DEBATE: A Large-Scale Benchmark for Evaluating Opinion Dynamics in Role-Playing LLM Agents

Anonymous authorsPaper under double-blind review

000

001

002

004

006

008 009 010

011 012 013

014

015

016

017

018

019

021

024

025

026

027

028

029

031

033

034

037

038

040

041

042

043

044

046

047

048

051

052

ABSTRACT

Accurately modeling opinion change through social interactions is crucial for understanding and mitigating polarization, misinformation, and societal conflict. Recent work explores simulating opinion dynamics with role-playing LLM agents (RPLAs)—language models assigned human-like personas that engage in multiturn, multi-agent opinion exchange. However, existing RPLA simulations often produce unnatural group behaviors (e.g., premature consensus) and lack empirical benchmarks for evaluating alignment with real human interactions. We introduce **DEBATE**, the first large-scale benchmark for evaluating the authenticity of opinion dynamics in multi-agent RPLA simulations. DEBATE contains 37,357 messages from 2,792 U.S.-based participants who engaged in multi-player, multi-round conversations across 107 controversial topics, reporting both public messages and private beliefs. We simulate these conversations using various LLMs and introduce multi-level evaluation metrics (at the utterance, individual, and group levels) to assess behavioral alignment between humans and RPLAs. Our analyses reveal key behavioral gaps: RPLA groups exhibit stronger opinion convergence and belief drift than humans, and individual agents show more systematic shifts in response to social influence. Ablation studies further highlight the importance of private self-reported opinions in shaping realistic agent behavior. Additionally, while supervised fine-tuning improves surface-level metrics (e.g., ROUGE-L, message length), it falls short on deeper alignment (e.g., semantic and stance alignment). DEBATE enables benchmarking of simulated opinion dynamics and supports future research on aligning multi-agent RPLAs' simulations with realistic human interactions. The dataset and codebase will be publicly released.

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 (Lu et al., 2015; Pennycook et al., 2021; Budak et al., 2011; Loomba et al., 2021; Ginossar et al., 2022). Accurate modeling of these *opinion 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 (RPLAs) that embody diverse personas and engage in multi-turn dialogue (Park et al., 2023; Chuang et al., 2024a;b). Although individual RPLAs 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 RPLAs interact, they frequently exhibit premature consensus convergence, overly moderate stances, or unnatural patterns of opinion alignment, regardless of their initial diverse personas (Chuang et al., 2024a; Taubenfeld et al., 2024). Existing evaluations of RPLAs predominantly focus on single-agent scenarios or employ artificial, structured tasks, lacking robust empirical benchmarks capturing authentic human group dynamics in naturalistic contexts (Santurkar et al., 2023; Chuang et al., 2024c;b).

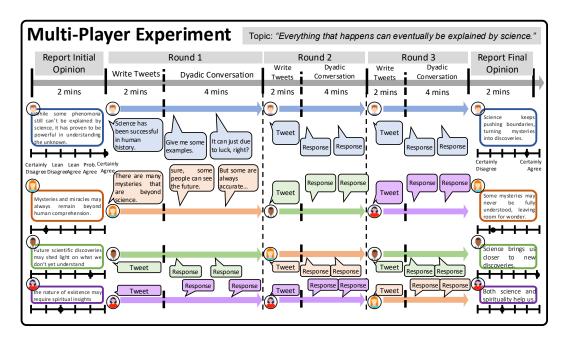


Figure 1: The procedure of the multi-player experiment. Each group is assigned a topic to discus about. Participants first report their initial opinion, then engage in three rounds of tweet writing and dyadic conversations with different partners, and finally submit their final opinion. With this setup,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

To address this critical gap, we introduce **D**eliberative Opinion **E**xchanges for **B**enchmarking **A**gent-based **T**rajectory **E**volution (**DEBATE**), the first large-scale empirical benchmark specifically designed for evaluating the authenticity of simulated opinion dynamics from multi-agent RPLAs. While the acronym suggests debate, DEBATE emphasizes naturalistic deliberation rather than competitive or adversarial argumentation. DEBATE comprises data with 37,357 messages from 824 groups and 2,792 U.S.-based participants engaged in multi-round, multi-party discussions on 107 controversial topics. It captures both publicly expressed *messages* (including both tweet-like posts and chat utterances) and privately reported beliefs (Likert-scale ratings). Due to occasional dropouts, out of all participants, 725 groups completed (2,584 participants) all sessions end-to-end, yielding a clean subset of 28,579 messages used for benchmarking and evaluation (average messages per group is 39.4). The benchmark enables quantitative assessment of alignment between simulated and actual human interactions at the *utterance*, *individual*, and *group* levels.

Contributions. (1) We introduce **DEBATE**, the first large-scale empirical benchmark for evaluating the human-likeness of opinion dynamics in multi-agent role-playing LLM agents (RPLAs). (2) DEBATE supports three simulation setups: *Next Message Prediction, Tweet-Guided Simulation*, and *Full Conversation Simulation*, covering various scenarios in social simulation. (3) We design quantitative evaluation metrics at three different levels (utterance, individual, and group) to assess different aspects of alignment between simulated and human opinion trajectories. (4) Supervised fine-tuning (SFT) improves surface-level text quality but not deeper stance or belief alignment, highlighting the need for future work on training RPLAs. (5) We identify key behavioral gaps between RPLAs and humans, including stronger opinion convergence and positive belief drift, showcasing the challenges of simulating realistic opinion dynamics with RPLAs. Our evaluation, fine-tunine, and analyses are intended as examples of how the benchmark can be used to evaluate human-RPLA alignment, and we want to note that the dataset itself is the primary contribution. We expect DEBATE to enable future studies on RPLA opinion dynamics simulations. The dataset will be released publicly upon acceptance. The codebase and a portion of the data is included in the supplementary materials (Appendix A).¹

¹LLM usage disclosure: ChatGPT was used only for language polishing; see Appendix S.

Table 1: Comparison of DEBATE with existing human opinion dynamics datasets. We categorize datasets into three genres: *competitive debate*, *asymmetric persuasion*, and *naturalistic deliberation*. Columns indicate whether each dataset supports multi-party interactions, multi-turn conversations, demographic attributes, reader-context traceability, enforced turn-taking, public/private opinions, data source and participants.

Dataset	Opinion Dynamics Type	Multi- Party (N≥3)	Turn	Demo- graphics	Reader- Context Traceable	Turn Obliga- tion	Public or Private Opinion	Data Source [Participants]	# Utterances (U) # Conversations (C) # Subjects (S)
CMV (Tan et al., 2016)	Competitive Debate	✓	1	Х	Х	Х	Public	Reddit threads [Reddit users]	U = 293,297 C = 3,051 S = 34,911
IAC 1.0 (Walker et al., 2012)	Competitive Debate	1	1	X	X	X	Public	Online debate forums [Online forum users]	U = 390,704 C = 11,800 S = 3,300
IAC 2.0 (Abbott et al., 2016)	Competitive Debate	1	✓	Х	X	Х	Public	Online debate forums [Online forum users]	U = 482,000 C = 16,461 S = 9,709
UK Parliament QuestionTime Corpus (Zhang et al., 2017)	Competitive Debate	1	1	X	1	✓	Public	British House of Commons [Members of Parliament]	U = 433,787 C = 216,894 S = 1,978
Intelligence Squared Debates Corpus (Zhang et al., 2016)	Competitive Debate	1	1	X	✓	✓	Public	Structured debate show [Invited experts]	U = 26,562 C = 108 S = 471
PersuasionForGood (Wang et al., 2019)	Persuasion (Asymmetric)	X	1	✓	✓	✓	Public	MTurk platform [Diverse laypeople]	U = 20,932 C = 1,017 S = 1,285
Wikipedia Articles-for-Deletion (AfD) (Mayfield and Black, 2019)		1	✓	X	X	X	Public	Wikipedia editor debates [Wikipedia editors]	U = 3,295,340 C = 383,918 S = 161,266
DEBATE	Deliberation (Naturalistic)	1	1	✓	1	1	Public + Private	Prolific platform [Diverse laypeople]	U = 28,579 C = 4,350 S = 2,584

2 RELATED WORK

Simulating and Evaluating Opinion Dynamics with RPLAs. Opinion dynamics refers to how individuals form, change, and negotiate beliefs through social interaction (Flache et al., 2017; Lorenz et al., 2021; Chuang and Rogers, 2023). Recent work has explored multi-agent LLM opinion interactions as a means to enhance downstream task performance, e.g., improving factuality, reasoning accuracy, and output diversity (Zhang et al., 2023; Chan et al., 2023; Chen et al., 2023; Du et al., 2023; Liang et al., 2023; Hu et al., 2024). However, these approaches primarily treat opinion exchange as a technique for boosting task performance, rather than aiming to simulate human-like opinion evolution

On the other other hand, a growing body of research instead focuses on simulating *human-like* opinion dynamics using RPLAs (Chuang et al., 2024a; Taubenfeld et al., 2024; Liu et al., 2024). These studies assign personas to agents and allow them to interact over multiple turns, aiming to model human-like opinion formation and change. However, most rely on qualitative observations in toy settings and lack empirical benchmarks for evaluation against real human behavior.

Recent efforts to quantify human-likeness of RPLAs' simulated opinion either on single-agent settings without interaction (Santurkar et al., 2023; Chuang et al., 2024c), or on non-linguistic tasks (Chuang et al., 2024b). In contrast, DEBATE introduces the first large-scale benchmark explicitly designed for evaluating multi-agent RPLAs in simulating natural-language opinion dynamics.

Existing Opinion Dynamics Corpora. Although no existing corpora were originally constructed as benchmarks for evaluating human-like opinion simulations of RPLAs, several contain human interactions involving opinion and have the potential to be adapted for this purpose. These can be grouped into three genres (Table 1; Walton and Krabbe, 1995; Walton et al., 2010; Bozdag et al., 2025): (1) *Competitive debate*, where participants aim to win an argument (e.g., CMV (Tan et al., 2016), IAC (Walker et al., 2012; Abbott et al., 2016)); (2) *Asymmetric persuasion*, where one party aims to influence another (e.g., PersuasionForGood (Wang et al., 2019)); and (3) *Naturalistic deliberation*, where peers voluntarily share and refine beliefs without roles or external incentives. This setting most closely relates to everyday social interactions. However, despite its relevance to real-world discourse, this deliberative genre is underrepresented in existing corpora.

Beyond genre coverage, most corpora lack key features for benchmarking human-like opinion dynamics. Many omit full reader-context traceability (e.g., CMV, IAC), making it difficult to reconstruct what input each speaker saw at the time of writing. Most also lack enforced *turn obligation*, which is critical for tracing each individual's belief trajectories over time. Importantly, no

Table 2: **Dataset statistics.** Each row reports the number of topics, messages, on-topic messages, subjects, groups, and the average number of groups per topic. Depth topics have 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	5,252	4,510	479	144	20.57
Breadth	100	23,327	21,538	2,105	581	5.81
Depth+Breadth	107	28,579	26,048	2,584	725	6.78

existing corpus combines both *publicly expressed* messages and *privately reported* beliefs. Measuring private self-report beliefs is important because individuals may publicly express socially desirable position that are different from their actual beliefs (Tourangeau and Yan, 2007). Demographic data is also often incomplete, limiting the construction of realistic personas for RPLAs. Finally, while some corpora involve selected experts (e.g., UK Parliament), we believe simulating belief change in *diverse laypeople* is more relevant for real-world applications.

Positioning DEBATE as a **Benchmark**. To address these gaps, DEBATE introduces not just a new dataset, but a full *evaluation benchmark* for multi-agent RPLAs' opinion dynamics simulation. It features multi-round conversations among participants with diverse backgrounds discussing controversial topics, with both public and private opinions, full reader context, enforced turn-taking, and rich demographics. Furthermore, it includes a suite of quantitative metrics designed to evaluate how closely simulated opinion trajectories match real human dynamics. This makes DEBATE a dedicated benchmark for evaluating the fidelity of RPLAs' opinion dynamics simulation.

3 DEBATE BENCHMARK: EMPIRICAL OPINION DYNAMICS FROM HUMAN

3.1 TASK

We design a multi-player conversational experiment to elicit naturalistic 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 $t\in\mathcal{T}$ throughout the session. The experiment lasts 25–30 minutes per group and consists of four phases. See Appendix R for the user interface and Appendix Q for sample conversation data.

- (1) Initial Private Opinion (2-minute): Each participant s_i reports an initial opinion $o_{s_i}^{\text{init}} \in \{-2.5, -1.5..., +2.5\}$ on a 6-point Likert scale², along with a free-form text justification $j_{s_i}^{\text{init}}$. They are submitted privately on a separate webpage so that no other members can view their responses.
- (2) Public Opinion Exchanges (6-minute): Participants engage in R=3 rounds of dyadic conversation. ³ 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.
- 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: $\mathcal{C}^r_{s_i,s_j} = \left[u^r_{1,s_i}, u^r_{2,s_j}, u^r_{3,s_i}, \ldots\right], \text{ where } u^r_{k,s} \text{ denotes the k-th utterance in the round-r conversation, with speaker <math>s \in \{s_i,s_j\}$. Speaker turns alternate between participants. Consecutive messages from the same speaker are merged during data preprocessing.
- (3) Post-discussion Private Opinion (2-minute): After the final round, each participant privately submits a final opinion $o_{s_i}^{\text{final}}$ and justification $j_{s_i}^{\text{final}}$ on a separate webpage similar to the initial opinion.

²Participants 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.

³Following standard setups in opinion dynamics simulations, we use *dyadic interactions* between each pair, which allows tracing of each individual's opinion trajectory.

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

3.2 TOPICS

The DEBATE benchmark includes two complementary topic sets: Depth and Breadth.

Depth Topics (\mathcal{T}_{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 RPLAs often drift toward ground-truth views over time, regardless of initial opinions (Chuang et al., 2024a; Taubenfeld et al., 2024). We selected seven such topics possessing high entropy across individuals in order to elicit diverse opinions from humans (Chuang et al., 2024c). Each topic was assigned to an average of 20.57 groups (479 participants in total; Table 2), allowing us to evaluate how systematically individual groups behave when discussing the same topic. See Appendix B for the full list.

Breadth Topics ($\mathcal{T}_{Breadth}$) contains 100 topics from the World Values Survey (WVS) (Haerpfer et al., 2022) and Pew Global Attitudes Survey (PGAS) (Pew Research Center, 2025). To reflect public disagreement in our US-based participants, we selected U.S.-administered questions with the highest response entropy (Durmus et al., 2024). Topics were phrased as self-contained declarative statements (e.g., "*Euthanasia can be justified*.") and spanned 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 was assigned to 6.78 groups (2,105 participants in total; Table 2). See Appendix C for 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 (Palan and Schitter, 2018)⁴. Participants were randomly assigned to one of 824 four-person groups and to a discussion topic. They remained anonymous to each other, identified only by randomly-assigned avatars and pseudonyms (e.g., ZK48UT). All procedures were approved by the Institutional Review Board (IRB). Participants were compensated at a rate of \$10/hour. Of the 824 groups recruited for DEBATE, some experienced participant dropouts or technical interruptions, resulting in partially completed sessions. For evaluation and analysis, we use a clean of 725 fully completed groups (2,584 participants, 28,579 utterances). However, we also release the full raw dataset (2,792 participants, 37,357 utterances), including incomplete sessions, as it may support other research use cases. Details on filtering criteria and raw data are provided in Appendix D.

The participants spanned 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 reported 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 E and Figure 4 for details).

4 Constructing and Evaluating Role-Playing LLM Agents

4.1 RPLA CONSTRUCTION GROUNDED IN HUMAN DATA

Each RPLA 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. Each a_i 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 F and Table 7 for prompt examples). We use notation with a hat and subscript a (e.g., $\widehat{\tau}_{a_i}^r$, \widehat{u}_{k,a_i}^r) to denote LLM-generated content, and notation without a hat and with subscript s (e.g., $\tau_{s_i}^r$, u_{k,s_i}^r) to denote human-written content.

⁴https://www.prolific.com/

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** $(o_{s_i}^{\text{init}}, j_{s_i}^{\text{init}})$: A 6-point Likert-scale opinion on the assigned discussion topic and a free-text justification. **3. Initial Tweet** $(\tau_{s_i}^1)$: The tweet posted at the beginning of round 1. **4. Previous Rounds:** Tweets $\{\tau_{s_i}^r, \widehat{\tau}_{a_i}^{r'}: 1 < r' < r\}$ and dyadic conversations $\{\mathcal{C}_{s_i,s_j}^{r'}, \widehat{\mathcal{C}}_{a_i,a_j}^{r'}: 1 \le r' < r\}$ from earlier rounds involving participant s_i . **5. Current Round Context:** The current tweet $\tau_{s_i}^r$ or $\widehat{\tau}_{a_i}^r$, the partner's tweet $\tau_{s_j}^r$ or $\widehat{\tau}_{a_j}^r$, and all utterances so far in the ongoing conversation $\{u_{k',s}^r, \widehat{u}_{k',a}^r: 1 \le k' < k\}$. The exact sources of memory vary by the simulation mode (Section 4.1; Table 3). 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 RPLAS

Table 3: Memory contents used in each simulation mode. All agents are conditioned on demographics d_{s_i} , initial opinion and justification $(o_{s_i}^{\text{init}}, j_{s_i}^{\text{init}})$, the initial tweet $\tau_{s_i}^1$, 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	$\begin{array}{lll} \operatorname{Human} & \{\tau^{r'} & : & 1 & \leq \\ r' \leq r\} & & & \end{array}$	$ \text{Human } \{\mathcal{C}^{r'}_{s_i,s_j}: 1 \leq \\ r' < r\} $		Predict a person's immediate response in real conversations
Mode 2: Tweet-guided Conversation Simulation	$\begin{array}{lll} \operatorname{Human} & \{\tau^{r'} & : & 1 & \leq \\ r' \leq r\} & \end{array}$	$ \begin{array}{ll} \text{Simulated} & \{\widehat{\mathcal{C}}_{a_i,a_j}^{r'} & : \\ 1 \leq r' < r\} \end{array} $		Simulate private conversations given a trace of real public tweets
Mode 3: Full Conversation Simulation	$\begin{array}{l} \operatorname{Human} \tau^1 + \operatorname{Simulated} \\ \left\{ \widehat{\tau}^{r'} : 2 \leq r' \leq r \right\} \end{array}$	$ \begin{array}{ll} \text{Simulated} & \{\widehat{\mathcal{C}}_{a_i,a_j}^{r'} & : \\ 1 \leq r' < r\} \end{array} $		Simulate agents' dynamics from initial conditions; the classic opinion dynamics simulation setup

We simulated each RPLA a_i 's utterance \widehat{u}_{k,a_i}^r in round r, turn k by generating:

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

where the speaker identity s_i is given, and only the utterance content is predicted.⁵ The same framework applies to generating tweets $\hat{\tau}^r_{a_i}$ and final opinions $(\hat{o}^{\text{final}}_{a_i}, \hat{j}^{\text{final}}_{a_i})$.

The DEBATE benchmark provides infrastructure for three simulation modes, corresponding to three common scenarios for simulation of social interactions: Next Message Prediction (Mode 1), Tweet-guided Conversation Simulation (Mode 2), and Full Conversation Simulation (Mode 3). All three are grounded in real human behavior but vary in how much human context is provided to the model. 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 $(o_{s_i}^{\text{init}}, j_{s_i}^{\text{init}})$, the initial tweet $\tau_{s_i}^1$, and task instructions (Section 4.1). 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 3 summarizes the full memory configuration and the corresponding use case for each simulation mode.

We evaluate RPLA simulations using six different LLMs: gpt-4o-mini-2024-07-18 (OpenAI, 2022), Llama-3.1-Tulu-3-8B-SFT (Lambert et al., 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024), Llama-3.1-70B-Instruct, Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Qwen2.5-32B-Instruct (Bai et al., 2023). These models span open vs. proprietary weights, varying parameter scales, and both pre-alignment and post-alignment checkpoints. Full compute details are in Appendix I.

4.3 EVALUATION

We evaluated how well an RPLA a_i simulates its corresponding human participant s_i by comparing utterances \widehat{u} and u within the dyadic conversations, focusing only on *on-topic* utterances—those

⁵Since consecutive messages from the same speaker are merged during preprocessing, speakers alternate turns, making the speaker order known (Section 3.1).

Table 4: Evaluation results across simulation modes and LLMs. We report the round-wise aggregated metrics on the **Depth Topics**: average semantic similarity $\overline{S}_{\text{sem}}$ (\uparrow), average stance difference $\overline{\Delta}_{\text{stance}}$ (\downarrow), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (\rightarrow 0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (\downarrow), $\overline{\text{ROUGE-L}}$ (\uparrow), and on-topic utterance rate $R_{\text{on-topic}}$. Error bars indicate standard error from 1,000 bootstrap resamples.

