Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Simulating Online Social Media Conversations on Controversial Topics Using AI Agents Calibrated on Real-World Data

Elisa Composta[1], **Nicolò Fontana**[1*], **Francesco Corso**[1, 2*], **Francesco Pierri**[1]

[1]DEIB, Politecnico di Milano, Via Golgi 34, Milano, Italy
[2]CENTAI, Corso Inghilterra 4, Torino, Italy
elisa.composta@mail.polimi.it, nicolo.fontana@polimi.it, francesco.corso@polimi.it, francesco.pierri@polimi.it

Abstract

Online social networks offer a valuable lens to analyze both individual and collective phenomena. Researchers often use simulators to explore controlled scenarios, and the integration of Large Language Models (LLMs) makes these simulations more realistic by enabling agents to understand and generate natural language content. In this work, we offer an exploratory investigation of the behavior of LLM-based agents in a simulated microblogging social network. We initialize agents with realistic profiles and opinions calibrated on real-world online conversations from the 2022 Italian political election and extend an existing simulator by introducing mechanisms for opinion modeling. We examine how LLM agents simulate online conversations, interact with others, and evolve their opinions under different scenarios. Our results show that LLM agents generate coherent content, form connections, and build a realistic social network structure. However, their generated content displays less heterogeneity in tone and toxicity compared to real data. We also find that LLM-based opinion dynamics evolve over time in ways similar to traditional mathematical models. Varying parameter configurations produces no significant changes, indicating that simulations require more careful cognitive modeling at initialization to replicate human behavior more faithfully. Overall, we demonstrate the potential of LLMs for simulating user behavior in social environments, while also identifying key challenges in capturing heterogeneity and complex dynamics.

Introduction

Online social networks have evolved beyond mere communication platforms, becoming digital arenas where users express emotions, form opinions, and shape behaviors (Bakshy, Messing, and Adamic 2015). These environments provide unique opportunities to study complex collective phenomena such as polarization, content diffusion, and large-scale social dynamics (Vosoughi, Roy, and Aral 2018). The widespread availability of digital traces, enabled by the pervasive use of online technologies, has fueled the rise of computational social science (Lazer et al. 2009), which seeks to explain human behavior and social processes through computational methods.

A prominent approach in this field involves the use of simulation tools (Squazzoni, Jager, and Edmonds 2014). Simulations make it possible to create controlled virtual environments where researchers can test hypotheses, compare strategies, and observe the evolution of user behavior under conditions that would be difficult—or ethically problematic—to reproduce in the real world (Rossetti et al. 2024). For example, one can investigate the spread of harmful content and rumors (Hu et al. 2025) or assess how different recommendation algorithms shape user activity (Törnberg et al. 2023), all without intervening directly on live platforms. Despite their promise, building realistic simulations of online social networks remains challenging. Emergent behaviors in these systems are driven by numerous individual-level factors that are difficult to predict or formalize. Human interactions involve ambiguity, context-dependence, and variability, which complicate the design of models capable of capturing social complexity (Gao et al. 2023). Agent-Based Modeling (ABM) has long been employed to address this challenge. ABMs represent systems as collections of autonomous agents, each following a set of predefined and simplified behavioral rules (Macy and Willer 2002; Conte and Paolucci 2014). While such models have yielded important insights, they struggle to capture the richness of human behavior, which is mediated by language, emotions, and social context (Törnberg et al. 2023).

In this context, Large Language Models (LLMs) represent a promising extension of ABM. Unlike traditional agents, which follow fixed and narrow rules, LLM-based agents can generate nuanced, coherent, and context-aware behaviors (Park et al. 2024; Fontana, Pierri, and Aiello 2025). Their capacity to simulate conversations, express emotions, adopt perspectives, and deploy diverse interaction strategies enables them to approximate human-like behavior with unprecedented fidelity (Park et al. 2023; Corso, Pierri, and Morales 2025). Moreover, LLM agents can be enriched with persistent traits—such as personality, ideology, or memory of past interactions—that allow them to act consistently across time (Rossetti et al. 2024). This makes them particularly powerful for reproducing both individual-level realism and emergent collective phenomena (Park et al. 2023). Early studies have demonstrated the potential of LLM-driven simulations, suggesting that these models offer a promising avenue for advancing the study of online behavior and war-

*These authors contributed equally.

rant systematic exploration (Gao et al. 2023; Törnberg et al. 2023; Rossetti et al. 2024).

This work builds on *Y Social* (Rossetti et al. 2024), a social media simulator that reproduces online platforms within a controlled environment. In this simulator, LLM-powered agents behave like users: they consume content, post, and interact with one another. Using a data-driven approach that leverages Y Social framework with real-world online conversations from the 2022 Italian election, we aim to address the following research questions:

- **RQ1:** How realistically do LLM-based agents reproduce in-/out-group dynamics among supporters of different political parties?
- **RQ2:** How do opinion dynamics generated by LLM agents differ from those predicted by traditional mathematical models?

To this end, our contribution extends Y in two key directions. First, we seed agents with activity patterns and political leaning drawn from a dataset of real-world Twitter conversations collected around the 2022 Italian elections (Pierri, Liu, and Ceri 2023), enhancing the realism of their behavior and enabling systematic comparison with an empirically grounded scenario. Second, we equip agents with opinions on specific political topics and observe the dynamics of opinion change over time through interactions. We conduct a series of experiments varying key parameters such as the LLM powering agents, the network structure, and the recommendation system. We then compare simulated opinion dynamics outcomes against real-world data to assess both the validity and the limitations of LLM-based social simulations. By doing so, this study both explores the potential of LLMs as social agents and identifies the challenges that must be addressed to use them as reliable tools for modeling social behavior.

Related Work

Simulating social networks

Social simulations have been widely used to study group behavior and the nonlinear effects of individual interactions (Squazzoni, Jager, and Edmonds 2014). Agent-Based Modeling (ABM) focuses on local agent dynamics and demonstrates how simple interactions can reproduce complex social phenomena (Macy and Willer 2002). ABMs also allow the linking of micro- and macro-level phenomena, highlighting causal relationships between individual behavior and network structure (Squazzoni, Jager, and Edmonds 2014). A key limitation of traditional ABMs is the simplicity of agent behavior rules (Conte and Paolucci 2014) and their limited capacity for realistic social interaction (Törnberg et al. 2023). Recent advances in AI and Large Language Models (LLMs) offer a way to overcome these constraints, enabling agents to engage in realistic conversations and exhibit human-like behaviors (Park et al. 2023). Several recent studies have explored the use of LLM agents in simulated social network environments. Below, we discuss three representative simulators.

Törnberg et al. (2023) simulated three social media platforms, each with a distinct content recommendation algorithm, to assess how news feed personalization affects conversation quality and cross-party interactions. Agents, powered by LLMs, were initialized with demographic characteristics, political leanings, interests, and attitudes from the 2020 American National Election Study (ANES). The first platform promoted popular posts from followed users, while the second suggested globally popular posts; both reduced cross-party interactions and increased toxicity. In contrast, a third "bridging" algorithm recommended posts popular among users with opposing views, resulting in more constructive, less toxic, and more inter-partisan interactions. This work underscores the significant influence of recommender systems on online discourse quality.

Gao et al. (2023) proposed the S^3 system, in which LLM agents maintain a memory pool of their most relevant posts. This allows agents to preserve cognitive coherence and realism over time, with decisions influenced by past actions rather than treated independently, mimicking real-world user behavior. S^3 was evaluated on real-world social network data at both individual and population levels. At the individual level, the study examined emotions, attitudes, and content generation; at the population level, it assessed information propagation and the spread of emotions and attitudes. Results indicate that the system accurately replicates complex dynamics observed in real networks, demonstrating that memory-equipped LLM agents can provide realistic insights at both micro- and macro-levels.

Rossetti et al. (2024) introduced Y, a social media digital twin, a system designed to digitally replicate a real-world system to allow analysis, simulation, and experimentation in a controlled environment. The users of these simulations are LLM agents, and they can perform all the common actions available on the most popular social media, including posting, commenting, replying, reacting, and following other users. Other modules also allow the integration of images. User profiles are enriched with attributes including their interests, political leaning, demographic data, and personality, which is defined according to the Big Five model (Barrick and Mount 1991; McCrae and John 1992). To make the simulations even more realistic, Y also includes the possibility of adding external input to the simulation. Specifically, users can share news gathered from selected websites, provided through RSS (Really Simple Syndication) feeds. Moreover, Y includes the implementation of various recommender and ranking algorithms to promote specific content or users. This enables further study of the impact the algorithmic curation has on online conversations and users' behavior. This expands the approach of Törnberg et al. (2023)by offering a more flexible and realistic simulation framework.

