Beyond Static Artifacts: An Evolutionary Framework for Synthetic Claim Generation

Yeqing Teng[1,3†], Jiasheng Si[1,3†], Shuxia Lin[2], Linhai Zhang[4], Weiyu Zhang[1,3], Wenpeng Lu[1,3], Deyu Zhou[2], Xiaoming Wu[1,3*]

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]School of Computer Science and Engineering, Southeast University, China
[3]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China
[4]King's College London, UK

Abstract

With the generative capabilities of large language models (LLMs) reshaping the information ecosystem, the concern with the sociological validity of claim detection benchmarks is increasing. Current claim detection benchmarks predominantly treat claims as static textual artifacts, overlooking the sociological etiology of how information naturally emerges and mutates. In this paper, we propose an evolutionary paradigm that models claims as socially evolving entities. In specific, we introduce a socially generative framework for synthetic claim generation, a multi-agent simulation grounded in the Open Claims Model. By decomposing claims into context, utterance, and proposition, our approach enables the precise simulation of unmitigated propagation to capture truth decay, and intervened propagation with multi-auditor oversight for targeted generation. Furthermore, we propose the *background-user-perspective* (BUP) framework, which reformulates check-worthiness as a condition-dependent probability rooted in social environment. Experiments on our datasets verify the data quality and reveal how network topology and user attributes systematically shape veracity drift[1].

1 Introduction

Claim detection is a crucial gatekeeping task within fact-checking systems that entails identifying verifiable semantic units from a massive stream of online information (Panchendrarajan and Zubiaga, 2024). However, with the generative capabilities of large language models reshaping the information ecosystem, the concern with the "sociological" validity of existing benchmarks is increasing. Existing data construction research reflects a deficiency in understanding the etiology of claims—how they emerge, mutate, and polarize within heterogeneous social networks. This prompts us to develop a social paradigm that incorporates the explicit "evolutionary" process into the claim generation.

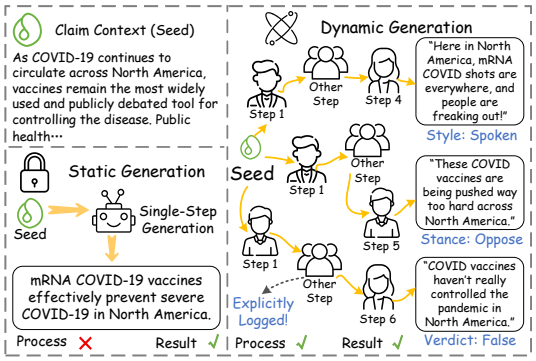

Figure 1: An example illustrating different claim generation methods. Dynamic generation enables multi-dimensional claim evolution with the generation path explicitly logged.

Specifically, substantial progress has been made on common claim detection benchmarks (Arslan et al., 2020; Ni et al., 2024; Dutta et al., 2022; Faramarzi et al., 2023). Nevertheless, existing benchmark research presents inherent limitations. These datasets typically follow a static *extract-rewriting paradigm*, which passively mines textual fragments from pre-existing sources (e.g., political debates) or performs template-filling generation using LLMs (Lucas et al., 2023; Bussotti et al., 2024; Ghosh et al., 2025). In this context, these works tend to produce claims that are linguistically diverse but lack sociological validity, limiting their potential to represent how claims are continuously reshaped by diffusion dynamics and cognitive biases. In the real information life-cycle (i.e., *created→propagated→transformed→repropagated*),

† Equal contribution.
* Corresponding author (wuxm@qlu.edu.cn).
[1]Code and data are available at https://github.com/yeqingteng/Beyond-Static-Artifacts

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33540–33570
July 2-7, 2026

claims evolve within human society, not merely linguistically generated (Xu et al., 2025; Ma et al., 2025).

In this paper, we draw inspiration from *generative social science* (Piao et al., 2025; Liu et al., 2025; Li et al., 2025a) and advocate a paradigm shift for claim generation: *moving beyond passive artifact extraction towards active information evolution*. Unlike prior claim generation that relies on predefined rules for substituting specific semantic fragments (Ortega and Gomez-Perez, 2025; Zhou et al., 2025; Bussotti et al., 2024), as shown in Fig. 1, we posit that claims are modeled not as static extracted sentences, but grow as a traceably evolving object shaped by the interplay between individual cognition and social propagation (Ghosh et al., 2025). This allows us to synthesize claim data that is no longer locked to static source corpora, but enables to trace and analyze their social trajectory.

To this end, we propose a traceable and intervenable social-evolutionary generative workflow for constructing evolved-claims in a multi-agent environment governed by sociological and cognitive principles. We ground this simulation in the *Open Claims Model* (Boland et al., 2022), which theoretically decomposes a claim into *Context* (the narrative anchor), *Utterance* (the linguistic form), and *Proposition* (the logical assertion). Building on this foundation, we explicitly model how a claim mutates via modular design, enabling our dataset to include not only the standard distinction between claims (i.e., *claim* vs. *non-claim*), but also cover the generation with different target properties (e.g., *Veracity Level*). Specifically,

(I) **Context-Anchored Initialization:** Building *Context* anchored to real-world narratives and generating *seed-claims* together with the necessary *non-claims* as the starting point of evolution.

(II) **Socialized Propagation and Evolution**: Simulating the claim life-cycle based on the dimensional decomposition of *Utterance* and *Proposition* through two mechanisms: (1) **unmitigated propagation** simulates the unconstrained sequential rewriting along propagation paths, which allows veracity shifts to naturally emerge and be amplified (Maurya et al., 2025); (2) **intervened propagation** introduces external auditor agents into a heterogeneous agent network to perceive and intervene in the specific direction of evolution.

Furthermore, *check-worthiness* is the intrinsic characteristic for defining the value of claims. However, due to its subjective nature, current research commonly presents conceptual ambiguity surrounding how to evaluate this property. Within our paradigm, we extend the definition of check-worthiness from hard labeling to a tunable judgment rooted in social propagation, and introduce the *background-user-perspective* (BUP) framework. This redefines check-worthiness as a condition-dependent probability, which is shaped by the function of the social backgrounds, user types, and evaluation perspectives.

We utilize this framework to construct the *Seed*, *Middle*, and *Evolved* datasets, which capture distinct stages of the information life-cycle. Our empirical results demonstrate that claims evolved through social simulation are more challenging to detect than their static seed counterparts, effectively mirroring the semantic complexity and noise introduced by real-world propagation. Furthermore, our analytical experiments confirm that our simulation effectively reproduces the dynamics of "truth decay", validating the sociological fidelity of our generation paradigm, and verifying the necessity of the proposed tunable check-worthiness assessments.