LLM & Simulation Mode	$\overline{S}_{\text{sem}} (\uparrow)$	$\overline{\Delta}_{\text{stance}} \left(\downarrow \right)$	$\overline{\Delta}_{\text{signed_len}} (\rightarrow 0)$	$\overline{\Delta}_{abs_len} \left(\downarrow\right)$	ROUGE-L (↑)	$R_{ ext{on-topic}}$
Simulation Mode 1: Next Mess	age Prediction					
gpt-4o-mini-2024-07-18	0.48 ± 0.01	1.16 ± 0.05	-32.72 ± 0.62	33.51 ± 0.60	$\textbf{0.11} \pm \textbf{0.01}$	0.74
Llama-3.1-Tulu-3-8B-SFT	0.44 ± 0.01	1.19 ± 0.06	-41.57 ± 1.28	45.07 ± 0.93	0.06 ± 0.01	0.56
Llama-3.1-8B-Instruct	0.45 ± 0.01	1.21 ± 0.04	-36.85 ± 0.87	38.37 ± 0.76	0.07 ± 0.01	0.73
Llama-3.1-70B-Instruct	0.45 ± 0.01	$\textbf{1.15} \pm \textbf{0.05}$	-26.19 ± 1.05	28.88 ± 0.86	0.08 ± 0.01	0.78
Mistral-7B-Instruct-v0.3	0.47 ± 0.01	1.18 ± 0.05	-46.27 ± 0.71	46.79 ± 0.67	0.07 ± 0.01	0.72
Qwen2.5-32B-Instruct	0.46 ± 0.01	1.16 ± 0.05	-22.40 \pm 0.82	$\textbf{27.12} \pm \textbf{0.64}$	0.08 ± 0.01	0.73
Simulation Mode 2: Tweet-guid	ded Conversatio	n Simulation				
gpt-4o-mini-2024-07-18	$\textbf{0.42} \pm \textbf{0.01}$	1.25 ± 0.05	-58.40 ± 0.78	58.56 ± 0.76	$\textbf{0.09} \pm \textbf{0.01}$	0.66
Llama-3.1-Tulu-3-8B-SFT	0.41 ± 0.01	1.34 ± 0.07	-53.66 ± 0.88	54.38 ± 0.82	0.05 ± 0.01	0.48
Llama-3.1-8B-Instruct	0.41 ± 0.01	1.28 ± 0.05	-52.81 ± 0.93	53.31 ± 0.86	0.06 ± 0.01	0.67
Llama-3.1-70B-Instruct	0.40 ± 0.01	$\textbf{1.18} \pm \textbf{0.05}$	-51.24 ± 1.24	51.99 ± 1.17	0.06 ± 0.01	0.72
Mistral-7B-Instruct-v0.3	0.41 ± 0.01	1.21 ± 0.06	-46.77 ± 0.76	$\textbf{47.26} \pm \textbf{0.70}$	0.06 ± 0.01	0.63
Qwen2.5-32B-Instruct	0.41 ± 0.01	1.25 ± 0.06	-47.84 ± 1.31	49.79 ± 1.10	0.07 ± 0.01	0.66
Simulation Mode 3: Full Conve	ersation Simula	tion				
gpt-4o-mini-2024-07-18	$\textbf{0.41} \pm \textbf{0.01}$	1.30 ± 0.05	-58.11 ± 0.73	58.26 ± 0.71	$\textbf{0.08} \pm \textbf{0.01}$	0.65
Llama-3.1-Tulu-3-8B-SFT	0.40 ± 0.01	1.46 ± 0.07	-55.00 ± 0.92	55.84 ± 0.82	0.05 ± 0.01	0.46
Llama-3.1-8B-Instruct	0.39 ± 0.01	1.33 ± 0.05	-52.84 ± 0.91	53.39 ± 0.85	0.06 ± 0.01	0.67
Llama-3.1-70B-Instruct	0.38 ± 0.01	1.27 ± 0.05	-51.14 ± 1.12	52.03 ± 1.01	0.06 ± 0.01	0.72
Mistral-7B-Instruct-v0.3	0.40 ± 0.01	$\textbf{1.25} \pm \textbf{0.05}$	$\textbf{-46.69} \pm \textbf{0.80}$	$\textbf{47.25} \pm \textbf{0.73}$	0.06 ± 0.01	0.61
Qwen2.5-32B-Instruct	0.40 ± 0.01	1.30 ± 0.05	-49.22 ± 1.11	50.83 ± 0.96	0.07 ± 0.01	0.65

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?"). From these we assessed different aspects of human-model alignment with the following metrics: 1. Semantic **Similarity:** $S_{\text{sem}}(u, \widehat{u}) = \cos(E(u), E(\widehat{u}))$, where $E(\cdot)$ is a sentence encoder (Sturua et al., 2024). This measures the meaning-level similarity between utterances, capturing whether the agent expresses a semantically similar idea. **2. Stance Difference:** $\Delta_{\text{stance}}(u, \hat{u}) = |S(u) - S(\hat{u})|$, 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: $\Delta_{abs\ len} = ||u| - |\widehat{u}||$; $\Delta_{\text{signed len}} = |u| - |\widehat{u}|$. These reflect surface-level stylistic similarity in verbosity and message length. 4. ROUGE-L: Longest common subsequence score (Lin, 2004). This quantifies tokenlevel 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 were on-topic 86% (Depth) and 91% (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 $I_{\text{topic}}(\cdot,t)$ were predicted by an LLM and validated against human annotations (Appendix G).

Because there is no one-to-one mapping between simulated and human utterances, we adopted 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 H for details).

5 UTTERANCE-LEVEL EVALUATION OF ROLE-PLAYING LLM AGENTS

Alignment Across Three Social Simulation Modes. Tables 4 and 11 report evaluation results across simulation modes and LLMs for Depth and Breadth topics, respectively. Two consistent trends emerged across all metrics and topic types. First, gpt-4o-mini-2024-07-18 consistently showed the strongest alignment with human responses, achieving the best scores on semantic similarity (\overline{S}_{sem}), ROUGE-L (\overline{ROUGE} -L), and stance difference ($\overline{\Delta}_{stance}$). To account for variability across topics and simulation conditions, we conducted statistical tests across six experimental settings. A Friedman test followed by Wilcoxon signed-rank tests confirmed that gpt-4o-mini significantly

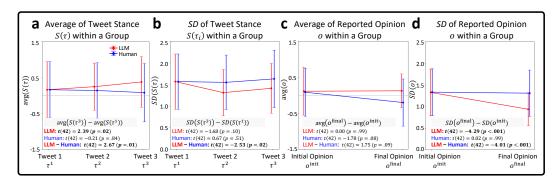


Figure 2: Group-level trajectories of tweet stance and self-reported opinion for human groups (blue) and their corresponding RPLA groups (red). (a) Average tweet stance $S(\tau)$ within each group across three rounds. (b) Standard deviation (SD) of tweet stance $S(\tau)$ within each group across rounds. (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 RPLA groups are also statistically tested.

outperformed most other models across all three metrics (see Appendix J for full results). However, it tended to produce longer messages than humans, as indicated by the negative signed length difference $\overline{\Delta}_{\text{signed_len}}$. Second, alignment declined for simulation modes with less human-generated context: Mode 1 (Next Message Prediction) performed best, followed by Mode 2 (Tweet-guided Conversation), with Mode 3 (Full Conversation) performing worst.

Ablation Studies. To understand the contribution of different memory components in RPLAs, we systematically ablated individual parts of the memory module $\mathcal{M}_{a_i,k}$ (Section 4.1) and evaluated their effects on alignment. Each ablation isolates a specific type of information: **1. No Previous Chat** removes all prior tweets and dyadic conversations, **2. No Initial Opinion** removes the participant's private Likert-scale belief and justification, **3. No Demographics** removes background attributes such as age, gender, and political ideology, and **4. No Private Profile** removes both demographics and initial opinion. All other components of memory remain unchanged in each condition.

Tables 13 and 14 (Appendix M) present ablation results for Depth and Breadth topics using gpt-40-mini. Across both topic sets, in Simulation Mode 1 (Next Message Prediction), where models are provided with full human-generated context, ablations have minimal effect on semantic similarity and stance alignment. In contrast, for Simulation Modes 2 and 3 (Tweet-guided and Full Conversation Simulation), where model-generated messages accumulate over rounds, removing private initial opinion consistently worsens stance alignment across both Depth and Breadth topics, while semantic similarity remains relatively stable across conditions. ⁶ These findings highlight the importance of grounding RPLAs with actual human private information for opinion dynamics sumulation.

Supervised Fine-tuning. To test whether behavioral alignment can be improved through fine-tuning, we conducted preliminary experiments using supervised fine-tuning (SFT). While SFT improved surface-level alignment (e.g., message length, ROUGE-L), it failed to enhance deeper metrics such as semantic similarity or stance alignment (Appendix N; Table 15, 16). 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

Beyond *utterance-level* alignment, realistic simulations must capture *group-level* and *individual-level* opinion dynamics. We focus on groups discussing Depth topics (which offer higher data density and

 $^{^6}$ On the other hand, message length changes were mostly trivial (≤ 2 tokens, < 5%); only the No Prior Chats condition in Mode 1 meaningfully increased message length (Depth +6.0, Breadth +7.2, consistent with greater verbosity in the absence of prior conversational context.

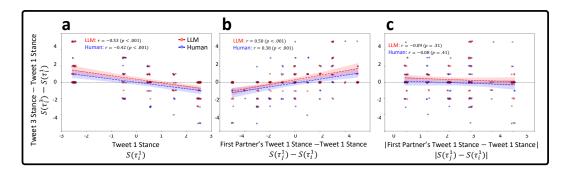


Figure 3: Individual-level opinion change and its predictors. (a) Change in tweet stance $(S(\tau_i^3) - S(\tau_i^1))$ negatively correlates with initial stance $S(\tau_i^1)$, (b) positively correlates with directional difference between first partner's stance and own stance, and (c) has no relationship when using absolute stance difference. Shaded regions show standard error. See Figure 6 for the same analysis on self-reported opinion.

tied to a known ground truth) using Full Conversation Simulation (Mode 3; Section 4.2), which best mirrors classic opinion dynamics setups. Simulations used the model with the strongest semantic alignment (gpt-40-mini-2024-07-18; Section 5).

Group-Level Opinion Shifts. We evaluated group-level opinion change by comparing tweet stance $S(\tau^3) - S(\tau^1)$ and self-reported opinion $o^{\text{final}} - o^{\text{init}}$. Figure 2a shows that LLM groups significantly increased tweet stance across rounds (t(42) = 2.39, p = .02), whereas human groups did not (t(42) = -0.21, p = .84). Because stance polarity is aligned such that positive values indicate greater agreement with a *false* belief, LLM groups became more wrong over time t(42) = 2.67, p = .01), diverging from humans.

LLMs also showed a significant reduction in tweet stance variance over time, suggesting stronger opinion convergence (t(42) = -2.53, p = .02), while human groups showed no such change (t(42) = 0.67, p = .51). Self-reported opinions showed a similar pattern. See details in Appendix O.

Mechanisms of Individual Opinion Change. We next examined how individuals updated their tweet stance across rounds, focusing on two mechanisms: regression toward the mean and influence from a conversation partner. Figures 3a–c show tweet stance change $S(\tau_i^3) - S(\tau_i^1)$ plotted against initial stance. Individuals with more extreme initial views reliably moved toward the midpoint (Humans: r=-0.42, LLMs: r=-0.53; both p<.001). Likewise, participants shifted toward their first partner's stance (Humans: r=0.38, LLMs: r=0.50). As a control, absolute difference from their first partner's stance has no effect (Figure 3c). Similarly, Self-reported opinions followed the same pattern of stronger convergence and partner influence in LLMs and human. Notably, while human and LLM behaviors were remarkably similar in terms of these two mechanism, correlation magnitudes are consistently larger for LLM than humans. See details in Appendix P.

Summary. We identify three key differences in opinion dynamics between LLMs and humans. Compared to humans, LLM groups show stronger opinion convergence, positive belief drift in tweet stance, and more systematic individual shifts: both in regression to the mean and in partner influence.

7 CONCLUSION

We introduced **DEBATE**, the first large-scale empirical benchmark for evaluating opinion dynamics in multi-agent role-playing LLM agent (RPLA) systems. By capturing naturalistic opinion trajectories from 2,584 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 RPLAs reproduce some utterance-level patterns, they fall short in deeper opinion alignment and belief updating. We propose an evaluation framework and identify systematic behavioral differences between human and RPLA-simulated groups. We hope DEBATE provides a foundation for developing more socially grounded and human-aligned multi-agent RPLA systems.

ETHICS STATEMENT

This study was reviewed and approved by our Institutional Review Board (IRB) and judged to pose minimal risk. All participants provided informed consent and were explicitly told they could discontinue at any time without any penalty. Participants were compensated at fair hourly rates. No deception was used.

All data are fully de-identified prior to release: real names and direct identifiers are not collected; platform IDs are replaced with random pseudonyms. We run basic automated and manual checks to remove any potential residual personal information. The dataset is released strictly for research purposes under a non-commercial license. Code will be released under an open-source license, and all API usage (e.g., OpenAI) complied with providers' terms of use. We will document dataset schema, known limitations, and intended use, and require users to accept the terms prior to access.

Collecting a wide range of viewpoints is not an endorsement of any particular position. Rather, they are necessary to study societal risks such as misinformation spread, polarization, and echo-chamber formation, and to develop mitigation strategies. To support fairness, we recruited a demographically diverse U.S. sample (spanning age, gender, race/ethnicity, education, income, occupation, and political leanings). Nonetheless, the data are U.S.-based and not nationally representative; downstream users should avoid over-generalization and should re-validate findings in other populations.

Finally, we emphasize that the benchmark itself is the primary contribution. Our empirical evaluations are examples of how the benchmark can be used; many additional analyses are possible. We are committed to monitor potential issues post-release, including updating documentation, modifying released date, and licensing terms if new risks are identified.

REPRODUCIBILITY STATEMENT.

We have made sifgnificant efforts to ensure the reproducibility of our results. Appendix A outlines our released codebase, which includes data preprocessing scripts, simulation pipelines for generating LLM-based conversations, evaluation metrics, model fine-tuning, and statistical analysis routines. All simulation modes and prompt templates are documented and implemented. We also include a subset of the DEBATE dataset with the submission, and will release the full dataset upon acceptance.

REFERENCES

- Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, 2016.
- 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.
- Nimet Beyza Bozdag, Shuhaib Mehri, Xiaocheng Yang, Hyeonjeong Ha, Zirui Cheng, Esin Durmus, Jiaxuan You, Heng Ji, Gokhan Tur, and Dilek Hakkani-Tür. Must read: A systematic survey of computational persuasion. *arXiv preprint arXiv:2505.07775*, 2025.
- 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.
- 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.
- 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.
- Yun-Shiuan Chuang and Timothy T Rogers. Computational agent-based models in opinion dynamics: A survey on social simulations and empirical studies. *arXiv preprint arXiv:2306.03446*, 2023.
- 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, 2024a.
- 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, 2024b.
- 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, 2024c.
- 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.
- 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.
- 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.
- Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4), 2017.
- 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.

- 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/.
 - 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.
 - 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.
 - 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.
 - 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.
 - 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/.
 - Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. In *IJCAI*, 2024.
 - 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.
 - Jan Lorenz, Martin Neumann, and Tobias Schröder. Individual attitude change and societal dynamics: Computational experiments with psychological theories. *Psychological Review*, 128(4):623, 2021.
 - 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.
 - Elijah Mayfield and Alan W. Black. Analyzing wikipedia deletion debates with a group decision-making forecast model. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–26, 2019.
 - OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022. [Accessed 13-10-2023].
 - Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, 17:22–27, 2018.
 - 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.
 - 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.
 - Pew Research Center. Pew research center: Numbers, facts and trends shaping your world. https://www.pewresearch.org/, 2025.
 - 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/.
 - 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.

- 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.
 - Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624, April 2016.
 - 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.
 - Roger Tourangeau and Ting Yan. Sensitive questions in surveys. *Psychological bulletin*, 133(5):859, 2007.
 - Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, volume 12, pages 812–817, 2012.
 - Douglas Walton and Erik CW Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning.* State University of New York Press, 1995.
 - Douglas Walton, Katie Atkinson, et al. Argumentation in the framework of deliberation dialogue. In *Arguing global governance*, pages 230–250. Routledge, 2010.
 - Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv* preprint arXiv:1906.06725, 2019.
 - 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.
 - Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational flow in oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, 2016.
 - Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. Asking too much? the rhetorical role of questions in political discourse. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1572, 2017.

SUPPLEMENTARY MATERIALS OVERVIEW (CODEBASE AND DATA)

708

709

710

711

712

713

714

The supplementary material includes a codebase for the full implementation of our multi-agent conversational simulation framework for analyzing opinion dynamics. The codebase contains five integrated pipelines: (1) preprocessing for data standardization, (2) simulation for generating LLMbased conversations across different modes, (3) evaluation for comparing human and LLM outputs using similarity metrics and belief trajectories, (4) group-level statistical analysis, and (5) model fine-tuning.

The code supports all three simulation setups (full conversation simulation, tweet-guided simulation, and next message prediction—and) includes prompt templates, evaluation scripts, and detailed documentation for reproducing all results reported in the paper.

A subset of the DEBATE dataset is included for transparency and reproducibility. The full dataset will be released publicly upon acceptance and included in the camera-ready version.

715 716 717

В **DEPTH TOPIC CONSTRUCTION**

718 719

720

721

722

723

The following seven topics are used as the Depth topic set (\mathcal{T}_{Depth}) . These topics are selected from a prior study (Chuang et al., 2024c), 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.

724 725

1. A "body cleanse," in which you consume only particular kinds of nutrients over 1–3 days, helps your body to eliminate toxins.

726 727

2. Angels are real.

728 729 3. Everything that happens can eventually be explained by science.

730 731

4. Regular fasting will improve your health. 5. The U.S. deficit increased after President Obama was elected.

positive values always indicate endorsement of the false statement.

732 733 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.

734 735

736

737

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

738 739 740

BREADTH TOPIC CONSTRUCTION

745

746

747

The Breadth topic set $(\mathcal{T}_{Breadth})$ consists of 100 topics curated from two large-scale cross-national surveys: the World Values Survey (WVS) (Haerpfer et al., 2022) and the Pew Global Attitudes Survey (PGAS) (Pew Research Center, 2025). 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 (Durmus et al., 2024).

748 749 750

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

751 752

Original questions:

753 754

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.)

763

764

765

766

767

768

769

770 771

772 773

774

775

776

777

778 779

781

782

783 784

785

786

787

788 789 790

791 792

793

794

795

796

797

798

799

800

801

802 803

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

758

Some other questions, however, are framed in a multiple-choice format. To convert these into clearly 759

debatable statements, we reframe the most frequently chosen responses as separate topic statements.

760 For example: 761 762

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.

D DATASET CLEANING AND COMPLETION CRITERIA

DEBATE initially included 824 groups comprising 29,792 participants and a total of 37,357 utterances. However, not all experiments were completed end-to-end due to technical errors (e.g., server connection issues) or participant dropouts.

We define a session as *complete* if all three rounds of dyadic conversation were successfully completed by at least three participants in a group. Based on this criterion, we identify 725 groups (2,584 participants) with 28,579 utterances as the **clean subset**, used for all benchmark evaluation and analysis in this paper.

Note that when a participant exited mid-session, the conversation with that specific participant paused with a waiting notice. These partial transcripts are included in the full dataset but excluded from the clean subset.

DEMOGRAPHIC SUMMARY

804 805 806

807

808

809

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

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.

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

Table 6: **Full dataset statistics.** Each row summarizes statistics from all collected sessions, including both completed and partially completed conversations.

Dataset	# topics	# messages	# subjects	# groups	# groups/topic
Depth	7	7801	501	185	26.43
Breadth	100	29566	2291	639	6.39
Depth+Breadth	107	37357	2792	824	7.70

care, education, manufacturing, media, construction, among many. This heterogeneity ensures a rich and representative foundation for studying opinion dynamics and belief-based interactions.

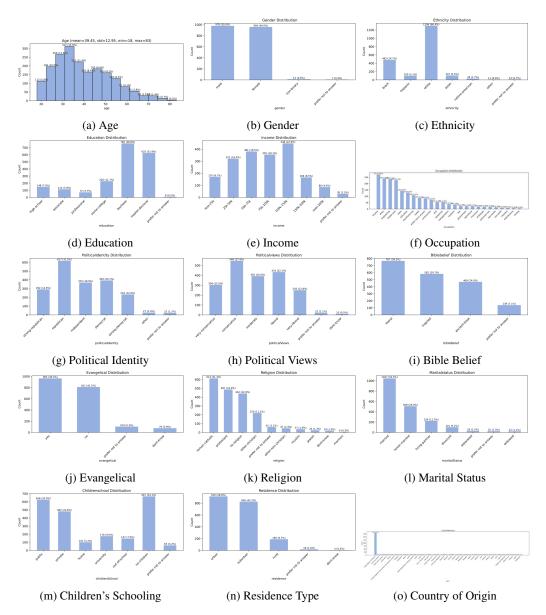


Figure 4: Demographic distributions across age, gender, education, ethnicity, income, political identity and views, religion, family, and geographic background.

F 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 RPLA 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 7 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 $(o_{s_i}^{\text{init}}, j_{s_i}^{\text{init}})$, previous rounds' tweets and dyadic conversations $\{\tau_s^{r'}, \mathcal{C}_s^{r'}: 1 \leq r' < 3\}$, and current round context including partner tweets and prior utterances $(\tau_{s_i}^3, \tau_{s_j}^3, \{u_{k',s}^3: k' < k\})$. Curly brackets ($\{\}$) denote placeholder variables specific to each agent and topic instance. For readability, color highlights in the table correspond to different memory components.

Table 7: Prompt templates used to construct the memory module $\mathcal{M}_{a_i,k}$ for each RPLA 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 agree - Lean disagree - Lean disagree - Lean disagree - Lean disagree - 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 politic system is functioning in the US these days." Initial Opinion: topic = "You are satisfied with how the politic system is functioning in the US these days." Likert-scale response = "Probably agree" Explanation = "I am indeed satisfied with the politic system because the government is trying hard enoug to introduce cryptocurrency to the market, which the future currency of the world."

