DEBATE: A Large-Scale Benchmark for Role-Playing LLM Agents in Multi-Agent, Long-Form Debates

Yun-Shiuan Chuang[1] **Ruixuan Tu**[1†] **Chengtao Dai**[1†] **Smit Vasani**[1†] **Binwei Yao**[1]

Michael Henry Tessler[3] **Sijia Yang**[1] **Dhavan Shah**[1] **Robert Hawkins**[2]

Junjie Hu[1] **Timothy T. Rogers**[1]

[1]University of Wisconsin–Madison [2]Stanford University [3]Google DeepMind
{yunshiuan.chuang, ruixuan.tu, cdai53, svasani}@wisc.edu
{binwei.yao, syang84, dshah, junjie.hu, ttrogers}@wisc.edu
rdhawkins@stanford.edu mhtessler@google.com

Abstract

Accurately modeling opinion change through social interactions is crucial for addressing issues like misinformation and polarization. While role-playing large language models (LLMs) offer a promising way to simulate human-like interactions, existing research shows that single-agent alignment does not guarantee authentic multi-agent group dynamics. Current LLM role-play setups often produce unnatural dynamics (e.g., premature convergence), without an empirical benchmark to measure authentic human opinion trajectories. To bridge this gap, we introduce DEBATE, the first large-scale empirical benchmark explicitly designed to evaluate the authenticity of the *interaction* between multi-agent role-playing LLMs. DEBATE contains 29,417 messages from multi-round debate conversations among over 2,792 U.S.-based participants discussing 107 controversial topics, capturing both publicly-expressed messages and privately-reported opinions. Using DEBATE, we systematically evaluate and identify critical discrepancies between simulated and authentic group dynamics. We further demonstrate DEBATE's utility for aligning LLMs with human behavior through supervised fine-tuning, achieving improvements in surface-level metrics (e.g., ROUGE-L and message length) while highlighting limitations in deeper semantic alignment (e.g., semantic similarity). Our findings highlight both the potential and current limitations of role-playing LLM agents for realistically simulating human-like social dynamics.

1 Introduction

Understanding how individual opinions change through social interactions is crucial across numerous domains, e.g., public health campaigns, conflict resolution, and misinformation mitigation [19, 23, 2, 18, 11]. Accurate modeling of these dynamics not only helps predict critical societal phenomena like opinion polarization but also informs effective interventions to mitigate adverse outcomes.

Recent advances in large language models (LLMs) have unlocked new possibilities for simulating human social interactions, particularly through the use of role-playing LLM agents that embody diverse personas and engage in multi-turn dialogue [22, 5, 6]. Although individual LLM agents

[†]Joint second authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Scaling Environments for Agents (SEA)

can often convincingly emulate human-like behaviors, prior research indicates that this single-agent authenticity does not guarantee realistic emergent dynamics in multi-agent settings. Specifically, when multiple role-playing LLM agents interact, they frequently exhibit premature consensus convergence, overly moderate stances, or unnatural patterns of opinion alignment, regardless of their initial diverse personas [5, 28]. Existing evaluations of role-playing LLM agents predominantly focus on single-agent scenarios or employ artificial, structured tasks, lacking robust empirical benchmarks capturing authentic human group dynamics in naturalistic contexts [26, 7, 6].

To address this critical gap, we introduce **D**atas**E**t for **B**enchmarking Multi-**A**gent Opinion **T**rajectories and **E**volution (**DEBATE**), the first large-scale empirical benchmark specifically designed for evaluating the authenticity of multi-agent role-playing LLM systems. DEBATE comprises data from 2,792 U.S.-based participants, organized into small groups engaged in multi-round, dyadic debates on a diverse set of controversial topics. Capturing both publicly expressed opinions (tweet-like messages) and privately reported beliefs (Likert-scale ratings), DEBATE provides fine-grained records of naturalistic opinion trajectories, enabling quantitative assessment of alignment between simulated and actual human interactions. Below is a summary of our contributions:

- **Empirical benchmark for multi-agent opinion dynamics.** We present **DEBATE**, the first large-scale benchmark for evaluating the *human-likeness* of multi-agent role-playing LLMs. It contains over 29,000 messages from 2,792 U.S. participants engaging in multi-round, multi-player debates across 107 controversial topics, with both tweet-like posts and messages, as well as private self-reported Likert-scale opinions. The dataset supports both in-depth and broad coverage analyses, as well as quantitative and qualitative evaluation of multi-agent role-playing LLM system.
- **Support for three common simulation setups.** DEBATE enables three simulation scenarios: (1) *Next Message Prediction*, (2) *Tweet-guided Conversation Simulation*, and (3) *Full Conversation Simulation*, covering a range of applications such as social forecasting, simulated social experiments, persuasion strategy development.
- **Evaluation at utterance, individual, and group levels.** We introduce metrics that assess alignment (*human-likeness*) of the multi-agent role-playing LLM system across three levels: *utterance-level* (e.g., semantic similarity, stance alignment), *individual-level* (e.g., regression to the mean, partner influence), and *group-level* (e.g., opinion convergence and shift).
- **Exploration of supervised fine-tuning (SFT).** Fine-tuning LLM agents on human data improves surface-level metrics (e.g., ROUGE-L, response length), but fails to enhance deeper semantic or stance alignment. This highlights the need for future work on better training methods.
- **Behavioral gaps in simulated opinion dynamics.** Compared to humans, LLM agents' groups show (1) stronger convergence in both public and private opinions, (2) positive drift in tweet stance, and (3) greater regression to the mean and partner's influence. These gaps point to challenges in emulating human-like opinion dynamics with multi-agent LLM systems.

2 Related Work

Multi-Agent Debate with LLMs. Recent work explores multi-agent debate as a means to improve task performance in LLMs, such as reasoning accuracy, factuality, and diversity of outputs [29, 3, 4, 8, 16, 13]. However, these approaches primarily use debate as a technique to boost performance rather than to authentically model human-like interactions. A separate line of work investigates emergent behaviors in multi-agent LLM simulations, attempting to reproduce human-like debate using role-playing LLM agents [5, 28]. Yet these studies lack empirical benchmarks to systematically evaluate the human-likeness of the simulated debates. In contrast, our work focuses on the fidelity of multi-agent simulations, using real human conversational trajectories as ground truth for evaluation.

Evaluating Human-Likeness in Role-Playing LLMs. Recent benchmarks have evaluated the human-likeness of role-playing LLM agents using empirical human data, but most focus on single-agent settings without interaction [26, 7], or on non-linguistic, artificial setups such as numeric guesstimation [6]. Our work extends this line by introducing a benchmark that captures multi-turn, multi-agent opinion trajectories in natural language, enabling evaluation of both individual- and group-level dynamics.

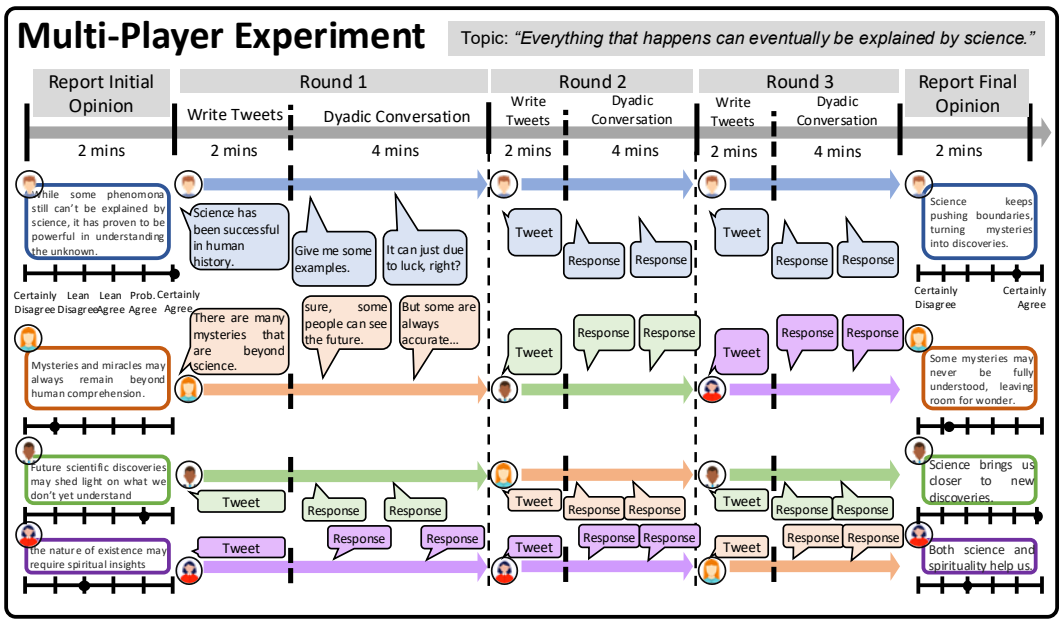


Figure 1: The procedure of the multi-player experiment. Each group is assigned a topic to discus about. Participants first report their initial opinion, then engage in three rounds of tweet writing and dyadic conversations with different partners, and finally submit their final opinion. With this setup, we collects naturalist opinion exchanges among groups.

Table 1: Dataset statistics. Each row reports the number of topics, messages, on-topic messages, subjects, conversation groups, and the average number of groups per topic. Depth topics has more groups per topic, while breadth topics span a wider range of themes with fewer groups per topic.

Dataset	# topics	# messages	# on-topic messages	# subjects	# groups	# groups/topic
Depth	7	5906	5252	501	185	26.43
Breadth	100	23511	23327	2291	612	6.12
Depth+Breadth	107	29417	28579	2792	797	7.45

3 DEBATE Benchmark: Empirical Opinion Dynamics from Human Groups

3.1 Task

We design a multi-agent conversational experiment to elicit naturalistic opinion exchanges and track empirical opinion dynamics (Figure 1). The dataset comprises G groups, each consisting of $N = 4$ participants $\{s_1, s_2, s_3, s_4\}$. Each group is randomly assigned a single controversial discussion topic $t \in \mathcal{T}$, which remains fixed throughout the session. For each four-person experiment, the experiment lasts approximately 25–30 minutes per group and consists of four phases:

(1) Initial Opinion: Each participant s_i provides an initial opinion $o_{s_i}^{\text{init}} \in \{-2.5, -1.5 \ldots, +2.5\}$ on a 6-point Likert scale[1], along with a free-text justification $j_{s_i}^{\text{init}}$, submitted within a 2-minute window.

(2) Opinion Exchanges: Participants engage in $R = 3$ rounds of dyadic conversation. In each round r, participants are randomly paired with one of the other group members who they haven't interacted with yet. Across three rounds, each participant interacts with every other group member exactly once. For each pair of distinct participants (s_i, s_j), where $i \neq j$:

- Each participant first writes a tweet-like post $\tau_{s_i}^r$ within 2 minutes, summarizing their opinion on the assigned topic.

[1]Participants selected from the six labels displayed in the interface: *(−2.5) Certainly disagree*, *(−1.5) Probably disagree*, *(−0.5) Lean disagree*, *(+0.5) Lean agree*, *(+1.5) Probably agree*, *(+2.5) Certainly agree*.

- After submitting their tweets, participants view each other's post and engage in a 4-minute real-time conversation via a chatbox interface. The conversation is represented as an ordered sequence: $\mathcal{C}^r_{s_i,s_j} = \left[u^r_{1,s_i}, u^r_{2,s_j}, u^r_{3,s_i}, \dots\right]$, where $u^r_{k,s}$ denotes the k-th utterance in the round-r conversation, with speaker $s \in \{s_i, s_j\}$. Speaker turns alternate between participants. Consecutive messages from the same speaker are merged during data preprocessing.

(3) Post-discussion Opinion: After the final round, each participant submits a final opinion $o^{\text{final}}_{s_i}$ and justification $j^{\text{final}}_{s_i}$ within a 2-minute response window.

(4) Demographic Survey: Finally, participants report demographic attributes d_{s_i} (e.g., age, gender, education, political orientation), with no time limit.

3.2 Topic

The DEBATE benchmark includes two complementary topic sets: *Depth* ($\mathcal{T}_{\text{Depth}}$) and *Breadth* ($\mathcal{T}_{\text{Breadth}}$). Depth topics support in-depth analysis of opinion dynamics, while Breadth topics enable generalization assessment across diverse issues.

Depth Topics. $\mathcal{T}_{\text{Depth}}$ comprises seven topics selected from a prior study, each tied to a known scientific consensus or "ground truth." An example is: *"The position of the planets at the time of your birth can influence your personality."* Prior work shows that LLM agents often drift toward ground-truth views over time, regardless of initial opinions [5, 28]. To reproduce this challenge, we select these topics with known ground-truth and with high-entropy topics to elicit diverse opinions from human [7]. Each topic is assigned to an average of 16.14 groups (456 participants in total; Table 1). See Appendix A for the full list. Below is a summary of our contributions.

