Anticipating Narrative Impact through Collaborative LLM-based Multi-Agent System

Maria Di Gisi*, Giuseppe Fenza†, Domenico Furno†, Mariacristina Gallo†, Pio Pasquale Trotta*
*IMT School for Advanced Studies Lucca, Lucca, Italy
Emails: maria.digisi@imtlucca.it, piopasquale.trotta@imtlucca.it
†University of Salerno, Fisciano (SA), Italy
Emails: gfenza@unisa.it, dfurno@unisa.it, mgallo@unisa.it

***Abstract*—The online information ecosystem is shaped by collaborative and adaptive processes in which narratives evolve through collective interpretation and reframing. Anticipating these dynamics is essential for developing systems that can provide proactive, interpretable support to human decision-makers before problematic narratives reach critical diffusion. However, most existing approaches rely on post-hoc analyses, offering limited utility for early intervention and pre-bunking strategies. This paper presents a collaborative multi-agent simulation framework based on Large Language Models (LLMs) to study the evolution of news narratives. Heterogeneous agents interact within a controlled network to iteratively generate and reframe news headlines, producing structured trajectories of narrative transformation. A notion of simulated *Narrative Impact* to quantify a news item's capacity to organize, sustain, and attract collective narrative activity over time, is formulated. Building on this formulation, an early alerting task predicts the structural impact of a narrative from partial observations of the simulation, enabling the identification of high-impact narratives at early stages. Experimental results, obtained using an XGBoost-based predictive model, demonstrate that the *Narrative Impact* can be accurately identified ($R^2 = 0.71$) at early stages, even in the absence of intentional misinformation.**

***Index Terms*—Collaborative LLM Agents, Disinformation, Reframing, Pre-bunking, Narrative Impact**

I. Introduction

Online information ecosystems are increasingly shaped by collaborative and socio-technical processes in which narratives are not passively consumed but actively transformed through collective interpretation, reframing, and contextual adaptation. News items evolve as they circulate across heterogeneous actors and platforms, giving rise to emergent narrative structures that may differ substantially from their original formulation. Importantly, such transformations do not necessarily stem from intentional misinformation, they can emerge endogenously from interaction dynamics, linguistic ambiguity, and the structural properties of the content itself [1].

For AI systems operating in these environments, the ability to anticipate narrative dynamics rather than merely react to them [2] has become a central requirement for trustworthiness and effective human-centered support. Much of the existing literature on misinformation, framing, and narrative analysis focuses on post-hoc detection, aiming to identify problematic content after it has already diffused widely [3]. While valuable for retrospective analysis, these approaches face intrinsic limitations. They provide little insight into how narratives may evolve before their impact materializes, and they struggle to scale to the volume and velocity of online information streams. As a result, they offer limited support for proactive decision-making and early intervention.

In response to these challenges, pre-bunking and anticipatory strategies have emerged as promising alternatives. Rather than evaluating the factual correctness of individual claims, these approaches aim to identify news items that are likely to generate impactful, unstable, or self-reinforcing narrative trajectories. Achieving this requires models that can explore plausible futures of narrative evolution under controlled conditions and do so in a way that is interpretable and aligned with human decision-making processes.

Agent-based simulation provides a natural methodological foundation for this goal. Recent advances in Large Language Models (LLMs) enable the construction of artificial agents capable of generating, interpreting, and reframing textual content in a human-like manner [4]. When embedded in interaction networks, LLM-based agents can be used to simulate collaborative narrative dynamics, allowing researchers to observe how news items may evolve through endogenous interactions alone. Such simulations offer a hybrid AI setting in which generative models, interaction structures, and predictive components jointly contribute to the analysis of complex socio-technical phenomena. However, simulation alone is not sufficient. To support anticipatory and collaborative AI systems, it is necessary to define quantitative and interpretable measures that capture how narratives develop and acquire relevance over time. In particular, beyond semantic variation, there is a need to model the extent to which a news item structures, sustains, and attracts collective narrative activity. These properties are crucial for comparing different items and for prioritizing those that may warrant early attention.

To address these issues, this paper introduces a multi-agent simulation framework based on LLMs in which heterogeneous agents, categorized into skeptical and ideologically polarized profiles, collaboratively generate and reframe real news headlines within an artificial network. A notion of simulated *Narrative Impact* is defined, which captures the structural effect of a news item in terms of its ability to organize and

*Corresponding author

sustain narrative activity over time.

