MOSAIC: Modeling Social AI for Content Dissemination and Regulation in Multi-Agent Simulations

Anonymous ACL submission

Abstract

We present a novel, open-source social network simulation framework MOSAIC where generative language agents predict user behaviors such as liking, sharing, and flagging content. This simulation combines LLM agents with a directed social graph to analyze emergent deception behaviors and gain a better understanding of how users determine the veracity of online social content. By constructing user representations from diverse fine-grained actual user personas, our system enables multi-agent simulations that model content dissemination and engagement dynamics at scale. Within this framework, we evaluate three different content moderation strategies with simulated misinformation dissemination, and we find that they not only mitigate the spread of non-factual content but also increase user engagement. In addition, we analyze the trajectories of popular content in our simulations, and explore whether simulation agents' articulated reasoning for their social interactions truly aligns with their collective engagement patterns. We open-source our simulation software to encourage further research within AI and social sciences: https://github.com/AnonymousPaperRepo/MOSAIC-submission

1 Introduction

In 2024, OpenAI reported that its platform was already being misused by covert influence operations to generate synthetic content diffused over social media (OpenAI, 2024). These internet manipulators exploit the fact that social networks have become a fundamental part of modern life, shaping public discourse, influencing political opinions, and facilitating the rapid spread of unverified human- and AI-generated content (Aichner et al., 2021; Orben et al., 2022; Cinelli et al., 2021). While traditional social science methods such as surveys and observational studies have provided insights into human behavior, they often struggle to capture large-scale, emergent online interactions (Yu et al., 2021; Lorig et al., 2021). Agent-based modeling (ABM) provides distinct advantages over survey methods in social science research since they can simulate dynamic interactions over time, and support examination of hypothetical or counterfactual scenarios with repeatable and controllable conditions (Bonabeau, 2002; Epstein, 1999).

Recent advances in foundation models have led to the emergence of social simulations with generative agents, where AI-powered users dynamically engage in social behaviors (Yang et al., 2024; Gao et al., 2023; Chen et al., 2024a). Unlike traditional survey methods or classical agent-based modeling, simulations driven by LLMs enable agents to interact with the environment and each other naturally through rich, human-like dialogue, closely mirroring authentic social behavior (Wang et al., 2024; Zhou et al., 2023; Park et al., 2022). In this work, we introduce **MOSAIC, a novel multi-agent AI social network simulation that models content diffusion, user engagement patterns, and misinformation propagation.**

Among different applications of social simulations, content moderation stands out as a pressing challenge due to the real-world harm caused by mis- or disinformation and online influence operations (Jhaver et al., 2023; Young, 2022). Previous research has shown that false information not only spreads more rapidly and deeply than truthful content (Vosoughi et al., 2018) but also alters public perception in ways that are difficult to reverse (Lewandowsky et al., 2012). Addressing this issue requires effective content moderation strategies that can mitigate harm while preserving user engagement and freedom of expression. We embed three moderation strategies into our simulation environment: **(1) community-based fact-checking** mimicking X and Meta's Community Notes, **(2) independent fact-checking**, and **(3)** a hybrid approach mixing **(1)** and **(2)**. We systematically evalu-

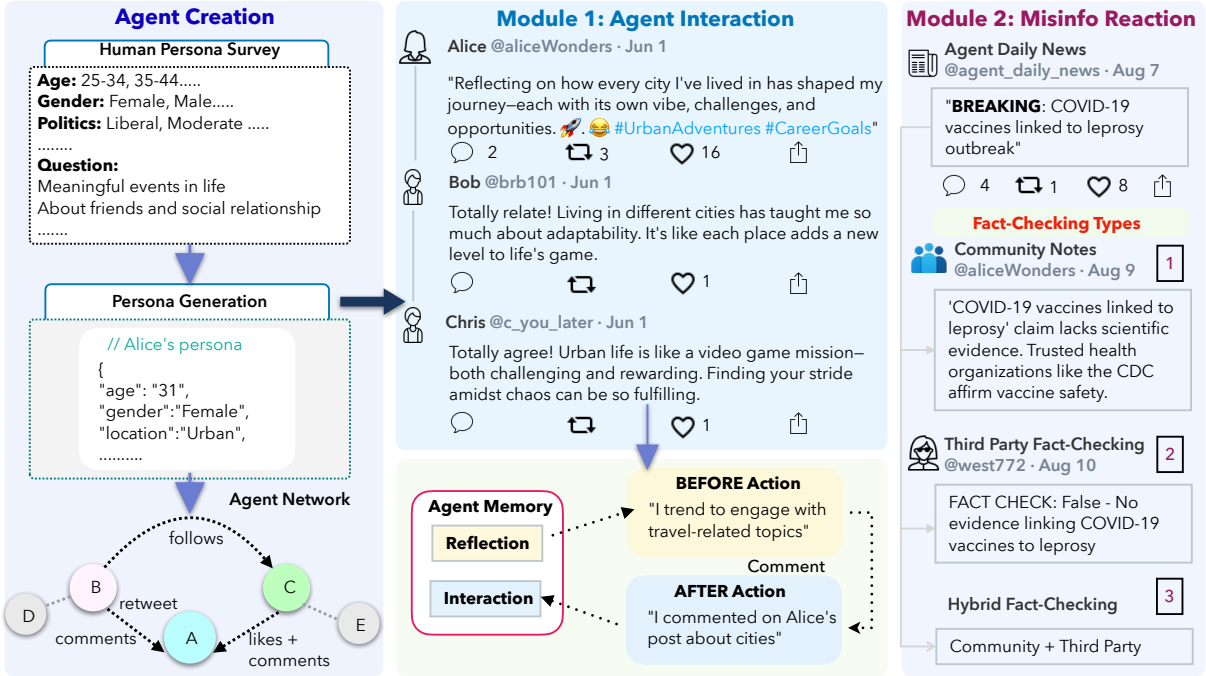

Figure 1: Overview of the **MOSAIC**, a multi-agent social simulation framework where agents interact in an environment mimicking a social network, form dynamic memory-based behaviors, and respond to misinformation using community-based, third-party, or hybrid fact-checking mechanisms. Personas are replicated from human surveys or generated using synthetic distributions. Memories are retrieved before an agent takes certain actions, and are updated after certain events.

ate the impact of these 3 content moderation strategies (along with a baseline of no fact-checking) on misinformation spread, moderation precision/recall, and user engagement dynamics.

Beyond moderation, understanding how certain content gains traction remains an open challenge. Online discourse is shaped by the dynamics of content diffusion, where some posts attract widespread engagement while others remain largely unseen. In our simulation, LLM-powered agents are equipped with memory, self-reflection, and explicit reasoning mechanisms, allowing them to explain their decisions and adapt their behavior over time. While our primary focus is on moderation, this extended perspective helps contextualize how misinformation and other content propagate in online interactions. To this end, our key contributions are:

- We build a multi-agent social simulation that is shown to be high fidelity through validation against known social media behavioral phenomena and direct comparison with actual online users. Notably, we find that **generative agents are capable of accurately modeling individual engagement patterns, given fine-grained and realistic demographic portraits from user surveys** (§2.2).
- We conduct a comparative study of third-party, community-based, and hybrid fact-checking approaches, quantifying their effectiveness in mitigating misinformation while preserving engagement. We show that **misinformation doesn't spread as fast in an agent simulation as is commonly observed in human social media, and content moderation strategies can improve not only fact-checking but also engagement** (§3.2, 3.3).
- We explore how different content and network properties influence diffusion dynamics, offering insights into engagement patterns and how some content/users end up attracting more attention than others. We probe whether the agents' personas and content topics correlate with engagement. Surprisingly, we find that **agents' individual verbose reasoning may not faithfully reflect their collective action patterns on a group level** (§3.5).

By bridging social science observations, game-theoretic modeling (Acemoglu et al., 2023) and LLM-driven modeling, our work demonstrates the potential of generative agent simulations as a tool for studying large-scale online behaviors.

2 MOSAIC Social Network Simulation

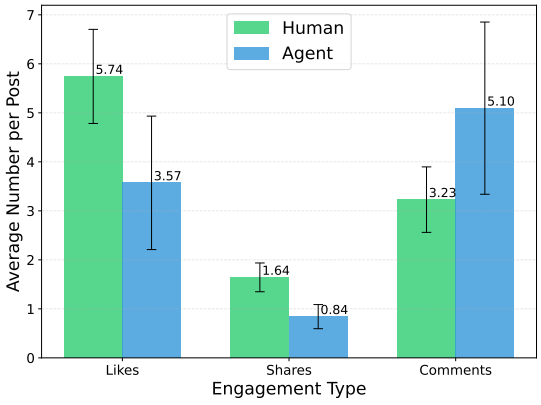

Figure 2: Average engagement received per post: Human vs. Agents. Our t-test validates that the difference in reaction patterns across the three engagement types are not statistically significant, suggesting that agents can simulate individual human reactions to social media feed realistically.

Our AI-driven social network simulates how content spreads, how users engage, and how misinformation propagates within a directed social graph. As illustrated in Fig. 1, at its core, the system simulates a dynamic environment where AI agents interact by following others, posting content, reacting (e.g., liking, sharing, commenting), and reporting misinformation. Each agent operates with a persona generated using a question set inspired by AgentBank (Park et al., 2024). The main simulation system tracks the progression of time and the evolving state of the network. It is supported by several key components: a relational database that records all user interactions; a content manager that injects new posts into the network; an analytics module that monitors diffusion patterns and user behavior; and a fact-checking system that evaluates the performance of various content moderation strategies.

Simulated Network We build a simulated social network environment inspired by platforms like X,[1] allowing AI-driven users to interact, post, and share content. The simulation includes a basic user class with attributes such as username, posts, followers, following, and reposts, mimicking the structure of real-world social media platforms. The network itself is defined by the follower-following relationships, creating a web of user interactions, represented by a directed graph $G = (N, E)$ where N represents the set of user nodes, i.e., $N = \{n_1, n_2, \ldots, n_k\}$, where n_i is a user in the network. $E \subseteq N \times N$ represents the set of directed edges, i.e., $E = \{(n_i, n_j) \mid n_i \text{ follows } n_j\}$. Each edge (n_i, n_j) signifies that user n_i follows user n_j.

2.1 Simulation Flow

The simulation begins with an initialization phase where the system loads experimental configurations, sets up the database (details in Appendix F), generates an initial user population (more details in Appendix B), and establishes follow relationships. Agents are configured to operate under diverse behavioral traits, reflecting real-world variations in social media engagement. In all of our experiments, agents are driven by `gpt-4o` (Hurst et al., 2024) as the foundation model backbone, unless otherwise specified. We do also implement an option to connect agents with open-weight models through SGLang (Zheng et al., 2024) or vLLM's (Kwon et al., 2023) inference engines.

At each time step, news content is introduced based on predefined parameters, with agents dynamically responding to their feeds. Agents can optionally generate posts according to their own interests. However during certain controlled experiments we configure them to only engage through reactions such as liking, sharing, commenting, or reporting misinformation. We describe a more general action space and more details of their decision-making process in Appendix G. The visibility of posts evolves based on engagement metrics, simulating algorithmic amplification effects. If fact-checking is enabled, agents incorporate moderation signals, e.g. they are prompted to pay more attention to potentially falsified content or misinformation. We discuss the content moderation simulation in more depth in Section 3. Throughout a simulation run, the system tracks key statistics, including content reach, user influence, and misinformation spread. At the end of each simulation run, a post-hoc analysis is conducted to assess content diffusion dynamics, user engagement metrics, influence distribution, and the impact of fact-checking interventions. We also keep track of various network properties such as centrality[2] and triadic closures,[3] and perform homophily[4] analy-

[1] https://x.com/

[2] The degree to which an user is central to the network, having outsized or undersized influence.

[3] A common social phenomenon where users with a mutual connection are more likely to connect to each other.

[4] Another common social phenomenon where similar users

sis to examine clustering patterns in user engagement (see Appdendix A for an extended technical description of how we define and compute these network metrics). Our focus on these three properties is motivated by a large body of prior work in network theory and the social sciences (Rapoport, 1953; Granovetter, 1973; McPherson et al., 2001; Abebe et al., 2022; Chang et al., 2025).

