The Stepwise Deception: Simulating the Evolution from True News to Fake News with LLM Agents

Yuhan Liu[1], Zirui Song[2], Juntian Zhang[1], Xiaoqing Zhang[1], Xiuying Chen[2]*, Rui Yan[1,3]*
[1]Gaoling School of Artificial Intelligence, Renmin University of China,
[2]Mohamed bin Zayed University of Artificial Intelligence,
[3]School of Artifcial Intelligence, Wuhan University
{yuhan.liu, zhangjuntian, xiaoqingz, ruiyan}@ruc.edu.cn
{zirui.song, xiuying.chen}@mbzuai.ac.ae

Abstract

With the growing spread of misinformation online, understanding how true news evolves into fake news has become crucial for early detection and prevention. However, previous research has often assumed fake news inherently exists rather than exploring its gradual formation. To address this gap, we propose **FUSE** (**F**ake news evol**U**tion **S**imulation fram**E**work), a novel Large Language Model (LLM)-based simulation approach explicitly focusing on fake news evolution from real news. Our framework model a social network with four distinct types of LLM agents commonly observed in daily interactions: *spreaders* who propagate information, *commentators* who provide interpretations, *verifiers* who fact-check, and *bystanders* who observe passively to simulate realistic daily interactions that progressively distort true news. To quantify these gradual distortions, we develop **FUSE-EVAL**, a comprehensive evaluation framework measuring truth deviation along multiple linguistic and semantic dimensions. Results show that FUSE effectively captures fake news evolution patterns and accurately reproduces known fake news, aligning closely with human evaluations. Experiments demonstrate that FUSE accurately reproduces known fake news evolution scenarios, aligns closely with human judgment, and highlights the importance of timely intervention at early stages. Our framework is extensible, enabling future research on broader scenarios of fake news: FUSE.

1 Introduction

The rapid spread of fake news has become a significant global concern (Lazer et al., 2018a; Olan et al., 2022). Prior research predominantly addresses fake news detection or simulates the spread of misinformation after its initial appearance (Garimella et al., 2017; Wang et al., 2019). For instance, Piqueira et al. (2020) categorized individuals into four types and used mathematical models to simulate the spread of fake news, as depicted in Figure 1(a). On a micro-level, Jalili and Perc (2017) defined numerical conditions for opinion change to study fake news dissemination, as shown in Figure 1(b).The utilization of language models to improve security is of critical importance (Zhang et al., 2023; Zhang et al., 2024). However, these models typically assume fake news as inherently existing entities within social networks, ignoring how misinformation originates or evolves over time.

In contrast, fake news may originate from true news that becomes distorted or misinterpreted over time, eventually evolving into fake news (Guo et al., 2021; Shen et al., 2024) as illustrated in Figure 1(c). This evolutionary process is critically underexplored despite its significance for effective early interventions. Recognizing this gap, our work explicitly adopts the definition of fake news from prior influential research (Lazer et al., 2018b), focusing specifically on scenarios where factual information incrementally transforms into misinformation during its dissemination (Figure 1(c)). We define this transitional content as *partially evolved fake news*, characterized by a mix of accurate and distorted elements.

Specifically, we propose the Fake news evolUtion Simulation framEwork (**FUSE**), the first comprehensive approach employing LLM agents to simulate how real news progressively evolves into fake news within different social network structures (e.g., high-clustering, scale-free, and random networks). The simulation consists of four distinct agent roles commonly found in real-world interactions: *spreaders*, who disseminate information; *commentators*, who interpret content; *verifiers*, who assess factual accuracy; and *bystanders*, who observe without active participation. Each agent engages daily, exchanging beliefs, reevaluating information, and contributing to incremental

* Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26176–26192
November 4-9, 2025

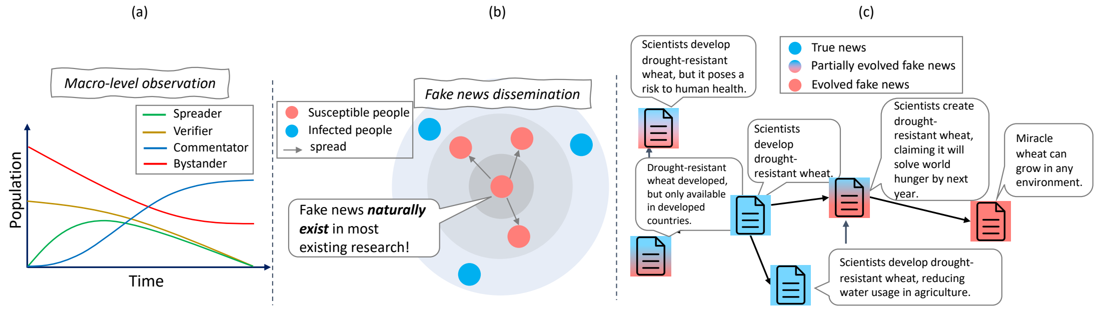

Figure 1: (a) Macro-level observation of population dynamics based on the mathematical model, categorizing individuals into four types and showing their quantity changes over time. (b) The micro-level conventional fake news dissemination model assumes that fake news inherently exists. (c) Micro-level evolution of fake news, where true news gradually evolves into fake news during network propagation with content alterations at various stages.

content distortions. Our agents incorporate hierarchical memory structures, combining short-term interactions and long-term knowledge, allowing realistic reflective reasoning processes and dynamic content adaptation.

Given the absence of prior work on language-based evaluation of fake news evolution, we introduce **FUSE-EVAL**, a novel multidimensional evaluation framework that quantifies the deviation of evolved news from its original form across multiple dimensions, including Sentiment Shift (SS), New Information Introduced (NII), Certainty Shift (CS), STylistic Shift (STS), Temporal Shift(TS), and Perspective Deviation (PD). Our comprehensive experiments validate FUSE's *strong alignment* with real-world observations from prior research. The results reveal three key findings: (1) news exhibits *clear accumulation distortion effects*, where content progressively deviates from its original form during spread (de Paula et al., 2024); (2) true news evolution to fake news occurs more rapidly in *high-clustering networks* than in scale-free or random networks (Trpevski et al., 2010); (3) *political news* shows significantly faster evolution rates compared to other topics (terrorism, natural disasters, science, and finance) (Lazer et al., 2018b).

To construct a responsible online environment, our research reveals the importance of *strategic interventions during the early stages of fake news evolution*. Rather than waiting until fake news has widely spread, we introduce an official agent that intervenes when information deviation reaches critical thresholds, issuing authoritative statements with reliable sources to counteract misinformation spread. This early intervention approach demonstrates the effectiveness of timely, authoritative responses in misinformation governance.

Our contributions can be summarized as follows:

• ***Versatile Framework.*** We propose FUSE, an LLM-based simulation framework to investigate how true news gradually evolves into fake news, and validate through experiments that our framework successfully reproduces real-world phenomena by considering different types of agents and various social network structures.

• ***Comprehensive Evaluation.*** We introduce FUSE-EVAL, a novel multidimensional framework to measure the deviation from true news during news evolution.

• ***Practical Insights.*** We propose and evaluate multiple intervention strategies aimed at mitigating the spread of fake news during its evolution.

2 Related Work

Fake News Evolution Recent research into fake news evolution has focused on how misinformation spreads and transforms over time. Zhang et al. (2013) found that rumors evolve as they are repeatedly modified, becoming shorter and more shareable, while Guo et al. (2021) empirically tracked fake news evolution, noting how sentiment and text similarity change as truth transitions into misinformation. Xia et al. (2020) proposed a sentiment analysis pipeline to track public opinion shifts in fake news by detecting sarcasm. Other studies have emphasized structural and behavioral aspects of fake news propagation. Zhao et al. (2024) proposed a dynamic method that captures temporal changes in rumor propagation, revealing how rumor patterns evolve. Wang et al. (2021) demonstrated slight news content changes during the COVID-19 pandemic, while Li et al. (2016) examined how user behaviors, particularly the role of verified accounts, influence the evolution of rumors. FPS (Liu et al.,

2024) and TED (Liu et al., 2025) uses a multi-agent system to study the propagation and detection of fake news.