Experimental results, involving a corpus of news headlines, demonstrate that the model can effectively anticipate the evolution of narrative trajectories with high accuracy, suggesting that early signals in collaborative reframing are strong predictors of a news item's structural influence.

The remainder of this manuscript is organized as follows: Section II reviews related work, Section III presents the framework, Section IV outlines the experimental setup and prediction model, and Section V discusses implications, limitations, and future directions.

II. Related Works

The study of information diffusion on online social networks has a long tradition, ranging from classical agent-based models to recent LLM-enhanced simulations. Early work has focused on modeling information propagation using simulated networks to capture patterns observed in real-world platforms [5]. For instance, studies using NetLogo demonstrated that paired optimal propagation parameters can reproduce Twitter diffusion dynamics, highlighting the role of network structure in shaping information flow [6]. Similarly, large-scale agent-based models on Facebook using AnyLogic provided a micro-level understanding of diffusion mechanisms and proposed correction procedures to handle incomplete network data [7]. An emergent and growing body of literature has also introduced large language models (LLMs) as dynamic agents capable of generating human-like content [8], reasoning over received messages, and autonomously deciding whether to propagate information. Frameworks such as LAIDSim leverage LLMs to model user decision-making during diffusion, integrating user profiles and message content to guide the generation of subsequent messages [4]. Similarly, FUSE [9] and other LLM-based simulations [10] explore the evolution of news from factual to false, demonstrating that LLM agents can reproduce real-world phenomena by interacting within structured social networks. Beyond network topology, recent research has emphasized the importance of LLM-agent characteristics and behavioral heterogeneity in diffusion dynamics. Factors such as agent personality, decision-making strategies, and interaction patterns have been shown to significantly influence the spread of content, news, and misinformation. Simulations incorporating personality traits (e.g., the Big Five) and realistic user personas populated from survey data demonstrated how agent heterogeneity and platform-specific algorithms can shape propagation outcomes and disinformation dynamics [11], [12]. In addition to propagation dynamics, LLM-enhanced simulations have been applied to tasks such as early fake news detection. Frameworks like AVOID generate virtual propagation trajectories with differentiated agent roles, providing complementary social signals to improve content-based detection methods [13].

Despite recent advances in multi-agent and LLM-based simulations of information diffusion, most existing approaches focus on the spread of factual or false content and emphasize static properties such as network structure or agent-level traits. In contrast, the proposed framework models news narratives as dynamic constructs that evolve through repeated reinterpretation and interaction among LLM-driven agents. By capturing how reframing and local social feedback shape engagement over time, the approach enables early, predictive assessment of *Narrative Impact* and the identification of emerging amplification patterns before they fully unfold.

III. Methodology

The proposed framework uses a graph-based network in which multiple LLM-based agents interact to simulate the spread of news in a social network. The main goal is to use the simulation as a basis to predict the impact of a news in a real-world scenario. The process, illustrated in Figure 1, is based on the components described below.

A. Simulation

The simulation relies on an LLM-based multi-agent architecture in which each agent is tasked with producing a reaction to a given news headline. The available reactions are designed to capture varying degrees of impact and influence within the network. Also, a numerical score (s) is assigned to each reaction to then compute the overall impact:

1) **Award (without Comment):** *Maximum Impact* ($s = 6$). Represents the highest level of belief and active promotion, involving maximum emotional or financial investment.
2) **Share (with Comment):** *Very-High Impact* ($s = 5$). The agent actively amplifies the content. The generated comment is a *reframed version* of the headline, adapted to the agent's persona and ideological stance.
3) **Share (without Comment):** *High Impact* ($s = 4$). Passive amplification of the content's reach, indicating belief in the message without additional framing.
4) **Comment (Support/Agreement):** *Medium-High Impact* ($s = 3$). Reinforces the message's credibility without directly amplifying its reach to new clusters.
5) **Upvote:** *Medium-Low Impact* ($s = 1$). Signals passive agreement with minimal contribution to content diffusion.
6) **Comment (Dissent/Skepticism):** *Minimum Impact* ($s = -1$). Expresses opposition or doubt; while critical, the interaction marginally increases visibility.
7) **Ignore (Read without Action):** *Zero Impact* ($s = 0$). No observable influence on content propagation.
8) **Downvote:** *Negative Impact* ($s = -3$). Active rejection of the content results in decreased visibility and reduced amplification.

