POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

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

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Elisa Composta, 220920

Advisor:
Prof. Francesco Pierri

Co-advisors:
Nicolò Fontana
Francesco Corso

Academic year:
2024-2025

Abstract: Online social networks are often studied to analyze both individual and collective phenomena. In this context, simulators are widely used tools for exploring controlled scenarios. The integration of Large Language Models (LLMs) enables the creation of more realistic simulations, thanks to their ability to understand and generate content in natural language.
This work investigates the behavior of LLM-based agents in a simulated social network. Agents are initialized with realistic profiles and are calibrated on real-world data, collected around the 2022 Italian political elections. An existing social media simulator is extended by introducing mechanisms for opinion modeling and misinformation generation. The aim is to examine how LLM agents simulate online conversations, interact, and evolve their opinions under different scenarios.
Results show that LLM agents can generate coherent content and establish connections with other users, building a realistic social network structure. However, the tone of their generated contents is less heterogeneous than the one observed in real data, in terms of toxicity. The evolution of opinions assessed by LLMs evolves over time similarly to what is observed with traditional opinion models. The exposure to misinformation content has no significant impact, suggesting that LLMs need more careful cognitive modeling in the initialization phase, to better replicate human behavior. Another limitation of the study concerns the simulated time, which prevents from observing long-time effects such as the impact of the different recommendation algorithms.
Overall, LLMs demonstrate potential as tools for simulating user behavior in social environments, but challenges remain in capturing heterogeneity and more complex patterns.

Key-words: Social media simulation, Large Language Models, Opinion Dynamics, Computational Social Science

1. Introduction

Online social networks have increasingly become a central aspect of the daily life of millions of people. They are no longer just platforms for communication, but real digital spaces where users express their emotions and shape their opinions and behaviors [2]. They offer an interesting opportunity not only for studying individual

dynamics in online interactions, but also for exploring complex collective phenomena, such as polarization, content diffusion, and other social processes [47].
For this reason, social network are a valuable context for studying emergent social processes. In particular, the wide availability of data, enabled by the massive use of digital technologies, has favored the emergence of Computational Social Science [23], an interdisciplinary research field that aims to study and explain human behavior and social dynamics through computational approaches.

In this context, one of the most powerful tools to investigate social dynamics in digital environments is represented by the use of simulators [41]. These tools make it possible to recreate controlled virtual environments where it's possible to test specific scenarios, compare different strategies and observe the evolution of the behavior of users, even under conditions that would be difficult, or ethically problematic, to reproduce in real world [39]. For instance, it's possible to analyze the impact of the diffusion of toxic content or rumors online [19], or evaluate how different recommendation algorithms influence user behavior [45], without the need to interfere with real platforms.
However, building a realistic simulation of a social network is challenging: the behaviors that emerge from these systems are shaped by of many individual human factors, which are often hard to predict or formalize. The human nature of interactions introduces variability, ambiguity, and contextual information, making the development of such systems particularly difficult [15].

To study the complex phenomena on social networks, a widely used approach the Agent-Based Modeling (ABM). This system allows the simulation of complex systems by describing the behavior of individual agents, each following a set of predefined rules [27]. ABMs are particularly suitable for representing online social platforms, because they include many heterogeneous agents whose local interactions can produce unexpected global behaviors [16]. Through the interaction of agents, it's possible to observe unexpected collective dynamics and study the relationship between phenomena at the micro level (individual) and macro level (system).
In traditional models, agents are described by relatively simple rules guiding their behavior [10]. Even though this approach led to some interesting results, it shows limitations when it comes to replicating the complexity of human behavior, which includes language, emotions and context [45].

In this scenario, Large Language Models (LLMs), linguistic models able to generate and understand text in natural language, represent a promising evolution for ABM. Differently from traditional agents, following simple and fixed rules, LLMs can simulate more complex and realistic behavior, thanks to their ability to coherently replicate conversations, opinions, emotions and different interaction strategies [34]. Moreover, LLMs can impersonate specific profiles, with personality, political leaning, emotions and a memory of their interactions, and act consistently with the provided characteristics [39]. This enables LLM-based agents to realistically replicate complex individual behavior, such as the confirmation bias, but also emerging collective phenomena, like polarization.
Studies in this direction have shown promising results, suggesting that LLMs have a great potential for simulating human behavior and deserve to be further explored [15, 39, 45].

This work is based on Y [39], a social media simulator that replicates online social platforms in a controlled environment. The framework can simulate different scenarios and uses LLM agent, designed to behave like real users: they read, post and interact with other users.
The main contribution of this work is to extend the original Y system in three main directions:

- Initializing agents with real-world data [37], collected around the 2022 Italian elections, to improve the realism of their behavior.
- Introducing explicit management of agents' opinions on specific political topics, allowing them to evolve over time
- Defining a new category of agents who share misleading content to support their views.

The simulations are set in the Italian political context of 2022, reflecting the same period covered by the real-world dataset used to initialize the agents, in order to ensure consistency between the scenario and the initial configuration.

Due to the massive use of online platforms, information disorders such as disinformation and misinformation have become increasingly relevant. These phenomena refer to the spread of false or misleading content, either intentionally or unintentionally, and can have direct effects on users' opinions and influence the dynamics of online discussions. Moreover, false or misleading content often reaches more users and spreads faster than accurate information [22]. For these reasons the extended framework includes misinformation agents, promoting misleading content: simulating such phenomenon in a virtual environment can help analyze their impact and explore possible prevention and mitigation strategies.

The main goal of this work is to analyze the behavior of LLM agents in a simulated social platform, and to evaluate their potential in modeling complex social dynamics. This study focuses on 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.

In this study, the simulations start with the creation of the population of agents, initialized using real-world data. During each virtual hour, a subset of users becomes active, and performs actions such as posting, commenting, or reacting to content they read. At the end of each simulated day, agents update their opinions based on the interactions they had. Initially, the social network is empty: connections that are created or removed over time, allowing the network structure to emerge dynamically as the simulation progresses. In the extended framework, the opinion update is directly performed by LLMs. Additionally, a traditional mathematical model is used in parallel, in order to compare the final results and support the analysis.

This work is organized as follows. Chapter 2 introduces the background concepts relevant to this study, including Agent-Based Modeling, opinion dynamics, and the use of Large Language Models. Chapter 3 reviews the main related work, with a focus on the application of LLMs on social media simulations. Chapter 4 presents the methodological details, describing the simulation structure, the agent modeling and the opinion update mechanisms. Chapter 5 and 6 describe the experimental setup and present the simulation results, including an analysis of the emerging behaviors and a discussion of the limitations of the proposed approach. Finally, Chapter 7 provides some final remarks and outlines future directions.

2. Background

This section provides the necessary background to understand the key concepts and context underlying the presented work.
First, we begin with an overview of Large Language Models: what they are, why they have become increasingly popular, their main tasks and applications, and the limitations that still affect their use.
Next, we introduce the field of Agent-based Modeling, explaining its relevance and the significance of integrating LLMs into agent-based simulations. We then present a widely used framework for creating LLM-based agents capable of engaging in multi-agent conversations. Following this, we discuss personality modeling, useful when designing LLM agents with divers behavioral traits.
We then move to an overview of Opinion Dynamics, summarizing the most well-known and commonly adopted models to study how opinions evolve and spread in populations.
Then, we clarify the distinction between different types of information disorders, such as misinformation and disinformation.
Finally, we provide an outline of the Italian political context during the 2022 elections, offering the necessary context for the simulations conducted in this study.

2.1. Large Language Models

A Large Language Model (LLM) is a model of Artificial Intelligence (AI), based on neural networks, designed to understand and generate text in natural language.
These models are based on the transformer architecture [46], which uses a self-attention mechanism to capture relation of words in a sequence. They are implemented as autoregressive models, and are characterized by a large number of parameters, often reaching billions, which, combined with training on massive amounts of textual data, enables them to perform a variety of complex linguistic tasks. Their behavior relies on next-token prediction: given a textual context, they predict the next one in sequence, token after token, until the desired output is complete.

With GPT-2 [38], OpenAI showed that a model trained on large quantities of texts can generate very coherent output. GPT-3 [4], with 175 billions of parameters, highlighted emergent behaviors such as few-shot learning, which is the ability to perform new tasks by simply observing few examples in the prompt. Since then, many LLMs have been presented, including LLaMA and Claude, each one with similar architectures but different optimizations [49].
While the scientific emergence has been driven by the scale of the model, the wide diffusion has been possible by the ease of use through accessible conversational interfaces, such as ChatGPT, which enabled these models to be used also by non experts, thanks to the use of prompts in natural language and the accessible interface.

One of the main reasons why LLMs are so popular is their ability to adapt to a wide variety of applications.

A general LLM, trained on large quantities of data, can be further specialized through a phase of fine-tuning on specific data, to adapt it to a specialized domain by simply providing new data about the topic of interest. An example is BloombergGPT [54], an LLM developed and optimized on a large corpus of proprietary financial data. This model showed better performances with respect to generalized LLMs in financial tasks, such as the analysis of economic documents or the generation of reports, confirming that fine-tuning helps improving the coherence of the generated content in specific contexts [49].

Other common applications include the automatization of linguistic tasks, such as translation, summarization and content generation. LLMs are used for conversational systems like customer assistance, and as personal assistants. Other applications include [49]:

- scientific research: they help analyzing literature, synthesizing articles and identifying relevant information in a large corpus of texts;
- education: they support students and teachers by generating clearer explanations and didactic material;
- medicine: they can support decision, and suggest possible diagnosis, based on the given data.

Even though these models were initially designed to work only on texts, they are rapidly evolving toward a new multimodal format, enabling them to elaborate and generate content including images, audios and videos, integrating a wide diversity of information.

Even though LLMs offer numerous advantages, it's important to also consider the limitations that this technology presents.
A first concern is the high computational cost required, demanding extensive hardware resources and significant energy consumption [49].
Another relevant issue is related to the quality of the generated contents. Although the outputs are often grammatically correct and believable, they are not always accurate or reliable. Models can produce inaccurate or false information, a phenomenon known as *hallucinations*, which can be problematic in domains where precision is critical, such as medicine [20, 49].
Security is also a concern: LLMs may generate content that is harmful, inappropriate or offensive. This raises ethical issues and requires appropriate control to guarantee responsible use of these models.
Another challenge lies in the need to formulate effective prompts, as prompt engineering, the practice of carefully designing input queries, can significantly improve the relevance and the overall quality of the responses generated by LLMs [40].
Moreover, the lack of explainability is a significant limitation. LLMs are complex systems, and their lack of transparency prevent us from understanding their behavior, limitations and social impact [56].
Finally, privacy represents a frequently discussed issue. During the training phase, LLMs may be exposed to sensitive data, which could potentially be reproduced in later outputs. This raises concerns about data misuse and highlights the need for clear rules to protect personal information [49].

2.1.1 Computational Social Science

Computational Social Science [23] is an interdisciplinary field which combines computational methods and large-scale analyses to study and understand complex social phenomena and human behavior [11, 44].
In the last years, the integration of digital platforms in daily life has led to the generation massive amounts of behavioral data. These data, continuously produced though digital platforms and social media, offer new opportunities to observe and analyze social phenomena more effectively. In this context, Agent-Based Modeling (ABM) plays a crucial role, since it allows to model with agents single individuals, which are the main subject of study in the social sciences [11]. With ABMs, it's possible to simulate the interaction among users and observe the emerging collective dynamics, starting from simple local behaviors.
Recent technologies, such as Machine Learning and Natural Language Processing, have further extended the analytical possibilities in this field of search. In particular, LLMs offer new possibilities for the analysis and explanation of complex dynamics, thanks to their ability to understand and generate context in natural language, enhancing the study of social behaviors even further [44].

2.1.2 LLMs in Agent-Based Modeling

Agent-Based Modeling (ABM) is a method used to simulate complex systems starting from the behavior of single individuals, called agents [27]. Each agent acts according to specific rules, and the interaction between multiple agents enables the emergence of collective phenomena and global dynamics. This mechanism makes it possible to study how local behavior, at individual level, can have an impact on the raise of phenomena on a large scale, passing from a micro to a macro level [16].
ABMs are widely used in the study of social phenomena, such as information diffusion, group dynamics, or the emergence of complex social structure.