Breadth Topics. $\mathcal{T}_{\text{Breadth}}$ contains 100 topics from the World Values Survey (WVS) [12] and Pew Global Attitudes Survey (PGAS) [24]. To reflect public disagreement in our US-based participants, we select U.S.-administered questions with the highest response entropy [10]. Topics are phrased as self-contained declarative statements (e.g., *"Euthanasia can be justified."*) and span domains such as science, policy, and social values. These topics are not linked to ground truths but reflect a wide range of viewpoints. On average, each topic is assigned to 3.90 groups (1560 participants in total; Table 1). See Appendix B and Table 5 for examples and construction details.

3.3 Human Data Collection and Dataset Summary

We recruited 2,792 unique participants who reside in the U.S. via the Prolific platform [21][2]. Participants were randomly assigned to one of 797 four-person groups and to a discussion topic. They remained anonymous to each other, identified only by avatars and randomly generated pseudonyms (e.g., `ZK48UT`). All procedures were approved by an Institutional Review Board (IRB), and participants were compensated at a rate of $10/hour. The Depth subset comprises 7 topics, each discussed by an average of 26.43 groups (501 participants total), while the Breadth subset spans 100 topics with 6.12 groups per topic (2,291 participants total). Across both subsets, the dataset includes 29,417 messages. Table 1 summarizes key statistics of the dataset.

The participants span a broad range of ages (18–83, $M = 39.5$, $SD = 13.0$), genders (50.2% male, 49.0% female), ethnicities (e.g., 66.4% White, 24.7% Black, 5.5% Asian, 5.1% Hispanic), educational backgrounds (ranging from high school to doctoral degrees), and income levels (from under $25k to over $200k). Participants also report a wide variety of occupations (e.g., finance, engineering, healthcare, manufacturing). This diversity provides a robust foundation for modeling opinion dynamics across varied social perspectives (see Appendix C and Figure 4 for details). Upon acceptance, the dataset will be released and a link will be included in the final version of the paper.

[2] `https://www.prolific.com/`

4 From Human Data to Role-Playing LLM Agents: Agent Construction, Interaction Simulation, and Evaluation

4.1 LLM Role-play Agent Construction Grounded in Human Data

Each role-playing LLM agent a_i is designed as a *digital twin* of a human participant s_i, simulating s_i's conversational behavior throughout the multi-round interaction. To achieve this, the agent is conditioned on a memory module $\mathcal{M}_{a_i,k}$ that aims to reflect s_i's first-person perspective right before producing the k-th utterance in round r. The memory is dynamically updated as tweets and utterances are exchanged.

The memory module $\mathcal{M}_{a_i,k}$ is instantiated via prompt templates that convert structured information into natural language inputs for the LLM (see Appendix D and Table 6 for prompt examples). We use notation with a hat and subscript a (e.g., $\widehat{\tau}^r_{a_i}$, $\widehat{u}^r_{k,a_i}$) to denote LLM-generated content, and notation without a hat and with subscript s (e.g., $\tau^r_{s_i}$, u^r_{k,s_i}) to denote human-written content.

At each turn k in round r, the agent memory $\mathcal{M}_{a_i,k}$ includes: **1. Demographic Profile (d_{s_i}):** Age, gender, education, income, ethnicity, marital status, residence, parental status, political ideology, religiosity, and occupation. **2. Initial Opinion ($o^{\text{init}}_{s_i}, j^{\text{init}}_{s_i}$):** A 6-point Likert-scale opinion on the assigned discussion topic and a free-text justification. **3. Initial Tweet ($\tau^1_{s_i}$):** The tweet posted at the beginning of round 1. **4. Previous Rounds:** Tweets $\{\tau^{r'}_{s_i}, \widehat{\tau}^{r'}_{a_i} : 1 < r' < r\}$ and dyadic conversations $\{\mathcal{C}^{r'}_{s_i,s_j}, \widehat{\mathcal{C}}^{r'}_{a_i,a_j} : 1 \leq r' < r\}$ from earlier rounds involving participant s_i. **5. Current Round Context:** The current tweet $\tau^r_{s_i}$ or $\widehat{\tau}^r_{a_i}$, the partner's tweet $\tau^r_{s_j}$ or $\widehat{\tau}^r_{a_j}$, and all utterances so far in the ongoing conversation $\{u^r_{k',s}, \widehat{u}^r_{k',a} : 1 \leq k' < k\}$. The exact sources of memory vary by the simulation mode (Section 4.1; Table 2). For example, the conversation history may come from real human (Mode 1), LLM simulation (Mode 3), or a mix of both (Mode 2).

4.2 Simulating Social Interactions with Role-playing LLM Agents

Table 2: Memory contents used in each simulation mode. All agents are conditioned on demographics d_{s_i}, initial opinion and justification $(o^{\text{init}}_{s_i}, j^{\text{init}}_{s_i})$, the initial tweet $\tau^1_{s_i}$, and task instructions. Blue entries indicate simulated content recursively generated by the model and added to the memory.

Simulation Mode	Tweets in Memory	Utterances from Prior Rounds	Utterances from Current Round	Application and Scenario
Mode 1: Next Message Prediction	Human $\{\tau^{r'} : 1 \leq r' \leq r\}$	Human $\{\mathcal{C}^{r'}_{s_i,s_j} : 1 \leq r' < r\}$	Human $\{u^r_m : 1 \leq m < k\}$	Predict a person's immediate response in real conversations
Mode 2: Tweet-guided Conversation Simulation	Human $\{\tau^{r'} : 1 \leq r' \leq r\}$	Simulated $\{\widehat{\mathcal{C}}^{r'}_{a_i,a_j} : 1 \leq r' < r\}$	Simulated $\{\widehat{u}^r_m : 1 \leq m < k\}$	Simulate private conversations given a trace of real public tweets
Mode 3: Full Conversation Simulation	Human τ^1 + Simulated $\{\widehat{\tau}^{r'} : 2 \leq r' \leq r\}$	Simulated $\{\widehat{\mathcal{C}}^{r'}_{a_i,a_j} : 1 \leq r' < r\}$	Simulated $\{\widehat{u}^r_m : 1 \leq m < k\}$	Simulate agents' dynamics from initial conditions; the **classic opinion dynamics simulation** setup

We simulate each role-playing LLM agent a_i's utterance $\widehat{u}^r_{k,a_i}$ in round r, turn k by generating:

$$\widehat{u}^r_{k,a_i} \sim P\left(u^r_{k,s_i} \,\middle|\, \mathcal{M}_{a_i,k}\right), \tag{1}$$

where the speaker identity s_i is given, and only the utterance content is predicted.[3] The same framework applies to generating tweets $\widehat{\tau}^r_{a_i}$ and final opinions $(\widehat{o}^{\text{final}}_{a_i}, \widehat{j}^{\text{final}}_{a_i})$.

We define three simulation modes corresponding to three real-world scenarios in social interaction simulations.The DEBATE dataset enables these three modes, all grounded in real human behavior data but varying in how much human context is provided to the model: Mode 1 (*Next Message Prediction*), Mode 2 (*Tweet-guided Conversation Simulation*), and Mode 3 (*Full Conversation Simulation*). This setup allows researchers to study different aspects of multi-agent communication, from immediate message prediction to end-to-end full trajectory generation from initial state. Each simulation is conditioned on the memory module $\mathcal{M}_{a_i,k}$, which includes basic information such as demographics d_{s_i}, initial opinion and justification $(o^{\text{init}}_{s_i}, j^{\text{init}}_{s_i})$, the initial tweet $\tau^1_{s_i}$, and task instructions (Section 4.1).

[3] Since consecutive messages from the same speaker are merged during preprocessing, speakers alternate turns, making the speaker order known (Section 3.1).

What varies across simulation modes is the source of tweets and conversational history—whether they come from real human data or are recursively generated and added to the memory. Table 2 summarizes the full memory configuration and the corresponding use case for each simulation mode.

4.3 Evaluation

We evaluate how well a role-playing LLM agent a_i simulates its corresponding human participant s_i by comparing utterances $\widehat{u}$ and u within the dyadic conversations. Evaluation is conducted only on *on-topic* utterances—those directly addressing the discussion topic t—excluding conversational fillers (e.g., *"hello"*, *"what do you think?"*) or unrelated remarks (e.g., *"which football team do you support?"*). We compute the following metrics to assess different aspects of alignment: **1. Semantic Similarity:** $S_{\text{sem}}(u, \widehat{u}) = \cos(E(u), E(\widehat{u}))$, where $E(\cdot)$ is a sentence encoder. This measures the meaning-level similarity between utterances, capturing whether the agent expresses a semantically similar idea. **2. Stance Difference:** $\Delta_{\text{stance}}(u, \widehat{u}) = |S(u) - S(\widehat{u})|$, using scalar stance scores in $[-2.5, -1.5, -0.5, +0.5, +1.5, +2.5]$. This captures alignment in opinion polarity, assessing whether the agent expresses a similar stance. **3. Length Metrics:** $\Delta_{\text{abs_len}} = \big||u| - |\widehat{u}|\big|$; $\Delta_{\text{signed_len}} = |u| - |\widehat{u}|$. These reflect surface-level stylistic similarity in verbosity and message length. **4. ROUGE-L:** Longest common subsequence score [17]. This quantifies token-level overlap, capturing whether the agent reuses similar lexical structures. **5. On-topic Utterance Rate ($R_{\text{on-topic}}$):** We also report the proportion of generated utterances that are judged on-topic: $R_{\text{on-topic}} = \frac{1}{|\widehat{\mathcal{U}}|} \sum_{\widehat{u} \in \widehat{\mathcal{U}}} I_{\text{topic}}(\widehat{u}, t)$. For reference, human utterances are on-topic 58% (Depth) and 66% (Breadth) of the time. While $R_{\text{on-topic}}$ does not directly reflect alignment, it offers insight into how focused the simulated agents remain. Note that stance scores $S(\cdot)$ and topic relevance indicators $I_{\text{topic}}(\cdot, t)$ are predicted by `gpt-4o-mini-2024-07-18`, validated against human annotations (Appendix E).

Because there is no one-to-one mapping between simulated and human utterances, we adopt a *round-wise aggregated* evaluation: each simulated utterance $\widehat{u}$ is compared to all on-topic human utterances u from the same round and speaker. We average metric scores across utterances, agents, and rounds, yielding $\overline{S}_{\text{sem}}$, $\overline{\Delta}_{\text{stance}}$, $\overline{\Delta}_{\text{abs_len}}$, $\overline{\Delta}_{\text{signed_len}}$, and $\overline{\text{ROUGE-L}}$ (see Appendix F for details).

4.4 Large Language Models (LLMs)

We evaluate six LLMs: `gpt-4o-mini-2024-07-18` [20], `Llama-3.1-Tulu-3-8B-SFT` [15], `Llama-3.1-8B-Instruct` [9], `Llama-3.1-70B-Instruct`, `Mistral-7B-Instruct-v0.3` [14], and `Qwen2.5-32B-Instruct` [1]. Our selection covers variation in license (e.g., `gpt-4o-mini` vs. open-weight models), model scale (8B vs. 70B), pre- vs. post-RLHF checkpoints (`Tulu-3` vs. `Llama-3.1-Instruct`), and reasoning model (`Qwen2.5`). See Appendix G for compute details.

5 Utterance-level Evaluation of Role-playing LLM Agents

Alignment Across Three Social Simulation Modes. Tables 3 and 9 report evaluation results across simulation modes and LLMs for Depth and Breadth topics, respectively. Two consistent trends emerge across all metrics and topic types. First, `gpt-4o-mini-2024-07-18` consistently shows the strongest alignment with human responses, achieving the best scores on semantic similarity ($\overline{S}_{\text{sem}}$), ROUGE-L ($\overline{\text{ROUGE-L}}$), and stance difference ($\overline{\Delta}_{\text{stance}}$). To account for variability across topics and simulation conditions, we conduct statistical tests across six experimental settings. A Friedman test followed by Wilcoxon signed-rank tests confirms that `gpt-4o-mini` significantly outperforms most other models across all three metrics (see Appendix H for full results). However, it tends to produce longer messages than humans, as indicated by the negative signed length difference $\overline{\Delta}_{\text{signed_len}}$. Second, as expected, alignment performance declines across simulation modes: Mode 1 (Next Message Prediction) performs best, followed by Mode 2 (Tweet-guided Conversation), and Mode 3 (Full Conversation) performs worst. This trend shows that providing more human-grounded context improves behavioral alignment, with alignment degrading as the simulation relies increasingly on model-generated conversation history.