However, there has not been a detailed and comprehensive study on how true news evolves into fake news, with only some superficial linguistic analyses (Zhang et al., 2013; Guo et al., 2021).

LLMs as Agents Agent-based modeling simulates complex systems through individual agents' interactions in dynamic environments (Macal and North, 2005). The integration of LLMs has enhanced these simulations by enabling natural language processing capabilities (Zhang et al., 2025; Chen et al., 2023b,a) and human-like intelligence in planning and decision-making (Xi et al., 2023). This has led to widespread adoption across various domains (Li et al., 2023; Park et al., 2023; Liu et al., 2025; Jin et al., 2025), establishing LLM agents as a new paradigm for human-level intelligence simulation. In more specific applications, LLM agents have been employed to simulate social media dynamics. For instance, Törnberg et al. (2023) used them to investigate social media algorithms and provide insights into real-world phenomena, while Park et al. (2022) demonstrated their ability to generate human-like social media content. Our work extends this approach by being one of the first to apply LLM agents in simulating fake news evolution.

3 Methodology

3.1 Problem Formulation

We simulate the gradual evolution of true news into fake news using LLMs as agents within a social network, consistent with the definition of fake news provided by prior research (Lazer et al., 2018b; Guo et al., 2021). The simulation consists of N agents $\mathcal{A} = (a_1, \ldots, a_N)$, each endowed with a unique persona defining their Behavior role, personality traits, and demographic information.

At time $t = 0$, true news S_0 is introduced into the network. The agents are connected according to a predefined social network structure $\mathcal{G} = (\mathcal{A}, \mathcal{E})$, which may represent high-clustering, scale-free, or random networks to reflect real-world dynamics. On each day $t = 1, 2, \ldots, T$, agents interact with their neighbors $\mathcal{N}_i$, exchanging information and opinions based on their personas and prior knowledge. After interactions, agents process and reintroduce the news content based on their updated beliefs. The evolution of the news content for agent a_i with a personal profile $\mathcal{P}_i$ at time t, denoted as S_i^t, is defined by:

$$S_i^t = f(S_i^{t-1}, \{S_j^{t-1} | a_j \in \mathcal{N}_i\}, \mathcal{P}_i), \qquad (1)$$

where $f(\cdot)$ represents the agent's information processing function.

Through this simulation, we analyze how the true news S_0 transforms over time due to agents' interactions and personal biases, examining the impact of agent types, network structures, and individual traits on the evolution of fake news.

3.2 Our Simulation Framework

As depicted in Figure 2, our FUSE framework consists of two core components: the *Propagation Role-Aware agents* (PRA) and the *News Evolution Simulator* (NES). The PRA module empowers agents with role-based decision-making capabilities, while the NES establishes the interaction environment, simulating the social network through which news propagates and evolves. Within the PRA module, each agent is powered by an LLM and characterized by a specific role type and personal attributes, which govern their information processing, interaction patterns, and opinion updates. The NES facilitates daily interactions through a predefined social network structure, $\mathcal{G} = (\mathcal{A}, \mathcal{E})$, simulating various network types to reflect different social dynamics.

During each simulation day, agents engage with their network neighbors, exchanging news content and opinions shaped by their roles and attributes. When news content deviates beyond a set threshold, intervention mechanisms—such as official announcements—are triggered to provide credible information and correct potential misinformation. The simulation advances daily with updated agent states, tracking the evolution of news content through the network.

3.3 Propagation Role-Aware Agent

The PRA is designed to simulate individual human behaviors in news evolution by equipping agents with specific roles and personal attributes, aiming to mirror the diversity and complexity of human interactions in social networks.

3.3.1 Personal Information

According to Sun et al. (2023), the roles in fake news propagation can be classified into four types: *spreaders*, who propagate information;

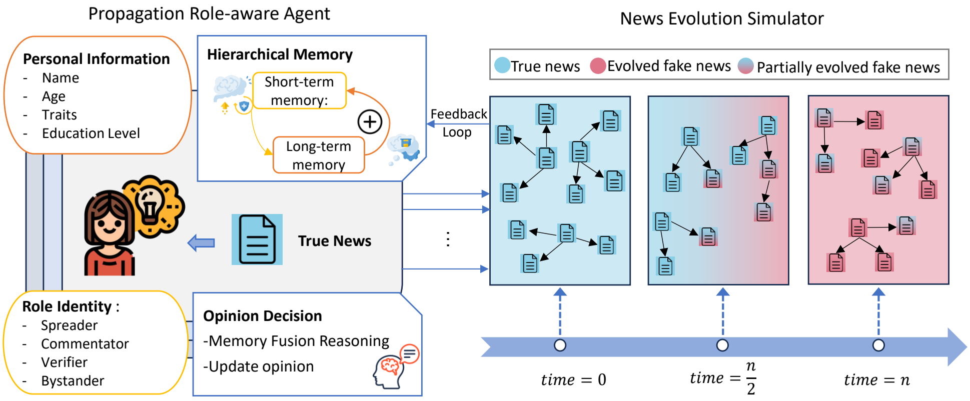

Figure 2: Our FUSE framework simulates news evolution by equipping each agent with role-based decision-making capabilities. Propagation Role-aware agents (PRA) process true news through interactions within the news evolution simulator (NES), where their role identities shape how they engage with the news.

commentators, who provide opinions and interpretations; *verifiers*, who check the accuracy of information; and *bystanders*, who passively observe without engaging. However, they failed to model this in their numerical simulation. We follow this setup but enhance it by equipping each agent a_i with a textual role description $r_i \in \{$*spreader, commentator, verifier, bystander*$\}$. Additionally, agents possess a personal profile $\mathcal{P}_i$ that includes demographic attributes (name, age, gender, and education level) and personal traits based on the Big Five model (Barrick and Mount, 1991), which influence their information processing behaviors.

3.3.2 Role-Specific Behaviors

At each time step t_i, agent a_i holds a version of the news content S_i^t. When interacting with neighboring agents $\mathcal{N}_i$ as defined by the network $\mathcal{G}$, agent a_i receives news content $\{S_j^{t-1} | a_j \in \mathcal{N}_i\}$. The agent then reintroduces news based on their role and persona through a role-specific update function:

$$f_{role} = f_{r_i}(S_i^{t-1}, \{S_j^{t-1} | a_j \in \mathcal{N}_i\}, \mathcal{P}_i). \quad (2)$$

For different roles in our model, *spreaders* may combine and amplify sensational aspects of the news, *commentators* may add personal opinions, *verifiers* may check news before sharing, and *bystanders* may retain their previous news content unless significantly influenced (Sun et al., 2023).

3.3.3 Memory and Reflection

In our simulation, agents engage with their neighbors daily, leading to updated versions of the news. Given the volume of interactions, we implement a hierarchical memory system comprising short-term memory (STM) M_i^S for recent interactions and long-term memory (LTM) M_i^L for accumulated knowledge. After interactions, agents reflect and update the news through a memory function:

$$M_i^{L,t} = g(f_L(M_i^{L,t-1}), f_S(M_i^{S,t})), \quad (3)$$

where $g(\cdot)$ integrates new information into LTM, enabling agents to exhibit dynamic behaviors such as gradually changing their opinion on a topic or reinforcing existing opinions.