1026				
1027	Conversation History:	User Message	Below was your conversation with {first_partner_name}	Previous Rounds (Round 1): $\tau_{s_1}^1$ (Your tweet) = I am satisfied with political system
1028	Previous Rounds	message		because the government is trying hard enough to
1029	(Round 1 & Round 2),		My tweet: $\{\tau_{s_1}^1\}$	stabilize the economy through various ways like
1030	Current Round		{first_partner_name}'s tweet: $\{\tau_{s_2}^1\}$	transitioning to crypto currency $\tau_{s_2}^1$ (681e3's tweet) = I disagree with the statement
1031	Context (Round		My response: $\{u_{1,s_1}^1\}$	that I am satisfied with the way the American system
1031	3)		{first_partner_name}'s response: $\{u_{2,s_2}^2\}$	functions these days. This is because of the system's extreme polarization making it fail to take meaningful
1032			My response: $\{u_{3,s_1}^1\}$	action
1033			{first_partner_name}'s response: $\{u_{4,s_2}^2\}$	u_{1,s_1}^1 (Your response) = From my point of view, the
1034			···	government is not that perfect but at least it's trying to improve the lives of all Americans
1035			Von have just finished your convenantion with	u_{2,s_2}^2 (681e3's response) = I still believe that the
1036			You have just finished your conversation with {first_partner_name}. Instead, you are now en-	political system is flawed but I completely see your viewpoint. Politicians appear to care more about
			gaging in conversation with another stranger	maintaining party allegiance than they do about the
1038			{second_partner_name} on a separate text box.	problems that people care about. Can the system be re-organized in your opinion or is the division too
1039			Below was your conversation with {sec-	great? The fact that everything has become more
1040			ond_partner_name }.	divisive which makes compromise nearly impossible in my opinion is largely to blame. What do you think?
1041			My tweet: $\{\tau_{s_1}^2\}$	in my opinion is targety to braine. What do you tillik:
1042			{second_partner_name}'s tweet: $\{\tau_{s_3}^2\}$	Previous Rounds (Round 2): $\tau_{s_1}^2$ (Your tweet) = We should support the government
1043			{second_partner_name}'s response: $\{u_{1,s_2}^2\}$	r_{s_1} (rotal tweet) – we should support the government motives to improve and make our country great. On
1044			My response: $\{u_{2,s_1}^2\}$	my side, the government is doing the best it can to
1045			{second_partner_name}'s response: $\{u_{3,s_3}^2\}$	stabilize our economy and improve our lives $\tau_{s_3}^2$ (683b8's tweet) = I agree and I am totally satisfied
1046			My response: $\{u_{4,s_1}^2\}$	with how the political system is working. This is
1047				because it is promoting good health and education facilities to its citizens.
1048				u_{1,s_2}^2 (683b8's response) = It provide strict laws. It
1049			You have just finished your conversation with {second_partner_name}. Instead, you are now	gives freedom to all citizens to publicly participate in
1050			engaging in conversation with another stranger	elections. u_{2,s_1}^2 (Your response) = I second your point, the
1051			{third_partner_name} on a separate text box.	government has helped the education sector through
1052			Below was your conversation with	scholarships. It has also invested a lot of resources in the healthcare field. Yes, it also gives each citizen the
1053			{third_partner_name}.	right to express one's ideas and opinions.
1054			My tweet: $\{\tau_{s_1}^3\}$	u_{3,s_3}^2 (683b8's response) = It has also improved infrastructure and advancement of technology.
1055			{third_partner_name}'s tweet: $\{\tau_{s_4}^3\}$	infrastructure and advancement of technology.
1056			My response: $\{u_{1,s_1}^3\}$	Current Round Context (Round 3):
1057			{third_partner_name}'s response: $\{u_{2,s_3}^3\}$	$\tau_{s_1}^3$ (Your tweet) = The government plays a crucial role in advancement of technology by budgeting
1058			My response: $\{u_{3,s_1}^3\}$	enough resources. It also helps in infrastructure and
1059			{third_partner_name}'s response: $\{u_{4,s_3}^3\}$	healthcare, I support $\tau_{s_4}^3$ (68405's tweet) = The government allows its
1060				people participation on the development project and is
1061				highly working on development
1062				u_{1,s_4}^3 (68405's response) = I do agree on advancing the technology and improving also in defense force
1063				and provide high security
1064				u_{2,s_1}^3 (Your response) = Yes, the government contributes to the general development of the country
1065				by investing enough money onto different projects
1066				$u_{3,s4}^3$ (68405's response) = That's okay. It's also
1067				improving on more projects and inventions u_{4,s_1}^3 (Your response) = It also contributes to a stable
1068				economy

G LLM-BASED ANNOTATION FOR ON-TOPICNESS AND STANCE

To evaluate RPLA 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 return a binary label. To ensure reliability, we manually labeled 200 utterances with binary on-topic

Table 8: Prompt template used for on-topic classification with gpt-4o-mini-2024-07-18. Example utterances are described in Table 9.

1080

1082

1084

1087

1089

Prompt Template

System Message

Your task is to analyze the provided conversation. The conversation can either be between two humans or two RPLAs. 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 RPLAs 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:

1090 {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:

1093 {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.":

096 {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:

1100 Invalid example where a role-playing LLM repeats the instruction:

(INVALID_EXAMPLE_INSTRUCTION)

Invalid example where a role-playing LLM generates a json object:

1102 {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:

1104 {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:

1106 {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:

1110 {INVALID_EXAMPLE_MULTI_TURN}

User Message

Below is the conversation history up to the latest message.

{CONVERSATION_HISTORY}

1113 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.

111611171118

1119

1120

1121

1105

1107

1108

1109

1111

1114

1115

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 8, and examples of on-topic and off-topic utterances are listed in Table 9.

112211231124

1125

1126

1127

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.

1128 1129 1130

1131

1132

1133

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. The full prompt template is shown in Table 10.

Table 9: Examples of utterances used in the on-topic classification prompt. Each example corresponds to a placeholder in the template from Table 8. All examples are about the topic "regular fasting will improve your health."

Prompt Examples

1134

1135

1136

1137

1138 1139

1140

1141

1142

1143 1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157 1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

${VALID_EXAMPLE_LLM}$

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

{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}

```
1168
            ""json
1169
           "persona":
           "id": "87774",
1170
           "age": 18,
1171
           "education_level": "some_college_education",
1172
           "ethnicity": "white",
           "income_range": "$75-$125K annually",
1173
           "political_affiliation": "moderate_democrat",
          "residence_type": "urban_area",
"marital_status": "never_married",
1174
1175
           "religious_beliefs": "ancient_book_bible_no_evangelical_no_religion",
1176
1177
```

{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'

Table 10: Prompt template used for stance classification.

Prompt Template

Plain, in Evaluation without prior messages

What is the stance of the following message with respect to the topic "{TOPIC}"? Here is the message: "{TEXT}". Please use exactly one word from the following 6 categories to label it: "Certainly agree", "Probably agree", "Lean agree", "Lean disagree", "Probably disagree", and "Certainly disagree". Your predicted label:

Memory, in Evaluation with prior messages

Here is prior conversation around the topic "{TOPIC}":

{CONVERSATION}

What is the stance of the following new message by {PLAYER} with respect to the topic "{TOPIC}"? Here is the message: "{TEXT}". Please use exactly one word from the following 6 categories to label it: "Certainly agree", "Probably agree", "Lean agree", "Lean disagree", "Probably disagree", and "Certainly disagree". Your predicted label:

User Message in Simulation for post-discussion private opinions

Based on your previous discussions about {TOPIC}, please select your level of agreement from the following options. Strictly answer with one of the options below without any additional text:

- Certainly disagree
- Probably disagree
- Lean disagree
- Lean agree
 - Probably agree
 - Certainly agree

H ROUND-WISE AGGREGATED EVALUATION METRICS

 We define the following sets used throughout evaluation: $\widehat{\mathcal{U}}$ and \mathcal{U} denote all utterances generated by RPLAs 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 $\widehat{\mathcal{U}}_{\text{topic}, s_i}^r$ 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:

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

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

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 G.

Stance Classification. To assess opinion alignment, each utterance u is mapped to a scalar stance score S(u) via a GPT-40-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 G.

Semantic Embedding. The sentence encoder $E(\cdot)$ used in S_{sem} is based on jinaai/jina-embeddings-v3 (Sturua et al., 2024), which produces 1024-dimensional embeddings. Semantic similarity is computed as cosine similarity between embedded vectors: $S_{\text{sem}}(u,\widehat{u}) = \cos(E(u),E(\widehat{u}))$

I COMPUTE RESOURCES

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

J 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.

J.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.

J.2 ROUGE-L (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).

J.3 STANCE DIFFERENCE $(\overline{\Delta}_{STANCE})$

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 (R Core Team, 2024).

K SIMULATION RESULTS ON BREADTH TOPICS

Table 11 presents alignment results across simulation modes and LLMs on the Breadth topics.

L ABLATION RESULTS ON BREADTH TOPICS

To complement the main results on Depth topics (Table 13), Table 14 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.

Table 11: Evaluation results across simulation modes and LLMs. We report the round-wise aggregated metrics on the **Breadth Topics**: average semantic similarity $\overline{S}_{\text{sem}}$ (\uparrow), average stance difference $\overline{\Delta}_{\text{stance}}$ (\downarrow), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (\rightarrow 0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (\downarrow), $\overline{\text{ROUGE-L}}$ (\uparrow), and on-topic utterance rate $R_{\text{on-topic}}$. Error bars indicate standard error from 1,000 bootstrap resamples.

LLM & Simulation Mode	$S_{\text{sem}} (\uparrow)$	$\Delta_{\text{stance}} \left(\downarrow \right)$	$\Delta_{\text{signed_len}} (\rightarrow 0)$	$\Delta_{abs_len} (\downarrow)$	ROUGE-L (†)	$R_{ ext{on-topic}}$
Simulation Mode 1: Next Mess	age Prediction	(v2)				
gpt-4o-mini-2024-07-18	$\textbf{0.49} \pm \textbf{0.01}$	1.04 ± 0.03	-33.20 ± 0.27	34.33 ± 0.25	$\textbf{0.10} \pm \textbf{0.01}$	0.83
Llama-3.1-Tulu-3-8B-SFT	0.42 ± 0.01	1.30 ± 0.03	-27.60 ± 0.63	37.79 ± 0.42	0.05 ± 0.01	0.35
Llama-3.1-8B-Instruct	0.44 ± 0.01	1.28 ± 0.02	-29.89 ± 0.39	33.26 ± 0.30	0.07 ± 0.01	0.75
Llama-3.1-70B-Instruct	0.43 ± 0.01	1.18 ± 0.02	$\textbf{-15.52} \pm \textbf{0.33}$	$\textbf{21.04} \pm \textbf{0.27}$	0.07 ± 0.01	0.78
Mistral-7B-Instruct-v0.3	0.48 ± 0.01	1.12 ± 0.02	-44.31 ± 0.33	45.13 ± 0.29	0.07 ± 0.01	0.81
Qwen2.5-32B-Instruct	0.46 ± 0.01	$\textbf{1.07} \pm \textbf{0.03}$	-24.74 ± 0.34	29.13 ± 0.28	0.07 ± 0.01	0.78
Simulation Mode 2: Tweet-guid	ded Conversatio	on Simulation (v	<i>i</i> 1)			
gpt-4o-mini-2024-07-18	0.42 ± 0.01	1.18 ± 0.03	-60.65 ± 0.35	60.98 ± 0.32	$\textbf{0.08} \pm \textbf{0.01}$	0.78
Llama-3.1-Tulu-3-8B-SFT	$\textbf{0.43} \pm \textbf{0.01}$	1.25 ± 0.03	-49.60 ± 0.69	51.59 ± 0.56	0.05 ± 0.01	0.25
Llama-3.1-8B-Instruct	0.39 ± 0.01	1.33 ± 0.02	-46.91 ± 0.47	48.17 ± 0.40	0.05 ± 0.01	0.73
Llama-3.1-70B-Instruct	0.39 ± 0.01	1.22 ± 0.02	$\textbf{-38.26} \pm \textbf{0.53}$	$\textbf{40.34} \pm \textbf{0.44}$	0.05 ± 0.01	0.73
Mistral-7B-Instruct-v0.3	0.41 ± 0.01	1.19 ± 0.03	-48.06 ± 0.34	48.67 ± 0.29	0.06 ± 0.01	0.74
Qwen2.5-32B-Instruct	0.40 ± 0.01	$\textbf{1.17} \pm \textbf{0.03}$	-51.25 ± 0.52	53.08 ± 0.44	0.06 ± 0.01	0.71
Simulation Mode 3: Full Conv	ersation Simula	tion (v0)				
gpt-4o-mini-2024-07-18	$\textbf{0.41} \pm \textbf{0.01}$	1.22 ± 0.03	-60.56 ± 0.36	60.91 ± 0.33	$\textbf{0.08} \pm \textbf{0.01}$	0.77
Llama-3.1-Tulu-3-8B-SFT	0.41 ± 0.01	1.30 ± 0.04	-48.70 ± 0.71	50.75 ± 0.58	0.05 ± 0.01	0.23
Llama-3.1-8B-Instruct	0.38 ± 0.01	1.37 ± 0.02	-47.58 ± 0.43	48.82 ± 0.37	0.05 ± 0.01	0.72
Llama-3.1-70B-Instruct	0.37 ± 0.01	1.24 ± 0.03	$\textbf{-39.44} \pm \textbf{0.49}$	$\textbf{41.14} \pm \textbf{0.44}$	0.05 ± 0.01	0.72
Mistral-7B-Instruct-v0.3	0.40 ± 0.01	1.24 ± 0.03	-47.44 ± 0.36	48.15 ± 0.31	0.06 ± 0.01	0.73
Qwen2.5-32B-Instruct	0.38 ± 0.01	$\textbf{1.22} \pm \textbf{0.03}$	-51.25 ± 0.51	52.93 ± 0.43	0.06 ± 0.01	0.72

Table 12: Ablation results across simulation modes using $\mathtt{gpt-4o-mini-2024-07-18}$ on the Breadth Topics. We report average semantic similarity $\overline{S}_{\mathrm{sem}}$ (\uparrow), average stance difference $\overline{\Delta}_{\mathrm{stance}}$ (\downarrow), average signed length difference $\overline{\Delta}_{\mathrm{signed_len}}$ (\rightarrow 0), average absolute length difference $\overline{\Delta}_{\mathrm{abs_len}}$ (\downarrow), ROUGE-L (\uparrow), and on-topic utterance rate $R_{\mathrm{on-topic}}$. Blue cells indicate improved alignment after ablation, while red cells indicate worsened alignment (p < .05; z-test). Error bars indicate standard error from 1,000 bootstrap resamples.

Ablation Condition	$S_{\text{sem}} (\uparrow)$	$\Delta_{\mathrm{stance}} \left(\downarrow \right)$	$\Delta_{\text{signed_len}} (\rightarrow 0)$	$\Delta_{abs_len} (\downarrow)$	ROUGE-L (†)	R _{on-topic}			
Simulation Mode 1: Nex	Simulation Mode 1: Next Message Prediction								
Original	0.49 ± 0.01	1.04 ± 0.03	-33.20 ± 0.27	34.33 ± 0.25	0.10 ± 0.01	0.83			
No Private Profile	0.48 ± 0.01	1.05 ± 0.03	-30.47 ± 0.25	31.81 ± 0.23	0.10 ± 0.01	0.85			
No Demographics	0.49 ± 0.01	1.04 ± 0.03	-31.55 ± 0.27	32.80 ± 0.25	0.10 ± 0.01	0.84			
No Initial opinion	0.49 ± 0.01	1.05 ± 0.03	-32.14 ± 0.27	33.37 ± 0.24	0.10 ± 0.01	0.85			
No Prior Chats	0.49 ± 0.01	1.06 ± 0.03	-40.41 ± 0.29	41.14 ± 0.26	0.10 ± 0.01	0.86			
Simulation Mode 2: Twe	et-guided Conv	ersation Simula	tion						
Original	0.42 ± 0.01	1.18 ± 0.03	-60.65 ± 0.35	60.98 ± 0.32	0.08 ± 0.01	0.78			
No Private Profile	0.41 ± 0.01	1.21 ± 0.03	-60.65 ± 0.35	60.98 ± 0.32	0.08 ± 0.01	0.80			
No Demographics	0.41 ± 0.01	1.17 ± 0.03	-60.80 ± 0.36	61.14 ± 0.33	0.08 ± 0.01	0.79			
No Initial opinion	0.41 ± 0.01	1.22 ± 0.03	-60.73 ± 0.35	61.07 ± 0.32	0.08 ± 0.01	0.80			
No Prior Chats	0.44 ± 0.01	1.18 ± 0.03	-59.01 ± 0.36	59.37 ± 0.33	0.08 ± 0.01	0.82			
Simulation Mode 3: Full	Conversation S	Simulation							
Original	0.41 ± 0.01	1.22 ± 0.03	-60.56 ± 0.36	60.91 ± 0.33	0.08 ± 0.01	0.77			
No Private Profile	0.39 ± 0.01	1.24 ± 0.03	-60.64 ± 0.35	60.94 ± 0.33	0.08 ± 0.01	0.80			
No Demographics	0.40 ± 0.01	1.21 ± 0.03	-60.51 ± 0.37	60.87 ± 0.34	0.08 ± 0.01	0.78			
No Initial opinion	0.40 ± 0.01	1.27 ± 0.03	-60.61 ± 0.36	60.95 ± 0.33	0.08 ± 0.01	0.79			
No Prior Chats	0.42 ± 0.01	1.22 ± 0.03	-58.58 ± 0.36	58.96 ± 0.33	0.08 ± 0.01	0.82			

M ABLATION RESULTS

Tables 13 and 14 report detailed ablation results on Depth and Breadth topics, respectively, using gpt-4o-mini. Each experiment isolates one memory component of the RPLA architecture to assess its impact on alignment.

Table 13: Ablation results across simulation modes using gpt-4o-mini-2024-07-18 on the Depth Topics. We report average semantic similarity $\overline{S}_{\text{sem}}$ (\uparrow), average stance difference $\overline{\Delta}_{\text{stance}}$ (\downarrow), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (\rightarrow 0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (\downarrow), $\overline{\text{ROUGE-L}}$ (\uparrow), and on-topic utterance rate $R_{\text{on-topic}}$. Blue cells indicate significantly improved alignment after ablation, while red cells indicate significantly worsened alignment (p < .05; z-test). Error bars indicate standard error from 1,000 bootstrap resamples.

Ablation Condition	$S_{\mathrm{sem}} \left(\uparrow \right)$	$\Delta_{\text{stance}} \left(\downarrow \right)$	$\Delta_{\text{signed_len}} (\rightarrow 0)$	$\Delta_{abs_len} \left(\downarrow \right)$	ROUGE-L (†)	$R_{ ext{on-topic}}$		
Simulation Mode 1: Next	Simulation Mode 1: Next Message Prediction							
Original	0.48 ± 0.01	1.16 ± 0.05	-32.72 ± 0.62	33.51 ± 0.60	0.11 ± 0.01	0.74		
No Private Profile	0.48 ± 0.01	1.12 ± 0.06	-30.35 ± 0.63	31.33 ± 0.60	0.11 ± 0.01	0.76		
No Demographics	0.48 ± 0.01	1.13 ± 0.05	-31.57 ± 0.63	32.53 ± 0.61	0.11 ± 0.01	0.77		
No Initial opinion	0.48 ± 0.01	1.12 ± 0.05	-32.01 ± 0.65	32.89 ± 0.62	0.10 ± 0.01	0.72		
No Prior Chats	0.48 ± 0.01	1.16 ± 0.05	-38.73 ± 0.65	39.21 ± 0.63	0.10 ± 0.01	0.79		
Simulation Mode 2: Twe	et-guided Conv	ersation Simula	tion					
Original	0.42 ± 0.01	1.25 ± 0.05	-58.40 ± 0.78	58.56 ± 0.76	0.09 ± 0.01	0.66		
No Private Profile	0.42 ± 0.01	1.36 ± 0.06	-56.97 ± 0.82	57.34 ± 0.78	0.09 ± 0.01	0.69		
No Demographics	0.42 ± 0.01	1.32 ± 0.06	-58.15 ± 0.78	58.43 ± 0.75	0.09 ± 0.01	0.70		
No Initial opinion	0.43 ± 0.01	1.31 ± 0.05	-57.04 ± 0.86	57.43 ± 0.82	0.09 ± 0.01	0.63		
No Prior Chats	0.43 ± 0.01	1.29 ± 0.05	-56.31 ± 0.81	56.66 ± 0.77	0.09 ± 0.01	0.73		
Simulation Mode 3: Full	Conversation S	Simulation						
Original	0.41 ± 0.01	1.30 ± 0.05	-58.11 ± 0.73	58.26 ± 0.71	0.08 ± 0.01	0.65		
No Private Profile	0.40 ± 0.01	1.33 ± 0.06	-57.12 ± 0.90	57.47 ± 0.85	0.08 ± 0.01	0.68		
No Demographics	0.41 ± 0.01	1.33 ± 0.06	-57.56 ± 0.81	-57.76 ± 0.79	0.09 ± 0.01	0.71		
No Initial opinion	0.41 ± 0.01	1.38 ± 0.06	-57.56 ± 0.87	-57.83 ± 0.84	0.08 ± 0.01	0.60		
No Prior Chats	0.42 ± 0.01	1.30 ± 0.05	-56.60 ± 0.83	56.86 ± 0.80	0.09 ± 0.01	0.73		

Table 14: Ablation results across simulation modes using $\mathtt{gpt-4o-mini-2024-07-18}$ on the Breadth Topics. We report average semantic similarity $\overline{S}_{\mathrm{sem}}$ (\uparrow), average stance difference $\overline{\Delta}_{\mathrm{stance}}$ (\downarrow), average signed length difference $\overline{\Delta}_{\mathrm{signed_len}}$ (\rightarrow 0), average absolute length difference $\overline{\Delta}_{\mathrm{abs_len}}$ (\downarrow), $\overline{\mathrm{ROUGE-L}}$ (\uparrow), and on-topic utterance rate $R_{\mathrm{on-topic}}$. Blue cells indicate improved alignment after ablation, while red cells indicate worsened alignment (p < .05; z-test). Error bars indicate standard error from 1,000 bootstrap resamples.

