POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Simulating online social media conversations with AI agents calibrated on real-world data

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Elisa Composta

Advisor: Prof. Francesco Pierri

Co-advisors: Nicolò Fontana, Francesco Corso

Academic year: 2024-2025

1. Introduction

Online social networks are a central aspect of the daily life of millions of people. They are not just communication platforms but digital spaces where users express their emotions and shape opinions. For this reason, they offer a valuable opportunity for studying social dynamics. This is the main focus of Computational Social Science, interdisciplinary field that uses computational methods to study and explain human behavior and social interactions.

To study these social processes in controlled conditions, simulators are used to recreate virtual environments. Among these, Agent-Based Modeling (ABM) simulates social systems by defining individual agents that follow simple rules, and whose interactions lead to collective behaviors. However, traditional ABMs struggle to capture the full complexity of human behavior, such as language tone and emotions.

Large Language Models (LLMs) offer an improvement by enabling agents that simulate conversations in natural language, and express opinions and emotions.

A particularly important aspect of online social media is the spread of misinformation and disinformation, which can influence opinions and social dynamics. Simulating these phenomena can help us better understand their impact and explore possible mitigation strategies.

This work explores the behavior of LLM-based agents in a simulated social network, by extending an existing simulator into three main directions: (i) integrating real-world data for initializing agents to improve the realism of their behavior; (ii) introducing an explicit opinion model that allows opinions to evolve over time; (iii) defining a new category of agents that generate misleading content to support their views.

Based on this framework, the study addresses the following research questions:

- Can LLM-based agents realistically simulate social dynamics in online platforms, including phenomena such as opinion evolution, misinformation diffusion, and network formation?
- What's the impact of misinformation on opinion shift?
- What are the current limitations of using LLM agents?

By addressing these questions, this work aims to evaluate the potential of LLMs as social agents and identify the key challenges of using them to model social dynamics.

2. Related work

Many studies have explored the simulation of social dynamics through Agent-Based Modeling (ABM), a widely used approach where complex collective behavior emerges from simple interactions among individuals. While traditional ABMs are limited by the simplicity of the behavioral rules, recent advancements in Large-Language-Models (LLMs) allow the creation of agents with more realistic and coherent behavior.
Simulations based on LLMs have been used to test different recommendation algorithms, showing for example that promoting the interaction between opposing views can reduce toxicity [6]. Y [4] is another example of simulation framework, designed as a digital twin of a social media platform, where LLM-based agents can post, reply, react and follow other users, supporting controlled experimentation of online behavior.

Rumor dissemination has long existed, even through traditional media, but the raise of online social networks has dramatically increased the speed and scale at which fake news can spread, making it an interesting phenomenon for simulation-based studies. Traditional ABMs tried to replicate it, but they lacked the complexity of human interactions. The use of LLM agents, instead, enables the simulation of more realistic dynamics, thanks to their ability to generate realistic and persuasive content, even in the context of disinformation. Some studies show that both agents' personalities and the underlying network structure influence the disinformation propagation, while other frameworks assign specific roles (e.g., *spreaders*, *verifiers*) to better analyze agent behavior [2].

Another key challenge in social simulations is opinion modeling. Traditional mathematical models, such as DeGroot or Friedkin-Johnsen, formalize social influence, but simplify the complexity of human communication and interaction.
To overcome these limitations, recent opinion dynamics studies introduced LLM agents, leveraging their ability to role-play and interact through natural language [1].
This makes them particularly suitable for modeling opinion change in settings where language and social interaction are crucial.

3. Methods

3.1. Simulation workflow

The results discussed in this work are based on Y [4], social media simulator with LLM agents. Each simulated day consists of multiple rounds during which a sample of active agents performs actions, such as posting and reacting to content. Agents start with no predefined social connections, allowing the network structure to emerge and evolve over time based on their interactions. The system is highly configurable, as it allows to specify parameters such as hourly activity, recommendation algorithms, and misinformation level. In addition to the existing simulator, at the end of each day an additional phase enables agents to update their opinions on the discussed topics.

3.2. Agents

This section explores how agents are modeled, one the most crucial aspects in social simulations.