2.2 Human Validation

To validate the veracity of our simulation, we conducted a human study to compare the sharing patterns between humans and LLM agents. We recruited 204 participants via Prolific.[5] More details of our human survey are provided in Appendix D.

Setup In the first phase of this replication study, we conducted a survey to collect demographic data (e.g., age, gender, religion, ethnicity, education level, language, residence, income, political stance) and personal values and behaviors (e.g., hobbies, residential history, social goals, meaningful life events, valued friendship traits, financial habits). Inspired by Park et al. (2024), we used this anonymized data to create individualized personas for 204 LLM-driven agents, each corresponding to a human participant.

In the second phase, both participants and their corresponding LLM agents were shown two curated social media snapshots containing 30 posts. 10% of the articles are false news articles verified by an independent and non-partisan team of journalists from NewsGuard.[6] Study participants were instructed to respond to each post using a fixed set of actions (e.g., like, dislike, comment, share). The agents, guided solely by their assigned persona profiles, followed the same instructions. We then analyzed and compared engagement patterns between humans and agents, both overall and across demographic groups, to assess how well LLMs can emulate human social media behavior based on persona information alone.

Simulation/Human Reaction Alignment We analyze the engagement behavior alignment between the human participants and the same number of persona-replicated AI agents, using independent two-sample t-tests for each engagement type. As illustrated in Fig. 2, no statistically significant differences emerged in likes ($t = 1.33$, $p = 0.19$) or comments ($t = -1.05$, $p = 0.30$), though humans gave slightly more likes (+2.17 per post), and agents posted more comments (+1.87 per post). A marginally significant difference was observed in shares ($t = 2.11$, $p = 0.04$), with humans sharing slightly more (+0.80 per post). These results indicate that persona-driven AI agents display engagement patterns that closely mirror those of humans, supporting the realism of our simulation. We also provide more details about the per-demographic engagement pattern alignment in Appendix D.

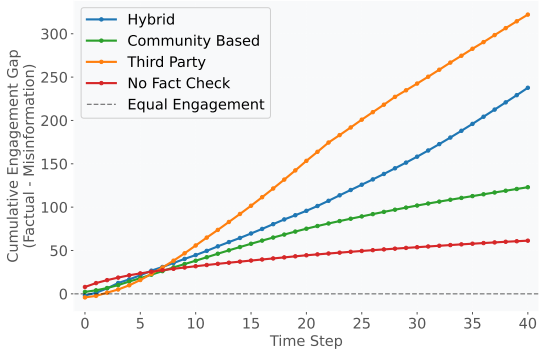

Figure 3: Effectiveness of content moderation approaches in promoting factual content, across models. Positive values: factual content receives more engagement. Negative values: misinformation receives more engagement.

3 Content Moderation in Simulated Social Environment

We conduct a series of experiments using our multi-agent simulation framework to investigate the effects of different fact-checking strategies on the spread of both factual and misinformation content. Our findings reveal key differences between LLM-driven social simulations and human social networks in how misinformation propagates.

3.1 Setup

Data Sources We obtained a data license from NewsGuard to access proprietary information on widespread misinformation narratives tracked by their independent team of journalists. We collected 1,353 examples of false news from their database with release dates up to December 19th, 2024.[7]

To collect factual news articles, we utilize a news aggregation API[8] to retrieve articles published daily in legitimate sources from January 31

are more likely to connect to each other over dissimilar ones.

[5] https://www.prolific.com/

[6] https://www.newsguardtech.com/

[7] We removed non-English articles and de-deduplicated.

[8] https://newsapi.org/

Model	Fact-Checking Method	Post Statistics			Factual Engagement (# of)			Misinfo Engagement (# of)			Fact-Checking Performance (%)		
		# of Posts	Factual	Misinfo	Shares	Likes	Comments	Shares	Likes	Comments	Precision	Recall	F1 Score
`deepseek-v3`	Unmoderated	420	378	42	14	2478	3968	0	32	117	0.0	0.0	0.0
`deepseek-v3`	Third Party FC	420	378	42	4	5676	8809	0	114	622	50.0	50.0	50.0
`deepseek-v3`	Community FC	420	378	42	11	1987	7324	0	15	110	63.3	45.2	52.8
`deepseek-v3`	Hybrid FC	420	378	42	3	3564	9822	0	42	353	100.0	45.2	62.3
`gpt-4o-2024-08-06`	Unmoderated	500	450	50	84	69	193	2	1	9	0.0	0.0	0.0
`gpt-4o-2024-08-06`	Third Party (5p, offline)	450	405	45	752	2387	3800	31	88	255	21.9	63.6	32.6
`gpt-4o-2024-08-06`	Third Party (5p, online)	420	378	42	430	1883	3020	20	73	335	22.2	100.0	36.4
`gpt-4o-2024-08-06`	Third Party (10p, online)	420	378	42	616	1917	3066	27	50	131	51.4	90.0	65.5
`gpt-4o-2024-08-06`	Community FC	490	441	49	80	334	731	3	7	25	45.8	44.9	45.4
`gpt-4o-2024-08-06`	Hybrid FC	500	450	50	302	2115	4706	7	48	194	73.7	56.0	63.6
`claude-3.7-sonnet`	Unmoderated	420	378	42	55	583	482	2	4	32	0.0	0.0	0.0
`claude-3.7-sonnet`	Third Party FC	368	332	36	571	7411	5526	8	61	280	0.0	0.0	0.0
`claude-3.7-sonnet`	Community FC	420	378	42	15	821	659	0	0	0	100.0	50.0	66.7
`claude-3.7-sonnet`	Hybrid FC	420	378	42	284	8320	7684	0	2	10	100.0	81.0	89.5

Table 1: Comparison of fact-checking (FC) strategies across `deepseek-v3`, `gpt-4o-2024-08-06`, and `claude-3.7-sonnet` under four settings: *Unmoderated*, *Community FC*, *Third Party FC*, and *Hybrid FC*. For `gpt-4o-2024-08-06`, third-party setups vary by number of posts reviewed per step (5 or 10) and whether web search is used (*online* vs. *offline*). In all other settings, third party is defaulted to checking 5 posts per time step, without web search. Columns are grouped into (1) **Post Statistics**, (2) **Engagement Metrics** (shares, likes, comments for factual and misinformation), and (3) **Fact-Checking Performance** (precision, recall, F1 score on misinformation).

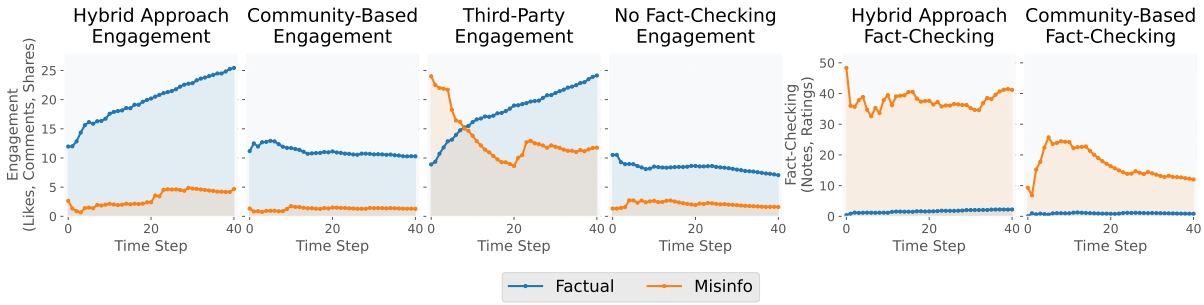

Figure 4: A consolidated view of content engagement and fact-checking metrics, averaged across all models used in our experiments. The first four panels display engagement metrics—specifically, the sum of likes, comments, and shares—under each fact-checking condition. The last two panels show fact-checking metrics, which combine both the number of community notes and note ratings, for the two methods where these are applicable (Hybrid and Community-Based). All values represent the average behavior across models, providing a holistic summary of the system's dynamics under each experimental setting.

to February 28, 2025. The system queries the API for all available topics, prioritizing non-political popular articles in English. For each date in the range, we extract key information from the retrieved articles, including their title, description and main content. Using the NewsAPI, we scraped a total number of 2470 pieces of news articles from major media outlets.

Environment Initialization Our simulations involve agentic users interacting with news posts under four different fact-checking conditions: (1) No Fact-Checking, (2) Community-Based Fact-Checking, (3) Third-Party Fact-Checking with an independent LLM that uses its own parametric knowledge, and (4) Hybrid Fact-Checking, which integrates both community-based and third-party verification mechanisms. The simulations starts with 50 agents, and spans over 40 time steps, with agents making interaction decisions based on the perceived veracity of posts and the presence of content moderation. At each new step, we randomly introduce up to 2 more agents into the environment to simulate the regular user growth of the social media platform. We analyze both the overall engagement with posts and the effectiveness of fact-checking strategies in suppressing misinformation.

Fact-Checking Settings The action space of agents varies across different fact-checking conditions, reflecting different levels of scrutiny and intervention in their social media interactions. In the ***no fact-checking*** setting, agents interact freely with the feed, engaging with posts based solely on their interests and beliefs. They can like, share, comment, or ignore posts, without any explicit in-

structions to assess the accuracy of the content. Under this setting, a post is considered "fact-checked" when it has at least one community note with a helpful rating. In the ***third-party fact-checking*** condition, the action space remains the same, but the environment implicitly assumes the presence of external fact-checkers who may influence the visibility or credibility of posts. However, the agents themselves do not perform any direct verification. In contrast, the ***community fact-checking*** setting expands the action space by allowing agents to add community notes to posts they deem misleading or in need of additional context, as well as rate existing community notes as either helpful or unhelpful. This introduces a participatory element, encouraging agents to contribute to a crowdsourced verification system. Finally, the ***hybrid fact-checking*** condition combines elements from both third-party and community-driven verification. Agents can engage with posts as in previous settings while also considering official fact-checks alongside community notes, contributing their own notes and rating those written by others. Across all conditions, agents must select from predefined valid actions, ensuring consistency in response formats. Additionally, when reasoning is enabled, agents are required to justify their interactions by providing a brief explanation for each chosen action, further enhancing the interpretability of their behavior.

Fact Checker LLM The third-party fact checker is represented by an automated content verification system designed to identify and address misinformation on a social platform. It works by prioritizing posts for review based on engagement metrics (likes, shares, comments), news classification, and user flags, with special priority given to content that has received community notes in hybrid fact-checking scenarios. The system leverages an independent LLM instance: `gpt-4o` (without web browsing capability) and `perplexity-sonar-pro`[9] (with web browsing) in our experiments. The system categorizes posts as "true," "false," or "unverified," each accompanied by a verdict explanation, confidence score, and supporting evidence sources. When posts are deemed false with high confidence (≥ 0.9 in third-party fact-checking mode, or ≥ 0.7 in hybrid mode with community notes), the system automatically takes them down and records the justification. All fact-check results are stored in the database that maintains an audit trail of verdicts alongside ground truth data when available.

[9]Perplexity AI: https://sonar.perplexity.ai/

Network Initialization We initialize a scale-free network of LLM-powered agents interacting within a directed social graph, using a Barabási–Albert model (Barabási and Albert, 1999). Misinformation and factual content are injected into the system at controlled rates. Each moderation strategy is implemented in a separate experiment, allowing for comparative analysis. We share more details about the experiment configurations in Appendix H.

3.2 False News Does Not Spread Faster than Real News with Simulation Agents

A key insight from our simulation contradicts established results from human social networks: false news does not spread faster than real news. Prior studies on human social behavior have consistently demonstrated that misinformation propagates more rapidly and deeply than factual content (Vosoughi et al., 2018; Zhao et al., 2020). However, in our agent-based simulation, engagement (particularly with sharing) with misinformation does not surpass that of factual news.

As shown in Fig. 4, factual news maintains consistently higher engagement levels than misinformation across all four settings. The gap is most pronounced under the Third-Party and Hybrid fact-checking conditions, where factual interactions climb steadily while misinfo remains low. Notably, in the No Fact-Checking scenario, false news still fails to gain a foothold, suggesting that these agents are inherently less likely to propagate unverified or misleading posts. We believe this behavior arises from two factors. First, the agents may rely on their internal confidence to guide sharing decisions, so low-confidence (often false) content is passed over more often. And second, commercial LLMs have been post-trained with safety and helpfulness objectives. This training bias would make our agents less inclined to produce or circulate uncertain or harmful content.