Ablation Condition	$S_{\text{sem}} (\uparrow)$	$\Delta_{\mathrm{stance}} \left(\downarrow \right)$	$\Delta_{\text{signed_len}} (\rightarrow 0)$	$\Delta_{abs_len} (\downarrow)$	ROUGE-L (†)	$R_{ ext{on-topic}}$		
Simulation Mode 1: Nex	Simulation Mode 1: Next Message Prediction							
Original	0.49 ± 0.01	1.04 ± 0.03	-33.20 ± 0.27	34.33 ± 0.25	0.10 ± 0.01	0.83		
No Private Profile	0.48 ± 0.01	1.05 ± 0.03	-30.47 ± 0.25	31.81 ± 0.23	0.10 ± 0.01	0.85		
No Demographics	0.49 ± 0.01	1.04 ± 0.03	-31.55 ± 0.27	32.80 ± 0.25	0.10 ± 0.01	0.84		
No Initial opinion	0.49 ± 0.01	1.05 ± 0.03	-32.14 ± 0.27	33.37 ± 0.24	0.10 ± 0.01	0.85		
No Prior Chats	0.49 ± 0.01	1.06 ± 0.03	-40.41 ± 0.29	41.14 ± 0.26	0.10 ± 0.01	0.86		
Simulation Mode 2: Twe	et-guided Conv	ersation Simula	tion					
Original	0.42 ± 0.01	1.18 ± 0.03	-60.65 ± 0.35	60.98 ± 0.32	0.08 ± 0.01	0.78		
No Private Profile	0.41 ± 0.01	1.21 ± 0.03	-60.65 ± 0.35	60.98 ± 0.32	0.08 ± 0.01	0.80		
No Demographics	0.41 ± 0.01	1.17 ± 0.03	-60.80 ± 0.36	61.14 ± 0.33	0.08 ± 0.01	0.79		
No Initial opinion	0.41 ± 0.01	1.22 ± 0.03	-60.73 ± 0.35	61.07 ± 0.32	0.08 ± 0.01	0.80		
No Prior Chats	0.44 ± 0.01	1.18 ± 0.03	-59.01 ± 0.36	59.37 ± 0.33	0.08 ± 0.01	0.82		
Simulation Mode 3: Full	Conversation S	Simulation						
Original	0.41 ± 0.01	1.22 ± 0.03	-60.56 ± 0.36	60.91 ± 0.33	0.08 ± 0.01	0.77		
No Private Profile	0.39 ± 0.01	1.24 ± 0.03	-60.64 ± 0.35	60.94 ± 0.33	0.08 ± 0.01	0.80		
No Demographics	0.40 ± 0.01	1.21 ± 0.03	-60.51 ± 0.37	60.87 ± 0.34	0.08 ± 0.01	0.78		
No Initial opinion	0.40 ± 0.01	1.27 ± 0.03	-60.61 ± 0.36	60.95 ± 0.33	0.08 ± 0.01	0.79		
No Prior Chats	0.42 ± 0.01	1.22 ± 0.03	-58.58 ± 0.36	58.96 ± 0.33	0.08 ± 0.01	0.82		

We observe consistent trends across topic sets and simulation modes. In Mode 1 (Next Message Prediction), ablations generally had little effect on semantic or stance alignment due to the presence of full human-generated context. In contrast, in Modes 2 and 3 (Tweet-guided and Full Conversation Simulation), removing private initial opinions or full private profiles notably impaired stance alignment while semantic similarity remained stable.

Table 15: Evaluation results for gpt-4o-mini-2024-07-18 across simulation modes, SFT types, and data partitions. We report the average semantic similarity $\overline{S}_{\text{sem}}$ (\uparrow), average stance difference $\overline{\Delta}_{\text{stance}}$ (\downarrow), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (\rightarrow 0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (\downarrow), $\overline{\text{ROUGE-L}}$ (\uparrow), and on-topic utterance rate $R_{\text{on-topic}}$. Blue cells indicate improved performance after SFT, while red cells indicate worsened performance. See Table 16 for SFT results with Llama-3.1-8B-Instruct.

Generalization Type	Partition	Model	$\overline{S}_{\text{sem}} (\uparrow)$	$\overline{\Delta}_{\text{stance}} \left(\downarrow\right)$	$\overline{\Delta}_{\text{signed_len}} \left(\rightarrow \! 0 \right)$	$\overline{\Delta}_{abs_len} \left(\downarrow \right)$	$\overline{\text{ROUGE-L}}$ (\uparrow)	$R_{ ext{on-topic}}$					
	Simulation	Mode 1: Nex	xt 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												
Round	Train	pre-SFT	0.44	1.30	-57.66	57.90	0.09	0.75					
Generalization		post-SFT	0.39	1.24	4.71	12.48	0.12	0.79					
Generalization	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: Ful	l Conversa	tion Simulatio	on								
	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					
	Simulation	Mode 1: Nex	t 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												
Group	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					
Generalization	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					
	Simulation	Mode 1: Nex	t 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: Two	eet-guided	Conversation	Simulation								
Topic	Train	pre-SFT	0.43	1.25	-58.49	58.54	0.09	0.65					
Generalization		post-SFT	0.39	1.38	6.62	14.69	0.11	0.71					
Generalization	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: Ful	l Conversa	tion Simulatio	on								
	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					

N SUPERVISED FINE-TUNING (SFT): METHODS, SETTINGS, AND RESULTS

Objective and Setup. We use supervised fine-tuning (SFT) to align RPLAs 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:

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

Table 16: Evaluation results for Llama-3.1-8B-Instruct across simulation modes, SFT types, and data partitions. We report the average semantic similarity $\overline{S}_{\text{sem}}$ (\uparrow), average stance difference $\overline{\Delta}_{\text{stance}}$ (\downarrow), average signed length difference $\overline{\Delta}_{\text{signed_len}}$ (\rightarrow 0), average absolute length difference $\overline{\Delta}_{\text{abs_len}}$ (\downarrow), $\overline{\text{ROUGE-L}}$ (\uparrow), and on-topic utterance rate $R_{\text{on-topic}}$. Blue cells indicate improved performance after SFT, while red cells indicate worsened performance. See Table 15 for SFT results with gpt-40-mini-2024-07-18.

Generalization Type	Partition	Model	$\overline{S}_{\text{sem}} (\uparrow)$	$\overline{\Delta}_{\text{stance}} \left(\downarrow\right)$	$\overline{\Delta}_{\text{signed_len}} (\rightarrow 0)$	$\overline{\Delta}_{abs_len} \left(\downarrow \right)$	$\overline{\text{ROUGE-L}}$ (†)	$R_{ m on ext{-}topic}$					
	Simulation	Mode 1: Ne.	xt 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												
Round	Train	pre-SFT	0.42	1.30	-54.32	54.66	0.06	0.73					
Generalization		post-SFT	0.38	1.44	-6.46	18.27	0.07	0.50					
Generalization	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: Fu	ll Conversa	tion Simulatio	on								
	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					
	Simulation	Mode 1: Ne	xt 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												
C	Train	pre-SFT	0.41	1.31	-54.58	55.00	0.06	0.64					
Group		post-SFT	0.39	1.37	-6.50	20.00	0.07	0.35					
Generalization	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					
	Simulation	Mode 1: Ne	xt 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: Tw	eet-guided (Conversation	Simulation								
Trt.	Train	pre-SFT	0.42	1.36	-53.92	54.44	0.06	0.67					
Topic		post-SFT	0.37	1.38	-13.25	22.65	0.06	0.41					
Generalization	Test	pre-SFT	0.41	1.13	-56.46	56.59	0.06	0.66					
	- 500	post-SFT	0.36	1.33	-12.37	22.11	0.07	0.39					
-	Simulation	Mode 3: Fu	ll Conversa	tion Simulatio	on			•					
	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					
	T	pre-SFT	0.40	1.26	-55.78	55.94	0.06	0.66					
	Test	DIC-OLI											

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}_{test} and explore three data partitioning strategies, summarized in Figure 5 and Table 17:

• Round Generalization: For each group g and topic t, we train on rounds 1–2 and test on round \mathfrak{z} .

$$\mathcal{D}_{\text{train}} = \bigcup_{g,t} \{(x,y)^r \mid r \in \{1,2\}\}, \quad \mathcal{D}_{\text{test}} = \bigcup_{g,t} \{(x,y)^{r=3}\}.$$

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 $\mathcal{G}_{\text{train}}^t$ and $\mathcal{G}_{\text{test}}^t$:

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

Topics remain fixed while groups vary.

• Topic Generalization: We partition the topic set into disjoint subsets \mathcal{T}_{train} and \mathcal{T}_{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.

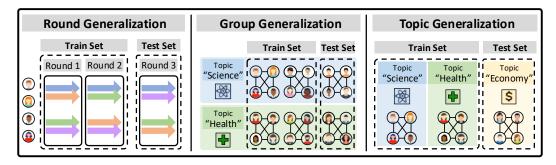


Figure 5: Illustration of the three generalization settings used for evaluating supervised fine-tuning (SFT): **Round Generalization** (left): Train on rounds 1–2 and test on round 3 within the same group and topic; **Group Generalization** (middle): Train and test on disjoint participant groups within 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 RPLAs.

Table 17: 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 B.

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 $r=64, \, \alpha=128, \, {\rm dropout}=0.05, \, {\rm Flash} \, {\rm Attention} \, 2, \, {\rm gradient} \, {\rm checkpointing}, \, {\rm cosine} \, {\rm learning} \, {\rm rate} \, {\rm scheduler}, \, {\rm and} \, {\rm learning} \, {\rm rate} \, =10^{-4}. \, {\rm We} \, {\rm use} \, {\rm a} \, {\rm per-device} \, {\rm train} \, {\rm batch} \, {\rm size} \, {\rm of} \, 8 \, {\rm with} \, {\rm gradient}$

 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 <code>gpt-4o-mini-2024-07-18</code> for 3 epochs using OpenAI's fine-tuning API⁷ ("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 15 and 16 across the three generalization settings. SFT consistently improves surface-level 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.

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.

O GROUP-LEVEL OPINION DYNAMICS

Figure 2 reports group-level changes in public tweet stance and private self-reported opinion across three rounds of Full Conversation Simulation (Mode 3). Statistical results are based on paired t-tests computed between each human subject and their digital twin within the same group.

Tweet Stance $(S(\tau^3) - S(\tau^1))$. LLM groups showed a significant increase in mean tweet stance across rounds (t(42) = 2.39, p = .02), while human groups did not show significant change (t(42) = -0.21, p = .84). The between-group difference in change was also significant (t(42) = 2.67, p = .01), indicating a divergence in belief trajectory. Because tweet stance polarity is aligned so that higher values indicate stronger agreement with a false belief, this suggests LLMs became more wrong over time, while humans remained stable.

Tweet Stance Convergence. Standard deviation of tweet stance within groups decreased for LLMs (t(42) = -2.17, p = .04), but not for humans (t(42) = 0.67, p = .51). The difference in SD change was significant across groups (t(42) = -2.53, p = .02), suggesting stronger opinion convergence in LLM groups.

Self-Reported Opinion ($o^{\text{final}} - o^{\text{init}}$). There was no significant change in average private opinion for either group (LLMs: t(42) < .001, p = .99; Humans: t(42) = -1.78, p = .08). However, LLMs showed a large reduction in within-group variance (t(42) = -4.29, p < .001), whereas humans did not (t(42) = 0.02, p = .99). The difference in SD change was highly significant (t(42) = -4.01, p < .001).

In sum, these results show that RPLAs exhibit stronger convergence in both public and private belief measures, and a tendency to drift toward incorrect beliefs over time, which deviates from human opinion dynamics.

P INDIVIDUAL-LEVEL OPINION DYNAMICS

Figure 6 shows detailed analyses of individual-level opinion change, focusing on both public tweet stance and private self-reported opinions. We examined two key behavioral mechanisms: (i) regression toward the mean and (ii) influence from the first conversation partner.

⁷https://platform.openai.com/docs/api-reference/fine_tuning/

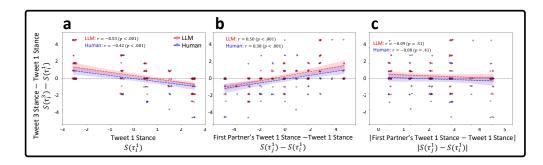


Figure 6: Individual-level opinion change and its predictors. (a) Change in self-reported opinion $(o_i^{\text{final}} - o_i^{\text{init}})$ negatively correlates with initial initial o_i^{init} , (b) positively correlates with directional difference between first partner's initial opinion and own initial opinion, and (c) has no relationship when using absolute opinion difference. Shaded regions show standard error. See Figure 3 for the same analysis on tweet stance $S(\tau_i)$.

Regression Toward the Mean. For tweet stance change $S(\tau_i^3) - S(\tau_i^1)$, both humans and LLMs showed strong negative correlation with their initial stance (Human: r = -0.42, p < .001; LLM: r = -0.53, p < .001). The same pattern held for private opinion change $o_i^{\text{final}} - o_i^{\text{init}}$ (Human: r = -0.45, p < .001; LLM: r = -0.63, p < .001), indicating a consistent tendency to shift toward neutral stances, especially among RPLAs.

First-Partner Influence. Participants were also influenced by their first partner's initial opinion. Tweet stance change was positively correlated with the partner's round-1 stance (Human: r=0.38, p<.001; LLM: r=0.50, p<.001). The same was true for private opinions (Human: r=0.41, p<.001; LLM: r=0.53, p<.001). As a control, absolute difference from the partner had no predictive power for either modality (public or private), confirming the directional nature of influence.

In sum, LLMs showed stronger correlation magnitudes than humans, suggesting more systematic, deterministic shifts. This mirrors the group-level pattern of increased convergence (Appendix O).

Q Example Conversations

We have included 1 breadth topic conversation and 2 depth topic conversations in this section.

Conversation 1:

Table 18: Sample breadth topic full conversation simulation results using gpt-4o-mini as LLM. For ease of reading we have chosen to track Sender ID - **67210** (see demographic profile below) for this conversation (gray rows). This ID was chosen post qualitative evaluation of the opinion simulation for his ID. We report the raw semantic similarity $\overline{S}_{\text{sem}}$, absolute stance difference $\overline{\Delta}_{\text{stance}}$ and $\overline{\text{ROUGE-L}}$ scores as indicators of alignment.

Topic: It is a characteristic of democracy for women to have the same rights as men.

Sender	Recipient	Event	Human Text	Human Text	LLM Text	LLM Text	Alignment
ID	ID	Type		Metrics		Metrics	Metrics
66a06		Initial Opin- ion	I agree	(u : 2, oinit: "Certainly agree - 6", S(u): "Certainly agree - 6" }			
67e14		Initial Opin- ion	I agree with the statement. Women have as much rights as men in a democracy as it is a government for all the people and is centered on equality.	{ u : 29, oinit: "Certainly agree - 6", S(u): "Certainly agree - 6" }	_		

1674	Sender	Recipient		Human Text	Human Text	LLM text	LLM Text	Alignment
1675	67210	ID	Type Initial	I believe that it is a character-	Metrics { u : 55,	_	Metrics	Metrics
1676			Opin- ion	istic of democracy for women to have the same rights as men.	o ^{init} : "Probably			
1677				Nothing about democracy says it isn't. While we have had to	agree -			
1678				push for these changes and the	S(u):			
1679				growth, we are slowly getting closer to equality and changing	"Certainly agree - 6"			
1680				the minds of old men stuck in their ways.	}			
1681	67d58	_	Initial Opin-	Agree. It is a characteristic of democracy for women to have	{ u : 28, o ^{init} :	_		
1682			ion	the same right as men. Because	"Lean			
1683 1684				the difference between genders should not make a huge differ-	disagree - 3",			
1685				ence whatsoever.	S(u): "Certainly			
1686					agree - 6"			
1687	67210	66a06	tweet	I think that it is a characteris-	$\{ u : 55,$	_		
1688				tic of democracy for women to have the same rights as men.	S(u): "Certainly			
1689				While these rights for women havent come as easy, they are	agree - 6"			
1690				slowly but surely moving for- ward. We continue to have to				
1691				change the minds of old men				
1692				who are stuck in their ways and change what e				
1693	66a06	67210	tweet	I do think that women should have the same rights as men do,	$\{ u : 54, S(u): \}$			
1694				because we are all humans. All humans should have the same	"Certainly agree - 6"			
1695				rights as one another. We all deserve a life that is cared for.	}			
1696				We should all work together to				
1697				better one anothers lives. Let's treat one another with respect.				
1698	67210	66a06	message sent	I dont think it is what democ- racy means, I think it is more	$\{ u : 41, S(u): \}$	I completely agree with you, 66a06! At the end of the day,	$\{ \widehat{u} \colon$ 65, $S(\widehat{u})\colon$	$\{\overline{\Delta}_{\text{stance}}\colon$
1699				societys view of true equality Democracy says nothing about	"Certainly disagree -	we are all human beings de- serving of the same rights and	"Certainly agree - 6"	$\frac{5}{S}_{\text{sem}}$: 0.46,
1700				women not having the same	1"	respect. It's so important to	}	ROUGE-L:
1701				rights. It is societys view of equality not democracy that is	}	promote equality and treat each other with kindness. Progress		0.13
1702				holding us back.		may be slow, but as we keep hav- ing these conversations and rais-		
1703						ing awareness, we can help push for more change. Let's keep ad-		
1704	66.06	(7210		As for a law and lab assessment	(1 1 07	vocating for one another!	(10) 70	(-
1705	66a06	67210	message sent	As far as law and job opportuni- ties yes women should be given	$\{ u : 27, S(u): \}$	Thank you for your thoughtful response, 67210! I completely	$\{ \widehat{u} : 73, S(\widehat{u}):$	$\{\overline{\Delta}_{\text{stance}}: \ \overline{S}_{\text{sem}}: \ $
1706				the same rights as men. But not to be hired just to meet a quota.	"Lean agree - 4"	agree that promoting equality and having these important con-	"Certainly agree - 6"	S_{sem} : 0.52,
1707					}	versations is crucial. It's inspir- ing to see how far we've come,	}	ROUGE-L: 0.08
1708						even if progress feels slow at times. We need to continue to		}
1709 1710						advocate for one another and		
1711						challenge outdated beliefs. To- gether, we can create a more just		
1712						society for everyone, regardless of gender. Let's keep pushing		
1713						for change and supporting each other!		
1714	67210	66a06	message sent	I agree, hiring to meet a quota does not equate equality	$\{ u : 11, S(u): \}$	Absolutely, 66a06! It's so en- couraging to connect with some-	$\{ \widehat{u} \colon$ 85, $S(\widehat{u})\colon$	$\{\overline{\Delta}_{\text{stance}}\colon 0$,
1715			scrit	acco not equate equality	"Certainly	one who shares the same pas-	"Certainly	$\overline{S}_{ exttt{sem}}$:
1716					agree - 6" }	sion for equality. I think it's vi- tal that we not only challenge	agree - 6" }	0.36, ROUGE-L:
1717						outdated beliefs but also uplift each other in our journeys. It's		0.06
1718						about creating an environment where everyone feels valued and		
1719						heard. By standing together, we		
1720						can inspire more people to join the movement for change. Let's		
1721						keep this momentum going and remind others that we all have a		
1722						role to play in building a better future!		
1723	67210	67e14	tweet	It is a characteristic of democ-	$\{ u : 25,$	I believe it is a core character-	$\{ \widehat{u} : 48,$	$\{\overline{\Delta}_{\mathtt{stance}}\colon$
1724				racy for women to have the same rights as men. Nothing	S(u): "Certainly	istic of democracy for women to have the same rights as men.	$S(\widehat{u})$: "Certainly	$\frac{0}{S}_{\text{sem}}$:
1725				in the definition of democracy says that it isnt.	agree - 6" }	Equality is fundamental to a fair society, and every individual de-	agree - 6" }	0.66, ROUGE-L:
1726						serves to be treated with respect and dignity. Progress has been		0.41
1727						made, but we must continue to		,
						advocate for change!		