3.2.1 Initialization and behavior

When creating the population, each agent receives a detailed profile built from a mix of randomly generated features, real-world data, and information sampled from real-world distributions.
The randomly sampled dimensions include name, surname, and personality, which follows the Big Five model, allowing up to 32 combinations of distinct personalities. Age and gender are assigned using weighted probabilities based on 2024 Twitter statistics in Italy, restricted to users aged 18-60 [5].
All agents are set with Italian nationality and share four main interests, corresponding to the political topics analyzed in this study: *Civil rights*, *Immigration*, *Nuclear energy*, *Reddito di Cittadinanza*.
A Twitter dataset collected around the 2022 Italian political elections [3] has been used to initialize some features: supported political leaning, writing toxicity, and activity level, normalized in the range $[0, 1]$ using a logarithmic transformation.
Before performing an action, each agent receives a role prompt describing its persona, its opinions, and a description of the topics, ensuring

it has the necessary background knowledge and stance definition.

When an agent is active, it performs one of the following actions: *post*, *comment* on an existing conversation, or just *read* a tweet. The action is selected based on two activity values (in $[0, 1]$) that define how likely the agent it to post or comment. If their sum is less then 1, the remaining probability is automatically assigned to the read action. This setup was designed to leverage the available data of real-world user activities.
The topic to be discussed in a post is randomly picked from the user's interests, among those active in the configured time window. Although the agents represent Italian users and topics, all generated content in the simulation is in English, reflecting the default language setting of the simulation environment. When the agent comments or reads a post, it can also decide to add a reaction (*like* or *dislike*) and to *follow* or *unfollow* the author. These additional behaviors contribute to shaping the network over time.
To decide which content the agent interacts with, a recommendation system selects the posts to show. Two algorithms are used: *ReverseChronoFollowersPopularity*, that recommends popular recent content mainly from followed users, and *ContentRecSys*, selecting random posts. For suggesting users to follow, the default algorithm, *PreferentialAttachment*, is used, which ranks users according to the product of the agent's neighbor set size and that of the candidate user.

3.2.2 Misinformation agents

This work introduces a new agent type that generates misleading content. These are designed like normal users, but are prompted to spread misinformation. They act individually, without coordination or harmful intent, and are therefore considered as misinformation agents.
Unlike other individuals, they are not initialized using directly real user data. Their political leaning is assigned ensuring a uniform distribution across coalitions. Toxicity and activity levels are generated using statistical distributions fitted on the dataset: for toxicity the best-fitting distribution is identified per coalition; for activity, posts and comments are modeled with discrete distributions and then converted into normalized scores, consistently with other agents.

3.3. Opinion modeling and update

To better represent individual behavior, this work introduces an explicit opinion model. Opinions are represented by a numerical score from -1 (strongly opposed) to +1 (strongly supported), along with a textual explanation. The initial opinion of each agent reflects the stance of their supported political coalition, and can evolve during the simulation, using either a mathematical model or an LLM-based approach.

Among the mathematical approaches, the implementation includes the classical Friedkin-Johnsen model, where agents update their opinion by combining their initial stance with those of their neighbors. The model used throughout the simulation is based on [7], and updates opinions based on the agent's current state rather than the initial one:

$$x_i(t+1) = (1-\lambda_i)x_i(t) + \lambda_i \sum_{j \in N_i(t)} w_{ij} x_j(t)$$

where $x_i(t)$ is the opinion of individual i at time t, N_i is the set of following users, λ_i is the user's susceptibility to other users, and w_{ij} is the impact user j has on i. Susceptibility is computed from personality traits, and the neighbors are weighted according to the type of interactions they had (*like*, *dislike*, *follow*).
These scores are only used for analysis, as all actions in the simulations are entirely driven by the LLM agents.

In the LLM-based opinion model, LLMs use the textual description of their opinions to act coherently with their views during the simulations.
In the update phase, the agent receives its own profile, the topics to update, its current opinions and the beliefs of its coalition, and a memory of recent interactions, including the content read/written, reactions, follow changes. The model then outputs both a textual explanation and a stance label, later mapped to a numerical score.
To simulate a slight resistance to change, the prompt also includes a confirmation bias, which is stronger for misinformation agents, to reflect their stronger attachment to their beliefs.

