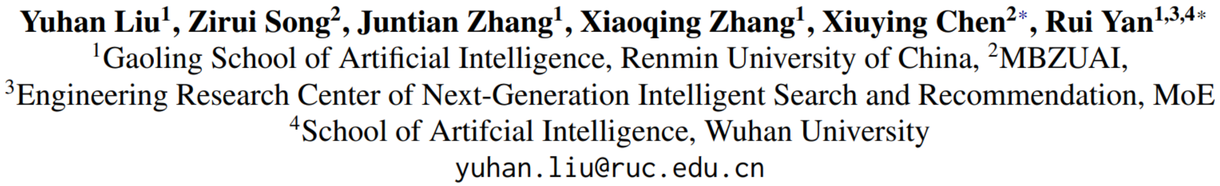
HUMANE Paper Review

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[1,3,4*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, [2]MBZUAI,
[3]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MoE
[4]School of Artifcial Intelligence, Wuhan University
yuhan.liu@ruc.edu.cn

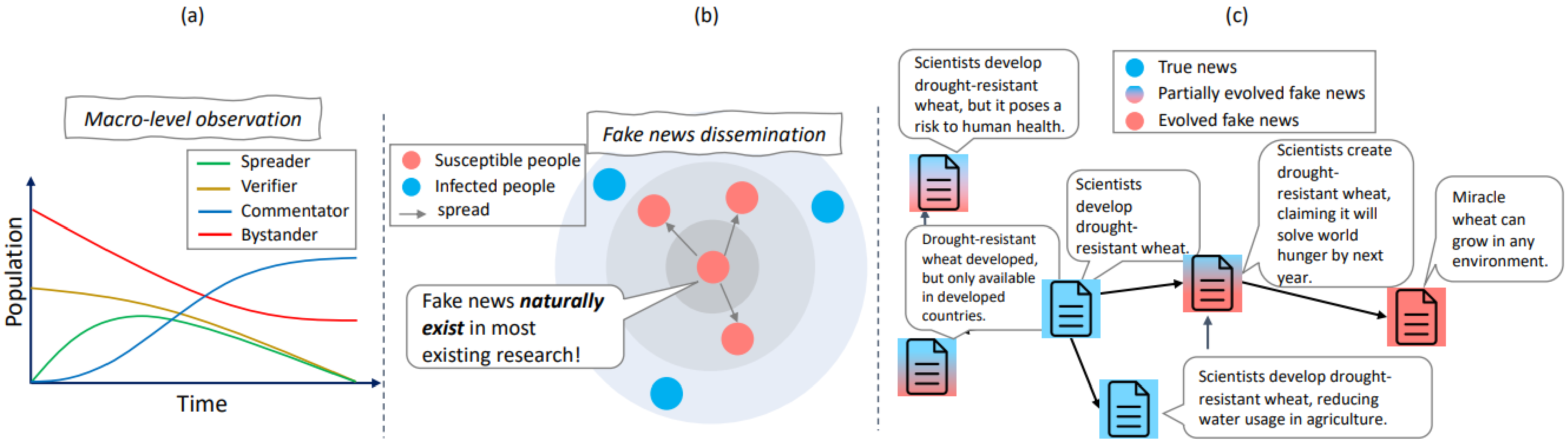
숭실대학교 문화콘텐츠학과, 석사과정생 이다현

EMNLP 2025

2025.10.16

Research Gap

- Prior research addresses misinformation after its initial appearance
  - Assume fake news as existing entities
  - Ignore how misinformation originates or evolves over time
- **In contrast, fake news may originate from true news**



Fake news ***naturally exist*** in most existing research!

Scientists develop drought-resistant wheat, but it poses a risk to human health.

Drought-resistant wheat developed, but only available in developed countries.

Scientists develop drought-resistant wheat.

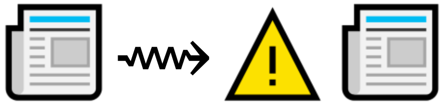
Scientists create drought-resistant wheat, claiming it will solve world hunger by next year.

Miracle wheat can grow in any environment.

Scientists develop drought-resistant wheat, reducing water usage in agriculture.

Research Gap

- Prior research addresses misinformation after its initial appearance
  - Assume fake news as existing entities
  - Ignore how misinformation originates or evolves over time
- **In contrast, fake news may originate from true news**



This work focuses on **how** *facts* gradually become *misinformation* during dissemination