2 Related Work

Claim Generation Existing claim generation research is largely text-centered and follows three methodological orientations: (i) rewriting and controlled manipulation via paraphrasing (Lucas et al., 2023), style transfer (Wu et al., 2024), and semantic perturbation (Wang et al., 2024); (ii) knowledge- and evidence-based generation through knowledge graphs (Ghosh et al., 2025), tabular evidence (Bussotti et al., 2024), and QA signals (Pan et al., 2021); and (iii) narrative- or domain-structured generation using high-level representations such as summaries or topics (Vykopal et al., 2024) for domain-specific content (Ortega and Gomez-Perez, 2025; Zhou et al., 2025). In contrast, we use multi-agent propagation to capture socialized trajectories and generate diverse claim types.

Claim Detection and Check-Worthiness Detection Existing claim detection datasets are largely extracted from real-world corpora, ranging from sentence-level political statements (Arslan et al., 2020) to social media posts from breaking events (Shaar et al., 2021; Nakov et al., 2022), and more recently to minimal verifiable units (Dutta et al., 2022) and LLM-refined claims from weak-context and long-form sources (Metropolitansky and Lar-

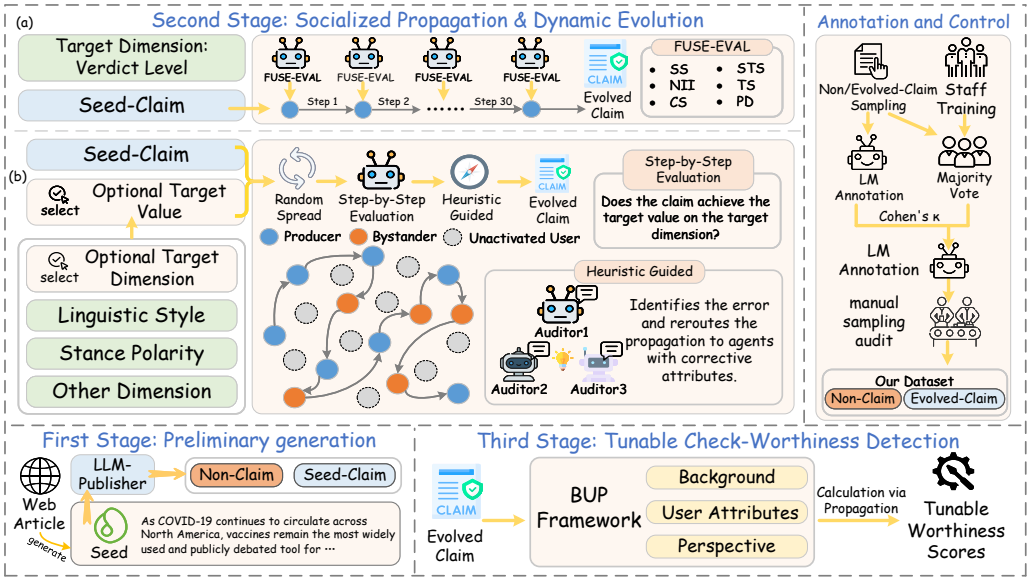

Figure 2: A social simulation framework for claim evolution and tunable check-worthiness detection.

son, 2025). In contrast to this extractive paradigm, we generate claims from narrative seeds via socially simulated evolution to obtain a diverse and controllable detection set. As the preceding task within the fact-checking pipeline, check-worthy detection evaluates the verification priority of claims and is driven by extra-textual factors such as novelty and controversy (Ni et al., 2024; Panchendrarajan and Zubiaga, 2024), making it inherently subjective and dynamic (Guo et al., 2022). This motivates us to introduce an interactive and tunable evaluation framework through social simulation.

Social Simulation Social simulation research is moving toward LLM-driven agents with enhanced cognition, enabling more controllable and traceable simulations (Piao et al., 2024), while incorporating behavioral science theories further improves generalization and interpretability across scenarios (Yan et al., 2025). Meanwhile, information ecosystem studies treat text as a dynamic entity continuously rephrased through social interactions, examining how its interaction with network structure drives diffusion (Yang et al., 2025; Sukiennik et al., 2025), as well as how distortion accumulates and can be controlled during misinformation propagation (Liu et al., 2024, 2025; Zhang et al., 2025b). Building on these advances, we seek to model propagative content evolution for claim generation.

3 Preliminary

Our framework is grounded in the theoretical literature on semantic claim representation (Boland et al., 2022). Thus, we begin by outlining the Open Claims Model (OCM) and then present the specific configuration of our simulated social environment.

Open Claims Model The OCM is the scheme for systematically representing the anatomy of online discourse. It theoretically decomposes a claim unit into three hierarchical layers to disentangle narrative background from linguistic variation: (i) **Context**: The narrative anchor that provides the background scenario (e.g., a specific event or topic), which serves as the immutable reference point for generation; (ii) **Utterance**: The linguistic form or surface realization of the claim, which concerns how the information is verbally encoded; (iii) **Proposition**: The logical assertion or semantic content of the claim, which concerns what is asserted as true or false.

To achieve diversity control over claim evolution, we further expand the latter two layers into six manipulable dimensions. Specifically, for Utterance, we define *Linguistic Style* (e.g., Rational vs. Emotional), *Expression Type* (e.g., Fact vs. Opinion), and *Message Granularity* (e.g., Macro vs. Micro). For Proposition, we define *Causal Structure*, *Stance Polarity*, and *Veracity Level*. This dimensional decomposition allows us to purposefully trace the trajectory of a claim as it mutates through social networks. More details can be found in Appendix C.

Simulated Social Environment The evolution of claims in our framework follows a social propaga-

tion mechanism, where the mutation of information is driven by the cognitive biases of heterogeneous agents. Specifically, we simulate the basic *user* agents by selecting individuals from the real-world user pool X (Zhang et al., 2025a), with each agent being instantiated with psychometric traits (e.g., *big five*). Moreover, following the *Inglehart-Welzel World Culture Map* (Inglehart and Welzel, 2005), we blend sociological groups into our agent design to simulate realistic cognitive filtering: Traditional Conservative, Technocratic Moderate, and Liberal Elite. Furthermore, to simulate the claim evolution, user agents are connected via specific network topologies that symbolize different social interaction patterns, wherein we adopt the Erdős–Rényi topology (Liu et al., 2025) as the default connection mode, and construct the clustering network and the scale-free network (Liu et al., 2025) to analyze the impact of different topologies on claim evolution. Details can be found in Appendix H.

