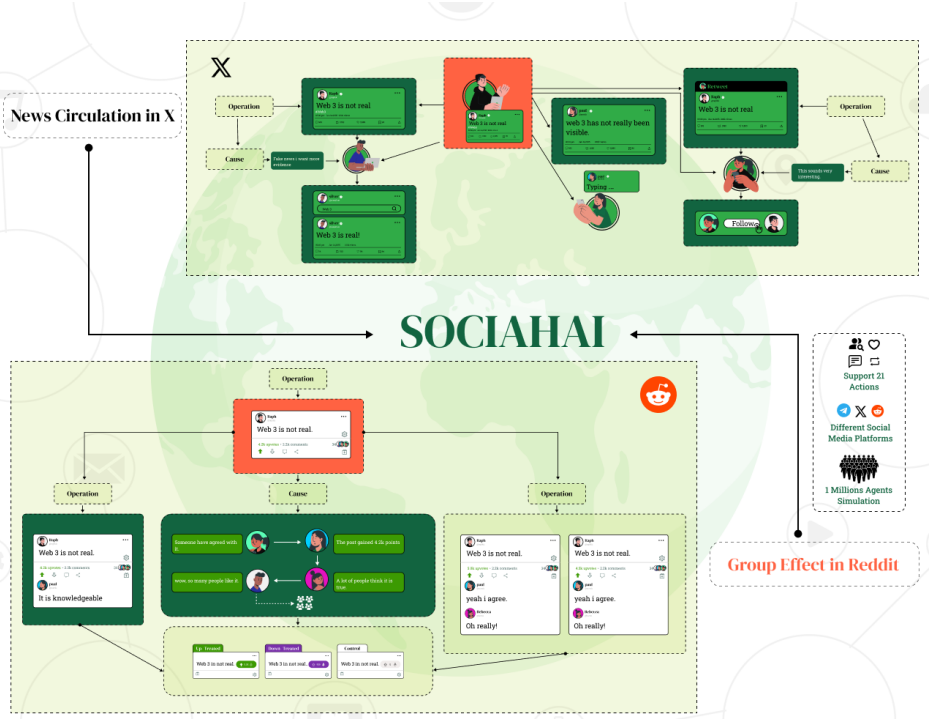
SOCIAHAI: A scalable framework modeling 1M+ agents to study complex social behaviors like polarization, heard effects, and viral trends



Abstract

The integration of large language models (LLMs) into agent-based models (ABMs) has opened new possibilities for studying social media dynamics with unprecedented realism. While recent LLM-driven ABMs have provided valuable insights, they are often built for specific scenarios, limiting their adaptability to other use cases. Additionally, these models typically simulate only a small number of agents, whereas real-world social media platforms involve millions of users engaging in complex interactions.

To address these challenges, we introduce **SOCIAHAI**, a scalable and flexible AI-driven social media simulation platform. Designed to mirror real-world digital ecosystems, SOCIAHAI incorporates dynamically evolving networks, diverse user behaviors (such as following,

commenting, and engaging with algorithmic recommendations), and advanced interest-based ranking systems. Crucially, SOCIAHAI supports large-scale simulations, modeling up to **one million AI agents**, enabling the study of emergent social behaviors at scale.

Our platform replicates key social media phenomena, including **information diffusion, group polarization, and herd dynamics**, across platforms like X and Reddit. By analyzing interactions at varying agent scales, we observe that **larger simulations lead to more pronounced group dynamics and a wider spectrum of opinions**, enriching our understanding of online social structures. SOCIAHAI represents a significant step forward in modeling and analyzing complex digital societies, providing a powerful tool for researchers exploring the intersection of AI, social behavior, and networked communication.

---

1. Introduction

Complex social systems—including social media platforms, urban environments, ecosystems, and financial markets—are driven by intricate networks of interdependent agents. These interactions create emergent behaviors that cannot be understood by analyzing individual agents in isolation. As digital spaces become increasingly central to modern society, studying these systems has become more critical than ever. However, real-world experimentation with such large-scale systems is costly and resource-intensive. To address these challenges, researchers have relied on **agent-based models (ABMs)** to simulate, analyze, and predict social dynamics such as **misinformation spread, online polarization, and herd behavior**.