Contributions

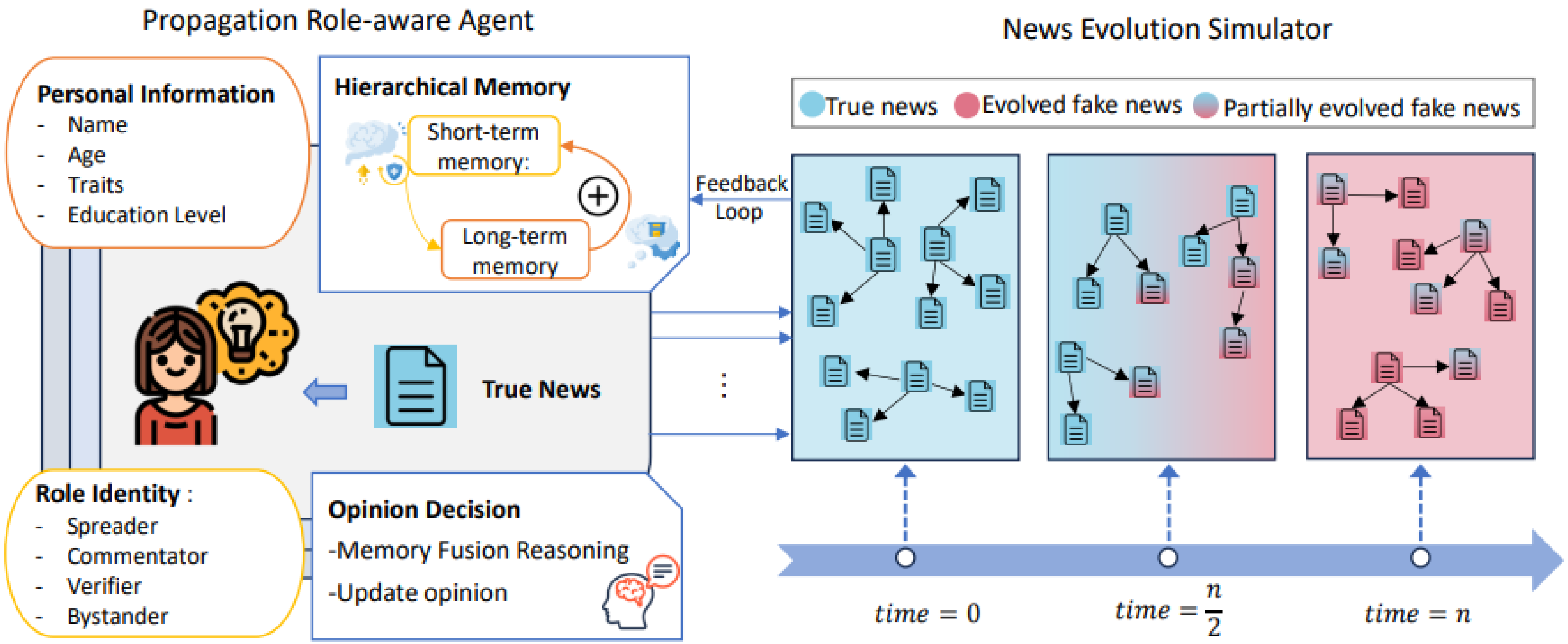
1. **Fake news evolUtion Simulation framEwork** (FUSE)
    - Employs LLM agents to simulate how *real news* becomes *fake news*
    - Four distinct agent roles

1. **FUSE-EVAL**
    - Quantifies the deviation of evolved news from its original
    - Use multiple dimensions

1. **Practical Insights**

FUSE Framework for Fake News Evolution

- Each agent with role-based decision-making capabilities



Propagation Role-aware Agent

✓ Personal Information

- Role r_i in fake news propagation *Sun et al. (2023)*
  - **Spreaders**: propagate information
  - **Commentators**: provide opinions and interpretations
  - **Verifiers**: check information
  - **Bystanders**: passively observe without engaging
- Personal Profile P_i
  - Demographic attributes
  - Personal traits based on the *Big five model*

Propagation Role-aware Agent

✓ Role-Specific Behaviors

- Each day $(t = 1, 2, \ldots, T)$, agents **interact** with their neighboring agents N_i
- How agents **reintroduce** news based on their role and persona

$$f_{role} = f_{r_i}\left(\underbrace{\boxed{S_i^{t-1}}}_{\text{Agent's news at t-1}}, \underbrace{\boxed{\{\, S_j^{t-1} \mid a_j \in N_i\}}}_{\text{Neighbor's news at t-1}}, \underbrace{\boxed{P^i}}_{\text{Agent's personal profile}}\right)$$

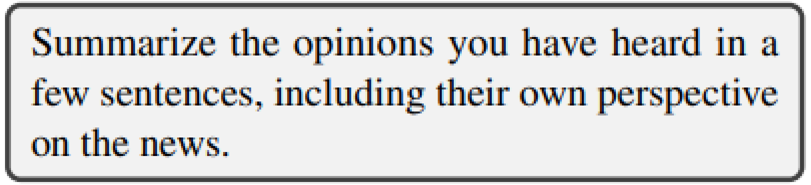
Propagation Role-aware Agent

✓ Memory and Reflection

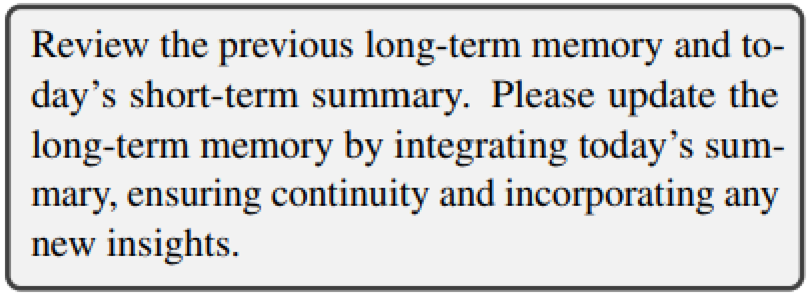
- Each Agent’s Short-term memory M_i^S and Long-term Memory M_i^L

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

Integrates new information to LTM



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



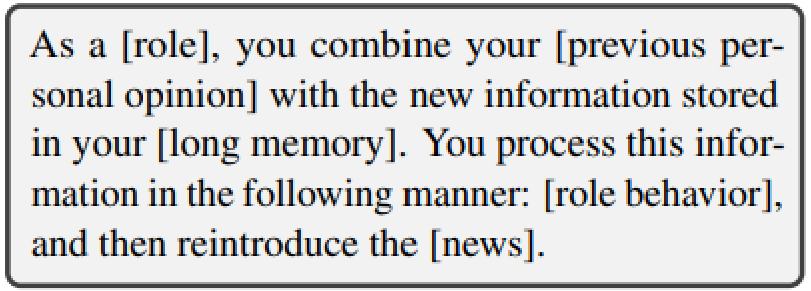
Review the previous long-term memory and today’s short-term summary. Please update the long-term memory by integrating today’s summary, ensuring continuity and incorporating any new insights.