In addition to the original news headline, agents observe the reactions of their neighbors, forming a shared and continuously evolving social context. This context includes interaction signals (e.g., shares, awards, upvotes) as well as, when available, the textual content of comments generated by other agents. Within this setting, agents do not act in isolation but participate in a collaborative decision-making process, where individual reactions both influence and are influenced by local collective

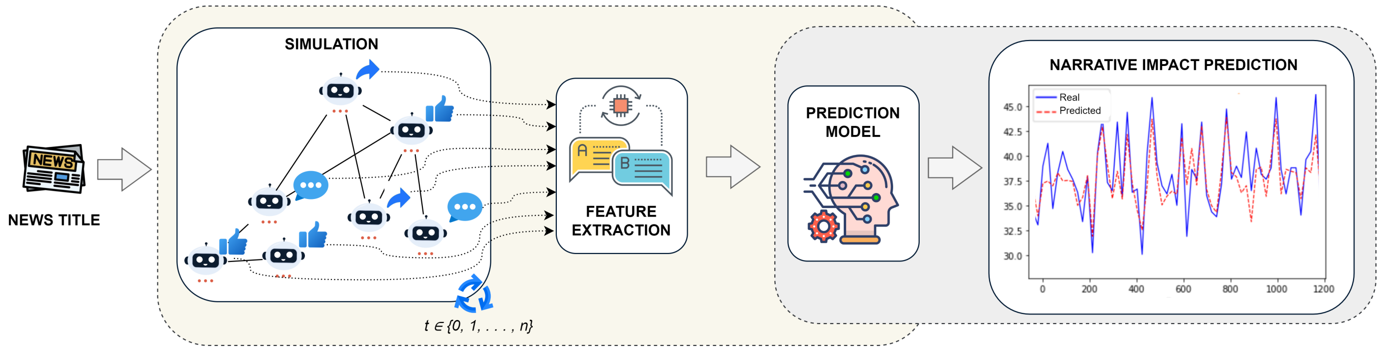

Fig. 1. Methodology

behavior. Reframed headlines produced through shares with comments function as collaborative signals that can alter the interpretation, emphasis, or ideological framing of the original news item. Through the accumulation of these interdependent contributions, narrative structures emerge, stabilize, or shift over time.

The collaboration in the proposed framework is realized at the signal level rather than at the model-optimization level. Each agent contributes partial, decentralized observations derived from its local interactions and reframing behavior.

B. Feature Extraction

The simulation enables the systematic tracking of agent interactions and reframing dynamics for each news item across multiple time steps, including its overall impact on the network at time t. The resulting data are used to construct a heterogeneous feature set capturing impact dynamics, user reactions, linguistic properties, and contextual information. Specifically, the features include: (i) normalized indicators of Narrative Impact over time; (ii) behavioral signals derived from reframing and commenting activity; (iii) sentiment- and bias-related linguistic features computed over both headlines and reframing comments; and (iv) categorical contextual information associated with the news topic. All features are computed at discrete time steps and capture both static properties and short-term temporal variations. These variables allow the model to assess how the impact at time t relates to reaction patterns and linguistic shifts observed at the previous step.

C. Prediction Model

A prediction model is trained to estimate the *Narrative Impact* of a news item at time $t+1$ from features observed at time t. By learning these incremental dependencies, the model enables the inference of narrative trends from partial observations of the simulation. By evaluating prediction performance across successive timesteps, the framework supports early alerting by identifying news items whose simulated impact is expected to intensify at early stages of narrative evolution, providing a principled basis for pre-bunking-oriented prioritization.

IV. Experimentation

The experimental evaluation investigates whether early signals from agent interactions and engagement patterns at time t can predict short-term Narrative Impact at time $t+1$. The following subsections describe the dataset, setup, and prediction strategy.