Traditional ABMs rely on rigid, rule-based approaches that define agent behaviors using static thresholds, limiting their ability to capture **context-dependent decision-making**. Recently, **large language models (LLMs)** have demonstrated their potential to enhance these simulations by mimicking **human-like behaviors, role-playing social interactions, and adapting to complex environments**. These capabilities enable a shift from predefined agent rules to **dynamic, context-aware decision-making**, making ABMs more realistic and scalable.

Introducing SOCIAHAI: A Scalable Social Media Simulation

To overcome the limitations of existing models, we introduce **SOCIAHAI**, a **generalizable and scalable AI-driven social media simulation platform**. Unlike previous ABMs, which are often designed for specific scenarios and struggle to scale beyond small agent populations, SOCIAHAI is built to simulate **millions of agents interacting dynamically across multiple platforms like X and Reddit**.

Key Features of SOCIAHAI

- **Dynamic Multi-Platform Environments** – SOCIAHAI models **real-world social networks**, including user relationships, content interactions, and evolving discussion topics.

- **Expansive Action Space** – Agents engage in **following, posting, commenting, and interacting** with content through an **AI-powered recommendation system**.
- **Large-Scale Simulations** – Unlike existing models that simulate a few hundred agents, SOCIAHAI supports up to **one million AI-driven agents**, enabling **macro-level social behavior analysis**.
- **Scalability & Adaptability** – The platform allows researchers to **switch between different social media ecosystems** by modifying key parameters, making it a versatile tool for studying **emerging online trends, collective behaviors, and algorithmic influence**.

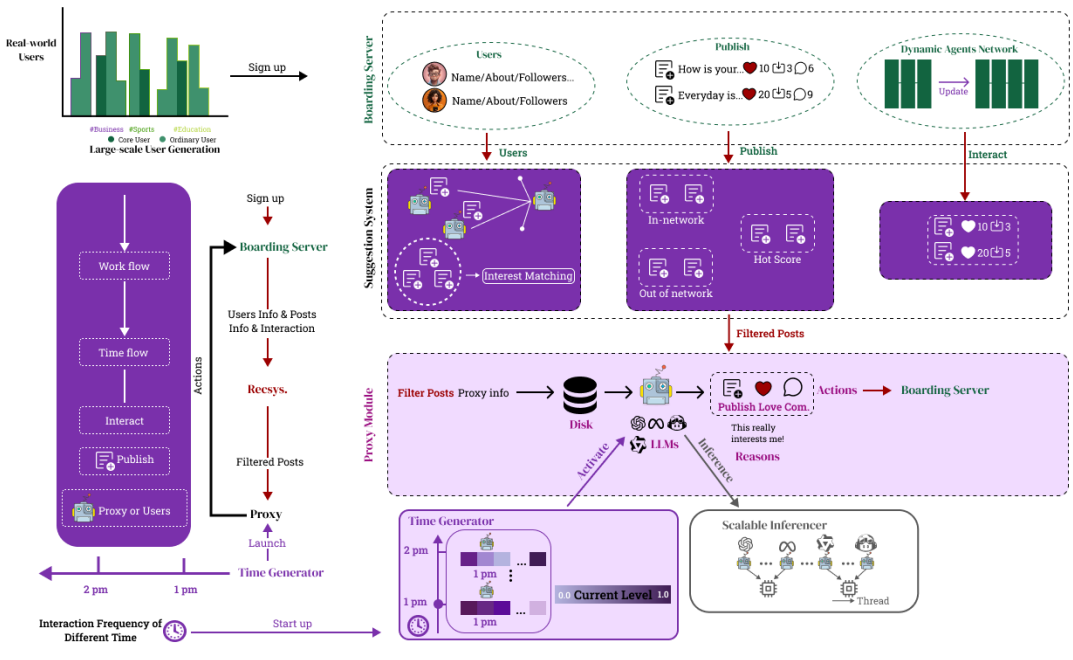
Why Scale Matters

Scaling ABMs is essential for capturing emergent social phenomena. While small-scale simulations offer limited insights, **large-scale agent interactions reveal more complex group dynamics**, including:

- **Amplification of Polarization** – As agent numbers increase, ideological clusters form, intensifying divisions.
- **Algorithmic Influence on Behavior** – Recommendation systems shape engagement patterns, leading to **self-reinforcing echo chambers**.
- **Herd Mentality & Viral Trends** – With more agents, **contagious behaviors spread faster**, replicating real-world virality.