Propagation Role-aware Agent

✓ Decision-Making Process

- How each agent's opinion evolves through a reasoning process
- Agents reflect on their news content everyday (time-step)

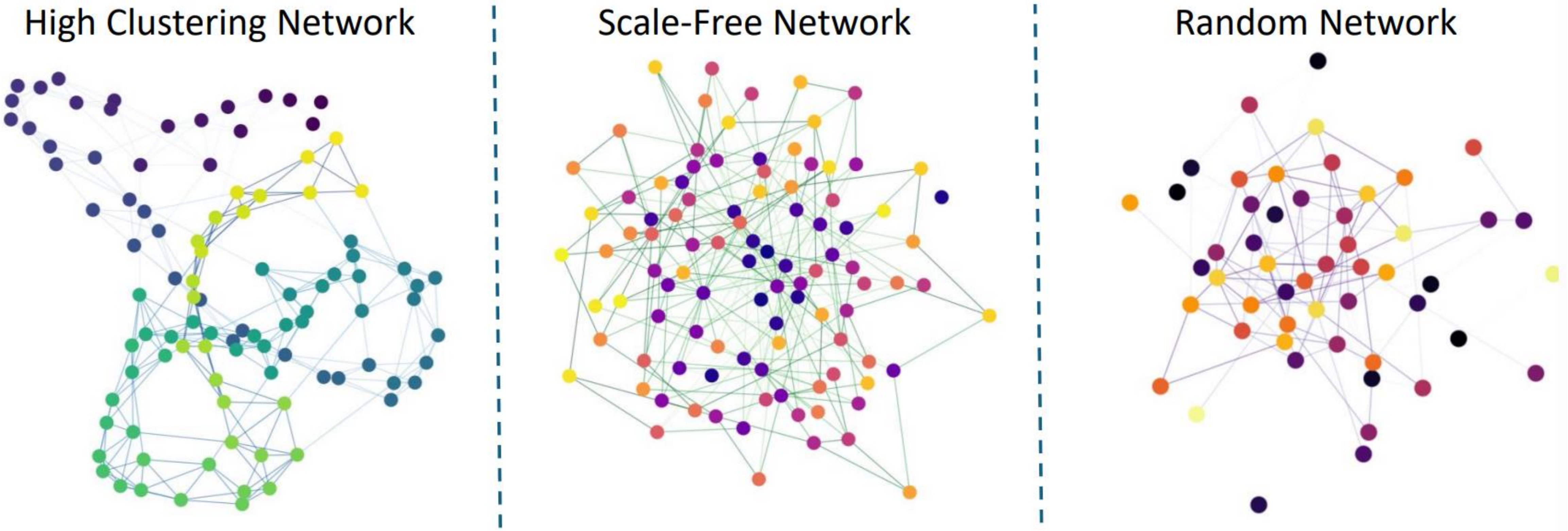
$$S_i^t = f_{dm}(S_i^{t-1}, m_i^{L,t-1}, r_i, P_i)$$



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].

News Evolution Simulator

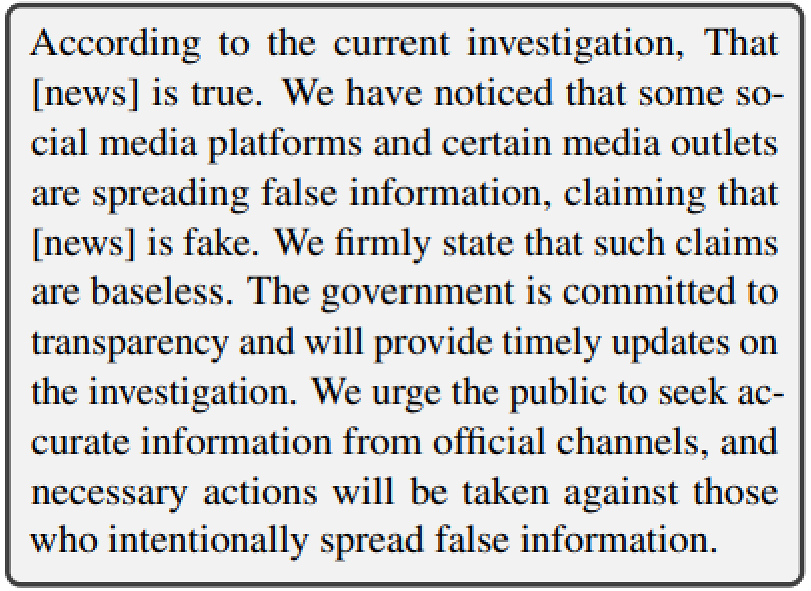
- Provides environment where news evolves over time
  - Social network structure $G = (A, E)$



News Evolution Simulator

✓ Intervention Mechanisms

- To simulate interventions to <u>**counter**</u> fake news evolution
- Works when the deviation between S_0 and S_i^t exceeds a **threshold**
- Official agent issues official announcements



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.

FUSE-EVAL

✓ Content Deviation Metrics

- Assess the **deviation** across multiple dimensions (1 to 10)
  - Sentiment
  - Information
  - Certainty
  - Style
  - Time reference
  - Paraphrasing

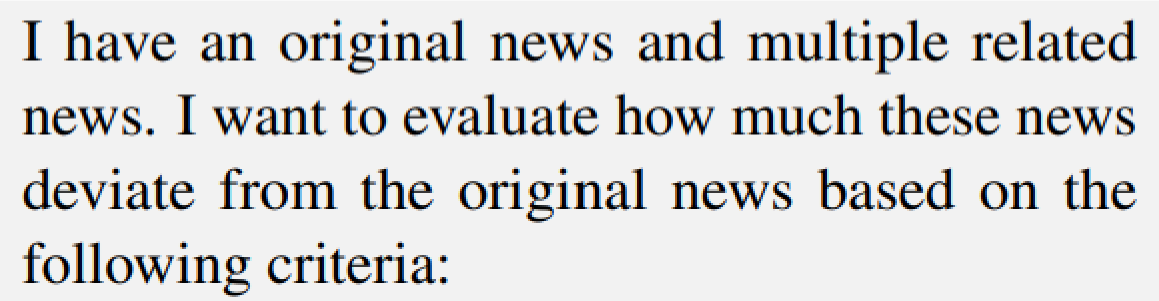
$$\boxed{T_i^t = \frac{1}{6}\sum_{d=1}^{6} D_{i,d}^t}$$