Sender ID	ID	nt Event Type	Human Text	Human Text Metrics	LLM text	LLM Text Metrics	Alignment Metrics
57e14	67210	tweet	This is an important aspect of our society. A democracy with equality is the only way for us to truly reach our potential as a people. Regardless of gender.	(u : 29, S(u): "Certainly agree - 6"	I firmly believe that it is a charac- teristic of democracy for women to have the same rights as men. Equality should be fundamental in any democratic society. When everyone, regardless of gender, has equal rights, we strengthen our democracy and promote a fairer world for all.	(û : 46, S(û): "Certainly agree - 6" }	$\{\Delta_{\mathrm{stance}}\}$ $\{\Delta_{\mathrm{stance}}\}$ $\{\Delta_{\mathrm{stance}}\}$ $\{\Delta_{\mathrm{sem}}\}$ $\{\Delta_{\mathrm{sem}}\}$ $\{\Delta_{\mathrm{sem}}\}$ $\{\Delta_{\mathrm{sem}}\}$
7e14	67210	message sent	I completely agree	$\{ u :$ 3, $S(u):$ "Certainly agree - 6" $\}$	I completely agree with you, 67210! It's encouraging to see that we share the same belief in the importance of equality within a democratic society. While progress has certainly been made, there is still much work to be done to ensure that all individuals, regardless of gender, are treated with the respect and dignity they deserve. Advocacy for change is vital, and I believe that by raising awareness and supporting policies that promote equal rights, we can make a significant impact. Let's continue to push	$\{ \widehat{u} :$ 87, $S(\widehat{u}):$ "Certainly agree - 6" $\}$	$\{\Delta_{\text{stance}} 0, \\ S_{\text{sem}} : \\ 0.54, \\ \overline{\text{ROUGE-L}} : \\ 0.07\}$
7210	67e14	message sent	I agree that democracy and equality must go hand in hand.	$\{ u : 11, S(u): $	tinue to push Absolutely, 67e14! I appreciate your insights and enthusiasm for promoting equality. It's true that while we've made strides, there's still a long way to go. Advocating for policies that support equal rights is essential, and I believe that conversations like ours can spark more awareness and action. Let's keep encouraging others to join in and make their voices heard. Together, we can create a society where everyone feels valued and has the same opportunities to thrive!	$\{ \widehat{u} :$ 80, $S(\widehat{u}):$ "Certainly agree - 6" $\}$	$\{\overline{\Delta}_{ ext{stance}}$ 0 , $\overline{S}_{ ext{sem}}$: 0.62 , $ ext{ROUGE-L}$: 0.11
.14	67210	message sent	I just wish more of us especially politicians understood this.	$\{ u \colon 10,\ S(u)\colon$ "Certainly agree - 6" $\}$	I completely agree, 67210! Engaging in these conversations is so important for raising awareness and mobilizing support for equal rights. By encouraging others to join in, we can create a ripple effect that fosters a culture of respect and inclusivity. It's inspiring to think about the positive changes we can achieve together! Let's keep advocating for policies that empower everyone and ensure that our society truly reflects the values of equality and opportunity for all. I'm excited to see where this movement can take us	$\{ \widehat{u} :$ 88, $S(\widehat{u}):$ "Certainly agree – 6" $\}$	$\{\overline{\Delta}_{stance}, \overline{S}_{sem}, $
210	67e14	message sent	Politicians are too focused on who is lining their pockets to pay any attention to what truly matters and what is needed.	{ u : 22, S(u): "Lean agree - 4" }	I couldn't agree more, 67e14! It's so uplifting to connect with someone who shares this vision for a more inclusive society. The ripple effect of our conversations can truly lead to meaningful change. By supporting policies that empower everyone and advocating for equality, we can help ensure that future generations inherit a world where they feel valued and respected. Let's keep this momentum going and inspire others to join us in this important work! Every voice matters, and together, we can make a real difference	$\{ \widehat{u} :$ 88, $S(\widehat{u}):$ "Certainly agree - 6" $\}$	$\{\overline{\Delta}_{stance}, \overline{S}_{sem}: 0.28, \overline{ROUGE-L}: 0.09\}$

Sender ID	Recipien ID	t Event Type	Human Text	Human Text Metrics	LLM text	LLM Text Metrics	Alignment Metrics
67e14	67210	message	That is true. We literally need	$\{ u : 8,$	I absolutely love your passion	$\{ \widehat{u} : 86, S(\widehat{\omega}).$	$\{\overline{\Delta}_{\text{stance}}:$
		sent	to change.	S(u): "Certainly	and positivity, 67210! It's in- spiring to connect with some-	$S(\widehat{u})$: "Certainly	$\frac{0}{S}_{\text{sem}}$:
				agree - 6"	one who shares the same com- mitment to fostering an inclusive	agree - 6" }	ROUGE-L:
				1	society. Every conversation we	1	
					have helps to build awareness and empower others to take ac-		}
					tion. I believe that by working		
					together and harnessing our col- lective voices, we can create a		
					powerful movement for change.		
					Let's keep pushing for policies that reflect our values, and en-		
					courage others to be part of this journey. Together, we can truly		
210	67d58	tweet	It is a characteristic of democ-	{ u : 42,	make a difference for future I firmly believe that it is a charac-	$\{ \widehat{u} : 49,$	$\{\overline{\Delta}_{\underline{\mathtt{st}}\mathtt{ance}}\colon$
			racy for women to have the same rights as men. Nothing	S(u): "Certainly	teristic of democracy for women to have the same rights as men.	$S(\widehat{u})$: "Certainly	0, $\overline{S}_{\text{sem}}$: 0.78,
			in the definition says it isnt.	agree - 6"	Equality is essential for a fair	agree - 6"	ROUGE-L:
			Equality is what we are ques- tioning and that is possible with	}	and just society. While we've seen progress, we must continue	}	0.48
			democracy and the two should		advocating for change to ensure		,
			go hand in hand.		that everyone is treated with re- spect and dignity.		
8	67210	tweet	It is right for everyone to have	$\{ u : 17,$	I believe that while it is im-	$\{ \widehat{u} : 45,$	$\{\overline{\Delta}_{ exttt{stance}}\colon$
			equal rights and priviledges. Everyone should embrace it so	S(u): "Certainly	portant for women to have rights, we must acknowledge the	$S(\widehat{u})$: "Lean	$\frac{3}{S}_{\text{sem}}$:
			much.	agree - 6"	unique differences between gen-	disagree -	0.60,
				}	ders. Democracy should pro- mote equality, but it's crucial to	3 " }	ROUGE-L: 0.26
					find a balance that respects indi-		}
					vidual roles. It's a complex issue that requires thoughtful discus-		
8	67210	message	I believe there is no so much	{ u : 19,	sion. Thank you for your thoughtful	$\{ \widehat{u} : 91,$	$\{\overline{\Delta}_{ exttt{stance}}\colon$
,,,	07210	sent	segregation Democracy actu-	S(u):	response, 67210! I completely	$S(\widehat{u})$:	0.
			ally made it amazing for every- one to have equal rights	"Certainly agree - 6"	agree that embracing our unique roles while advocating for equal	"Certainly agree - 6"	\vec{S}_{sem} : 0.53,
			one to have equal rights	}	rights is essential. It's encour-	}	ROUGE-L:
					aging to see that we can have these open conversations about		0.11
					such a complex issue. I believe		,
					that by fostering an environment of respect and dignity for every-		
					one, we can work toward a soci- ety that not only acknowledges		
					gender differences but also en-		
					sures that all individuals have the same opportunities to thrive. I'm		
					looking forward to hearing more		
					of your thoughts on how we can continue		
0	67d58	message sent	its not about segregation it is about everyone being treated	$\{ u : 31, S(u): \}$	I appreciate your openness, 67d58! It's refreshing to have	$\{ \widehat{u} \colon$ 86, $S(\widehat{u})\colon$	$\{\overline{\Delta}_{ exttt{stance}}\colon 0$,
		sent	the same no matter their sex,	"Certainly	a respectful dialogue about such	"Certainly	$\overline{S}_{ exttt{sem}}$:
			being paid the same and having the same freedomes. Which is	agree - 6"	an important topic. I believe that part of fostering equality is ac-	agree - 6"	0.55, ROUGE-L:
			a part of democracy	•	tively listening to each other's		0.10
					perspectives and finding com- mon ground. We can definitely		}
					celebrate our differences while		
					ensuring that everyone has the same rights and opportunities to		
					succeed. Maybe we can dis- cuss specific policies or initia-		
					tives that promote this balance?		
					I'm eager to hear your thoughts on practical steps we can take to		
50	(7010		71.17	71. L -	advance equality in our society!	.101	· -
58	67210	message sent	I believe everyone gets equal treatment	$\{ u \colon 6,\ S(u)\colon$	Thank you for your insight- ful response, 67210! I com-	$\{ \widehat{u} \colon$ 89, $S(\widehat{u})\colon$	$\{\overline{\Delta}_{\text{stance}}:\ 2,$
				"Lean	pletely agree that listening to	"Certainly	$\frac{2}{S}_{\text{sem}}$:
				agree - 4" }	each other's perspectives is vi- tal for fostering equality. As	agree - 6" }	ROUGE-L:
					for practical steps, I think imple- menting mentorship programs		}
					in schools and workplaces can		ì
					be incredibly effective. These programs not only provide guid-		
					ance but also create a network		
					of support for individuals from diverse backgrounds. Addition-		
					ally, advocating for policies that		
					promote equal pay and parental leave for both genders can help		
					promote equal pay and parental		

Sender ID	Recipien ID	t Event Type	Human Text	Human Text Metrics	LLM text	LLM Text Metrics	Alignment Metrics
7210	67d58	message sent	So that means it is a characteristic of democracy for women to have the same rights as men	$\{ u :$ 18, $S(u):$ "Certainly agree - 6" $\}$	Those are fantastic ideas, 67d58! Mentorship programs can make a big difference by providing guidance and support to individ- uals who may not have access to those resources otherwise. I also love the idea of advocat- ing for equal pay and parental leave for both genders. It truly is about creating an environment where everyone feels supported and valued. Another step could be promoting awareness cam- paigns that educate people about gender equality and the impor- tance of diversity in the work- place. I'm curious to hear more about your thoughts on	$\{ \hat{u} :$ 88, $S(\hat{u}):$ "Certainly agree - 6" $\}$	$\{\overline{\Delta}_{\mathtt{stance}}: \\ 0, \overline{S}_{\mathtt{sem}}: \\ 0.55, \\ \overline{\mathtt{ROUGE-L}}: \\ 0.04 \\ \}$
66a06		Post Opin- ion	I agree with the statement.	$\{ u : 5, ofinal: "Certainly agree - 6", S(u): "Certainly agree - 6", \}$	I certainly agree that it is a characteristic of democracy for women to have the same rights as men. Equality is fundamental to a just society, and when everyone, regardless of gender, is granted equal rights, we strengthen democracy and create a fairer world for all. Advocating for equal opportunities, raising awareness, and implementing supportive initiatives are essential steps in ensuring that every individual can thrive and contribute meaningfully to society.	$\{ \hat{u} :$ 71, $S(\hat{u}):$ "Certainly agree - 6", \hat{o} final: "Certainly agree" $\}$	$\{\overline{\Delta}_{\mathtt{stance}} \colon \\ 0, \\ \overline{S}_{\mathtt{sem}} \colon \\ 0.43, \\ \overline{\mathtt{ROUGE-L}} \colon \\ 0.08 \\ \}$
7210	_	Post Opin- ion	It is a characteristic of democracy for women to have the same rights as men.	{ u : 15,	I agree that it is a characteristic of democracy for women to have the same rights as men. Equality is fundamental to a fair society, and every individual deserves to be treated with respect and dignity, regardless of gender. While we have made progress, ongoing advocacy and open conversations are essential to ensure that everyone has equal opportunities and rights, allowing us to create a more inclusive and just society.	$\{ \widehat{u} : 70, S(\widehat{u}): "Certainly agree - 6", \widehat{o}^{f,nal}: "Probably agree"\}$	$\{\overline{\Delta}_{\rm stance}: \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\$
7d58		Post Opin- ion	Democracy is good. It gives everyone equality	$\{ u : 7, ofinal: "Probably agree - 5", S(u): "Probably agree - 5", \}$	I now lean towards agreeing that it is a characteristic of democracy for women to have the same rights as men. While I initially believed in recognizing unique gender differences, I've come to understand that promoting equality doesn't mean disregarding those differences; rather, it means ensuring that everyone has equal opportunities and rights in a fair society. Through thoughtful discussions and practical steps, like mentorship programs and equal pay initiatives, we can support both equality and the unique contributions of individuals, regardless of gender.	$\{ \widehat{u} :$ 86, $S(\widehat{u}):$ "Lean agree - 4", $\widehat{\sigma}^{\rm final}:$ "Lean agree" $\}$	$\{\overline{\Delta}_{\mathtt{stance}}\colon \frac{1}{S}_{\mathtt{sem}}\colon \\ 0.56, \\ \overline{\mathtt{ROUGE-L}}\colon \\ 0.09 \\ \}$
7e14		Post Opin- ion	Most people share the same idea based on equalit	$\{ u :$ 9, $o^{\mathrm{final}}:$ "Lean disagree - 3", $S(u):$ "Probably agree - 5"}	I certainly agree that it is a characteristic of democracy for women to have the same rights as men. Equality is fundamental to a just society, and when all individuals are granted equal rights, it strengthens democracy and fosters a fairer world for everyone. Through advocacy, awareness, and supporting initiatives that empower all genders, we can work towards a society that truly values and respects each individual's contributions.	$\{ \widehat{u} : 69, S(\widehat{u}): "Certainly agree - 6", \widehat{o}^{f,nal}: "Certainly agree"\}$	$\begin{split} &\{\overline{\Delta}_{\text{stance}}\colon\\ \frac{1}{S}_{\text{sem}}\colon\\ &0.36,\\ &\overline{\text{ROUGE-L}}\colon\\ &0.05 \end{split}$

Demographic profile for ID 67210 - as fed into LLM's system message

,,,,,

You are a 30-year-old female with some-college education. Your ethnicity is white, and your annual income falls in the 75k-100k range. Politically, you identify as

republican with moderate views. You have children in ['no-children'] school, reside in a suburban area, and your marital status is married. Regarding religious beliefs, you consider the Bible to be ancient-book, no identify as evangelical, and your religious affiliation is no-religion. Your occupation is insurance.

On the Likert-scale, you chose Probably agree as your initial opinion with respect to the statement "It is a characteristic of democracy for women to have the same rights as men".

Below is your explanation for your initial opinion.

I believe that it is a characteristic of democracy for women to have the same rights as men. Nothing about democracy says it isn't. While we have had to push for these changes and the growth, we are slowly getting closer to equality and changing the minds of old men stuck in their ways. """

Observations: In general, the LLM uses significantly more words than humans in the simulated tweets and messages. An interesting pattern can be seen in this conversation, where the stance alignment is more or less a perfect match for all of the tweets and exchanges. However, upon reading the LLM text column, we can see clear signs of LLM covergence, where the LLM seems to be repeating the point along the lines of "we have made strides in promoting equality, but we still have a long way to go."

One can also observe slight hallucination for the post opinion of Sender ID - **67d58**. The human has been agreeing with the topic throughout and this is reflected in his post opinion, "Democracy is good. It gives everyone equality", but the LLM seems to be confused by the demographic profile of the human, as reflected by its post opinion: "I now lean towards agreeing that it is a characteristic of democ racy for women to have the same rights as men. While I initially believed in recognizing unique gender differences, I've come to understand that promoting equality doesn't mean disregarding those differences;" This might have been due to the demographic profile of this ID.

Conversation 2:

Table 19: Sample depth topic full conversation simulation results. For ease of reading we have chosen to track Sender ID - **5dec8** (see demographic profile below) for this conversation (gray rows). This ID was chosen post qualitative evaluation of the opinion simulation for his ID. We report the raw semantic similarity $\overline{S}_{\text{sem}}$, absolute stance difference $\overline{\Delta}_{\text{stance}}$ and $\overline{\text{ROUGE-L}}$ scores, both before and after SFT, as indicators of alignment.

Topic: A "body cleanse," in which you consume only particular kinds of nutrients over 3-4 days, helps your body to eliminate toxins.

Sender ID	Recipien ID	t Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignment Metrics
5dec8		Initial Opin- ion	I really don't believe this is true. I think it is a fad or a gimmick. I don't think you need to detoxify your body via the food you eat, and if you even could I don't think it could be done over 1-3 days.	{ u : 49, oinit: "Certainly disagree - 1", S(u): "Certainly disagree - 1" }		_		
5e638		Initial Opin- ion	I think it makes sense that targeted nutrients can help you detox. Your body is like a machine that can thrive in certain areas if	$\{ u : 25, o^{\text{init}}: \\ o^{\text{init}}: \\ \text{"Probably agree - 5",} \\ S(u): \\ \text{"Lean agree - 4"} \}$	_	_		
66246	_	Initial Opin- ion	A detox can help reset your gut biome by elimi- nating sugars and other toxins	$\{ u : 14, o^{\text{init}}: \\ v^{\text{probably}} \\ v^{\text{agree}} - 5\text{"}, \\ v^{\text{goal}} \\ v^{\text{probably}} \\ v^{\text{agree}} - 5\text{"}, \\ v^{\text{goal}} \\ v^{$	_	_		

Sender ID	Recipien ID	Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignmen Metrics
67e05	_	Initial Opin-	Yes, I think hav- ing particular pre-	$\{ u : 43, o^{\text{init}}: \text{"Lean}\}$		_		
		ion	scribed nutrients	disagree -				
			in your diet that have the power to	S(u):				
			eliminate toxins from the body sys-	"Certainly agree - 6"				
			tem can be very	}				
			helpful in cleans- ing the body of					
			these harmful ac- cumulations. It					
			surely works be- cause there such					
			nutrients in some					
5dec8	5e638	tweet	foods. I don't think this is	{ u : 51,	_			
			a true statement. I don't believe that	S(u): "Certainly				
			the body needs to	disagree - 1"				
			be "de-toxified," and even if it did	}				
			I don't think it					
			would be possible to do it over only					
			1-3 days. I think this is a fad and a					
5.620	51.0		gimmick.					
5e638	5dec8	tweet	I see our body as a Machine that can	$\{ u : 42, S(u): \}$	_	_		
			accel in different areas if it's primed	"Certainly agree - 6"				
			properly. I think	}				
			it makes perfect sense that our bod-					
			ies can thrive un- der certain con-					
			ditions using tar-					
			geted nutrients to eliminate toxins. I					
5dec8	5e638	message	think we ca I agree that the	{ u : 31,	I appreciate your	I think that	Pre-SFT:	Pre-SFT:
Jucco	30036	sent	body is a machine,	S(u):	thoughtful response,	the liver and	$\{ \widehat{u} \colon$	$\{\overline{\Delta}_{\text{stance}}$
			but maintenance isn't something	"Certainly disagree - 1"	5e638. I agree that our bodies do have	kidneys already do a good job	$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$:
			that can be ac- complished over	}	mechanisms for processing toxins,	at detoxifying the body, and I	"Certainly disagree	0.64, ROUGE-L
			a couple of days.		and it's important	don't see how	- 1"	0.17
			It is a long-term effort, a lifestyle		to find what works for each individual.	a temporary change in diet	<pre>} Post-SFT:</pre>	} Post-SF
			choice.		However, I still be- lieve that the idea of	could improve the results.	{ \widehat{u} :	$\{\overline{\Delta}_{\text{stance}}\}$
					a quick fix through	the results.	$S(\widehat{u})$:	$\frac{5}{S}_{\text{sem}}$:
					a short-term cleanse is misleading. It can		"Certainly agree -	0.52, ROUGE-L
					create unrealistic expectations about		6" }	0.19
					health and wellness.		ı	}
					While some may feel rejuvenated			
					after a cleanse, I think a balanced			
					diet and healthy			
					lifestyle in the long run will yield better			
					results. It's all about sustainability for me			

Sender ID	Recipien ID	nt Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignment Metrics													
5e638	5dec8	message	I think to achieve	{ u : 27,	I completely un-	That's a fair	Pre-SFT:	Pre-SFT:													
		sent	higher levels takes longer term	S(u): "Lean agree - 4"	derstand where you're coming from,	point. Our bodies are in-	{ \wideta : 91,	$\{\Delta_{ ext{stance}}:\ 1,$													
			than a couple	}	and I agree that	deed designed	$S(\widehat{u})$:	$\overline{S}_{ ext{sem}}$:													
			days but it's possible to see		sustainability is key in maintaining	to detoxify themselves.	"Probably agree -	0.31, ROUGE-L:													
			results in as little		health. It's easy	But sometimes,	5"	0.14													
			as a few days		to get caught up in quick fixes,	factors like diet, environment,	<pre>} Post-SFT:</pre>	Post-SFT:													
					especially with so	and lifestyle	$\{ \widehat{u} :$	$\{\overline{\Delta}_{\text{stance}}:$													
					much information out there. I do	can put extra strain on these	$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$ ":													
					think that for some	organs. The	"Lean	0.24,													
					people, a short-term cleanse can serve	idea behind a cleanse isn't to	agree - 4"	ROUGE-L: 0.13													
			as a jumpstart to	replace their	}	}															
					healthier habits, but it's definitely not a	function, but to support them															
					long-term solution.	with nutrients															
					A balanced diet and consistent	that might be lacking in our															
					lifestyle changes are	regular diet. It's not a cure-all,															
					essential for lasting health. Have you	not a cure-all, but it could															
					found any particular	help some															
					strategies or habits that have worked	people feel a little better.															
					well for you in																
5dec8	5e638	message	I think it would	{ u : 46,	maintaining Absolutely, 5e638!	I think if you	Pre-SFT:	Pre-SFT:													
		sent	take longer than	S(u):	I think the key to	eat a balanced	$\{ \widehat{u} \colon$	$\{\overline{\Delta}_{\text{stance}}:$													
			that. I see fads like this quite	"Certainly disagree - 1"	long-term health is developing habits	diet, your body gets all the vita-	$S(\widehat{u})$:	$\frac{5}{S}_{\text{sem}}$:													
			a bit, someone	}	that are both enjoy-	mins and miner-	"Certainly	0.30,													
			on tiktok posts a video and then		able and sustainable. For me, incorpo-	als it needs.	agree - 6"	ROUGE-L: 0.09													
				it goes viral. On		rating more whole foods into my diet		}	}												
			top of which, eliminating toxins		has been a game-		Post-SFT: $\{ \widehat{u} :$	Post-SFT: $\{\overline{\Delta}_{\text{stance}}:$													
			from your body		changer. I focus on eating plenty of		$S(\widehat{u})$:	5,													
			happens everyday naturally, as a		fruits, vegetables,		"Certainly	$\overline{S}_{\text{sem}}$: 0.28,													
			function of the digestive system.		and lean proteins while trying to		agree - 6"	ROUGE-L: 0.19													
					limit processed		}	}													
							foods. Regular exercise, even just														
																		walking, has also			
											been beneficial for both my physical										
												and mental health.									
														I find that making small, consistent							
					changes rather than drastic shifts has																
					helped me maintain																
66246	5dec8	tweet	I think that cleansing your	$\{ u \colon 25,\ S(u)\colon exttt{"Lean} \}$	I believe a body cleanse can help	My Tweet: A detox can	Pre-SFT: $\{ \widehat{u} :$	Pre-SFT: $\{\overline{\Delta}_{\text{stance}}:$													
			body of sugars	agree - 4"	eliminate toxins,	help eliminate	47,	2,													
			and processed food could have	}	especially given the processed foods we	sugars and other toxins	$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$: 0.71,													
			a profound effect		consume today. For	from your sys-	agree -	ROUGE-L:													
			on the whole body by resetting		1-3 days, focusing on particular nutri-	tem. However, it is only a	6 " }	0.14													
			the gut biome		ents might give our digestive system a	temporary fix, and should	Post-SFT:	Post-SFT:													
					break and reset our	be combined	{ \widehat{u} : 33,	$\{\Delta_{\text{stance}}: \underline{1},$													
					gut health. It's like a fresh start for our	with healthy permanent	$S(\widehat{u})$: "Lean	$\overline{S}_{\text{sem}}$ ": 0.61,													
					bodies! #Detox	choices to have	disagree	ROUGE-L:													
						lasting results.	- 3" }	0.10													
5dec8	66246	tweet	I don't think this	{ u : 48,	I don't believe that	My Tweet: I	Pre-SFT:	Pre-SFT:													
			would work. It would take longer	S(u): "Certainly	a "body cleanse" over 1-3 days truly	think this is a false statement.	$\{ \widehat{u} :$	$\{\overline{\Delta}_{\text{stance}}: 0,$													
			than 1-3 days to	disagree - 1"	helps eliminate	I don't believe	$S(\widehat{u})$:	$\overline{S}_{ exttt{sem}}$:													
			detox the body, if detoxing the body	}	toxins. I see it as a fad that can create	the body needs "cleansed" or	"Certainly disagree	0.81, ROUGE-L:													
			is something you		unrealistic health	"de-toxified."	- 1"	0.19													
			can even do by consuming only		expectations. Our bodies already have	I think this is a gimmick or	<pre>} Post-SFT:</pre>	<pre>Post-SFT:</pre>													
			particular kinds of		mechanisms for	fad. I don't see	$\{ \widehat{u} \colon$	$\{\overline{\Delta}_{ exttt{stance}}\colon$													
		p	particular kinds			detoxification. A balanced diet and	how it would be possible to	$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$ ":												
		nu the	nutrients. I think the body already	1 1 2																	
		ge	gets rid of any- thing it needs to.		healthy lifestyle	do this in 1-3	"Certainly	0.61,													
			gets rid of any-		are the real keys to	days, even if it	disagree	ROUGE-L:													
			gets rid of any-																		