Ablation Studies. To assess the contribution of different memory components in role-playing LLM agents, we conduct a set of ablation experiments by systematically removing individual parts of the memory module $\mathcal{M}_{a_i,k}$ (Section 4.1). Each ablation isolates the effect of a specific type of information

Table 3: Evaluation results across simulation modes and LLMs. We report the round-wise aggregated metrics on the **Depth Topics**: average semantic similarity $\overline{S}_{\text{sem}}$ (↑), average stance difference $\overline{\Delta}_{\text{stance}}$ (↓), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (→0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (↓), average ROUGE-L $\overline{\text{ROUGE-L}}$ (↑), and on-topic utterance rate $R_{\text{on-topic}}$.

LLM & Simulation Mode	$\overline{S}_{\text{sem}}$ (↑)	$\overline{\Delta}_{\text{stance}}$ (↓)	$\overline{\Delta}_{\text{signed_len}}$ (→0)	$\overline{\Delta}_{\text{abs_len}}$ (↓)	$\overline{\text{ROUGE-L}}$ (↑)	$R_{\text{on-topic}}$
Simulation Mode 1: Next Message Prediction						
gpt-4o-mini-2024-07-18	**0.49**	**1.13**	-33.44	34.21	**0.11**	0.91
Llama-3.1-Tulu-3-8B-SFT	0.45	1.20	-45.00	46.69	0.06	0.77
Llama-3.1-8B-Instruct	0.46	1.22	-39.02	40.04	0.08	0.88
Llama-3.1-70B-Instruct	0.46	1.15	-30.80	32.23	0.08	0.92
Mistral-7B-Instruct-v0.3	0.48	1.17	-47.20	47.59	0.08	0.90
Qwen2.5-32B-Instruct	0.46	1.16	**-23.75**	**28.22**	0.08	0.90
Simulation Mode 2: Tweet-guided Conversation Simulation						
gpt-4o-mini-2024-07-18	**0.42**	1.25	-58.93	59.10	**0.09**	0.71
Llama-3.1-Tulu-3-8B-SFT	0.41	1.33	-54.33	54.99	0.05	0.58
Llama-3.1-8B-Instruct	0.41	1.28	-54.62	55.03	0.06	0.68
Llama-3.1-70B-Instruct	0.40	**1.18**	-56.67	57.04	0.06	0.75
Mistral-7B-Instruct-v0.3	0.41	1.21	**-47.71**	**48.09**	0.06	0.67
Qwen2.5-32B-Instruct	0.41	1.25	-49.67	51.32	0.07	0.69
Simulation Mode 3: Full Conversation Simulation						
gpt-4o-mini-2024-07-18	**0.41**	1.30	-58.48	58.65	**0.08**	0.69
Llama-3.1-Tulu-3-8B-SFT	0.40	1.47	-55.36	56.23	0.05	0.54
Llama-3.1-8B-Instruct	0.40	1.33	-54.46	54.96	0.06	0.67
Llama-3.1-70B-Instruct	0.38	1.28	-55.61	55.95	0.06	0.75
Mistral-7B-Instruct-v0.3	0.40	**1.26**	**-47.68**	**48.11**	0.06	0.63
Qwen2.5-32B-Instruct	0.40	1.30	-49.66	51.22	0.07	0.67

on agent behavior. We consider the following ablation conditions: **1. No Previous Chat:** Removes all conversational history from previous rounds, including tweets $\{\tau_{s_i}^{r'}, \widehat{\tau}_{a_i}^{r'} : 1 \leq r' < r\}$ and dyadic conversations $\{\mathcal{C}_{s_i,s_j}^{r'}, \widehat{\mathcal{C}}_{a_i,a_j}^{r'} : 1 \leq r' < r\}$. **2. No Initial Opinion:** Excludes the participant's initial Likert-scale opinion $o_{s_i}^{\text{init}}$ and justification $j_{s_i}^{\text{init}}$ from memory. **3. No Demographics:** Omits the demographic profile d_{s_i}, removing background information such as age, gender, education, and political ideology. **4. No Private Profile:** Omits both the demographic profile and the initial opinion (d_{s_i}, $o_{s_i}^{\text{init}}$, $j_{s_i}^{\text{init}}$), removing all private information about the human participant. All other components of memory remain unchanged in each condition.

Tables 4 and 10 show the ablation results on Depth topics and Breadth topics, respectively, using `gpt-4o-mini`. We observe that in Simulation Mode 1, where the full human conversation history is available, ablating prior chats or private profile information has minimal impact on semantic alignment. In contrast, for Modes 2 and 3, where the conversation history is simulated and recursively added to agents' memory, removing private profile information consistently degrades performance in terms of both semantic similarity and stance alignment. This highlights the importance of grounding agents with actual human private information when we want to simulate conversation recursively.

Supervised Fine-tuning (Appendix K). To test whether behavioral alignment can be improved through fine-tuning, we conducted preliminary experiments using supervised fine-tuning (SFT) on the Depth subset. While SFT improves surface-level alignment (e.g., message length, ROUGE-L), it fails to enhance deeper metrics such as semantic similarity or stance alignment. The mixed results suggest that naive SFT does not robustly improve simulated opinion trajectories. Developing training methods that explicitly target alignment in opinion trajectory remains an important direction for future work.

6 Opinion Dynamics: Evaluating Group and Individual Opinion Alignment

Beyond *utterance-level* alignment, realistic simulations must capture *group-level* and *individual-level* opinion dynamics. Utterance similarity alone does not ensure emergent group behaviors (e.g., opinion convergence) are reproduced. We focus on the groups discussing Depth topics, using Simulation

Table 4: Ablation results across simulation modes using `gpt-4o-mini-2024-07-18` on the Depth Topics. We report average semantic similarity $\overline{S}_{\text{sem}}$ (↑), average stance difference $\overline{\Delta}_{\text{stance}}$ (↓), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (→ 0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (↓), average ROUGE-L $\overline{\text{ROUGE-L}}$ (↑), and on-topic utterance rate $R_{\text{on-topic}}$. Blue cells indicate improved performance after ablation, while red cells indicate worsened performance after ablation.

Ablation Condition	S_{sem} (↑)	Δ_{stance} (↓)	$\Delta_{\text{signed_len}}$ (→ 0)	$\Delta_{\text{abs_len}}$ (↓)	ROUGE-L (↑)	$R_{\text{on-topic}}$
Simulation Mode 1: Next Message Prediction						
Original	0.49	1.12	-34.01	34.80	0.11	0.91
No Private Profile	0.49	1.09	-32.68	33.48	0.11	0.92
No Demographics	0.49	1.07	-31.28	32.23	0.11	0.92
No Initial opinion	0.48	1.08	-33.23	34.04	0.11	0.91
No Prior Chats	0.49	1.12	-39.58	40.09	0.10	0.91
Simulation Mode 2: Tweet-guided Conversation Simulation						
Original	0.43	1.24	-58.31	58.47	0.09	0.71
No Private Profile	0.42	1.34	-58.12	58.31	0.09	0.74
No Demographics	0.42	1.37	-57.12	57.41	0.09	0.75
No Initial opinion	0.43	1.32	-57.28	57.62	0.09	0.72
No Prior Chats	0.44	1.29	-56.52	56.87	0.09	0.78
Simulation Mode 3: Full Conversation Simulation						
Original	0.42	1.33	-57.71	57.91	0.09	0.69
No Private Profile	0.42	1.35	-57.21	57.43	0.09	0.76
No Demographics	0.40	1.35	-57.41	57.77	0.08	0.72
No Initial opinion	0.41	1.36	-57.69	58.02	0.08	0.65
No Prior Chats	0.43	1.35	-56.43	56.75	0.09	0.77

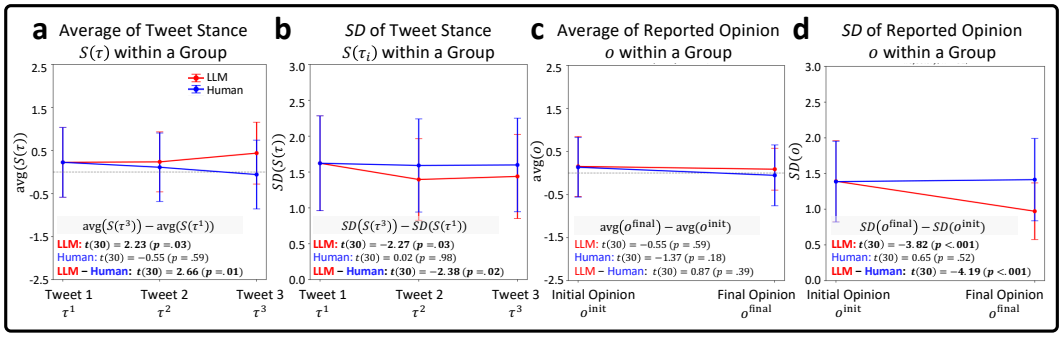


Figure 2: Group-level trajectories of tweet stance and self-reported opinion for human groups (blue) and their corresponding LLM groups (red). (a) Average tweet stance $S(\tau)$ within each group across three rounds. (b) Standard deviation (*SD*) of tweet stance $S(\tau)$ within each group across rounds. (c) Average self-reported opinion o within each group from initial to final measurement. (d) *SD* of self-reported opinion o within each group. Values are averaged across all Depth-topic groups. Error bars indicate the standard error across groups. Below each panel, paired t-test results assess whether the change from Tweet 1 to Tweet 3 (or from initial to final opinion) is significant; significant results are **boldfaced**. Differences in change between human and LLM groups are also statistically tested.

Mode 3 (Full Conversation Simulation; Section 4.2) because it mirrors classic opinion dynamics setups and the model with best semantic alginemnt (`gpt-4o-mini-2024-07-18`; Section 5).

Group-Level Opinion Shifts: Public Tweet Stance vs. Private Self-Reported Opinion. We evaluate group-level opinion change by comparing (public) tweet stance $S(\tau^3) - S(\tau^1)$ and (private) self-reported opinion $o^{\text{final}} - o^{\text{init}}$ across rounds, using paired t-tests between each human group and its corresponding LLM-simulated group (i.e., digital twins). Figure 2a shows that LLM groups significantly increase their average tweet stance from round 1 to round 3 ($t(30) = 2.23$, $p = .03$), while human groups show no significant change ($t(30) = -0.55$, $p = .59$). The difference between these changes is statistically significant ($t(30) = 2.66$, $p = .01$), indicating that LLM groups tend to become more positive in their stance (i.e., endorsing the false statements in the Depth topics). On the other hand, Figure 2b shows that LLM groups exhibit a significant reduction in the standard deviation (*SD*) of tweet stance over time ($t(30) = -2.27$, $p = .03$), suggesting opinion convergence within the

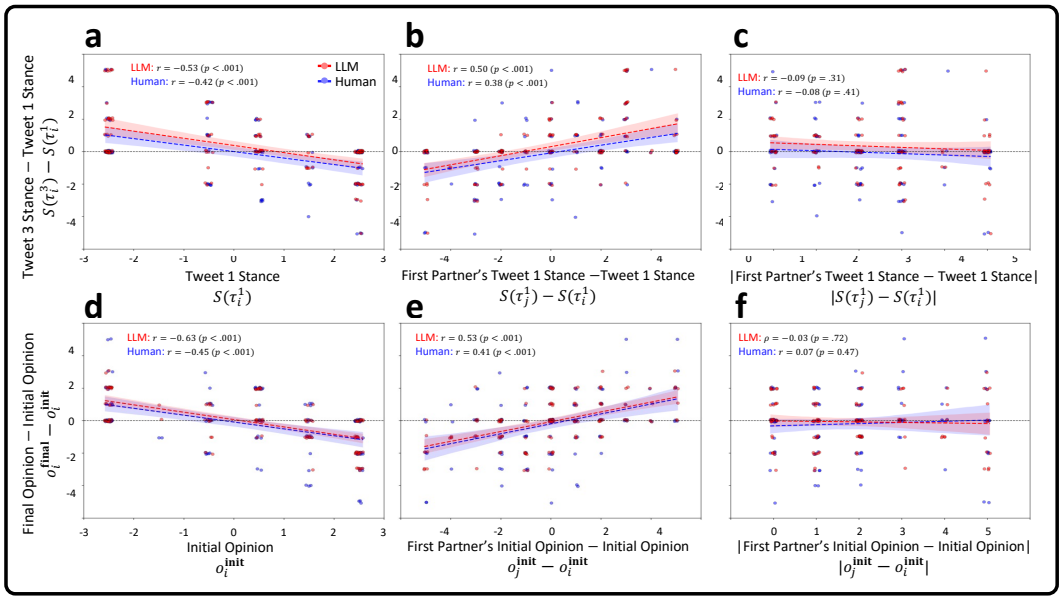


Figure 3: Individual-level opinion change and its predictors. (a) Change in tweet stance ($S(\tau_i^3) - S(\tau_i^1)$) negatively correlates with initial stance $S(\tau_i^1)$, (b) positively correlates with directional difference between first partner's stance and own stance, and (c) has no relationship when using absolute stance difference. (d–f) Replicate the above with self-reported opinions ($o_i^{\text{final}} - o_i^{\text{init}}$). Shaded regions show standard error.

group. In contrast, human groups show no change in *SD* ($t(30) = 0.02$, $p = .98$), and the difference in *SD* change between LLM and human groups is significant ($t(30) = -2.28$, $p = .02$).