Total Deviation (TD)

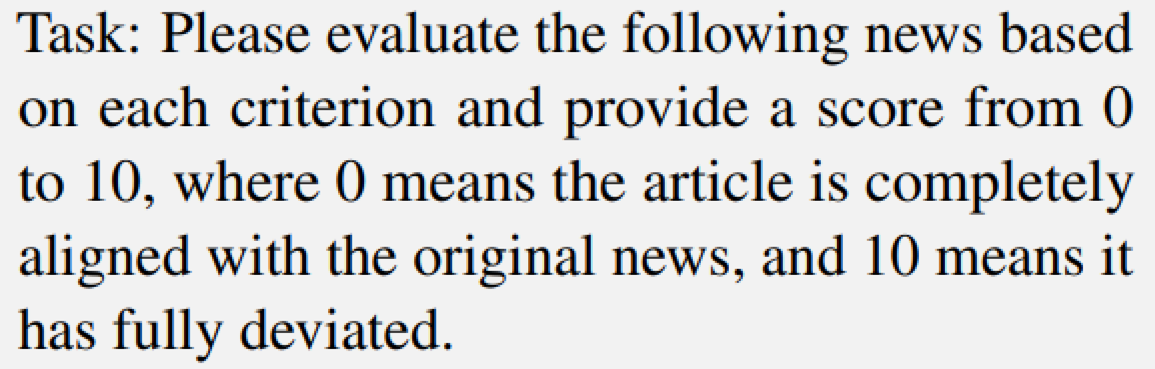
FUSE-EVAL

✓ Content Deviation Metrics

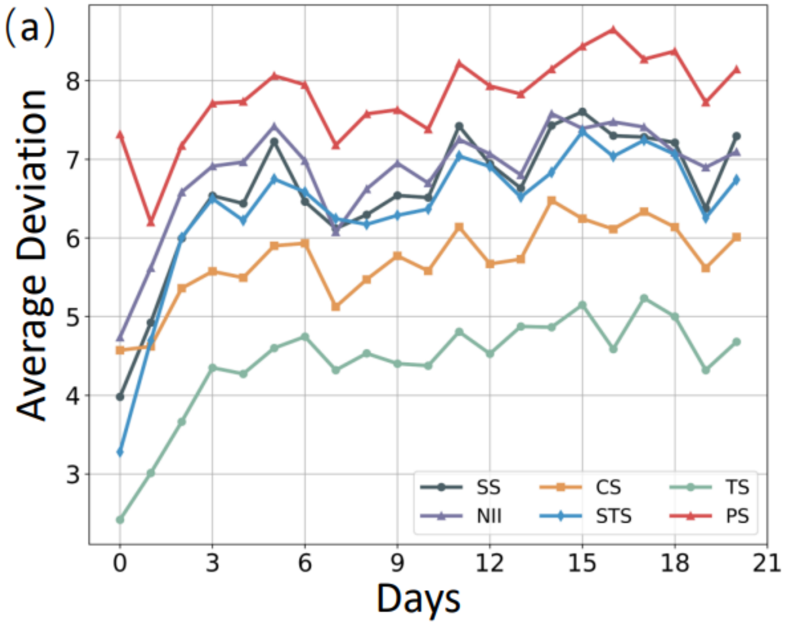
- Assess the **deviation** across multiple dimensions (1 to 10)



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:



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.



FUSE-EVAL

✓ Statistical Deviation Metrics

- Provide insights into the overall patterns of news evolution
- Key metrics
  - Δ Deviation
  - Average Deviation
  - Deviation Variance
  - Final Deviation
  - Maximum and Minimum Deviation
  - Peak Deviation Time
  - Half deviation Time

➡

✓ How much

✓ How evenly across agents

✓ How quickly

Validation of FUSE Framework

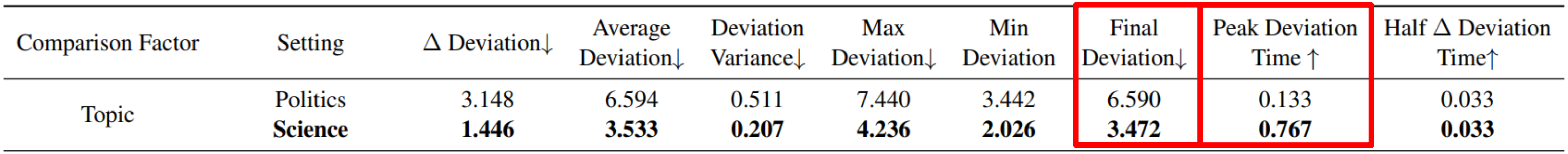
- **How well does FUSE align with real-world patterns?**
  - Topic
  - Social Network structure
  - Spread Type
  - Personality Traits
- **To what extent can FUSE reproduce real-world fake news?**
  - Analyze how generated content is similar with reality
  - Across various topics

❖ GPT-4o-mini, 40 agents

Validation of FUSE Framework

✓ Topic comparison

- **Five topics:** politics, science, finance, terrorism, and urban legends
- Political fake news spread **fastest**
- Final deviation for political news is **90%** higher than science news