4 Methodology

As illustrated in Figure 2, we adopt a theory-informed social simulation workflow to construct our dataset, with each claim labeled as {*claim*, *non-claim*}. Meanwhile, we assign a controllable attribute to each claim based on the six manipulable dimensions within the OCM, allowing for purposeful propagation within the social environment. This process encompasses three distinct phases: (I) Context-Anchored Initialization (§4.1), which grounds claims in real-world narratives; (II) Socialized Propagation (§4.2), which evolves claims through multi-agent interactions; and (III) Tunable Evaluation (§4.3), which assesses check-worthiness via the BUP framework.

4.1 Context-Anchored Initialization

• **Data Preparation** To ensure the sociological validity of our seed-claims, inspired by generating disinformation from abstracts (Vykopal et al., 2024), we collected authoritative news abstracts as *claim context* C_{ctx} from news outlets (e.g., Reuters, The Guardian) spanning five high-stakes topics, including COVID-19, health, Russia-Ukraine, US-Elections, and Regional-Affairs. Within these abstracts, we initially filtered out information lacking logical propositions and extracted atomic semantic units. More details can be found in Appendix D.

• **Seed-Claim and Non-Claim Generation** By taking the extracted C_{ctx} as input, we generate Seed-Claims (labeled as *claim*) and Non-Claims (labeled as *non-claim*) by prompting GPT-4o with distinct templates. Notably, we strictly adhere to the fine-grained taxonomy proposed by Konstantinovskiy et al. (2021): only statements verifiable by public evidence (e.g., quantity, causation) are retained as claims, while personal experiences and "not a claim" (e.g., questions, directives) are categorized as non-claims. The details are listed in Appendix F. Furthermore, to emulate the ambiguity inherent in real-world discourse, we introduce "gray areas" into a subset of the initial data prior to the evolution. Specifically, *non-claims* are rewritten to adopt professional phrasing while remaining semantically empty, whereas *seed-claims* are pre-processed with referential ambiguity, non-semantic symbols (e.g., emojis), and emotional modifiers to inject a realistic signal-to-noise ratio.

4.2 Socialized Propagation and Evolution

We consider the six dimensions encompassed by *Utterance* and *Proposition* as manipulable dimensions in the propagation evolution, with each dimension d corresponding to a set of discrete selectable values (e.g., *Veracity Level* includes True, Half-True, etc.). More details about the authenticity values can be found in Table 13. In this context, we design two distinct propagation mechanisms to capture different facets of information dynamics.

Unmitigated Propagation: Simulating Truth Decay For the dimension of *Veracity Level*, following (Maurya et al., 2025), we model unmitigated propagation as a sequential rewriting process, allowing veracity shifts to emerge naturally without conditional constraints. This process simulates the "telephone game" effect, where subtle semantic shifts accumulate over time. Specifically, given the *seed-claim*, each user agent rewrites the claim based on its cognitive biases and demographic profiles, and propagates it randomly to its neighboring users following specific social networks. Specifically: (i) Erdős–Rényi (Random) network: user agent performs unconstrained random propagation to its neighbors; (ii) Scale-free network: a user agent propagates the claim based on its attributes and the influence of the hub-user, simulating the amplification effect of opinion leaders on information evolution; (iii) Clustering network: propagation occurs only among users within homogeneous groups (e.g., *Liberal Elite*), reflecting the impact of community homogeneity on evolution trajectory.

Model	CheckThat(2022)		AFaCTA(2024)		Claimify(2025)		Ours(seed)		Ours(middle)		Ours(evolved)	
	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)
RoBERTa	71.67	71.15	77.50	77.49	90.83	90.76	**88.54**	**88.05**	86.67	86.15	84.79	84.15
DeBERTa	75.83	75.63	80.00	79.80	90.00	90.32	**88.33**	**87.83**	86.98	86.50	85.63	85.17
GPT-4o	62.92	72.27	80.31	75.61	89.24	90.11	**90.00**	**89.10**	85.83	84.89	82.29	81.89
Qwen3-Max	61.46	72.05	81.35	80.09	84.21	88.39	**85.73**	**85.32**	81.25	80.66	79.38	78.66
Deepseek-V3.2	67.95	74.94	80.83	80.79	86.16	87.13	**88.13**	**87.76**	81.77	80.70	77.29	76.70

Table 1: Claim detection performance across multiple datasets. This table compares model performance on three public datasets and on our *Seed*, *Middle* and *Evolved* datasets, which represent different stages of claim evolution.

Dimension	Seed-Claim		Evolved-Claim	
	coverage	length	coverage	length
Style	0.14	18.81	**0.16**	**26.95**
Granularity	0.15	18.12	**0.18**	**34.50**
Type	0.13	19.14	**0.17**	**35.69**
Structure	0.13	19.02	**0.15**	**27.64**
Stance	0.14	18.57	**0.16**	**29.20**
Veracity	0.14	18.93	**0.16**	**26.82**

Table 2: Length and information coverage of *seed-claims* and *evolved-claims* across six evolutionary dimensions.

Train	vanilla test	style	stance	granularity
vanilla	80.00%	75.83%	77.50%	72.50%
enhanced	80.67%	78.30%	79.17%	77.50%

Table 3: Classification accuracy of models trained on vanilla and enhanced datasets under original and shifted test conditions.

Intervened Propagation: Controllable Diffusion To guide the synthetic claim generation to cover specific linguistic or logical features, we introduce intervened propagation for the remaining dimensions within the OCM (except for *Veracity Level*). By combining multi-expert assessment with topology-aware directed propagation, intervened propagation enables the manipulation of claim evolution. Specifically, given a target evolution dimension (e.g., *Linguistic Style*) and its corresponding target value (e.g., Emotional): (i) **Producer Dynamics**: User agents in the networks dynamically assess their own attributes and current claims against the target value. Those with matching profiles act as *producers* to rewrite the claim and propagate it to their neighbors within the three social networks, while others act as *bystanders* to propagate it without modification. (ii) **Multi-Auditor Oversight**: To guide the claim toward the target value, we introduce a multi-auditor mechanism to monitor the claim evolution process. Specifically,

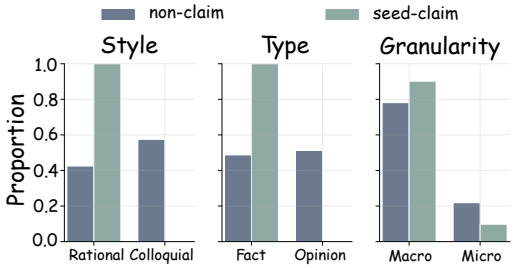

Figure 3: Separability of *non-claims* and *seed-claims* across three key dimensions.

after each propagation step, three auditors evaluate whether the claim meets the target value via a majority voting strategy: if a claim deviates from this target, a synthesized diagnosis from the auditors identifies the error and reroutes the propagation to agents with corrective attributes.