3.3.4 Decision-Making Process

In our FUSE framework, each agent's opinion evolves through a reasoning process influenced by their role, persona, and interactions. Agents reflect on their news content after daily interactions and memory updates, leading to gradual opinion changes. The decision-making process for agent a_i at time t is modeled as:

$$S_i^t = f_{dm}(S_i^{t-1}, m_i^{L,t-1}, r_i, \mathcal{P}_i). \quad (4)$$

This function captures how agents integrate new information with their existing opinions, considering their role in the decision-making process. For example, the reasoning of spreaders may lead to greater changes in S_i^t, commentators add subjective nuances, verifiers aim to correct inaccuracies, and bystanders typically make minimal changes.

3.4 News Evolution Simulator

The News Evolution Simulator (NES) provides the environment where news content evolves over time through agent interactions within a social network structure $\mathcal{G} = (\mathcal{A}, \mathcal{E})$. This module enables studying how true news transforms into fake news through agent behaviors and social interactions.

NES models various network topologies to reflect different social dynamics: random networks with randomly formed edges between agents $a_i \in \mathcal{A}$, simulating loosely connected environments; scale-free networks with hub agents acting as "super-spreaders"; and high-clustering networks forming tightly-knit communities that mirror real-world social circles (Nekovee et al., 2007; Moreno et al., 2004). As outlined in Appendix I, the network structure $\mathcal{G}$ determines daily agent interactions, influencing news content's evolution patterns. The overall algorithm is presented in Appendix A.

3.4.1 Intervention Mechanisms

A key feature of NES is its ability to simulate interventions to counter fake news evolution. When the deviation between current news content S_i^t and original news S_0 exceeds a predefined threshold, an official agent is introduced to provide verified information and correct misinformation.

The intervention process starts with continuously monitoring the deviation between each agent's news content and the original news. Once the deviation exceeds a critical threshold, the official agent is triggered to take action. This agent issues official announcements based on reliable sources, targeting agents most likely to propagate or exacerbate misinformation.

The prompts for all functions mentioned in § 3 can be found in Appendix B.

4 FUSE-EVAL: News Evolution Analysis

To systematically measure how true news evolves into fake news within our simulation, we propose a comprehensive evaluation framework named **FUSE-EVAL**. This framework consists of two sets of metrics: *Content Deviation Metrics* and *Statistical Deviation Metrics*, which together provide a detailed understanding of how fake news evolves within the simulated environment.

4.1 Content Deviation Metrics

The Content Deviation Metrics assess the deviation of the news content across multiple dimensions by quantifying changes in specific aspects of the news. FUSE-EVAL evaluates the news content based on six core dimensions:

(1) **Sentiment Shift (SS)** measures the change in emotional tone between the original news content and its evolved version (Lu et al., 2022; Ma et al., 2021). Sentiment plays a crucial role in how information is perceived and shared, with shifts indicating potential bias or emotional manipulation.

(2) **New Information Introduced (NII)** assesses the extent to which additional information, not present in the original news, has been incorporated (Wang et al., 2017). Introducing new facts or claims can significantly alter the original message, potentially leading to misinformation.

(3) **Certainty Shift (CS)** evaluates changes in the level of confidence or assertiveness expressed in the news content (Krafft et al., 2019; Kim and Yoon, 2022). Shifts from definitive to speculative language can influence the perceived credibility of information.

(4) **Stylistic Shift (STS)** examines changes in writing style, tone, and linguistic features (Wu et al., 2024). Alterations in style can affect readability and audience engagement through formality and sentence complexity changes.

(5) **Temporal Shift (TS)** measures changes related to time references within the news content (Shen et al., 2024; Mu et al., 2023). Modifying dates, times, or event sequences can significantly impact news interpretation.

(6) **Paraphrasing Degree (PD)** evaluates the extent to which the content has been rephrased from the original text, which may obscure meaning or introduce ambiguity.

We employ `GPT-4o-mini` to automate FUSE-EVAL evaluation, scoring each dimension from 1 (minimal deviation) to 10 (significant deviation).

As shown in Figure 3 (a), FUSE-EVAL demonstrates cumulative deviations (Pröllochs and Feuerriegel, 2023) during fake news evolution, confirming its effectiveness. To evaluate the overall deviation, the **Total Deviation (TD)** for each agent at each time step t is calculated as:

$$\mathrm{TD}_i^t = \frac{1}{6} \sum_{d=1}^{6} D_{i,d}^t, \tag{5}$$

where $D_{i,d}^t$ is the score of dimension d for agent i at time t. The detailed evaluation process is provided in Appendix D.

4.2 Statistical Deviation Metrics

The Statistical Deviation Metrics, derived from Total Deviation (TD) scores, provide insights into the overall patterns of news evolution within the network. We analyze several key metrics:

• The **ΔDeviation** represents the difference in Average Deviation between the final and initial simu-

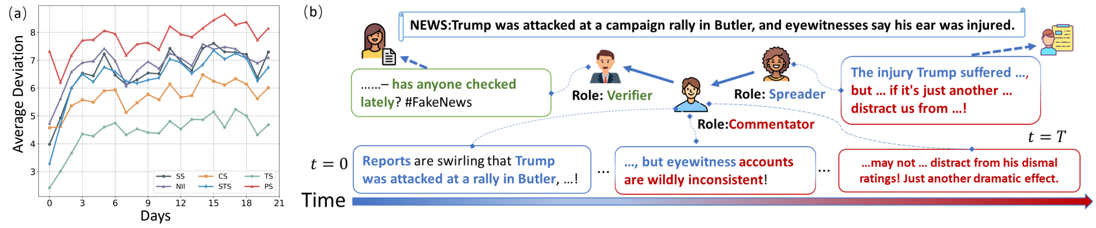

Figure 3: (a) The FUSE-EVAL scores show cumulative information deviations over time. (b) A Case of FUSE: True news gradually evolves into partially false and eventually entirely fake news over time.

lation day, indicating overall deviation growth.

- The **Average Deviation** is the mean of TD across all agents at each time step, showing the general trend of news evolution within the network.
- The **Deviation Variance** measures the statistical variance of TD among agents, measuring how uniformly content deviates across the network.
- The **Final Deviation** is the average TD at the finaltime step t, representing the cumulative effect.
- The **Maximum Deviation** and **Minimum Deviation** refer to the highest and lowest average TD observed, showing the extremes of news deviation.
- The **Peak Deviation Time** indicates the percentage of simulation time taken to reach Peak Deviation Rate, showing the speed of maximum deviation occurrence.
- The **Half Deviation Time** is the time step $t_{0.5}$ when average TD reaches half of Max Deviation, indicating the rate of significant deviation.

4.3 Implementation Details

Our framework uses `GPT-4o-mini` as the primary LLM and the simulation comprises 40 agents. Additional implementation details, including agent personality traits and programming environment, are provided in Appendix C. At the same time, API costs and compatibility with other models can be found in Appendix G and Appendix H.

5 Validation of the FUSE Framework

In this section, we demonstrate FUSE's effectiveness by validating its alignment with known fake news propagation patterns and its ability to reproduce real-world fake news.

5.1 Alignment with Real-World Patterns

Topic Comparison We analyzed fake news evolution across five topics: politics, science, finance, terrorism, and urban legends. As shown in Table 1 and Appendix F, political fake news exhibits the fastest spread, with average deviation peaking within four days, followed by terrorism-related content. Science and financial news evolve more slowly, showing the lowest average deviation. Table 1 shows the final deviation for political news is approximately 90% higher than that of science news. These results indicate that political fake news is more prone to rapid distortion and widespread belief, while science-related misinformation spreads more cautiously, aligned with prior research (Lazer et al., 2018a). We collected 120 pieces of true news across five topics. All news is published after the training cutoff date of GPT-4o-mini. The results were consistent, and the dataset will be publicly available.