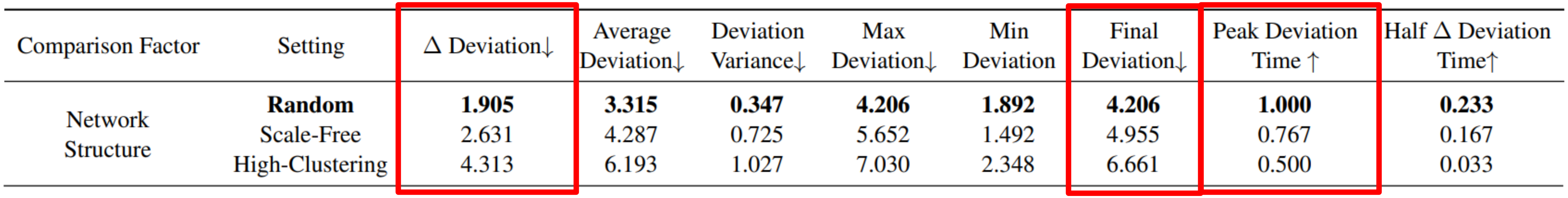
Comparison Factor	Setting	Δ Deviation↓	Average Deviation↓	Deviation Variance↓	Max Deviation↓	Min Deviation	Final Deviation↓	Peak Deviation Time ↑	Half Δ Deviation Time↑
Topic	Politics	3.148	6.594	0.511	7.440	3.442	6.590	0.133	0.033
	Science	**1.446**	**3.533**	**0.207**	**4.236**	**2.026**	**3.472**	**0.767**	**0.033**

political fake news is more prone to rapid distortion and widespread belief

Validation of FUSE Framework

✓ Social Network Comparison

- High-clustering networks: **fastest** and most **extensive** fake news spread



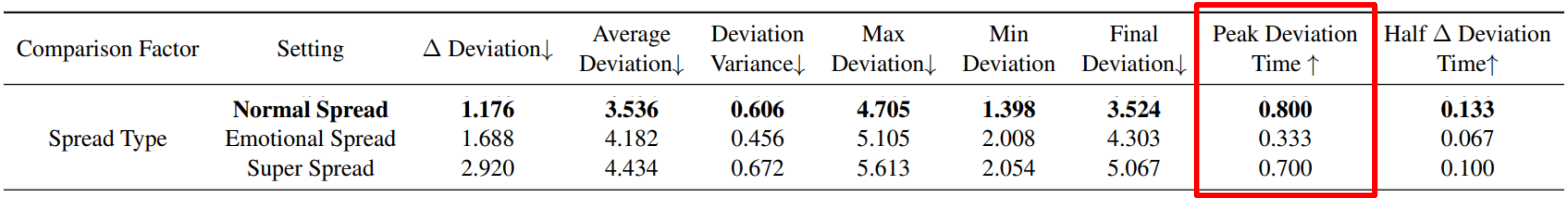
Comparison Factor	Setting	Δ Deviation↓	Average Deviation↓	Deviation Variance↓	Max Deviation↓	Min Deviation	Final Deviation↓	Peak Deviation Time ↑	Half Δ Deviation Time↑
Network Structure	**Random**	**1.905**	**3.315**	**0.347**	**4.206**	**1.892**	**4.206**	**1.000**	**0.233**
	Scale-Free	2.631	4.287	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

Echo chamber →

Validation of FUSE Framework

✓ Spread Type Comparison

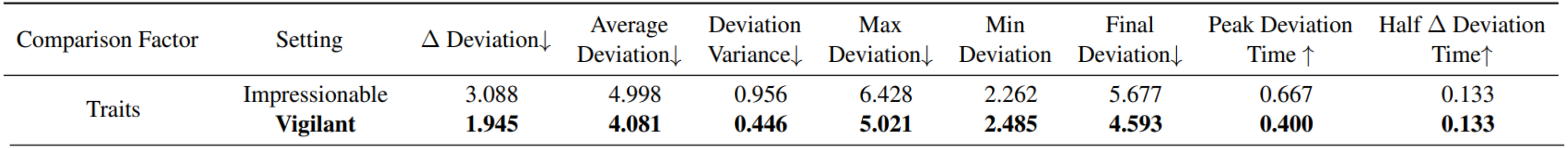
- Super spread: the highest misinformation level



Comparison Factor	Setting	Δ Deviation↓	Average Deviation↓	Deviation Variance↓	Max Deviation↓	Min Deviation	Final Deviation↓	Peak Deviation Time ↑	Half Δ Deviation Time↑
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