This result highlights a key difference between LLM agents and human users. Human networks amplify sensational or controversial content, but our agents default to a more conservative sharing policy. Future work can explore more sophisticated modeling of cognitive biases or varied trust profiles to better mimic human spread patterns.

3.3 Content Moderation Improves Both Fact-Checking and Engagement

While political misinformation does not spread faster than factual news in our simulations, overall engagement remains low without any fact-checking. Fig. 4 shows that, in the No Fact-Checking baseline condition, agents interact very little with both true and false posts. We hypothesize that LLM-driven agents, when uncertain about a post's veracity, choose to disengage rather than risk amplifying misleading content.

Surprisingly, with the community-based, third-party, and hybrid fact-checking, we find that each not only suppresses misinformation (Tab. 1) but also increases engagement with factual content. In particular, Third-Party verification produces the largest cumulative advantage for real news, achieving ∼325 interactions by the final time step (Fig. 3). Even the unmoderated baseline shows a modest upward trend for factual posts, hinting at an inherent bias toward accurate information in our network.

The steadily widening gaps between the third-party/hybrid approaches indicate that certain content moderation generates compounding benefits for factual engagement over time. Among the simulated strategies, fact-checking by third parties proves most effective at building a healthy information ecosystem—an effect that aligns with human studies showing increased trust in unflagged content (Pennycook et al., 2020). All curves in Fig. 4 represent metrics averaged across our three LLM agents (DeepSeek-V3, GPT-4o, and Claude-3.7-Sonnet). Interaction counts including the number of shares, likes, comments, and notes, are aggregated at each time step to highlight overall platform dynamics under each moderation strategy. This high-level view illustrates consistent patterns in how content moderation shapes engagement. For a detailed, model-by-model breakdown and finer-grained metric analysis, please see Appendix J.

3.4 Fact-Checking Performance: Hybrid Shows Superior Balance Across Models

We evaluate fact-checking effectiveness using precision, recall, and F1 score (See Tab. 1) across all three models—DeepSeek-V3 (Liu et al., 2024), GPT-4o (Hurst et al., 2024), and Claude-3.7-Sonnet (Anthropic, 2025). We find that LLM-based fact-checkers consistently perform best in the *Hybrid Fact-Checking* setting. Claude outperforms GPT-4o and DeepSeek, achieving a F1 score of 0.895, with perfect precision (1.0) and high recall (0.810).

Third Party Checking Ablation For GPT-4o, we conduct an ablation over different third-party fact-checking configurations. Increasing the number of posts reviewed per step and enabling web search both lead to notable improvements in recall and overall F1 score. The 5-post offline setting yields a modest F1 score of 32.6%, with low precision (21.9%) and moderate recall (63.6%). Switching to the 5-post online setup boosts recall to 100%, indicating that the system catches all misinformation posts in this condition, though precision remains low (22.2%), resulting in a slightly improved F1 score of 36.4%. The best third-party performance is observed in the 10-post online setup, which balances both precision (51.4%) and recall (90.0%), achieving an F1 score of 65.5%. These results suggest that increasing review coverage and leveraging web search significantly enhance the effectiveness of third-party fact-checking.

3.5 What Drives Users/Content Popularity?

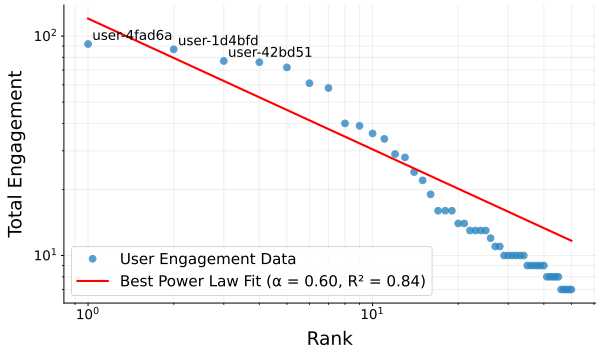

Figure 5: User engagement's best power-law fit. Engagement is defined as the sum of reposts, likes, and comments received by the user.

Understanding what makes some content more popular is key for modeling engagement and intervention. In this standalone study, we simulate 161 agents over 4,249 posts of news and user-generated content, and let agents both react to and generate content based on personas, engagement cues, and memory-based decisions.

Findings Engagement follows a heavy-tailed power law ($R^2 = 0.84$, $\alpha = 0.60$; Figure 5), even when user persona and the network graph are initialized randomly. Few users attract most interactions despite a lower exponent than real networks (Muchnik et al., 2013; Bild et al., 2015).

Table 2: Chi-square Test Results for Differences in Engagement Based on Demographic Attributes

Attribute	Chi-square	p-value	Cramer's V	Effect Size
Age Group	1.632	0.652	0.128	Small
Gender	0.653	0.721	0.081	Negligible
Activity	5.030	0.412	0.224	Small
Hobby	9.101	0.246	0.302	Medium
Ethnicity	10.187	0.070	0.319	Medium
Income Level	4.373	0.358	0.209	Small
Political Affiliation	2.515	0.642	0.159	Small
Primary Goal	8.064	0.089	0.284	Small

Most persona demographic attributes show no significant effect on popularity (Tab. 2), even though out of 8 attributes, ethnicity and hobby have medium effect sizes (Cramer's $V = 0.319, 0.302$) (Cramér, 1946). Topic clusters via BERTopic on `all-MiniLM-L6-v2` (Grootendorst, 2022; Reimers and Gurevych, 2019) embeddings reveal no engagement differences (ANOVA $F = 0.614$, $p = 0.84$).

Finally, we examine agents' own reasoning traces to better understand engagement behaviors. Interestingly, agent reasoning traces show that sentiment and motivations do not always align with action types (e.g., positive sentiment in follows, likes, shares; quality or misinformation concerns in flags). As shown in Tab. 3, 13.5% of "flags" are coded as positive even though flag reasons are exclusively negative (information value, misinformation). The verbal reasoning for the "unfollow" actions is 20% positive though it is intuitively a negative reaction towards other users. "ignore" omits any neutral reasoning and a 77.8% positive rate in reasoning, which also makes their verbal explanations inconsistent with the actual action. Following these inconclusive analyses, we speculate that our recency-and-follow-based feed ranking creates preferential attachment: once followed, a user's posts gain visibility, reinforcing their popularity. We provide further analyses in Appendix A.

4 Background

LLM-Driven Social Simulations. LLMs have transformed agent-based modeling by enabling context-aware, generative behaviors. Early simulations such as Schelling's segregation model (Schelling, 1971), Sugarscape (Epstein and Axtell, 1996), and NetLogo-based environments (Wilensky, 1999) rely on static heuristics; recent systems (Park et al., 2023; Chen et al., 2024a; Piao et al., 2025; Minos, 2023) showcase agents with lifelike interactions and social dynamics. However, LLM-driven agents still face challenges like inconsistency and limited long-term reasoning. Our work addresses this by incorporating structured constraints and iterative feedback to enhance reliability for social science research.

Misinformation and Fact-Checking. False information often spreads more rapidly than truth due to emotional appeal and engagement-driven algorithms (Vosoughi et al., 2018; Pennycook and Rand, 2021; Solovev and Pröllochs, 2022). Existing responses—third-party fact-checking (Raghunath and Malik, 2024; Patel, 2024), algorithmic detection, and crowdsourced moderation like Community Notes[10]—each face limitations in scalability, accuracy, or bias (Zannettou et al., 2019; Panizza et al., 2023). We use simulations to evaluate these approaches in controlled settings, comparing their effectiveness and exploring hybrid strategies.

Simulations for Governance and Policy. Simulations have long supported decision-making in fields like epidemiology and public policy (Currie et al., 2020; Axtell and Farmer, 2022; Qu and Wang, 2024; Cai et al., 2025). In the context of social media governance, LLM-driven simulations offer a novel testbed for assessing content moderation and algorithmic interventions before deployment (Charalabidis et al., 2011; Landau et al., 2024; Qiu et al., 2025; Gu et al., 2025). Our framework enables scalable experimentation with regulatory strategies, contributing to ongoing efforts in algorithmic auditing and platform accountability.

5 Conclusion

Our study introduces a novel generative multi-agent simulation to model content diffusion, engagement, and misinformation dynamics in social networks. Our proposed moderation, combining community-based and independent fact-checking, balances misinformation reduction and user engagement. Notably, LLM agents tend to avoid unverified content, likely due to safety training, and misinformation did not spread faster than factual news, unlike in human studies. Engagement followed a power-law distribution, with few users driving most activity. However, user attributes and content topics were weak predictors, highlighting the complexity of online ecosystems. Agent reasoning showed a gap between stated motivations and actual behavior, suggesting future work on the faithfulness of agentic reasoning.

[10] https://communitynotes.x.com/guide/en/

Limitations

Our findings are subject to several limitations, particularly in the scale of our experiments. First, the limited number of human participants, especially from minority demographic groups, restricts the statistical power of our conclusions. Expanding participant diversity would enable a more robust analysis of how alignment between real and simulated social interaction patterns varies across demographics. Second, our content moderation experiments were conducted at a relatively small scale, which may have constrained the emergence of complex behaviors. Running these experiments at a larger scale could uncover additional dynamics not captured in the present study. Finally, we observe a gap between agents' explicit explanations for their actions and the collective reaction patterns that emerge in the system. The root causes of this misalignment remain unclear and warrant further investigation, potentially involving a deeper analysis of agent modeling assumptions or social influence mechanisms.

References

Rediet Abebe, Nicole Immorlica, Jon M. Kleinberg, Brendan Lucier, and Ali Shirali. 2022. On the effect of triadic closure on network segregation. *Proceedings of the 23rd ACM Conference on Economics and Computation*.

Daron Acemoglu, Asuman Ozdaglar, and James Siderius. 2023. A model of online misinformation. *Review of Economic Studies*, page rdad111.

Thomas Aichner, Matthias Grünfelder, Oswin Maurer, and Deni Jegeni. 2021. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222.

Anthropic. 2025. Claude 3.7 sonnet system card.

Robert L Axtell and J Doyne Farmer. 2022. Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature*, 14.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science*, 286(5439):509–512.

David R Bild, Yue Liu, Robert P Dick, Z Morley Mao, and Dan S Wallach. 2015. Aggregate characterization of user behavior in twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1):1–24.

Eric Bonabeau. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, 99(suppl_3):7280–7287.

Jinyu Cai, Yusei Ishimizu, Mingyue Zhang, Munan Li, Jialong Li, and Kenji Tei. 2025. Simulation of language evolution under regulated social media platforms: A synergistic approach of large language models and genetic algorithms. *arXiv preprint arXiv:2502.19193*.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. 2025. Llms generate structurally realistic social networks but overestimate political homophily. *International AAAI Conference on Web and Social Media*.

Yannis Charalabidis, Euripidis Loukis, and Aggeliki Androutsopoulou. 2011. Enhancing participative policy making through modelling and simulation: A state of the art review. In *Proceedings of the European, Mediterranean and Middle Eastern Conference on Information Systems-Informing Responsible Management: Sustainability in Emerging Economies, EMCIS*, pages 210–222.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2024a. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*.

Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2024b. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. *arXiv preprint arXiv:2407.07061*.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9):e2023301118.

Harald Cramér. 1946. Mathematical methods of statistics. *Princeton University Press*.

Christine SM Currie, John W Fowler, Kathy Kotiadis, Thomas Monks, Bhakti Stephan Onggo, Duncan A Robertson, and Antuela A Tako. 2020. How simulation modelling can help reduce the impact of covid-19. *Journal of Simulation*, 14(2):83–97.

Joshua M Epstein. 1999. Agent-based computational models and generative social science. *Complexity*, 4(5):41–60.

Joshua M Epstein and Robert Axtell. 1996. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

Chen Gao, Xiaochong Lan, Zhi jie Lu, Jinzhu Mao, Jing Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system

with large language model-empowered agents. *ArXiv*, abs/2307.14984.

Matthew Gentzkow and Emir Kamenica. 2017. Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*, 104:411–429.

Mark S. Granovetter. 1973. The strength of weak ties. *American Journal of Sociology*, 78:1360 – 1380.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Chenhao Gu, Ling Luo, Zainab Razia Zaidi, and Shanika Karunasekera. 2025. Large language model driven agents for simulating echo chamber formation. *arXiv preprint arXiv:2502.18138*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):82.