4.3 Tunable Check-Worthiness Evaluation

Inspired by the Social Amplification of Risk Framework (Kasperson et al., 1988), we introduce the *Background–User–Perspective* (BUP) framework. This framework redefines check-worthiness not as an intrinsic static property, but as a dynamic evaluative potential shaped by the interplay of social context and evaluation goals. Formally, we model the check-worthiness of a claim c as a conditional function dependent on three key coordinates: the **Social Background** (S), the **User Attribute** (A), and the **Evaluation Perspective** (P):

$$CW(c \mid S, A, P) = f\big(c, S_{\text{soc}}, A_{\text{attr}}, P_{\text{persp}}\big) \quad (1)$$

where S_{soc} denotes the socio-environmental parameters (e.g., Trust Climate), A_{attr} represents the interacting user attributes (e.g., Occupation), and P_{persp} specifies the evaluation dimensions and their weights. Specifically, P determines the balance between two complementary perspectives: **propagation ability** (macro-level diffusion depth and

audience size) and **content value** (micro-level user ratings). Under a specific instantiation of (S, A, P), the function f calculates the final soft-label score as a weighted aggregation of the simulated propagation metrics and content ratings yielded by the user agents within the given environment. Details can be found in Appendix B.

4.4 Label Annotation and Quality Control

To mitigate LLM hallucinations, we adopt a closed-loop human–machine annotation workflow. Specifically, **(i) Automated Annotation:** We employ DeBERTa-V3 for binary verifiability classification, while GPT-5.1, Claude-4.5-Sonnet, and Llama-3.1-405B are utilized for multi-class evolution labeling (e.g., *Veracity Level*). To ensure reliability, we use PolitiFact-style human annotations as a reference and calibrate model agreement with Cohen's κ. **(ii) Manual Batch Checking:** Adopting a batch-wise sampling strategy (100 instances per batch), we manually inspect 10% of samples. Batches with <70% accuracy are discarded, those between 70%–90% are regenerated, and only batches exceeding 90% accuracy are retained. Complete details of this annotation process can be found in Appendix A. And in Tables 11, 12, and 13, we present examples of two types of *evolved-claims* and *non-claims*; in Tables 14 and 15, we provide example traces of two types of propagation.

5 Experiments and Analysis

5.1 Experiment Settings

• **Datasets:** We use three public claim detection datasets: CheckThat (Nakov et al., 2022), AFaCTA (Ni et al., 2024), and Claimify (Metropolitansky and Larson, 2025), covering verifiable and non-verifiable claims from tweets, political speech, and LLM-generated texts. Based on our framework, we further obtain *Seed*, *Middle* and *Evolved* datasets, containing seed-claims (pre-evolution), middle-claims (mid-evolution) or evolved-claims (post-evolution) together with non-claims.

• **Evaluation Metrics:** As shown in Table 2, we evaluate the two claim types using *Length* and *Coverage*, measuring token length and information preservation. *Coverage* is defined based on the element-level perspective introduced in (Metropolitansky and Larson, 2025), uses an NLI model to test how well a claim entails atomic units extracted from the basis seed. Both metrics are reported at the dimension level as the mean over claims.

• **Implementation Details:** We compute *Coverage* with mDeBERTa-v3-XNLI. GPT-4o, Qwen3-Max, and DeepSeek-V3.2 serve as auxiliary auditors in intervened propagation, while GPT-4o handles context-generation and FUSE-EVAL evaluation. Unmitigated and intervened propagations run up to 30 steps within a simulated 90-agent network.

5.2 Claim Detection Performance

As shown in Table 1, our datasets demonstrate a certain level of detection difficulty compared to existing benchmarks (especially *evolved* dataset). By deliberately introducing "gray areas" during the Context-Anchored Initialization stage for *seed-claims*, which are naturally retained or even amplified during the propagation process, our generated data presents a realistic challenge. Consequently, model performance on our datasets is notably lower (indicating higher detection difficulty) than on Claimify dataset, which relies on LLM-refined claims from weak-context and long-form sources. Instead, the detection difficulty approaches that of the real-world AFaCTA dataset, which is sourced from U.S. political speeches. This alignment with real-world ambiguity effectively demonstrates the practical applicability and realism of our dataset construction paradigm.

5.3 Propagation Effect on Claim Detectability

We further investigate why models achieve higher detection performance on *seed-claims* than on *middle-claims* and *evolved-claims*. As shown in Figure 3, before propagation, *seed-claims* are clearly separable from *non-claims* along *Language Style*, *Expression Type*, and *Message Granularity*, but intervened propagation shifts these dimensions through user-driven rewrites, weakening the discriminative signals learned by RoBERTa and DeBERTa. Table 2 further shows that propagated claims increase in both *Length* and *Coverage*, with *Length* growing much faster, resulting in "factual dilution" at semantic and structural levels. Together, these two factors indicate that the class polarization and high fact density of *seed-claims* are the key reasons why they are more easily distinguished by detection models.

5.4 Robustness under Controlled Distribution Shifts

To evaluate the practical value of *evolved-claims*, we analyze model performance under controlled

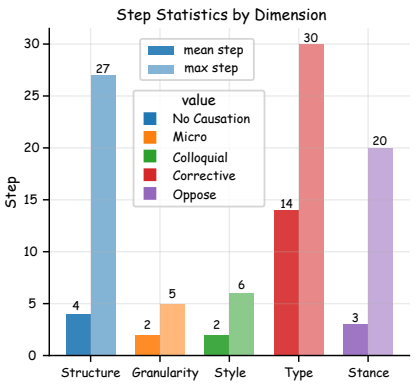

Figure 4: The target value requiring the highest average number of propagation steps for successful evolution within each dimension, alongside its maximum step.

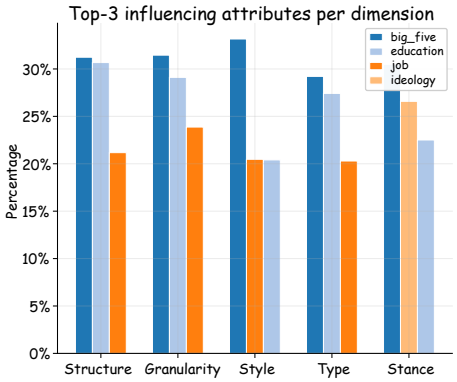

Figure 5: The influence of key attributes that drive users to act as *producers* during propagation across different dimensions.

distribution shifts (*style*, *stance*, and *granularity*) using the AFaCTA dataset and a DeBERTa classifier. As shown in Table 3, augmenting the *vanilla* training set with *evolved-claims* (*enhanced*) maintains high compatibility with the original performance, slightly improving accuracy on the *vanilla* test set (80.67% vs. 80.00%). Furthermore, the enhanced model consistently outperforms the *vanilla* baseline across all three shifted scenarios. Notably, under the *granularity* shift, the addition of evolved data yields a significant 5.00% absolute performance gain (77.50% vs. 72.50%). These results demonstrate that incorporating *evolved-claims* encourages the model to learn more generalizable semantic features, successfully mitigating performance degradation when confronted with real-world data shifts.