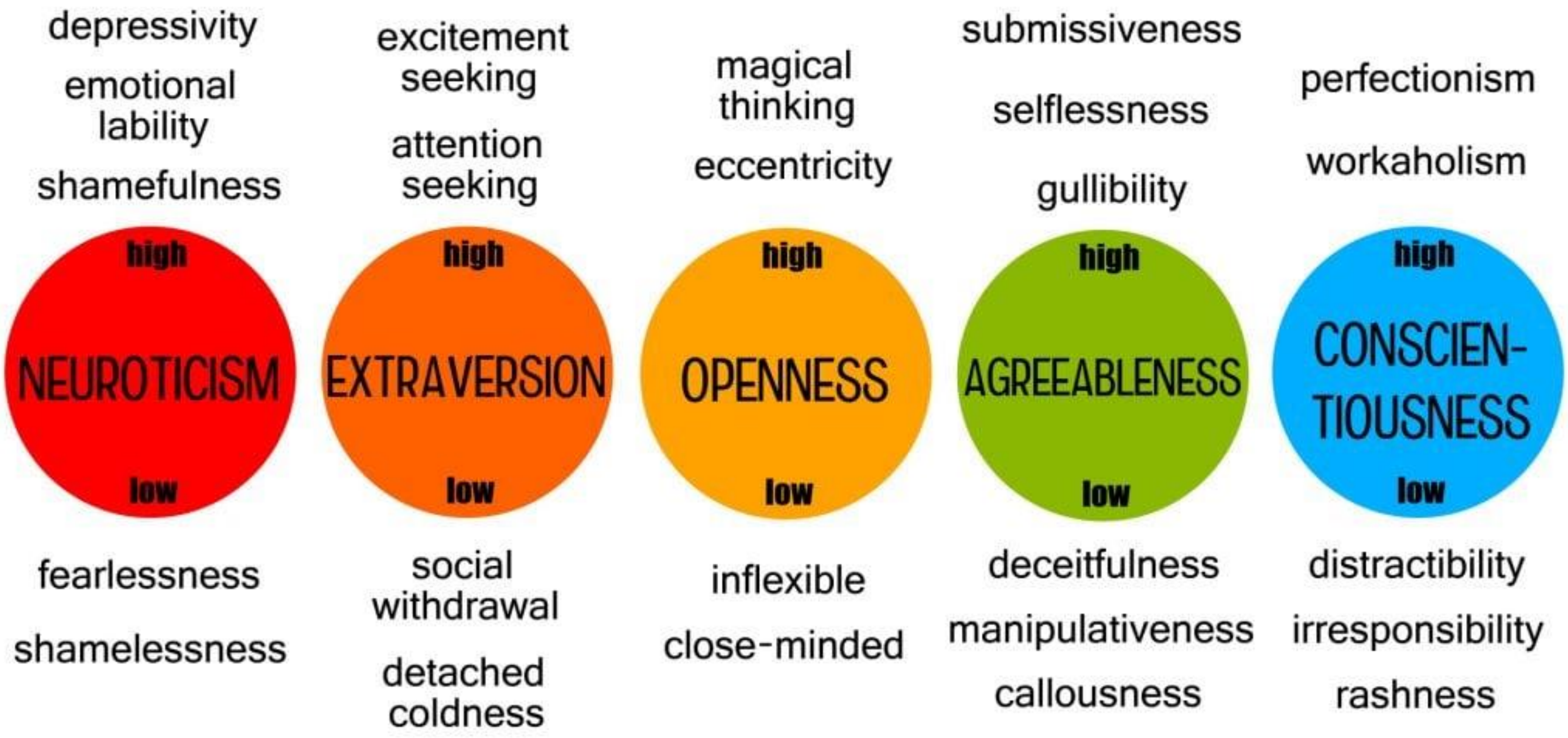
✓ Personality Traits Comparison

- Impressionable agents are more **prone** to accepting and spreading misinformation
- Vigilant agents maintain more stable beliefs



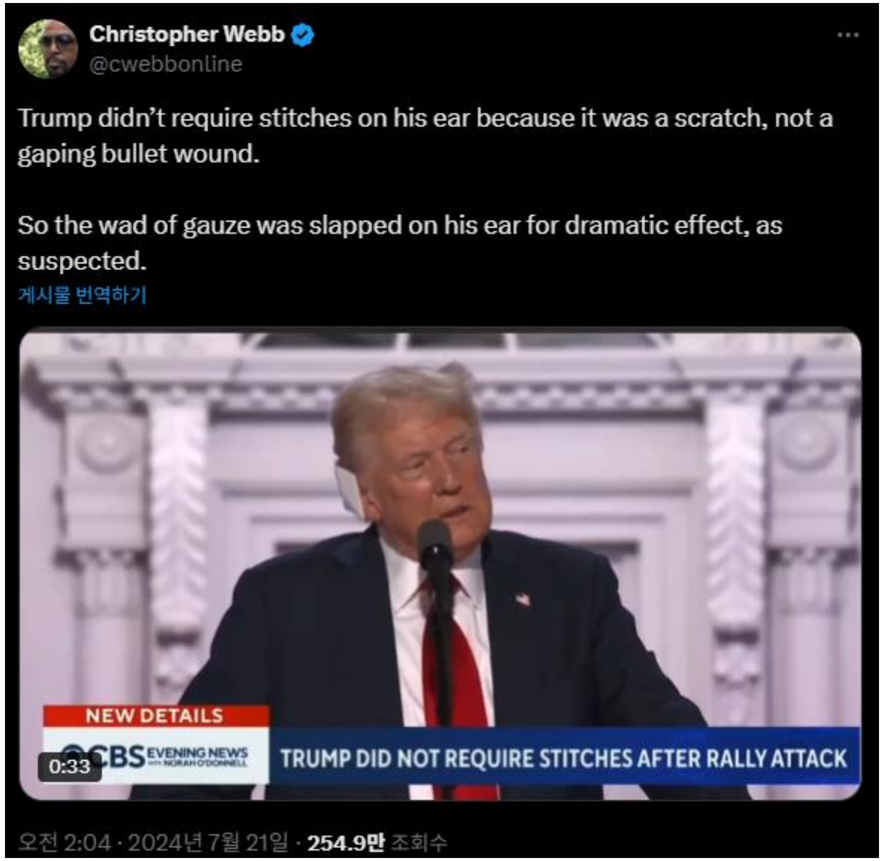
Comparison Factor	Setting	Δ Deviation↓	Average Deviation↓	Deviation Variance↓	Max Deviation↓	Min Deviation	Final Deviation↓	Peak Deviation Time ↑	Half Δ Deviation Time↑
Traits	Impressionable	3.088	4.998	0.956	6.428	2.262	5.677	0.667	0.133
	Vigilant	**1.945**	**4.081**	**0.446**	**5.021**	**2.485**	**4.593**	**0.400**	**0.133**

Big Five personality dimensions



Validation of FUSE Framework

✓ **To what extent can FUSE reproduce real-world fake news?**



Christopher Webb
@cwebbonline

Trump didn't require stitches on his ear because it was a scratch, not a gaping bullet wound.

So the wad of gauze was slapped on his ear for dramatic effect, as suspected.