A. Datasets

To conduct the experimentation, two starting data sources were used. The primary dataset is the **BBC News Dataset**[1], a news corpus collected by scraping RSS feeds from the BBC News website. From this dataset, a random subset of 300 entries was extracted, consisting exclusively of news headlines published between December 4, 2023, and December 4, 2024. The selected headlines span multiple topical categories, including politics, sports, and general news events. This dataset serves as the source of input news items that trigger and drive the simulation dynamics.

The **Republican Community Dataset** consists of user-generated messages collected from the Reddit community r/Republican[2]. This subreddit represents an online space where Republican users discuss conservative values, political issues, and current events. The corpus was obtained through web scraping and includes approximately 1000 messages posted between August 4, 2018, and January 11, 2025. This dataset serves as a reference to characterize the ideological stance and guide the behavior of politically polarized agents in the simulation.

B. Experimentation Setup

This subsection details the experimental setup, beginning with a description of the simulation environment, including the implementation of the LLM-based multi-agent system and the agents' characteristics, followed by the prediction modeling component.

1) Simulation Details: To conduct the experimentation and coordinate the interaction workflow among agents, the CrewAI framework[3] was adopted. Each agent is driven by the same large language model, `gemini-2.0-flash`. To encourage response diversity and reduce deterministic behavior, the `temperature` parameter was set to 0.9. The network topology was configured according to a *Real-World* structure

[1] https://www.kaggle.com/datasets/gpreda/bbc-news
[2] https://www.reddit.com/r/Republican/
[3] https://docs.crewai.com/

that mimics the modular organization of social networks [14]. This topology is characterized by high connectivity within communities and sparse connections across different groups, thereby modeling the emergence of partially isolated clusters and echo chambers.

The simulated network consists of 20 agents, divided into two distinct user profiles: 10 *Skeptical Users* and 10 *Polarized Users*. Each agent is instantiated with an individual profile composed of a $backstory$, a $goal$, and a $personal\ identity$, including a full name and an occupation randomly generated using the Faker library[4]. The two agent types are defined as follows.

Skeptical User

- *Goal*: To question incoming information and attempt to verify factual accuracy.
- *Backstory*: A cautious and analytical individual who prioritizes evidence and reliability over emotional reactions.

Polarized User

- *Goal*: To react to news content primarily based on its ideological alignment.
- *Backstory*: A user strongly embedded in a specific political ideology, exhibiting selective exposure and motivated reasoning.

In addition to the shared attributes, **Polarized Users** possess a *worldview* that reflects their pre-existing beliefs and ideological biases. This worldview is constructed using 100 randomly selected messages from the Republican Community Dataset to guide the agents' interpretation and framing of news content.

The simulation is modeled as a discrete-time process where $t \in \{0, 1, \ldots, n\}$ with defined stopping criteria:

- $t = 0$: At the starting point, each agent in the network sees the input (i.e., a not reframed news), not considering the possible and different connections with other agents, and provides their own specific reaction to the given content. At this stage the provided reaction depends on the provided content and on the user profile. Also, the reaction for polarized users is filtered through the worldview lens.
- *From* $t = 1$ *to* $t = n$: In the following phases, agents continue to process the information, influenced by two additional and critical memory components:
  1) *Internal Consistency*: Agents retain memory of their own previous reactions to ensure a stable and coherent narrative identity over time.
  2) *Social Context*: Agents observe the reactions produced in $t-1$ by their connected peers. This peer-to-peer influence drives the narrative's emergent properties, as agents might moderate or radicalize their reframing in response to social feedback.

To evaluate the temporal evolution and define precise stopping criteria, each agent's categorical reaction is mapped to a numerical score (e.g., $s = 6$ for Award). The aggregate impact at a given timestep t is calculated as the sum of individual scores: $NIS_t = \sum_{a \in \mathcal{A}} s_a$, where $\mathcal{A}$ denotes the set of all agents in the network and s_a represents the numerical score assigned to the reaction of agent $a \in \mathcal{A}$. It was set $n = 5$ and define three specific stopping criteria:

- The simulation terminates if every agent in the pool chooses to ignore the content. This indicates that the narrative has lost all traction within the specific network cluster.
- The process stops if the variation in total impact between two consecutive steps falls below a sensitivity threshold α (i.e., 0.5). This represents a state where the *Narrative Impact* has reached a steady state.