5.5 Propagation Steps and User Attributes

We analyze intervened propagation, a core claim-generation mechanism, in terms of spread steps and user attributes. As shown in Figure 4, the average number of spread steps for *Structure*, *Granularity*, *Style*, and *Stance* remains relatively low, indicating that these dimensions can be guided toward their target values primarily through local expressive or structural adjustments. In contrast, *Type* requires longer spread paths, as content-level reorganization in its rewriting imposes higher alignment demands between user attributes and expressive capacity. We also discovered that target values such as *Oppose* and *Corrective* are mainly introduced through *producer*-driven rewrites and are relatively rare in content directly generated by the model. Figure 5 further shows that controllable rewriting is driven by specific attributes, with *Big Five* and *education* playing prominent roles across five dimensions, highlighting the role of cognitive and expressive traits in shaping users' likelihood of becoming a *producer*.

5.6 Unmitigated Truth Drift

We examine the evolution of veracity for *seed-claims*. As shown in Figure 6, initially true seed-claims drift more toward half true and false in scale-free networks than in random networks. In clustered networks, *Moderates* and *Liberals* partially limit drift based on their relatively high cognitive levels, and *Conservatives* accumulate greater deviations through value filtering. FUSE-EVAL results (Figure 7) show a "rapid-then-plateau" pattern across five topics, indicating that early propagation drives most drift (Jin et al., 2013). And we can see larger increases for the more controversial topics of *COVID-19*, *Russia–Ukraine*, and *US-Elections*. Across dimensions, *Temporal Shift* and *Paraphrasing Degree* dominate overall distortion (Figure 12). Overall, veracity drift is shaped by propagation depth, network topology, group type, and topic. Further analysis of unmitigated propagation are provided in Appendix E.

5.7 Alignment Between Simulated and Real-World

We aligned the simulation results with real-world empirical observations in multiple dimensions (see Appendix I for details). First, the cognitive load theory explains why convergence at the *Style* occurs more rapidly than at the deeper *Type*, which is consistent with empirical studies on social media users' tendency to perform low-cost, shallow edits (Adamic et al., 2016; Zubiaga et al., 2016).

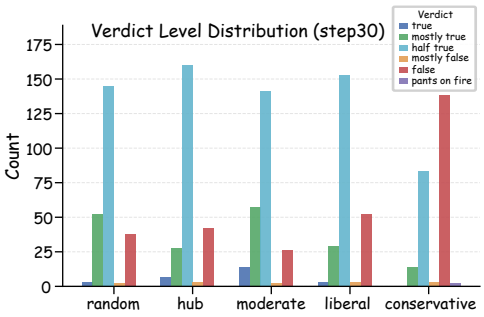

Figure 6: Verdict drift under unmitigated propagation after 30 steps across network topologies and user groups.

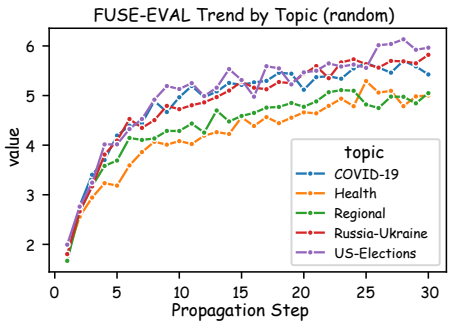

Figure 7: The trajectories across topics during propagation in the random network.

Model	Dimension	Cohen's κ
DeBERTa-V3	Claim / Non-claim	0.77
GPT-5.1	Language Style	0.75
	Expression Type	0.71
	Message Granularity	0.73
	Causal Structure	0.66
	Stance Polarity	0.68
	Veracity Level	0.66

Table 4: Human-machine annotation agreement across dimensions.

Target Dimension	Direct	One Auditor	Ours
Language Style	92.92%	95.21%	**96.46%**
Message Granularity	78.33%	93.50%	**94.58%**
Expression Type	72.22%	84.83%	**85.17%**
Causal Structure	85.83%	89.09%	**91.13%**
Stance Polarity	77.29%	90.54%	**93.36%**

Table 5: Target-value attainment across intervened propagation dimensions.

Second, bias amplification by hub nodes in scale-free networks supports the role of opinion leaders in accelerating truth decay. (Grinberg et al., 2019; Vosoughi et al., 2018). Additionally, our findings regarding group behavior and topic sensitivity, namely that conservative groups exhibit stronger value-based filtering tendencies, and controversial topics (e.g., *COVID-19*) lead to higher distortion rates in information propagation, are strongly supported by extensive sociological and communication data (Guess et al., 2019; Bakshy et al., 2015; S. et al., 2020; Pulido et al., 2020). Moreover, relevant psychological studies (Pennycook and Rand, 2021; Lawson and Kakkar, 2022) show that psychological attributes are more influential on information manipulation than some basic demographic factors, which is consistent with our simulation.

To further enhance the alignment between our simulation and reality, we introduced experiments involving real human participants. For example, we conducted human validation of the evolution trajectory and a "Human Turing Test" on our synthetic claims (also detailed in Appendix I).

5.8 Human-Machine Annotation Results

For **automated annotation** described in Section 4.4, Table 4 shows that DeBERTa-V3 and GPT-5.1 achieve Cohen's κ of 0.77 and 0.66–0.75 across the annotation tasks, respectively, reaching *substantial agreement*. For **manual batch checking** described in Section 4.4, we regard claims and non-claims as invalid if their labels are inconsistent with human annotations. Specifically, (i) for *seed-claims* and *non-claims*, the **claim detection labels** (*claim* vs. *non-claim*) annotated by the model must match the human annotations; (ii) for *evolved-claims*, the **target values** annotated by the model (e.g., *Half-True*) in the corresponding evolution dimensions (e.g., *Verdict Level*) must be consistent with human annotations. Based on this, Table 6 indicates that all three types of data achieve retention rates above 95% after closed-loop filtering, demonstrating stable label quality control at scale. Table 8 further shows that discard numbers vary across evolution dimensions, but each dimension ultimately retains a substantial amount of quality-controlled data for downstream analysis. More detailed results and analysis are provided in Appendix A.