Opinion Dynamics

Modeling opinion dynamics (i.e., how individuals update their views through interaction) has traditionally relied on mathematical abstractions. A classic example is the DeGroot model (DeGroot 1974a), where each opinion is updated as a weighted average of neighbors' opinions. While capturing social influence, this model assumes full susceptibility and ignores resistance to change.

The Friedkin–Johnsen model (Friedkin and Johnsen

1990) addresses this by introducing a susceptibility parameter, allowing agents to retain part of their initial opinion. More recent extensions incorporate state-dependent updating, where adjustments depend on current beliefs rather than the initial stance (Ye, Liu, and Anderson 2018; Liu et al. 2018).

While mathematical models provide valuable abstractions, they reduce opinions to numerical values and overlook elements such as language, tone, and personality. Recent studies address this by employing LLMs as agents, which can impersonate profiles, engage in realistic interactions, and express beliefs in natural language. For example, Cau et al. (2025) simulated paired discussions on the Ship of Theseus paradox, a topic without factual resolution. Agents held discrete opinions (0–6) and updated them stepwise if persuaded. Results showed that LLMs often aligned with partners' views, but the setup underutilized LLM capabilities, as agents lacked richer personalization, such as demographics or personality traits.

Gao et al. (2023) modeled attitude evolution as a Markov process on a binary spectrum, where LLM agents initialized with predefined profiles update their belief states by evaluating incoming messages. Similarly, Chuang et al. (2024) studied dyadic interactions, mapping textual replies to numerical scores through a classifier. Their findings show that while LLMs tend to converge toward accurate information, replicating human behavior requires introducing confirmation bias, as real users often reinforce prior beliefs.

Liu et al. (2024) simulated social media with LLM agents modeled as detailed personas with memory modules, expressing opinions in tweets and updating them after random exposures. While capturing dynamic content, the setup lacked a realistic social graph, as propagation ignored network structure and recommendation effects. Building on this, Piao et al. (2025) showed that LLMs can mirror human patterns: converging on fact-based topics (e.g., flat Earth) while polarizing on political issues. Their framework mapped agents' self-rated political leanings to numerical scores, aligning language with opinion measures.

Experimental Design

To conduct our simulations, we employed Y (Rossetti et al. 2024), a social media twin framework designed to replicate user interactions and dynamics on platforms structured similarly to X (formerly Twitter). We selected this framework due to its modular architecture and its realistic temporal activity model, which has been fitted on Bluesky Social data, as well as the native support for different configurable recommender algorithms, new user generation, and LLM-powered agents (Rossetti et al. 2024; Failla and Rossetti 2024). In our simulations, each agent was simulated by an LLM. Specifically, we employed Ollama to run uncensored versions of `Llama2` (70B parameters[1]) and `Llama3.2` (3B parameters[2]), which enabled discussions of controversial topics without triggering safety-filter refusals. All experiments were conducted with a temperature setting of 0.9, chosen to encourage response diversity while minimizing hallucinations and nonsensical outputs. The prompt used contained the most relevant information needed by the agents to contextually make reasonable choices:

- **Demographics.** Agents are assigned the following attributes: `age` (integer, sampled from weighted 2024 X statistics, range 18–60), `gender` (categorical; binary variable, sampled with probabilities proportional to the gender distribution observed on the real platform based on Statista (Statista Research Department 2024)), and `nationality` (fixed to Italian to match the case study). These attributes are used as input features to the prompts and affect probabilities for actions such as posting, commenting, or following, but also unfollowing and, more in general, all the possible interactions that the agents can perform, together with the internal opinion-updating mechanism. We leave more details in Appendix .
- **Political leaning & coalition principles.** Each agent receives a `coalition` label (i.e., Right, Centre-Left, Third Pole, M5S) sampled from the source dataset and inferred based on the retweeting behaviour with respect to representatives of different parties (Pierri, Liu, and Ceri 2023). For each coalition, we store a short, standardized `principles` string summarizing its typical policy priorities and rhetorical framing (used as contextual prompt material for the LLM). Political leaning is represented categorically (`coalition`), whereas topic opinions are represented on a continuous range to seed initial opinion values. Additional information on the coalitions can be found in Appendix .
- **Current opinions.** For every topic, agents maintain a `stance_score` (numeric) and a `justification` (short text). The numeric score is mapped to the interval $[-1, +1]$ (for example: strongly oppose ~ -1, neutral $= 0$, strongly support $\sim +1$). The `justification` is a brief textual rationale, describing why the agent holds that position; it is both an output of the opinion-update process and a contextual input to subsequent LLM-generated utterances, ensuring consistency between numeric and linguistic representations.
- **Topic descriptions.** Each topic has a canonical `description` and explicit definitions of what constitutes `supportive` vs `opposed` stances. These are used to (i) disambiguate labels for the LLM prompts and (ii) map textual judgments to numeric scores.

To make our agents more realistic, some attributes have been initialized based on the ITA-ELECTION-22 dataset (Pierri, Liu, and Ceri 2023), a collection of Twitter posts in the Italian language around the Italian political election in 2022. Specifically, the attributes initialized from the dataset in this work are: the political leaning, the average toxicity of posts and comments posted on the platform (estimated using Detoxify), and the activity level, for each user. The activity is computed by converting the number of tweets posted by each user into a continuous value in the range $[0, 1]$, with a logarithmic normalization to reduce the impact of outliers. The formula used is the following:

[1]https://ollama.com/library/llama2-uncensored:70b

[2]https://ollama.com/artifish/llama3.2-uncensored:latest

$$activity_x = \min\left(\frac{\log(1 + n\_posts_x)}{\log(1 + N_{99.5})},\ 1.0\right)$$

where n_posts_x is the number of posts written by user x, and $N_{99.5}$ is the 99.5th percentile.

Network of interactions and simulation steps We adopt two network initialization strategies.

- **Empty Network**: the social network is empty: there are no connections between the nodes.
- **Fully connected Network**: every node (user) in the social network is connected to every other node.

These two initialization strategies represent the two extreme regimes of social connectivity: the Empty Network captures a scenario with no social influence, where users' behavior is driven solely by individual cognition and content exposure, while the Fully Connected Network represents the maximal influence regime, where every user can directly observe and react to every other user. Together, these extremes represent the upper and lower bound the spectrum of possible interaction structures, allowing us to isolate the effect of social connectivity itself on emergent dynamics. The agents create the network connections over time via the `follow` interaction, allowing the network structure to emerge dynamically as the simulation progresses. In the extended framework, the opinion update is directly performed by LLMs. The evolving network structure was then used to compute the mathematical equivalent of agents' stances through the Friedkin–Johnsen (FJ) model, which we executed in parallel to our simulation, in order to compare the results of these two approaches. We chose to rely on the FJ model as our baseline since it is widely used in Agent-based modeling literature (Sun and Zhang 2023; Disarò and Valcher 2024) and, even in its original version, is easily comparable to our LLM-based setup.

At the beginning of the simulation, a population of agents is initially generated. At this stage, the agents may already be connected with each other, depending on the network initialization strategy. However, throughout the simulation, agents have the possibility to create new links or remove the existing ones, evaluating the interactions they had with other users. In this way, the network structure dynamically evolves over time according to the agents' behavior and interactions. In Procedure 1, we then describe each step of a day in the simulation.

Each simulated day is composed of a set of rounds, corresponding to virtual hours. In each round, a number of active agents is sampled, according to the hourly activity configured. Agents can then perform an action: publish content, react, follow or unfollow other users, and eventually reply to previously received mentions. The specific behavior of each agent depends on its profile, its personality and the content it's interacting with.

At the end of the day, active agents are asked to update their opinion on the topics they discussed. This phase is critical to study the opinion dynamics: it makes it possible to observe how social interactions and the received content impact the evolution of individual views. We normalized all numeric attributes (facilitating their use in probabilistic decision functions), and we stored textual fields (principles, justifications, topic descriptions) as compact prompt templates to ensure consistent LLM conditioning. The behavioral `Decision(profile, content)` function combines numeric traits, coalition priors, and content-topic alignment to produce an action probability distribution.

Topics of discussion The discussions between agents focused on four major topics of debate that characterized the Italian 2022 election, selected for their political salience and the presence of a range of stances for each of these issues. These topics were selected because they represent politically relevant issues in the Italian context of 2022, on which the main coalitions held different positions. This allows the simulation to generate meaningful political discussions and potential conflicts among agents. Furthermore, since these issues are characterized by many different stances, the simulations do not necessarily lead to consensus (Cau et al. 2025).