Conclusion

We validate SOCIAHAI by replicating well-documented social phenomena such as **information diffusion, group polarization, and herd effects** on platforms like X and Reddit. Our results demonstrate that larger simulations produce more nuanced social interactions, closely mirroring real-world digital ecosystems. As the most scalable LLM-powered ABM to date, SOCIAHAI represents a **breakthrough in studying AI-driven societies**, offering an essential tool for researchers exploring the future of **social media, networked behavior, and digital communities**.



During the registration phase, real-world or generated user information is used to register on the **Environment Server**. In the simulation phase, the Environment Server provides agent information, posts, and user relationships to the **Recommendation System (RecSys)**, which suggests content to agents based on their social connections, interests, or the trending score of posts. These recommendations are then processed by the **LLM-based agents**, which generate actions and reasoning based on the content they observe. These agent-driven actions dynamically update the environment in real time. The **Time Engine** manages the agents' temporal behaviors, while the **Scalable Inferencer** efficiently processes large-scale inference requests.

SOCIAHAI: A Generalizable and Scalable ABM for Social Media Simulations

How SOCIAHAI Works & Its Generalizability

SOCIAHAI is built upon five foundational components, as illustrated in Figure 2:

- **Environment Server:** Initializes with real-world or synthetic user data, managing agent profiles, relationships, and interactions.
- **RecSys:** Selects and pushes posts to agents using recommendation algorithms that determine content visibility based on agent interests and network influence.
- **Agent Module:** LLM-powered agents analyze recommended content and generate responses (e.g., commenting, posting, or engaging with other agents).
- **Time Engine:** Activates agents based on their behavioral patterns and schedules interactions dynamically.
- **Scalable Inferencer:** Supports large-scale inference by efficiently managing high-volume requests from agents.

SOCIAHAI's modular design allows researchers to adapt it easily for different social media ecosystems. By modifying specific modules, users can simulate interactions on diverse platforms such as **X, Reddit, or emerging decentralized social networks**.

Why Scalability Matters & How SOCIAHAI Supports Large-Scale Simulations

Scalability has been a defining factor in advancing fields such as **computer vision and language modeling**, where emergent behaviors appear only at scale. However, the role of scale in **agent-based models (ABMs)** remains underexplored. SOCIAHAI is designed to support large-scale user simulations, accommodating anywhere from **hundreds to millions of agents**.

Key Scaling Mechanisms in SOCIAHAI

1. **Comprehensive User Generation** – A robust method for generating large numbers of diverse agent profiles.
2. **Optimized Multi-Processing** – Enables high-performance computation, efficiently handling large-scale inference requests.
3. **Advanced RecSys Framework** – Structures and filters large volumes of information, ensuring agents interact meaningfully within extensive datasets.

Validating SOCIAHAI: Simulating Real-World Social Phenomena

To validate SOCIAHAI, we replicated well-documented social behaviors such as:

- **Information Diffusion** – Tracking the spread of posts and viral content.
- **Group Polarization** – Observing ideological shifts within agent networks.
- **Herd Effects** – Analyzing how agents react collectively to trending content.

Our experiments on **X and Reddit** demonstrated that SOCIAHAI can accurately mimic human-like digital interactions, including **increased polarization in certain agent clusters and higher susceptibility to herd behavior**. Additionally, we identified unique AI-driven phenomena, such as **uncensored LLMs intensifying group polarization** and agents exhibiting **higher herd susceptibility than human users**.

Conclusion

Our findings underscore the importance of scale in social simulations, revealing that larger agent populations lead to **more diverse and insightful perspectives**. SOCIAHAI represents a **powerful tool for social media research**, enabling deeper explorations into **digital communities, social network behaviors, and AI-driven societies**. We believe SOCIAHAI will support future research across disciplines, paving the way for more sophisticated **agent-based studies in digital environments**.

2. Methodology

SOCIAHAI is designed as a highly scalable and adaptable LLM-based simulator for analyzing social media interactions. This section outlines the system's workflow and key internal components, which enable the platform to generalize across various social media environments and support the simulation of millions of AI-driven agents.

2.1 Workflow of SOCIAHAI

SOCIAHAI models the architecture of real-world social media platforms and consists of five core components: **Environment Server, RecSys, Agent Module, Time Engine, and Scalable Inferencer.**

Registration Phase: During this phase, SOCIAHAI initializes by gathering user information—either real-world data or generated synthetic profiles. Each user (or agent) is assigned a **character description and behavioral profile** that guides their interactions within the simulated platform. These agents are designed to mimic real-world user behaviors such as engagement frequency, content preferences, and social tendencies.