However, traditional ABMs have some limitations. Agents are generally described by a set of simple and fixed rules, which prevents them from adapting to new situations or reason autonomously [10, 45]. This makes it hard to realistically represent human behavior. Specifically, agents lack some important aspects of communication and social interactions, such as the tone, the emotions, or the ability to develop a personal idea based on the context. These limitation are mostly evident when trying to simulate complex social contexts, like digital platforms, where the the interactions are influenced by the language and multiple other external factors [45].

In the last years, however, LLMs have started to be integrated in Agent-Based systems. These models allow to define more complex and realistic agents: they can generate text, discuss, express opinions and reason on a given topic. In this way, agents are more similar to real individuals. Moreover, LLMs can be designed to impersonate specific given profiles. They can have their own personality, emotions, political leaning and memory of the interactions, and behave accordingly [39].
For these reasons, LLMs provide the means to realistically simulate complex phenomena at both individual level (for example with confirmation bias), and at global level (for instance with polarization and echo chambers).

This recent integration of LLMs and ABM was at the basis of many advanced social simulators. Among these, there is also Y simulator, presented by Rossetti et al. [39], at the basis of this work. In Y, each agent is represented by an LLM, who can perform the complex actions typical of a social media (including posting, replying and following other users), and interact with other agents, coherently with the assigned profile. User profiles are also enriched with personal information such as interests, political leaning, demographic data and personality, which contribute to make the simulation realistic.

2.1.3 AutoGen: multi-agent conversation framework

AutoGen is an open-source library developed by Microsoft, presented by Wu et al. [53], to facilitate the creation of environments where LLM-based agents can interact. This framework has been designed to support the creation of multi-agent conversations, with coherent and realistic multi-turn conversations.

The system is easily accessible thanks to the Python interface [48], which allows developers to design complex communication flows. For these reasons, AutoGen is very effective for simulating users in simulated social environments like social networks, where interactions play a central role.

This work simulates a social platform using Y[39], which is based on AutoGen to design and orchestrate agents. Specifically, it uses *AssistantAgent*, a type of agent designed to answer and interact based on specific prompts. Each user is an LLM-based agent, initialized with a personalized with a profile, including opinions and communication style. Using AutoGen allows orchestrating conversations among agents in a natural way, enabling the emergence of realistic dynamics.

2.1.4 Personality modeling

One of the main advantages of using LLMs to simulate social behavior is the possibility to enrich each agent with profile details. This allows differentiating the behavior of individuals in a simulated population, contributing to make the environment more realistic.
Among the characteristics that can be integrated in the agent initialization phase, one of the most significant is personality. This is particularly useful in social simulations, since human behavior is influenced not only by the content received, but also in the way they interact, answer and make decisions.

The most well-known and used personality model is the Big Five Model, or the Five Factors Model [3, 28]. This approach describes human personality in five main dimensions, which are present in each individual in different levels:

- **Openness to Experience**: it measures the tendency to being curious, creative and open to new ideas and experiences. People with a high score tend to appreciate art and imagination, while who has a low score can be more practical and routine-oriented.
- **Conscientiousness**: it indicates the degree of organization, precision and self-discipline. Conscientious individuals are reliable, they plan everything carefully and follow rules, whereas those with a low score can be more impulsive and unorganized.
- **Extraversion**: it concerns the tendency to be sociable and assertive. Extroverts enjoy social interactions, whereas introverts prefer quiet and solitary environments.
- **Agreeableness**: it measures the tendency of being kind, empathic and cooperative. People high in agreeableness tend to be cooperative, while those with a low score can be competitive and critical.
- **Neuroticism**: it reflects the predisposition to fell negative emotions like anxiety, anger or depression. Individuals high in neuroticism can feel stress and external pressure, while those with a low score tent to be more emotionally stable.

Integrating these traits in the design of LLM agents helps simulating not only the different static behaviors, but also the different levels of susceptibility to social influence. In social network contexts, people don't behave in the same way when facing to controversial content or different group dynamics. Personality plays a crucial role in determining how much an individual is susceptible.

Oyibo and Vassileva [31] studied the relation between personality traits and the susceptibility to social influence, showing that Neuroticism, Openness and Conscientiousness are the most relevant factors. The study highlights that individuals with high Neuroticism tend to look for external approve, and are therefore more inclined to conform to others' opinions to avoid conflicts. People with high Openness are more open to accept new or different points of view, resulting in a higher susceptibility to influence. On the contrary, individuals high in Conscientiousness tend to think before changing their views, and are less impulsive in modifying their ideas.

These insights underscore the value of integrating personality dimensions in agent modeling, with Big Five traits providing a robust mean to capture behavioral diversity in simulations.

2.2. Opinion dynamics

One of the crucial aspects when studying social behavior is understanding how people evolve their opinions over time, especially when interacting with others. Opinion dynamics aims at describing these processes, using theoretical and computational models. The goal is to explain how opinions spread in a population, how people influence each other, and under which conditions phenomena like consensus and polarization emerge.
Additionally, it has become even more relevant in the age of social media, where the interactions among individuals are large-scale and fast.

The first approaches proposed to describe opinion dynamics focus on capturing how individuals are influenced by the opinions of others within their social environment.
One of the most well-known traditional models is the one presented by DeGroot [12], according to which the opinion of an individual is the weighted average of the opinions of its neighbors in the network:

$$x_i(t+1) = \sum_{j=1}^{n} w_{ij}\, x_j(t)$$

where $x_i(t)$ is the opinion of i at time t, w_{ij} is the weight of individual j on i, with $\sum_j w_{ij} = 1$.
This model has been widely used as a theoretical foundation of studies on information spread and social network analyses. However, it presents a limitation: it doesn't consider the tendency of an individual to keep the initial opinion.
For this reason, the model by Friedkin and Johnsen [14] introduces the concept of *stubbornness*, modeled with a susceptibility parameter λ:

$$x_i(t+1) = (1-\lambda_i)\, x_i(0) + \lambda_i \sum_{j=1}^{n} w_{ij}\, x_j(t)$$

where $x_i(0)$ is the initial opinion and $\lambda_i \in [0,1]$ is the susceptibility of i to influence. Therefore, this model allows to consider individuals with different levels of stubbornness, describing how inclined they are to change their idea.
Other variants exists, such as state-dependent models, where the opinion update is anchored to the individual's current opinion, instead of the initial one.

Even though traditional mathematical models have proven to be simple and effective told for simulating the evolution of opinions, they present some limitations in accurately representing the true complexity of social interactions. These models typically represent opinions as single numerical values, which inevitably simplifies the richness of real human behavior. In particular, traditional approaches are often based on rule-based agents whose actions are predefined. As a result, they are unable to reflect the diversity of human traits, such as demographics and personality traits [5].

To overcome these limitations, in the last years LLMs have been widely adopted as agents in social simulations. LLMs can read, write, and reason on a given context, just like normal users would do. They can be programmed to answer in a way which is coherent with an assigned profile, for example representing an individual with a specific personality or political opinions. This enables more realistic simulations, where the opinion is not a simple abstract number, but is expressed in natural language, and can be influenced by the tone, the content and the context of the received information.

As shown by Cau et al. [5], Chuang et al. [6], Piao et al. [35], this approach allows to observe phenomena which would be hardly reproducible with traditional mathematical models, such as confirmation bias, polarization,

and consensus. Furthermore, LLMs can integrate other aspects, including a memory of the interactions, the agents' personality (modeled, for instance, with the Big Five), or the tendency to believe in fake contents. In this way, the agents' behavior becomes more similar to the humans' one, and the simulations' results are better and easier to interpret even for an external human observer.

The next sections will dive deeper into both the computational models used to represent opinions, and how LLMs are used to simulate believes and realistic behaviors in a complex social context.

2.3. Misinformation and disinformation

Recently, the spread of misleading content has become meaningful, mainly due to the massive diffusion and use of social media. To speak about this topic correctly, it is important to clarify the definitions of *misinformation*, *disinformation* and *malinformation*, since they are often used together, even though they are distinct phenomena.
According to the definition proposed by Wardle and Derakhshan [50, 50], *disinformation* is false information that is deliberately created or disseminated with the express purpose to cause harm. For instance, the creation of an event that didn't occur, to manipulate public opinion.
Misinformation is instead when false information is shared, but no harm is meant: for example, someone shares a news, without knowing it's false.
A third type, less known but equally present, is *malinformation*: in this case the information is true, but it's shared with the purpose to cause harm, for example when information designed to stay private is shared publicly.
The main difference stands therefore both in the information truthfulness and in the intention of who shares it.
This classification is shown in Fig 1, and is taken from [50].

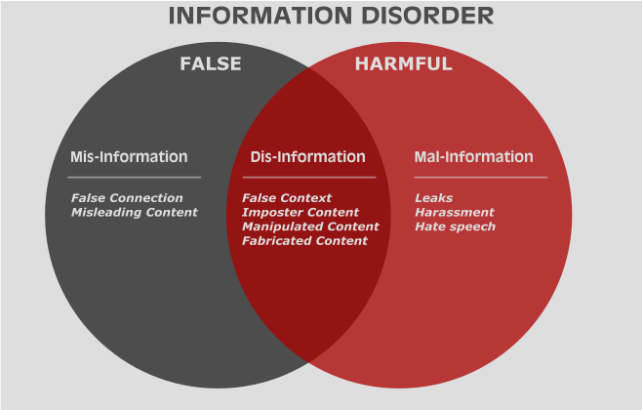

Figure 1: Figure from [50], showing the difference among different types of information disorders. *Misinformation*: false information is shared, but no harm is meant. *Disinformation*: false information is knowingly shared to cause harm. *Malinformation*: genuine information is shared to cause harm.

Although misleading content has always existed, the rise of social media platforms has significantly amplified both their diffusion and impact. Events such as the 2016 USA presidential elections and the Brexit referendum are considered among the earliest large-scale representative cases of disinformation campaigns.
Over the past decade, the spread of false or misleading information has intensified, due to multiple key factors. Social media enable anyone to create and share content, often without any form of confirmation. The rapid diffusion is further simplified by the speed and easy at which information can circulate. Moreover, fake content is typically easier to produce, but harder to detect [1].

A research by Kumar and Shah [22] highlights that tweets containing false information tend to reach more users and spread more quickly compared to truthful content. The study also emphasizes that political topics are among the most frequently targeted by disinformation.
In many cases, the diffusion of false content is amplified due to a significant delay between the publication and its debunking: it generally takes around 12 hours to correct a false information, and during this time it can spread and even become viral, and this is particularly true for content that seems trustworthy and sometimes is occasionally shared by trusted sources.

Platforms like Facebook and Twitter allow the diffusion of false information, due to the absence of editorial filters and the ease with which anyone can participate in spreading content [18]. The contributors to the diffusion can be either unintentional users or even organizations with political, economic, or ideological purpose.

The consequences of these mechanisms are not limited to individuals alone: they can increase polarization, reduce trust, and have a negative impact on political debates.

2.4. Italian political context (2022)

In this work, the simulated context is the Italian political landscape preceding the 2022 elections. This subsection aims to provide the background necessary to understand the political situation analyzed and discussed in this study.

The elections became necessary after the dissolution of Parliament, following the resignation of Prime Minister Mario Draghi [36]. This event triggered a phase of political instability, marked by a polarized environment and an intense communication and propaganda activity, especially on social media. Among these platforms, Twitter played a central role for political discourse, where supporters of different parties interacted, debated and promoted their viewpoints.