4. Experiments and discussion

The simulations had 100 agents interacting over 21 virtual days, initialized with profiles, personalities and opinions. The model used is *artifish/llama3.2-uncensored*, available on *Ollama*, chosen for its lack of filters, essential when dealing with political topics or controversial opinions.
Different scenarios were tested by varying two parameters: the level of misinformation (0%, 5%, 10%, and 50%), and the content recommendation system. Specifically, two recommender algorithms were used:

- *ReverseChronoFollowersPopularity*, which shows recent posts from followed users, with some content from non-followed users to ensure exposure to different views
- *ContentRecSys*, suggesting random content

Each scenario was run 10 to 20 times with new agents, to ensure the robustness of the results.

The analysis is structured on multiple levels to provide a overview of the behavior of LLMs agents within a simulated social media.

4.1. Interactions

The comparison of interaction activity by user type, in Figure 1, shows that misinformation agents are more active in content generation, as they post and comment more than base users. However, they engage less in building connections, as *follows* are significantly lower. *Likes* are used similarly by the two groups, while base users are more active in expressing disagreement through *dislikes*. Thanks to the *follow* action, users form connections across coalitions and between different agent types, allowing a structured network to emerge. The *unfollow* action, instead, is almost absent, suggesting that longer simulations may be needed to see a observe the evolution of the connections over time.

To analyze how users interact within and across coalitions, Figure 2 shows, for each coalition, the percentage of in-group interactions, categorizing the actions into positive (*like*, *follow*) and negative (*dislike*, *unfollow*).
Looking at positive interactions, Centre-Left and Third Pole show a balanced behavior, with about a half of their *likes* and *follows* directed at in-group users. In contrast, M5S and Right show fewer in-group positive interactions. For M5S,

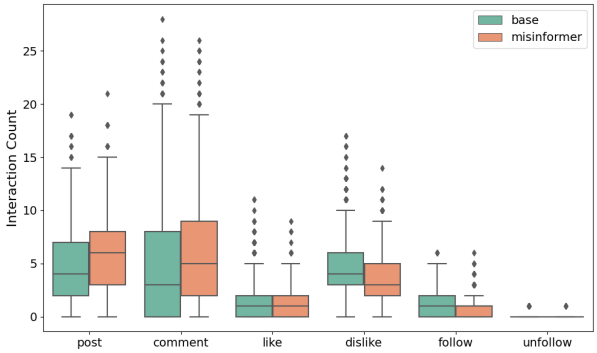

Figure 1: Number of interactions per user by agent type, with each point representing a single user from a simulation run.

this may be due to its smaller size in the simulated populations, which increases the chance of out-group interactions. The Right, even though it's one of the largest group, still shows a strong preference for out-group positive interactions.
Negative interactions are mostly directed toward other coalitions, and the Right stands out for the higher rate of in-group negative interactions. This might indicate that the Right may have a greater level of internal fragmentation, with more conflict among users, even if sharing the same political views.

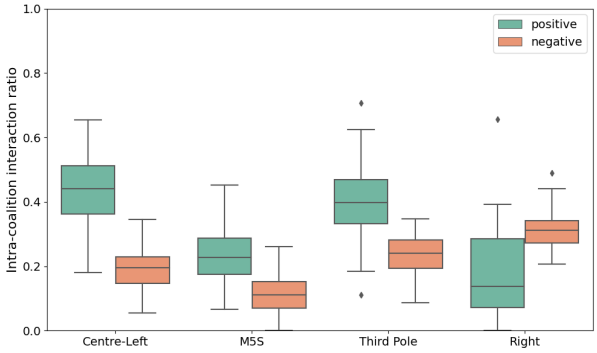

Figure 2: Percentage of in-group interactions, divided into positive and negative interactions, for each coalition, with each point representing the ratio from a single simulation run.

4.2. Opinion evolution

One important aspect introduced in this work in extension to the existing simulator is opinion modeling. Figure 3 shows that the opinion evolution of scores assigned by LLMs have the same trends as traditional opinion dynamics models. This confirms that LLMs can effectively model

opinion change at population level. Coalitions that start with the same opinion tend to evolve in parallel, suggesting that their initial view is more influential than the coalition itself. Moreover, across all topics, there's a gradual convergence toward neutral values, indicating a general reduction in polarization over time.