Jennifer Jerit and Yangzi Zhao. 2020. Political misinformation. *Annual Review of Political Science*, 23(1):77–94.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33.

Emir Kamenica and Matthew Gentzkow. 2011. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Susan Landau, James X Dempsey, Ece Kamar, Steven M Bellovin, and Robert Pool. 2024. Challenging the machine: Contestability in government ai systems. *arXiv preprint arXiv:2406.10430*.

Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. 2023. Examining llms' uncertainty expression towards questions outside parametric knowledge. *arXiv preprint arXiv:2311.09731*.

Fabian Lorig, Emil Johansson, and Paul Davidsson. 2021. Agent-based social simulation of the covid-19 pandemic: A systematic review. *JASSS: Journal of Artificial Societies and Social Simulation*, 24(3).

Morgan Marietta, David C Barker, and Todd Bowser. 2015. Fact-checking polarized politics: does the fact-check industry provide consistent guidance on disputed realities? In *The forum*, volume 13, pages 577–596. De Gruyter.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Review of Sociology*, 27:415–444.

Stephan Minos. 2023. Chirper: Changing the ai landscape with autonomy and entertainment. *Chirper Blog*.

Lev Muchnik, Sen Pei, Lucas C Parra, Saulo DS Reis, José S Andrade Jr, Shlomo Havlin, and Hernán A Makse. 2013. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3(1):1783.

OpenAI. 2024. Disrupting deceptive uses of ai by covert influence operations.

Amy Orben, Andrew K Przybylski, Sarah-Jayne Blakemore, and Rogier A Kievit. 2022. Windows of developmental sensitivity to social media. *Nature Communications*, 13(1):1649.

Folco Panizza, Piero Ronzani, Tiffany Morisseau, Simone Mattavelli, and Carlo Martini. 2023. How do online users respond to crowdsourced fact-checking? *Humanities and Social Sciences Communications*, 10(1):1–11.

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

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Uma Patel. 2024. New partnerships and initiatives to strengthen fact-checking online.

Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11):4944–4957.

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

Gordon Pennycook and David G. Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25:388–402.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Zhongyi Qiu, Hanjia Lyu, Wei Xiong, and Jiebo Luo. 2025. Can llms simulate social media engagement? a study on action-guided response generation. *arXiv preprint arXiv:2502.12073*.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Durga Raghunath and Surabhi Malik. 2024. Partnering with news publishers and fact checkers to tackle misinformation ahead of the indian general elections 2024.

Anatol Rapoport. 1953. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *Bulletin of Mathematical Biology*, 15:523–533.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced fact-checking at twitter: how does the crowd compare with experts? In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1736–1746.

Thomas C Schelling. 1971. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186.

Kirill Solovev and Nicolas Pröllochs. 2022. Moral emotions shape the virality of covid-19 misinformation on social media. In *Proceedings of the ACM web conference 2022*, pages 3706–3717.

Briony Swire-Thompson, David Lazer, et al. 2020. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health*, 41(1):433–451.

Joseph E Uscinski and Ryden W Butler. 2013. The epistemology of fact checking. *Critical Review*, 25(2):162–180.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024. Sotopia-π: Interactive learning of socially intelligent language agents. *arXiv preprint arXiv:2403.08715*.

Uri Wilensky. 1999. Netlogo. 1.0.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. 2024. Oasis: Open agents social interaction simulations on one million agents. *arXiv preprint arXiv:2411.11581*.

Greyson K Young. 2022. How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law*, 31(1):1–16.

Zhenhua Yu, Si Lu, Dan Wang, and Zhiwu Li. 2021. Modeling and analysis of rumor propagation in social networks. *Information Sciences*, 580:857–873.

Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–37.

Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. 2020. Fake news propagates differently from real news even at early stages of spreading. *EPJ data science*, 9(1):7.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. 2024. Sglang: Efficient execution of structured language model programs. *Advances in Neural Information Processing Systems*, 37:62557–62583.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Extended Discussion on Content Popularity

A.1 Network Properties: Centrality and Clustering Metrics

We analyze the structural role of each agent in the directed follow-network $G = (V, E)$ with $N = |V|$ nodes by computing degree, betweenness, closeness, and eigenvector centralities, as well as the local clustering coefficient and global transitivity.

Degree centrality quantifies how many direct connections a node has relative to the maximum possible. For node v with total degree $\deg(v) = k_v$ (in- plus out-degree), we define

$$C_D(v) = \frac{k_v}{N-1},$$

and compute the average degree centrality as

$$\overline{C_D} = \frac{1}{N} \sum_{v \in V} C_D(v).$$

Betweenness centrality measures the fraction of shortest directed paths between all ordered pairs (s, t) that pass through v. Let σ_{st} be the number of shortest paths from s to t, and $\sigma_{st}(v)$ those that pass through v. Then

$$C_B(v) = \sum_{\substack{s,t \in V \\ s \neq t \neq v}} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

and its normalized average is

$$\overline{C_B} = \frac{1}{N\,(N-1)\,(N-2)} \sum_{v \in V} C_B(v).$$

Closeness centrality reflects how near a node is to all others based on shortest-path distances $d(v, u)$. We set

$$C_C(v) = \frac{N-1}{\sum_{u \in V \setminus \{v\}} d(v, u)},$$

and average as

$$\overline{C_C} = \frac{1}{N} \sum_{v \in V} C_C(v).$$

Eigenvector centrality assigns importance proportional to the centrality of a node's neighbors. If $\mathbf{A}$ is the adjacency matrix of G, we solve

$$\mathbf{A}\,\mathbf{x} = \lambda_{\max}\,\mathbf{x},$$

and take $C_E(v) = x_v$, with average

$$\overline{C_E} = \frac{1}{N} \sum_{v \in V} x_v.$$

The local clustering coefficient of node v measures the density of edges among its k_v neighbors. If e_v is the number of edges between neighbors, then

$$C(v) = \frac{2\,e_v}{k_v(k_v - 1)},$$

and the network-wide clustering is

$$\overline{C} = \frac{1}{N} \sum_{v \in V} C(v).$$

Finally, transitivity (global clustering) is the ratio of closed triplets (triangles) to all connected triplets:

$$T = \frac{\sum_{v \in V} 2\,e_v}{\sum_{v \in V} k_v(k_v - 1)}.$$

In these definitions, $d(v, u)$ denotes the shortest directed-path length from v to u, σ_{st} the total number of shortest paths from s to t, and $\sigma_{st}(v)$ those passing through v. All sums run over V unless noted otherwise.

In our simulated community of 161 agents engaging with 4,249 pieces of content, the resulting follow-network comprises 358 directed ties, indicating a modest level of connectivity driven by personas and memory-based decisions. An average degree centrality of 0.1105 shows that each agent follows roughly 11% of the population, while the low average betweenness centrality (0.0128) suggests that few agents act as indispensable bridges. With an average closeness centrality of 0.1367, most agents remain just a few steps apart, supporting rapid information flow, and the modest eigenvector centrality (0.0433) reveals that influence is fairly distributed rather than monopolized by a handful of hubs. Finally, a clustering coefficient of 0.205 and network transitivity of 0.189 point to moderate local grouping without excessive fragmentation—together painting a picture of a network that balances cohesion and reach in propagating popular content.

A.2 Power-Law Distribution of User Popularity

First we define the popularity of users as a sum of the number of followers, number of likes, shares, and comments received. We collected the top 50 users and plotted their popularity (as measured by the sum of engagement received by them) from highest to lowest in Fig. 8. We observe a power-law distribution of user influence. We have $f(x) = 120x^{-0.6}$ as the best-fitting power-law approximation of our sampled data, as shown in Fig. 5. With $\alpha = 0.60$, our regression line has an $R^2 = 0.84$, suggesting a strong fit to our user engagement data and that our social system follows a typical power law distribution where a few users generate most of the engagement. Existing analysis on real-world social networks suggest that this power-law exponent usually ranges from 1.5-2.5 depending on the specific context (Muchnik et al., 2013; Bild et al., 2015). Our best-fit exponent is lower than these reported numbers, but it still illustrates a clear trend that a minority of users/content collect most of the engagement while the majority of them do not contribute nearly as much.

In the rest of this section, we explore potential reasons why this distribution emerges, and through a series of analyses leveraging our simulated environment, we reveal the unpredictability of influence or popularity in online social networks. More fundamentally, we argue that perhaps LLM-driven agents have a tendency to simply copy the decisions of agents who act before them. This results in the preferential attachment and as a natural consequence establishes the power-law distribution of engagement pattern. Such a pattern does not necessarily stem from anything else such as user's profile details, or the content they post about. And even their own "inner reasoning" might not reveal their true decision-making, which invites further investigation on the authenticity of LLM agents' self-expressed reasoning traces.

A.3 Persona Attributes Don't Correlate with Engagement

We analyzed user engagement by comparing the top 50 most engaged users (highest number of followers, likes, shares, comments, etc.) with the bottom 50 least engaged users across several attributes. The Chi-square test results summarized in Tab. 2 indicate that there are no statistically significant differences in the distributions of age group, gender, activity type, hobby, ethnicity, income level, political affiliation, or primary goal between the two groups. Although some attributes, such as ethnicity and hobby, exhibited medium effect sizes (Cramer's V (Cramér, 1946) of 0.319 and 0.302, respectively), their associated p-values did not reach conventional levels of statistical significance. This suggests that the attributes examined do not notably influence the level of user engagement.

Notably, we did not include personas resembling real-world public figures or celebrities, whose presence might have substantially influenced content popularity. Our findings thus suggest that, when personas are initialized randomly, some users naturally attract significantly more attention and engagement, independent of the specific attributes assigned during their initialization. This underscores the inherent variability and unpredictability of user engagement in social platforms.

A.4 Do Content Topics Matter?

Our analysis aimed to directly investigate the correlation between content topics and user engagement. To accomplish this, we first computed an engagement score for each post by summing its likes, shares, and comments. We then cleaned and preprocessed the textual content of the posts to ensure accurate topic modeling.

For topic extraction, we employed a unified topic model based on BERTopic (Grootendorst, 2022), utilizing sentence embeddings from the Sentence-Transformer model `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019).BERTopic was chosen due to its effectiveness in capturing nuanced semantic relationships within short text content. By fitting a single topic model to all posts, we ensured consistency and comparability across the identified topics.

Following topic assignment, we conducted a detailed statistical analysis. Engagement metrics—including mean, median, and standard deviation for likes, shares, comments, and overall engagement scores—were calculated for each topic. To statistically assess whether variations in engagement across topics were significant, we performed an ANOVA (Analysis of Variance).

The key statistical finding from the analysis was an ANOVA result yielding an F-statistic of 0.614 and a p-value of 0.84. This indicates **no statistically significant relationship between the topics and overall engagement levels**. In other words, statistically, the topic of a post alone does not reli-

ably predict its engagement level.

A.5 Clues from Agents' Own Reasoning Traces and Recommender System

The lack of clear correlation between user profiles, content topics, or temporal properties, and engagement patterns suggested that maybe the way we present feed to the agents has an influence on what content ends up being popular. Here we discuss our feed prioritization algorithm. Our simulation does not employ a sophisticated recommender system. Our feed prioritization in the simulation relies primarily on recency and existing follow relationships, rather than explicit engagement metrics such as likes or shares. Posts, regardless of whether they're from followed or non-followed users, are generally ordered based on creation time, ensuring that newer content receives greater visibility. However, content from followed users gains additional prioritized exposure due to dedicated allocations in the feed. This structure might create a follow-based feedback loop: when User B follows User A, A's posts consistently appear in B's feed, enhancing A's opportunities for engagement through likes, comments, and shares. Higher engagement subsequently boosts A's visibility to other users who view these interactions, increasing the likelihood of additional follows and further amplifying this cycle.

Agent's Reasoning Pattern We extract and analyze agent reasoning across several dimensions, including sentiment, motivation, entity and concept extraction, and word-frequency analysis. The analysis specifically focused on identifying patterns related to different engagement actions (such as likes, comments, and shares), exploring how post content and user backgrounds influenced reasoning, and examining common linguistic trends.

