness":            {"score": 0, "reason": ""},
56   "overall_score": 0.00,
57   "summary": ""
58 }
```

Table 15: A reference prompt for seven-dimension semantic evaluation.