2) Extracted Features: The simulation, conducted on the selected subset of the BBC News Dataset, enabled the systematic collection of the heterogeneous features introduced preliminarily in Section III. The feature set includes the following variables:

- **Impact**: *Narrative Impact score* at time t (the aggregate score was normalized such that the maximum observed value is set to 100).
- **N_reframings**: The number of reframing actions performed by agents (i.e., reaction = *Share (with Comment)*).
- **Reframing_Sentiment_t**: The average emotional polarity of the reframing activities measured with the `spacytextblob`[5] python library.
- **N_comm_sk**: The count of comments skeptical about the news.
- **N_comm_ag**: The count of comments that agree with the news.
- **Lexical_Bias_Title**: The probability of lexical bias measured within the headline. This probability has been measured through the model `mediabiasgroup/roberta-babe-ft`[6].
- **Mean_Lexical_Bias**: The average probability of lexical bias within all the reframing comments at time *t*.
- **Delta_Lexical_Bias**: The difference between *Mean_Lexical_Bias* at time t and *Mean_Lexical_Bias* at time $t-1$. It is useful to determine whether there is a shift in the lexicon used in the reframings between two subsequent steps.
- **Topic_Class**: The thematic category of the news item (e.g., politics, technology), treated as a categorical predictor, which has been identified by the model `WebOrganizer/TopicClassifier-NoURL`[7].

Additional information on features and their statistical distributions is presented in Table I.

3) Prediction Model: The prediction model leverages the XGBoost[8] (eXtreme Gradient Boosting) algorithm, an ensemble learning method chosen for its high efficiency and ability to capture non-linear relationships. The gradient boosting configuration is characterized by a large number of weak learners

[4]https://pypi.org/project/Faker/0.7.4/

[5]https://spacy.io/universe/project/spacy-textblob

[6]https://huggingface.co/mediabiasgroup/roberta-babe-ft

[7]https://huggingface.co/WebOrganizer/TopicClassifier-NoURL

[8]https://xgboost.readthedocs.io/en/stable/

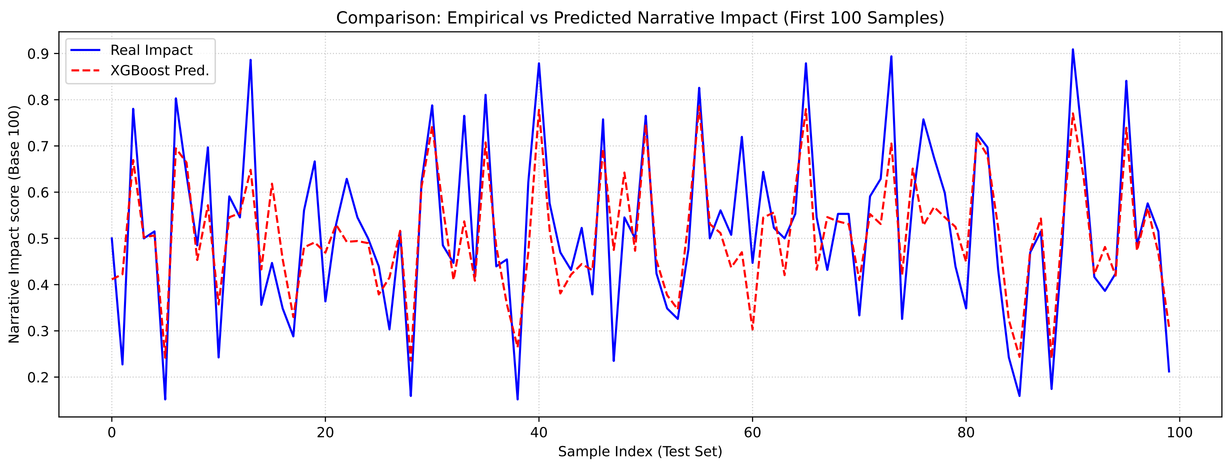

Fig. 2. Comparison between simulated (Real) and predicted *Narrative Impact* scores for the first 100 test samples.