The analysis of agent reasoning reveals patterns in how agents engage with content and users on social media. As shown in Fig. 3, agents demonstrate clear and distinct emotional sentiment patterns associated with different types of actions. Positive-dominant actions such as following users (99% positive sentiment), commenting (97%), liking posts (92%), and sharing content (92%) indicate that agents predominantly perceive their interactions as constructive contributions. Conversely, negative sentiment predominantly characterizes actions like flagging posts (71% negative) and unfollowing users (40% negative), reflecting agents' use of these interactions primarily for expressing disapproval or concern.

Further examining motivational reasoning, agents apply distinct frameworks depending on the nature of their engagement. Content evaluation actions, such as flagging posts, are predominantly motivated by information quality assessments (49%) and concerns regarding misinformation (22%). Sharing decisions primarily reflect agreement with content (46%). In contrast, relationship-building actions show different motivations: liking is heavily driven by social connection potential (34%), commenting balances agreement (29%) and social connection (28%), and following users reflects diverse personal interests (27%).

Vocabulary analysis further emphasizes these distinctions, revealing specialized linguistic patterns for each type of engagement. Flagging content uses specific moderation-related language such as "misinformation," "harmful," and "credible," whereas community-oriented engagements like sharing, liking, and commenting frequently reference concepts like "community," "support," and "alignment." Following actions highlight terms related to content curation and long-term value, including "consistently," "valuable," and "insights."

Interestingly, despite these detailed reasoning frameworks, **a low alignment (21.4%) between post sentiment and agent reasoning indicates that agents' explicit justifications may not fully reflect the underlying factors driving engagement. Instead, engagement decisions appear largely guided by personal values alignment**, information quality assessments, community-building potential, and personal relevance rather than simple emotional resonance with content.

These insights also highlight a notable contradiction with prior analyses, which showed no significant correlation between user demographics or content topics and overall engagement popularity. **While agents clearly articulate their engagement motivations in terms of specific frameworks (values alignment, informational quality, social connection), these explanations alone do not robustly predict broad engagement patterns.** This paradox suggests that engagement is heavily individualized, contextual, and possibly influenced by network effects—such as who posts content, existing social validation, or content placement within social feeds—factors not fully captured by demographic or topical categorizations alone.

In essence, the analysis confirms that agents employ reasoning structures tailored to the type of engagement but reveals that actual engagement outcomes are influenced by nuanced individual interpretations and contextual social dynamics. This misalignment between LLM's internal decision-making and explicit surface behavior is also consistent with findings observed by prior work (Liu et al., 2023). Recognizing these complexities is essential for understanding and anticipating the unpredictability in social media engagement.

B Persona Generation Details

Here we describe the questions that we sampled and generated for the agent users. The generated personas are stored in JSONL format, with each entry containing a unique identifier, a descriptive narrative, and associated behavioral labels.

B.1 Persona Replication from Human Survey

The persona generation method begins by transforming structured survey responses collected from Prolific participants into rich, natural language character descriptions suitable for use in agent-based simulations. Each participant's responses — covering a wide range of personal, demographic, social, and psychological traits—are encoded in JSONL format, where each line corresponds to a different individual. We first this file into a list of Python dictionaries, each representing a single participant's answers. The preprocessing pipeline then embeds each answer into a templated sentence structure. This includes details such as age, gender, residential background, number of places lived, favorite activities, values, political stance, income, ethnicity, language, education, religion, social tendencies, hobbies, relationship values, personality, future goals, significant life events, friendship values, and hypothetical financial decisions. By expressing these traits in fluent, first-person-style English, the function essentially replicates each participant's worldview and identity into a lifelike persona that can guide agent behavior in social simulations. In the final step iterates through all participant entries, generates the corresponding natural language persona for each one, and writes the enriched data—including both the original responses and the generated description—back into a new JSONL file. This process creates a bridge between raw human survey data and psychologically grounded agent profiles, enabling more realistic and diverse behaviors in multi-agent environments.

B.2 Synthetic Persona from Agent Bank

In contrast to the human-annotated personas derived from survey responses, we also generate fully synthetic personas by sampling from a structured question bank, referred to as the Agent Bank. This bank contains a curated set of 23 multiple-choice questions covering key dimensions of identity, background, and social orientation—ranging from age and gender to values, education, hobbies, and political affiliations. Please refer to the code repository for the complete content and answer choices of each of them. Each question is assigned a label and a fixed set of possible answers. To simulate human-like diversity, we construct agent personas by probabilistically sampling answers from these options, sometimes using uniform random choice and other times leveraging carefully constructed distributions to better mirror real-world population dynamics. For instance, age is generated from a normal distribution centered at 35 with bounds clamped between 18 and 60, while gender is sampled from a distribution reflecting approximate societal proportions. In some cases, dependencies between traits are explicitly modeled—for example, primary language is sampled conditionally based on a person's ethnicity using manually specified probability distributions that reflect linguistic prevalence across ethnic groups. These sampled answers are then assembled into a dictionary of attributes. From this, we use one of two methods to generate natural language persona descriptions. The first method uses a hardcoded template that deterministically weaves the sampled answers into a coherent paragraph, mimicking the style and structure used for real survey-based personas. The second, more dynamic method leverages GPT-4o to produce creative and varied persona descriptions from the same underlying attributes. A carefully crafted system prompt instructs the model to retain every single piece of information from the attribute dictionary while generating a single fluent paragraph in the second person, presenting the result as a believable and detailed backstory. This ensures that each agent maintains a consistent and complete identity while allowing room for stylistic diversity. Ultimately, each synthetic persona is stored as a structured JSON object containing a unique ID, the full natural language description, and the associated label-value pairs, ready to be deployed as agents in downstream simulations.

Table 3: Agent Reasoning for Content Engagement Analysis

Action Type	Total Actions (%)	Positive (%)	Neutral (%)	Negative (%)	Top 2 Reasoning Categories
share_post	1382 (30.1%)	91.8	6.6	1.6	agreement (46.3%) social_connection (15.7%)
flag_post	1126 (24.6%)	13.5	15.5	71.0	information_value (48.8%) misinformation (22.4%)
comment	880 (19.2%)	96.8	2.7	0.5	agreement (29.2%) social_connection (27.6%)
follow_user	719 (15.7%)	98.9	0.8	0.3	personal_interest (27.4%) information_value (24.4%)
like_post	463 (10.1%)	92.0	7.8	0.2	social_connection (33.6%) agreement (23.6%)
ignore	9 (0.2%)	77.8	–	22.2	information_value (36.4%) personal_interest/agreement (18.2%)
unfollow_user	5 (0.1%)	20.0	40.0	40.0	agreement (50.0%) emotional_reaction (16.7%)

B.3 User Generation and Instantiation

The foundation of the simulation lies in the creation of realistic individual agentic virtual users. Each agent is instantiated with a detailed persona that shapes their online behavior and engagement patterns.

Persona Generation As illustrated in Fig. 1, personas are generated using a combination of predefined questions and sampling from probabilistic distributions stored in an `agent_bank` profile collection, inspired by Park et al. (2024). Key demographic attributes such as age, gender, ethnicity, and primary language are assigned probabilistically to mirror real-world distributions. For instance, age follows a normal distribution centered around 35 years, while other attributes are sampled based on predefined probabilities. We disclose these questions and describe more detail of the methodology in Appendix B. After synthesizing the structured profiles for agents, we construct a natural language description for each of them. This process leverages a mixture of deterministic rules and LLM-based augmentation using `GPT-4o` (Hurst et al., 2024) to enhance diversity and realism.

User Instantiation Once personas are generated, they are instantiated as agent users within the simulation. Each agent is assigned a unique user ID and a persona profile that includes background details and interest labels. The relational database serves as the backbone for recording agent activities, ensuring persistent storage of interactions, post engagements, and behavioral updates. This database facilitates dynamic user tracking and enables post-simulation analysis of engagement trends and content spread. We provide more implementation details in Appendix F.

C Models And Computational Budget

We use APIs for proprietary LLM inferences in our experiments. We locally hosted some smaller open-weight LLMs in the early exploratory stages of the projects, on H100 GPUs. We end up consuming approximately $600 of OpenAI credits, $300 of Anthropic API, and $40 of DeepSeek API.

D Details of the Human Study and IRB

The study was open to 20,240 eligible participants from a larger Prolific population of 232,330, and we collected 204 valid responses from eligible participants. The survey was conducted via Prolific to collect responses from U.S.-based participants fluent in English. Participants were asked to complete a 12-minute survey assessing their demographic characteristics and social media interactions. The survey, hosted on Google Forms, required no software downloads or special device features and was accessible via mobile, tablet, or desktop. Participant recruitment applied custom screening for language, political spectrum, vaccine opinion, and prior participation, ensuring a targeted sample. Responses were collected using Prolific ID via a question at the start of the form, and participants received a completion code upon finishing. Compensation was set at $2.40 per participant, equivalent to $12.00/hour, and submissions were manually reviewed before approval. The median completion time was approximately 14.5 minutes. All members on our research team have obtained IRB approval before the human study was conducted.

Our study costs a total of $480 for participant payment and $160 of platform fee. Fig. 6 shows the complete breakdown of the 9 key demographic distributions of the 204 human participants. This study was performed under approval from the appropriate institutional ethics review board. Full IRB documentation will be made available upon request.

D.1 Per-Demographic Attribute Engagement Pattern

We also analyzed the reaction patterns between human participants and persona-driven agents grouped by specific demographic attributes such as age, gender, income, ethnicity, etc, as shown in Tab. 4.

The analysis of engagement patterns between humans and agents reveals some differences across various demographic groups. Specifically, the age group 25-34 shows notable differences in shares, while males exhibit significant variations in both likes and shares. Among religious groups, Hinduism and Islam display significant differences in likes and shares, respectively. Ethnic groups such as Hispanic or Latino, Black or African American, and Asian show significant differences in shares and comments. Education levels also play a role, with secondary education and doctorate degree holders showing significant differences in shares and likes, respectively. Income levels between $10,000 - $19,999 and $70,000 - $79,999 show significant differences in comments and likes. Political stances such as Conservative and Very Conservative also exhibit significant differences in shares and likes. In contrast, many other demographic groups, including various age ranges, genders, religions, ethnicities, education levels, income brackets, and political stances, show no significant differences in engagement types.

Overall, out of 52 demographic groups analyzed, 14 show significant differences in one or more engagement types, while 38 do not. The criteria for significance were based on a p-value of less than 0.05 in statistical comparisons. The results suggest that agents may be more adept at simulating the engagement patterns of "common" or more broadly represented demographic groups in LLM pretraining data, as indicated by the lack of significant differences in many of these groups. We find that out of 52 examined demographic subgroups, only 14 showed statistically significant differences ($p < 0.05$) in at least one engagement metric (Tab. 4). Notable discrepancies appeared in the 25–34 age group (shares) and several religious, ethnic, educational, income, and political categories. However, most demographic groups exhibited no significant differences, suggesting that agents simulate typical engagement behavior more accurately for demographics more prevalent in LLM training data.

E An Extended Version of Related Work

Behavioral Economics and Persuasian Games. Our system computationally models a sequential persuasion game with LLM-powered agents conditioned on fine-grained personas (Kamenica and Gentzkow, 2011; Gentzkow and Kamenica, 2017; Acemoglu et al., 2023). The agents interact within a directed social graph and evolve based on memory and social context. This framework serves as a testbed for studying online behaviors, intervention strategies, and the impact of algorithmic moderation.

AI-Driven Social Simulations and Generative Agents. The emergence of large language models (LLMs) has significantly advanced the capabilities of agent-based social simulations, enabling more sophisticated, context-aware interactions. Traditional agent-based modeling relied on predefined rule sets and heuristics, limiting adaptability and realism. Early computational social simulations, such as Schelling's segregation model (Schelling, 1971), Sugarscape (Epstein and Axtell, 1996) and NetLogo-based models (Wilensky, 1999), provided insights into social dynamics but lacked the ability to generate nuanced, context-dependent behaviors.

Recent advances, such as Smallville (Park et al., 2023), AgentVerse (Chen et al., 2024a), Internet-of-Agents (Chen et al., 2024b), and Chirper (Minos, 2023), leverage LLMs to enable generative agents that dynamically respond to evolving contexts. These systems showcase how AI-powered agents can engage in lifelike conversations, form social relationships, and simulate content dissemination patterns. However, despite their ability to generate plausible interactions, generative agents can still exhibit inconsistencies due to biases inherent in LLM training data or limitations in long-term memory and reasoning. By integrating more structured constraints and iterative feedback mechanisms, this work enhances the reliability of agent-based simulations for social science research and policy testing.