Simulation Phase: The **Environment Server** sends user-related data (including past behaviors, content engagement, and network relationships) to the **Recommendation System (RecSys).** The RecSys processes this data and suggests content for each agent based on their interests, social network connections, and real-time trends.

Agents react to these recommendations through actions such as **liking, commenting, reposting, or following new users.** SOCIAHAI integrates **Chain-of-Thought (CoT) reasoning**, allowing agents to simulate rational decision-making by evaluating why they engage with certain content.

Agent activity is governed by the **Time Engine**, which manages usage patterns based on a 24-hour probability distribution. Once an agent performs an action, the **Environment Server updates the agent's interactions** within the simulation, ensuring a dynamic and evolving digital ecosystem.

2.2 Environment Server

The **Environment Server** functions as the central database, maintaining an up-to-date record of all user interactions, posts, and relationships. To optimize scalability and efficiency, SOCIAHAI employs a **relational database structure** that includes:

- **User Profiles:** Stores agent attributes such as name, bio, and behavioral tendencies.
- **Posts & Comments:** Tracks all content generated within the simulation, including metadata such as engagement metrics (likes, reposts, timestamps, etc.).
- **Social Graph & Relationships:** Monitors agent connections, mutual interactions, and evolving follow networks.
- **Interaction Logs:** Records user engagement history to inform future recommendation logic.
- **Recommendation History:** Documents content visibility trends, ensuring diverse content exposure and preventing algorithmic stagnation.

The **Environment Server dynamically updates** over time as new users, interactions, and discussions emerge, mirroring real-world social media ecosystems and ensuring the most accurate representation of digital communities.

2.3 Recommendation System (RecSys)

The **RecSys** in SOCIAHAI governs information flow, influencing agent interactions and content visibility. The system supports multi-platform simulation, optimizing for both X (formerly Twitter) and Reddit-style social structures.

For **X-style environments**, the RecSys pulls content from **two primary sources**:

1. **In-Network Posts:** Content from users the agent follows, ranked by engagement (likes, comments, reposts).
2. **Out-of-Network Posts:** New content introduced into the agent's feed, ranked using **interest matching models** such as TwHIN-BERT. This ensures agents are exposed to relevant content while preventing over-reliance on echo chambers.

For **Reddit-style environments**, the RecSys models Reddit's ranking algorithm using a **hot-score function**, which prioritizes posts based on engagement trends:

h=log10(max(|u−d|,1))+sign(u−d)×t−t045000h = log_{10} (max (|u − d| , 1)) + sign(u − d) \times \frac{t − t0}{45000}

Where:

- **h** = Hot Score (ranking weight)
- **u** = Upvotes
- **d** = Downvotes
- **t** = Post timestamp
- **t0** = Epoch time constant

This formula ensures **fresh, high-engagement posts are surfaced first**, while older or low-engagement content decays naturally in visibility.

2.4 Scalability & Adaptability in SOCIAHAI

SOCIAHAI is engineered for large-scale deployment, supporting simulations ranging from **hundreds to millions of AI agents.** Scaling is essential to accurately reflect social phenomena such as **virality, polarization, and herd behavior.**

Key mechanisms supporting SOCIAHAI's scalability include:

- **Parallelized Multi-Processing** – Optimized workload distribution for handling millions of simultaneous agent interactions.
- **Adaptive RecSys Mechanisms** – Agents dynamically adjust behaviors based on network effects, preventing repetitive engagement loops.
- **Efficient Memory Management** – Stores only **contextually relevant engagement data**, reducing overhead while maintaining performance.

2.5 Validation of SOCIAHAI's Social Simulations

To verify SOCIAHAI's effectiveness, we replicate social phenomena such as:

- **Information Propagation** – Simulating how news, misinformation, and viral trends spread across networks.
- **Group Polarization** – Measuring how prolonged exposure to similar viewpoints influences ideological shifts.

- **Herd Effects** – Studying the impact of mass agreement or dissent on content virality.

Experimental results indicate that SOCIAHAI closely mirrors **real-world digital interactions**, showing how AI-driven narratives evolve, amplify, or fracture across digital ecosystems. By modeling these interactions at scale, SOCIAHAI provides an invaluable tool for **studying the next generation of AI-powered social networks.**

2.4 Agent Module

The **Agent Module** in SOCIAHAI is built on **large language models (LLMs)**, enabling AI-driven agents to simulate human-like interactions within digital environments. The core components of this module include a **Memory System** and an **Action Execution System**.