We conduct the same analyses for self-reported opinions (Figures 2c,d). Both LLM and human groups exhibit no significant change in average self-reported opinions (LLM: $t(30) = -0.55$, $p = .59$; Human: $t(30) = -1.37$, $p = .18$). In terms of *SD*, LLM groups show a significant reduction in *SD* ($t(30) = -4.65$, $p < .001$), while human groups do not ($t(30) = 0.65$, $p = .52$). The difference in *SD* change between LLM and human groups is significant ($t(30) = -4.19$, $p < .001$).

In sum, these results reveal two key findings. First, compared to humans, LLM groups exhibit stronger opinion convergence in both public tweet stance and private self-reported opinion. Second, within LLM groups, public tweet stance becomes increasingly positive over time, while private self-reported opinions remain unchanged, indicating a dissociation between the two measurements.

Mechanisms of Individual Opinion Change. We examine how individuals update their public tweet stance and private self-reported opinion across rounds, focusing on two mechanisms: regression to the mean and partner influence. Figures 3a–c analyze tweet stance change $S(\tau_i^3) - S(\tau_i^1)$. Individuals with more extreme initial stances tend to move toward the midpoint (Figure 3a), with significant negative correlations for humans ($r = -0.42$, $p < .001$) and LLMs ($r = -0.53$, $p < .001$). Individuals also shift toward their first partner's tweet stance (Figure 3b; Human: $r = 0.38$, $p < .001$; LLM: $r = 0.50$, $p < .001$). As a control, absolute difference from their first partner's stance has no effect (Figure 3c). Simiarly, Figures 3d–f show analogous findings for self-reported opinion change $o_i^{\text{final}} - o_i^{\text{init}}$. Individuals with more extreme initial opinions shift toward the midpoint (Figure 3d; Human: $r = -0.45$, $p < .001$; LLM: $r = -0.63$, $p < .001$) and toward their first partner's initial opinion (Figure 3e; Human: $r = 0.41$, $p < .001$; LLM: $r = 0.53$, $p < .001$), again with no effect of absolute difference (Figure 3f; Human: $r = 0.07$, $p = .47$; LLM: $r = -0.03$, $p = .72$).

Across both public tweets and private opinions, LLM agents consistently show stronger correlations than humans, both in their tendency to regress toward the mean and in how they get influenced by their partner's opinion. This suggests that LLM agents exhibit more systematic shifts than their human counterparts, consistent with the group-level findings.

Summary. We identify three key differences in opinion dynamics between LLM agents and humans. LLM groups show stronger opinion convergence, positive drift in public tweet stance, and more systematic individual shifts including stronger regression to the mean and greater susceptibility to

partner influence. These gaps highlight the difficulty of modeling human-like opinion dynamics using role-playing LLMs.

7 Conclusion

We introduced **DEBATE**, the first large-scale empirical benchmark for evaluating multi-agent role-playing LLM systems. By capturing rich, naturalistic opinion trajectories from 2,792 U.S.-based participants across multi-round, multi-party interactions, DEBATE enables fine-grained evaluation of simulated opinion dynamics at the utterance-, individual-, and group-levels. Our experiments reveal both promising capabilities and persistent challenges: while current LLM agents reproduce some utterance-level patterns, they fall short in deeper opinion alignment and belief updating. We propose a evaluation framework and identify systematic behavioral differences between human and LLM-simulated groups. We hope DEBATE provides a foundation for developing more socially grounded and human-aligned multi-agent LLM systems.

Acknowledgements

We thank the reviewers and the area chair for their feedback. This work was funded by the Multi University Research Initiative grant from the Department of Defense, W911NF2110317 (with Rogers as Co-I) and a Research Forward award from the University of Wisconsin-Madison (Rogers, PI).

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 665–674, 2011.

[3] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[4] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6, 2023.

[5] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, 2024.

[6] Yun-Shiuan Chuang, Nikunj Harlalka, Siddharth Suresh, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[7] Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Beyond demographics: Aligning role-playing llm-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14010–14026, 2024.

[8] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[10] Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *First Conference on Language Modeling*, 2024.

[11] Tamar Ginossar, Iain J Cruickshank, Elena Zheleva, Jason Sulskis, and Tanya Berger-Wolf. Cross-platform spread: vaccine-related content, sources, and conspiracy theories in youtube videos shared in early twitter covid-19 conversations. *Human vaccines & immunotherapeutics*, 18(1):1–13, 2022.

[12] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, et al. World values survey: Round seven–country-pooled datafile version 5.0, 2022. URL `https://www.worldvaluessurvey.org/`.

[13] Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. *arXiv preprint arXiv:2406.19643*, 2024.

[14] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[15] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[16] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[17] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013/`.

[18] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

[19] Wei Lu, Wei Chen, and Laks VS Lakshmanan. From competition to complementarity: comparative influence diffusion and maximization. *Proceedings of the VLDB Endowment*, 9(2):60–71, 2015.

[20] OpenAI. Introducing ChatGPT. `https://openai.com/blog/chatgpt`, 2022. [Accessed 13-10-2023].

[21] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, 17:22–27, 2018.

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

[23] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592 (7855):590–595, 2021.

[24] Pew Research Center. Pew research center: Numbers, facts and trends shaping your world. `https://www.pewresearch.org/`, 2025.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

[26] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[27] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task lora, 2024. URL `https://arxiv.org/abs/2409.10173`.

[28] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, 2024.

[29] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.

A Depth Topic Construction

The following seven topics are used as the Depth topic set ($\mathcal{T}_{\text{Depth}}$). These topics are selected from a prior study [7], which introduced a set of 64 topics, all associated with claims that are supported by scientific or factual evidence. We choose a subset of topics that exhibit high entropy in opinion (i.e., people tend to disagree with each other), making them suitable for evaluating opinion dynamics in human groups.

1. A "body cleanse," in which you consume only particular kinds of nutrients over 1–3 days, helps your body to eliminate toxins.
2. Angels are real.
3. Everything that happens can eventually be explained by science.
4. Regular fasting will improve your health.
5. The U.S. deficit increased after President Obama was elected.
6. The United States has the highest federal income tax rate of any Western country.
7. The position of the planets at the time of your birth can influence your personality.

All topics except one are framed using *false-framing*, meaning that disagreement with the statement aligns with the ground truth. The only exception is *"Everything that happens can eventually be explained by science."*, which is truth-framed. To ensure consistency in analysis, we reverse-coded stance polarity and Likert scores for this topic in Section 6 by multiplying them by -1, so that positive values always indicate endorsement of the false statement.

B Breadth Topic Construction

The Breadth topic set ($\mathcal{T}_{\text{Breadth}}$) consists of 100 topics curated from two large-scale cross-national surveys: the World Values Survey (WVS) [12] and the Pew Global Attitudes Survey (PGAS) [24]. Because our study only recruited participants based in the United States, we filtered and selected survey questions that were assigned to U.S. respondents. To ensure the topics naturally elicit divergent human views, we selected questions that have the highest entropy in response distributions among U.S. participants, as measured in prior work [10].

Most original questions are already framed as evaluative statements rated on a Likert scale. For example:

- **Original questions:**
  Please tell me for each of the following statements whether you think it can always be justified, never be justified, or something in between.
  Euthanasia can always be justified. (Presented along with a 10-point Likert scale.)

In these cases, we use the original statement directly as a debate topic (e.g., *"Euthanasia can be justified."*).

Some other questions, however, are framed in a multiple-choice format. To convert these into clearly debatable statements, we reframe the most frequently chosen responses as separate topic statements. For example:

- **Original questions:**
  In your opinion, what is the most important problem facing this country today?
  (Options: **Economic problems (19.59%)**, Children and education (4.12%), Crime (3.09%), Health (4.12%), Housing (1.03%), People (11.34%), Politics (14.43%), **International affairs (36.08%)**, Science (1.03%), Others (5.15%))
- **Reframed as two separate debatable topics:**
  - *International affairs is the most important problem facing the U.S. today.*
  - *Economic problems are the most important problem facing the U.S. today.*

We also revised certain phrasings to reflect the present-day political context. For instance:

- **Original questions:**
  How confident are you that Joe Biden can make good decisions about the use of military force?

- **Revised topic statement:**
  Donald J. Trump can make good decisions about the use of military force.

These modifications ensure that all topics are relevant, interpretable, and debate-worthy, while remaining faithful to the spirit of the original survey questions. Each topic statement was manually reviewed to confirm that it is clearly phrased as a 1) self-contained declarative sentence, 2) framed in a way that invites disagreement, and 3) suitable for eliciting meaningful opinion exchanges in multi-party conversations.

The full list of all 100 Breadth topics will be included in the released dataset upon paper acceptance. Table 5 provides 43 representative examples, along with tentatively assigned category labels. These categories are introduced solely to help readers understand the topic diversity and are not derived from the original WVS or PGAS surveys. They are not used in any part of our simulation, evaluation, modeling, or analysis.

Table 5: Categorization of 43 representative Breadth topics used in our study.

Topic Category	Topic Statement
Governance & Democracy	A democratic system where citizens, not elected officials, vote directly on major national issues to decide what becomes law is a good way of governing the US. It is a characteristic of democracy for the state to make people's incomes equal. Living in a country that is governed democratically is important. The United States is being governed democratically today. The army taking over when the government is incompetent is a characteristic of democracy.
Science & Technology	Science and technology are making our lives healthier, easier, and more comfortable. The world is better off because of science and technology. It is important for people to know about science in their daily life. We depend too much on science and not enough on faith. Because of science and technology, there will be more opportunities for the next generation.
Morality & Social Norms	Sex before marriage can be justified. Suicide can be justified. Homosexuality can be justified. Abortion can be justified. Having casual sex can be justified. Violence against other people can be justified in some cases.
Economic Inequality & Social Mobility	Incomes should be made more equal. The growing gap between the rich and poor poses the greatest threat to the world. The fact that some people work harder than others is the most important reason for the gap between the rich and the poor in the United States. Knowing the right people is important for getting ahead in life. Belonging to a wealthy family is important for getting ahead in life.
Media & Trust in Institutions	Journalists provide fair coverage of elections in the US. TV news favors the governing party in general. News organizations are doing well at reporting different positions on political issues fairly. There is abundant corruption in the United States. Most politicians in the United States are corrupt.
International Relations & Trade	Donald J. Trump can deal effectively with China. The North American Free Trade Agreement (NAFTA) has been good for the US. The United States benefits a lot from the World Health Organization. Overall, increased tariffs on imported goods from foreign countries are good for the US. International affairs is the most important problem facing the US today.
Public Policy & Government Role	The government should take more responsibility to ensure that everyone is provided for, rather than leaving it to individuals. Public debt is the most important issue for the government to address first. The lack of employment opportunities is the most important issue for the government to address first. Government ownership of business should be increased.
Religion & Belief	We depend too much on science and not enough on faith. Religious and ethnic hatred poses the greatest threat to the world. It is an essential characteristic of democracy for religious authorities to interpret the laws.
US Identity & Society	Being born in the United States is important for truly being American. The United States has the best quality of universities. The United States is a place where a young person could lead a good life. I'm worried about a civil war in the United States.