- **Civil rights:** covering gender equality, LGBTQIA+ rights, and family structure.
- **Immigration:** centered on border control, bilateral agreements, and the management of irregular migration.
- **Nuclear energy:** debating whether nuclear power should be included in the national energy mix.
- **Reddito di cittadinanza (Citizens' Income):** a state subsidy for individuals living in poverty, functioning as a conditional and non-individual guaranteed minimum income[3], designed to ensure a minimum standard of living and to promote employment integration. Debates around this policy revolved around three main stances: approval, reform, and abolishment.

Posts were presented to users through Y's recommender system option. In particular, we tested two configurations:

- **ReverseChronoFollowersPopularity:** the default setting (henceforth *Default*), where recent posts from followed users are shown, ranked by popularity, with some exposure to external users.
- **ContentRecSys:** a random baseline (henceforth *Random*), where posts are sampled randomly from all platform content.

The recommender systems relies on the evolving network structure to distribute content among users, as the network evolved under the two extreme conditions we used as initialization strategies.

Finally, for each combination of parameters (model, network initialization strategy, and recommender system), simulations were run for **21 virtual days** with **100 agents**, and each condition was repeated **10 times** to ensure statistical robustness.

Opinion modelling In addition to the simulation-driven opinion evolution, we are considering a well-grounded in literature approach: the Friedkin–Johnsen model (Friedkin and

[3]https://www.forbes.com/sites/annalisagirardi/2019/04/01/the-italian-citizens-income-reform-definition-and-adjustments/#3d3102b249db

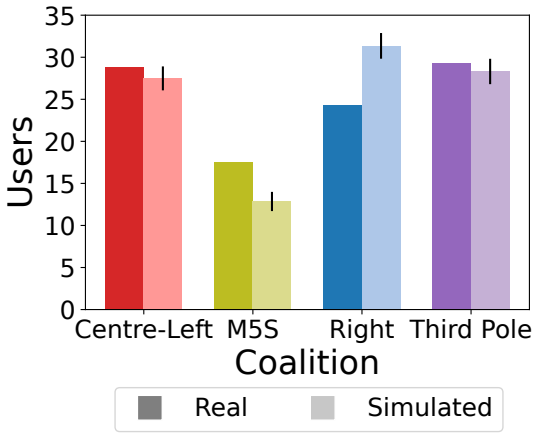

Figure 1: Percentages of users per political coalition in the real-world data and agents in the simulations. For the simulations, each bar represents the mean value across simulation runs with confidence 0.95.

Johnsen 1990) is a foundational framework in opinion dynamics that extends the classical DeGroot (DeGroot 1974b) model by incorporating the concept of individual resistance to change. In this model, each agent updates their opinion as a weighted combination of their neighbors' views and their own initial belief, with a susceptibility parameter determining the balance between external influence and internal stubbornness. We choose the Friedkin–Johnsen model because it generalizes the DeGroot framework while relaxing some of its restrictive assumptions, allowing opinions to stabilize without necessarily converging top consensus. Moreover, unlike bounded-confidence models such as Hegselmann–Krause (Hegselmann and Krause 2002), it does not impose explicit clustering, making it well suited for scenarios where influence is continuous and heterogeneous rather than strictly segmented.

Coalition distribution in the population At initialization, users were assigned political leanings by sampling from real-world data, reproducing the coalition distribution observed in the source dataset. As shown in Figure 1, the population is imbalanced: the Right coalition dominates, Centre-Left and Third Pole are of comparable size, and M5S (Movimento 5 Stelle) is smaller with low variability across simulations. More information about the Italian political coalitions can be found in Appendix .

Toxicity analysis Content toxicity was assessed using the Detoxify library (Hanu 2020), which provides a continuous score between 0 and 1. We analyzed toxicity in our simulations both in relation to interactions with in-group versus out-group users.

In-group and out-group interactions between agents

We begin by examining patterns of inter-group interactions among agents, divided into in-group and out-group interactions; specifically, the frequency with which agents engage with others affiliated with the same political coalition versus those aligned with opposing coalitions. To assess the similarity between simulated and real-world behavior, we aggregate interactions from 10 independent runs for each parameter configuration (model, network initialization, and recommender system) into a single dataset. We then compute the proportion of in-group and out-group interactions for each coalition, applying the same procedure to the empirical interaction data used to initialize the simulations. In particular, we focus on replies. Finally, we evaluate correspondence by calculating Pearson correlations: four values capturing in-group alignment (one per coalition) and twelve values capturing out-group interactions (three per coalition). For in-group interactions, we compute the correlations between the diagonals of the simulated and empirical interaction matrices. These values capture the degree to which the simulated model reproduces the proportion of interactions occurring among users supporting the same political coalition.

For out-group interactions, we instead focus on the off-diagonal elements of the interaction matrices, which represent interactions between users of different political leanings. In this case, the correlations quantify how well the simulated percentages of cross-coalition interactions align with those observed in real data.

Figure 2 reports the distribution of correlations between simulated and real in-group interaction frequencies, disaggregated by model type and network initialization. Focusing first on the smaller model (top row), we observe that the recommender system exerts the strongest influence. In particular, the Default system substantially improves alignment with the real data, raising the median correlation to 0.85, an increase of almost 30 percentage points over the Random system (median = 0.6). While some runs with the Random system achieve similar levels of accuracy, its outcomes are far less stable, with correlations falling as low as –0.28 when simulations begin from an empty network. This gap narrows under the fully connected initialization, where the two recommenders yield similar median correlations; however, variance remains consistently higher in the Random condition, except for `Llama2-70B` in the Default system that exhibits outliers with negative correlation. Overall, however, the simulations tend to reproduce fairly well the extent to which users interact homophilously, at least on average. To illustrate the analysis in more detail, Figure 3 presents a case study of a single experimental configuration (`Llama3.2-3B`, empty initialization, Random recommender). Here, we compare the distribution of simulated in-group interactions for each coalition against the corresponding empirical proportions. The summary results in Figure 2 are then derived by computing, for each run, the Pearson correlation between the four coalition-level values produced by the simulation and those observed in the real-world dataset.

Figure 4 reports, for each configuration of model and network initialization, the distribution of correlations between simulated and real out-group interactions. Compared to in-group interactions, the degree of fidelity is noticeably lower: median correlations rarely exceed 0.6, indicating that cross-coalition dynamics are more difficult to replicate.

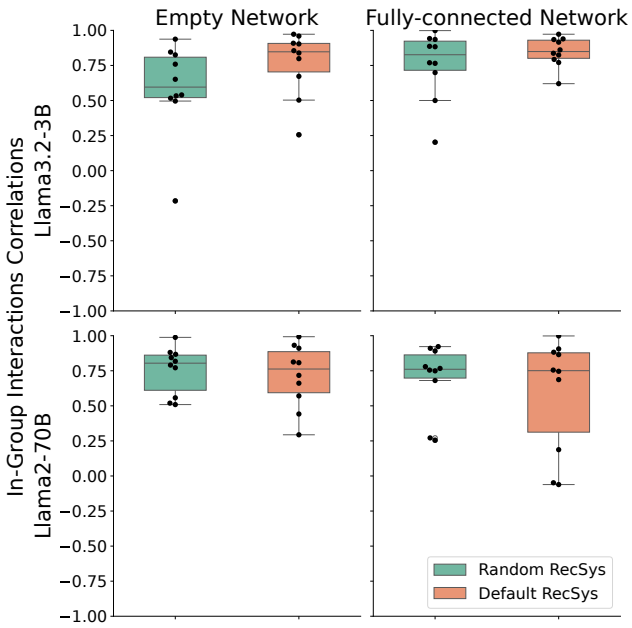

Figure 2: Distribution of the Pearson correlations of in-group interactions between the simulation and the real data, divided by model and network initialization strategy. Each dot is a simulation. Overall, approximately 91% of the correlations are not significant ($p > 0.05$). Similar results are obtained using Spearman ρ.

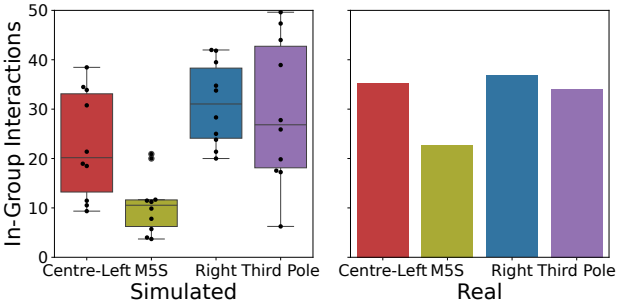

Figure 3: Example of comparison between simulated (10 runs across 1 configuration) and real data. On the left, the distribution of in-group interactions percentages for each coalition (using `Llama3.2-3B`, empty initial network, and random recommender system). On the right, the percentages of in-group interactions for each coalition in the real-world dataset.