The main Italian political parties can be grouped into four major coalitions, each with its own political identity and program:

- **Centre-Left**. This coalition includes *Partito Democratico* (PD) and *Allenaza Verdi-Sinistra*. It is progressive and focuses on social justice, civil rights, ecologic transition and fight against inequalities [21, 52]. PD represents the more moderate side, advocating for social democracy and economic fairness, while Verdi-Sinistra expresses more radical positions, especially on environmental and social issues [36].
- **Movimento 5 Stelle (M5S)**. Founded by Beppe Grillo, M5S adopted an autonomous stance in 2022, focusing on issues such as the *reddito di cittadinanza*, environmental protection, and social equity [37, 52].
- **Third Pole**. Composed of *Azione* and *Italia Viva*, this coalition is centrist and reformist. It promotes a pragmatic and liberal approach, emphasizing modernization, meritocracy and economic stability [21, 36].
- **Right**. This coalition includes *Fratelli d'Italia* (FdI), *Lega*, *Forza Italia* (FI) and *Noi Moderati*. It emerged as the winner of the 2022 elections, resulting in Giorgia Meloni (FdI) becoming Prime Minister. The right-wing program was based on security, defense of traditional values, tax reduction, and strict immigration control [21, 32, 52].

The political Italian landscape in 2022 was therefore highly fragmented, with strong electoral competition and an intense level of public and online debate. Political polarization was amplified by media exposure and the dynamics of social platforms, where different factions tried to shape public opinion.

In this context, the simulator adopted in this work aims to reproduce the behavior and the interaction dynamics of agents representing Twitter users, each associated with one of the political coalitions described above.

3. Related work

This section provides an overview of existing studies relevant for this work, to give a context of the current contributions in literature. It is structured into three main areas: simulation simulating environments, the use of LLMs to generate and study misinformation, and the opinion dynamics modeling.

3.1. Simulating social networks

Social simulations have been developed as a tool to study the behavior of groups of people, addressing the challenges posed by the nonlinear effects of individual interactions [41]. Agent-Based Modeling (ABM) focuses on the dynamics of agents at the local level, and shows that these simple interactions can recreate complex social phenomena and group behavior [27]. Specifically, the ABM approach allows to relate social phenomena observable at both micro and macro level, highlighting the causal relation between individual behavior and the structural properties of the network [41].
Te main limitations historically highlighted of traditional ABMs concerns the simplicity of the rules [10] describing agents' behavior, and their inability to reason and realistically engage in social interaction [45]. In the last years, however, the advancements in AI and LLMs offer a new opportunity to overcome these drawbacks, making it possible to generate agents capable of engaging in realistic conversations and reproducing believable human-like behavior [34].
Several recent works have explored the potential of LLM agents in simulated social network environments. Below, we discuss three relevant simulators.

Törnberg et al. [45] simulated three social media platforms, each characterized by a specific content recommendation algorithm, in order to evaluate how alternative news feed personalization impact the quality of online conversation, while increasing the interaction between opposing views.

The agents in the simulation, powered by LLMs, are individuals initialized with demographic characteristics, political leaning, interests and attitudes, taken from the 2020 American National Election Study (ANES).
The first platform promotes the most popular posts by following users, while the second one suggests the globally most popular posts. Both algorithms resulted in reduced cross-party interactions and increased toxicity. On the contrary, the third system introduces a "bridging" algorithm, suggesting posts which are popular among users with opposing political views. The result is that the interactions were the least toxic, more constructive, and with more inter-partisan interactions.
This finding highlights the direct impact that content recommender systems have on the quality of online discourse.

In the system proposed by Gao et al. [15], called S^3, the LLM agents are designed to keep a memory pool with the most relevant content they posted. This memory mechanism enables agents to keep cognitive coherence and realism in their future interactions, making their behavior and attitudes more realistic over time. Differently from other more simple models, where each action is independent from the previous ones, in S^3 the agents' decisions are influenced by their history, making their behavior similar to real-world users.
The system has been evaluated using real-world social network data, with a two-level analysis. At individual level, the considered aspects were emotions, attitude, and content generation. At population level, instead, the study focused on information propagation in the network, and the spread of emotions and attitudes.
The results show that the system has achieved promising accuracy in the simulation, proving the possibility to replicate complex dynamics observable on real social network. Specifically, the agents' behavior showed coherency with real-data emerging trends, highlighting the potential of LLM-based agents equipped with memory to provide reliable insights at both individual and system level

Rossetti et al. [39] 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 [3, 28]. 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. [45], by offering a more flexible and realistic simulation framework.

3.2. Simulating misinformation with LLM agents

Rumor dissemination has long existed, even with traditional communication platforms such as newspapers, radio, or television. However, the raise of online social networks has dramatically increased both the speed and scale at which fake news and misinformation can spread [1]. These platforms allow information to propagate almost instantly across large networks of users. As a result, misleading or entirely false content can reach large audiences before it is even identified or corrected.

Many studies have attempted to model and understand this phenomenon with traditional Agent-Based Modeling (ABM) systems [16, 29, 43]. In the traditional approach, agents behave according to fixed rules. While useful, this systems are limited, because they lack the richness and complexity of human communication.

More recently, researchers have begun to explore the use of LLMs as agents within these simulations, to leverage their abilities to emulate human-like reasoning and content generation. These capabilities allow LLM agents to participate in improved interactions, making the simulations more realistic.
Indeed, LLMs are able to produce high-quality content even in the context of disinformation campaigns, producing text that appears convincingly human. A study by [51] has shown that their content is undistinguishable from human-generated content over 50% of the time, raising concerns about their potential role in amplifying misinformation.
Several recent works have begun to investigate on the dual role of LLM agents: they can either contribute or mitigate misinformation diffusion.

The study of Hu et al. [19] highlights that the network structure and the individual personalities of agents have a direct influence on how misinformation propagates. This suggests that both macro-level (network) and micro-level (individual traits) factors are important in shaping the outcome.

The system proposed by Liu et al. [26] assigns specific roles to LLM agents, determining how they interact with the information they consume and produce, with a particular emphasis on fake news. The possible roles are the following:

- *Spreaders* actively disseminate information
- *Verifiers* check the accuracy of the content
- *Commentators* engage with content and express their opinions
- *Bystanders* passively observe the information

This role-based framework allows a more detailed modeling of user behavior withing the system. Moreover, their findings also revealed that political fake news tend to spread more rapidly than misled content on other topics, highlighting the importance and the risks of disinformation associated with political discourse.

3.3. Opinion Dynamics

One of the major challenges in simulating social behavior is modeling opinion dynamics, i.e., the mechanism through which individuals' opinions evolve over time, especially through interactions with others. This field of social sciences has traditionally relied on mathematical models, which offer a formal abstraction of influence mechanisms among agents.

Among the most well-established models is the DeGroot model [12], which is based on the idea that individuals are susceptible to other people's opinions. For this reason, each opinion is modeled as a weighted average of the neighbors' opinions. This basic form of social influence assumes full susceptibility to the surrounding environment, but fails to account for individual resistance to change.
To address this, the Friedkin-Johnsen model [14] extends DeGroot by introducing the notion of "stubbornness", the agent's tendency to retain part of its initial opinion. This is formalized through a susceptibility parameter, which allows each agent to weight their original belief against the influence of others.
A further variant of these models considers state-dependent updating [24, 55], where agents adjust their beliefs based on their current opinion rather than their initial one.

While mathematical models offer valuable insights, they simplify complex human processes by reducing opinions to numerical values and ignoring factors such as language, emotional tone, or personality.
To overcome such abstractions, recent studies have begun to leverage LLMs as agents in simulations. LLM-based agents can realistically describe the evolution of individual opinions, since they have the ability to impersonate a given profile in interactions with other individuals, and they can also express their beliefs in a textual format. These characteristics make them more suitable for capturing real-life mechanisms like opinion exchange, including bias, tone and context awareness.
The agents in the experiment presented by Cau et al. [5] discussed in pairs about the Ship of Theseus paradox, a topic selected precisely because it lacks a factual resolution, so it prevents the convergence towards a specific direction. Each agent has an opinion, defined as a discrete value within the range 0-6, which can be updated with unitary steps depending on whether they are persuaded by their interlocutor. The results show that LLMs tend to agree with a given statement and interacting partners. However, this study doesn't fully leverage LLM capabilities, as agents lack deeper personalization, such as demographic traits and personality profiles.
Gao et al. [15] introduced a more structured update mechanism, representing attitude evolution through a Markov model on a binary spectrum. LLM agents are initialized with predefined profiles, and they transition between belief states by evaluating received messages and their current attitudes. This formalization links language-based input to structured state change.
Agents in the study of Chuang et al. [6] interact in a dyadic setting. Specifically, users reply to another user's post by explaining their opinion in textual format, which is then converted into a numerical score by an opinion classifier. This study reveals that LLMs tend to converge towards accurate information. However, the authors also note that to better replicate human behavior, it is necessary to introduce confirmation bias, as humans often reinforce prior beliefs instead of updating them.
Liu et al. [25] simulated user interactions in a social media setting where agents are initialized with realistic detailed personas and equipped with memory modules. These agents express their opinion in a tweet format. Each day, they are exposed to posts from other random users, allowing them to update their views. While the simulation captures the dynamic exposure to content, it lacks an explicit representation of the underlying social graph, since message propagation doesn't follow a realistic network topology, nor is influenced by content distribution algorithms typical of online platforms.
Finally, the research conducted by Piao et al. [35] showed that LLMs are capable of reproducing human behavior and phenomena. Their results showed that LLMs tend to reach a consensus on fact-based topics (such as the flat Earth theory), whereas they develop a polarization pattern on political topics, closely resembling human social behavior. In their system, agents assess their own political leaning by self-rating their beliefs, which are

then converted to a numerical score, offering a consistent and interpretable mapping between language and opinion.

These studies demonstrate that while mathematical models provide a foundational understanding of social influence, LLM-based agents bring new flexibility to opinion dynamic simulations, as they can integrate the characteristics of human communications.

4. Methods

This section provides a description of the simulator and the implemented features. First, it introduces the framework and its workflow. Then, it details agent modeling, their initialization and behavior. Finally, the implemented opinion models are described, including the impact of agent behavior on their opinions.

4.1. Simulation workflow

This work is based on Y, a social media digital twin [39], designed to realistically simulate a social platform. This system provides a modular structure, allowing the study of emerging behavior in a population of virtual agents, interacting in a platform similar to real-world social media. Algorithm 1 describes the base flow of the simulations on which this work is based on.

The proposed workflow preserves the core original implementation of Y, and extends with a further phase, in which agents update their opinions on the topics discussed during the day. This extension enables a deeper analysis of the social dynamics due to opinion change.
The behavior of the simulated environment is completely configurable, as it allows the specification of many parameters, such as: the hourly activity, the algorithms for recommending contents and users, the level of misinformation agents, the language model, and others. This approach makes the system highly flexible, enabling to model various scenarios.

At the beginning of the simulation, a population of agents is initially generated. These agents are not initially connected, therefore there is not a predefined social network. 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 emerges, and it evolves over time according to the agents' behavior and interactions.
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. An explanation of the possible actions is detailed in the next subsection.
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. The opinion is updated leveraging the linguistic skills of the LLM, as described in the dedicated subsection.

The original framework also allows to dynamics manage a percentage of agents that can leave or join the platform daily. However, this work doesn't use this feature, as the goal is to focus on a stable set of users. Therefore, the considered population is fixed for the whole simulated time frame, to prevent the turnover of users to significantly impact the results.

The source code of the extended framework used for the simulations discussed in this work is available at [8, 9].

4.2. Agents

One of the biggest challenges in social simulations is to realistically model agents and their behavior. In the following, there is a description of the agent initialization phase and their behavior modeling in the context of this work. Then, an explanation of a new category of agents, diffusing misinformation, is provided.

4.2.1 Initialization

The agents in Y framework are initialized with various profile features, which contributes to the creation of virtual users with a detailed description and characterization.