C Demographic Summary

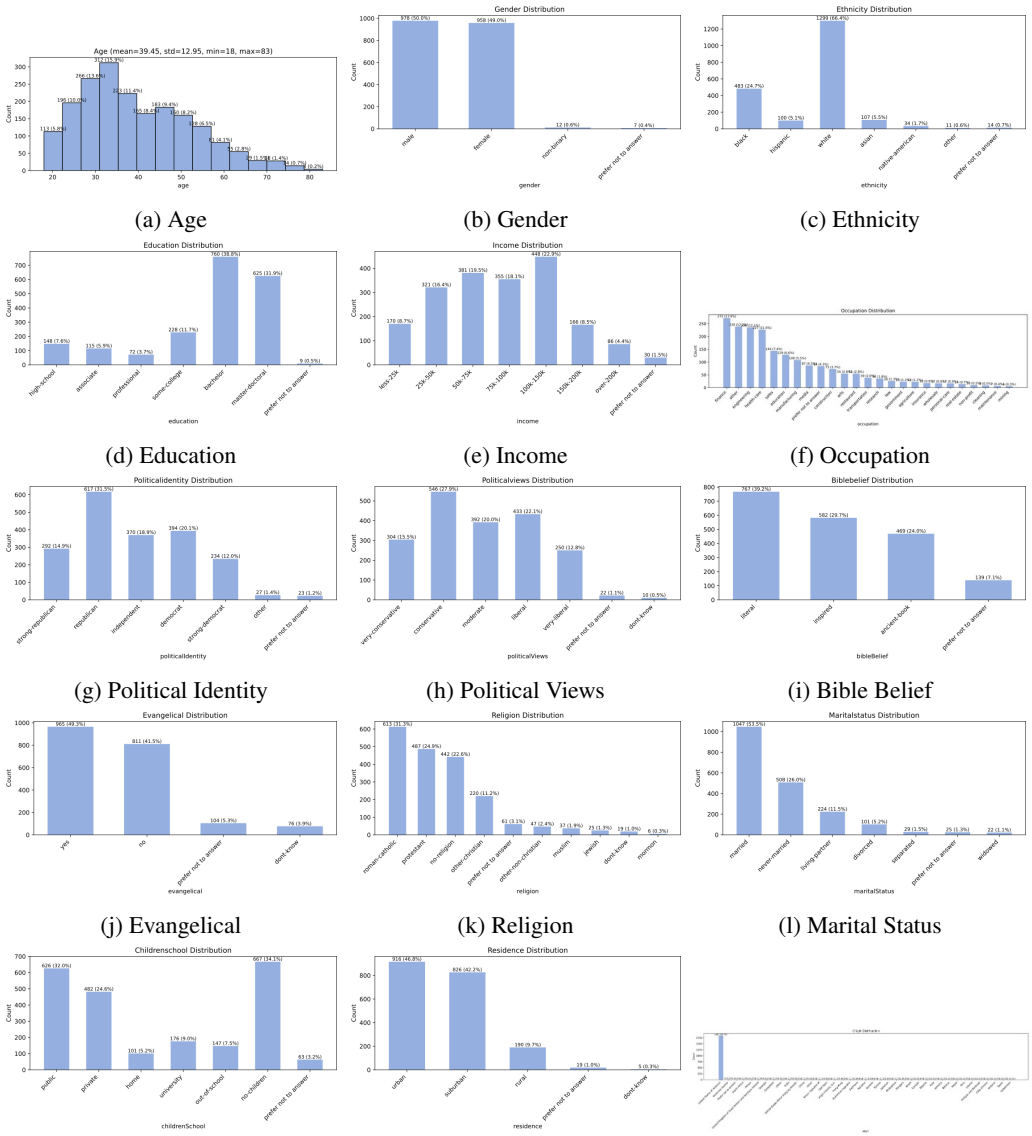
Of the 2,012 total participants in our study, 1,955 (97.2%) completed the demographic questionnaire; the remainder exited the experiment early. The resulting sample reflects substantial demographic diversity across multiple dimensions (Figure 4). Participants range in age from 18 to 83 (M = 39.5, SD = 13.0) and span a broad spectrum of gender identities, education levels, ethnic backgrounds, and income brackets. The cohort includes individuals with high school to doctoral-level education, and income levels range from under $25k to over $200k. Racial and ethnic diversity is well represented, with participants identifying as Black, Hispanic, White, Asian, Native American, and multiracial. Political identities and views are distributed across the ideological spectrum, and respondents report a wide variety of religious affiliations and Bible interpretations. Participants also vary in marital and parental status, geographic residence (urban, suburban, rural), and religious orientation (with nearly half identifying as evangelical and others expressing secular or alternative beliefs). Occupation is similarly diverse, with respondents employed across sectors including finance, engineering, health care, education, manufacturing, media, construction, among many. This heterogeneity ensures a rich and representative foundation for studying opinion dynamics and belief-based interactions.

D Prompt Templates for LLM Role-play Simulation

We detail the prompt templates used to construct the memory module $\mathcal{M}_{a_i,k}$ for each role-playing LLM agent a_i in our multi-agent opinion exchange setup. Each agent simulates a human participant and is prompted with information that mirrors the participant's first-person memory before producing the k-th utterance in a given round.

Each simulation begins with a system prompt that defines the agent's persona and task framing, followed by a sequence of user prompts corresponding to different memory components. All simulations adhere to the closed-world assumption (see Section 4.1) and are structured to match the human task instructions (see Section 3.1).

Table 6 illustrates an example prompt used in Simulation Mode 1: Next Message Prediction (Section 4.2). This example reflects the memory state of agent a_i at the beginning of Round 3, where all prior tweets and utterances are written by humans and added to the prompt as input. Each user prompt corresponds to one component of the memory module $\mathcal{M}_{a_i,k}$: demographic profile d_{s_i}, task instruction, initial opinion $(o^{\text{init}}_{s_i}, j^{\text{init}}_{s_i})$, previous rounds' tweets and dyadic conversations $\{\tau^{r'}_s, \mathcal{C}^{r'}_s : 1 \leq r' < 3\}$, and current round context including partner tweets and prior utterances $(\tau^3_{s_i}, \tau^3_{s_j}, \{u^3_{k',s} : k' < k\})$. Curly brackets ({}) denote placeholder variables specific to each agent and topic instance. For readability, color highlights in the table correspond to different memory components.



(a) Age (b) Gender (c) Ethnicity

(d) Education (e) Income (f) Occupation

(g) Political Identity (h) Political Views (i) Bible Belief

(j) Evangelical (k) Religion (l) Marital Status

(m) Children's Schooling (n) Residence Type (o) Country of Origin

Figure 4: Demographic distributions across age, gender, education, ethnicity, income, political identity and views, religion, family, and geographic background.

Table 6: Prompt templates used to construct the memory module $\mathcal{M}_{a_i,k}$ for each LLM agent a_i during role-play (Section 4.1). This example reflects the memory state of agent a_i at the beginning of Round 3 under Mode 1: Next Message Prediction (Section 4.2), where prior tweets and utterances written by humans were added to the memory. Each prompt governs one component of memory: demographic profile d_{s_i}, task instruction, initial opinion $(o_{s_i}^{\text{init}}, j_{s_i}^{\text{init}})$, previous rounds $\{\tau_s^{r'}, \mathcal{C}_s^{r'} : 1 \leq r' < 3\}$, and current round context $(\tau_{s_i}^3, \tau_{s_j}^3, \{u_{k',s}^3 : k' < k\})$. Curly brackets ({}) denote placeholder variables that are different for each agent and topic. Color highlights correspond to different memory components.

Prompt Type	**Message Type**	**Prompt Template**	**Example Values for Placeholders**
Agent Initialization: Demographic Profile (d_i), Task Instruction, Initial Opinion ($o_i^{\text{init}}, j_i^{\text{init}}$)	*System Message*	Role play this person: You are a {age}-year-old {gender} with {education} education. Your ethnicity is {ethnicity}, and your annual income falls in the {income bracket} range. Politically, you identify as {party ID} with {ideology} views. You have children in {children_school_status}, reside in a {urbanicity} area, and your marital status is {marital status}. Regarding religious beliefs, you consider the Bible to be {bible view}, {yes/no} identify as evangelical, and your religious affiliation is {religious affiliation}. Your occupation is {occupation}. You have been interacting with other strangers on Twitter. You can decide to change or maintain your belief about the topic {topic}. You would first write a tweet about the topic {topic} that reflected your opinion. You would then engage in a private conversation through a textbox with a different stranger. In the conversation, you would first see the tweet the stranger wrote along with your own tweet. After seeing both tweets, you would be asked to read and respond to the stranger about the topic {topic}. Throughout the interactions, you are alone in your room with limited access to the Internet. You cannot search for information about the topic {topic}, nor go out to ask other people. To form your belief, you can only rely on your initial belief and the information shared by others on Twitter. Before interacting with other people, below is your initial opinion on {topic} using a 6-point Likert scale: - Certainly disagree - Probably disagree - Lean disagree - Lean agree - Probably agree - Certainly agree On the Likert scale, you chose {Likert-scale opinion} as your initial opinion regarding the statement {topic}. Below is your explanation for your initial opinion: {free-text justification} This opinion represents your starting point. It's based on your current understanding, personal experiences, and the beliefs that have shaped your perspective. As you engage in discussions, your views may evolve, but this is where you begin.	**Demographic Profile:** age = 41 gender = female education = master ethnicity = white income bracket = 50k-75k party ID = republican ideology = conservative children_school_status = ['private', 'university'] urbanicity = rural marital status = married bible view = literal evangelical = yes religious affiliation = protestant occupation = finance **Task Instruction:** topic = "You are satisfied with how the political system is functioning in the US these days." **Initial Opinion:** topic = "You are satisfied with how the political system is functioning in the US these days." Likert-scale response = "Probably agree" Explanation = "I am indeed satisfied with the political system because the government is trying hard enough to introduce cryptocurrency to the market, which is the future currency of the world."

Conversation History: Previous Rounds (Round 1 & Round 2), Current Round Context (Round 3)	*User Message*	Below was your conversation with {first_partner_name} My tweet: $\{\tau^1_{s_1}\}$ {first_partner_name}'s tweet: $\{\tau^1_{s_2}\}$ My response: $\{u^1_{1,s_1}\}$ {first_partner_name}'s response: $\{u^2_{2,s_2}\}$ My response: $\{u^1_{3,s_1}\}$ {first_partner_name}'s response: $\{u^2_{4,s_2}\}$ ... ... You have just finished your conversation with {first_partner_name}. Instead, you are now engaging in conversation with another stranger {second_partner_name} on a separate text box. Below was your conversation with {second_partner_name}. My tweet: $\{\tau^2_{s_1}\}$ {second_partner_name}'s tweet: $\{\tau^2_{s_3}\}$ {second_partner_name}'s response: $\{u^2_{1,s_3}\}$ My response: $\{u^2_{2,s_1}\}$ {second_partner_name}'s response: $\{u^2_{3,s_3}\}$ My response: $\{u^2_{4,s_1}\}$ ... ... You have just finished your conversation with {second_partner_name}. Instead, you are now engaging in conversation with another stranger {third_partner_name} on a separate text box. Below was your conversation with {third_partner_name}. My tweet: $\{\tau^3_{s_1}\}$ {third_partner_name}'s tweet: $\{\tau^3_{s_4}\}$ My response: $\{u^3_{1,s_1}\}$ {third_partner_name}'s response: $\{u^3_{2,s_3}\}$ My response: $\{u^3_{3,s_1}\}$ {third_partner_name}'s response: $\{u^3_{4,s_3}\}$	**Previous Rounds (Round 1):** $\tau^1_{s_1}$ (Your tweet) = I am satisfied with political system because the government is trying hard enough to stabilize the economy through various ways like transitioning to crypto currency $\tau^1_{s_2}$ (681e3's tweet) = I disagree with the statement that I am satisfied with the way the American system functions these days. This is because of the system's extreme polarization making it fail to take meaningful action u^1_{1,s_1} (Your response) = From my point of view, the government is not that perfect but at least it's trying to improve the lives of all Americans u^2_{2,s_2} (681e3's response) = I still believe that the political system is flawed but I completely see your viewpoint. Politicians appear to care more about maintaining party allegiance than they do about the problems that people care about. Can the system be re-organized in your opinion or is the division too great? The fact that everything has become more divisive which makes compromise nearly impossible in my opinion is largely to blame. What do you think? **Previous Rounds (Round 2):** $\tau^2_{s_1}$ (Your tweet) = We should support the government motives to improve and make our country great. On my side, the government is doing the best it can to stabilize our economy and improve our lives $\tau^2_{s_3}$ (683b8's tweet) = I agree and I am totally satisfied with how the political system is working. This is because it is promoting good health and education facilities to its citizens. u^2_{1,s_3} (683b8's response) = It provide strict laws. It gives freedom to all citizens to publicly participate in elections. u^2_{2,s_1} (Your response) = I second your point, the government has helped the education sector through scholarships. It has also invested a lot of resources in the healthcare field. Yes, it also gives each citizen the right to express one's ideas and opinions. u^2_{3,s_3} (683b8's response) = It has also improved infrastructure and advancement of technology. **Current Round Context (Round 3):** $\tau^3_{s_1}$ (Your tweet) = The government plays a crucial role in advancement of technology by budgeting enough resources. It also helps in infrastructure and healthcare, I support $\tau^3_{s_4}$ (68405's tweet) = The government allows its people participation on the development project and is highly working on development u^3_{1,s_4} (68405's response) = I do agree on advancing the technology and improving also in defense force and provide high security u^3_{2,s_1} (Your response) = Yes, the government contributes to the general development of the country by investing enough money onto different projects u^3_{3,s_4} (68405's response) = That's okay. It's also improving on more projects and inventions u^3_{4,s_1} (Your response) = It also contributes to a stable economy

E LLM-Based Annotation for On-topicness and Stance

To evaluate LLM agent alignment with human behavior, we use `gpt-4o-mini-2024-07-18` as a classifier: one for identifying on-topic utterances and another for mapping stance to a scalar value. Both classifiers are implemented using prompting.