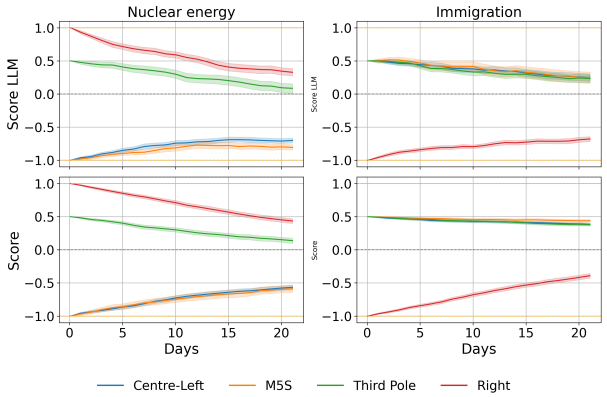

Figure 3: Opinion evolution, with scores assigned by the LLM and a traditional model, aggregated across all simulation runs of a single experimental setup.

4.3. Misinformation

To evaluate the impact of misinformation, we considered opinion shift, which is the difference between each user's final and initial opinion. Results in Figure 4 show that increasing the amount of misinformation in the system doesn't significantly affect how agents update their views. Even in the extreme scenario with 50% of agents producing misleading content, opinion shifts remains similar to those in setups with lower misinformation.

A possible question is whether the lack of misinformation impact might be related to the confirmation bias, which was explicitly introduced in this work. While the bias is visible, in the narrow distributions indicating a resistance to opinion change, it affects all conditions equally, and doesn't explain the absence of differences across misinformation levels. Agents do evolve their opinions over time, but the dynamics of change are independent from the misinformation exposure.

These findings reveal a limitation: even though LLMs can simulate realistic opinion change, they don't replicate the real-world susceptibility to misinformation. This may be due to missing factors such as emotional reasoning or social signals such as popularity or perceived credibility. Adding these factors could improve future simulations.

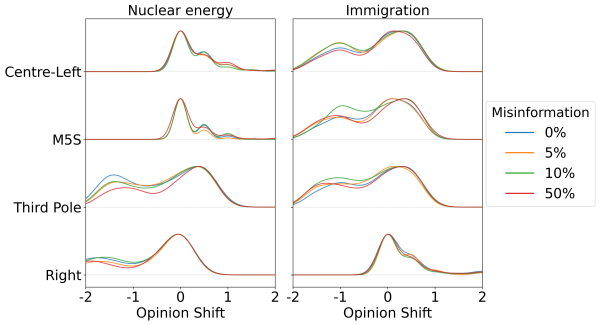

Figure 4: Distribution of individual users' opinion shifts by topic and coalition across all simulation runs for each misinformation level.

4.4. Toxicity analysis

To evaluate how agents express toxicity toward different groups, we compute the delta of the logarithms of the toxicity of comments directed to out-group and in-group users. Figure 5 shows that both real and simulated data are centered around zero, indicating that agents don't have a specific preference in toxicity direction.

However, the real data have a wider distribution, indicating that real users show greater variability in the toxicity toward the two groups. This suggests that the simulations fail to reproduce the diversity of behavior of real world in how toxicity is distributed.

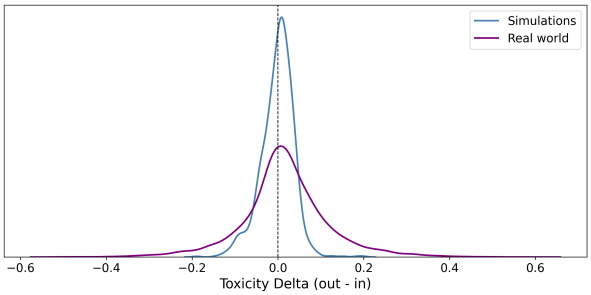

Figure 5: Distribution of the difference in mean toxicity toward out-group and in-group comments for each user in each simulation.

Figure 6 investigates how toxicity varies across political coalition and content types.

In general, posts tend to be more toxic than comments, except for the Right coalition, which

may indicate a more conflictual style in replying. M5S has the highest average toxicity in posts, whereas Centre-Left and Third Pole maintain a more moderate and stable tone in both types of contents.
Despite the low average values, all distributions have a positive skew, suggesting that LLMs can produce highly toxic content, even though at lower frequency. This behavior, possibly facilitated by the use of an uncensored model, contributes to a more realistic simulation of online conversations.