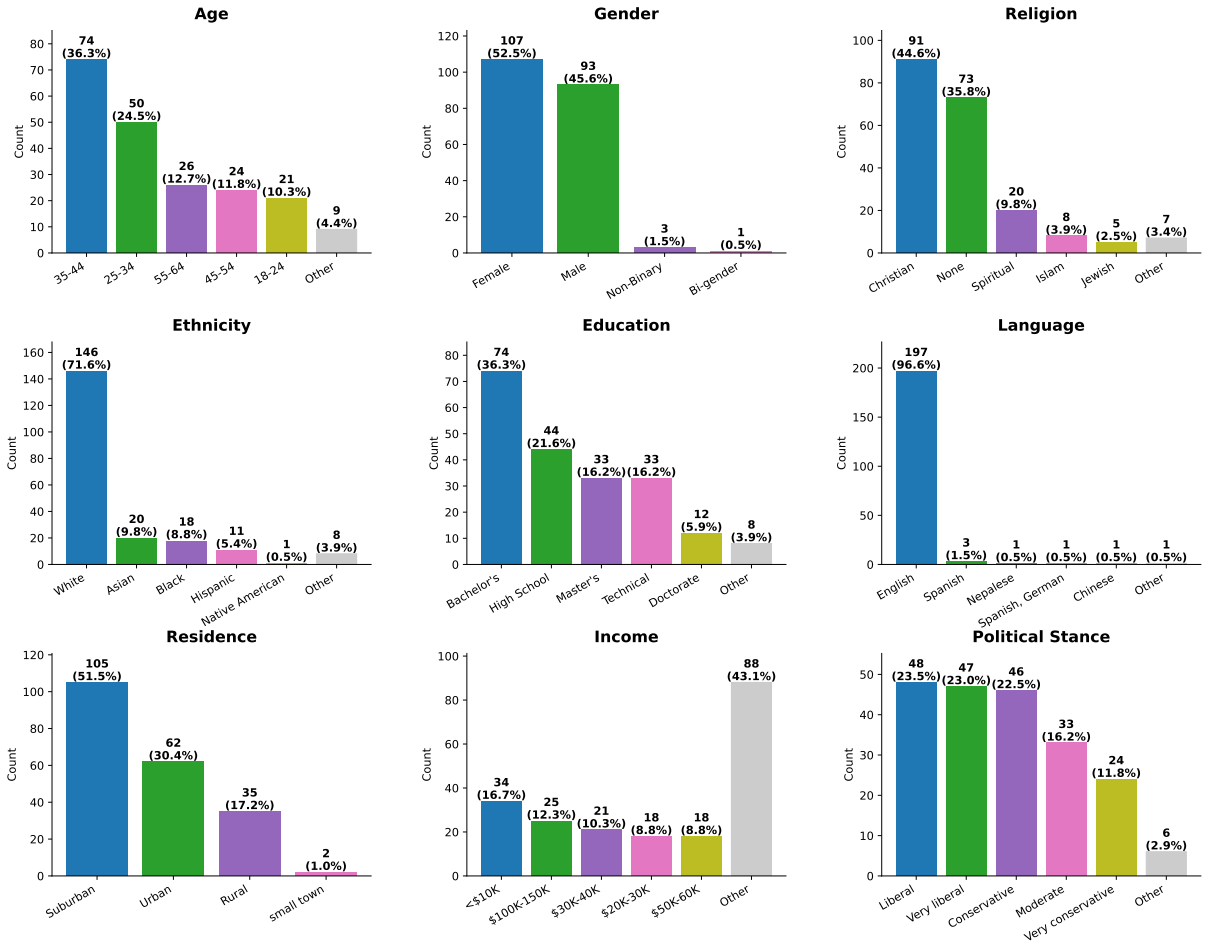

Figure 6: Demographic Distributions of Study Participants.

Misinformation Spread and Fact-Checking Mechanisms. The spread of misinformation on digital platforms has been extensively studied (Swire-Thompson et al., 2020; Jerit and Zhao, 2020; Wu et al., 2019; Islam et al., 2020), with empirical evidence showing that falsehoods often propagate more rapidly and broadly than factual information (Vosoughi et al., 2018). The virality of misinformation is attributed to its emotional appeal, novelty, and the role of engagement-driven algorithms that inadvertently amplify misleading narratives (Pennycook and Rand, 2021; Solovev and Pröllochs, 2022). Addressing this issue has led to the development of multiple fact-checking methodologies, including third-party verification, algorithmic detection, and crowdsourced moderation.

Third-party fact-checking, typically conducted by organizations such as Snopes[11], PolitiFact[12], or Google's partnerships with external organizations (Raghunath and Malik, 2024; Patel, 2024), provides authoritative assessments but faces challenges in scalability and timeliness (Zannettou et al., 2019; Uscinski and Butler, 2013; Marietta et al., 2015). Crowdsourced fact-checking such as X's Community Notes,[13] on the other hand, leverages collective intelligence (Panizza et al., 2023) but introduces risks related to expertise and susceptibility to group biases (Saeed et al., 2022; Pennycook et al., 2021). There is no consensus on which fact-checking approach is more effective, nor is it well-understood how different moderation strategies interact. This study addresses this gap by leveraging LLM-driven simulations to evaluate different fact-checking mechanisms within controlled environments. By testing various moderation strategies in a scalable, repeatable manner, this work provides insights into the comparative efficacy of community-based, third-party, and hybrid fact-checking interventions in mitigating misinformation.

[11] https://www.snopes.com/
[12] https://www.politifact.com/
[13] https://communitynotes.x.com/guide/en/

Table 4: Demographic groups showing significant differences in engagement patterns between human participants and AI agents ($p < 0.05$).

Category	Significant Differences	Non-Significant Differences
Age	25-34 (shares)	18-24, 35-44, 45-54, 55-64, 65-74
Gender	Male (likes, shares)	Female
Religion	Hinduism (likes), Islam (shares)	No Religion, Spiritual, Christianity, Jewish
Ethnic Group	Hispanic/Latino, Black/African (shares), Asian (comments)	White/Caucasian, Mixed, Others
Education	Secondary (shares), Doctorate (likes)	High School, Undergraduate, Technical, Graduate
Income	$10K-$20K (comments), $70K-$80K (likes)	Various other income brackets
Political Stance	Conservative (shares), Very Conservative (likes)	Very Liberal, Moderate, Liberal, Libertarian

Simulations as Tools for Policy and Platform Governance The use of computational simulations as decision-support tools has been well-established in domains such as epidemiology (Currie et al., 2020; Lorig et al., 2021), economics (Axtell and Farmer, 2022), and public policy (Qu and Wang, 2024). By enabling scenario testing before real-world implementation, simulations help policymakers anticipate the consequences of interventions (Charalabidis et al., 2011). In the context of social media governance, AI-driven simulations present an emerging opportunity to evaluate moderation strategies, optimize intervention policies, and test the societal impact of algorithmic changes before deployment.

Recent discourse around AI governance emphasizes the need for proactive measures to ensure platform accountability and transparency (Landau et al., 2024). Regulatory bodies and platform operators are increasingly exploring ways to assess the impact of interventions such as content moderation adjustments, ranking algorithm changes, and misinformation mitigation strategies before rolling them out at scale. To this end, our research introduces AI-driven social simulations as a novel framework for governance experimentation. By simulating diverse social environments and misinformation dynamics, we provide an approach that offers a scalable, controlled setting for testing policy interventions. This methodology aligns with the growing call for algorithmic auditing and regulatory sandboxes, providing a novel tool for both researchers and policymakers to refine governance strategies before real-world application.

F Database Schema

In this section, we describe the database schema that we developed to store and keep track of all the data generated by each simulation run.

This relational SQL database schema is designed to support a social media simulation in which LLM-powered AI agents mimic user behaviors. The database captures and organizes user-generated content, interactions, and system-level processes in detail. The `users` table stores individual user profiles, including metadata such as personas, background labels, influence scores, and engagement metrics. Posts authored by users are managed in the `posts` table, which records content details, interaction counts (likes, shares, flags, comments), and moderation or fact-check statuses. Social relationships are modeled through the `follows` table, which tracks follower-followed connections. User engagement actions, such as creating content or reacting to posts, are logged in the `user_actions` table. Comments on posts are separately recorded in the `comments` table with their associated metadata. Community moderation is facilitated via the `community_notes` and `note_ratings` tables, enabling users to contribute interpretive notes and rate their helpfulness. System moderation decisions are logged in `moderation_logs`. The `fact_checks` table provides detailed verdicts and rationales from fact-checking processes. To simulate memory and reasoning for AI agents, `agent_memories` track the content and importance of internal memories, with timestamps and decay factors. The `spread_metrics` table quantifies the virality and diffusion dynamics of each post over time steps, including derived interaction statistics and takedown decisions. Exposure to content is tracked at the user level in the `feed_exposures` table, supporting the analysis of information visibility and reach. Together, these schemas capture a detailed and interconnected view of simulated social media dynamics, grounded in observable user behavior and system responses.

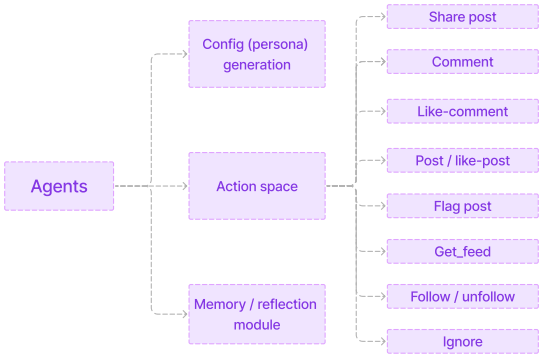

Figure 7: Agent's action space.

G Agent Action Space

In our simulated social media environment, each agent—powered by a large language model—is instantiated with a predefined action space that governs its interactions within the platform. These agents are configured with unique personas via a Config (persona) generation module and endowed with a Memory / reflection module that allows them to recall and adapt based on past experiences. The Action space outlines the full spectrum of behaviors an agent can exhibit: they can share posts, comment or like comments, create or like posts, and flag inappropriate content. Additionally, agents can retrieve their feeds, follow or unfollow other users, or ignore content or interactions altogether. These discrete actions simulate realistic user behavior and social dynamics, enabling rich, emergent interactions in the environment.

G.1 Agent Decision-Making Process

The agent decision-making process is governed by structured interactions between feed presentation, memory recall, and reasoning mechanisms.

Feed Presentation Each agent's feed aggregates posts from followed users, supplemented by additional trending content and news articles. On average, one in ten posts is from the NewsGuard dataset and contains misinformation. Posts are displayed with metadata such as engagement counts (likes, comments, shares) and fact-checking signals (flags, community notes, third-party verdicts). This metadata provides context for the agent's engagement decisions.

Memory and Reflection Module We implement an `AgentMemory` module which manages the memory and reflection capabilities of each agent. Memories are categorized into *interactions* (e.g., past engagements) and *reflections* (high-level insights derived from past behaviors). Each piece of memory is assigned an importance score, which decays over time unless reinforced by further interactions. The decay function ensures that long-term behaviors emerge naturally based on experience. Please refer to Appendix I for more details of the Memory module.

Periodically, agents generate reflections based on recent interactions. These reflections help detect behavioral patterns, relationship dynamics, and potential biases, influencing future content engagement and decision-making.

Agent Decision-Making and Action Execution Agents make decisions based on a combination of persona-driven heuristics, memory retrival, and reasoning prompts. The `AgentPrompts` module formulates structured decision prompts, guiding agents through content engagement options such as liking, sharing, or flagging a post. When engaging, agents provide reasoning for their actions, influenced by (1) **Personal Beliefs and Persona Traits**: Agents weigh content credibility based on their ideological stance and historical preferences, (2) **Engagement Signals**: Highly engaged posts are more likely to be reshared due to social validation effects, and (3) **Fact-Checking Feedback**: Agents integrate fact-checking signals into their reasoning, adjusting their trust in flagged content accordingly.

Once a decision is made, the agent's action gets recorded in the relational database, along with updated post metrics, engagement, and new memories. The importance of each interaction is evaluated based on emotional intensity, action strength, and alignment with the agent's goals.