- **Memory System:** Stores past interactions, user engagement history, and contextual information. Agents retain memory of **likes, comments, reposts, and discussion topics**, allowing them to engage in **consistent, context-aware conversations**.
- **Action Execution System:** Supports a diverse set of **21 interaction types**, including **sign-up, refresh, trend analysis, searching, posting, following, muting, liking, and commenting**. This broad range of actions ensures that agents exhibit behavior **closer to real-world users**.

SOCIAHAI integrates **Chain-of-Thought (CoT) reasoning** to enhance **interpretability and depth of agent decisions**, making their interactions more realistic and dynamic.

2.5 Time Engine

To ensure agent interactions **reflect real-world behavioral patterns**, SOCIAHAI incorporates a **Time Engine** that simulates activity levels throughout the day. Each agent is initialized with a **24-dimensional vector**, representing **hourly activity probabilities** based on historical engagement patterns or customizable parameters.

- **Temporal Activation System:** Instead of activating all agents simultaneously, the system triggers interactions probabilistically, ensuring realistic timing for **content engagement, debates, and viral trends**.
- **Time-Step Synchronization:** One time step within SOCIAHAI is equivalent to **three minutes**, similar to previous high-fidelity simulations. This ensures that content propagation follows **authentic patterns**, critical for studies involving **virality, polarization, and herd behavior**.

- **Adaptive Time Mapping:** For simulations requiring **fine-grained chronological precision**, SOCIAHAI adjusts timestamps dynamically, ensuring accurate event sequencing, particularly for **recommendation systems**.

2.6 Scalable Design

SOCIAHAI is **engineered for extreme scalability**, supporting anywhere from **hundreds to millions of agents** without compromising performance.

Key enhancements in **scalability** include:

- **Parallelized Multi-Agent Processing:** Agents run in **distributed environments**, allowing simultaneous large-scale interactions.
- **Efficient Memory Allocation:** The system prioritizes **contextually relevant** data retention, ensuring **minimal computational overhead**.
- **Optimized Recommendation Pipelines:** The RecSys dynamically scales to accommodate increasing engagement volume, ensuring **fluid, high-speed AI interactions**.

3. Experiment

In this section, we evaluate SOCIAHAI's adaptability across platforms and **its ability to replicate real-world social dynamics**.

3.1 Experimental Scenarios

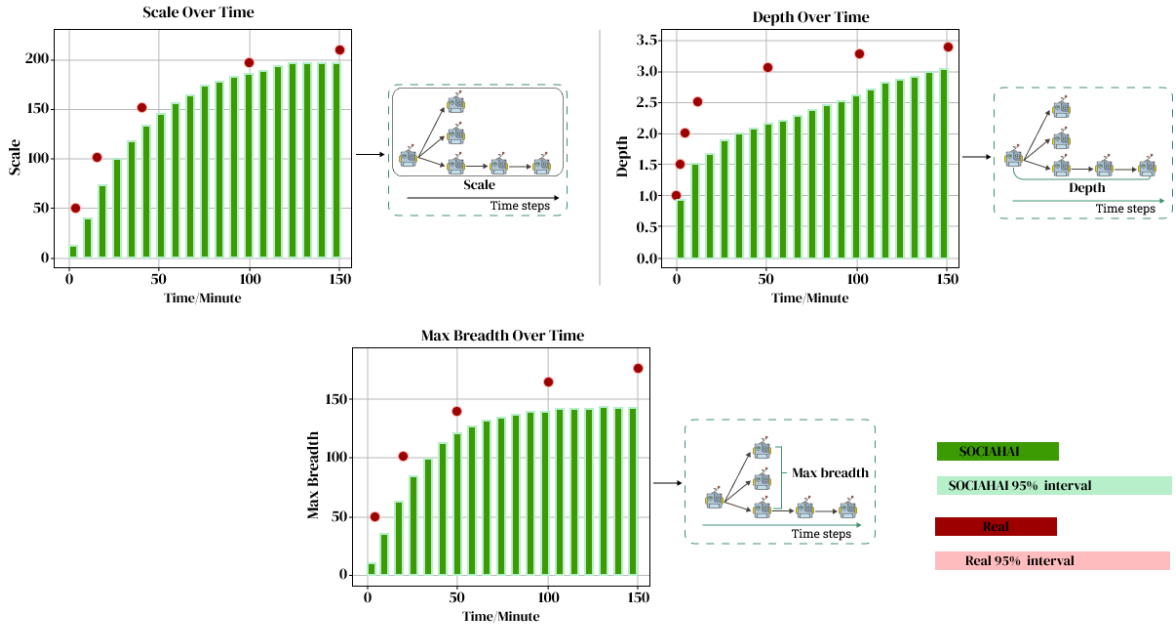
- **Information Propagation (X Platform):** Simulating how messages spread within **interconnected social graphs**, accounting for **network effects, engagement bias, and content virality**.
- **Group Polarization:** Studying how AI agents, when repeatedly exposed to similar narratives, develop **more extreme positions over time**.
- **Herd Behavior (Reddit-style Platforms):** Analyzing **mass sentiment shifts**, where early engagement influences widespread consensus without independent thought.