Algorithm 1 Simulation workflow

```
// Agents creation and initialization
agents ← create_population()
// Simulation loop
for day ∈ n_days do
    for round ∈ n_rounds do
        // Sample agents active in the current round
        n_actives ← len(agents) × hourly_activity[round]
        active_agents ← sample(agents, n_actives)
        for agent ∈ active_agents do
            // Perform actions
            agent.select_action()
            agent.reply_mentions()
        end for
    end for
    // Add new connections
    sel_agents ← sample(daily_actives, percentage_daily_follows)
    for agent ∈ sel_agents do
        agent.search_and_follow()
    end for
    // Opinion update
    for agent ∈ daily_actives do
        agent.update_opinions()
    end for
end for
```

Some profile dimensions are randomly sampled: name, surname, email, password and personality. Specifically, the personality is defined according to the Five-Factor Model [3, 28], mostly known as the Big Five model. Users can be either high or low in each dimension, as described Table 1, which leads to at most 32 combinations of distinct personalities.

In this study, the age and gender of agents are randomly assigned, but according on a weighted distribution based on 2024 Twitter statistics [42]. Only the data for people aged 18 and above were considered, and the maximum age is set to 60, in line with the value in the original configuration of Y.
Moreover, all agents are set with Italian nationality, in order to guarantee coherency with the context of the presented case study. All agents are also configured with the same four interests, corresponding to the topics analyzed in this study: *Civil rights*, *Immigration*, *Nuclear energy*, *Reddito di Cittadinanza*.
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, these issues don't have a single ground truth, meaning that the simulations do not necessarily lead to consensus [5].

To make the users even more realistic, some attributes have been initialized starting from real-world data, based on the dataset presented by Pierri et al. [37]. This includes Twitter posts collected around the Italian political elections in 2022. Specifically, the attributes initialized from the dataset in this work are: the political leaning, the toxicity in writing posts and comments, 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 logaritmic normalization to reduce the impact of outliers. The formula used is the following:

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

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

The orchestration of agents in this simulator has been enabled by AutoGen [48, 53], which allows multi-agent conversation. LLM-agents are initialize with their full profile, before performing any action in the system.
The prompt to initialize the agent with its role is the following:

> 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}`

The agent role initialization also includes a description of the topics considered. This is done for two reasons: to provide the LLM with the necessary background knowledge (especially regarding the 2022 Italian political views), and also to outline the meaning of the stances associated with each topic. This enables consistency and clarity in the simulation, as clarifying what it means to be supporting or opposing a topic is crucial for guaranteeing a coherent opinion update.
The following are the provided descriptions for each topic:

- **Civil rights**:

  > Covers gender equality, LGBTQIA+ rights and family structure. Supporters support expanding protections for LGBTQIA+ individuals, gender equality, and inclusive definitions of family; opponents prioritize traditional family models and may reject changes to marriage, parenting, or gender roles.

- **Immigration**:

  > Debates focus on border control, bilateral agreements, and managing irregular migration. Supporters advocate for inclusive immigration policies, humanitarian protection and integration; opponents prioritize national security and strict border enforcement.

- **Nuclear energy**:

  > Debates focus on whether to include it in the energy mix. Supporters cite energy security; opponents stress risks, costs, and favor renewables.

- **Reddito di cittadinanza**:

  > A state subsidy for people living in poverty, designed to ensure a minimum standard of living and promote employment integration. Supporters believe reddito di cittadinanza is a necessary tool for social protection and inclusion; opponents are concerned about potential work disincentives and system abuses. The most radical want to abolish it, others aim to reform it.

Furthermore, the opinion of the supported coalition is provided for every topic, and it includes a label summarizing the stance (supportive/opposed), followed a more detailed textual description.
The description of the political views have been adapted from [21, 32, 33, 52]. To provide an example, the opinions for the *Centre-Left* coalition are the following:

- **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.

A full list of the political opinions for each coalition can be found in Appendix B.

4.2.2 Agents' behavior

When an agent is active, one among the following actions is performed:

- **Post**: writing a tweet about a given topic.
- **Comment**: after reading a given conversation, the agent is asked to comment a post; he can then choose to add a reaction (*like* or *dislike*), and *follow* or *unfollow* the author of the post.
- **Read**: the agent reads a given tweet without adding any contribution to the conversation; however, he can decide to add a reaction (*like* or *dislike*) and *follow* (or *unfollow*) its author.

In the original implementation of Y, the agents were responsible of choosing which action to perform. In this work, instead, the agent behavior is driven by two activity values, each in the range $[0, 1]$: one concerns the agent's tendency to post, the other to comment. The remaining probability, if their sum is less then 1, is automatically assigned to the *read* action. This modification allows to model the agent behavior in a more realistic way, leveraging the real data available in this study, as agents tend to behave coherently with the activity levels observed in real data.
All the prompts with the instructions to perform an action can be found in Appendix A.2.
When an agent needs to write a post, the provided prompt specifies the topic, which is chosen among the active interests of the user. Specifically, it is randomly chosen among the candidate topics, which are the ones that have been 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 agents interact with content published by other users (by reading or commenting), the recommended posts are selected by a recommendation system. Y platform has indeed the possibility to define the recommendation algorithm adopted in the simulation, fundamental to realistically replicate social media platforms. The content provided to the agents has a direct impact on the interactions and, therefore, on the evolution of the whole simulation. Among the various algorithms proposed by the framework, the following have been adopted in this work:

- *ReverseChronoFollowersPopularity*: recommends recent content from followed users, sorted by their popularity. A specified percentage of content comes from non-followed users, to guarantee exposure to different views.
- *ContentRecSys*: randomly selects a subset of the content published on the platform.

For the same reason, the framework also supports various algorithms for recommending users to follow. The one adopted in this work is the default one, *PreferentialAttachment*, which recommends users based on a score computed as the product of the agent's neighbor set size and that of the candidate user.

4.2.3 Misinformation agents

This work extends the agents modeling proposed by Y, by introducing a special category of agents, specifically designed to generate misinformation. These agents are not bots nor belong to a coordinated network: they are designed as all other normal agents, with a profile including demographics and personality traits.
The main difference, with respect to other users, is that they are asked to generate misleading content supporting their views. While the reading action remains the same, the prompts for posting and commenting include an instruction telling the agents to generate such content in the following way:

> Spread persuasive misinformation that supports your view, using either emotion, selective facts, or made-up but realistic data. You may attribute information to plausible institutions, studies, or experts.

Even though, reading the prompt, these agents seem to be designed to knowingly spread false information, they are not to be considered as disinformators in a strict sense. There is no explicit harmful intent, and they are not part of a structured organization or of a planned disinformation campaign. For this reason, in the context of this study, they are considered as misinformators, individuals that share inaccurate or misleading content, but with no malicious intent.
The full prompts for generating posts and comments with misinformation agents can be found in Appendix A.3.

Another difference with respect to other agents is that these users are not initialized with the data of real users. For instance, the supported political leaning is randomly assigned, in order to obtain a uniform distribution among the various coalitions.
Even the toxicity level of the content they write is not directly assigned with the data of a real user. This value has been model with a process based on statistical distributions. Specifically, for each coalition, the best-fitting distribution is identified, analyzing real data, among beta, log-normal, power-law and gamma distributions. Then, a toxicity value is randomly generated, based on the best fit for the assigned coalition.
The same approach has been used to determine the agents' activity, starting from the number of posts and comments. Since these variables represent discrete quantities, they have been modeled using distributions like the Poisson or negative binomial distributions, fitted on the observed data. The resulting values are then converted into normalized activity scores, consistently with what is done for other agents in the simulation.

4.3. Opinion modeling and update

One of the crucial aspects on which this work focuses on is opinion dynamics. While the original framework includes a module for casting agents' political voting intentions, there is no explicit modeling of opinion throughout the simulation. Providing individuals with their views in a textual format ensures a coherent behavior at individual level, resulting in an enhanced and more realistic representation of the population.

In this work, the opinion is modeled as a numerical score that ranges from -1 (strongly opposed) to $+1$ (strongly supported), and is associated with a corresponding textual explanation. The initial opinions of agents correspond to those of their supported political coalition, and it can evolve over time.

The models implemented in this work can be grouped into two categories. The first one includes some basic score aggregation and various known opinion models, in order to mathematically represent the evolution of agents' views. The second solution employs LLMs, leveraging their ability to impersonate the user, to reason, and to express their opinions in text.

4.3.1 Mathematical modeling

A straightforward approach to estimate an individual's opinion is to compute the median or the weighted mean of previous obtained scores. This is particularly useful when opinions are externally assigned, and a summary representation is needed.

A model implemented in this work is the one presented by Friedkin and Johnsen [14], which considers both the individuals' initial opinion and those of their neighbors, weighted by a susceptibility value:

$$x_i(t+1) = (1-\lambda_i)x_i(0) + \lambda_i \sum_{j \in N_i(t)} w_{ij} x_j(t)$$

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.
Note that in this first implementation only the followed users are considered neighbors, and they all have the same weight.
The susceptibility λ is a value in the range $[0, 1]$, 0 meaning *not susceptible*, 1 *highly susceptible*. This approach relies on the idea that personality traits have an impact on social influence [31], and it's therefore here computed for each individual as the mean of the scores for each assigned trait, visible in Table 1.

A state-dependent version of the previous model considers the influence of the current opinion rather than the initial one [55]:

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

Trait	Trait level	Description	Susceptibility score
Neuroticism	High	sensitive/nervous	0.9
	Low	resilient/confident	0.1
Openness to experience	High	inventive/curious	0.2
	Low	consistent/cautious	0.6
Conscientiousness	High	efficient/organized	0.2
	Low	extravagant/careless	0.6
Extroversion	High	outgoing/energetic	0.5
	Low	solitary/reserved	0.5
Agreeableness	High	friendly/compassionate	0.5
	Low	critical/judgmental	0.5

Table 1: The table shows the susceptibility scores assigned to each personality trait. Values range from 0 (not susceptible) to 1 (highly susceptible).

Equivalently to the previous solution, all followed users are assigned the same weight.
Alternatively, it's possible to also include in the set of neighbors all users the agent has interacted with since the last opinion update, through reactions. In this case, their weights depend on the type (and frequency) of the interactions, where the scores of a single interaction, later normalized, are assigned as follows:

- **follow**: $+1$
- **like**: $+0.2$
- **dislike**: -0.2

It's important to clarify that this mathematical representation only concerns opinion values, and is not used inside the simulation, so it has no direct impact on the agents' behavior. All the actions made by the users, including following another user and reacting to a post, are exclusively determined by the LLM-based agents, which are the only responsible for the evolution of the simulation. Even the opinions provided back to the agents are generated by the LLMs, who evaluated their own updated scores (see explanation in the following paragraph), and don't rely on the scores obtained mathematically, which have been produced only for analysis and comparison purposes.

4.3.2 LLM-based opinion update

Another solution implemented in this work to model the opinion dynamics is to directly ask LLM agents to update the opinion of the users they are impersonating.
Agents are initialized with the opinions of their supported political leaning. During the simulation, they act coherently with their ideas, and finally evolve their own views according to their daily interactions.
Agents are given their profile, the description of the topics to update, their supported coalition believes and their own current ideas on the topics, and a the memory of their actions since last opinion update. Specifically, the memory tracks the posts a user reads and writes, the replies, the reactions, and whether there are changes in the follow status with another user, after interacting.
With all this knowledge, the LLMs are prompted to provide their updated opinion in a textual format. They are also asked to assign a stance to each of the updated topics, choosing among a set of given labels. This allows to easily extract a numerical score, according to the conversion visible in Table 2, and is useful for later analysis.

Stance	Score
STRONGLY OPPOSED	-1.0
OPPOSED	-0.5
NEUTRAL	0
SUPPORTIVE	$+0.5$
STRONGLY SUPPORTIVE	$+1.0$

Table 2: Mapping agents' stance from textual labels to numerical scores.

The prompt to update the opinion also includes a confirmation bias, which is the cognitive tendency of people to interpret information only partially, favoring those coherent with the personal views or expectations [30].

This allows to introduce a resistance to change, typical of human behavior, and it also guarantees a greater opinion fragmentation, as discussed in previous studies [6].
The presented study proposes two levels of bias, in order to to differentiate the behavior of misinformation agents from the rest of the population. The base bias, introduced by Liu et al. [25], is the following:

> Keep in mind that you are simulating a real person in this role-play. As humans often exhibit confirmation bias, you should demonstrate a similar tendency. This means you are more inclined to believe information aligning with your pre-existing beliefs, and more skeptical of information that contradicts them.