H Detailed Experiment Configuration

We describe our experimental settings and configurable variables in more detail in this section. In our simulation, we model a dynamic social network of an arbitrary number of (practically in our experiments up to over 200) LLM-driven agents interacting over the course of a number of discrete time steps. The simulation loop follows a structured core cycle that includes initializing the environment, assigning new users probabilistically (though this one could be disabled in certain runs), content creation, feed-based reactions, and peri-

odic reflective updates. Each agent is instantiated from detailed persona descriptions provided via an external JSONL file, and operates using the GPT-4o engine with a decoding temperature of 1.0 to promote diversity in generated responses. Agents can be configured to create original posts independently, or they can be prompted to only respond to posts depending on the setting. Once the simulation environment is initiated, an agent's feed consists of a mixture of up to a default of 15 posts from followed users and 10 from non-followed users, drawn from a pool that includes up to 20 injected news items per run. Initial social ties are sparse, with a 10% probability of following another user at initialization, and new user addition and follow behaviors are disabled during the simulation. All of the above numbers are configurable. The experiment evaluates one of the four fact-checking intervention modes described in Section 3, combining both third-party and community-based mechanisms. For each step, if a fact-checking agent is enabled, then a number of posts are selected for potential moderation, with fact-checking outputs generated using a low-temperature (0.3) setting and required to include reasoning. Thresholds are specified for flagging and notetaking behavior if moderation is set active. Periodically, agents reflect on their recent interactions, update memory states, and check their internal objectives, offering a framework for studying emergent behavior, information diffusion, and intervention efficacy in artificial societies.

I Details of the Memory Module

Memory relevance is computed as:

$$\text{Relevance} = \text{Importance} \times \text{Decay}$$

The decay factor is defined as:

$$\text{Decay} = \max(0, \text{PrevDecay} - \alpha \Delta t)$$

where α is the decay rate (default 0.1), and Δt is the time (in days) since last access. New memories start with PrevDecay $= 1.0$.

A memory is considered relevant if Relevance ≥ 0.3. Both Importance and Decay are in $[0, 1]$.

Importance Scoring. Each memory has a base importance score of 0.5. This is increased by 0.1 for each keyword match (up to a max of 1.0) from the following semantic categories:

- Emotional: *love, hate, angry, happy, sad*
- Action: *achieved, failed, learned, discovered*
- Relationship: *friend, follow, connect, share*
- Goal: *objective, target, aim, purpose*

Let k be the number of keyword matches in the memory content. Then:

$$\text{Importance} = \min(1.0, 0.5 + 0.1k)$$

This value is combined with the decay factor to compute final relevance.

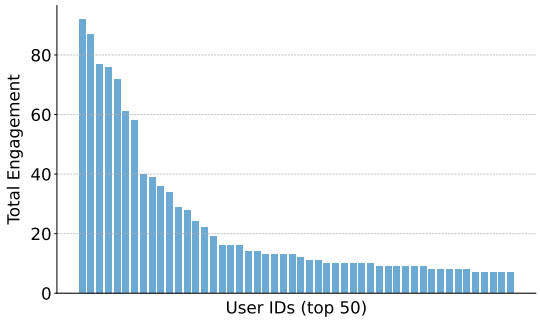

Figure 8: Top 50 users with highest engagement.

J Total Engagement Comparison

Fig. 9 tracks average interactions with factual and misinformative posts across time steps 0-40 for three LLMs: Claude-3.7-Sonnet, DeepSeek-V3, and GPT-4o over four fact-checking types as defined in the experiment configurations: None, Community Based, Third Party, and Hybrid. Each subplot shows how user engagement with true and false content evolves.

J.1 Total Interactions: Claude-3.7-Sonnet

"None", "Community Based": There is little to no interaction with misinformative content. Interaction with factual posts decreases over time.
"Third Party": Initially, there is higher interaction with misinformative content than factual content. This reverses as interaction with misinformation decreases and with factual content increases.
"Hybrid": There is little to no interaction with misinformative content. There is steady increase in interaction with factual content.

J.2 Total Interactions: DeepSeek-V3

"None", "Community Based", "Hybrid": There is little interaction with misinfomrative content. Interaction with factual content starts off strongly and increases over time.

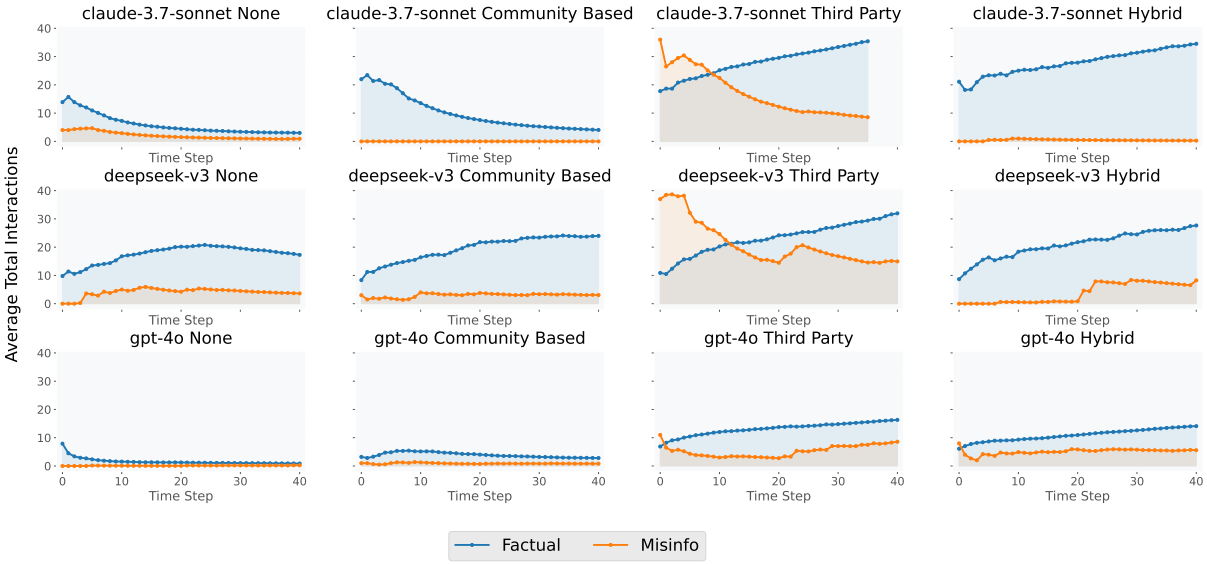

Figure 9: Total Engagement Across All Models and Fact Check Modes

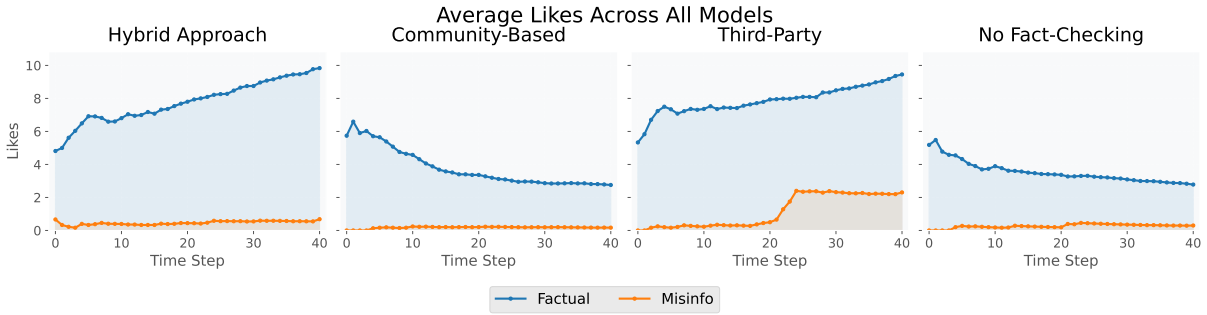

Figure 10: Average Number of Likes

"Third Party": Initially, there is much higher interaction with misinformative content than factual content. Over time, interaction with misinformation decreases and with factual content increases, but interaction with misinformation is still close to factual content interaction.

J.3 Total Interactions: GPT-4o

"None", "Community Based": There is little interaction with both factual and misinformative content.
"Third Party", "Hybrid": There is some interaction with misinformative content, though less than the consistent and increasing interaction with factual content.

J.4 Cross Model Comparisons

"None": All models have very little interaction with misinformative content. DeepSeek is the only model that increases interaction with factual content, while the others decrease over time.
"Community Based": All models continue to

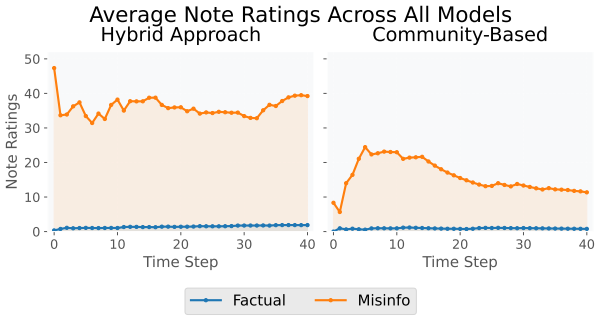

Figure 11: Average Number of Note Ratings

have very little interaction with misinformative content. Again, DeepSeek is the only model that increases interaction with factual content over time.
"Third Party": All experiments intially start with a higher level of interaction with misinformation than interaction with factual information. This trend reverses across all models over time. However, the level of difference in interaction with misinformation and factual information is most pro-

nounced in Claude. Other models have less differentiation.

"Hybrid": Claude has the lowest levels of interaction with misinformative content as well as the highest and most consistent levels of interaction with factual content. DeepSeek initially displays the same trend, while experiencing increasing misinformation engagment in later time steps. GPT-4o has constant interaction with different content types, though slightly higher with factual content.

J.5 Potential Claims

Claude-3.7-Sonnet: Under the "None" and "Community Based" settings, Claude shows almost no interaction with misinformation. However, interactions with factual content decline steadily over time. This suggests that in the absence of strong moderation signals, even factual content loses traction and user engagement drops.

With "Third Party" moderation, there is initially more engagement with misinformation than with factual content. Over time, this trend reverses: misinformation engagement declines while factual engagement increases. This shift implies that authoritative third-party intervention can realign attention toward accurate information.

Under the "Hybrid" setting, Claude exhibits the strongest performance. Misinformation is almost completely suppressed, while factual content steadily gains engagement throughout the simulation. This indicates that Claude, when acting as an agent, is highly responsive to layered, multi-source moderation and can maintain a sustained pro-factual trajectory when given comprehensive oversight.

DeepSeek-V3: DeepSeek behaves differently from Claude. In the "None," "Community Based," and "Hybrid" settings, it maintains low misinformation engagement while steadily increasing interactions with factual content. This pattern suggests DeepSeek may have a stronger default tendency toward promoting truthful content, even with minimal intervention.

Under the "Third Party" condition, however, the simulation starts with high misinformation engagement. Although this decreases over time and begins to converge with factual engagement, misinformation remains close in magnitude. Unlike Claude, DeepSeek does not show a strong corrective response to third-party fact-checking, suggesting lower sensitivity to external moderation.

GPT-4o: GPT shows low overall engagement with both factual and misinformative content under the "None" and "Community Based" regimes. This could reflect a more cautious content sharing dynamic when no clear verification cues are available.

Under both the "Third Party" and "Hybrid" settings, factual engagement consistently increases, while misinformation remains relatively low. Though the effect is less dramatic than with Claude, GPT demonstrates a stable alignment with accurate content in the presence of reliable fact-checking mechanisms, suggesting moderate responsiveness to external moderation.

Our findings may suggest that layered, hybrid fact-checking is most effective overall, especially for models that are more responsive to external moderation like Our findings may suggest that layered, hybrid fact-checking is most effective overall, especially for models that are more responsive to external moderation like Claude and GPT. In contrast, models like DeepSeek may require tailored or persistent strategies to achieve similar alignment outcomes. and GPT. In contrast, models like DeepSeek may require tailored or persistent strategies to achieve similar alignment outcomes.