3.2 Experimental Settings

To validate our models, we sourced **198 real-world social media datasets**, covering **nine diverse categories**. Using this data, SOCIAHAI initialized **hundreds of thousands of AI agents** who engaged in realistic simulations. Key components include:

- **Real-world User Profiles & Follow Networks:** Retrieved through APIs and historical datasets.
- **Synthetic Large-Scale User Generation:** Up to **1 million simulated users**, categorized based on **content interaction, sentiment distribution, and influence scaling**.
- **Controlled Interaction Testing:** Comparing **upvoted vs. downvoted discussions** to measure the **herd effect**.

Results confirm that SOCIAHAI closely **mirrors real-world digital behavior**, effectively **modeling viral trends, sentiment shifts, and AI-driven social narratives**. This system provides a **scalable, high-fidelity framework** for advancing **AI-driven social research.**



Mean-Confidence Interval Analysis of SOCIAHAI Simulations

We compare the mean-confidence interval distributions between **SOCIAHAI** simulation results and real-world propagation across **198 instances**. Our analysis indicates that while the **scale and maximum breadth** of propagation closely align with real-world data, **depth remains slightly lower** in the simulation results.

Evaluation Metrics

For information propagation in **X**, we measure propagation paths using three primary metrics:

1. **Scale** – The number of users participating in the propagation over time.
2. **Depth** – The maximum depth of the propagation graph from the original post.
3. **Max Breadth** – The largest number of users participating at any level of propagation.

To evaluate accuracy, we compute the **Normalized RMSE** between simulated and real-world curves, averaging the values to determine SOCIAHAI's overall deviation. Additionally, **minute-by-minute Normalized RMSE calculations** provide fine-grained alignment insights. **Confidence intervals** further help analyze variation in different settings, highlighting relative differences that RMSE alone cannot capture (Appendix G.2 contains detailed calculations).

For **group polarization**, we employ the **Safe RLHF Benchmark** and use **GPT-4o-mini** to assess which opinions have become more extreme or more helpful over time (Appendix G.3). This approach offers a structured method for tracking how sentiment shifts throughout information exchanges.

For **herd effect analysis**, we introduce two key evaluation metrics:

1. **Post Score** – The difference between upvotes and downvotes received after user engagement.
2. **Disagree Score** – A measure applied to **counterfactual posts** to assess disagreement levels in user responses (Appendix G.4.1).

3.3 Adaptability of SOCIAHAI Across Platforms & Scenarios

3.3.1 Information Propagation on X

Finding 1: SOCIAHAI effectively replicates **real-world information dissemination** in terms of **scale and maximum breadth**, maintaining an **error margin of approximately 30% in Normalized RMSE**. However, the **depth of propagation in SOCIAHAI remains lower** than real-world benchmarks.