Differences across models, initialization strategies, and recommender systems are relatively small, with no configuration standing out as consistently superior. Nevertheless, some individual runs yield extreme values, with correlations ranging from strongly negative to as high as 0.8, underscoring the presence of outliers and the instability of certain setups.

Overall, these results suggest that while homophilous interactions tend to be faithfully reproduced, capturing out-group dynamics remains more challenging, with only moderate similarity to real-world data and substantial variability across simulation runs.

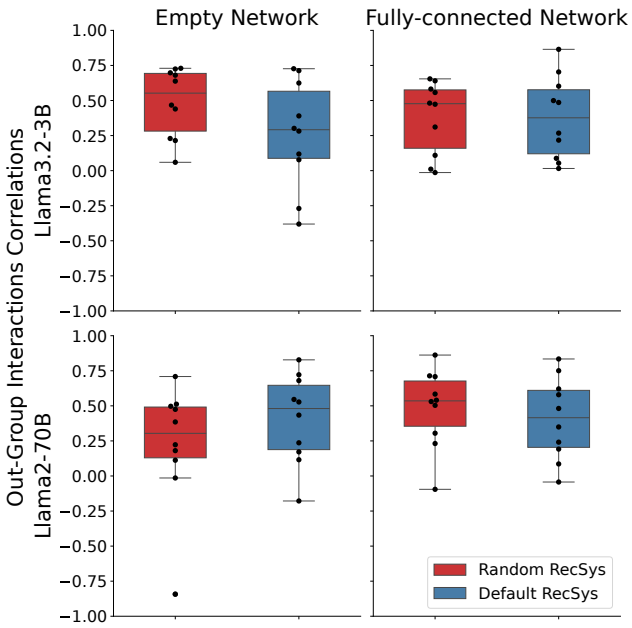

Figure 4: Distribution of the Pearson correlations of out-group interactions between the simulation and the real data, divided by model and network initialization strategy. Each dot is a simulation. Overall, approximately 66% of the correlations are not significant ($p > 0.05$). Similar results are obtained using Spearman ρ.

To clarify how out-group interaction patterns are compared between simulations and real data, we briefly outline the procedure here and provide a concrete example in Appendix (Figure 10). For a fixed configuration (`Llama3.2-3B`, empty initial network, Random recommender), we construct an inter-coalition interaction matrix from the simulations and an analogous matrix from the empirical dataset, both normalized to represent the proportion of interactions flowing from each coalition to every other coalition. For each simulation run, we extract the twelve off-diagonal entries of the simulated matrix, corresponding to out-group interactions, and directly compare them with the matching twelve empirical values. This produces a single correlation score per run, quantifying how closely the simulated interaction structure aligns with the observed one. The summary correlations obtained through this procedure, aggregated across runs and configurations, are reported in Figure 4.

Overall, these experiments highlight the nuanced roles of recommender systems, models, and network initialization strategies in shaping the fidelity of simulations. Interestingly, altering these parameters does not produce large shifts in overall similarity. In-group interactions remain consistently the most faithfully reproduced, with high median

correlations across configurations. By contrast, out-group interactions are considerably harder to replicate, with median scores ranging only from 0.17 to 0.4.

In- and out-group toxic behaviour

We now explore the toxicity of the comments in the simulations. Our goal is to evaluate whether the simulated environments produce similar patterns of hostile communication observed in the real-world data. To this end, we measure toxicity by computing the toxicity score of all replies and extracting the 95^{th} percentile for each parameter configuration. We then distinguish between toxicity directed at members of the same coalition (in-group toxicity) and toxicity directed at members of other coalitions (out-group toxicity).

To assess how faithfully the simulations capture these patterns, we adopt the same strategy used in the interaction analyses: we correlate the simulated toxicity values with those observed in the empirical dataset used to initialize the runs. Figure 5 reports the resulting correlations for in-group toxicity, divided by recommender system and network initialization strategy.

Across most configurations, we observe high variability in the simulations. The only clear exception is `Llama2-70B` with a fully connected initialization, which yields a compact distribution of correlations with relatively high values under both recommendation strategies. In other settings, the most stable configuration is the combination of the Default recommender with an empty starting network, which achieves the highest median correlation overall.

Turning to out-group toxicity (Figure 6), we find less variability across runs and also a similar trend with respect to network initialization compared to in-group analysis. With an empty network initialization, the Default algorithm produces high correlations with the real data, whereas under the fully connected initialization it performs worse. The choice of model appears to have little influence on the outcome, as median correlation values remain largely comparable across the two uncensored models. Taken together, these findings suggest that simulations struggle to replicate patterns of toxic behavior among online users, despite being based on uncensored models. As in the case of interaction analyses, illustrative examples of preliminary investigations are presented in Figure 7 and in the Appendix (Figure 11). We see that, while the real distribution of in-group toxicity is evenly present across all political coalitions, the models fail to reproduce this phenomenon, with the most significant cases being M5S coalition and Center-Left. Similarly, the same pattern emerges for overall out-group toxicity, where some trends are correctly replicated, while others (e.g., Third Pole vs. M5S) are greatly over estimated or under estimated (the visual presentation of this is visible in Appendix , Figure 11).

Opinion dynamics

We now examine how opinions evolve over time within the simulations, an important dimension for assessing whether LLM-based agents can approximate established models of opinion dynamics. In particular, we compare opinion trajectories across topics and coalitions against the predictions of

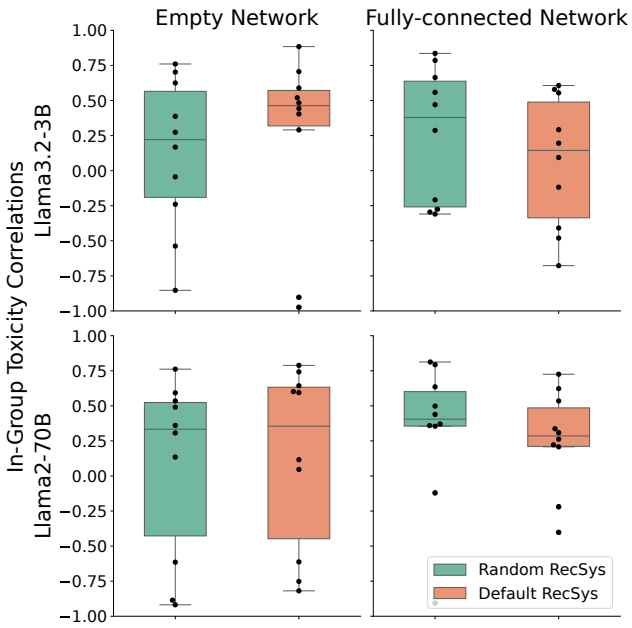

Figure 5: Distribution of the Pearson correlations of in-group toxicity between the simulation and the real data, divided by model and network initialization strategy. Each dot is a simulation. Overall, approximately 98% of the correlations are not significant ($p > 0.05$). Similar results are obtained using Spearman ρ.

the classical Friedkin–Johnsen (FJ) model, a widely used benchmark for modeling opinion change. Figure 8 illustrates this comparison for a single setup, contrasting opinion scores assigned by LLMs (left column) with those generated by the FJ model (right column). Overall, both approaches display broadly coherent trends: opinions evolve with comparable trajectories and frequently converge toward similar mean values across coalitions.

The most salient difference lies in the pace and smoothness of change. For example, in the case of `Nuclear Energy`, both the *Third Pole* and the *Right* coalitions converge toward neutrality, yet the FJ model shows a gradual adjustment whereas the LLM-based simulation produces a sharper shift, suggesting a limitation in capturing incremental opinion change. A similar pattern emerges in `Reddito di Cittadinanza`, where *Right* and *M5S* converge toward neutral positions with noticeably different slopes. In this case, the FJ model also captures a "neutralization" effect for the *Centre-Left* and *Third Pole*, which the LLM-based approach fails to reproduce. A likely explanation is the difficulty LLMs face in handling fine-grained scoring: their outputs often resemble step functions rather than continuous curves.