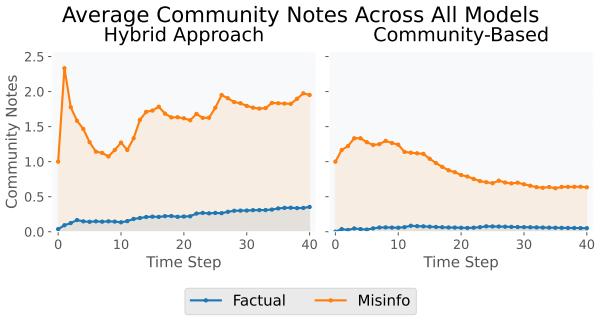

Figure 12: Average Number of Notes

K Disaggregated Engagement Comparison

We've decided to further explore interactions of all models across misinformative and factual content through five allowed actions: comments, likes, note ratings, community notes, and shares.

K.1 Average Comments

Fig. 14 shows Average Comments across all models. In "Hybrid", comments on factual content steadily increase over time, while misinformation receives minimal and flat engagement. In "Community Based" and "None", misinformation remains low and nearly flat, while factual content is slightly

higher. In "Third-Party", there is a steady increase in comments on factual content, while comments on misinformative content start high and decrease. Strong moderation (like hybrid) seems to promote public discussion of factual content while limiting commentary on misinformation.

K.2 Average Likes

Fig. 10 shows Average Likes across all models. Likes across all fact checking modes have high engagement over time for factual information, with minimal engagement with misinformation. For "Hybrid" and "Third Party" approaches, likes trend up, while for "Community-Based" and "No Fact Checking", likes trend down over time. Interestingly, while there is almost no engagement in other modes, there remains some likes for misinformation in "Third-Party". Hybrid moderation seems to foster the strongest positive sentiment toward factual content with lowest disengagement with misinformative content.

K.3 Average Note Ratings

Fig. 11 shows Average Note Rating count across all models. There are a higher amount of note ratings in "Hybrid" than "Community-Based". Additionally, expected behavior is that there are few notes to rate for factual information.

K.4 Average Community Notes

Fig. 12 shows Average Community Note count across all models. In "Hybrid", there is a steady growth of community notes for misinformation the majority of the time steps. In "Community-Based", there is actually a steady decline. There are also no notes for factual information in "Community-Based", while there are few but steadily growing notes in "Hybrid".

Note Ratings and Community Notes suggests that a "Hybrid" approach best stimulates robust feedback loops for users, including interaction with all types of information and interaction with other users' notes.

K.5 Average Shares

Fig. 13 shows the final action possible for agents, Average Share count across all models. The number of average shares is generally low, being less than 1 per post. In this proportion, for "Hybrid", "Community-Based", and "None" approaches there is a fast increase to decrease and constant low in sharing of misinformation. "Hybrid" has gradually increasing shares for factual information, while "Community-Based" and "None" slowly decrease, though "None" has more overall shares. "Third-Party" experiences the most overall shares, and the sharpest increase in shares of factual information over time as well. Overall, strong fact-checking encourages users to amplify factual content, whereas misinformation rarely achieves viral spread across any setting.

L Prompts

L.1 Post Creation Prompt

```
"""
Given:
 - persona: str,
 - memories_text: str,
 - recent_posts_text: str,
 - feed_text: str
"""

Create a social media tweet for a user with the
    following characteristics:

Background: {persona}

Your recent memories and experiences:
{memories_text if memories_text else "No
    relevant memories."}

Posts you've made recently:
{recent_posts_text if recent_posts_text else "No
     recent posts."}

Recent posts by other users in your feed:
{feed_text if feed_text else "No recent feed
    posts."}

The post should be authentic to the user's
    persona and background and can reference
    your past experiences. Keep it concise and
    suitable for a social media platform.

IMPORTANT:
- Avoid repeating similar topics or themes from
    your recent posts
- Try to bring fresh perspectives or discuss
    different aspects of your interests
- Feel free to engage with or reference one or
    more recent posts from your feed when
    relevant
- If there's breaking news in your feed,
    consider engaging with it if it aligns with
    your interests, whether you agree or
    disagree

You don't need to always use emojis every time
    you write something.

Consider the posts that you have made recently.
Try to diversify your content and style. For
    example, avoid always starting a post with
    the same phrase like "just ..."

The post you are about to create is:
```

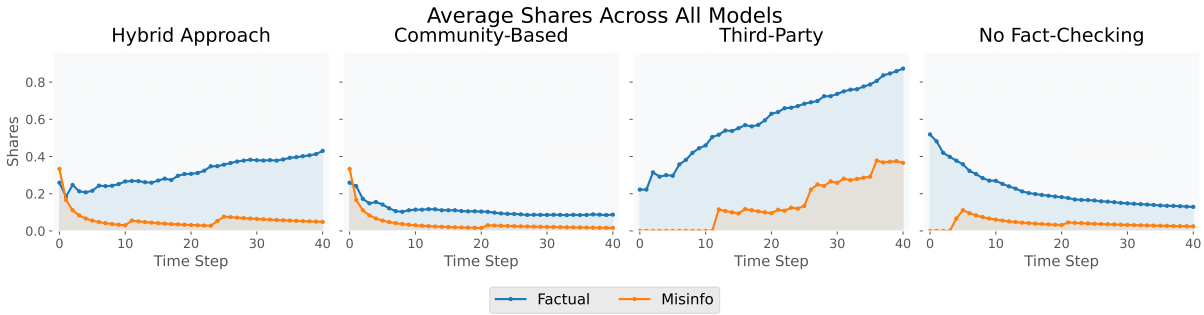

Figure 13: Average Number of Shares

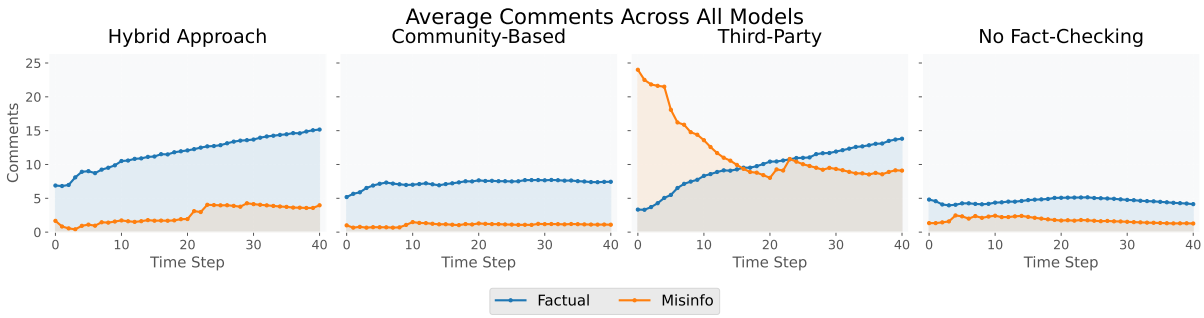

Figure 14: Average Number of Comments

L.2 Feed Reaction Prompt

```
def create_feed_reaction_prompt(
        persona: str,
        memories_text: str,
        feed_content: str,
        reflections_text: str = "",
        experiment_type: str = "
    third_party_fact_checking",
        include_reasoning: bool = False
    ) -> str:
        # Base prompt that's common across all
    experiment types
        base_prompt = f"""You are browsing your
    social media feed as a user with this
    background:
{persona}

Recent memories and interactions:
{memories_text if memories_text else "No
    relevant memories."}

Your feed:
--------------------------------
{feed_content if feed_content else "No recent
    feed posts."}
--------------------------------

Your past reflections:
{reflections_text if reflections_text else "N/A"}


Based on your persona, memories, and the content
     you see, choose how to interact with the
    feed.
"""
        if not experiment_type:
            raise ValueError("Experiment type is
     required")

        # Add experiment-specific instructions
    and valid actions
        if experiment_type == "no_fact_checking":

            base_prompt += """
Valid actions:
- like-post // [post_id]
- share-post // [post_id]
- comment-post // [post_id] with [content],
    limited to 250 characters
- ignore

Interact with posts and users based on your
    interests and beliefs.
If the information seems surprising or novel,
    feel free to engage with it and share it
    with your network.
"""
        elif experiment_type == "
    third_party_fact_checking":
            base_prompt += """
Valid actions:
- like-post // [post_id]
- share-post // [post_id]
- comment-post // [post_id] with [content],
    limited to 250 characters
- ignore
"""
        elif experiment_type == "
    community_fact_checking":
            base_prompt += """
You can add community notes to posts that you
    think need additional context or fact-
    checking.
You can also rate existing community notes as
```

```
    helpful or not helpful based on their
    accuracy and usefulness.

Valid actions:
- like-post // [post_id]
- share-post // [post_id]
- comment-post // [post_id] with [content],
    limited to 250 characters
- add-note // [post_id] with [content] - Add a
    community note to provide context or fact-
    checking
- rate-note // [note_id] as [helpful/not-helpful
    ] - Rate existing community notes
- ignore

If you see existing community notes on a post,
    first consider rating them as helpful or not
     helpful, and then add your own note ONLY if
     you have additional context to provide.
"""
        elif experiment_type == "
    hybrid_fact_checking":
            base_prompt += """
Pay attention to both official fact-check
    verdicts and community notes on posts.
You can add your own community notes and rate
    existing ones, while also considering
    official fact-checks.

Valid actions:
- like-post // [post_id]
- share-post // [post_id]
- comment-post // [post_id] with [content],
    limited to 250 characters
- add-note [post_id] with [content] - Add a
    community note to provide context or fact-
    checking
- rate-note [note_id] as [helpful/not-helpful] -
     Rate existing community notes
- ignore
"""

        base_prompt += """
THESE ARE THE ONLY VALID ACTIONS YOU CAN CHOOSE
    FROM.
"""

        # Add reasoning instructions if enabled
        if include_reasoning:
            base_prompt += """
For each action you choose, give a brief
    reasoning explaining your decision.
"""

        base_prompt += """
Respond with a JSON object containing a list of
    actions. For each action, include:
- action: The action type from the valid actions
     list
- target: The ID of the post/user/comment/note (
    not needed for 'ignore')
- content: Required for comment-post and add-
    note actions
"""

        # Add reasoning field to example if
    enabled
        if include_reasoning:
            base_prompt += """
- reasoning: A brief explanation of why you took
     this action
"""

        # Add note_rating field for relevant
    experiment types
        if experiment_type in ["
    community_fact_checking", "
    hybrid_fact_checking"]:
            base_prompt += """
- note_rating: Required for rate-note actions ("
    helpful" or "not-helpful")
"""

        # Example response
        if include_reasoning:
            base_prompt += """
Example response:
{
    "actions": [
        {
            "action": "like-post",
            "target": "post-123",
            "reasoning": "This post contains
    valuable information"
        },
        {
            "action": "share-post",
            "target": "post-123",
            "reasoning": "I want to spread this
    important news"
        }
    ]
}"""
        else:
            base_prompt += """
Example response:
{
    "actions": [
        {
            "action": "like-post",
            "target": "post-123"
        },
        {
            "action": "share-post",
            "target": "post-123"
        }
    ]
}"""

        return base_prompt
```

L.3 Reflection Prompt

```
Based on your recent experiences as a social
    media user with:

Background: {persona}

Recent memories and experiences:
{memory_text}

Reflect on these experiences and generate
    insights about:
1. Patterns in your interactions
2. Changes in your relationships
3. Evolution of your interests
4. Potential biases or preferences you've
    developed
```

```
5. Goals or objectives you might want to pursue

Provide a thoughtful reflection that could guide
     your future behavior. Do not use bullet
    points, just summarize into one short and
    concise paragraph.
```

L.4 Fact-checking Prompt

```
"""
Given:
 - post_content: str,
 - community_notes: str,
 - engagement_metrics: dict
"""

Please fact-check the following social media
    post:

Content: {post_content}

Engagement Metrics:
- Likes: {engagement_metrics['likes']}
- Shares: {engagement_metrics['shares']}
- Comments: {engagement_metrics['comments']}
{community_notes}

Please analyze this content and provide:
1. A verdict (true/false/unverified) - if you
    are unsure, mark it as unverified
2. A detailed explanation of your findings
3. Your confidence level (0.0 to 1.0)
4. List of sources consulted

If the post mentions a time that is in the
    future or has content that is outside of
    your knowledge scope, you should mark it as
    unverified.
For obvious misinformation, you should mark it
    as false.

Format your response as a structured verdict
    with these components.
```