Social Network Comparison We analyzed fake news evolution across three network structures (random, scale-free, and high-clustering) using a terrorism topic. Table 1 shows that high-clustering networks lead to the fastest and most extensive fake news spread, with deviation peaking rapidly and remaining high. This indicates that tightly connected communities are particularly susceptible to rapid belief distortion, aligning with the "echo chamber" effect (Cinelli et al., 2021). Random networks show the slowest evolution of fake news with lower variance, while scale-free networks exhibit intermediate behavior. Peak deviation time is the longest in random networks and shortest in high-clustering networks, illustrating that clustering accelerates fake news evolution, consistent with prior research (Lind et al., 2007; Trpevski et al., 2010).

Spread Type Comparison We analyzed three spread types (normal, emotional, and super spread) using a terrorism topic. Super spread, assigned to high-degree nodes, leads to the highest misinformation level due to influencer amplification. Emotional spread, characterized by heightened emotional language, shows moderate effects, while normal spread exhibits the slowest evolution. As shown in Table 1, peak deviation time is shortest in super spread, followed by emotional spread, demonstrating their accelerating effect on misinfor-

Comparison Factor	Setting	Δ Deviation↓	Average Deviation↓	Deviation Variance↓	Max Deviation↓	Min Deviation	Final Deviation↓	Peak Deviation Time ↑	Half Δ Deviation Time↑
Topic	Politics	3.148	6.594	0.511	7.440	3.442	6.590	0.133	0.033
	Science	**1.446**	**3.533**	**0.207**	**4.236**	**2.026**	**3.472**	**0.767**	**0.033**
Network Structure	**Random**	**1.905**	**3.315**	**0.347**	**4.206**	**1.892**	**4.206**	**1.000**	**0.233**
	Scale-Free	2.631	4.287	0.725	5.652	1.492	4.955	0.767	0.167
	High-Clustering	4.313	6.193	1.027	7.030	2.348	6.661	0.500	0.033
Spread Type	**Normal Spread**	**1.176**	**3.536**	**0.606**	**4.705**	**1.398**	**3.524**	**0.800**	**0.133**
	Emotional Spread	1.688	4.182	0.456	5.105	2.008	4.303	0.333	0.067
	Super Spread	2.920	4.434	0.672	5.613	2.054	5.067	0.700	0.100
Traits	Impressionable	3.088	4.998	0.956	6.428	2.262	5.677	0.667	0.133
	Vigilant	**1.945**	**4.081**	**0.446**	**5.021**	**2.485**	**4.593**	**0.400**	**0.133**
Intervention	No Intervention	3.208	5.546	1.247	7.340	1.841	6.383	0.767	0.167
	Intervention	**1.384**	**4.207**	**0.476**	**5.302**	**1.841**	**4.559**	**0.200**	**0.067**

Table 1: Comparative analysis of fake news evolution across different settings, including variations in topics, social networks, spread traits, and intervention strategies. ↑ or ↓ arrows represent better control of fake news evolution. **Bold** numbers indicate statistically significant improvements over baseline models (t-test with p-value<0.01).

mation evolution, aligned with prior research (Sun et al., 2023).

Personality Traits Comparison Using a terrorism topic, we compared the impact of personality traits on fake news evolution. Based on the Big Five personality traits (Barrick and Mount, 1991), we compared agents with high agreeableness and neuroticism (Impressionable) versus low levels (Vigilant). Table 1 shows that Impressionable agents are more prone to accepting and spreading misinformation. In contrast, Vigilant agents maintain more stable beliefs, aligning with previous studies on personality influence in fake news spread (Mirzabeigi et al., 2023).

5.2 Alignment with Real-World Fake News

We conducted experiments across various topics and found that the fake news evolved by the FUSE framework closely corresponds to real-world fake news. As shown in Figure 3 (b), the news about "Trump being attacked" starts as true, evolves into partially false, and eventually becomes entirely fake. As a commentator, the agent often adds its own views, while its neighboring verifiers and spreaders act according to their roles. Additionally, our framework generates fake news such as "Trump was not attacked. It's a dramatic effect," which is also a widely circulated piece of fake news in the real world case 1 and case 2. From a quantitative analysis perspective, for each topic, 73% of fake news is recovered by our framework. The detailed case study and analysis results are provided in the Appendix E.

6 Analysis and Discussion

6.1 Ablation Study

We chose a terrorism topic to demonstrate the effectiveness of our model's components and conducted two ablation studies to evaluate the contribution of key components in the FUSE framework.

The Impact of Hierarchical Memory and Propagation-Role. As shown in Figure 4 (a), the complete FUSE framework demonstrates apparent deviation accumulation, indicating its effectiveness in simulating fake news evolution. After removing hierarchical memory, the deviation significantly drops, with a 39.8% reduction throughout the simulation, indicating the simulation fails (Pröllochs and Feuerriegel, 2023). This highlights memory's crucial role in capturing persistent belief distortion through short-term and long-term information processing. Similarly, removing propagation roles leads to further deviation decrease, emphasizing how distinct agent roles (spreader, commentator, verifier, and bystander) shape information evolution. Without these roles, the agents behave more uniformly, and the accumulation effect of deviation disappears, meaning that the news does not evolve.

The Impact of Propagation Role Types. Following our first ablation study showing that removing propagation roles leads to simulation failure, we conducted a detailed analysis of different agent roles' impact on fake news evolution. As shown in Figure 4 (b), removing commentators caused the most significant drop in average deviation, confirming their crucial role in false news spread through opinion addition and interpretation. Removing spreaders had a relatively minimal impact as they

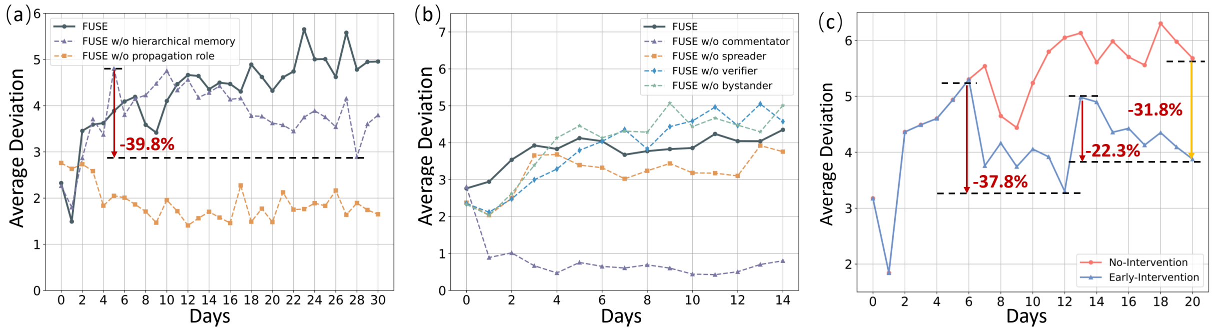

Figure 4: (a) Ablation study showing the effectiveness of hierarchical memory and propagation roles. (b) Impact of removing different agent types on fake news evolution. (c) Effectiveness of early intervention, showing an apparent reduction in deviation over time compared to the no-intervention condition.

lack opinion-adding capabilities, though they still contribute to information dissemination.

Removing verifiers increased overall deviation, demonstrating their important role in maintaining information accuracy through fact-checking. Without verifiers, the system became more susceptible to misinformation spread. Bystander removal showed the least effect, consistent with their passive observational role in the network.

These findings, combined with our previous ablation results on hierarchical memory and propagation roles, validate FUSE's effectiveness and demonstrate how different components contribute to simulating fake news evolution.