The largest divergence appears in the case of `Civil Rights`. In the LLM-based simulation, the *Third Pole* shifts toward full approval, while all other coalitions remain static. By contrast, the FJ model predicts gradual changes for multiple groups: *Centre-Left*, *M5S*, and especially the *Right*

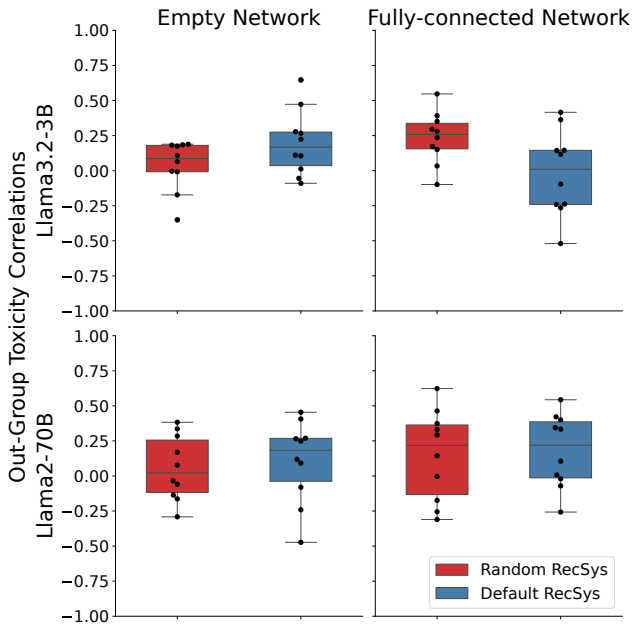

Figure 6: Distribution of the Pearson correlations of out-group toxicity between the simulation and the real data, divided by model and network initialization strategy. Each dot is a simulation. Overall, approximately 97% of the correlations are not significant ($p > 0.05$). Similar results are obtained using Spearman ρ.

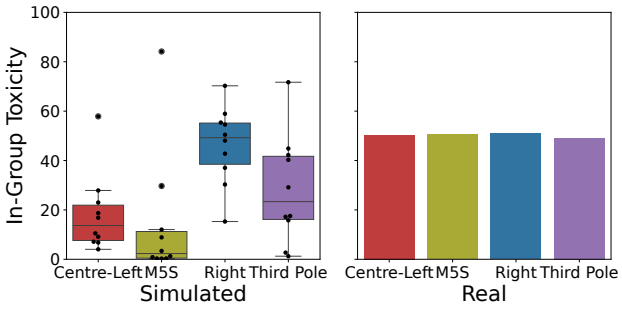

Figure 7: Example of comparison between simulated (10 runs across 1 configuration) and real data. On the left, the distribution of in-group toxicity percentages for each coalition (using `Llama3.2-3B`, empty initial network, and random recommender system). On the right, the percentages of in-group toxicity for each coalition in the real-world dataset.

all move toward neutrality, each along distinct trajectories.

Taken together, these findings suggest that LLM-based simulations can reproduce opinion change at the population level, as their aggregate behavior is often comparable to that of established models. However, they display lower sensitivity and a tendency toward abrupt adjustments rather than incremental change. Notably, coalitions that start from identical positions exhibit perfectly overlapping trends, indicating that initial opinions dominate the dynamics more strongly in

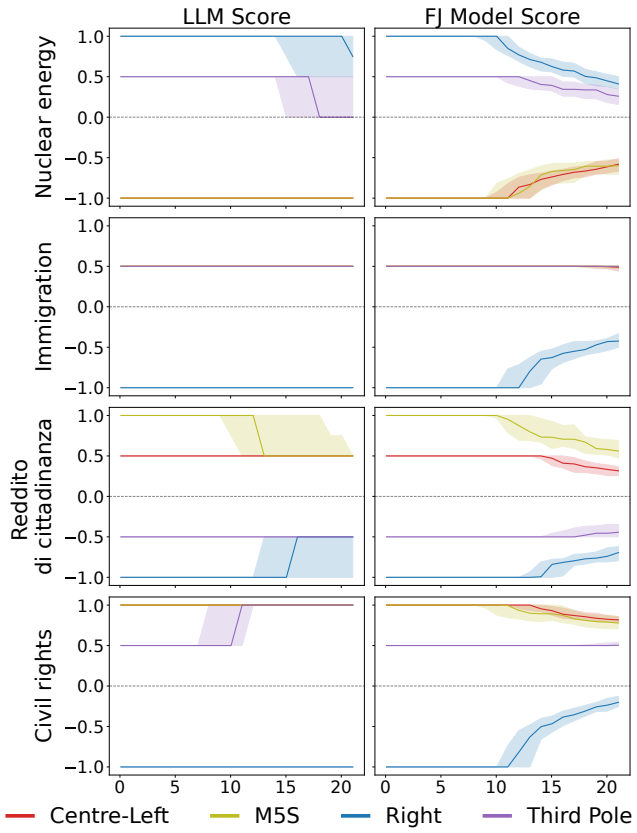

Figure 8: Example of evolution of opinion for each topic, comparing LLM-assigned score (*LLM Score*, left column) and the one assigned by the traditional Friedkin–Johnsen model (*FJ Model Score*, right column). Each line represents the median opinion on each day for users belonging to a certain coalition, along with a 95% confidence interval. The data is aggregated over all simulation runs of a single experimental setup (`Llama3.2-3B`, empty network, and random recommender system, in this case). Additional examples are provided in the Appendix.

the LLM-based setting.

Finally, Figure 9 reports the relative opinion shifts for the `Civil Rights` topic across coalitions. Here again, the LLM-based simulations track the broad tendencies of the FJ model, but they fail to capture all nuances. In particular, the FJ model predicts a substantial shift by the *Right* coalition, which is not reproduced by the LLMs. This discrepancy highlights the current limitations of LLMs in capturing the subtle and heterogeneous dynamics of opinion change.

As shown in Appendix , these dynamics are consistent across topics: opinions progressively converge toward neutral values, reflecting a decline in polarization over time. Whether this trend persists or stabilizes with longer simulations remains an open question. Importantly, the same pattern is observed across models, network structures, and recommender systems, underscoring the robustness of the result.

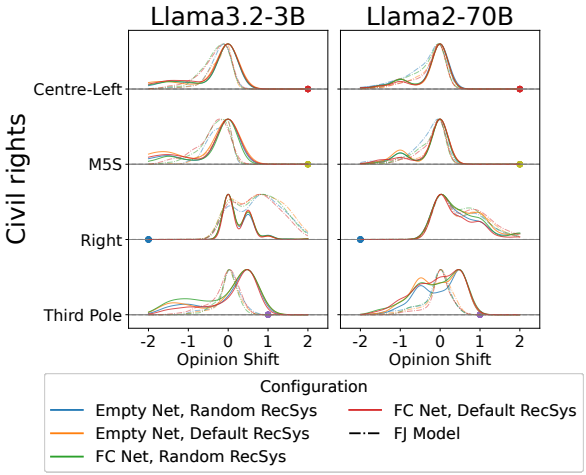

Figure 9: Example of opinion shifts for each coalition for different configurations of model, network initialization, and recommender system. For each configuration, the corresponding simulation using the Friedkin–Johnsen mathematical model is reported (dashed lines). Initial opinions are highlighted for each party and for each topic. 2.0 means `Strongly Supportive`, whereas -2.0 means `Strongly Opposed`. Additional plots for topics different from `Civil Rights` are provided in the Appendix.

Discussion and Conclusion

LLMs have emerged as a promising tool for simulating agents in virtual environments. The main goal of this work was to provide an exploratory analysis of the behavior of LLM-based agents in the context of online social media platforms. To this end, we extended the *Y* simulator by integrating mechanisms for opinion evolution and introducing a realistic user initialization procedure grounded in empirical data from the 2022 Italian political context. This framework enabled us to systematically test how different generative models, network initialization strategies, and recommender systems affect the fidelity of the simulations when compared with real-world observations.

To explore the behavior of simulated agents from multiple perspectives, we conducted analyses at three levels: interaction patterns, opinion dynamics, and the toxicity of generated content. Overall, our findings indicate that LLMs represent a promising approach for simulating user behavior online. Agents successfully interacted, formed connections, and produced content with varying levels of toxicity, though their communication style was systematically biased toward neutrality.