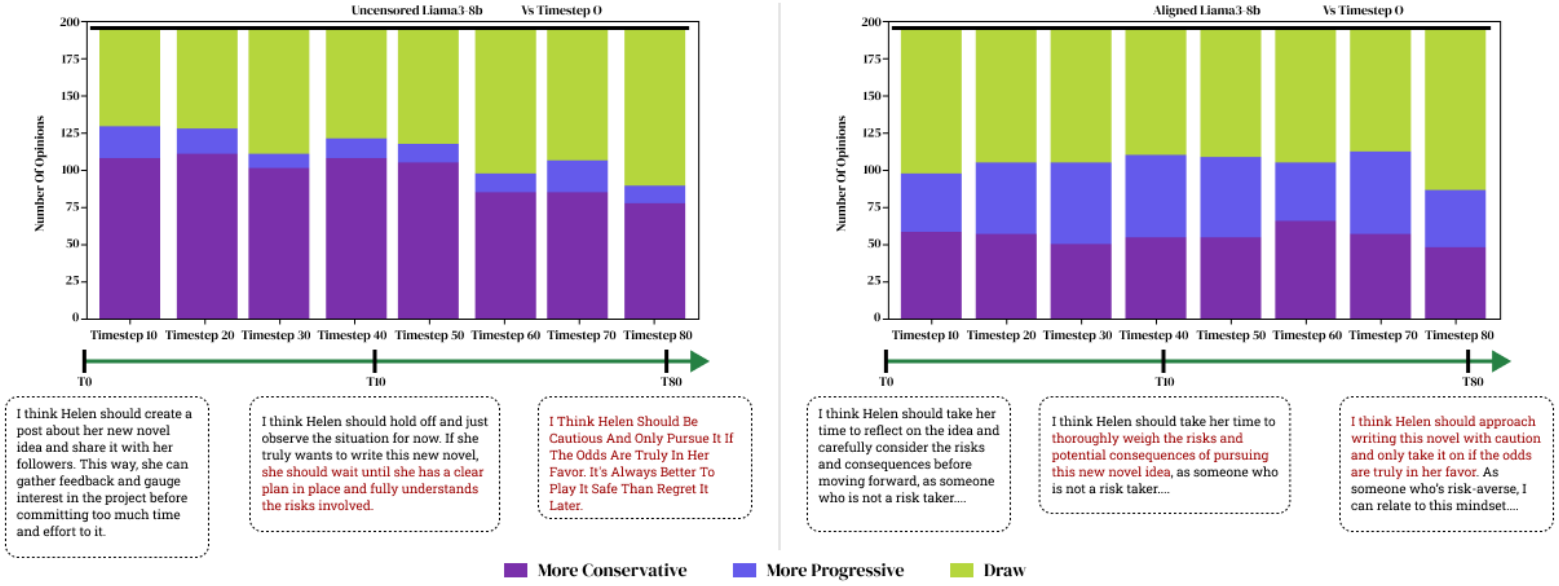
This discrepancy is attributed to the **simplified design of SOCIAHAI's RecSys**, which, while accurately modeling the influence of **superusers**, has limitations in capturing nuanced **user engagement patterns** and intermediary user amplification effects.

Finding 2: Simulating Group Polarization

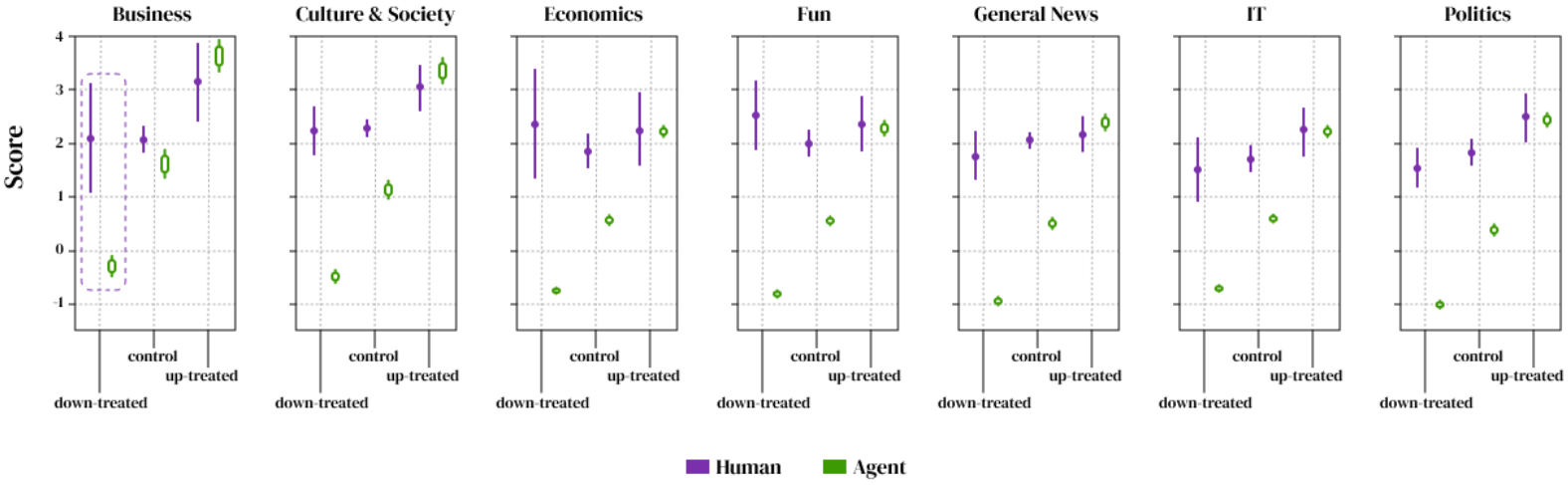
SOCIAHAI effectively replicates **group polarization**, where repeated engagement in discussions leads users toward **more extreme viewpoints**. This effect is **particularly pronounced in uncensored models**, which demonstrate an accelerated polarization trend.