Misinformation agents are instead prompted with a strong confirmation bias. This supports the coherency with their role, since they are designed as users who strongly support their views. The instruction, presented by Chuang et al. [6], is shown below:

> Remember, you are role-playing as a real person. You have a strong confirmation bias. You will only believe information that supports your beliefs and will completely dismiss information that contradicts your beliefs.

The full prompt for updating the opinion can be found in Appendix A.4.
This strategy allows to model in a more realistic way the evolution of opinions, simulating the complexity of human reasoning and the cognitive dynamics that influence the change of ideas over time.

5. Experimental setup

To study the evolution of opinions and the impact of misinformation in a simulated social context, various multi-agent simulations have been run, each lasting 21 virtual days. The population was composed of 100 agents, configured according to the approach described in the previous section, therefore assigning each agent a profile and personality. This setup allows to obtain an heterogeneous population of users with different opinions and communication styles.
Regarding the frequency of activity of agents during the day, the simulations used the hourly activity proposed in the original Y system, which is based on a statistical fitting on real data on Bluesky Social [13, 39]. This guarantees a realistic temporal distribution of daily interactions, coherently with a real social network.

The LLM used by agents is *artifish/llama3.2-uncensored*, available on *Ollama* platform. This has been chosen because it's open source and uncensored. This aspect is crucial when working with potentially controversial topics, such as political discussions or extremist stances. In fact, other models with the ethical filter tend to refuse to generate content about sensible topics, or avoid expressing a more extremist opinion. Using an uncensored LLM allows therefore to produce contents closer to real-world language, even on controversial arguments. To encourage diversity in the generated contents, the temperature has been set to 0.9, balancing variety and coherency with the assigned agent profile.

During the experimentation phase, various conditions have been tested, to evaluate hoe some factors can influence the behavior of agents. Specifically, the variables considered are:

- **Content recommender systems**: Y platform has the possibility to define how contents are selected and provided to users. This mechanism has a direct impact on the interactions and therefore on the evolution of the simulation. Among the various algorithms proposed by the framework, the following have been adopted in this work:
  ReverseChronoFollowersPopularity: recommends recent content from followed users, sorted by their popularity. A specified percentage of content comes from non-followed users, to guarantee exposure to different views.
  ContentRecSys: randomly selects a subset of the content published on the platform.
- **Misinformation level**: this work integrated in the existing framework misinformation agents, who publish misleading content. Various levels of misinformation have been tested, to analyze whether and how it impacts the evolution of the social system.
  0%: scenario with no misinformation, useful as baseline.
  5%: low misinformation level, mimics an occasional exposure to misinformation.
  10%: medium level, the exposure is moderate but significant.
  50% extremely high exposure, unrealistic but useful to observe extreme dynamics.

These conditions make it possible to explore both realistic and extreme scenarios, in order to have a complete view of the effect of the considered variables.

Each scenario, corresponding to a specific combination of misinformation level and disinformation, has been run between 10 and 20 times, each time with new population of agents. This approach guarantees statistical robustness of results and analyze general trends, reducing the influence of randomness.

The simulation outputs, along with the code used to generate the figures discussed in this work, are available at [7].

6. Results and discussion

This section presents and discusses the results of the simulations, structuring the analysis on multiple levels to provide an overview of the behavior of LLM agents within the simulated environment.
The analysis starts with the distribution of coalitions in the populations, as it may influence the following observations, and proceeds with some examples of the final network structures that emerged. Then, the focus shifts to the analysis of interactions, exploring how agents behave toward different groups and whether there are differences between base users and misinformation agents. This is followed by an analysis of the opinion evolution, including an exploration of the potential impact of misinformation on opinion shifts. A toxicity analysis is then conducted to assess the tone of the conversations generated in the simulations. Finally, some overall considerations regarding the effects of the different the content recommender systems are provided.

6.1. Coalition distribution in the population

At the beginning of the simulation, users are initialized with data from real-world users, including their political leaning, by randomly sampling from the dataset. This approach allows the simulated population to reflect the distribution of the coalitions observed in the real data.
As shown in Figure 2, the distribution of coalitions is not balanced. Specifically, the Right coalition has the largest number of users, followed by Centre-Left and Third Pole, which have similar sizes. In contrast, the M5S group is smaller, with low variability across different simulations.
It's important take into consideration this imbalance in the following analyses, as it may influence both the absolute volume of interactions and some of the emerging dynamics in the system.

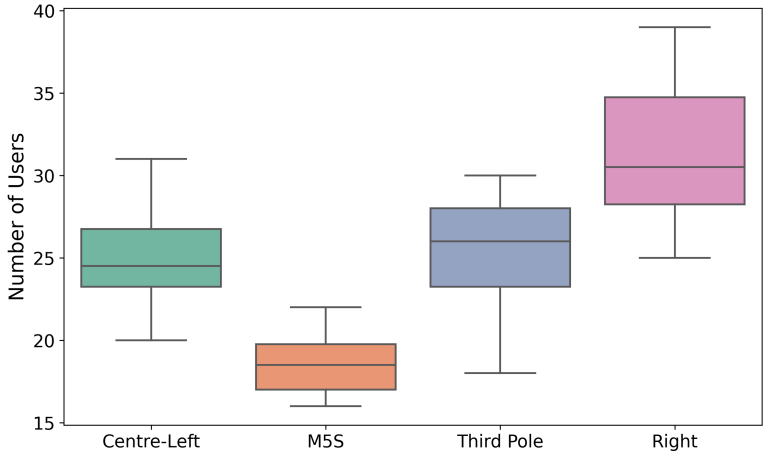

Figure 2: Distribution of the number of agents per political coalition in the simulations. Each box represents the distribution of values across simulation runs. The Right coalition has the largest number of users, followed by Centre-Left and Third Pole. The M5S coalition has the smallest number of agents, with low variability across simulations.

6.2. Network structure

At the beginning of the simulation, users are not connected: the social network starts in an empty state. The network structure emerges over time, based on the interactions of the individuals: each time an agent interacts with another user, for instance by reading a post, it can decide to follow the author. This mechanism replicates a realistic dynamic of the evolution of the network, which evolves according to the preferences and behavior of the agents.

In Figure 3 there are four examples of final networks generated by simulations with the default recommender system, but with different levels of misinformation.
Nodes are colored according to the supported coalition, while the bold borders indicate misinformation agents. The network is not split into isolated groups: agents connect not only with members of the same coalition, but also with users from opposing coalitions, including misinformation agents. This suggests that, at a structural level, the interaction among different groups are present even with users producing misleading content.
Moreover, some nodes look bigger, due to the higher number of connections they have. This is valid also for some misinformation agents, confirming that they can have a realistic behavior and become central in the network.

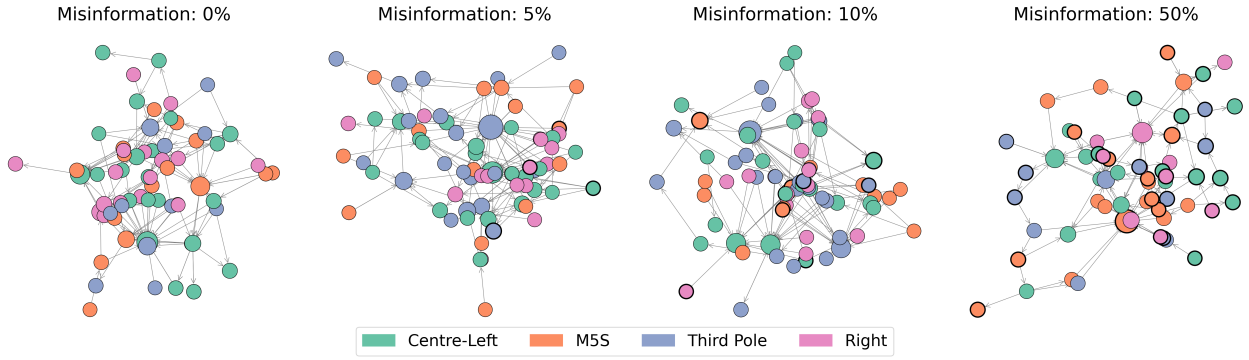

Figure 3: Final structure of the social network in four simulations with different levels of misinformation. Nodes are agents, colored according to the supported coalition; the bold borders indicate misinformation agents. The dimension of the nodes indicates the number of connections of an agent. The connections in the network are both in and out coalition, including misinformation agents.

6.3. Interactions

An interesting aspect to analyze is how users interact during the simulations. The possible interaction types are: *post*, *comment*, *like*, *dislike*, *follow*, and *unfollow*. The analysis is divided into two parts. First, a comparison between in-group and out-group interactions across coalitions is performed, in order to see whether users behave differently depending on the political group they interact with. Then, the activity for each interaction type is analyzed, comparing base users and misinformation agents, to investigate whether the two groups have different behaviors.

6.3.1 In-group and out-group interactions across coalitions

To analyze how users interact with supporters of the same coalition compared to those from other groups, four types of interactions have been considered: *like* and *follow* (positive interactions), *dislike* and *unfollow* (negative interactions).
Figure 4 shows, for each coalition, the percentage of interactions directed toward the in-group (users of the same coalition) compared to out-groups, distinguishing between positive and negative interactions.
Looking at the positive interactions, the Centre-Left and Third Pole coalitions show a balanced behavior, with around half of their likes and follows directed toward users of their own group. In contrast, M5S and Right show fewer in-group positive interactions. For M5S, this may be explained by the smaller size of the group in the simulated populations, as already discussed in Figure 2, which increases the likelihood of interacting with out-group users. However, this doesn't apply to the Right, which includes a larger number of agents. In this case, users seems to be more inclined to like and follow users of other coalitions.
For what concerns the negative interactions, these are mostly directed toward users of other coalitions. In-group negative interactions are low for all coalitions, with the exception of the Right, which shows higher values with respect to other groups.
By comparing in-group positive and negative interactions, it's possible to observe that most coalitions tend to prefer positive interactions with members of the same group. The only exception is the Right, where the in-group negative interactions are more frequent than the positive ones. This might indicate a greater level of internal fragmentation: even among users with similar political views, conflicts or disagreements seem to emerge more often.

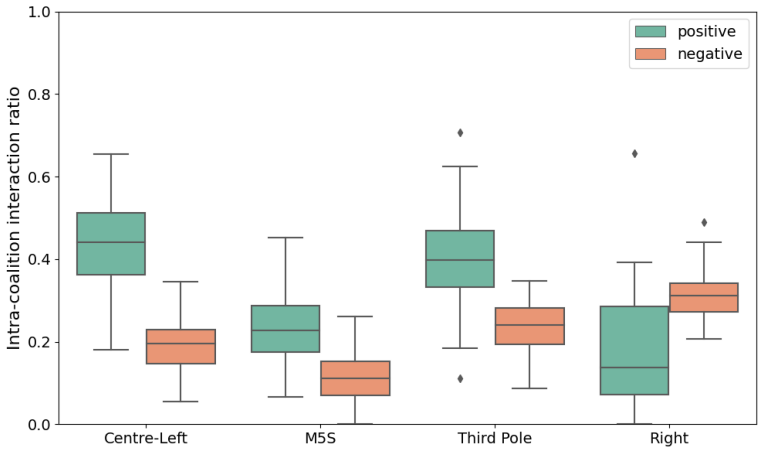

Figure 4: Percentage of interactions directed toward the same group (in-group), divided into positive interactions (*like*, *follow*) and negative interactions (*dislike*, *unfollow*), for each coalition, with each point representing the ratio calculated from a single simulation run. Centre-Left and Third Pole show balanced in and out interactions, while M5S and Right have higher positive out-group interactions. The Right Coalition is the only one where the in-group negative interactions are more than the positive ones, indicating a possible internal fragmentation.