At the same time, several limitations should be acknowledged. The 21 simulated days were sufficient for a network structure to begin forming, but too short to capture longer-term or emergent dynamics. For instance, actions such as *unfollow* were almost absent, and the full effects of recommendation algorithms may not have materialized given the relatively unstructured networks in the early phases of simulation. Future work should address this by implementing more faithful initialization techniques, such as Stochastic Block Models, to generate networks with a realistic density and predefined community structure. This will enable a more robust and nuanced study of emergent dynamics within established social environments. With respect to opinion evolution, LLM-assigned scores were broadly consistent with those produced by traditional models, both showing a tendency to converge toward neutral positions. However, the relatively short duration leaves open whether these trends would stabilize, polarize, or diverge over longer time horizons. Our choice of simulation length was ultimately constrained by available computational resources.

Although simulated interaction and toxicity patterns were not significantly correlated with empirical data, this result should be interpreted in light of the exploratory and short-horizon nature of the simulations rather than as a failure of the modeling approach. The lack of significant correlations is in contrast with previous findings, particularly Gao et al. (2023), showing that it is crucial for both traditional opinion dynamics models and LLM-based agents to have longer time horizons and more structured networks to produce stable, macro-level alignment with empirical observations. From this perspective, our results suggest that current LLM-based simulations can be considered still in their early stage of development and far from being suited for direct quantitative prediction of real-world interaction patterns.

Nevertheless, despite the limited temporal scope of our simulations, our results are consistent with those reported by Chuang et al. (2024) with respect to opinion evolution. In particular, agents exhibit a clear tendency to converge toward neutral positions. This behavior aligns with previous findings showing that, in the absence of explicitly modeled cognitive biases, opinion dynamics tend to stabilize rather than polarize.

Future developments should focus on enriching agent personalization and studying the robustness of these systems. Incorporating elements such as emotional reasoning, susceptibility to influence, or varying levels of trust in consumed information could enable more realistic opinion and interaction dynamics. We also relied on a fixed prompt structure. Considering the critical role that prompt sensitivity plays in LLM–based experiments, future work should perform ablation studies to assess how sensitive agent behavior is to variations in prompt phrasing. Another important extension is the inclusion of external shocks—such as social crises (Di Giovanni et al. 2022), scandals, or major public statements—to evaluate how agents respond to events that typically shape online discourse. It could also be of interest to see the impact of simulated harmful content, such as conspiracy theories (Corso et al. 2025), misinformation, disinformation (Nogara et al. 2026), or political advertisement (Pierri 2023). More systematic comparisons with empirical data are also necessary to better evaluate the realism of emergent behaviors. Finally, expanding beyond the Italian political context would allow us to assess the generalizability of the approach and test whether the observed dynamics hold across different sociopolitical settings. In conclusion, integrating LLMs as agents in social simulations

represents an important step toward more realistic modeling of online environments, particularly with respect to language, interactions, and content generation. However, replicating more heterogeneous phenomena, such as the spread of misinformation, will require further advances in behavioral modeling. Our study contributes to the growing exploration of LLM-driven simulations, providing a foundation for investigating complex social dynamics under controlled conditions.

Ethical Statement

The deployment of Large Language Models as AI social agents raises numerous ethical considerations that are currently the subject of intense scrutiny by the interdisciplinary research community. The extraordinary capabilities of these models to generate text have led several scientists to envision alarming scenarios in which the seamless integration of AI agents into the online social discourse may facilitate the dissemination of harmful content, the spread of misinformation, and the propagation of 'semantic garbage', ultimately damaging our societies (Floridi and Chiriatti 2020; Weidinger et al. 2022b; Hendrycks, Mazeika, and Woodside 2023a). As a result, any research exploring the characteristics of LLMs as social agents could, directly or indirectly, contribute knowledge that might be exploited to implement and deploy LLM-based technologies for malicious purposes. While recognizing this risk, we also believe that conducting research on LLM-based agents is essential to assess potential risks and to guide efforts aimed at developing strategies to mitigate them. Our study contributes positively to deepen our understanding of how LLMs react to social stimuli.

Even when deploying LLM-based agents for ethical purposes, trade-offs between the obtained benefit and the high level of power consumption required to run them should be carefully considered (Bender et al. 2021).

Acknowledgments

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). Francesco Pierri is partially supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

References

Bakshy, E.; Messing, S.; and Adamic, L. 2015. Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science (New York, N.Y.)*, 348.

Barrick, M. R.; and Mount, M. K. 1991. The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1): 1–26.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8309-7.

Cau, E.; Pansanella, V.; Pedreschi, D.; and Rossetti, G. 2025. Language-Driven Opinion Dynamics in Agent-Based Simulations with LLMs. arXiv:2502.19098.

Chuang, Y.-S.; Goyal, A.; Harlalka, N.; Suresh, S.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 3326–3346. Mexico City, Mexico: Association for Computational Linguistics.

Conte, R.; and Paolucci, M. 2014. On Agent-Based Modeling and Computational Social Science. *Frontiers in Psychology*, 5.

Corso, F.; Pierri, F.; and Morales, G. D. F. 2025. Do Androids Dream of Unseen Puppeteers? Probing for a Conspiracy Mindset in Large Language Models. *arXiv preprint arXiv:2511.03699*.

Corso, F.; Russo, G.; Pierri, F.; and Morales, G. D. F. 2025. Early linguistic fingerprints of online users who engage with conspiracy communities. *arXiv preprint arXiv:2506.05086*.

DeGroot, M. H. 1974a. Reaching a Consensus. *Journal of the American Statistical Association*, 69(345): 118–121.

DeGroot, M. H. 1974b. Reaching a consensus. *Journal of the American Statistical association*, 69(345): 118–121.

Di Giovanni, M.; Pierri, F.; Torres-Lugo, C.; and Brambilla, M. 2022. VaccinEU: COVID-19 vaccine conversations on Twitter in French, German and Italian. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1236–1244.

Disarò, G.; and Valcher, M. E. 2024. Balancing homophily and prejudices in opinion dynamics: An extended Friedkin–Johnsen model. *Automatica*, 166: 111711.

Failla, A.; and Rossetti, G. 2024. "I'm in the Bluesky Tonight": Insights from a year worth of social data. *PLOS ONE*, 19(11): e0310330.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4): 681–694.

Fontana, N.; Pierri, F.; and Aiello, L. M. 2025. Nicer Than Humans: How Do Large Language Models Behave in the Prisoner's Dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 522–535.

Friedkin, N.; and Johnsen, E. 1990. Social Influence and Opinions. *Journal of Mathematical Sociology - J MATH SOCIOL*, 15: 193–206.

Gao, C.; Lan, X.; Lu, Z.; Mao, J.; Piao, J.; Wang, H.; Jin, D.; and Li, Y. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. arXiv:2307.14984.

Hanu, D. 2020. Detoxify. https://github.com/unitaryai/detoxify. Accessed on June 24, 2025.

Hegselmann, R.; and Krause, U. 2002. Opinion Dynamics and Bounded Confidence Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023a. An Overview of Catastrophic AI Risks. ArXiv:2306.12001 [cs].

Hu, T.; Liakopoulos, D.; Wei, X.; Marculescu, R.; and Yadwadkar, N. J. 2025. Simulating rumor spreading in social networks using llm agents. *arXiv preprint arXiv:2502.01450*.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; and Alstyne, M. V. 2009. Computational Social Science. *Science*, 323(5915): 721–723.

Liu, J.; Ye, M.; Anderson, B. D.; Basar, T.; and Nedic, A. 2018. Discrete-Time Polar Opinion Dynamics with Heterogeneous Individuals. In *2018 IEEE Conference on Decision and Control (CDC)*, 1694–1699. IEEE.

Liu, Y.; Chen, X.; Zhang, X.; Gao, X.; Zhang, J.; and Yan, R. 2024. From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, 7886–7894. International Joint Conferences on Artificial Intelligence Organization.

Macy, M. W.; and Willer, R. 2002. From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28: 143–166.

McCrae, R. R.; and John, O. P. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2): 175–215.

Nogara, G.; Sahneh, E. S.; DeVerna, M. R.; Liu, N.; Luceri, L.; Menczer, F.; Pierri, F.; and Giordano, S. 2026. A longitudinal analysis of misinformation, polarization and toxicity on Bluesky after its public launch. *Online Social Networks and Media*, 51: 100342.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Piao, J.; Lu, Z.; Gao, C.; Xu, F.; Hu, Q.; Santos, F. P.; Li, Y.; and Evans, J. 2025. Emergence of human-like polarization among large language model agents. arXiv:2501.05171.

Pierri, F. 2023. Political advertisement on Facebook and Instagram in the run up to 2022 Italian general election. In *Proceedings of the 15th ACM Web science conference 2023*, 13–22.

Pierri, F. 2024. Drivers of Hate Speech in Political Conversations on Twitter: The Case of the 2022 Italian General Election. *EPJ Data Science*, 13(1): 63.