TABLE I
STATISTICS OF FEATURE SET

Feature	Mean	SD	Min	Max
Impact	51.28	12.44	14.44	87.78
N_reframings	5.02	3.20	0.00	18.00
Reframing_Sentiment_t	0.05	0.20	-0.78	1.00
N_comm_sk	4.21	3.94	0.00	17.00
N_comm_ag	1.90	2.13	0.00	16.00
Mean_Lexical_Bias	0.63	0.31	0.00	0.99
Lexical_Bias_Title	0.15	0.24	0.03	0.98
Delta_Lexical_Bias	0.08	0.29	-0.99	0.99

trained with a low learning rate (1000 estimators, learning rate 0.01). Tree depth is deliberately constrained (maximum depth of 5), and both row- and feature-level subsampling (0.8) are applied to control model complexity. Training is monitored via early stopping with a patience of 50 rounds. The model is optimized using a squared-error objective, and all experiments are conducted with a fixed random seed to ensure reproducibility. The implemented model uses a sliding-window approach, where the feature vector at time t is used to predict the impact at time $t+1$.

C. Results

Upon completing the training phase, the predictive performance of the XGBoost model was evaluated on the held-out test set to determine its efficacy in forecasting the *Narrative Impact* score. The model demonstrated substantial explanatory power, achieving a Coefficient of Determination (R^2) of 0.71. The precision of the model's estimations was quantified using the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE), which were 9.69 and 7.58, respectively.

Figure 2 illustrates the comparison between the real (i.e., obtained from the simulation) *Narrative Impact* score and the predicted one from the trained model for each of the first 100 samples of the test set. The figure shows values that are highly similar, confirming the model's accuracy. Additionally, the essential role of the aforementioned feature set is empirically validated by an ablation study. Results in Table II demonstrate that a model relying solely on static content features (i.e., *Lexical_Bias_Title* and *Topic_Class*) achieves limited predictive accuracy ($R^2 = 0.24$), while adding also the total impact (i.e., *Impact*) improves the total accuracy significantly ($R^2 = 0.61$). However, including also the social interaction signals described in Section IV leads to the best overall performance.

TABLE II
ABLATION STUDY RESULTS COMPARING MODEL PERFORMANCE WITH AND WITHOUT SOCIAL INTERACTION FEATURES.

Feature Set	$R^2 \uparrow$	$RMSE \downarrow$	$MAE \downarrow$
Static Content	0.24	11.48	8.93
Static Content + Impact	0.61	11.13	8.65
w/ Social Interactions	**0.71**	**9.69**	**7.58**

Original Title: Interest rates expected to be held at 5.25%.

Reframed Versions at time t:

- Stagnant rates – a reflection of Biden's failing economy.
- Biden's economy continues its stagnation: Interest rates remain unchanged.
- Interest rates steady – a potential sign of stability in Biden's economy?
- Biden's economy: Interest rates stuck, burdening Americans.

Fig. 3. Example of the evolution of a news from the original title to its reframed versions at time t.

As for qualitative results, Figure 3 illustrates the collaborative reframing process, showing how a neutral economic headline (i.e., *Interest rates expected to be held at 5.25%*) evolves as it passes through the ideological lenses of the simulated agents. The original headline is purely factual and lacks political attribution. However, the reframed versions demonstrate a clear "politicization" of the narrative. The agents do not change the underlying fact (that rates are steady), but they wrap it in a new interpretive frame for example by linking the interest rate decision to "Biden's economy". This example highlights the framework's ability to capture how narratives gain structural relevance not by changing facts, but by modifying the linguistic and ideological context through social interaction.

Figure 4 illustrates how two distinct narratives evolve in nearly opposite directions over time. The first headline, "*Weather record for warmest January set in Scottish Highlands*" (blue line), begins with a modest initial impact but exhibits steady and nearly linear growth, peaking at $t = 5$ with a score close to 70. In contrast, the news title "*Woods set for record Masters cut despite sandstorm*" (orange line) follows a different trend: it reaches its peak at $t = 3$ before undergoing a pronounced decline. This difference may be attributable to the specific content of the news headline: because the title refers to climate issues, it addresses themes that are highly salient to ideological communities. The second headline is a sport-centric news outlet that lacks the same level of ideological friction. Consequently, once the immediate curiosity about the event is satisfied, the story fails to sustain long-term resonance and quickly fades from public discourse.