On-topic Classification. For each simulated utterance $\widehat{u}$ and associated discussion topic t, we classify whether $\widehat{u}$ is on-topic. An utterance is considered on-topic if it directly addresses the content or implications of the assigned discussion topic t, rather than containing social talk or unrelated comments. The classifier uses a system prompt that defines "on-topicness" and asks the model to

Table 7: Prompt template used for on-topic classification with `gpt-4o-mini-2024-07-18`. Example utterances are described in Table 8.

Prompt Template
System Message Your task is to analyze the provided conversation. The conversation can either be between two humans or two role-playing LLMs. They are assigned a topic of interest, and are asked to discuss only that topic. You have to determine if the latest response in the conversation is "valid and relevant" to the topic of interest "{TOPIC}". To show what "valid and relevant" means, below are some "valid" example cases where either two role-playing LLMs or two humans are discussing another topic of interest: "{OTHER_TOPIC}". Valid example where a role-playing LLM generates a "valid and relevant" response: {VALID_EXAMPLE_LLM} Valid example where a human generates a "valid and relevant" response: {VALID_EXAMPLE_HUMAN} Another valid example where a human generates a "valid and relevant" response in context of the conversation: {VALID_EXAMPLE_CONTEXTUAL} Sometimes whether the response is relevant may be ambiguous, but the relevancy can be inferred from the conversation history. Here is a valid example where the response itself may be ambiguous, but is indeed relevant to the topic: {VALID_EXAMPLE_AMBIGUOUS} Sometimes a response may be too uninformative on its own to determine relevance, but its relevance can be inferred from the conversation history. Here is a valid example where the response itself may seem uninformative, yet it is indeed relevant to the topic because a person's perspective is likely to remain consistent with what they have previously expressed—especially when using affirming words like "yeah.": {VALID_EXAMPLE_YEAH} In some cases, the human or the role-playing LLM may generate some messages that are "invalid", "ill-formatted" or "irrelevant" to the topic. For example, the LLM may repeat the instruction, generate irrelevant response, output json object, or generate ill-formatted responses (responses that are not from the perspective of role-playing), among many. Similarly, a human can also utter irrelevant or invalid responses. For example, the humans may digress from the topic of interest in their conversation. Below are some concrete "invalid" examples of "invalid" or "irrelevant" response: Invalid example where a role-playing LLM repeats the instruction: {INVALID_EXAMPLE_INSTRUCTION} Invalid example where a role-playing LLM generates a json object: {INVALID_EXAMPLE_JSON} Invalid example where a role-playing LLM generates a response that is irrelevant to the topic of interest. Recall that in this conversation, the topic of interest is "{OTHER_TOPIC}". Below is the example: {INVALID_EXAMPLE_IRRELEVANT} Invalid example where a human generates a response that is irrelevant to the topic. Recall that in this conversation, the topic of interest is "{OTHER_TOPIC}". Below is the example: {INVALID_EXAMPLE_HUMAN} Invalid example where a role-playing LLM generates a response that is ill-formatted. The initial part is redundant. The role-playing LLM should directly generate a response to the other role-playing LLM, instead of a response to the user. Below is the example: {INVALID_EXAMPLE_REDUNDANT} Invalid example where a role-playing LLM generates a response that is ill-formatted. The role-playing LLM should not generate subsequent responses from the other role-playing LLM. Below is the example: {INVALID_EXAMPLE_MULTI_TURN}
User Message Below is the conversation history up to the latest message. {CONVERSATION_HISTORY} The latest message is: {LATEST_MESSAGE} Based on the provided conversation history, determine if the latest message is "valid" in the context of the conversation. Answer with "VALID" or "INVALID" only.

return a binary label. To ensure reliability, we manually labeled 200 utterances with binary on-topic judgments. We iteratively refined the prompt and verified that the LLM's outputs matched human judgment on this validation set with high consistency. The final classifier outputs a binary indicator $I_{\text{topic}}(\widehat{u}, t) \in \{0, 1\}$. The full prompt template is shown in Table 7, and examples of on-topic and off-topic utterances are listed in Table 8.

Stance Classification. To evaluate opinion alignment, we map each utterance u or $\widehat{u}$ to a scalar stance score $S(u) \in \{-2.5, -1.5, -0.5, +0.5, +1.5, +2.5\}$ also using `gpt-4o-mini-2024-07-18`. These scores correspond to the following six-point Likert categories: *(-2.5) Certainly disagree, (-1.5) Probably disagree, (-0.5) Lean disagree, ($+0.5$) Lean agree, ($+1.5$) Probably agree, ($+2.5$) Certainly agree*.

We manually annotated 200 utterances with stance labels and tuned the prompt to produce outputs that aligned with human judgment. The final prompt includes the topic, utterance, and the instruction to return one of the six stance categories, which are then mapped to scalar values.

For both classifiers, they were validated against the 200 human-labeled utterances per topic, and achieved 90% accuracy. Full prompts and examples are available in our code release.

Table 8: Examples utterances used in the on-topic classification prompt. Each example corresponds to a placeholder in the template from Table 7. All examples are about the topic "regular fasting will improve your health."

Prompt Examples
{VALID_EXAMPLE_LLM} "Absolutely! I think having a structured eating schedule can definitely help with planning and time management. It can create a sense of routine that makes it easier to make healthier choices. I like the idea of limiting eating to specific times rather than cutting out food entirely; it feels more sustainable and balanced!"
{VALID_EXAMPLE_HUMAN} "i dont know much research myself but i have tried it before and it felt pretty good. i would imagine there are both positives and negatives though."
{VALID_EXAMPLE_CONTEXTUAL} Previous message from another human: "My only caveat with regular fasting is that it doesn't work for everyone. While it suits me well, it may not provide the same benefits for someone else" Valid current message (from the person of interest): "yes i agree"
{VALID_EXAMPLE_AMBIGUOUS} Previous message from the person of interest: "I agree! It won't solve the root of the issues Many people don't eat healthy on a regular basis so a quick 1-3 day cleanse isn't going to fix years of unhealthy lifestyle choices It may be a good starting place tho, as long as it's done correctly with the help of a doctor" Previous message from another human: "Cooking your own meals instead of fast food, exercising, getting enough sleep, easing stress - those will help more in the long term" Seemingly ambiguous but actually valid current message (from the person of interest): "I agree 100% absolutely! And if you can take it a step further and grow your own food, even better! That's it...I'm starting a garden! Forget the cleanse!"
{VALID_EXAMPLE_YEAH} Previous message of the person of interest: "We need to get to the point to teach people how to know such a thing, it is very pointless, and we need doctors on board to educate us on it because it is very dangerous when you don't have any knowledge about it" Previous message from another human: "Some of my friends swear this is the best thing they've ever done, but I am still skeptical as to what benefits I will receive besides my body just feeling better. . . " Previous message of the person of interest: "What do you have in mind?" Previous message from another human: "That is the most sense I've heard about this topic yet. Well I mean, if it's meant for something that I don't need or suffer from, why would I want to use it, ya know? I just want to know more, and a doctor would be a great start" Previous message of the person of interest: "What makes put in that point" Previous message from another human: "I have heard of different types of cleanses and I would assume they all do different things, but also I haven't heard much about possible side effects and wonder how many studies have been done on these Have you?" Seemingly uninformative but actually valid current message (from the person of interest): "yeah"
{INVALID_EXAMPLE_INSTRUCTION} "Now reply again as if you are still participating in this conversation, making use of the input generated by the system previously. Note that there was one iteration before this. 89489's previous message: '"'" Yes, America is one of those few lucky nations with higher tax than others. I believe we should all pitch into help fight terrorism and crime, no matter how low-classed someone gets. At least they're still trying to contribute and pay their taxes. And those who aren't willing to"

{INVALID_EXAMPLE_JSON}
'"'"'json
{
"persona": {
"id": "87774",
"age": 18,
"education_level": "some_college_education",
"ethnicity": "white",
"income_range": "$75-$125K annually",
"political_affiliation": "moderate_democrat",
"residence_type": "urban_area",
"marital_status": "never_married",
"religious_beliefs": "ancient_book_bible_no_evangelical_no_religion",
}
}

{INVALID_EXAMPLE_IRRELEVANT} "Oh nice! Soccer is really popular. Do you have a favorite team or player?"
{INVALID_EXAMPLE_HUMAN} "u watch football?"
{INVALID_EXAMPLE_REDUNDANT} "Sure thing. Here's what my response could look like after considering all points brought forth by each participant: My Response: Absolutely, accountability plays a crucial role here — knowing when you last ate encourages mindful eating which can't hurt. And yeah, combining fasting with regular exercise indeed enhances its potential benefits. Like you pointed out, holding oneself responsible aids greatly towards maintaining a balanced diet. Yet, whether'regular' fasting improves overall health without any negative repercussions isn't something one-size-fits-all; everybody"
{INVALID_EXAMPLE_MULTI_TURN} "I both agree and disagree with this statement. I believe that regular fasting can be healthy depending on its implementation. 89218's response: One thing we need to make clear is whether we're talking about traditional forms of fasting such as the Muslim tradition during Ramadan or the Jewish Yom Kippur practice? Or do you mean something more modern/contemporary? My Response : With regards to my original comment about the relationship between religion and culture versus public policy, I guess I am commenting mainly under the"

F Round-wise Aggregated Evaluation Metrics

We define the following sets used throughout evaluation: $\widehat{\mathcal{U}}$ and $\mathcal{U}$ denote all utterances generated by LLM agents and humans, respectively. Their on-topic subsets with respect to discussion topic t are denoted $\widehat{\mathcal{U}}_{\text{topic}} \subseteq \widehat{\mathcal{U}}$ and $\mathcal{U}_{\text{topic}} \subseteq \mathcal{U}$. For each agent–participant pair (a_i, s_i) and round r, we denote $\widehat{\mathcal{U}}^r_{\text{topic},a_i}$ and $\mathcal{U}^r_{\text{topic},s_i}$ as their respective on-topic utterances in round r.

Round-wise Aggregation. For each simulated on-topic utterance $\widehat{u} \in \widehat{\mathcal{U}}^r_{\text{topic},a_i}$, we compare it against all human on-topic utterances $u \in \mathcal{U}^r_{\text{topic},s_i}$ produced by the corresponding human participant s_i in the same round. This yields the round-wise average metric score:

$$\overline{M}^{\text{round}} = \frac{1}{|\widehat{\mathcal{U}}_{\text{topic}}|} \sum_{i=1}^{N} \sum_{r=1}^{R} \sum_{\widehat{u} \in \widehat{\mathcal{U}}^r_{\text{topic},a_i}} \left(\frac{1}{|\mathcal{U}^r_{\text{topic},s_i}|} \sum_{u \in \mathcal{U}^r_{\text{topic},s_i}} M(\widehat{u}, u) \right), \tag{2}$$

where $M \in \{S_{\text{sem}}, \Delta_{\text{stance}}, \Delta_{\text{abs_len}}, \Delta_{\text{signed_len}}, \text{ROUGE-L}\}$ and $\widehat{\mathcal{U}}_{\text{topic}} = \bigcup_{i=1}^{N} \bigcup_{r=1}^{R} \widehat{\mathcal{U}}^r_{\text{topic},a_i}$.

On-topic Classification. We define an utterance $\widehat{u}$ as on-topic with respect to topic t if $I_{\text{topic}}(\widehat{u}, t) = 1$, where I_{topic} is predicted by `gpt-4o-mini-2024-07-18`. The classifier was validated against 200 human-labeled utterances per topic, achieving 90% accuracy. Utterances are deemed off-topic if they do not substantively address the assigned discussion topic. Common off-topic examples include greetings (e.g., *"hello"*), meta-remarks (*"what do you think?"*), or unrelated diversions (*"do you watch football?"*). For details of classification, see E.