5.9 Targeted Claim Control Success Rate

Table 5 reports the accuracy of intervened propagation in achieving **preset target values** across five controllable dimensions (*Linguistic Style*, *Message Granularity*, *Expression Type*, *Causal Structure*, and *Stance Polarity*). For each *evolved-claim*, the final annotation is compared with its generation-time target to compute the pass rate, following the anno-

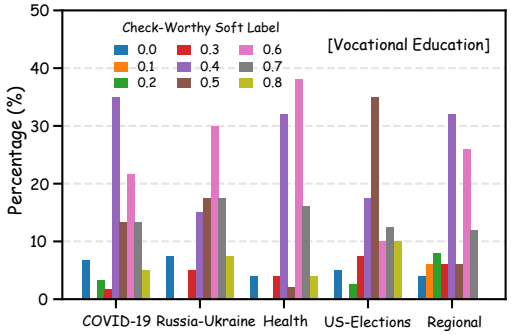

Figure 8: User-conditioned check-worthiness distributions across topics for users with *vocational education*.

tation procedure in Appendix A. We compare three settings: one-shot direct generation, single-auditor, and the full method with multi-auditor review. The results show that the full method achieves target-value attainment rates above 85% on all dimensions, with several exceeding 90%, indicating stable alignment with the intended targets. The gap between one-shot generation and single-auditor demonstrates that auditing effectively suppresses drift, while the further improvements of the full method over the single-auditor setting confirm the effectiveness of introducing multi-auditor review.

5.10 Check-Worthiness Sensitivity

We analyze check-worthiness distributions across different settings to demonstrate their tunable nature. Figure 8 shows that *vocational-education* users produce more evenly spread scores, while Figure 13 shows that *higher-education* users concentrate more on mid-range values, reflecting more cautious judgments. Figure 14 further shows that focusing only on *Controversy* and *Novelty* sharpens high-score concentration, whereas adding *propagation capacity* and other *content values* yields smoother distributions, reflecting how multi-dimensional value trade-offs attenuate the dominance of highly stimulating cues. We further validate the BUP framework through human-in-the-loop empirical evaluation in Appendix I.

6 Conclusion

This work advances the paradigm of synthetic claim generation from static textual manipulation to sociologically grounded evolution. By integrating the Open Claims Model with multi-agent simulation, we established a traceable framework to generate the claim data that reproduces the etiology of truth decay and semantic mutation within heterogeneous networks. Meanwhile, the proposed *Background–User–Perspective* (BUP) framework reconceptualizes check-worthiness: moving beyond binary labeling to a conditional probability shaped by the interplay of *background*, *user*, and *perspective*, capturing the subjective and dynamic nature of verification.

Limitations

Our study has several limitations that point to valuable future directions:

- **Scope of Human Validation:** While our closed-loop calibration ensures baseline quality, the massive scale of our social simulation necessitates heavy reliance on LLMs for both generation and annotation. The extent of pure human validation remains relatively limited, serving primarily as a calibration reference rather than the core evaluation metric. Exploring scalable frameworks for comprehensive human-in-the-loop validation—moving beyond corroborating consensus among different LLMs—remains a crucial direction for future propagation-based generation paradigms.

- **Static User Modeling:** The current framework approximates different types of individuals by assigning static attributes to LLMs, but does not incorporate explicit episodic and persistent memory, belief updating, or human-like reasoning processes. This, to some extent, limits its ability to capture the dynamic nature of real user behavior.

Ethics Statement

This study uses news abstracts as generation seeds, which are carefully processed to exclude any private or sensitive information. Individuals from a real-world user pool X are modeled as abstract agents with synthetic attributes, without involving or exposing any personal or sensitive data. Partially false or misleading claims generated through our evolutionary framework are produced within a controlled simulation environment, and any downstream use must comply with relevant ethical guidelines and regulations.

Acknowledgments

This work is supported by Key R&D Program of Shandong Province (No.2024CXGC010113), the Taishan Scholars Program (No.TSQN202507242,

TSPD20240814), the National Natural Science Foundation of China (No.62402258, No.62376130), the Shandong Provincial Natural Science Foundation (No.ZR2024QF099, ZR2024MF088), the Program of New Twenty Policies for Universities of Jinan (No.202333008), and the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2025ZDZX01).

References

Lada A. Adamic, Thomas M. Lento, Eytan Adar, and Pauline C. Ng. 2016. Information evolution in social networks. In *Proceedings of the ninth ACM international conference on web search and data mining*, pages 473–482.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223.

Furkan Arslan, Naeemul Hassan, Chengkai Li, Mark Tremayne, Bill Adair, and Di Xin. 2020. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 821–829.

Eytan Bakshy, Seth Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

K. Boland, P. Fafalios, A. Tchechmedjiev, and et al. 2022. Beyond facts: A survey and conceptualisation of claims in online discourse analysis. *Semantic Web*, 13(5):793–827.

Samantha Bradshaw and Philip N. Howard. 2018. The global organization of social media disinformation campaigns. *Journal of International Affairs*, 71(1.5):23–32.

Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, and Paolo Papotti. 2024. Unknown Claims: Generation of fact-checking training examples from unstructured and structured data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12105–12122.

Elizabeth Clark, Tal August, Sofia Serrano, Niket Tandon, Suchin Gururangan, and Noah A. Smith. 2021. All that's "human" is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 7282–7296.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Robin I. M. Dunbar. 1992. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493.

Subhabrata Dutta, Rudra Dhar, Prantik Guha, Arpan Murmu, and Dipankar Das. 2022. A multilingual dataset for identification of factual claims in indian twitter. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 88–92.

Noushin Salek Faramarzi, Fateme Hashemi Chaleshtori, Hossein Shirazi, Indrakshi Ray, and Ritwik Banerjee. 2023. Claim extraction and dynamic stance detection in COVID-19 tweets. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1059–1068.

Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, and Sriparna Saha. 2025. InfoGen: Generating complex statistical infographics from documents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 20552–20570.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363(6425):374–378.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5(1):1–8.

Zhiji Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the Eleventh International Conference on Learning Representations*, pages 1–16.

Ronald F. Inglehart and Christian Welzel. 2005. *Modernization, Cultural Change and Democracy: The Human Development Sequence*. Cambridge University Press, Cambridge, UK.

Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, pages 1–9.

Roger E. Kasperson, Ortwin Renn, Paul Slovic, Halina S. Brown, Jacque Emel, Robert Goble, Jeanne Kasperson, and Samuel Ratick. 1988. The social amplification of risk: A conceptual framework. *Risk Analysis*, 8(2):177–187.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–896.

Lev Konstantinovskiy, Oliver Price, Will Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, 2(2):1–16.

Matthew A. Lawson and Hemant K. Kakkar. 2022. Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. *Journal of Experimental Psychology: General*, 151(5):1–24.