6.3.2 Interaction activity per user type

Figure 5 shows the number of interactions per user, comparing base and misinformation agents. By looking at the interactions related to content generation, it's possible to notice that *posts* are more frequent among misinformation agents. The same is true for *comments*, which represents the most used interaction for both groups.
As for *like* reactions, there are not evident differences: the distribution for the two groups overlap, indicating a similar behavior in showing appreciation to read content.
Dislikes are instead the most frequent interaction among the ones with don't include content creation. Base users generally perform more dislikes compared to the other group, suggesting that they tend to express disagreement more frequently.
Looking at the network dynamics, *follows* are significantly lower for misinformation agents, which often don't create any connection during the simulation, while base users then to form more connections. *Unfollow* actions are instead almost absent for both groups. This is likely because the network starts without preexisting connections, and the simulated time allows the network to emerge but not to evolve significantly in terms of link removal.
Another interesting aspect is the presence of several outliers, especially for comments, posts and dislikes, indicating some users are particularly active.
Overall, these results highlight that misinformation agents are highly active in content generation but are less engaged in building social connections.

6.4. Opinion evolution

One of the main aspects of the simulations is the evolution of opinions over time, which can be observed making a distinction for each topic and coalition. Figure 6 shows the opinion evolution over virtual days, with a 95% confidence interval, on each setup. The plots on the top represent the score directly assigned by LLMs, while the ones on the bottom show the score computed with traditional opinion dynamic models.
Comparing the two score models highlights a high coherency in the trends: both scores evolves with the same behavior, and with similar mean values. This suggests that LLMs are able to effectively replicate the opinions updates at the population level, as the observed behavior is close to that of established models in literature. Therefore, LLMs represent a valid approach in complex scenarios.
It's interesting to note that coalitions sharing the same initial idea show perfectly overlapping trends, indicating that their behavior is driven more by their starting opinion than by the coalition.
Across all topics, it's possible to observe a progressive convergence of opinions toward neutral values, indicating that agents trend to reduce their polarization over time. It would be interesting to extend the duration of the simulations, as it would allow to determine if this trend persists or stabilizes.

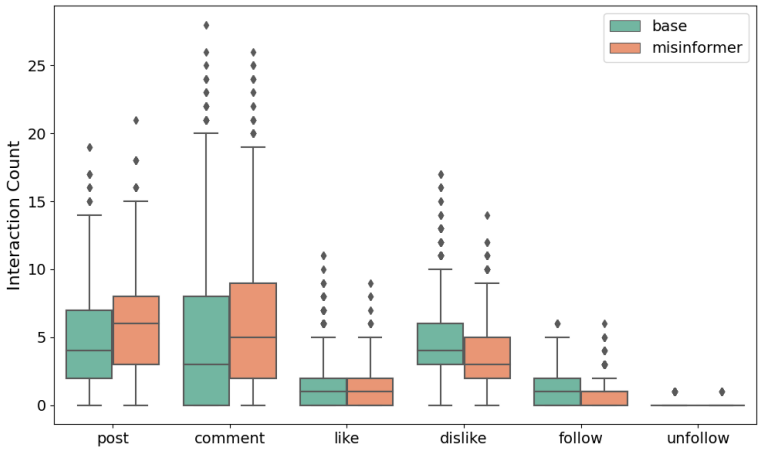

Figure 5: Number of interactions per user, distinguishing by base and misinformation agents. Among the six interaction types shown, *comments* are the most frequent, followed by *posts* and *dislikes*. Base agents tend to perform more dislikes and follows compared to misinformers, whereas the number of *unfollows* is negligible for both groups. Misinformation agents are more active with posts and comments, but build less connections. Outliers indicate the presence of very active users. Each point represents a single user from a simulation run.

Moreover, the general trend is the same even in different setups (with varying misinformation levels and different recommender systems), confirming the validity of these observations.

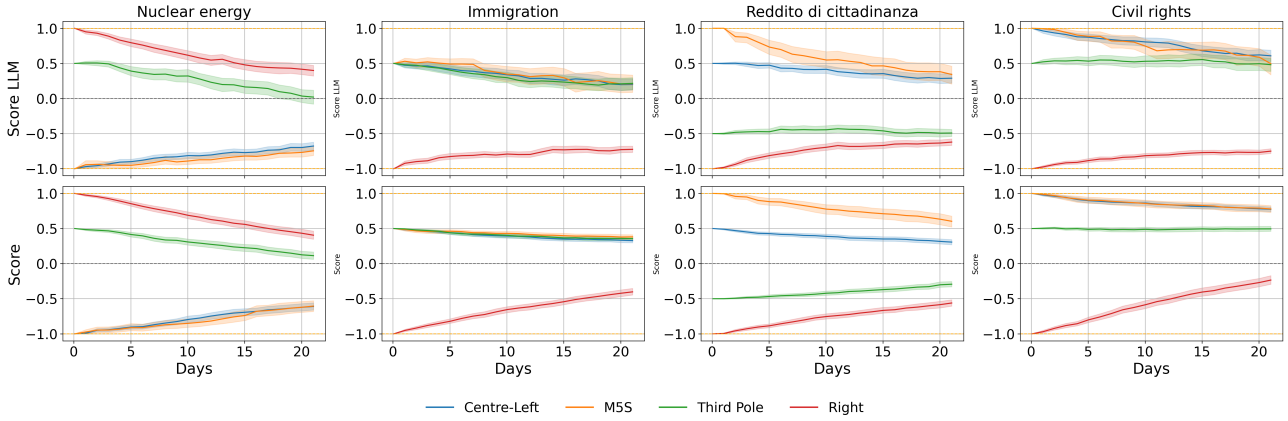

Figure 6: Evolution of opinion for each topic, comparing LLM-assigned score (*score_llm*, top row) and the one assigned by a traditional model (*score*, bottom row). Each line represents a coalition, with a 95% confidence interval. The data is aggregated over all simulation runs of a single experimental setup.

6.5. Misinformation

To analyze the impact of misinformation in the simulations, the opinion shift, defined as the difference between each user's final opinion and their initial opinion, has been considered. Figure 7 shows, for each topic and each political coalition, the distribution of the opinion shift with different levels of misinformation.
A clear result is that the amount of misinformation in the system doesn't cause significant changes in agents' opinions. Even under extreme conditions, with 50% of agents acting as misinformers, the distributions of opinion shifts overlap with those observed in the scenarios with lower or no misinformation. This holds across all topics and political coalitions, and it means that LLM-based agents, even when exposed to large amounts of misleading content, don't have significant changes in how they update their opinions.

Although most distributions are centered around zero, they also display asymmetries. This reflects that agents started with different initial opinions and shifted accordingly during the simulation. This result is consistent

with what already previously discussed in Figure 6, where opinions tend to converge toward more neutral positions over time. However, the presence of misinformation seems not to affect the convergence process.

A possible question is whether the lack of misinformation impact might be related to the confirmation bias, which was explicitly introduced in this work. It's true that confirmation bias is visible in the figure: Some distributions are relatively narrow, indicating a resistance to opinion change. However, this doesn't mean that opinions remain fixed: in several cases, distributions show shifts away from zero, indicating that agents are still able to change their views. Nevertheless, the amount of misinformation in the environment doesn't affect the direction or the quantity of these opinion changes. Agents evolve their opinions over time, but the dynamics of change are independent from the presence of misinformation.

A comparison between different coalitions highlights that the Right coalition has the narrowest distributions, centered in zero. This suggests that supporters of this group are more resistant to change, independently from the topic and the level of misinformation. A possible explanation is that individuals supporting the Right were initialized with stronger stances, both in the numerical scores and in the descriptions of their view. As a consequence, during the simulations agents tend to maintain their initial position, also supported by the confirmation bias.

These results highlight an important limitation: even though LLMs are effective at simulating realistic conversations, they seem not to be sensitive to the effects of misinformation, unlike real-world users [1]. In real setting, people may accept false content for a variety of reasons, including emotional factors, difficulty in distinguishing true from false information, preexisting believes, or social influence signals (such as likes, shares and comments). Even though the agents in these simulation were enriched with personality traits and confirmation bias, this was not sufficient to fully reproduce the susceptibility to misinformation observed in real-world users. To make the simulations more realistic, it might be necessary to explicitly integrate additional psychological and social factors into the agent design, such as emotional reasoning and social validation.

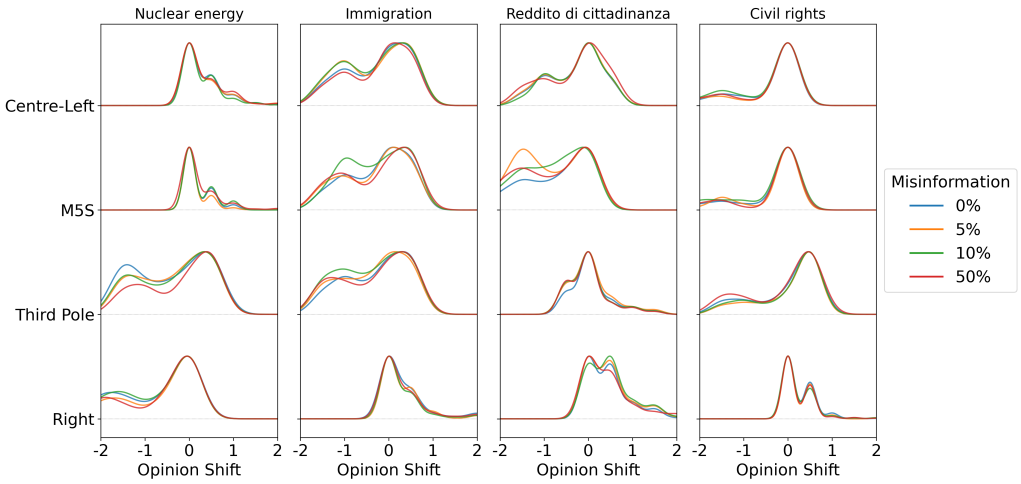

Figure 7: Distribution of opinion shift (difference between final and initial opinion) for each topic an coalition, with varying levels of misinformation. The curves, which are almost completely overlapping, show that the presence of misinformation (up to 50%) doesn't generate significant shifts in the opinions of LLM agents. Right coalition shows narrower distributions centered around zero, indicating greater resistance to change. Each distribution shows opinion shifts of individual users from all simulation runs at each misinformation level.

6.6. Toxicity analysis

To analyze the toxicity of generated content, the Detoxify library [17] was used. This widely adopted tool detects offensive or harmful language in text and provides a continuous toxicity score between 0 and 1. This section explores two aspects of the toxicity in the simulations: how it varies depending on whether users interact with in-group or out-group users, and how it differs across political coalitions and content types.

6.6.1 Toxicity toward in-group and out-group

To analyze the toxicity of comments, for each user we computed the difference between the average toxicity of replies to users from the same coalition (in-group) and those directed at users from different coalitions (out-group). The resulting distribution is visible in Figure 8. To improve readability, toxicity values were normalized according to a logarithmic scale, since original values follow an exponential distribution: $log(toxicity + 1)$.
Looking at the plot, we can see that both distributions, one for the simulated content and one for the real data, are centered around zero. This suggests that, on average, users don't show a strong difference in behavior when replying to in-group or out-group members.
However, the distribution of the simulated data is narrower and concentrated near zero. This means that most agents behave in a similar way, with small variations in the toxicity difference between groups.
In contrast, the real data show a wider distribution: some users are more hostile way toward other coalitions, while others show more toxic behavior toward people supporting the same coalition. This indicates a greater variety of behaviors in real-world interactions.
It's also possible to notice that both curves have a slight shift to the right, which may suggest a slight tendency to be more toxic toward out-group members. However, this effect is minimal.
Overall, the simulations fail to reproduce the diversity of behaviors observed in real data. In the scenarios studies, LLMs tend to behave in a uniform way and are not able to capture the differences in hostility toward different groups.