Stance Classification. To assess opinion alignment, each utterance u is mapped to a scalar stance score $S(u)$ via a GPT-4o-mini classifier. The model predicts one of six bins corresponding to a 6-point Likert scale, rescaled to real values $[-2.5, -1.5, -0.5, +0.5, +1.5, +2.5]$. The classifier was validated on a sample of 200 manually annotated utterances per topic, achieving 90% accuracy. For details of classification, see E.

Semantic Embedding. The sentence encoder $E(\cdot)$ used in S_{sem} is based on `jinaai/jina-embeddings-v3` [27], which produces 1024-dimensional embeddings. Semantic similarity is computed as cosine similarity between embedded vectors: $S_{\text{sem}}(u, \widehat{u}) = \cos(E(u), E(\widehat{u}))$

G Compute Resources

We ran all experiments (including simulations, fine-tuning, and evaluation) on a GPU machine equipped with 1x NVIDIA H100 PCIe (80GB).

H Statistical Tests for Utterance-Level Alignment Metrics

To assess whether the best-performing model (`gpt-4o-mini-2024-07-18`) consistently outperforms others, we conduct statistical tests across six experimental conditions (2 datasets × 3 simulation modes) for three metrics: semantic similarity $\overline{S}_{\text{sem}}$ (higher is better), ROUGE-L $\overline{\text{ROUGE-L}}$ (higher is better), and stance difference $\overline{\Delta}_{\text{stance}}$ (lower is better). For each metric, we apply a repeated-measures Friedman test to detect overall model differences, followed by one-sided, paired Wilcoxon signed-rank tests to test whether `gpt-4o-mini` outperforms each baseline. The Wilcoxon tests are conducted to test whether the best-performing model reliably outperforms the rest.

H.1 Semantic Similarity ($\overline{S}_{\text{sem}}$)

The Friedman test reveals a significant overall difference across the six models ($\chi^2 = 17.87$, $df = 5$, $p = .003$). Wilcoxon tests show that `gpt-4o-mini-2024-07-18` significantly outperforms `Llama-3.1-8B-Instruct` ($p = .018$), `Llama-3.1-70B-Instruct` ($p = .017$),

`Mistral-7B-Instruct-v0.3` ($p = .024$), and `Qwen2.5-32B-Instruct` ($p = .018$). The difference with `Llama-3.1-Tulu-3-8B-SFT` is not statistically significant ($p = .146$), but the trend still favors `gpt-4o-mini`.

H.2 ROUGE-L ($\overline{\text{ROUGE-L}}$)

The Friedman test also shows a significant difference in ROUGE-L scores ($\chi^2 = 26.35$, $df = 5$, $p < .001$). Wilcoxon tests confirm that `gpt-4o-mini-2024-07-18` significantly outperforms all baseline models: `Llama-3.1-Tulu-3-8B-SFT` ($p = .017$), `Llama-3.1-8B-Instruct` ($p = .016$), `Llama-3.1-70B-Instruct` ($p = .013$), `Mistral-7B-Instruct-v0.3` ($p = .016$), and `Qwen2.5-32B-Instruct` ($p = .018$).

H.3 Stance Difference ($\overline{\Delta}_{\text{stance}}$)

The Friedman test indicates a significant overall difference in stance alignment across models ($\chi^2 = 21.57$, $df = 5$, $p = .001$). Lower values indicate better alignment. Wilcoxon tests show that `gpt-4o-mini-2024-07-18` significantly outperforms `Llama-3.1-Tulu-3-8B-SFT` ($p = .018$) and `Llama-3.1-8B-Instruct` ($p = .018$). For `Llama-3.1-70B-Instruct` ($p = .300$), `Mistral-7B-Instruct-v0.3` ($p = .392$), and `Qwen2.5-32B-Instruct` ($p = .211$), the differences are not statistically significant but are still in the expected direction (underperforming compared to `gpt-4o-mini`).

Summary. Across all three metrics and six experimental settings, `gpt-4o-mini-2024-07-18` is the most consistently aligned with human responses.

All tests were conducted using R [25].

I Simulation Results on Breadth Topics

Table 9: Evaluation results across simulation modes and LLMs. We report the round-wise aggregated metrics on the **Breadth Topics**: average semantic similarity $\overline{S}_{\text{sem}}$ (↑), average stance difference $\overline{\Delta}_{\text{stance}}$ (↓), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (→0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (↓), average ROUGE-L $\overline{\text{ROUGE-L}}$ (↑), and on-topic utterance rate $R_{\text{on-topic}}$.

LLM & Simulation Mode	S_{sem} (↑)	Δ_{stance} (↓)	$\Delta_{\text{signed_len}}$ (→0)	$\Delta_{\text{abs_len}}$ (↓)	ROUGE-L (↑)	$R_{\text{on-topic}}$
Simulation Mode 1: Next Message Prediction						
gpt-4o-mini-2024-07-18	**0.49**	**1.01**	-33.26	34.26	**0.10**	0.94
Llama-3.1-Tulu-3-8B-SFT	0.42	1.30	-28.61	38.36	0.05	0.53
Llama-3.1-8B-Instruct	0.43	1.29	-30.40	33.50	0.06	0.80
Llama-3.1-70B-Instruct	0.43	1.16	**-15.37**	**20.44**	0.07	0.87
Mistral-7B-Instruct-v0.3	0.48	1.11	-44.87	45.54	0.07	0.95
Qwen2.5-32B-Instruct	0.45	1.05	-25.13	29.02	0.07	0.90
Simulation Mode 2: Tweet-guided Conversation Simulation						
gpt-4o-mini-2024-07-18	0.41	1.17	-61.30	61.58	**0.08**	0.81
Llama-3.1-Tulu-3-8B-SFT	**0.42**	1.25	-50.26	51.91	0.05	0.27
Llama-3.1-8B-Instruct	0.39	1.35	-47.90	48.94	0.05	0.75
Llama-3.1-70B-Instruct	0.38	1.22	**-38.46**	**40.39**	0.05	0.75
Mistral-7B-Instruct-v0.3	0.41	1.18	-48.47	48.99	0.06	0.76
Qwen2.5-32B-Instruct	0.40	**1.16**	-52.02	53.62	0.06	0.70
Simulation Mode 3: Full Conversation Simulation						
gpt-4o-mini-2024-07-18	0.40	**1.22**	-61.34	61.64	**0.08**	0.79
Llama-3.1-Tulu-3-8B-SFT	**0.41**	1.29	-48.87	50.99	0.05	0.21
Llama-3.1-8B-Instruct	0.38	1.39	-48.32	49.44	0.05	0.72
Llama-3.1-70B-Instruct	0.36	1.23	**-40.06**	**41.59**	0.05	0.73
Mistral-7B-Instruct-v0.3	0.39	1.22	-48.08	48.74	0.06	0.74
Qwen2.5-32B-Instruct	0.38	1.22	-52.01	53.59	0.06	0.72

Table 9 presents alignment results across simulation modes and LLMs on the Breadth topics.

Table 10: Ablation results across simulation modes using `gpt-4o-mini-2024-07-18` on the Breadth Topics. We report average semantic similarity $\overline{S}_{\text{sem}}$ (↑), average stance difference $\overline{\Delta}_{\text{stance}}$ (↓), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (→0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (↓), average ROUGE-L $\overline{\text{ROUGE-L}}$ (↑), and on-topic utterance rate $R_{\text{on-topic}}$. Blue cells indicate improved performance after ablation, while red cells indicate worsened performance after ablation.

Ablation Condition	S_{sem} (↑)	Δ_{stance} (↓)	$\Delta_{\text{signed_len}}$ (→ 0)	$\Delta_{\text{abs_len}}$ (↓)	ROUGE-L (↑)	$R_{\text{on-topic}}$
Simulation Mode 1: Next Message Prediction						
Original	0.49	1.01	-33.26	34.26	0.10	0.94
No Private Profile	0.48	1.01	-31.48	32.58	0.10	0.94
No Demographics	0.48	1.03	-30.27	31.49	0.10	0.95
No Initial opinion	0.48	1.02	-32.11	33.23	0.10	0.94
No Prior Chats	0.49	1.05	-40.68	41.32	0.10	0.94
Simulation Mode 2: Tweet-guided Conversation Simulation						
Original	0.41	1.17	-61.30	61.58	0.08	0.81
No Private Profile	0.41	1.16	-61.57	61.86	0.08	0.80
No Demographics	0.40	1.19	-61.40	61.71	0.08	0.81
No Initial opinion	0.41	1.18	-61.56	61.84	0.08	0.82
No Prior Chats	0.43	1.17	-59.62	59.96	0.08	0.84
Simulation Mode 3: Full Conversation Simulation						
Original	0.40	1.22	-61.34	61.64	0.08	0.79
No Private Profile	0.39	1.19	-61.44	61.73	0.08	0.80
No Demographics	0.39	1.21	-61.31	61.61	0.08	0.81
No Initial opinion	0.40	1.24	-61.29	61.62	0.08	0.81
No Prior Chats	0.42	1.21	-59.30	59.64	0.08	0.83

J Ablation Results on Breadth Topics

To complement the main results on Depth topics (Table 4), Table 10 presents ablation results on Breadth topics using `gpt-4o-mini`. We observe similar trends across simulation modes: in Mode 1, removing prior chat or private profile information has little effect on semantic alignment, whereas in Modes 2 and 3, ablating private profile leads to decreased semantic and stance alignment.

K Supervised Fine-Tuning (SFT): Methods, Settings, and Results

Objective and Setup. We use supervised fine-tuning (SFT) to align role-playing LLM agents with human opinion trajectories. Given a training set $\mathcal{D}_{\text{train}} = \{(x, y)\}$ of context–response pairs, where $x = \mathcal{M}_{a_i,k}$ is the agent's memory state and $y \in \{\tau^r_{s_i}, u^r_{k,s_i}, o^{\text{final}}_{s_i}, j^{\text{final}}_{s_i}\}$ is the human tweet, utterance, final opinion, or justification, we optimize the following log-likelihood objective:

$$\mathcal{L}_{\text{SFT}} = -\sum_{(x,y)\in\mathcal{D}_{\text{train}}} \log P_\theta(y \mid x).$$

This setup mirrors Simulation Mode 1 (Next Message Prediction), where the model is conditioned on actual human conversation history. As a proof of concept, we conduct SFT experiments only on the Depth topics.

Train/Test Partitioning. To evaluate generalization, we define a held-out test set $\mathcal{D}_{\text{test}}$ and explore three data partitioning strategies, summarized in Figure 5 and Table 11:

- **Round Generalization:** For each group g and topic t, we train on rounds 1–2 and test on round 3:
$$\mathcal{D}_{\text{train}} = \bigcup_{g,t} \{(x, y)^r \mid r \in \{1, 2\}\}, \quad \mathcal{D}_{\text{test}} = \bigcup_{g,t} \{(x, y)^{r=3}\}.$$
Participants and topics are shared between training and testing.
- **Group Generalization:** For each topic $t \in \mathcal{T}$, we partition participant groups into disjoint sets $\mathcal{G}^t_{\text{train}}$ and $\mathcal{G}^t_{\text{test}}$:
$$\mathcal{D}_{\text{train}} = \bigcup_{t} \bigcup_{g\in\mathcal{G}^t_{\text{train}}} \{(x, y)_{g,t}\}, \quad \mathcal{D}_{\text{test}} = \bigcup_{t} \bigcup_{g\in\mathcal{G}^t_{\text{test}}} \{(x, y)_{g,t}\}.$$

Topics remain fixed while groups vary.

- **Topic Generalization:** We partition the topic set into disjoint subsets $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{test}}$:

$$\mathcal{D}_{\text{train}} = \bigcup_{t \in \mathcal{T}_{\text{train}}} \{(x, y)_t\}, \quad \mathcal{D}_{\text{test}} = \bigcup_{t \in \mathcal{T}_{\text{test}}} \{(x, y)_t\}.$$

This requires generalization across unseen topics and new participant groups.