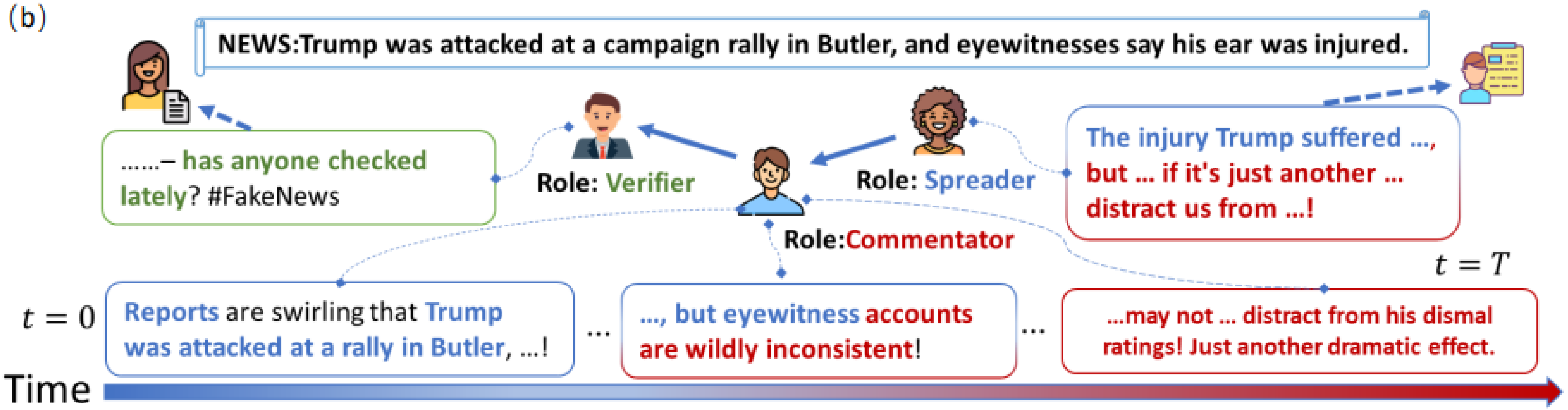
게시물 번역하기

NEW DETAILS
CBS EVENING NEWS WITH NORAH O'DONNELL
TRUMP DID NOT REQUIRE STITCHES AFTER RALLY ATTACK
0:33

오전 2:04 · 2024년 7월 21일 · **254.9만** 조회수

Validation of FUSE Framework

✓ To what extent can FUSE reproduce real-world fake news?



(b)
NEWS:Trump was attacked at a campaign rally in Butler, and eyewitnesses say his ear was injured.
……– has anyone checked lately? #FakeNews
Role: Verifier
Role: Spreader
Role:Commentator
The injury Trump suffered …, but … if it's just another … distract us from …!
$t = T$
$t = 0$
Reports are swirling that Trump was attacked at a rally in Butler, …!
…
…, but eyewitness accounts are wildly inconsistent!
…
…may not … distract from his dismal ratings! Just another dramatic effect.
Time

- For each topic, 73% of fake news was recovered

Ablation Study

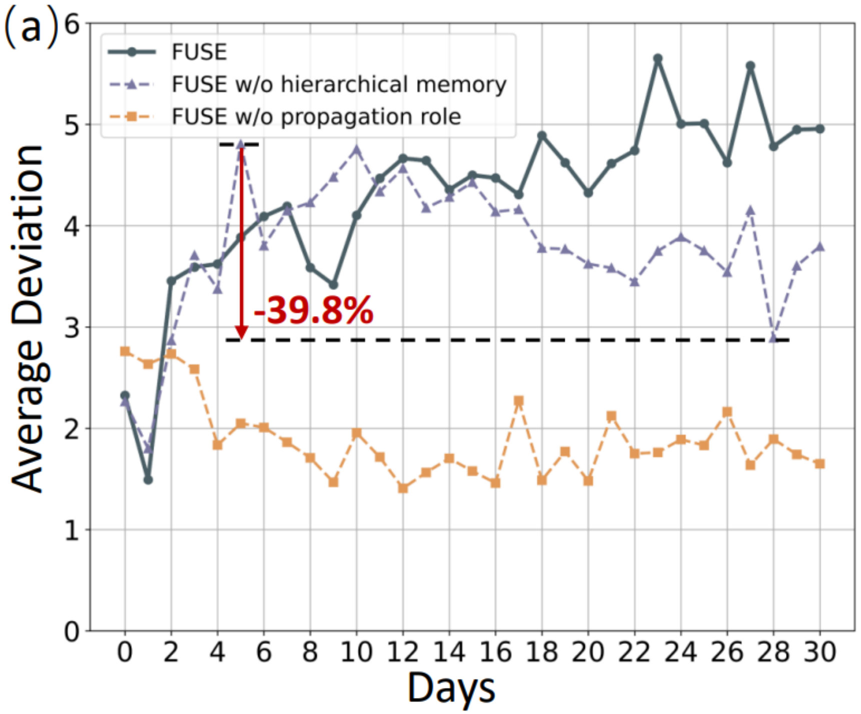
✓ The Impact of Hierarchical Memory and Propagation-Role

① Hierarchical Memory

- Removing hierarchical memory: 39.8% reduction
- Memory is crucial in capturing persistent distortion

② Propagation Role

- Removing propagation roles: critical
- Agents behave more **uniformly**, and the accumulation effect **disappears**