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 Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Pu Liang, Luis Alonso, and Kent Larson. 2025a. Simulating society requires simulating thought. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS), Position Paper Track*, pages 1–17.

Y. Li, Q. Zhang, W. Lu, and et al. 2025b. Time-aware medication recommendation via intervention of dynamic treatment regimes. In *Proceedings of the ACM Web Conference 2025*, pages 5161–5172.

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, Zirui Song, Juntian Zhang, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. 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*, pages 26187–26203.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305.

Rui Ma, Xue Wang, and Guang R. Yang. 2025. Fighting fake news in the age of generative ai: Strategic insights from multi-stakeholder interactions. *Technological Forecasting and Social Change*, 216:1–22.

Raj Gaurav Maurya, Vaibhav Shukla, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. 2025. Simulating misinformation propagation in social networks using large language models. *arXiv preprint arXiv:2511.10384*.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 6996–7045.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, Alex Nikolov, and Yavuz Selim Kartal. 2022. Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In *Proceedings of the 2022 Conference and Labs of the Evaluation Forum (CLEF 2022)*, pages 368–392.

Jingwei Ni, Mingyang Shi, Dominik Stammbach, Nathanael Schärli, Adina Williams, and Robert West. 2024. AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1890–1912.

Raúl Ortega and Jose Manuel Gomez-Perez. 2025. Sciclaims: An end-to-end generative system for biomedical claim analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 141–154.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483.

Rajalakshmi Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:1–15.

Gordon Pennycook and David G. Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402.

Jinghua Piao, Yuwei Yan, Nian Li, Jun Zhang, and Yong Li. 2024. Exploring large language model agents for piloting social experiments. In *Proceedings of the Second Conference on Language Modeling*, pages 1–25.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of

LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Cristina M. Pulido, Beatriz Villarejo-Carballido, Gisela Redondo-Sama, and Aitor Gómez. 2020. Covid-19 infodemic: More retweets for science-based information on coronavirus than for false information. *International Sociology*, 35(4):377–392.

Brennen J. S., Simon F. M., Howard P. N., and Nielsen R. K. 2020. Types, sources, and claims of covid-19 misinformation. *Reuters Institute for the Study of Journalism*, pages 1–13.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 82–92.

Jiasheng Si, Yao Zhu, Wenpeng Lu, and 1 others. 2024. Denoising rationalization for multi-hop fact verification via multi-granular explainer. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12593–12608.

Nicholas Sukiennik, Yichuan Xu, Yuqing Kan, Jinghua Piao, Yuwei Yan, Chen Gao, and Yong Li. 2025. The roots of international perceptions: Simulating US attitude changes towards china with LLM agents. *arXiv preprint arXiv:2508.08837*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14830–14847.

Lionel Z. Wang, Yiming Ma, Renfei Gao, Beichen Guo, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Ng. 2024. Megafake: A theory-driven dataset of fake news generated by large language models. *arXiv preprint arXiv:2408.11871*.

J. Wu, J. Guo, and B. 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.

Guiqiong Xu, Meng Qian, and Lei Meng. 2025. Misinformation dissemination on social media: key research themes and evolutionary paths between 2013 and 2023. *Humanities and Social Sciences Communications*, 12(1):1–15.

Yuwei Yan, Jinghua Piao, Xiaochong Lan, Chenyang Shao, Pan Hui, and Yong Li. 2025. Simulating generative social agents via theory-informed workflow design. *arXiv preprint arXiv:2508.08726*.

Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. 2025. Twinmarket: A scalable behavioral and social simulation for financial markets. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*, pages 1–34.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, and 2 others. 2025a. Socioverse: A world model for social simulation powered by LLM agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*.

Yunyao Zhang, Zikai Song, Hang Zhou, Wenfeng Ren, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025b. GA-S^3: Comprehensive social network simulation with group agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8950–8970.

X. Zhao, W. Lu, C. Zheng, and et al. 2025. Plan dynamically, express rhetorically: A debate-driven rhetorical framework for argumentative writing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9562–9584.

Siying Zhou, Yiquan Wu, Hui Chen, Xueyu Hu, Kun Kuang, Adam Jatowt, Chunyan Zheng, and Fei Wu. 2025. ClaimGen-CN: A large-scale chinese dataset for legal claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12296–12323.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS One*, 11(3):1–29.

A Detailed Annotation Protocol

A.1 Initial Annotation Workflow

For the *seed-claims* and *non-claims* in the context-anchored generation stage, we apply a trained DeBERTa-V3 (He et al., 2023) model to perform binary classification in order to determine their verifiability. For the *evolved-claims* produced in the propagation stage, GPT-5.1 is used to conduct multi-class annotation to determine the target value along the corresponding evolution dimensions. Specifically, for evolved-claims generated via unmitigated propagation, GPT-5.1 assigns

Data	Initial	Accepted	Reflowed	Discarded	Final	Ratio (%)
Non-Claim	5000	4600	200	200	4800	96.00
Seed-Claim	5000	4800	100	100	4900	98.00
Evolved-Claim	4900	4400	300	200	4700	95.92

Table 6: Quality control results for *non-claims*, *seed-claims*, and *evolved-claims* using DeBERTa-V3 and GPT-5.1.

Model	Initial	Accepted	Reflowed	Discarded	Final	Ratio (%)
Claude	4900	4200	400	300	4600	93.88
Llama	4900	4700	100	100	4800	97.96

Table 7: Quality control results for *evolved-claims* using Claude-4.5-Sonnet and Llama-3.1-405B.

Evolution Dimension	Discarded	Evolved-Claim
Language Style	44	480
Message Granularity	26	480
Expression Type	36	689
Causal Structure	19	946
Stance Polarity	25	955
Veracity Level	50	1150

Table 8: *Evolved-claim* counts and discarded samples across evolution dimensions.

Dimension	Llama-3.1	Claude-4.5
Language Style	0.72	0.73
Expression Type	0.69	0.75
Message Granularity	0.72	0.73
Causal Structure	0.65	0.68
Stance Polarity	0.66	0.69
Veracity Level	0.65	0.67

Table 9: Human-machine annotation agreement (Cohen's κ) across dimensions when using Llama-3.1 and Claude-4.5.

their *Veracity Level* (e.g., *True*, *False*); for *evolved-claims* generated via intervened propagation, GPT-5.1 assigns labels on the target dimension (e.g., *Official* or *Colloquial* for *Linguistic Style* dimension). Both types of model-based annotation share the same human–machine calibration workflow: on a randomly sampled subset jointly annotated by humans and models, Cohen's κ (Cohen, 1960) is used to verify inter-annotator agreement, after which the model annotates the remaining samples. The human annotation protocol follows the *PolitiFact* platform, with three-annotator majority voting used to form consensus labels as the "golden label". Tables 16 and 17 specify the annotation criteria used to define each possible target value within every dimension.