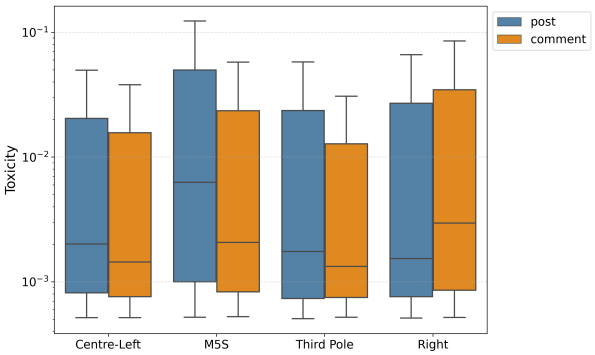

Figure 6: Toxicity of LLM-generated texts by political coalition, including all posts and comments from all simulations.

4.5. Content recommendations

The comparison on the two content recommendation algorithms doesn't reveal any significant behavioral difference. At the beginning of the simulation, agents are not yet connected, so the default algorithm, *ReverseChronoFollowersPopularity*, doesn't have follower information, and ends up behaving as the random recommender, *ContentRecSys*.
As shown in Figure 7, both the volume and the in-group ration of interactions is similar across the two algorithms. To observe more meaningful effects, simulations should either run for a longer virtual time or start from a network with pre-existing connections, allowing the recommender system to have a greater influence.

5. Conclusions

This work explored the use of LLMs as agents in social simulations, by extending the original simulator to integrate opinion modeling, misinformation agents, and a realistic initialization based on real-world data.

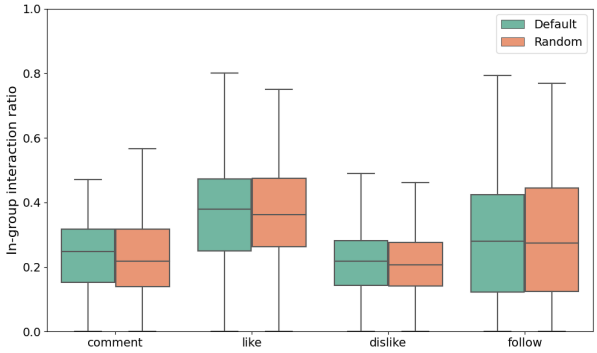

Figure 7: Percentage of in-group interactions by type for two recommendation algorithms, with each point representing a single simulation run.

Results showed that LLM-based agents are capable of interacting, generating content with different linguistic styles, and forming social connections. Opinion scores assigned by LLMs followed the trends of traditional opinion dynamics models, supporting their validity for population-level analysis.
However, some limitations emerged. The simulated time was not sufficient to observe long-term effects, and the network was not sufficiently structured for the recommendation systems to have an impact. Moreover, agents lacked the cognitive and emotional mechanisms necessary to reproduce the susceptibility to misinformation, which was negligible in the simulations.
Future research could explore a wider range of disinformation strategies (e.g., bots, coordinated groups), integrate multimodal content (text, images, videos), or introduce external events during the simulation timeline. Additionally, a comparison with real-world data would help assess the realism of the emergent behaviors and validate LLM-based agents.
Overall, this work shows that LLMs are a powerful tool for simulating social phenomena, enabling more realistic modeling of interactions in agent-based systems. However, replicating complex dynamics remains a challenge, and further studies are required to make these simulations closer to real-world dynamics.

References

[1] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulat-

ing opinion dynamics with networks of llm-based agents, 2024.

[2] Yuhan Liu, Zirui Song, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. From a tiny slip to a giant leap: An llm-based simulation for fake news evolution, 2024.

[3] Francesco Pierri, Geng Liu, and Stefano Ceri. Ita-election-2022: A multi-platform dataset of social media conversations around the 2022 italian general election, 2023.

[4] Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. Y social: an llm-powered social media digital twin, 2024.

[5] Statista Research Department. Distribution of users on twitter worldwide as of january 2024, by age group and gender, 2024. Accessed on June 7, 2025.

[6] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms, 2023.

[7] Mengbin Ye, Ji Liu, and Brian D. O. Anderson. Opinion dynamics with state-dependent susceptibility to influence. In *Proceedings of the 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS)*, 2018.