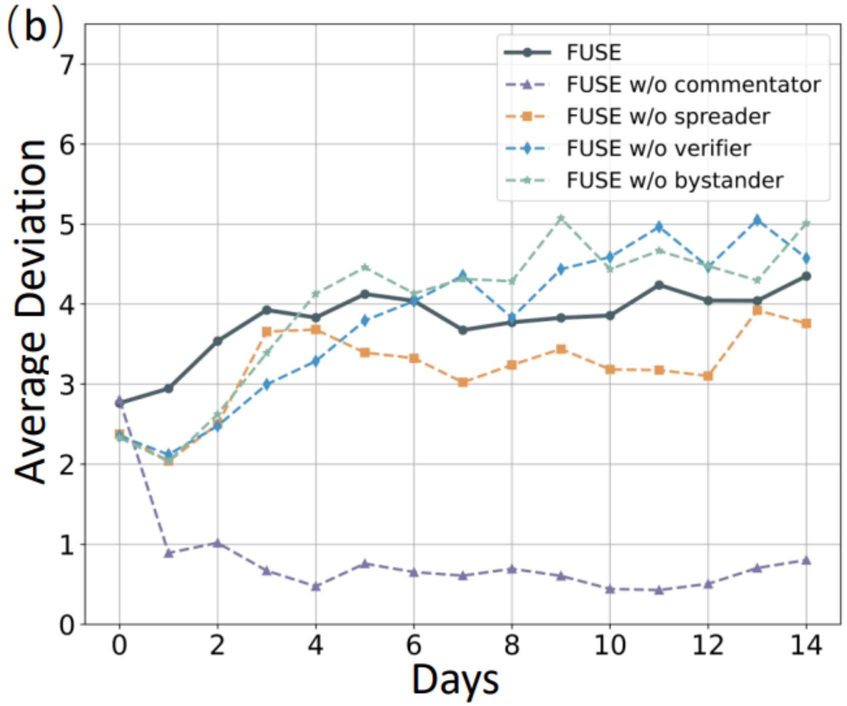


Ablation Study

✓ The Impact of Propagation Role Types

- Removing commentators: most significant drop

💣 *Demonstrate how different components contribute to simulating fake news evolution*

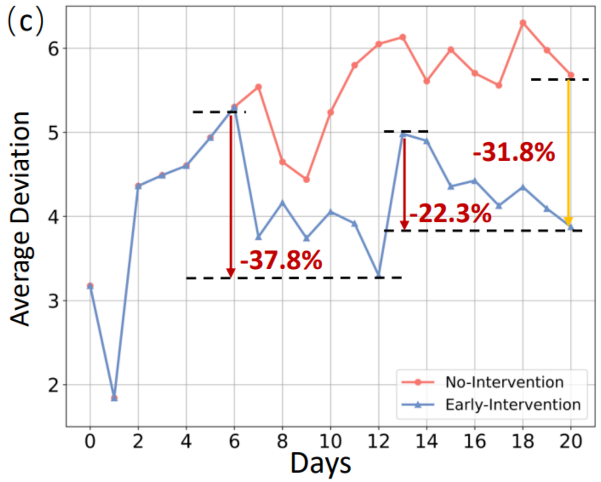


Discussion

✓ When and how can we intervene to reduce the spread of fake news?

- First intervention reduced deviation by 37.8%
- Effect gradually gets weakened through time
- Intervention strategy maintained lower deviation

💣 *Requires both early and regular interventions*

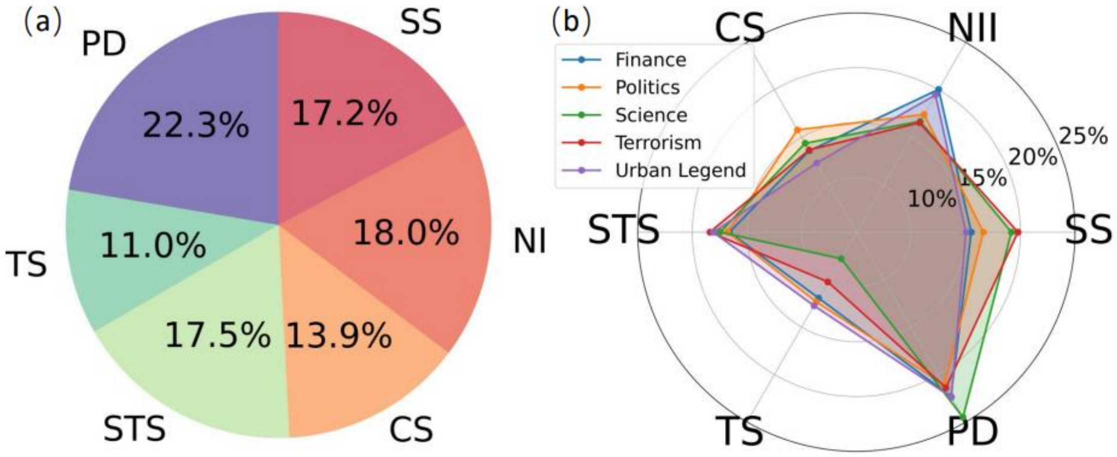


Comparison Factor	Setting	Δ Deviation↓	Average Deviation↓	Deviation Variance↓	Max Deviation↓	Min Deviation	Final Deviation↓	Peak Deviation Time ↑	Half Δ Deviation Time↑
Intervention	No Intervention	3.208	5.546	1.247	7.340	1.841	6.383	0.767	0.167
	Intervention	**1.384**	**4.207**	**0.476**	**5.302**	**1.841**	**4.559**	**0.200**	**0.067**

Discussion

✓ Factors in Fake News Evolution

- PD and NII are the main drivers of fake news evolution



💣 *Understanding these patterns can help developing targeted strategies*

Takeaway

- Key findings
  1) News exhibits accumulation distortion effects
  2) News distortion occurs rapidly in high clustering networks
  3) Political news evolves faster than other topics

Reveals the importance of early strategic intervention in fake news evolution

Review

- Strengths
  - Models how true news gradually evolves into fake news
  - Prevents oversimplification by adding roles, memory, and network dynamics
  - Uses multiple evaluation metrics for realistic validation

- Weaknesses
  - Possible data leakage despite GPT-4o-mini cutoff claim
  - Lacks modeling of intentional manipulation or disinformation
  - LLM bias and factual errors may affect simulation reliability