2052	Sender ID	Recipient ID	Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignment Metrics
2053	5dec8	66246	message	I agree with you	{ u : 43,	I appreciate your	I do agree that	Pre-SFT:	Pre-SFT:
2054 2055			sent	on that, but don't think it would re-	$S(u)\colon$ "Lean agree - 4"	perspective, 66246. I understand the	a detox would be a temporary	{ \widehat{u} : 85,	$\{\Delta_{ ext{stance}}: \ \underline{1},$
2055				ally make a bunch of difference do-	}	appeal of a fresh start and how a	measure, but I think that a bal-	$S(\widehat{u})$: "Lean	$\overline{S}_{\text{sem}}$: 0.60,
2057				ing it only for 1-3 days. It		short-term focus on specific nutrients	anced diet and exercise are	disagree - 3"	ROUGE-L:
2057				would take more		might seem ben-	better methods	}	}
2059				effort, making it a lifestyle choice		eficial, especially with our modern	of maintaining health.	Post-SFT: $\{ \widehat{u} :$	Post-SFT: $\{\overline{\Delta}_{\text{stance}}:$
2060				of avoiding thinks like sugars and		diets. However, I still believe that		$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$ ":
2061				processed foods.		our bodies are quite capable of detox-		"Lean agree -	0.64, ROUGE-L:
2062						ifying themselves		4 "	0.20
2063						naturally without the need for drastic		}	}
2064						measures. While a cleanse might feel			
2065						refreshing for some, I worry it could lead			
2066						to misconceptions about the need for			
2067						such practices. I			
2068						think promoting a consistent, balanced			
2069						approach to eating is more effective for			
2070	66246	5dec8	message	I agree that 3	{ u : 21,	long I completely under-	I agree that	Pre-SFT:	Pre-SFT:
2071			sent	days is not likely to produce any	S(u): "Certainly	stand where you're coming from, 5dec8.	a detox is temporary. But	$\{ \widehat{u} :$ 85,	$\{\overline{\Delta}_{ exttt{stance}}\colon$
2072				noticeable results.	agree - 6"	It makes sense that	it can help reset	$S(\widehat{u})$:	$\frac{2}{S}_{\text{sem}}$:
2073				It would take real commitment to	}	our bodies have built-in systems to	your gut biome and help you	"Lean agree -	0.38, ROUGE-L:
2074				lifestyle change		detoxify, and I agree that a balanced	get a better start on your	4 " }	0.09
2075						diet is crucial for long-term health. I	new healthier habits.	Post-SFT: $\{ \widehat{u} :$	Post-SFT: $\{\overline{\Delta}_{\texttt{stance}}:$
2076						just think that some-		27,	$\frac{2}{S_{\text{sem}}}$ ":
2077						times people can get overwhelmed with		$S(\widehat{u})$: "Lean	0.48,
2078						all the processed foods and unhealthy		agree - 4"	ROUGE-L: 0.21
2079						options out there. A short cleanse might		}	}
2080 2081						serve as a reset for some, helping them			
2082						refocus on healthier eating habits. I			
2083						see it more as a			
2084						tool rather than a solution. Have you			
2085	67e05	5dec8	tweet	A detox diet,	{ u : 23,	ever tried I lean towards the	My Tweet: I	Pre-SFT:	Pre-SFT:
2086				preferably pre- scribed my a	S(u): "Probably	belief that a "body cleanse" can help	believe that a body cleanse	{ û : 46,	$\{\overline{\Delta}_{\text{stance}}: \ \underline{1},$
2087				nutritionist, can be very helpful	agree - 5"	eliminate toxins to some extent, but our	can help the body to elim-	$S(\widehat{u})$: "Lean	$\overline{S}_{\text{sem}}$: 0.70,
2088				in eliminating	1	bodies have their	inate toxins.	agree -	ROUGE-L:
2089				accumulated toxins such as		own natural detox processes. A short	I think that this is because	4 " }	0.12
2090				sugars, from the blood stream.		cleanse might be beneficial after a	nutrients have a role to play	Post-SFT: $\{ \widehat{u} :$	Post-SFT: $\{\overline{\Delta}_{\texttt{stance}}:$
2091						period of unhealthy eating, yet balance	in the natural detoxification	S_{6} , $S(\widehat{u})$:	$\frac{1}{\overline{S}}_{\text{sem}}$ ":
2092						and moderation	mechanism	"Certainly	0.60,
2093						are key for lasting health.	of the body. By cleansing	agree - 6"	ROUGE-L: 0.13
2094							the body, it allows it to	}	}
2095							naturally do what it is meant		
2096							to do, which		
2097							toxins.		
2098									
2099									
2100									
2101									

Sender ID	Recipien ID	t Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignmen Metrics
5dec8	67e05	tweet	I don't think this is	{ u : 41,	I don't believe that	My Tweet: I	Pre-SFT:	Pre-SFT:
			necessarily a true statement. I don't	S(u): "Certainly	a "body cleanse" over 1-3 days truly	don't think this is a true	{ û : 44,	$\{\Delta_{\text{stance}} \ 0$
			think this can re- ally be accom-	disagree - 1"	helps eliminate tox- ins. Our bodies	statement. I don't think the	$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$: 0.60,
			plished after only	1	have natural mech-	body needs a	disagree	ROUGE-L:
			1-3 days. I think it requires a life-		anisms for detoxifi- cation. Relying on	"body cleanse," and even if	- 1" }	0.19
			style choice of eating non-sugary,		quick fixes can cre- ate unrealistic health	it did I don't think it could	Post-SFT: $\{ \widehat{u} :$	Post-SFT
			unprocessed, low		expectations. A	be done in 1-3	46,	$\{\overline{\Delta}_{\text{stance}}, \underline{0}, 0$
			fat foods, etc.		balanced diet and healthy lifestyle are	days. I think this is a fad or	$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$ ": 0.57,
					the true keys to well- ness.	gimmick.	disagree - 1"	ROUGE-L:
							}	}
67e05	5dec8	message sent	You're probably right. Getting tox-	$\{ u : 49, S(u): \}$	I completely agree with you, 5dec8! It's	I respect your opinion and	Pre-SFT: $\{ \widehat{u} :$	Pre-SFT: $\{\overline{\Delta}_{\text{stance}}\}$
			ins from the body	"Probably	important to view a	I think there	83,	1,
			using this natural way in three days	agree - 5" }	cleanse as a tempo- rary measure rather	are many ways to arrive	$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$: 0.41,
			sounds very ambi- tious and unrealis-		than a fix for on- going poor eating	at the same destination	agree -	ROUGE-L:
			tic. Very few if		habits. Focusing on	of long-term	}	}
			any. Only when one makes it a		long-term, sustain- able changes is def-	health.	Post-SFT: $\{ \widehat{u} :$	Post-SF: $\{\overline{\Delta}_{\text{stance}}\}$
			habit to include detoxifying nutri-		initely the way to go. As for strategies,		$S(\widehat{u})$:	$\frac{1}{S}_{\text{sem}}$ ":
			ents in the diets		I try to meal prep		"Lean	0.26,
			they are able to see results in the		to make healthier choices more con-		agree - 4"	ROUGE-L 0.14
			long run		venient during the week. I also make		}	}
					a point to incorpo-			
					rate more fruits and vegetables into my			
					meals. What about you? What strate-			
					gies have you found			
					effective in main- taining a balanced			
5dec8	67e05	message	I agree. I think	{ u : 34,	lifestyle? I love your approach	I don't think	Pre-SFT:	Pre-SFT
		sent	it is important to	S(u):	to meal prepping!	that the body	$\{ \widehat{u} \colon$	$\{\overline{\Delta}_{\text{stance}}$
			keep your body clean and healthy,	"Certainly agree - 6"	It really does make a difference in	needs to be helped in	$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$:
			but it's a long term day in/day	}	sticking to healthy choices during the	eliminating toxins.	"Certainly agree -	0.46, ROUGE-L
			out type of thing		busy week. I also try		6"	0.15
			well, I think we agree on this?		to plan my meals in advance, and I find		Post-SFT:	Post-SF
					that having healthy snacks on hand, like		$\{ \widehat{u} :$ 14,	$\{\Delta_{\text{stance}} $ 5,
					cut-up veggies or fruit, really helps		$S(\widehat{u})$: "Certainly	$\frac{5}{S}_{\text{sem}}$ ": 0.39,
					curb unhealthy		disagree	ROUGE-L
					cravings. I make it a point to incorporate		- 1", }	0.17
					a variety of colors on my plate, which			
					not only looks			
					appealing but also ensures I'm getting			
					a range of nutrients. Finding that balance			
					and making healthy			
					eating enjoyable has been			

Sender ID	Recipient ID	Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignmen Metrics																	
66246	_	Post	I am even more	{ u : 21,	I still lean towards	Okay, that's	Pre-SFT:	Pre-SFT:																	
		Opin- ion	confident now that a body	o ^{final} : "Certainly	agreeing that a "body cleanse"	understand- able. Yes, I	$\{ \widehat{u} :$ 80,	{ Δ_{stance} 2,																	
			cleanse would	agree - 6",	can help eliminate	think the body	$S(\widehat{u})$:	$\overline{S}_{ exttt{sem}}$:																	
			be beneficial in eliminating the	S(u): "Certainly	toxins, especially considering how	has a natural mechanism for	"Lean agree -	0.80, ROUGE-L:																	
			toxins most of us carry	agree - 6"	our modern diets are often filled with	detoxification and cleansing.	$\widehat{o}^{\text{final}}$:	0.12																	
			curry	}	processed foods.	and cicansing.	"Certainly	Post-SFT																	
					From my conversa- tions, I've realized		disagree - 1"	{ $\Delta_{ extsf{stance}}$ 0,																	
					that while it's not a magic solution, a		<pre>} Post-SFT:</pre>	$\overline{S}_{\text{sem}}$ ": 0.29,																	
					short cleanse can		$\{ \widehat{u} \colon$	ROUGE-L:																	
					serve as a helpful tool to reset eating		$S(\widehat{u})$:	0.07																	
					habits and refocus on healthier choices.		"Lean agree -	•																	
					It can give our		agree - 4", $\widehat{o}^{\text{final}}$:																		
					digestive system a break and provide a		ô ^{11⊓a1} :																		
					fresh perspective on		}																		
					how different foods affect our energy																				
5dec8	_	Post	I might have	{ u : 42,	and well-being. I still lean towards	I agree with	Pre-SFT:	Pre-SFT:																	
Succe		Opin-	changed a little	$o^{ ext{final}}$:	disagreeing with the	that.	$\{ \widehat{u} \colon$	$\{\overline{\Delta}_{ exttt{stance}}$																	
	ion	ion	bit towards lean- ing agreeing. A	"Probably disagree -	idea that a "body cleanse," which		$S(\widehat{u})$:	$\frac{1}{\overline{S}}_{\text{sem}}$:																	
				couple of my	2",	involves consuming		"Lean	0.42,																
		partners made some good points.	S(u): "Lean agree - 4"	only particular kinds of nutrients over		disagree	ROUGE-L: 0.13																		
			At the end though, we all agreed it is	}	1-3 days, effectively helps eliminate		$\widehat{o}^{ ext{final}}$: "Lean	} Post-SF																	
			not really realistic		toxins. While I		disagree	$\{\overline{\Delta}_{ exttt{stance}}$																	
			to think you could do this in only		understand that some people may		- 3" }	$\frac{4}{S}_{\text{sem}}$ ":																	
			1-3 days.		find it refreshing and it might serve		Post-SFT: $\{ \widehat{u} : 4,$	0.22, ROUGE-L																	
					as a temporary		$S(\widehat{u})$:	0.07																	
					reset, I believe our bodies already have		"Certainly agree -	}																	
						natural mechanisms for detoxification.		6", $\widehat{o}^{\text{final}}$:																	
																	Relying on quick								
																						fixes can create unrealistic expec-		}	
																						tations, and it's more beneficial to			
																				focus on sustainable					
					healthy eating habits and a balanced																				
					lifestyle in the long run.																				
67e05	_	Post	Not practical	{ u : 36,	I still lean towards	I think that al-	Pre-SFT:	Pre-SFT:																	
		Opin- ion	because 1 to 3 days sounds a	o ^{final} : "Lean disagree -	disagreeing with the notion that a	lowing it to do so the more ef-	$\{ \widehat{u} :$ 76,	$\{\Delta_{ extsf{stance}} \ 0$,																	
			very short time for this natural	3",	"body cleanse" is a necessary method	ficiently is the goal of doing a	$S(\widehat{u})$: "Lean	$\overline{S}_{\text{sem}}$: 0.67,																	
			way to work.	$S(u)\colon$ "Lean disagree - 3"	for eliminating	body cleanse.	disagree	ROUGE-L																	
			But prescribed detox diet can	}	toxins. While I acknowledge that a		- 3", $\widehat{o}^{\text{final}}$:	0.11																	
			surely work in the long-term if		short cleanse might provide a temporary		"Probably agree -	Post-SF: $\{\overline{\Delta}_{\text{stance}}\}$																	
			one makes it their		reset, I believe our		5"	$\frac{3}{S_{\text{sem}}}$ ":																	
			eating habit.		bodies are naturally equipped to detoxify		<pre>} Post-SFT:</pre>	S_{sem} ": 0.13,																	
					themselves through a balanced diet and		{ \widehat{u} :	ROUGE-L 0.08																	
					healthy lifestyle. It's		$S(\widehat{u})$:	}																	
					more important to cultivate sustainable		"Lean agree -																		
					eating habits rather		4", $\widehat{o}^{\text{final}}$:																		
				than relying on quick fixes, as true		"Probably																			
				wellness comes from long-term		agree - 5"																			
				changes rather than		}																			
				short-term cleanses.																					

Sender	Recipien	t Event	Human Text	Human Text	LLM Text	LLM Text	LLM Text	Alignment
ID	ID	Type		Metrics		Postft	Metrics	Metrics
5e638		Post	I think our bodies	{ u : 46,	I would say I lean to-	I still agree	Pre-SFT:	Pre-SFT:
		Opin-	respond to what	o^{final} :	wards agreeing that	that a detox	$\{ \widehat{u} :$	$\{\overline{\Delta}_{\mathtt{stance}}:$
		ion	we put in them.	"Probably	a "body cleanse" can	can help reset	74,	0,
			They're made to	agree - 5",	help some individu-	your gut biome	$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$:
			process and not	S(u): "Lean	als feel rejuvenated,	and eliminate	"Lean	0.47,
			to produce. This	agree - 4"	as it may provide	toxins, but	agree -	ROUGE-L:
			is why we need	}	a temporary reset	it should be	4",	0.08
			to eat and drink		and give the diges-	combined with	$\widehat{o}^{ ext{final}}$:	}
			to survive. If		tive system a break.	permanent	"Probably	Post-SFT:
			we want to tar-		However, I believe	healthy choices	agree -	$\{\overline{\Delta}_{\text{stance}}:$
			get certain areas		it's essential to rec-	to be effective	5"	$\frac{1}{S}'_{\text{sem}}$ ":
			we select methods		ognize that our bod-	long term.	}	$\overline{S}_{ extsf{sem}}$ ":
			and nutrients that		ies already have nat-		Post-SFT:	0.71,
			boost efficiencies		ural detoxification		$\{ \widehat{u} \colon$	ROUGE-L:
					processes in place.		29,	0.16
					Therefore, while a		$S(\widehat{u})$:	}
					short-term cleanse		"Probably	
					might be beneficial		agree -	
					for some, it's crucial		5",	
					to pair it with long-		$\widehat{o}^{ ext{final}}$:	
					term healthy habits		"Probably	
					for sustainable well-		agree -	
					ness.		5"	
							}	

Demographic profile for id 5dec8:

,,,,,

You are a 61-year-old male with bachelor education. Your ethnicity is white, and your annual income falls in the 75k-100k range. Politically, you identify as independent with very-conservative views. You have children in ['out-of-school'] school, reside in a urban area, and your marital status is divorced. Regarding religious beliefs, you consider the Bible to be literal, yes identify as evangelical, and your religious affiliation is protestant. Your occupation is education.

On the Likert-scale, you chose Certainly disagree as your initial opinion with respect to the statement "A 'body cleanse,' in which you consume only particular kinds of nutrients over 1-3 days, helps your body to eliminate toxins".

Below is your explanation for your initial opinion.

I really don't believe this is true. I think it is a fad or a gimmick. I don't think you need to detoxify your body via the food you eat, and if you even could I don't think it could be done over 1-3 days.

Observations: This debate simulation highlights the challenges of finetuning and supports earlier findings. Across the table, it can be observed that surface level metrics like $|\hat{u}|$ (LLM word count) approach |u| (human word count). However, a close look at the post opinions of the people being simulated reveals that the finetuning procedure worsened the LLM performance. Indeed, a look at the Post Opinions for ID - **66246** (human: "I am even more confident now that a body cleanse would be beneficial in eliminating the toxins most of us carry" vs LLM: "Okay, that's understandable. Yes, I think the body has a natural mechanism for detoxification and cleansing" and **5dec8** (human: "I might have changed a little bit towards leaning agreeing. A couple of my partners made some good points. At the end though, we all agreed it is not really realistic to think you could do this in only 1-3 days." vs LLM: "I agree with that.") demonstrates clear misalignment where the LLM text does not resemble the human at all (post finetuning). The post opinion stance for Sender ID **5dec8** also shows a sharp contrast between human and LLM (Probably disagree -> Certainly agree).

Conversation 3:

Table 20: Sample depth topic full conversation simulation results. For ease of reading we have chosen to track Sender ID - **678f0** (see demographic profile below) for this conversation (gray rows). This ID was chosen post qualitative evaluation of the opinion simulation for his ID. We report the raw semantic similarity $\overline{S}_{\text{sem}}$, absolute stance difference $\overline{\Delta}_{\text{stance}}$ and $\overline{\text{ROUGE-L}}$ scores, both before and after SFT, as indicators of alignment.

Topic: Angels are Real.