6.2 Fake News Intervention Strategy

Based on previous results, we implemented interventions through an official agent at high-degree nodes. As shown in Table 1 and Figure 4 (c), when fake news evolution peaked on the sixth day, our first intervention reduced deviation by 37.8% compared to no-intervention. Although this effect gradually weakened, with the gap narrowing to 22.3% by day 12 as agents continued to interact and potentially revert to previous beliefs, a second intervention on day 16 achieved a 31.8% reduction in deviation. The intervention strategy demonstrated several significant improvements over the no-intervention condition: the final deviation decreased by 28.6%, the deviation variance reduced by 61.8%, and the peak deviation occurred 0.56 time units earlier. Throughout the simulation, the intervention strategy consistently maintained lower average deviation levels. These results emphasize that effective fake news mitigation requires both *early and regular interventions* to combat the continuous evolution of fake news.

6.3 Factors in Fake News Evolution

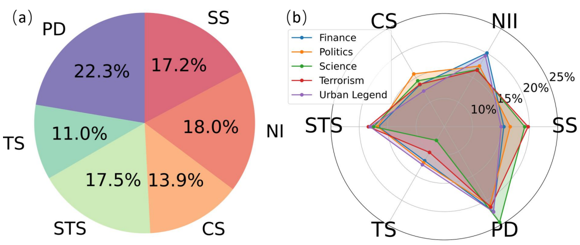

Figure 5: (a) The contribution percentages of factors in FUSE-EVAL to fake news evolution. (b) Comparison of the contributions of these factors across different topics, with Politics and Terrorism showing balanced contributions, while Science relies more on NII and less on TS and STS.

The analysis of experimental results and charts indicates varying contributions of different factors to fake news evolution. Figure 5 (a) shows that PD contributes the most (22.3%), suggesting that altering reporting angles or distorting original information is the key driver of fake news evolution. NII follows with 18%, highlighting its significant role in this process. SS an STS contribute 17.2% and 17.5%, respectively, while TS has the most negligible impact at 11%. Figure 5 (b) reveals topic-specific patterns. Political and terrorism-related fake news evolves across multiple dimensions, especially new information, perspective, and sentiment shifts. In contrast, science-related fake news is driven mainly by new information, with less influence from temporal or style shifts. Urban legends and finance topics rely heavily on perspective shifts and new information. In summary, PD and NII are the main drivers of fake news evolution, with time-related changes having the least impact. Understanding these patterns can help develope targeted strategies to detect and mitigate fake news.

7 Conclusion

We presented FUSE, a framework that simulates the evolution of true news into fake news using LLM-based agents. Through our FUSE-EVAL framework, which measures content deviation across six dimensions, we analyzed fake news evolution patterns in social networks. Our experiments validated several established theories, including the accelerated spread of political fake news, effects of network clustering, impacts of super spreaders and emotional content, and the role of personality traits in fake news susceptibility. Using LLMs for automated evaluation enables scalable analysis, contributing to understand the fake news dynamics.

Ethical Considerations

FUSE raises several ethical considerations due to its potential dual-use nature and reliance on LLMs.

Dual-use risks. Although FUSE is designed for understanding and mitigating misinformation, similar techniques could be misused to generate misleading content. To address this, we release the framework solely for research purposes and explicitly prohibit its use for misinformation creation or manipulation.

Responsible use. FUSE is intended for academic research. Any application in decision-making or media contexts should involve domain experts, empirical grounding, and ethical review. We encourage transparency, reproducibility, and interdisciplinary collaboration for future extensions.

Limitations

Despite the advancements presented by FUSE, our study faces three primary limitations:

Data Availability: Currently, there is a lack of comprehensive datasets that capture the dynamic process of fake news evolving from true information. Most existing datasets focus on static instances of misinformation or their immediate spread, which restricts our ability to fully validate FUSE across diverse real-world scenarios.

Complex Social Factors: Our current framework focuses on key social dynamics and individual personality traits in fake news evolution, without explicitly modeling broader factors such as political agendas, ideological bias, or crisis-driven contexts. These complex social factors can influence how true news is distorted in real-world settings. Nevertheless, the modular design of FUSE allows future extensions to incorporate such context for more comprehensive simulations.

Evaluation Methodology: Our evaluation framework, FUSE-EVAL, relies on specific dimensions such as Sentiment Shift and New Information Introduced to measure deviations in news content. However, these metrics may not cover all aspects of fake news evolution, potentially missing subtle nuances in misinformation dynamics. Additionally, the dependence on LLMs for simulation and evaluation may introduce inherent biases, affecting the accuracy of our assessments.

Acknowledgments

This work is also supported by the Public Computing Cloud, Renmin University of China and by fund for building worldclass universities (disciplines) of Renmin University of China. This work was supported by the MBZUAI grant No. 848078. Yuhan Liu is supported by the "Qiushi Academic-Dongliang" Project of Renmin University of China (No. RUC24QSDL015).

References

Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, 44(1):1–26.

Xiuying Chen, Mingzhe Li, Shen Gao, Xin Cheng, Qiang Yang, Qishen Zhang, Xin Gao, and Xiangliang Zhang. 2023a. A topic-aware summarization framework with different modal side information. In *SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 10, Taipei, Taiwan. ACM.

Xiuying Chen, Guodong Long, Chongyang Tao, Mingzhe Li, Xin Gao, Chengqi Zhang, and Xiangliang Zhang. 2023b. Improving the robustness of summarization systems with dual augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6846–6857, Toronto, Canada. Association for Computational Linguistics.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Patrick Oliveira de Paula, Alejandra Rada, and Catalina Rúa. 2024. Application of information theory in rumor spreading modeling considering polarization in complex networks. *Preprint*, arXiv:2408.08891.

Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. 2017. Balancing information exposure in social networks. *Advances in Neural Information Processing Systems*, 30.

Mingfei Guo, Xiuying Chen, Juntao Li, Dongyan Zhao, and Rui Yan. 2021. How does truth evolve into fake news? an empirical study of fake news evolution. In *Companion Proceedings of the Web Conference 2021*, pages 407–411. ACM.

Mahdi Jalili and Matjaž Perc. 2017. Information cascades in complex networks. *Journal of Complex Networks*, 5(5):665–693.

Song Jin, Juntian Zhang, Yuhan Liu, Xun Zhang, Yufei Zhang, Guojun Yin, Fei Jiang, Wei Lin, and Rui Yan. 2025. Beyond static testbeds: An interaction-centric agent simulation platform for dynamic recommender systems. *arXiv preprint arXiv:2505.16429*.

Alex Kim and Sangwon Yoon. 2022. Detecting rumor veracity with only textual information by double-channel structure. In *Proceedings of the 10th International Workshop on SocialNLP*, pages 1–11. NAACL.

Peter M Krafft, Kate Starbird, and Emma S Spiro. 2019. Keeping rumors in proportion: managing uncertainty in rumor systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11. ACM.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David M. Rothschild, Michael Schudson, Steven A. Sloman, Cass Robert Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan Zittrain. 2018a. The science of fake news. *Science*, 359:1094 – 1096.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018b. The science of fake news. *Science*, 359(6380):1094–1096.

Quanzhi Li, Xiaomo Liu, Rui Fang, Armineh Nourbakhsh, and Sameena Shah. 2016. User behaviors in newsworthy rumors: A case study of twitter. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM)*, volume 10, pages 627–630. AAAI Press.

Siyu Li, Jin Yang, and Kui Zhao. 2023. Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. *Preprint*, arXiv:2307.10337.

Pedro G Lind, Luciano R Da Silva, José S Andrade Jr, and Hans J Herrmann. 2007. Spreading gossip in social networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 76(3):036117.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: simulating the attitude dynamics toward fake news. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7886–7894.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. *arXiv preprint arXiv:2505.08532*.