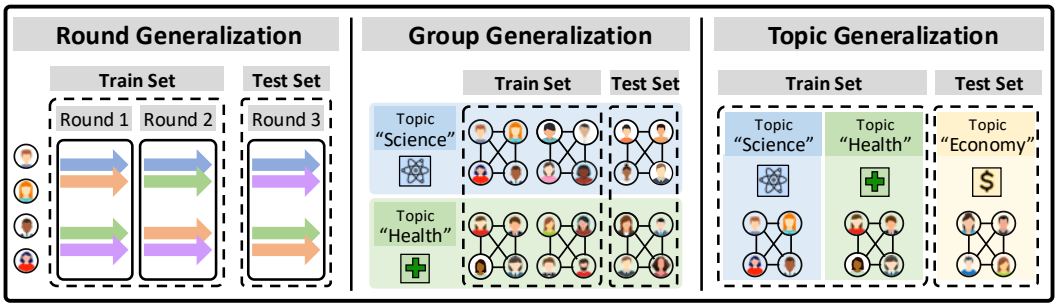


Figure 5: Illustration of the three generalization settings used for evaluating supervised fine-tuning (SFT): **Round Generalization** (left): Train on rounds 1–2 and test on round 3 within the same group and topic; **Group Generalization** (middle): Train and test on disjoint participant groups within the same topic; **Topic Generalization** (right): Train and test on disjoint sets of topics and participants. Each setting evaluates a different dimension of generalization for role-playing LLM agents.

Table 11: SFT dataset statistics for each generalization setting.

Data Type	Partition	(x, y) Pairs	On-topic (x, y) Pairs	Subjects
Round Generalization	Train	2256	1833	452
	Test	1645	1386	452
Group Generalization	Train	2588	2006	376
	Test	623	518	76
Topic Generalization	Train	2258	1759	340
	Test	983	786	112

In the Round and Group Generalization settings, the topic distribution is held constant across partitions. For Topic Generalization, we partition the Depth dataset by topic. Specifically, the held-out test topics are: *Regular fasting will improve your health* and *The U.S. deficit increased after President Obama was elected*, while the remaining five topics are used for training. The full Depth topic list is in Appendix A.

Fine-Tuning Details. **LLaMA-3.1-8B-Instruct.** We fine-tune `Llama-3.1-8B-Instruct` for 5 epochs using LoRA with 4-bit quantization (nf4) and the following configuration: LoRA rank $r = 64$, $\alpha = 128$, dropout $= 0.05$, Flash Attention 2, gradient checkpointing, cosine learning rate scheduler, and learning rate $= 10^{-4}$. We use a per-device train batch size of 8 with gradient accumulation steps of 32. Loss is computed only on the assistant's completion tokens. We enable model compilation with PyTorch using the Inductor backend. All models are fine-tuned using the `trl` library and `SFTTrainer`.

GPT-4o-mini. We fine-tune `gpt-4o-mini-2024-07-18` for 3 epochs using OpenAI's fine-tuning API[4] (`"type": "supervised"`) with automatic selection of batch size and learning rate multiplier. Loss is also computed in a completion-only setting.

Results and Limitations. We fine-tune both models on the Depth topics and report results in Tables 12 and 13 across the three generalization settings. SFT consistently improves surface-level

[4] `https://platform.openai.com/docs/api-reference/fine_tuning/`

alignment: the signed length difference $\overline{\Delta}_{\text{signed_len}}$ moves toward zero, absolute length difference $\overline{\Delta}_{\text{abs_len}}$ decreases, and ROUGE-L $\overline{\text{ROUGE-L}}$ improves across all settings.

However, deeper semantic and opinion-level metrics deteriorate. SFT reduces average semantic similarity $\overline{S}_{\text{sem}}$ and increases average stance difference $\overline{\Delta}_{\text{stance}}$, even on training data. This suggests SFT encourages surface-form mimicry without behavioral alignment, and may in fact harm deeper opinion-consistent modeling.

Table 12: Evaluation results for `gpt-4o-mini-2024-07-18` across simulation modes, SFT types, and data partitions. We report the average semantic similarity $\overline{S}_{\text{sem}}$ (↑), average stance difference $\overline{\Delta}_{\text{stance}}$ (↓), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (→0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (↓), average ROUGE-L $\overline{\text{ROUGE-L}}$ (↑), and on-topic utterance rate $R_{\text{on-topic}}$. Blue cells indicate improved performance after SFT, while red cells indicate worsened performance. See Table 13 for SFT results with `Llama-3.1-8B-Instruct`.

Generalization Type	Partition	Model	$\overline{S}_{\text{sem}}$ (↑)	$\overline{\Delta}_{\text{stance}}$ (↓)	$\overline{\Delta}_{\text{signed_len}}$ (→0)	$\overline{\Delta}_{\text{abs_len}}$ (↓)	$\overline{\text{ROUGE-L}}$ (↑)	$R_{\text{on-topic}}$
Round Generalization	*Simulation Mode 1: Next Message Prediction*							
	Train	pre-SFT	0.50	1.14	-36.17	36.71	0.11	0.93
		post-SFT	0.45	1.20	3.72	12.59	0.14	0.79
	Test	pre-SFT	0.47	1.06	-29.54	30.84	0.11	0.90
		post-SFT	0.44	1.21	4.65	12.04	0.14	0.79
	Simulation Mode 2: Tweet-guided Conversation Simulation							
	Train	pre-SFT	0.44	1.30	-57.66	57.90	0.09	0.75
		post-SFT	0.39	1.24	4.71	12.48	0.12	0.79
	Test	pre-SFT	0.40	1.11	-59.79	59.79	0.09	0.67
		post-SFT	0.38	1.39	6.56	13.73	0.12	0.79
	Simulation Mode 3: Full Conversation Simulation							
	Train	pre-SFT	0.43	1.34	-57.43	57.59	0.09	0.75
		post-SFT	0.38	1.40	3.44	13.51	0.11	0.81
	Test	pre-SFT	0.38	1.30	-58.39	58.68	0.08	0.64
		post-SFT	0.35	1.42	6.47	13.71	0.11	0.79
Group Generalization	*Simulation Mode 1: Next Message Prediction*							
	Train	pre-SFT	0.49	1.12	-34.09	34.85	0.11	0.89
		post-SFT	0.45	1.15	4.80	12.99	0.14	0.73
	Test	pre-SFT	0.49	1.11	-33.86	34.72	0.11	0.94
		post-SFT	0.46	1.14	3.76	12.55	0.15	0.83
	Simulation Mode 2: Tweet-guided Conversation Simulation							
	Train	pre-SFT	0.43	1.27	-58.30	58.42	0.09	0.68
		post-SFT	0.37	1.31	6.49	13.77	0.11	0.72
	Test	pre-SFT	0.44	1.14	-58.32	58.65	0.09	0.72
		post-SFT	0.41	1.42	3.53	13.70	0.11	0.87
	Simulation Mode 3: Full Conversation Simulation							
	Train	pre-SFT	0.41	1.34	-57.85	58.06	0.08	0.68
		post-SFT	0.35	1.40	7.05	14.25	0.11	0.73
	Test	pre-SFT	0.43	1.26	-57.14	57.30	0.09	0.70
		post-SFT	0.38	1.41	2.54	15.00	0.11	0.73
Topic Generalization	*Simulation Mode 1: Next Message Prediction*							
	Train	pre-SFT	0.50	1.17	-35.04	36.01	0.11	0.88
		post-SFT	0.46	1.26	5.09	13.11	0.15	0.73
	Test	pre-SFT	0.47	1.00	-32.00	32.42	0.11	0.96
		post-SFT	0.45	1.08	3.32	11.50	0.14	0.77
	Simulation Mode 2: Tweet-guided Conversation Simulation							
	Train	pre-SFT	0.43	1.25	-58.49	58.54	0.09	0.65
		post-SFT	0.39	1.38	6.62	14.69	0.11	0.71
	Test	pre-SFT	0.42	1.23	-58.00	58.35	0.09	0.82
		post-SFT	0.38	1.15	5.19	11.47	0.11	0.80
	Simulation Mode 3: Full Conversation Simulation							
	Train	pre-SFT	0.42	1.37	-56.97	57.16	0.09	0.63
		post-SFT	0.38	1.39	5.04	14.80	0.10	0.74
	Test	pre-SFT	0.42	1.26	-58.93	59.14	0.08	0.85
		post-SFT	0.35	1.17	6.41	12.83	0.11	0.81

Table 13: Evaluation results for `Llama-3.1-8B-Instruct` across simulation modes, SFT types, and data partitions. We report the average semantic similarity $\overline{S}_{\text{sem}}$ (↑), average stance difference $\overline{\Delta}_{\text{stance}}$ (↓), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (→0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (↓), average ROUGE-L $\overline{\text{ROUGE-L}}$ (↑), and on-topic utterance rate $R_{\text{on-topic}}$. Blue cells indicate improved performance after SFT, while red cells indicate worsened performance. See Table 12 for SFT results with `gpt-4o-mini-2024-07-18`.

Generalization Type	Partition	Model	$\overline{S}_{\text{sem}}$ (↑)	$\overline{\Delta}_{\text{stance}}$ (↓)	$\overline{\Delta}_{\text{signed_len}}$ (→0)	$\overline{\Delta}_{\text{abs_len}}$ (↓)	$\overline{\text{ROUGE-L}}$ (↑)	$R_{\text{on-topic}}$
Round Generalization	*Simulation Mode 1: Next Message Prediction*							
	Train	pre-SFT	0.46	1.22	-38.74	39.67	0.08	0.87
		post-SFT	0.39	1.51	-3.15	16.64	0.07	0.58
	Test	pre-SFT	0.44	1.21	-40.05	41.21	0.08	0.88
		post-SFT	0.38	1.36	0.63	19.70	0.07	0.56
	Simulation Mode 2: Tweet-guided Conversation Simulation							
	Train	pre-SFT	0.42	1.30	-54.32	54.66	0.06	0.73
		post-SFT	0.38	1.44	-6.46	18.27	0.07	0.50
	Test	pre-SFT	0.39	1.27	-55.62	56.17	0.06	0.64
		post-SFT	0.35	1.50	-9.35	20.67	0.07	0.38
	Simulation Mode 3: Full Conversation Simulation							
	Train	pre-SFT	0.41	1.32	-54.42	54.93	0.06	0.72
		post-SFT	0.38	1.43	-7.02	19.80	0.07	0.46
	Test	pre-SFT	0.35	1.35	-55.37	55.80	0.06	0.62
		post-SFT	0.36	1.42	-6.01	19.39	0.07	0.35
Group Generalization	*Simulation Mode 1: Next Message Prediction*							
	Train	pre-SFT	0.45	1.20	-38.35	39.39	0.07	0.85
		post-SFT	0.39	1.21	-2.89	19.31	0.07	0.51
	Test	pre-SFT	0.47	1.26	-42.17	43.03	0.08	0.90
		post-SFT	0.40	1.35	-7.23	19.44	0.08	0.60
	Simulation Mode 2: Tweet-guided Conversation Simulation							
	Train	pre-SFT	0.41	1.31	-54.58	55.00	0.06	0.64
		post-SFT	0.39	1.37	-6.50	20.00	0.07	0.35
	Test	pre-SFT	0.43	1.23	-55.17	55.49	0.06	0.77
		post-SFT	0.40	1.30	-8.08	19.51	0.07	0.48
	Simulation Mode 3: Full Conversation Simulation							
	Train	pre-SFT	0.39	1.34	-54.78	55.28	0.06	0.64
		post-SFT	0.38	1.38	-6.60	20.31	0.07	0.35
	Test	pre-SFT	0.42	1.30	-54.38	54.82	0.06	0.73
		post-SFT	0.40	1.57	-13.37	25.11	0.06	0.38
Topic Generalization	*Simulation Mode 1: Next Message Prediction*							
	Train	pre-SFT	0.46	1.27	-39.54	40.57	0.08	0.84
		post-SFT	0.40	1.34	-5.64	19.10	0.07	0.60
	Test	pre-SFT	0.45	1.11	-38.45	39.40	0.07	0.92
		post-SFT	0.38	1.19	-4.86	18.61	0.07	0.58
	Simulation Mode 2: Tweet-guided Conversation Simulation							
	Train	pre-SFT	0.42	1.36	-53.92	54.44	0.06	0.67
		post-SFT	0.37	1.38	-13.25	22.65	0.06	0.41
	Test	pre-SFT	0.41	1.13	-56.46	56.59	0.06	0.66
		post-SFT	0.36	1.33	-12.37	22.11	0.07	0.39
	Simulation Mode 3: Full Conversation Simulation							
	Train	pre-SFT	0.39	1.36	-54.19	54.83	0.06	0.66
		post-SFT	0.38	1.60	-10.60	21.93	0.07	0.42
	Test	pre-SFT	0.40	1.26	-55.78	55.94	0.06	0.66
		post-SFT	0.36	1.24	-17.12	25.48	0.06	0.32

Below is a compile-ready LaTeX block that inserts an explicit **"Advantage (GAE) Computation" step** between the Rollout and Training phases. Copy–paste into any LaTeX/Markdown math renderer for a paper-style screenshot.

Conclusion. While SFT improves surface-level imitation, it fails to capture opinion-level behavioral alignment. Designing fine-tuning objectives that align with deeper social dynamics remains an important area for future work.