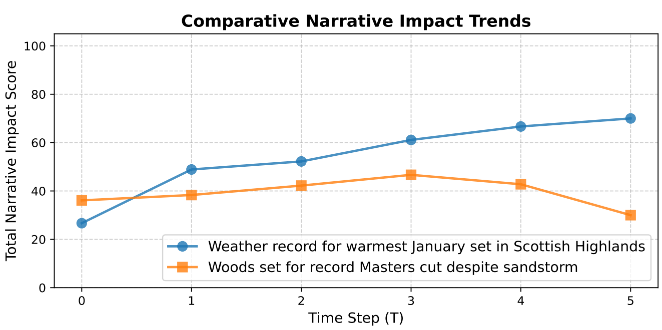

Fig. 4. Comparison between the impact of two different news stories across the various time steps (from 0 to 5).

V. Conclusion & Limitations

This paper introduces an LLM-based multi-agent framework designed to simulate and predict the evolution of news narratives in collaborative environments, demonstrating the feasibility of an anticipatory AI system capable of identifying impactful narrative trajectories. Experimental results validate the effectiveness of combining generative agent-based simulations with predictive modeling. The XGBoost regressor achieved high accuracy ($R^2 = 0.71$), proving that the structural impact of a news item is often encoded in its reframing dynamics, even in scenarios not characterized by intentional misinformation.

Despite these promising results, some limitations remain. The current simulation is constrained by a fixed network topology and a limited number of agent profiles. Furthermore, the numerical scoring system (s) for agent reactions is currently heuristic; while functional for this study, it lacks grounding in established psychological or sociological frameworks. Future research will integrate empirically validated metrics, explore more dynamic and large-scale network structures, and include a broader range of ideological worldviews.

Acknowledgment

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

[1] D. Cavaliere, M. Gallo, and C. Stanzione, "Propaganda detection robustness through adversarial attacks driven by explainable ai," in *World Conference on Explainable Artificial Intelligence*. Springer, 2023, pp. 405–419.

[2] D. Barman and O. Conlan, "Evaluating prebunking and nudge techniques in tackling misinformation: A between-subject study on social media platforms," in *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, 2024, pp. 167–177.

[3] M. Di Gisi, G. Fenza, D. Furno, M. Gallo, V. Loia, and P. P. Trotta, "Federated prompt tuning for news framing: A community-aware approach to narrative exploitability," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[4] Y. Hu, G. Sherpa, L. Zhang, W. Li, Q. Bai, Y. Wang, and X. Wang, "An llm-enhanced agent-based simulation tool for information propagation," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 8679–8682.

[5] S. J. Alam and A. Geller, "Networks in agent-based social simulation," in *Agent-based models of geographical systems*. Springer, 2011, pp. 199–216.

[6] Z. Chen, "An agent-based model for information diffusion over online social networks," *Papers in Applied Geography*, vol. 5, no. 1-2, pp. 77–97, 2019.

[7] H. R. Nasrinpour, M. R. Friesen *et al.*, "An agent-based model of message propagation in the facebook electronic social network," *arXiv preprint arXiv:1611.07454*, 2016.

[8] J. S. Park, L. Popowski, C. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Social simulacra: Creating populated prototypes for social computing systems," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3526113.3545616

[9] Y. Liu, Z. Song, J. Zhang, X. Zhang, X. Chen, and R. Yan, "The stepwise deception: Simulating the evolution from true news to fake news with LLM agents," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, nov 2025, pp. 26 176–26 192. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1330/

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

[11] X. Li, Y. Xu, Y. Zhang, and E. C. Malthouse, "Large language model-driven multi-agent simulation for news diffusion under different network structures," 2024. [Online]. Available: https://arxiv.org/abs/2410.13909

[12] Y. Liu, X. Chen, X. Zhang, X. Gao, J. Zhang, and R. Yan, "From skepticism to acceptance: simulating the attitude dynamics toward fake news," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 7886–7894.

[13] B. Gu, M. Gao, J. Yu, Z. Wang, Z. Liu, K. Shu, and H. Zhang, "Ahead of the spread: Agent-driven virtual propagation for early fake news detection," 2026. [Online]. Available: https://arxiv.org/abs/2601.02750

[14] J. Leskovec and J. Mcauley, "Learning to discover social circles in ego networks," *Advances in neural information processing systems*, vol. 25, 2012.