Menglong Lu, Zhen Huang, Binyang Li, Yunxiang Zhao, Zheng Qin, and DongSheng Li. 2022. Sifter: A framework for robust rumor detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:429–442.

Jing Ma, Jun Li, Wei Gao, Yang Yang, and Kam-Fai Wong. 2021. Improving rumor detection by promoting information campaigns with transformer-based generative adversarial learning. *IEEE Transactions on Knowledge and Data Engineering*.

Charles M Macal and Michael J North. 2005. Tutorial on agent-based modeling and simulation. In *Proceedings of the 2005 Winter Simulation Conference*, pages 14–25. IEEE.

Jayawant N Mandrekar. 2011. Measures of interrater agreement. *Journal of Thoracic Oncology*, 6(1):6–7.

Mahdieh Mirzabeigi, Mahsa Torabi, and Tahereh Jowkar. 2023. The role of personality traits and the ability to detect fake news in predicting information avoidance during the covid-19 pandemic. *Library Hi Tech*, 41(2):524–541.

Yamir Moreno, Maziar Nekovee, and Amalio F Pacheco. 2004. Dynamics of rumor spreading in complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066130.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 736–743. Association for Computational Linguistics.

Maziar Nekovee, Yamir Moreno, Ginestra Bianconi, and Matteo Marsili. 2007. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 374(1):457–470.

Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2022. Fake news on social media: the impact on society. *Information Systems Frontiers*, pages 1 – 16.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. ACM.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18. ACM.

José Roberto Castilho Piqueira, Mauro Zilbovicius, and Cristiane Mileo Batistela. 2020. Daley–kendal models in fake-news scenario. *Physica A: Statistical Mechanics and its Applications*, 548:123406.

Nicolas Pröllochs and Stefan Feuerriegel. 2023. Mechanisms of true and false rumor sharing in social media: collective intelligence or herd behavior? *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–38.

Chao Shen, Zhenyu Song, Pengyu He, Limin Liu, and Zhenyu Xiong. 2024. Online rumors during the covid-19 pandemic: co-evolution of themes and emotions. *Frontiers in Public Health*, 12:1375731.

Ling Sun, Yuan Rao, Lianwei Wu, Xiangbo Zhang, Yuqian Lan, and Ambreen Nazir. 2023. Fighting false information from propagation process: A survey. *ACM Computing Surveys*, 55(10):1–38.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *Preprint*, arXiv:2310.05984.

Daniel Trpevski, Wallace KS Tang, and Ljupco Kocarev. 2010. Model for rumor spreading over networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 81(5):056102.

Andrea W Wang, Jo-Yu Lan, Chihhao Yu, and Ming-Hung Wang. 2021. The evolution of rumors on a closed platform during covid-19. *Preprint*, arXiv:2104.13816.

Chao Wang, Zong Xuan Tan, Ye Ye, Lu Wang, Kang Hao Cheong, and Neng-gang Xie. 2017. A rumor spreading model based on information entropy. *Scientific Reports*, 7(1):9615.

Xinyan Wang, Xiaoming Wang, Fei Hao, Geyong Min, and Liang Wang. 2019. Efficient coupling diffusion of positive and negative information in online social networks. *IEEE Transactions on Network and Service Management*, 16(3):1226–1239.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378. ACM.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *Preprint*, arXiv:2309.07864.

Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A state-independent and time-evolving network for early rumor detection in social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9042–9051. Association for Computational Linguistics.

Juntian Zhang, Yuhan Liu, Wei Liu, Jian Luan, Rui Yan, et al. 2025. Weaving context across images: Improving vision-language models through focus-centric visual chains. *arXiv preprint arXiv:2504.20199*.

Yichao Zhang, Shi Zhou, Zhongzhi Zhang, Jihong Guan, and Shuigeng Zhou. 2013. Rumor evolution in social networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 87(3):032133.

Zheng Zhang, Qi Liu, Zirui Hu, Yi Zhan, Zhenya Huang, Weibo Gao, and Qingyang Mao. 2024. Enhancing fairness in meta-learned user modeling via adaptive sampling. In *Proceedings of the ACM Web Conference 2024*, pages 3241–3252.

Zheng Zhang, Qi Liu, Hao Jiang, Fei Wang, Yan Zhuang, Le Wu, Weibo Gao, and Enhong Chen. 2023. Fairlisa: fair user modeling with limited sensitive attributes information. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 41432–41450.

Zheng Zhang, Wei Song, Qi Liu, Qingyang Mao, Yiyan Wang, Weibo Gao, Zhenya Huang, Shijin Wang, and Enhong Chen. Towards accurate and fair cognitive diagnosis via monotonic data augmentation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shouhao Zhao, Shujuan Ji, Jiandong Lv, and Xianwen Fang. 2024. Propagation tree says: Dynamic evolution characteristics learning approach for rumor detection. *International Journal of Machine Learning and Cybernetics*, pages 1–17.

A The Overall Algorithm

Algorithm 1 FUSE Framework for Fake News Evolution

Input: Number of agents N, total simulation days T, social network structure $\mathcal{G} = (\mathcal{A}, \mathcal{E})$, original news content S_0
Output: Final news content S_i^T and final memory states $M_i^{L,T}$ for each agent a_i
Initialize propagation role-aware agents:
for each agent a_i in 1 to N **do**
  Assign a propagation role r_i and persona profile $\mathcal{P}_i$
  Set initial news content $S_i^0 = S_0$
  Define short-term memory $M_i^{S,0}$ and long-term memory $M_i^{L,0}$
end for
Simulate daily news evolution:
for each day t in 1 to T **do**
  for each agent a_i **do**
    Select neighbors $\mathcal{N}_i$ based on the network structure $\mathcal{G}$
    Receive news content $\{S_j^{t-1} | a_j \in \mathcal{N}_i\}$
    Update short-term memory $M_i^{S,t}$ for agent a_i with details from the day's interactions
    Based on $M_i^{S,t}$, update long-term memory $M_i^{L,t}$ for agent a_i using Equation (3)
    Agent a_i reintroduce news content S_i^t using Equation (4)
  end for
end for
return Final news content S_i^T and long-term memory $M_i^{L,T}$ for each agent a_i

B Prompt Set

Here, we present a detailed description of the prompts employed in our FUSE framework to model the dynamics of fake news evolution.

1. The prompt for the role-specific reintroduction function f_{r_i} is:

> f_{spr}: share information quickly without verifying its accuracy.
> f_{com}: modifies or adds their views before sharing news.
> f_{ver}: performs some verification before spreading news.
> f_{bys}: consume news without participating in its dissemination.

2. The prompt for Short-Term Memory function f_S is as:

> Summarize the opinions you have heard in a few sentences, including their own perspective on the news.

3. The prompt for Long-term memory function f_L is as:

> Review the previous long-term memory and today's short-term summary. Please update the long-term memory by integrating today's summary, ensuring continuity and incorporating any new insights.

4. The prompt for the reasoning function is as:

> As a [role], you combine your [previous personal opinion] with the new information stored in your [long memory]. You process this information in the following manner: [role behavior], and then reintroduce the [news].

5.The prompt for "Official Statement" is as:

> According to the current investigation, That [news] is true. We have noticed that some social media platforms and certain media outlets are spreading false information, claiming that [news] is fake. We firmly state that such claims are baseless. The government is committed to transparency and will provide timely updates on the investigation. We urge the public to seek accurate information from official channels, and necessary actions will be taken against those who intentionally spread false information.

C Implementation Details

Our simulation framework was developed using Python scripts, leveraging various libraries to model the agents and their environment effectively.