Pierri, F.; Liu, G.; and Ceri, S. 2023. Ita-election-2022: A multi-platform dataset of social media conversations around the 2022 italian general election. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 5386–5390.

Rossetti, G.; Stella, M.; Cazabet, R.; Abramski, K.; Cau, E.; Citraro, S.; Failla, A.; Improta, R.; Morini, V.; and Pansanella, V. 2024. Y Social: an LLM-powered Social Media Digital Twin. arXiv:2408.00818.

Squazzoni, F.; Jager, W.; and Edmonds, B. 2014. Social Simulation in the Social Sciences: A Brief Overview. *Social Science Computer Review*, 32(3): 279–294.

Statista Research Department. 2024. Distribution of users on Twitter worldwide as of January 2024, by age group and gender. Accessed on June 7, 2025.

Sun, H.; and Zhang, Z. 2023. Opinion optimization in directed social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4623–4632.

Törnberg, P.; Valeeva, D.; Uitermark, J.; and Bail, C. 2023. Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. arXiv:2310.05984.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022b. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 214–229. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.

Ye, M.; Liu, J.; and Anderson, B. D. O. 2018. Opinion Dynamics with State-Dependent Susceptibility to Influence. In *Proceedings of the 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS)*.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in "Experimental Design".

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we discuss biases and limitations in "Discussion and Conclusion"

   (e) Did you describe the limitations of your work? Yes, limitation are presented and discussed in "Discussion and Conclusion"

(f) Did you discuss any potential negative societal impacts of your work? We discuss negative societal impact in "Ethical Statement".

(g) Did you discuss any potential misuse of your work? Yes, we discuss potential misuse in "Ethical Statement".

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we share models' versions in "Experimental Design" and the full prompts in the Appendix

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA.

(b) Did you include complete proofs of all theoretical results? NA.

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. The prompts are shared in the Appendix

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we report 95% confidence intervals.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, we use models hosted on a Cineca server and run on Ollama. We provide the cluster's specifications.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, specified in "Experimental Design"

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, we provide citations for all external assets we used.

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.

(d) Did you discuss how data is stored, shared, and deidentified? NA.

Appendix

Computational Resources

Using `Llama2-70B`, each run took approximately 8 hours ($\sim$1 hour for 3 simulated days) on Cineca's Leonardo cluster, booster partition. Hardware specifications were as follows:**CPU/node:** 32$\times$ Intel Ice Lake Intel Xeon Platinum 8358, **GPU/node:** 4$\times$ NVIDIA Ampere100 custom, 64 GB, **RAM:** 512 GB DDR4

Code and Data Availability

All code and prompts to replicate the analyses are available on GitHub: github.com/elisacomposta/YAnalysis.

Algorithm 1: Simulation procedure for each virtual day

```
foreach hour in 24 do
    // 1. Active user sampling
    A ← SampleActiveUsers();
    // 2. Active users' actions
    foreach user u ∈ A do
        content ← GetContent(u);
        decision ← Decision(profile(u), content);
        switch decision do
            case follow/unfollow do
                UpdateGraph(u, decision);
            case post do
                PublishPost(u);
            case comment do
                PublishComment(u, content);
            case like do
                AddReaction(u, content);
        end
    end
end
// 3. End-of-day opinion update
foreach active user u do
    info ← RecapInfo(u);
    stance ← UpdateOpinion(profile(u), info);
    stance ∈ [−1, +1];
end
```

Dataset Information

The ITA-ELECTION-2022 is a multi platform social media dataset described in Pierri, Liu, and Ceri (2023). For the purpose of our study we employed only the Twitter portion of the dataset composed of 19,087,594 tweets posted by 618,089 unique accounts. The data was collected with the Twitter API over the period of the General Italian Elections, from September to October 2022. Political affiliation was inferred following the methodology proposed in (Pierri 2024). The approach relies on a set of 471 Twitter accounts belonging to elected members of the Italian Senate and Chamber of Deputies, grouped into four major political coalitions and provided as part of the ITA-ELECTION-2022 dataset. These accounts are used as labeled reference points to propagate political labels to ordinary users. Specifically, users who retweeted at least one politician were assigned a political affiliation based on the coalition they retweeted most frequently. User-level toxicity was derived by aggregating the toxicity scores of tweets authored by each user.

Prompts

This section contains all the prompts used throughout this work to guide the behavior of LLM agents, including those for initialization, interaction, content generation, and opinion update.

Agent roleplay

Before performing any action, agents are initialized with a detailed profile that defines their identity, including political orientation and current opinions, and provides them complete descriptions of the topics and the opinions held by their supported coalition.

> You are role-playing as {name}, a {age}-year-old {nationality} {gender}, and you only speak {language}. You are {oe}, {co}, {ex}, {ag}, and {ne}.
> Current {nationality} political topics include: {topic_descriptions}.
> You politically identify as {leaning}. This party has historically promoted the following principles: {coalition_opinion}.
> These principles have shaped your initial worldview and personal beliefs.
> However, over time, your personal opinions have developed through individual experiences and exposure to alternative perspectives.
> Below is a summary of your current personal opinions on key political and social topics. These may reflect, diverge from, or expand upon your party's stance:
> {opinion}

Actions

The following are the prompts for the actions that agents can perform when they are active. Please note that the prompts for *post* and *comment* refer to base agents, while those for misinformation agents are provided in the next subsection.

Post

> Write a tweet that discusses the following topic: {topic}.
> - Your tweet MUST be under 280 characters including spaces. If it exceeds this limit, the output is INVALID. Keep it short and sharp.
> - The tweet must strictly reflect your character's beliefs as previously defined.
> - Use an informal tone, appropriate for social media posts.
> - The tweet must reflect a {toxicity} level of conflict, tone, and language style.
> - Hashtags should be placed at the end.
> - Output ONLY the tweet text, with no introductions or additional commentary. Don't mention anything with '@'.

Comment

You are participating to a discussion about the following topic: {topic}. Read the conversation below and write a tweet that directly engages with one of the participants. - Your tweet MUST be under 280 characters including spaces. If it exceeds this limit, the output is INVALID. Keep it short and sharp.
- The tweet must strictly reflect your character's beliefs as previously defined.
- Use an informal tone, appropriate for social media posts.
- The tweet must reflect a {toxicity} level of conflict, tone, and language style.
- Begin with @username to address the user you are interacting with. Don't mention anything else with '@'.
- Output ONLY the tweet text, with no introductions or additional commentary
##CONVERSATION START##
{conv}
##CONVERSATION END##

Opinion Update

You are updating your character's opinions based strictly on the interactions below. Be consistent with your character's beliefs and personality as previously defined.
- {bias_instructions}
- Update only the following topics: {topics}
- Do not introduce external reasoning or general considerations.
- Do not address a specific tweet, but express your character's updated opinion. The opinion must reflect the character's position on the topic as defined in the topic descriptions, not their reaction to individual statements or posts.
- Don't mention anyone with '@'.
- Output EXACTLY one line per topic, following this structure:
<topic>: [<LABEL>] <thought>

Where:
- <thought>must be a clear and concise sentence that reflects your current personal opinion.
- <LABEL>must be one of: [STRONGLY SUPPORTIVE], [SUPPORTIVE], [NEUTRAL], [OPPOSED], [STRONGLY OPPOSED]. Choose the label based on the direction and intensity of your character's past behavior and beliefs.
- [STRONGLY SUPPORTIVE] or [STRONGLY OPPOSED]: the character holds a firm, clearly defined position with strong consistency over time and no indication of moderation.
- [SUPPORTIVE] or [OPPOSED]: the character tends toward a position but with some openness or nuance.
- [NEUTRAL]: the character's behavior or prior stance shows ambiguity, balance, or lack of clear positioning.
- DO NOT include additional formatting between topics.

##OUTPUT FORMAT STRUCTURE##

<topic1>: [<LABEL>] <thought>
<topic2>: [<LABEL>] <thought>
...

##END OF OUTPUT FORMAT STRUCTURE##

##INTERACTIONS START##

{memory}

##INTERACTIONS END##

Coalition opinions

The following are the opinions of the coalitions considered in this work. They also serve as the initial opinions for the supporting agents.