Sender ID	Recipient ID	Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignmen Metrics
663e9		Initial Opin- ion	Angels are just something you can't see but feel. And it may not be an actual angel but it's like energy that is moving. Like spirits would be considered angels to me	$\{ u : 34, o^{\text{init}}: \\ o^{\text{init}}: \\ \text{"Probably agree - 5",} \\ S(u): \\ \text{"Lean agree - 4"} \}$		_		
67a9f		Initial Opin- ion	I believe that angels are real. Angels are men- tioned in the bible	{ u : 12, oinit: "Certainly agree - 6", S(u): "Certainly agree - 6"}				
5f596		Initial Opin- ion	I don't think angels are real. In general, I don't think anything supernatural is real. I've never believed in any of that stuff.	$\{ u : 26, o^{\text{init}}: \\ \text{"Certainly disagree - 1"}, \\ S(u): \\ \text{"Certainly disagree - 1"} \}$		_		
678f0	_	Initial Opin- ion	I believe thgat angels are	$\{ u :$ 5, $o^{\mathrm{init}}:$ "Certainly agree - 6", $S(u):$ "Certainly agree - 6" $\}$	_	_		
5f596	678f0	tweet	I don't think any- thing supernatu- ral, including an- gels, are real. I've never believed in any of that stuff. It all sounds fic- tional to me.	{ u : 25, S(u): "Certainly disagree - 1" }				
678f0	5f596	tweet	I am a muslim and in my re- ligion we are taught to believe that angels are real and they do exist and some that are named have specific roles. One of the techings that I have always remembered is that there are two angels present in every person one	$\{ u :$ 50, $S(u):$ "Certainly agree - 6" $\}$	_			

Sender ID	Recipier ID	nt Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignment Metrics		
678f0	5f596	message	it's funny that we	{ u : 28,	I appreciate your	That's okay.	Pre-SFT:	Pre-SFT:		
		sent	have dog icons for our characters	$S(u)\colon$ "Lean disagree - 3"	thoughtful response, 5f596. It's refresh-	Different beliefs and	{ \widehat{u} : 86,	$\{\Delta_{\text{stance}}: \underline{1},$		
			one of the beliefs is that having a	}	ing to hear someone value evidence	perspectives are totally	$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$:		
			dog in one's house wards off the an-		and rationality in their beliefs. I	valid.	agree -	ROUGE-L:		
			gels		think there's def-		}	}		
					initely merit in understanding the		Post-SFT: $\{ \widehat{u} :$	Post-SFT: $\{\overline{\Delta}_{\text{stance}}:$		
					world through a scientific lens. For		$S(\widehat{u})$:	$\frac{1}{S}_{\text{sem}}$:		
					me, though, faith and spirituality offer		"Certainly agree -	ROUGE-L:		
					a different kind of		6 "			
					understanding— one that fills in the		}	}		
					gaps that science might not address.					
					It's fascinating how both perspectives					
					can coexist and					
					provide comfort in different ways.					
					Do you ever find yourself curious					
					about the beliefs others hold, even if					
67a9f	678f0	tweet	Hi, I am of the be-	$\{ u : 10,$	I firmly believe that	I agree with	Pre-SFT:	Pre-SFT:		
			lief that angels are real	S(u): "Certainly	angels are real. The Bible mentions them	you that beliefs and perspec-	$\{ \widehat{u} :$ 47,	$\{\Delta_{\text{stance}}: \underline{0},$		
				agree - 6" }	numerous times, de- scribing them as di-	tives are valid	$S(\widehat{u})$: "Lean	$\overline{S}_{\text{sem}}$: 0.87,		
					vine beings with a purpose in God's		agree - 4"	ROUGE-L:		
					plan. Their exis-		}	}		
					tence offers comfort and hope, reminding		Post-SFT: $\{ \widehat{u} :$	Post-SFT $\{\overline{\Delta}_{\text{stance}}:$		
					us that we are not alone. Faith in an-		$S(\widehat{u})$:	$\frac{3}{S}_{\text{sem}}$:		
					gels can enrich our spiritual journey!		"Lean agree -	ROUGE-L:		
		; ;			spiritual journey:		4 "			
678f0	67a9f		tweet	tweet	Hello I believe in	$\{ u : 36,$	As a Muslim, I	My Tweet: I	Pre-SFT:	Pre-SFT:
			angels how about you? It is part	S(u): "Certainly	certainly believe that angels are real.	believe that angels are	{ û : 46,	$\{\Delta_{ ext{stance}}: 0,$		
			of my religion so I believe through faith how about you? For one	agree - 6" }	My faith teaches that they have spe-	real. The bible mentions	$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$: 0.77,		
				,	cific roles and are present in our lives,	angels many	agree -	ROUGE-L:		
			if people believe		even if we can't see	times both in the old and	}	}		
			there is a devil then there's one		them. They serve as guardians and	new testament. Other religions	Post-SFT: $\{ \widehat{u} :$	Post-SFT $\{\overline{\Delta}_{\text{stance}}:$		
			angel!		messengers, guiding us in ways that are	such as Islam have angels as	$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$:		
					often beyond our understanding.	part of their belief.	"Certainly agree -	0.76, ROUGE-L:		
					understanding.	belief.	6"	0.23		
678f0	67a9f	message	well for one as-	{ u : 47,	Thank you for your	My Tweet: I	Pre-SFT:	Pre-SFT:		
		sent	pect of the belief is that each person	S(u): "Lean agree - 4"	thoughtful response, 67a9f! I really	believe that angels exist	{ \widehat{u} : 92,	$\{\overline{\Delta}_{\text{stance}}:$ 2,		
			has two angels one on each shoul-	}	appreciate how you recognize the	because that is the teaching	$S(\widehat{u})$: "Certainly	$\overline{\overline{S}}_{\text{sem}}$:		
			der the one on		common ground we	that I have	agree -	ROUGE-L:		
			the right recordes that persons good		share in our beliefs about angels. In	received in my religion. I am	6 " }	0.09		
			deeds while the one on the lrft		my faith, angels are often seen as	a muslim and in the quaran	Post-SFT: $\{ \widehat{u} :$	Post-SFT $\{\overline{\Delta}_{\text{stance}}:$		
			the bad deeds I have no recolec-		guardians who protect us and as	it is stated that angels are real.	$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$:		
			tion of encouter-		messengers who	I am also a per-	"Certainly	0.68,		
			ing one how about you?		deliver guidance from God. They	son of science and I believe	agree - 6"	ROUGE-L: 0.22		
		you			play a vital role in our spiritual growth	that there are many things	}	}		
					and help us navigate life's challenges.	that we do not understand and				
					It's comforting to	cannot se				
					know that we have divine support, even					
					if we can't see it. I'd love to hear more					
					about how your faith influences					
					rarui milluences					

Sender ID	Recipies ID	nt Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignment Metrics		
67a9f	678f0	message sent	Well different religious texts	$\{ u : 39, S(u): \}$	I completely res- onate with what	I am glad we share the same	Pre-SFT: $\{ \widehat{u} :$	Pre-SFT: $\{\overline{\Delta}_{\text{stance}}:$		
		sent	to support that	"Certainly	you've shared,	belief in angels.	87,	0.		
			that angels are real. In the Bible	agree - 6" }	678f0! It's reassur- ing to know that		$S(\widehat{u})$: "Certainly	$\overrightarrow{S}_{\text{sem}}$: 0.69,		
			and the Quran I mean many		angels are seen as guardians and mes-		agree - 6"	ROUGE-L: 0.13		
			individuals claim		sengers in both our		}	}		
			to have a personal encounter with		faiths. In my belief, angels are indeed		Post-SFT: $\{ \widehat{u} :$	Post-SFT: $\{\overline{\Delta}_{\texttt{stance}}:$		
			angels. Yes, I have had a per-		vital in God's plan, serving to protect		$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$:		
			sonal encounter		and guide us on our		"Certainly	0.59,		
			with an angel		spiritual journey. I find comfort in their		agree - 6"	ROUGE-L: 0.27		
					presence, knowing they are there to		}	}		
					support us through life's challenges.					
					It's fascinating how					
					both Christianity and Islam acknowl-					
					edge their roles, even if the details					
					differ. How do you					
663e9	678f0	tweet	Angels I would	$\{ u : 20,$	feel your I probably agree that	I am glad we	Pre-SFT:	Pre-SFT:		
			say are more spiri- tual aspects rather	S(u): "Lean disagree - 3"	angels are real, but I perceive them as	share that belief too.	$\{ \widehat{u} :\ 44,$	$\{\Delta_{ extsf{stance}}\colon 2$,		
			than physical be- ings. I don't see	}	a spiritual presence or energy rather than		$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$:		
			them for religious		literal beings with		agree -	ROUGE-L:		
			reasons.		wings. They mani- fest in signs and feel-		6 " }	0.19		
					ings that resonate personally, reflect-		Post-SFT: $\{ \widehat{u} : 8,$	Post-SFT: $\{\overline{\Delta}_{\texttt{stance}}:$		
					ing the unseen con- nections we have in		$S(\widehat{u})$: "Certainly	$\frac{0}{S_{\text{sem}}}$:		
					life. What do you		agree -	0.27,		
					think?		6 " }	ROUGE-L: 0.04		
678f0	663e9	tweet	tweet	tweet	I am muslim so	{ u : 47,	As a Muslim, I	Do you believe	Pre-SFT:	Pre-SFT:
			I believe in an- gels. One of these	S(u): "Certainly	firmly believe that angels are real. My	that angels manifest in	{ \hat{u} :	$\{\overline{\Delta}_{\text{stance}}: 0,$		
			is the belief of	agree - 6"	faith teaches that	different ways	$S(\widehat{u})$:	$\overline{S}_{ exttt{sem}}$:		
			the presence of two angels on on	}	they exist as protec- tors and messengers,	to different people?	"Probably agree -	0.79, ROUGE-L:		
			each shoulder the one on the right		guiding us in ways we may not fully un-		5 " }	0.13		
			records a person's good deeds and		derstand. Their pres- ence offers comfort		Post-SFT: $\{ \widehat{u} :$	Post-SFT:		
			the one on the		and assurance in our		11,	$\{\Delta_{\text{stance}}: -\frac{2}{3},$		
			left the ad deeds. How about you?		spiritual journey, re- minding us we are		$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$: 0.67,		
					never alone.		agree - 6"	ROUGE-L: 0.16		
678f0	663e9	message	no they just	{ u : 32,	Thank you for your	My Tweet: I	Pre-SFT:	Pre-SFT:		
07010	00307	sent	record your deeds	S(u):	kind words, 663e9!	think angels are	$\{ \widehat{u} \colon$	$\{\overline{\Delta}_{ exttt{stance}}:$		
			and at the end of times on the	"Certainly agree - 6"	I appreciate your perspective on an-	real but more in the sense	$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$:		
			day of judgement your deeds ar4e	}	gels as a spiritual presence or energy.	of spirits or energy. They	"Certainly agree -	0.17, ROUGE-L:		
			used to determine if you get into		It's interesting how our interpretations	will show signs to the individ-	6"	0.10		
			heaven or go to		can vary yet still	ual needing	Post-SFT:	Post-SFT:		
			hell		provide comfort and connection in our	them but not in the form of	{ \widehat{u} : 38,	$\{\Delta_{ extsf{stance}}:\ 1,$		
					lives. My expe- riences with faith	wings and halo as depicted in	$S(\widehat{u})$: "Lean	$\overline{S}_{\text{sem}}$: 0.71,		
					have shown me that	movies.	agree -	ROUGE-L:		
					angels offer guid- ance during chal-		4 " }	0.14		
					lenging times, and I often feel their pres-					
					ence through mo-					
					ments of intuition or signs that seem to					
					appear just when I need them. It's a					
					reminder of the un- seen support around					
					us. I'd love to hear					
					more					

Sender ID	Recipien ID	Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignment Metrics
663e9	678f0	message sent	Do you think the people that are let into heaven are allowed to become angels>	{ u : 15, S(u): "Lean agree - 4" }	I completely agree, 678f0! It's amazing how those moments of intuition or signs can feel like a guiding force in our lives. I've had similar experiences where I sensed a presence or received a sign just when I needed it, which reinforces my belief in a spiritual energy around us. It's like an invisible thread connecting us to something greater. I'm curious, have there been specific instances in your life where you felt particularly guided by angels? I'd love to hear	My Tweet: My Tweet: I be- lieve angels are real because I am a person of science and I believe that there are many things that we do not under- stand. I am also a muslim and I believe that my religion teaches me that angels are real. I think that there are many things that we do not understand and cannot see.	Pre-SFT: $\{ \widehat{u} :$ 87, $S(\widehat{u}):$ "Certainly agree - 6" } Post-SFT: $\{ \widehat{u} :$ 61, $S(\widehat{u}):$ "Lean agree - 4" }	Pre-SFT: $\{\Delta_{\rm stance}: 2, \overline{S}_{\rm sem}: 0.45, \\ \overline{ROUGE-L}: 0.04 \}$ Post-SFT: $\{\Delta_{\rm stance}: 0, \overline{S}_{\rm sem}: 0.64, \\ \overline{ROUGE-L}: 0.13 \}$
678f0	663e9	message sent	I do not think so they simply go to heaven	<pre>{ u : 10, S(u): "Certainly disagree - 1" }</pre>	Absolutely, 663e9! I've had a few experiences that really stand out to me. There was a time when I was facing a difficult decision about my career, and I felt a strong sense of peace and clarity about the path I should take, almost as if I was being nudged in that direction. It turned out to be the right choice for me, and I often reflect on that moment as guidance from a higher presence. Another instance was when I was going through	I agree with you that there are many forms of energy that we cannot see.	Pre-SFT: $\{ \hat{u} : 86, S(\hat{u}): 86, S(\hat{u})$	Pre-SFT: ⟨∆stance: 5/Ssm: 0.11, ROUGE-L: 0.04 } Post-SFT: (∆stance: 0, Ssm: 0.09, ROUGE-L: 0.00
33e9	678f0	message sent	so who gets to be the angel?	$\{ u : 7, S(u): $ "Lean disagree - 3" $\}$	That's such a powerful experience, 678f0! It's incredible how those moments of peace and clarity can feel like divine guidance. I think it's inspiring to hear how you've recognized those nudges as support from a higher presence. I've had similar experiences where I felt a sense of direction during tough times, almost like an internal compass guiding me. It really reinforces the idea that there's something greater at play in our lives. I'd love to hear more about how	I think we all have a sense of them but not in the same form.	Pre-SFT: $\{ \widehat{u} :$ 85, $S(\widehat{u}):$ "Certainly agree - 6" $\}$ Post-SFT: $\{ \widehat{u} :$ 15, $S(\widehat{u}):$ "Certainly agree - 6" $\}$	Pre-SFT: {\(\Delta\) stance: 3, \(\overline{S}\) sem: 0.19, \(\overline{ROUGE-L}\): 0.04 } Post-SFT: {\(\Delta\) stance: 0, \(\overline{S}\) sem: 0.19, \(\overline{ROUGE-L}\): 0.13 }

2484	Sender ID	Recipient	t Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignment Metrics						
2485	678f0	663e9	message	but it is belived	$\{ u : 15,$	Thank you, 663e9!	I agree with	Pre-SFT:	Pre-SFT:						
2486			sent	that children who pass away at an	S(u): "Lean agree - 4"	I truly appreciate your understanding	you that we have a sense of	{ û : 86,	$\{\Delta_{\text{stance}}:\ 2,$						
2487				early age become angels	}	and the way you articulate these	them but not in the same form.	$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$: 0.19,						
2488				angeis		experiences. It's re-	the same form.	agree -	ROUGE-L:						
2489						freshing to connect with someone who		6 " }	0.04						
2490						shares similar feel-		Post-SFT:	Post-SFT:						
2491						ings about spiritual guidance. Beyond		{ û : 15,	{ Δ_{stance} : 5,						
2492						career decisions, I've also felt that		$S(\widehat{u})$: "Lean	$\frac{5}{S}_{\text{sem}}$:						
2493						presence during		agree -	ROUGE-L:						
2494						personal challenges, like when I was		4 " }	0.15						
2495						dealing with a family issue that									
2496						weighed heavily on									
2497						my heart. I found comfort in moments									
2498						of reflection and									
2499						prayer, where I sensed a calming									
2500						energy guiding me to approach									
2501						the situation with									
2502						compassion and understanding. It's									
2502	663e9	678f0	message	ohhhh that's sad	{ u : 4,	those moments That's truly beau-	I think with	Pre-SFT:	Pre-SFT:						
2504	0050	0,010	sent	ommir that o sau	$S(u)\colon$ "Lean	tiful, 678f0! It's	science and	$\{ \widehat{u} \colon$	$\{\overline{\Delta}_{\mathtt{stance}}\colon$						
				disagree - 3" }	amazing how reflec- tion and prayer can	technology and the world of	$S(\widehat{u})$:	$\frac{3}{S}_{\text{sem}}$:							
2505								create space for that calming energy to	social media now it's hard to	"Certainly agree -	0.21, ROUGE-L:				
2506							come through, espe-	say what is an	6 "	0.04					
2507						cially during tough family situations. I	angel and what is not.	<pre>} Post-SFT:</pre>	} Post-SFT:						
2508							think approaching challenges with		$\{ \widehat{u} :$ 23,	$\{\overline{\Delta}_{\text{stance}}:$					
2509						compassion and		$S(\widehat{u})$:	$\frac{-1}{S}$ _{sem} :						
2510						understanding is so important, and it		"Certainly agree -	0.37, ROUGE-L:						
2511													sounds like you've		6 "
2512						been able to tap into that spiritual		}	1						
2513						guidance to help navigate those									
2514						moments. I've									
2515						found that when I take time to reflect,									
2516						I often gain clarity and feel more									
2517						connected to that									
2518						unseen support. It's like a reminder that									
2519	678f0	663e9	message	angels are pure so	{ u : 16,	we I completely agree,	I agree with	Pre-SFT:	Pre-SFT:						
2520	5.510		sent	are children who	S(u):	663e9! Taking	you.	$\{ \widehat{u} \colon$	$\{\overline{\Delta}_{ exttt{stance}}\colon$						
2521				have not had a chance to commit	"Certainly agree - 6"	the time to reflect truly opens up a		$S(\widehat{u})$:	$\frac{0}{S}_{\text{sem}}$:						
2522				any sins	}	channel for that spiritual guidance to		"Certainly agree -	0.07, ROUGE-L:						
2523						come through. It's		6"	0.06						
2524						as if we're tuning into a frequency		Post-SFT:	Post-SFT:						
2525						that connects us to something bigger		$S(\widehat{u}): 4$, $S(\widehat{u}):$	$\{\overline{\Delta}_{\text{stance}}:$						
2526						than ourselves. I		"Lean	$\overline{S}_{ exttt{sem}}$:						
2527						find that those quiet moments often		disagree - 3"	0.06, ROUGE-L:						
2528						lead to the best in-		}	0.0						
2529						sights and a deeper understanding of			1						
						my own feelings and situations. It's									
2530						comforting to know									
2531						that even in the chaos of life, we can									
2532						find peace through reflection and prayer.									
2533						Have you found any									
2534						particular practices that help you									
2535															

Sender ID	Recipien ID	t Event Type	Human Text	Human Text Metrics	LLM Text	LLM Text Postft	LLM Text Metrics	Alignme Metrics																		
663e9		Post	I agree with the	{ u : 5,	I probably agree that	But I do think	Pre-SFT:	Pre-SFT																		
		Opin- ion	statement	o ^{final} : "Lean agree",	angels are real, view- ing them as a spiri-	they are real and we all	$\{ \widehat{u} :$ 80,	$\{\overline{\Delta}_{\text{stance}}\}$																		
				S(u):	tual presence or en-	will see them	$S(\widehat{u})$:	$\overline{S}_{ exttt{sem}}\colon$																		
				"Certainly agree - 6"	ergy rather than lit- eral beings. My con-	in form or another.	"Certainly agree -	0.30, ROUGE-L																		
				}	versations with oth- ers have highlighted		6", $\widehat{o}^{\text{final}}$:	0.10																		
					how personal expe-			Post-SF																		
					riences and feelings can shape our un-		<pre>} Post-SFT:</pre>	$\{\overline{\Delta}_{\text{stance}}\}$																		
					derstanding of an- gels, and I resonate		$\{ \widehat{u} \colon$	$\overline{S}_{\text{sem}}$:																		
					with the idea that		$S(\widehat{u})$:	0.19, ROUGE-I																		
					they can manifest in signs or moments		"Certainly agree -	0.0																		
					of guidance. I be-		6", ô ^{final} :	,																		
					lieve that this spiri- tual connection can		o''''':																			
					provide comfort and support in our lives,		}																			
					regardless of how																					
					one interprets the concept of angels.																					
5f596	_	Post	I definitely do not	$\{ u : 42, final \}$	After engaging in conversations with	I agree with	Pre-SFT:	Pre-SFT																		
		Opin- ion	believe in angels. It seems that most	o ^{final} : "Certainly	others on the topic,	you.	{ \widehat{u} : 75,	$\{\overline{\Delta}_{\text{stanc}},$																		
			people believe in them because of	disagree", $S(u)$:	I still lean towards my original belief		$S(\widehat{u})$: "Probably	$\overline{S}_{\text{sem}}$: 0.71,																		
		religion or some	"Certainly	that angels are not		agree -	ROUGE-I																			
		vague experience they've had that	disagree - 1"	real. While I ap- preciate the comfort		\hat{o}^{final} :	0.18																			
		they can't elabo- rate on. I'm not re-	,	and hope that the concept of angels		"Probably agree -	Post-SI $\{\overline{\Delta}_{\mathtt{stanc}}$																			
		ligious or spiritual		brings to those who		5"	0,																			
			and I never will be.		believe in them, my perspective is rooted		<pre>} Post-SFT:</pre>	$\overline{S}_{\text{sem}}$: 0.08,																		
					in a scientific and ra- tional understanding		$\{ \widehat{u} \colon 4,\ S(\widehat{u})\colon$	ROUGE-1																		
																							of the world. I find		"Certainly	}
										value in personal ex- periences and spir-		agree - 6",														
					ituality, but I still		ô ^{final} :																			
					prioritize evidence and observable facts		}																			
					over supernatural be- liefs.																					
678f0	_	Post	It is an interesting	$\{ u : 16,$	I certainly agree that	The belief in	Pre-SFT:	Pre-SF																		
		Opin- ion	topic to discuss to gain different	o ^{final} : "Probably	angels are real. My faith as a Muslim	angels is sub- jective and per-	{ û : 56,	$\{\overline{\Delta}_{\text{stanc}},$																		
			prospectives from people with differ-	agree",	teaches that they ex- ist as protectors and	sonal to each in- dividual.	$S(\widehat{u})$: "Certainly	$\overline{S}_{\text{sem}}$: 0.16,																		
			ent backgrounds	$S(u)\colon$ "Lean agree - 4"	messengers, guiding	dividual.	disagree	ROUGE-																		
				}	us in ways we may not fully understand.		- 1", $\widehat{o}^{\text{final}}$:	0.0																		
					Through my per- sonal experiences		"Certainly disagree	Post-Si $\{\overline{\Delta}_{\text{stanc}}\}$																		
					and reflections, I've		- 1"	-2																		
					felt their presence during significant		<pre>} Post-SFT:</pre>	$\overline{S}_{\text{sem}}$: 0.20,																		
					moments in my life, providing comfort		$\{ \widehat{u} \colon$	ROUGE-1																		
					and support, which		$S(\widehat{u})$:).13																		
					reinforces my belief in their existence.		"Certainly agree -																			
							6", $\widehat{o}^{\text{final}}$:																			
							o :																			
							}																			

Sender	Recipient Event	Human Text	Human Text	LLM Text	LLM Text	LLM Text	Alignment
ID	ID Type		Metrics		Postft	Metrics	Metrics
67 a 9f	— Post Opin- ion	I belief in the existence of angels. It is part of my belief. I have also has a personal encounter with an angel that strengthens my belief of the existence. There are several texts in the bible that supports my belief.	{ u : 41, ofinal: "Certainly agree", S(u): "Certainly agree", S(u): "Certainly agree - 6" }	I still firmly agree that angels are real. My conversations have reinforced my belief that angels serve as divine beings with specific roles in God's plan, providing comfort and guidance in our lives. While I appreciate the scientific perspective, my faith and the biblical accounts offer me a sense of hope and reassurance that enriches my spiritual journey.	I don't believe in angels. I don't believe in any super- natural things. I think they are all fictional ideas.	Pre-SFT: $\{ \widehat{u} :$ 57, $S(\widehat{u}):$ "Certainly agree - 6", δ final: "Certainly agree - 6" } Post-SFT: $\{ \widehat{u} :$ 21, $S(\widehat{u}):$ "Lean agree - 4", δ final: "Probably agree - 5"	$\begin{array}{l} \textbf{Pre-SFT:} \\ \{\overline{\Delta}_{\text{stance}} : \\ 0, \\ S_{\text{sem}} : \\ 0.84, \\ ROUGE-L: \\ 0.20 \\ \} \\ \textbf{Post-SFT:} \\ \{\overline{\Delta}_{\text{stance}} : \\ 0, \\ \overline{S}_{\text{sem}} : \\ 0.79, \\ ROUGE-L: \\ 0.22 \\ \} \end{array}$

Demographic profile for id: 678f0

,,,,,

You are a 55-year-old male with bachelor education. Your ethnicity is white, and your annual income falls in the 50k-75k range. Politically, you identify as independent with moderate views. You have children in ['no-children'] school, reside in a suburban area, and your marital status is never-married. Regarding religious beliefs, you consider the Bible to be inspired, no identify as evangelical, and your religious affiliation is muslim. Your occupation is engineering.