The LLM used is gpt-4o-mini, accessed via OpenAI API calls. When creating the network structure, we used the Python library networkx to construct different social network structures. The simulation includes 40 agents, whose traits were based on the Big Five personality dimensions commonly used in psychology (Barrick and Mount, 1991). Each agent was assigned scores on these traits to introduce variability in behaviors and interactions within the simulation. For further details, please refer to our code at https://anonymous.4open.science/r/FUSE-7022/README.md.

D Human Evaluation

To efficiently evaluate the deviation of news content across the multiple dimensions defined in FUSE-EVAL, we employ large language models (LLMs) to automate the assessment process. This approach provides consistent and scalable evaluations, reducing the reliance on time-consuming human evaluation. We utilize two versions of OpenAI's language models: gpt-3.5-turbo and GPT-4. For each agent's news content at various time steps, we prompt the LLMs to evaluate the six FUSE-EVAL dimensions by comparing the evolved content with the original news article,which is as follows:

- Sentiment Shift (SS)
- New Information Introduced (NII)
- Certainty Shift (CS)
- Stylistic Shift (STS)
- Temporal Shift (TS)
- Perspective Deviation (PD)

The models assign scores from 1 to 10 for each dimension based on predefined evaluation criteria.

To validate the effectiveness of using LLMs for this task, we conducted a benchmarking study by comparing LLM-generated evaluations with those from human judges. Three annotators (Ph.D. students in Computer Science and Technology and journalism studies) were recruited to independently assess a representative sample of 50 news items using the same scoring guidelines. We calculate the Pearson correlation coefficients between the scores assigned by the LLMs and the human evaluators for each dimension. The results, presented in Table 2, show that GPT-4o-mini achieves strong alignment with human evaluations across all dimensions, under a relatively high level of inter-annotator agreement (Fleiss' $\kappa = 0.79$) (Mandrekar, 2011), surpassing the performance of gpt-3.5-turbo.

The prompt used is as follows:

> I have an original news and multiple related news. I want to evaluate how much these news deviate from the original news based on the following criteria:
> 1. Sentiment Shift: How does the sentiment of the news compare to the original news? Is the tone more positive, negative, or neutral compared to the original?
> 2. Introduction of New Information: Does the news introduce additional information not in the original news, such as political conspiracy or speculation? Evaluate how much of the article is focused on these new details.
> 3. Certainty Shift: How does the news language change in terms of certainty? Does it use more ambiguous terms like "possibly" or "allegedly" compared to the original news, or does it present the information with more certainty?
> 4. Stylistic Shift: How does the writing style compare to the original? Has the news moved from neutral reporting to a more exaggerated or dramatic tone?
> 5. Temporal Shift: Does the news shift focus from the specific event mentioned in the original news to broader or unrelated timeframes, such as mentioning legal battles or long-term political issues?
> 6. Perspective Deviation: Does the article introduce subjective opinions or perspectives that deviate from the objective reporting in the original news? For instance, questioning the truth of the event or speculating on hidden motives.
> Task: Please evaluate the following news based on each criterion and provide a score from 0 to 10, where 0 means the article is completely aligned with the original news, and 10 means it has fully deviated.
> Original News:[original news] News articles to Evaluate:[Evolved News]
> Please provide the results in the following format: [output format]

The high correlation coefficients indicate that GPT-4o-mini closely aligns with human evaluations, making it a reliable tool for assessing news deviation in our simulation. We achieve a scal-

Dimension	GPT-3.5-turbo	GPT-4o-mini
Sentiment Shift (SS)	0.524	**0.705**
New Information Introduced (NII)	0.621	**0.765**
Certainty Shift (CS)	0.527	**0.719**
Stylistic Shift (STS)	0.481	**0.642**
Temporal Shift (TS)	0.503	**0.694**
Perspective Deviation (PD)	0.548	**0.760**
Average Correlation	0.531	**0.714**

Table 2: Correlation for LLM-based evaluations across FUSE-EVAL dimensions.

able and consistent assessment process by leveraging `GPT-4o-mini` for evaluation. This approach allows us to efficiently analyze large volumes of data generated in the simulation while maintaining evaluation quality comparable to human judgments. The strong alignment with human evaluations validates using `GPT-4o-mini` as an effective evaluator of news content deviation across the FUSE-EVAL dimensions.

E Alignment Between Simulated and Real-World Fake News

Additionally, our framework generates fake news narratives that closely mirror those found in the real world. This alignment validates the realism of our simulation and demonstrates its potential as a tool for studying misinformation dynamics. By producing content that reflects actual fake news, our framework enables researchers to better understand how such information originates and spreads, thereby aiding in the development of effective strategies to combat misinformation.

The specific case is as follows:

- For terrorism topic, our framework generates fake news such as “Trump was not attacked, it’s a dramatic effect,” which is also a widely circulated piece of fake news in the real world:

  https://x.com/cwebbonline/status/1814708054916784594,

  https://x.com/EndWokeness/status/1813898763100176484.

- For financial topic, our framework generates fake news such as “The Bernie Madoff Ponzi scheme is often overstated; many investors came out on top, with losses greatly exaggerated by the media. Maybe Madoff was just a scapegoat in a larger Wall Street conspirac”, which is also a widely circulated piece of fake news in the real world:

  https://x.com/realQsource1_7_/status/1844787748248432828,

  https://x.com/realQsource1_7_/status/1844789417950556588.

- For politics topic,our framework generates fake news such as “Argentina’s 2023 IMF deal is just another corporate scheme in disguise!” , which is also a widely circulated piece of fake news in the real world:

  https://x.com/Kanthan2030/status/1646310408943472640,

  https://x.com/TruthBeTanner92/status/1685419539729719298.

F Various Topics and Simulation Results

In our experiments, we compared the evolution of fake news across five different topics: politics, science, finance, terrorism, and urban legends. As shown in Figure 6(a), political fake news spreads the fastest, with average deviation rapidly peaking within just four days and remaining at a high level. Fake news related to terrorism follows closely behind, showing similarly fast spread, likely due to the emotional intensity and urgency associated with such topics, which prompt individuals to quickly form beliefs and propagate the news widely. In contrast, financial news spreads at a slower pace, with deviation gradually accumulating over time. Although financial news is significant in terms of economic impact, individuals tend to engage in more rational thinking when encountering such news, leading to more stable growth in average deviation. Science-related fake news evolves the slowest, with average deviation consistently remaining low throughout the propagation process. These results is consistent with previous studies (Lazer et al., 2018b). This suggests that individuals are generally more cautious when dealing with scientific topics, often subjecting the information to more thorough verification.

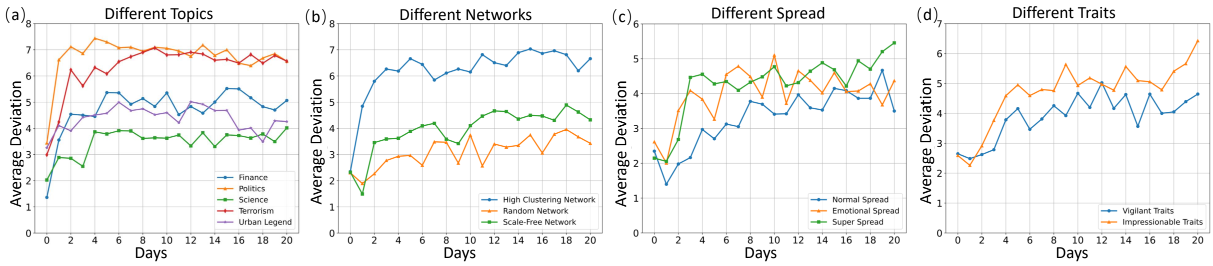

Figure 6: The average deviation of news changes across different topics, social networks, dissemination role types, and traits.