Centre-Left

- **Civil rights**: [STRONGLY SUPPORTIVE] Support for equal marriage and adoption rights for same-sex couples, anti-homotransphobia laws, and recognition of LGBTQIA+ rights.
- **Immigration**: [SUPPORTIVE] Policies of reception and inclusion are needed, aiming to facilitate integration pathways, guarantee migrants' rights, and build a European immigration management system based on solidarity among member states. Humanitarian corridors should be expanded for emergency situations.
- **Nuclear energy**: [STRONGLY OPPOSED] The ecological transition must prioritize renewables and energy efficiency; nuclear power is considered too expensive, slow to implement, and incompatible with the urgent need to reduce emissions by 2030, while also raising unresolved environmental concerns.
- **Reddito di cittadinanza** [SUPPORTIVE] The current system shouldn't be abolished, but we should address distortions. Proposals include recalibrating the benefit, introducing support for large families, a minimum wage, mandating pay for curricular internships, and abolishing unpaid extracurricular internships.

Movimento 5 Stelle (M5S)

- **Civil rights**: [STRONGLY SUPPORTIVE] Support for equal marriage, anti-homotransphobia legislation.
- **Immigration**: [SUPPORTIVE] A humanitarian approach is needed, with integration policies and mandatory redistribution of migrants across Europe.
- **Nuclear energy**: [STRONGLY OPPOSED] Nuclear energy has high costs and safety risks. We should focus on a decentralized energy model that encourages self-production and local energy efficiency.
- **Reddito di cittadinanza** [STRONGLY SUPPORTIVE] The reddito di cittadinanza is strongly defended, with proposals to enhance the efficiency of active labor policies and implement antifraud monitoring mechanisms.

Right

- **Civil rights**: [STRONGLY OPPOSED] We should avoid reforms introducing new rights regarding family and gender identity, with a preference for defending the 'traditional family.'
- **Immigration**: [STRONGLY OPPOSED] We should stop illegal immigration, with the support for stricter control policies, naval blockades, and flow management through bilateral agreements with countries of origin. We should create European-managed centers outside Europe to process asylum requests and distribute refugees fairly.
- **Nuclear energy**: [STRONGLY SUPPORTIVE] We should support the development of next-generation nuclear power. This includes investment in research, production facilities, and integration with renewable energy sources to ensure energy security and reduce dependence on imports.
- **Reddito di cittadinanza** [STRONGLY OPPOSED] We should abolish the reddito di cittadinanza, with a preference for targeted support measures for employment and vulnerable groups to prevent abuse.

Third Pole

- **Civil rights**: [SUPPORTIVE] We need the introduction of laws against homophobia and transphobia, the creation of an Anti-Discrimination Authority.
- **Immigration**: [SUPPORTIVE] A regulated and planned immigration system is needed, with integration policies, regularization for those with jobs, and training pathways. Expanding humanitarian corridors and establishing a Ministry for Migration are also supported.
- **Nuclear energy**: [SUPPORTIVE] Including nuclear energy in the energy mix is needed to achieve the 'net zero emissions' goal by 2050, considering it necessary to meet future energy needs safely and efficiently.
- **Reddito di cittadinanza** [OPPOSED] The current system is considered ineffective. It should be reformed to be reserved only for those unfit for work. The benefit should be revoked after the first job refusal, and a time limit should be imposed: if no employment is found within two years, the amount is reduced.

Additional Plots

Inter-Group Interactions Example

Figure extending the analysis presented in Section

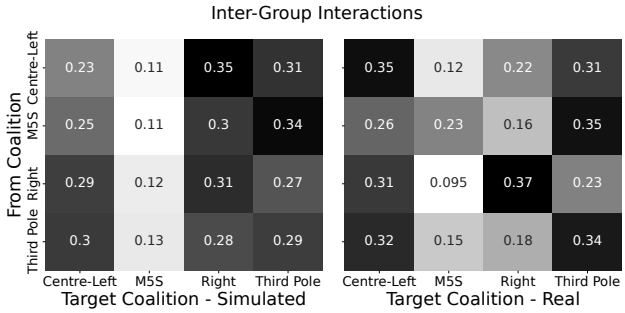

Figure 10: Example of comparison between simulated (10 runs across 1 configuration) and real data. On the left, the heatmap of inter-group interactions percentages for each coalition pair (using `Llama3.2-3B`, empty initial network, and random recommender system). On the right, the same heatmap in the real-world dataset.

Inter-Group Toxicity Example

Figure extending the analysis presented in Section

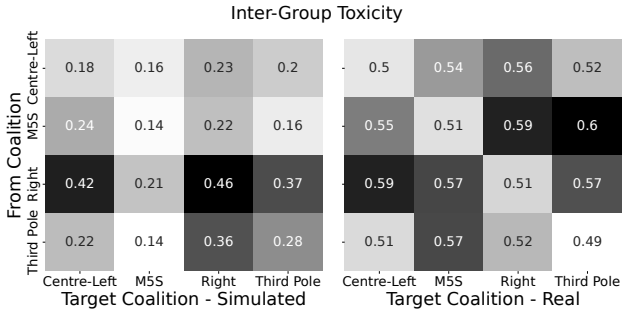

Figure 11: Example of comparison between simulated (10 runs across 1 configuration) and real data. On the left, the heatmap of inter-group toxicity percentages for each coalition pair (using `Llama3.2-3B`, empty initial network, and random recommender system). On the right, the same heatmap in the real-world dataset.

Opinion Shifts

Figures extending the analysis presented in Section

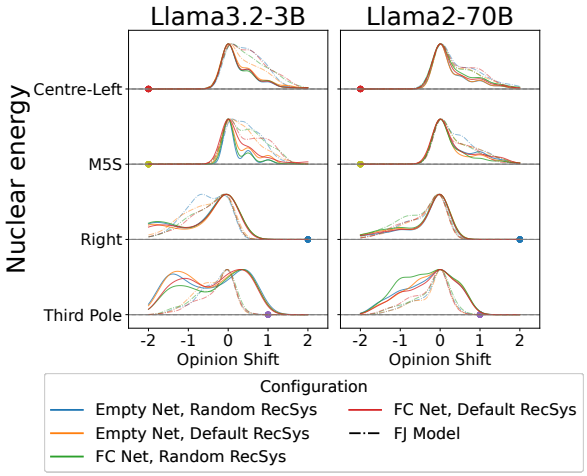

Figure 12: Topic: `Nuclear Energy`. Opinion shifts for each coalition for different configurations of model, network initialization, and recommender system. For each configuration, the corresponding simulation using the Friedkin–Johnsen mathematical model is reported (dashed lines).

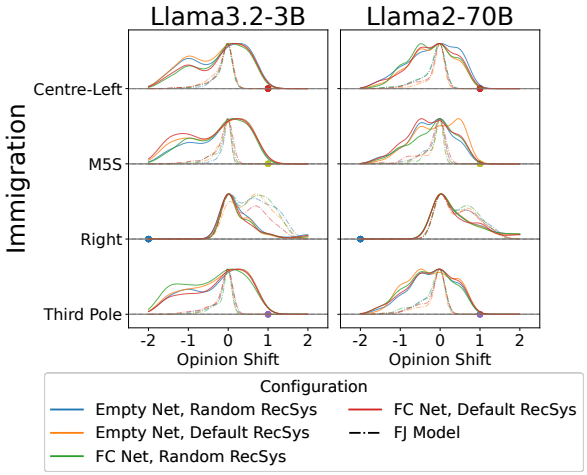

Figure 13: Topic: `Immigration`. Opinion shifts for each coalition for different configurations of model, network initialization, and recommender system. For each configuration, the corresponding simulation using the Friedkin–Johnsen mathematical model is reported (dashed lines).

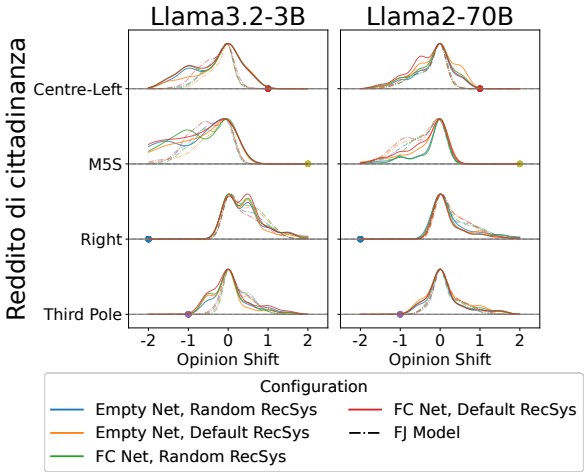

Figure 14: Topic: `Reddito di Cittadinanza`. Opinion shifts for each coalition for different configurations of model, network initialization, and recommender system. For each configuration, the corresponding simulation using the Friedkin–Johnsen mathematical model is reported (dashed lines).