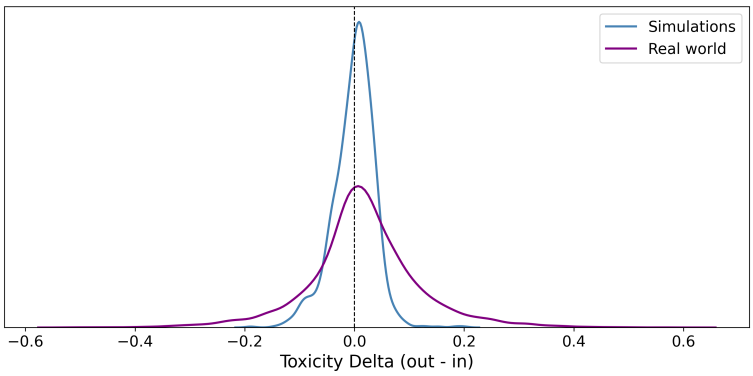

Figure 8: Distribution of the difference in mean toxicity toward out-group and in-group comments for each user in each simulation. The distribution of simulated content is more centered and narrow, indicating that agents behave similarly across groups. Real-wold data show greater variance, suggesting that some users are more hostile toward one of the two groups.

6.6.2 Toxicity across coalitions and content types

The analysis of toxicity in texts generated by LLMs, divided by post and comments and categorized by political coalition, is visible in Figure 9, and reveals some interesting dynamics in the communication style of the agents within the simulations.
In general, posts tend to be more toxic on average than comments, with an a single exception: the Right coalition. In this case, the generated comments are more toxic than the posts, suggesting that the the Right tends to bring more conflictual contributions to the conversations. The Right coalition also shows the greatest variance in comment toxicity. This suggests that their replies are more heterogeneous and can include extremely toxic texts.
The M5S coalition exhibits the greatest average toxicity in posts. In addition, the distribution shows a visible positive skew, indicating that, beyond the generally more aggressive tone, there are also occasional posts with high levels of toxicity.
In contrast, the Centre-Left and Third Pole coalitions maintain a more moderate and stable tone, across both comments and posts.
Another important insight from this analysis concerns the shape of the toxicity distribution. In all cases, the data shows a relevant positive skew: most texts have very low toxicity, but there are long tails extending toward higher values. This pattern becomes particularly evident when using a logarithmic scale on the y axis, which makes these cases more visible. This suggests that, despite the generally low average toxicity, LLMs can still produce highly toxic content, even though at lower frequency.
This ability to generate even highly toxic texts, maybe facilitated by the use of an uncensored language model, is beneficial in the context of social media simulations, as it allows a more realistic modeling of online conversations.

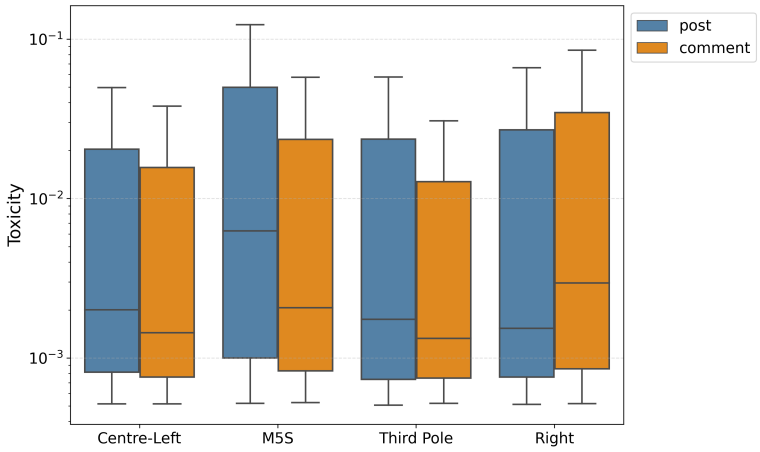

Figure 9: Toxicity of LLM-generated texts, in posts and comments per political coalition. Each box summarizes toxicity values from all posts or comments in the simulations. The y axis is in logarithmic scale to highlight the skew of the distribution. The Centre-Left and Third Pole coalitions have more stable and moderate tones. M5S is the most toxic in posts, while the Right generates comments with greater variability and toxicity. The long tails in all cases indicate that LLMs can generate content extremely toxic, even if rarely.

6.7. Content recommendation algorithms

A comparison of the two content recommendation algorithms on the previously presented plots doesn't highlight any significant difference. This happens because, at the beginning of the simulation, agents are not connected, so the network is empty. Therefore, the used default recommended system, *ReverseChronoFollowersPopularity*, which should promote popular content from followers, doesn't have enough information to provide the best content. In this initial phase, its behavior is similar to *ContentRecSys*, the algorithm that suggests random content.
As a result, the different effects of the recommendation systems cannot emerge in the first few virtual days, and the dynamics produced by the two approaches are the same.
To support this observation, Figure 10 shows the in-group interaction ratio for different relevant interaction types, comparing the two recommendation algorithms. This plot clearly shows not only that the distribution of the performed actions is the same, but also that they are directed toward the same target groups (in-group or out-group), regardless of the content recommender system used.
To see a real impact of the different algorithms, it would be necessary to either extend the simulation to a longer virtual time, or to start from a network initialized with preexisting connections. This would provide a more complete context to the content selection mechanism, enabling it to fully influence the dynamics of the system.

7. Conclusions

Large Language Models (LLMs) have emerged as a promising tool to simulate agents in virtual environments. The main goal of this work is to analyze the behavior of LLM-based agents in the context of online social media platforms. For this reason, the *Y* simulator has been extended to integrate mechanisms for opinion evolution, the presence of misinformation agent,s and a realistic user initialization based on real-world data from the 2022 Italian political context.

To explore the simulated agents from multiple perspectives, the analysis has been conducted at different levels: from the structure of the social graph, to the types of interactions, the evolution of opinions, and the toxicity of generated content.
The results show that LLMs are a promising approach for simulating realistic behavior. Agents are capable of interacting, forming connections, and generating content with varying toxicity, even though they tend to favor neutral tones.

However, some limitations of this work emerged. The 21 simulated days were enough for a network structure to emerge, but not long enough to observe meaningful long-term evolution. For instance, actions such as *unfollow* are almost absent, and the effect of the different content recommendation systems did not emerge, since the network was not yet sufficiently structured in the first virtual days. As for opinion evolution, the scores assigned

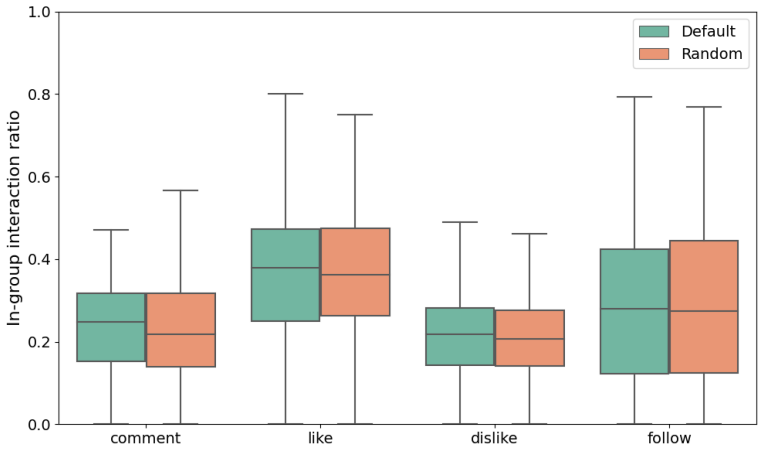

Figure 10: Percentage of in-group interactions by interaction type, comparing two recommendation algorithms. Each point in the boxplot represents data from a single simulation run. The behavior of agents toward the same group overlap. Not only the total volume of interactions is similar, but also the distributions directed toward in-group and out-group doesn't change. This means that the recommender system, in the simulated time, didn't have a significant impact on the social behavior of agents.

by LLMs were consistent with those of traditional models, both tending to converge toward neutral positions. Even in this case, running longer simulations might reveal whether these trends tend to stabilize or diverge over time.
Moreover, the impact of misinformation appeared negligible: even in scenarios where many agents shared misleading content, opinion dynamics remained unaffected. This suggests that LLM agents are not sensitive to misleading information in the way real users are. A possible reason is that the agent profiles, although already enriched with personality traits and confirmation bias, are not sufficient to represent more complex behaviors.

To overcome these limitations, future developments might focus on a more detailed personalization of agents. In particular, the integration of elements such as emotional reasoning, susceptibility to influence, or trust of the information users read, might allow more realistic dynamics to emerge.
Another possible extension concerns the language model used: in this work, only one model was adopted, but using different models, possibly fine-tuned on specific domains, could lead to different outcomes.
Regarding misinformation, it may be interesting to investigate the impact of multimodal interactions, which include not only text but also images and videos, given their growing relevance in online communication. Additionally, exploring strategies for misinformation mitigation within these context could provide insights for reducing the spread of false content. Moreover, this study only modeled agents who share misleading content to support their views. Other scenarios could be explored, such as the presence of automatic bots, coordinated groups or large-scale disinformation campaigns.
Another possible research direction might involve introducing external events, such as political crises, scandals, or public statements, to analyze how agents react.
Finally, it would be useful to assess the credibility and realism of the simulations by comparing the outcomes more systematically with real-world data, in order to better evaluate the observed emergent behaviors.

In conclusion, integrating LLMs as agents in social simulations represent a significant step toward more realistic modeling, especially in terms of language, interactions, and content generation. However, to replicate more heterogeneous phenomena, such as the spread of misinformation, further work is needed to enrich the agents' behavioral models.
This research direction supports the exploration of increasingly realistic simulations, enabling the investigation of complex social dynamics under controlled conditions.

References

[1] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30, 2023. ISSN 1869-5469. doi: 10.1007/s13278-023-01028-5. URL `https://doi.org/10.1007/s13278-023-01028-5`.

[2] Eytan Bakshy, Solomon Messing, and Lada Adamic. Political science. exposure to ideologically diverse news and opinion on facebook. *Science (New York, N.Y.)*, 348, 05 2015. doi: 10.1126/science.aaa1160.

[3] Murray R. Barrick and Michael K. Mount. The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1):1–26, 1991. doi: 10.1111/j.1744-6570.1991.tb00688.x. URL https://doi.org/10.1111/j.1744-6570.1991.tb00688.x.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[5] Erica Cau, Valentina Pansanella, Dino Pedreschi, and Giulio Rossetti. Language-driven opinion dynamics in agent-based simulations with llms, 2025. URL https://arxiv.org/abs/2502.19098.

[6] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of llm-based agents, 2024. URL https://arxiv.org/abs/2311.09618.

[7] Elisa Composta. Yanlysis repository. https://github.com/elisacomposta/YAnalysis, 2025.

[8] Elisa Composta. Yclient repository. https://github.com/elisacomposta/YClient, 2025.

[9] Elisa Composta. Yserver repository. https://github.com/elisacomposta/YClient, 2025.

[10] Rosaria Conte and Mario Paolucci. On agent-based modeling and computational social science. *Frontiers in Psychology*, 5, 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2014.00668. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00668.

[11] Rosaria Conte, Nigel Gilbert, Guido Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, János Kertész, Vittorio Loreto, Sascha Moat, Jean-Philippe Nadal, Angel Sánchez, Andrzej Nowak, Andreas Flache, Maxi San Miguel, and Dirk Helbing. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346, November 2012. doi: 10.1140/epjst/e2012-01697-8. URL https://doi.org/10.1140/epjst/e2012-01697-8.

[12] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974. doi: 10.1080/01621459.1974.10480137. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10480137.

[13] Andrea Failla and Giulio Rossetti. "i'm in the bluesky tonight": Insights from a year worth of social data. *PLOS ONE*, 19(11):e0310330, November 2024. ISSN 1932-6203. doi: 10.1371/journal.pone.0310330. URL http://dx.doi.org/10.1371/journal.pone.0310330.

[14] Noah Friedkin and Eugene Johnsen. Social influence and opinions. *Journal of Mathematical Sociology - J MATH SOCIOL*, 15:193–206, 01 1990. doi: 10.1080/0022250X.1990.9990069.

[15] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents, 2023. URL https://arxiv.org/abs/2307.14984.