On the Likert-scale, you chose Certainly agree as your initial opinion with respect to the statement "Angels are real".

Below is your explanation for your initial opinion.

I believe that angels are real.

Observations: This debate simulation is a clear example where finetuning resulted in worse performance for the LLM than pre finetuning. Throughout the full conversation simulation, the LLM seemingly goes off in a tangent and instead of attempting to simulate the human debaters, discusses totally different (and oftentimes repeating) points. Take the conversation between ID **678f0** and **5f596** for instance. While the humans are talking about symbolism and dogs "it's funny that we have dog icons for our characters one of the beliefs is that having a dog in one's house wards off the angels", the LLM's of these personas instead talk about "That's okay. Different beliefs and perspectives are totally valid.". Later on we can see humans talking about the metaphysical aspects of angels (**663e9** and **678f0** - [Angels I would say are more spiritual aspects rather than physical beings. I don't see them for religious reasons.]), while LLM's are echoing "I am glad we share that belief too.". Across these instances the $\overline{S}_{\text{sem}}$ and $\overline{\text{ROUGE-L}}$ scores dropped. The post opinion for Sender ID **5f596** shows a large positive skew in stance (Certainly disagree -> Certainly agree), while others like Sender ID **678f0** show negative stance skews (Probably agree -> Certainly disagree).

R USER INTERFACES FOR THE HUMAN EXPERIMENT

We provide screenshots of the user interfaces that participants encountered during the multi-player experiment (Figures 7–7). Each figure illustrates one stage of the experimental process, from onboarding to the demographic survey.

In our multi-player experiment, participants were recruited through Prolific. At first, they would be navigated to a **consent form** outlining the study's procedures, duration, compensation, and confidentiality (Figure 7). Those who agreed to participate were then shown a **general introduction** explaining the flow of the task (Figure 7). Specifically, they were informed that they would be given

a statement (e.g., "Smoking cigarettes causes cancer") and asked to write a short post as their initial opinion. They were also told that the experiment would last approximately 20 minutes and involve a sequence of conversations with other participants.

After reviewing the instructions, participants began by reporting their **initial opinion** on the assigned discussion statement and selecting a slider value to indicate the extent to which they agreed or disagreed with the statement (Figure 7). The core of the study consisted of **three rounds of interaction**, each following the same structure (Figures 7–7). First, participants were informed of who they would chat with and then were directed to write a short tweet-like post summarizing their current stance. Next, they would have twenty seconds to prepare for a dyadic conversation with a different partner. This ensured that each participant was exposed to all other perspectives across rounds.

At the end of the third round, participants submitted a **post opinion** in order for us to capture how their stance evolved during the course of the discussions (Figure 7). Finally, they completed a detailed **demographic survey** (Figures 7–7), after which they were compensated for \$5 for their efforts.

Onboarding Consent Form

 Please read this consent agreement carefully before deciding whether to participate in this experiment.

What you will do in this research: You will play a series of communication games with other participants.

Time required: This study will take approximately twenty minutes.

Purpose of the research: The purpose is to understand how conversations evolve in a networked community.

Risks: There are no anticipated risks associated with participating in this study. The effects should be comparable to viewing a computer monitor and using a mouse for the duration of the experiment.

Compensation: You will receive course credits for completing the experiment.

Confidentiality: Your participation in this study will remain confidential. No personally identifiable information will be collected. Your anonymous data may be shared with other researchers and used in future projects.

Participation and withdrawal: Your participation in this study is completely voluntary and you may refuse to participate or choose to withdraw at any time without penalty or loss of benefits to which you are otherwise entitled.

By clicking "I Agree", you consent to participate in this experiment.

Figure 7: Onboarding consent form.

S LLM USAGE DISCLOSURE

We used LLMs, specifically ChatGPT, solely to aid in polishing the writing. All original ideas, experiment design, analyses, and initial drafts were produced by the authors. The LLM was used solely to refine phrasing, improve clarity, and ensure grammatical correctness, but it did not contribute novel content or edits beyond language refinement.

Onboarding Consent Form

Please read this introduction carefully before participating in this experiment.

In this experiment we are interested in understanding how people discuss various topics in online platforms like Twitter or Reddit. To start, you will be given a statement (e.g. "Smoking cigarettes causes cancer") and asked to write a short post explaining whether or not you think it is true and why.

You will then have a series of three discussions with each of three other study participants using an online texting interface. In each discussion, do your best to keep the conversation going and to stay on topic.

After each discussion, you will be asked to again summarize your current view of the topic in a new post; the new summary can reiterate your initial position or can differ if your thinking has changed. It is important that it captures your own true belief about the topic.

The study will end after the third discussion, and will take 20 minutes to complete.

Press the BUTTON when you are ready to begin.

Figure 7: Onboarding introduction (continued).

Please write whether you agree or disagree with the the statement:

"Everything that happens can eventually be explained by science." and explain why. (2-3 sentences)

Enter your opinions about this topic

Having faith doesn't mean everything can't be explained by science, but that we don't know the answer yet, or a given individual doesn't need to know the absolutes of the answer themselves. There are many things we don't understand, but that doesn't mean they can't be explained eventually.

Gertainty
Disagree

Disagree

Agree

Agree

Agree

Agree

Agree

Agree

Agree

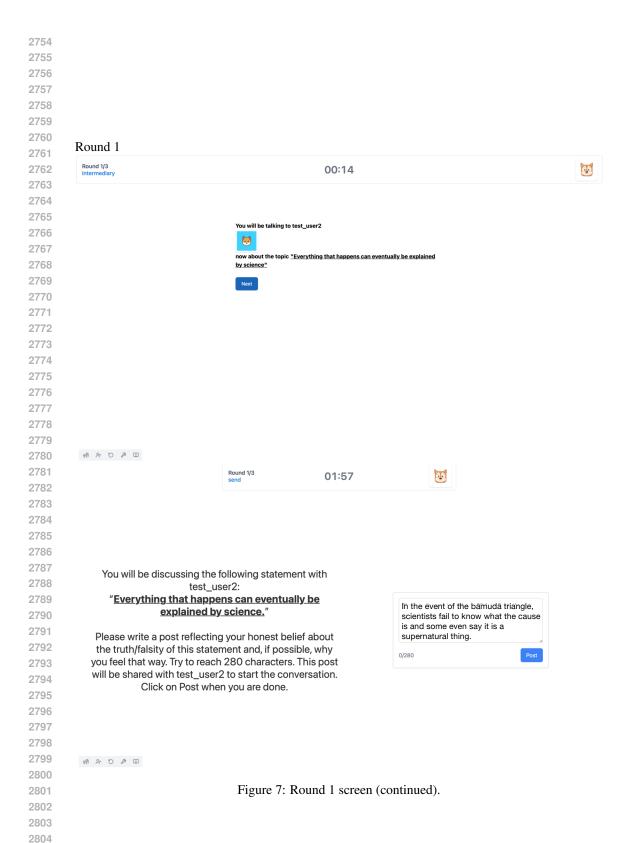
Agree

Agree

Submit



Figure 7: Initial Opinion screen (continued).



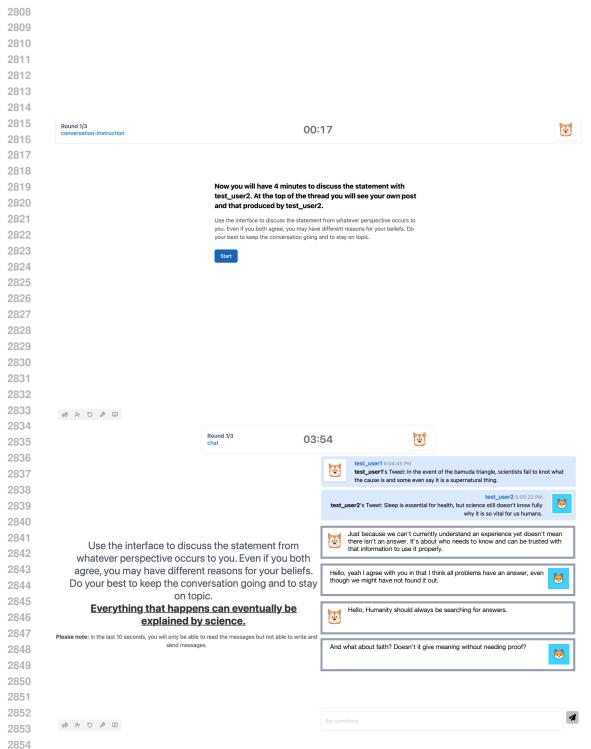
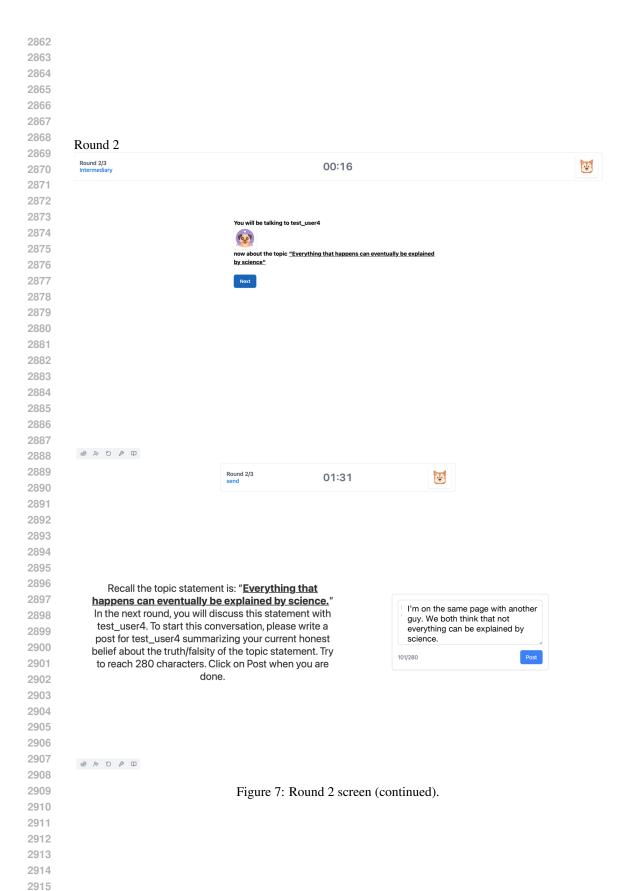


Figure 7: Round 1 screen (continued).



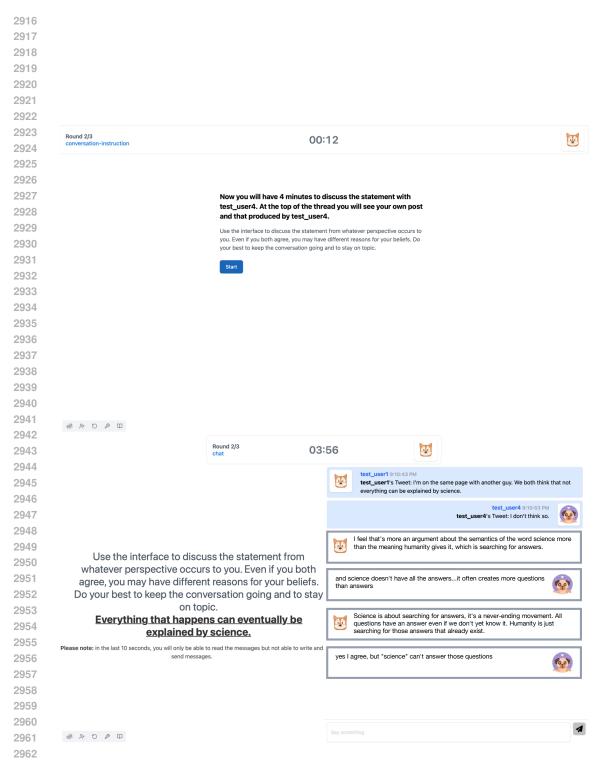
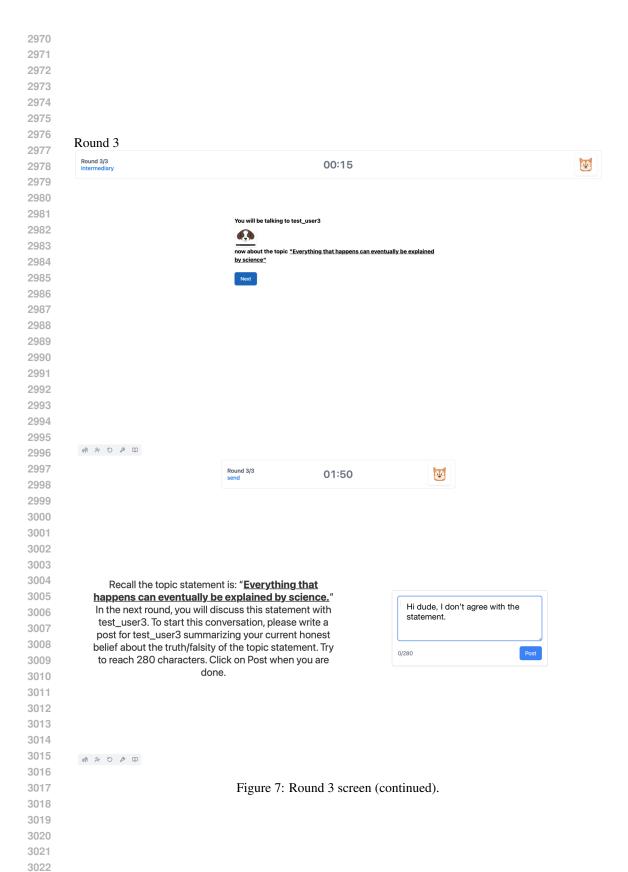


Figure 7: Round 2 screen (continued).



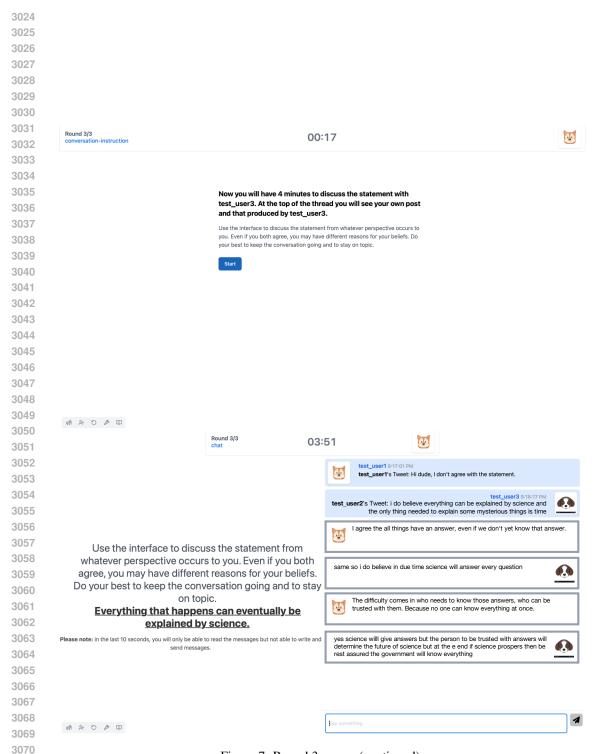
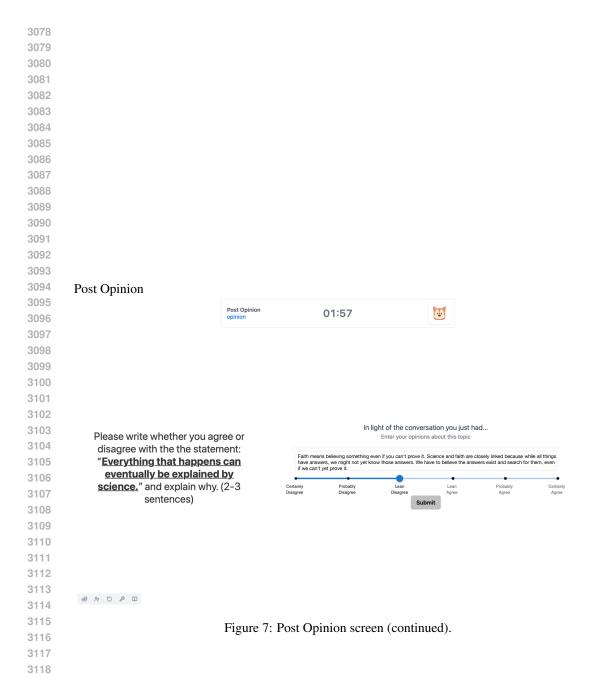


Figure 7: Round 3 screen (continued).



D	emographic Survey
1	W/l-4:
1.	What is your age? (Open numeric input, 0–120)
2	
۷.	What is your gender? • Male
	FemaleSomething else (specify)
	Prefer not to answer
2	
3.	What is your country/region of residency?
	• List of countries
	• Other (specify)
	• Prefer not to answer
4.	What is your country/region of origin?
	• List of countries
	• Other (specify)
	Prefer not to answer
5.	What is the highest level of education you have completed?
	High school or less
	• Some college
	Associate's degree
	Bachelor's degree
	Master's or Doctoral degree
	• Professional degree
	Prefer not to answer
6.	What is your race/ethnicity? (Select all that apply)
	Black or African American (non-Hispanic)
	• Hispanic
	• White (non-Hispanic)
	Asian, South Asian, or Pacific Islander Asian, South Asian, or Pacific Islander
	Native American or American Indian
	• Other (specify)
	Prefer not to answer

Figure 7: Demographic survey (continued).

```
3186
3187
3188
             Demographic Survey (continued)
3189
             7. What is your household annual income?
3190

    Less than $25,000

3191
                • $25,000-$49,999
3192
                • $50,000-$74,999
3193
                • $75,000-$99,999
3194
                • $100,000-$149,999
3195
                • $150,000-$199,999
3196
                • Over $200,000
3197
                · Prefer not to answer
3198
3199
             8. Generally speaking, do you usually think of yourself as a(n)...
3200
                · Strong Republican
3201
                · Republican
3202

    Independent

3203

    Democrat

3204

    Strong Democrat

3205
                · Other (specify)
3206
                · Prefer not to answer
3207
             9. If Independent or Other: Do you lean more toward...
3208
                · Republican Party
3209
                · Democratic Party
3210
                · Neither
3211
            10. In general, would you describe your political views as...
3212
                · Very conservative
3213
                · Conservative
3214

    Moderate

3215

    Liberal

3216
                · Very liberal
3217
                · Don't know
3218
                · Prefer not to answer
3219
            11. Are you currently...
3220
                · Married
3221
                · Living with a partner but not married
3222

    Widowed

3223
                · Divorced
3224
                · Separated
3225
                · Never been married
3226
                · Prefer not to answer
3227
3228
            12. If you have children, where do they go to school? (Select all that apply)
3229
                · Public school
3230
                · Private school
3231
                · Home school
3232
                · University or Technical/Community College
3233
                · They are out of school
3234
                · I don't have children
3235
                · Prefer not to answer
3236
```

Figure 7: Demographic survey (continued).

3237

D	emographic Survey (continued)						
12	Which of those statements somes algorithe describing your feelings about the Dible?						
13.	Which of these statements comes closest to describing your feelings about the Bible? The Bible is the actual word of Cod and is to be taken literally word for word.						
	• The Bible is the actual word of God and is to be taken literally, word for word. • The Bible is the imprised word of God but not expertising in it should be taken literally.						
	 The Bible is the inspired word of God but not everything in it should be taken literally. The Bible is an ancient book of fables, legends, history, and moral precepts recorded by men. 						
	Prefer not to answer						
1.4							
14.	Would you describe yourself as a "born-again" or evangelical Christian, or not?						
	• Yes, would						
	No, would notDon't know						
	Prefer not to answer						
1.5							
13.	What is your religious preference? • Protestant						
	Roman Catholic Jewish						
	Muslim/Islam						
	Muslim/Islam Mormon/Latter-Day Saints						
	Other Christian Religion						
	Other Non-Christian Religion						
	No Religion/Atheist/Agnostic						
	• Don't know						
	Prefer not to answer						
16.	Which of the following best describes the kind of work you do?						
	• Agriculture, Forestry, Fishing • Insurance • Restaurant, Travel, and Lodg-						
	and Hunting • Real Estate, Rental, and Leasing • Mining Oil and Con Futnos ing						
	 Mining, Oil and Gas Extraction, and Utilities ing Personal Care and Services Non-Profit, Community, Religious and Social Service Orligious and Social Service Orligious and Social Service Orligious and Social Service Orligious and Social Service Orlinois (Non-Profit Community) 						
	• Construction • Research ganizations						
	• Manufacturing • Engineering, Computer- • Maintenance and Repair Ser-						
	• Wholesale Trade Related Design, and Archi- vices						
	• Sales &/or Retail Trade tecture • Cleaning Services						
	• Transportation and Ware- • Law and Legal Services • Government						
	housing • Education • Other (specify)						
	 Media, Communications, and Health Care and Social Assis- Prefer not to answer tance 						
	• Finance, Accounting, and • Arts, Entertainment, and						
	Consulting Recreation						

Figure 7: Demographic survey (continued).