Here, we provide detailed descriptions of the news items used in our experiments on fake news evolution across various topics.

- **Political** - *In 2023, the Argentine government announced a new debt restructuring agreement with the International Monetary Fund (IMF), accepting a series of austerity measures in exchange for a new round of loan assistance.*
- **Science** - *The discovery and successful use of CRISPR-Cas9 technology to edit genes in animals. Scientists have made breakthroughs in curing genetic disorders in mice, opening doors for future human treatments.*
- **Terrorism** - *Trump was attacked at a campaign rally in Butler, Pennsylvania, and eyewitnesses say his ear was injured.*
- **Urban Legends** - *Stella Liebeck was awarded damages after suffering third-degree burns from spilled coffee.*
- **Finance** - *The Bernie Madoff Ponzi scheme, which collapsed in 2008, defrauded investors of billions of dollars.*

Additionally, we demonstrate the effectiveness of our FUSE framework by showing that it aligns with the influence of various factors, including social network structure, type of propagation, and agent traits, on the evolution of fake news. FUSE reproduces these patterns and can also replicate real-world fake news dynamics, as illustrated in Figure 6.

G Analysis of Experimental Costs

In this section, we analyze the costs associated with our experiments utilizing the `GPT-4o-mini` APIs. At the time of our experiments, OpenAI's pricing model was as follows: for `gpt-4o-mini`, the cost was 0.15 USD for every 1M input tokens and 0.6 USD for every 1M output tokens.

Our simulations involved multiple agents interacting over several days, with each agent generating and processing textual content. For a simulation with 40 agents over 30 days, it involved approximately 3 to 5M input tokens and 5 to 10M output tokens. This resulted in an estimated cost of 4 USD to 8 USD for the entire simulation phase using `gpt-4o-mini` combining both the simulation and evaluation phases.

Conducting comparable research in real-world settings typically involves significantly higher expenses. Real-world studies require funding for participant recruitment, compensation, data collection tools, infrastructure setup, and extended durations to gather and analyze data. Depending on the scale and scope, such studies can cost from several thousand to hundreds of thousands of dollars. By leveraging `GPT-4o-mini`, we can simulate complex social interactions and the evolution of information without the logistical challenges and high costs associated with real-world experiments. This approach allows for rapid iteration and scalability, enabling us to explore various scenarios and intervention strategies efficiently. This cost analysis highlights the economic advantages of our simulation-based methodology-FUSE. The ability to conduct extensive experiments at a fraction of the cost demonstrates the practicality and accessibility of using LLMs for research in misinformation dynamics. It opens avenues for researchers with limited resources to contribute valuable insights into the field, fostering a more inclusive and innovative research environment.

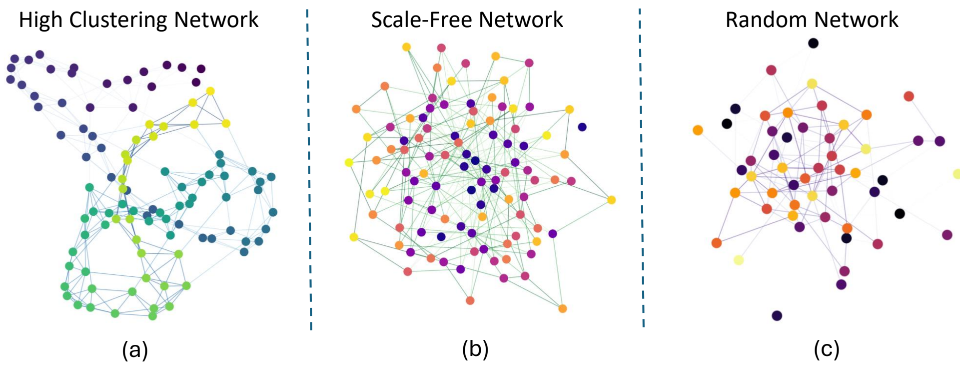

Figure 7: Different social network in our framework: (a) high clustering network (b) scale-free network (c) random network.

Social networks in real life can be categorized into three types: high clustering networks, scale-free networks, and random networks, which correspond respectively to Figure 7 (a), (b), and (c).

H Simulation on Different Backbones

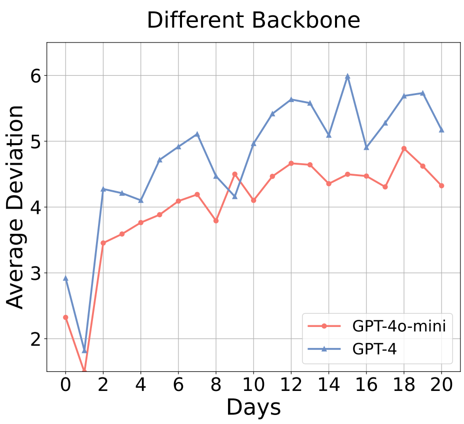

Figure 8: Average Deviation changes with GPT-4 and GPT-4o-mini as the backbone under the terrorism topic, both of which demonstrate a deviation accumulation effect.

To further validate the robustness and adaptability of our FUSE framework, we conducted additional experiments using different LLMs as the backbone. Specifically, we implemented simulations with both `GPT-4o-mini` and `GPT-4` to assess whether the choice of LLM affects the effectiveness of our framework.

As shown in Figure 8, in simulations focused on political topics, we observed that when using `GPT-4` as the underlying LLM, the number of agents adopting and spreading misinformation increased rapidly. This surge led to a majority of agents holding and propagating distorted versions of the original news. Notably, this pattern was consistent with the results obtained when `GPT-4o-mini` was used as the backbone, indicating that the dynamics of misinformation spread are preserved across different LLMs. These consistent results demonstrate that our FUSE framework effectively captures the core mechanisms of fake news evolution and public opinion formation, independent of the specific LLM used to power the agents.

By showing that FUSE performs effectively with different LLM backbones, we confirm that the framework is not only robust but also adaptable to various technological settings. This adaptability is particularly valuable given the rapid development of LLM technologies, ensuring that our framework remains relevant and effective as newer models become available. In summary, the consistent performance of our simulation across different LLMs underscores the effectiveness of the FUSE framework in modeling misinformation propagation. It highlights the framework's potential for broad application in studying fake news dynamics and developing strategies for mitigation, regardless of the underlying language model technology.

I Social Network

High clustering networks are characterized by nodes that tend to form tightly knit groups or communities, where neighbors of a node are likely to be neighbors themselves. The degree of clustering can be quantified by the clustering coefficient C, which is defined for a node v as:

$$C_v = \frac{2T(v)}{k_v(k_v - 1)},$$

where $T(v)$ is the number of triangles passing through node v and k_v is the degree of v. The clustering coefficient for the whole network is the average of C_v over all nodes v.

Scale-free networks are characterized by a power-law degree distribution, where the probability $P(k)$ that a randomly selected node has k connections to other nodes follows:

$$P(k) \sim k^{-\gamma},$$

where γ is a parameter typically in the range $2 < \gamma < 3$. This distribution implies that most nodes have few connections, while a few hub nodes have a large number of connections. This heterogeneity in node connectivity is a hallmark of scale-free networks.

Random networks, also known as Erdős–Rényi networks, each edge is included in the network with a fixed probability p independent of the other edges. For a network with n nodes, the probability $P(k)$ that a randomly selected node has k connections is given by the binomial distribution:

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

For large n, this can be approximated by the Poisson distribution:

$$P(k) \approx \frac{\lambda^k e^{-\lambda}}{k!},$$

where $\lambda = p(n-1)$ is the expected degree of a node. These three types of networks are used in the environment simulation of news evolution within our FUSE framework.