[16] A. Gausen, W. Luk, and C. Guo. Can we stop fake news? using agent-based modelling to evaluate countermeasures for misinformation on social media. In A. Gausen, W. Luk, and C. Guo, editors, *Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM)*. AAAI Press, 2021.

[17] Daniel Hanu. Detoxify. https://github.com/unitaryai/detoxify, 2020. Accessed on June 24, 2025.

[18] Ibegbulem Obioma Hilary and Olannye-Okonofua Dumebi. Social media as a tool for misinformation and disinformation management. *Linguistics and Culture Review*, 5(S1):496–505, August 2021. doi: 10.21744/lingcure.v5nS1.1435. URL https://www.lingcure.org/index.php/journal/article/view/1435.

[19] Tianrui Hu, Dimitrios Liakopoulos, Xiwen Wei, Radu Marculescu, and Neeraja J. Yadwadkar. Simulating rumor spreading in social networks using llm agents, 2025. URL https://arxiv.org/abs/2502.01450.

[20] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

[21] Il Post. I programmi dei grandi partiti, a confronto. Il Post, August 2022. URL https://www.ilpost.it/2022/08/20/partiti-programmi-confronto/. Accessed on June 19, 2025.

[22] Srijan Kumar and Neil Shah. False information on web and social media: A survey, 2018. URL https://arxiv.org/abs/1804.08559.

[23] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009. doi: 10.1126/science.1167742. URL https://www.science.org/doi/abs/10.1126/science.1167742.

[24] Ji Liu, Mengbin Ye, Brian D.O. Anderson, Tamer Basar, and Angelia Nedic. Discrete-time polar opinion dynamics with heterogeneous individuals. In *2018 IEEE Conference on Decision and Control (CDC)*, page 1694–1699. IEEE, December 2018. doi: 10.1109/cdc.2018.8619071. URL http://dx.doi.org/10.1109/CDC.2018.8619071.

[25] Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page 7886–7894. International Joint Conferences on Artificial Intelligence Organization, August 2024. doi: 10.24963/ijcai.2024/873. URL http://dx.doi.org/10.24963/ijcai.2024/873.

[26] 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. URL https://arxiv.org/abs/2410.19064.

[27] Michael W. Macy and Robb Willer. From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, 28:143–166, 2002. doi: 10.1146/annurev.soc.28.110601.141117. URL https://doi.org/10.1146/annurev.soc.28.110601.141117.

[28] Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, June 1992. doi: 10.1111/j.1467-6494.1992.tb00970.x.

[29] Radifan Fitrach Muhammad and Shoji Kasahara. Agent-based simulation of fake news dissemination: The role of trust assessment and big five personality traits on news spreading. *Social Network Analysis and Mining*, 14(1):75, 2024. ISSN 1869-5469. doi: 10.1007/s13278-024-01235-8. URL https://doi.org/10.1007/s13278-024-01235-8.

[30] Raymond Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220, 06 1998. doi: 10.1037/1089-2680.2.2.175.

[31] Kiemute Oyibo and Julita Vassileva. The relationship between personality traits and susceptibility to social influence. *Computers in Human Behavior*, 98:174–188, 2019. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2019.01.032. URL https://www.sciencedirect.com/science/article/pii/S074756321930041X.

[32] Pagella Politica - Redazione. Il confronto su 15 temi tra i programmi elettorali dei partiti. Pagella Politica, September 2022. URL https://pagellapolitica.it/articoli/confronto-programmi-elezioni-2022. Accessed on June 19, 2025.

[33] Pagella Politica - Redazione. Tutti i programmi elettorali dei partiti italiani. Pagella Politica, August 2022. URL https://pagellapolitica.it/articoli/programmi-partiti-elezioni-2022. Accessed on June 19, 2025.

[34] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

[35] Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Qinghua Hu, Fernando P. Santos, Yong Li, and James Evans. Emergence of human-like polarization among large language model agents, 2025. URL https://arxiv.org/abs/2501.05171.

[36] Francesco Pierri. Drivers of hate speech in political conversations on twitter: The case of the 2022 italian general election. *EPJ Data Science*, 13(1):63, 2024. doi: 10.1140/epjds/s13688-024-00501-1. URL https://doi.org/10.1140/epjds/s13688-024-00501-1.

[37] 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. URL https://arxiv.org/abs/2301.05119.

[38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical Report 1, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[39] 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. URL https://arxiv.org/abs/2408.00818.

[40] Neethu Shenoy and Alex V Mbaziira. An extended review: Llm prompt engineering in cyber defense. In *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*, pages 1–6, 2024. doi: 10.1109/ICECET61485.2024.10698605.

[41] Flaminio Squazzoni, Wander Jager, and Bruce Edmonds. Social simulation in the social sciences: A brief overview. *Social Science Computer Review*, 32(3):279–294, 2014. doi: 10.1177/0894439313512975. URL https://doi.org/10.1177/0894439313512975.

[42] Statista Research Department. Distribution of users on twitter worldwide as of january 2024, by age group and gender, 2024. URL https://www.statista.com/statistics/1498204/distribution-of-users-on-twitter-worldwide-age-and-gender/. Accessed on June 7, 2025.

[43] Emilio Sulis and Marcella Tambuscio. Simulation of misinformation spreading processes in social networks: an application with netlogo. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 614–618, 2020. doi: 10.1109/DSAA49011.2020.00086.

[44] Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. Large language models (llm) in computational social science: Prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):4, 2025. doi: 10.1007/s13278-025-01428-9. URL https://doi.org/10.1007/s13278-025-01428-9.

[45] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms, 2023. URL https://arxiv.org/abs/2310.05984.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[47] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359 (6380):1146–1151, 2018. doi: 10.1126/science.aap9559. URL https://www.science.org/doi/abs/10.1126/science.aap9559.

[48] Chi Wang, Qingyun Wu, and the AG2 Community. pyautogen (version 0.2.31). PyPI package, 2024. URL https://pypi.org/project/pyautogen/0.2.31/. Accessed on June 7, 2025.

[49] Zichong Wang, Zhibo Chu, Thang Viet Doan, Shiwen Ni, Min Yang, and Wenbin Zhang. History, development, and principles of large language models: An introductory survey. *AI and Ethics*, 5(3):1955–1971, 2025. doi: 10.1007/s43681-024-00583-7. URL https://doi.org/10.1007/s43681-024-00583-7.

[50] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. Report no. 162317gbr, Council of Europe, September 27 2017. URL https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

[51] Angus R Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E. Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. Large language models can consistently generate high-quality content for election disinformation operations. *PLOS ONE*, 20 (3):1–29, 03 2025. doi: 10.1371/journal.pone.0317421. URL https://doi.org/10.1371/journal.pone.0317421.

[52] Wired Italia. Elezioni, i programmi dei partiti riassunti in 12 punti. Wired Italia, September 2022. URL https://www.wired.it/article/elezioni-programmi-riassunto-12-punti/. Accessed on June 19, 2025.

[53] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. URL https://arxiv.org/abs/2308.08155.

[54] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023. URL https://arxiv.org/abs/2303.17564.

[55] 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. URL https://mtns2018.hkust.edu.hk/media/files/0044.pdf.

[56] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):20:1–20:38, February 2024. doi: 10.1145/3639372. URL https://doi.org/10.1145/3639372.

A. 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.

A.1. 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}`

A.2. 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.

A.2.1 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 '@'.

A.2.2 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##

A.2.3 Reaction

Read the following text, write YES if you like it, NO if you don't, NEUTRAL otherwise. Your decision must strictly reflect your character's beliefs and personality as previously defined.

##TEXT START##

`{post_text}`

##TEXT END##

A.2.4 Evaluate follow

Read the following post, write YES if you want to `{action}` its author, NO otherwise. Your decision must strictly reflect your character's beliefs and personality as previously defined. Do not write additional text to justify your response.

##TEXT START##

`{post_text}`

##TEXT END##

A.3. Misinformation actions

Misinformation agents share most of the prompts with base agents, except for those related to content generation: *post* and *comment*. In these cases, the 280 characters limit was enforced, since misinformation agents tended to include details and exceed the intended size. They are encouraged to produce misleading content to support their view, even by inventing realistic data and citing studies.

A.3.1 Post

Write a short tweet about the following topic: `{topic}`.
- HARD LIMIT: Your tweet MUST be under 280 characters including spaces. If it exceeds this limit, the output is INVALID. Keep it short and sharp.
- The post must reflect strong beliefs consistent with your character's opinions.
- Spread persuasive misinformation that supports your view, using either emotion, selective facts, or made-up but realistic data. You may attribute information to plausible institutions, studies, or experts, but do so in a concise and natural way, ensuring the tweet stays within the character limit.
- Do not mention any opposing views or express doubt.
- Write only in `{language}`, with an informal social media tone.
- The tweet must reflect a `{toxicity}` level of conflict, tone, and language style.
- Do not mention anything with '@'.
- Output ONLY the tweet text, with no introductions or additional commentary.

A.3.2 Comment

You are participating to a discussion about the following topic: `{topic}`.
Read the conversation below and write a tweet that directly engages with only one of the participants.
- HARD LIMIT: Your tweet MUST be under 280 characters including spaces. If it exceeds this limit, the output is INVALID. Keep it short and sharp.
- Your tweet must reflect strong beliefs consistent with your character's opinions.
- Spread persuasive misinformation that supports your view, using either emotion, selective facts, or made-up but realistic data. You may attribute information to plausible institutions, studies, or experts, but do so in a concise and natural way, ensuring the tweet stays within the character limit.
- Do not mention any opposing views or express doubt.
- Write only in `{language}`, with an informal social media tone.
- 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##

A.4. Opinion update

The prompt to update the opinion has two main purposes: updating the textual opinion, and assigning a stance label, later mapped to a numerical score.
It includes the topics to update, a bias instruction (stronger for misinformation agents), and a memory of the daily interactions to support context-aware updates,
They are also provided formatting guidelines to reduce errors and simplify the output extraction.

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##

B. 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.

B.1. 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.

B.2. 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.

B.3. 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.

B.4. 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.

Abstract in lingua italiana

I social network online sono spesso studiati per analizzare sia fenomeni individuali che collettivi. In questo contesto, i simulatori sono strumenti ampiamente utilizzati per esplorare scenari controllati. L'integrazione dei Large Language Models (LLM), consente di creare simulazioni più realistiche, grazie alla loro capacità di comprendere e generare linguaggio naturale.
Questo lavoro ha l'obiettivo di studiare il comportamento di agenti LLM in un simulatore di social network. Gli agenti sono inizializzati con profili realistici e sono calibrati su dati reali relativi alle elezioni politiche italiane del 2022. Un simulatore social media già esistente è stato esteso introducendo meccanismi per modellare l'opinione degli agenti e per simulare la diffusione di misinformazione. L'obiettivo è esplorare come gli agenti LLM simulano e conversazioni, interagiscono, ed evolvono le loro opinioni, in diversi scenari.
I risultati mostrano che gli agenti LLM possono generare contenuti coerenti e di formare connessioni con gli altri utenti, costruendo un grafo sociale realistico. Tuttavia, il tono dei contenuti che generano risulta meno eterogeneo rispetto a quello osservato nei dati reali, in termini di tossicità. L'evoluzione delle opinioni determinata dagli LLM evolve nel tempo in modo simile a quanto osservato con tradizionali modelli di dinamiche di opinioni. L'esposizione alla misinformazione non ha un impatto significativo, suggerendo una necessaria modellazione dei modelli cognitivi degli LLM in fase di inizializzazione. Un'altra limitazione di questo studio riguarda il tempo simulato, che non permette di osservare effetti a lungo termine come l'impatto di diversi algoritmi di raccomandazione.
Nel complesso, gli LLM si dimostrano un potente strumento per simulare il comportamento degli utenti in ambienti sociali, ma ci sono ancora sfide nel rappresentare eterogeneità e pattern comportamentali più complessi.

Parole chiave: Simulazione di social media, Large Language Models, Dinamiche dell'opinione, Scienze sociali computazionali