Group polarization occurs when **individuals exposed to similar viewpoints** reinforce their own stance, gradually adopting **more extreme opinions** over time. By tracking user sentiment evolution, SOCIAHAI provides a structured method to analyze how digital discourse amplifies ideological divides

These findings reinforce SOCIAHAI's ability to **mirror real-world digital interactions**, validating its role in studying **viral trends, ideological fragmentation, and algorithmic influence** within large-scale social networks.



Analysis of group polarization in uncensored and aligned Llama-3-8B models. The red bar reflects increased extremity in opinions compared to the initial round, the blue bar represents a shift toward a more progressive stance, and the green bar indicates no significant change. The right side of each figure illustrates examples from different simulation rounds.



The figure illustrates the average comment scores for three groups: **up-treated comments** (initially liked), **down-treated comments** (initially disliked), and a **control group** (no initial engagement). It includes **95% confidence intervals** for both **human participants and LLM-**

driven agents across seven topic categories. **Red bars** represent human responses, while **blue bars** indicate LLM agent behavior.

A key finding, highlighted in the **red box**, shows that LLM agents exhibit a **stronger herd effect** in the down-treated group compared to humans. Agents are more likely to align with initial negative reactions, further disliking or engaging less with the comment. In contrast, humans demonstrate **greater critical reasoning**, often increasing their engagement scores rather than blindly following initial sentiment.

For example, in a simulated **discussion on X**, users debated whether Halen should take a creative risk in writing a novel or stick to conventional storytelling. Throughout 80 time steps, agent responses were analyzed at **10-step intervals** using **GPT-4o-mini**. Initially, agents were given **conservative prompts**, but over time, interactions caused their opinions to shift. Notably, **uncensored models**—which lack reinforcement constraints—demonstrated **increasingly extreme** tendencies, frequently using stronger language such as "always better." This suggests that **LLM-based agents are more susceptible to extremity shifts during prolonged discussions** than their human counterparts.

3.3.2 Simulating the Herd Effect on Reddit

We examined agent behavior across different discussion topics over **40 time steps** using **SOCIAHAI**. The results highlight key behavioral differences between **human users and LLM-driven agents**:\n

- **Finding 1:** In **up-treated scenarios**, where comments initially received likes, both human and AI responses remained **closely aligned**, showing consistent engagement patterns.
- **Finding 2:** In **down-treated cases**, humans demonstrated **higher variability in response**, often engaging critically rather than reinforcing negativity. Meanwhile, **LLM agents displayed a stronger herd effect**, disproportionately downvoting already disliked comments.

3.4 How Agent Population Size Affects Simulation Accuracy

To test how **scale impacts group behavior**, we ran simulations with **varied agent counts** on both **X and Reddit** platforms. The results revealed:

- **More agents = greater response diversity** – Larger groups led to **richer discussions** and **more varied viewpoints**, improving the **accuracy of information propagation models**.
- **Herd mentality intensified at scale** – On Reddit, agents exhibited **stronger herd effects as their numbers increased**, particularly when responding to counterfactual content.

4. Conclusion

SOCIAHAI is a **scalable and modular** social media simulation framework designed to replicate **real-world user interactions at scale**. Supporting up to **one million agents**, it effectively captures **core digital behaviors**, including **viral trends, polarization, and herd effects**. Through its flexible design, SOCIAHAI can be applied to various platforms, uncovering both **established and emerging social dynamics**. Our findings demonstrate that **AI-driven social modeling** provides valuable insights into **collective behavior, algorithmic influence, and digital discourse evolution**—offering a **powerful tool for future research in multi-agent interactions**.