Model	Accuracy (%)
Claude-4.5-Sonnet	93.20
Llama-3.1-405B	91.40
GPT-5.1	94.40

Table 10: Performance of different LLMs in further human spot-check validation.

A.2 Batch-Based Quality Control

On this basis, to ensure data quality under large-scale annotation, we further adopt a batch-based sampling quality control and closed-loop correction mechanism (Klie et al., 2024). All automatically annotated data are divided into 100-instance batches, with 10% of each batch randomly sampled for manual inspection. Batches with an inspection accuracy of above 90% are directly accepted; those with accuracy between 70% and 90% are returned to the generation and re-annotation pipeline; and those with accuracy below 70% are discarded. This batch-level closed-loop process is iteratively executed until all retained batches satisfy the acceptance criteria. The resulting dataset covers both non-claims and multiple types of verifiable claims.

A.3 Cross-Model Validation and Reliability

It is worth noting that to verify our "silver standard" labels are OpenAI-agnostic, we conducted a cross-model annotation validation using two additional models (Claude-4.5-Sonnet and Llama-3.1-405B) (Tables 7 and 9). Human spot-checks demonstrate that their label passing rates are highly comparable to those of GPT-5.1. Furthermore, to robustly verify the label reliability of the synthetic data, we conducted an additional human spot-check validation. Comparing human and model annotations (Claude-4.5-Sonnet, Llama-3.1-405B, and GPT-5.1) across 500 previously uninspected samples reveals an average model accuracy of 93% (Table 10), further

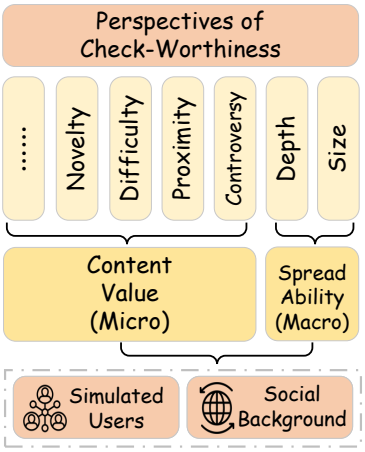

Figure 9: An overview of the BUP framework.

demonstrating the reliability of our synthetic claim labels.

A.4 Human Annotator Setup

The above human–machine annotation procedures are supported by a human annotation team. From 10 students with NLP backgrounds, we selected three annotators after training and evaluation. Each sample requires 3–10 minutes of human annotation time, with an average of approximately 5 minutes, and the compensation rate is USD 3.15 per hour.

B Check-Worthiness Evaluation and Propagation Modeling

B.1 Background and Perspective in the BUP Framework

As shown in Figure 9, we quantify the check-worthiness of a claim through its propagation from two complementary perspectives: *propagation ability* and *content value*. *Propagation ability* characterizes the diffusion potential of a claim within a given user population and is defined by two metrics. *Depth* measures the maximum number of hops that the claim reaches in the propagation cascade, while *Size* denotes the total number of distinct users reached by the claim during propagation. Together, these two metrics reflect the extent to which a claim can gain public visibility and amplify agenda-setting effects within the social system. *Content value* captures the importance of a claim at the cognitive and societal-risk levels and is evaluated along six interpretable dimensions: *relevance to the public interest*, *controversy or potential harm*, *involvement of elite individuals or organizations*, *geographic or cultural proximity*, *novelty*, and *verification difficulty* (Si et al., 2024). These dimensions constitute the primary content-level drivers of check-worthiness. Furthermore, we contextualize the evaluation within the *social background*. We incorporate three representative environmental dimensions that modulate check-worthiness: *time node* (sensitive vs. normal) (Vosoughi et al., 2018; Li et al., 2025b), *regulatory pressure* (high vs. low) (Bradshaw and Howard, 2018), and *trust climate* (high vs. low) (Lazer et al., 2018).

B.2 Construction of the Propagation Network

In tunable check-worthiness detection, we do not adopt the random, scale-free, or clustering networks used in the generation stage. Instead, we compute semantic similarities among 90 simulated users using *bge-large-en-v1.5*. Guided by *Dunbar's social layering theory* (Dunbar, 1992), we then construct three tiers of social ties for each user, namely *strong*, *medium*, and *weak*. This process thereby yields a propagation network that encodes relational layers. This design constrains information flow by social proximity, thereby more realistically capturing the tendency that users preferentially transmit information to socially closer contacts.

B.3 Check-Worthiness Scoring via Propagation

For each claim, we simulate its propagation within the selected user subgroup and selected social background. Each user who receives the claim assigns a continuous score in the range [0,1] on each *content value* dimension, representing their subjective judgment of how check-worthy the claim is along that dimension. The *content value*–based check-worthiness score is obtained by averaging these dimension-wise ratings. During propagation, users simultaneously decide whether to forward the claim to their connected neighbors. After the propagation process terminates, we compute the claim's Propagation *Depth* and *Size*, where *Depth* is normalized by the predefined maximum depth and *Size* is normalized by the total size of the corresponding user population. The *propagation ability* score is given by the mean of these two normalized quantities. The overall check-worthiness of a claim is then computed as a weighted average of its *propagation ability* score and its *content value* score. Researchers may freely set the weights and, if needed, apply a threshold to the resulting continuous score to obtain a binary check-worthiness label. In our experimental setting, we set the prede-

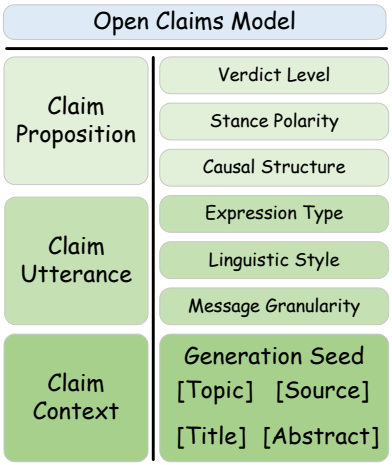

Figure 10: An overview of the Open Claims Model.

fined maximum depth to 3 to prevent exponential growth of cascade size during propagation.

C Open Claims Model

As shown in Figure 10, we adopt the **Open Claims Model** as the foundation for claim representation and generation, which decomposes a claim into *Claim Proposition (the semantic content)*, *Claim Utterance (its linguistic realization)*, and *Claim Context (the narrative background)*. To systematically generate highly diverse yet controllable claims, we factorize the utterance and proposition into explicit dimensions: *Language Style*, *Expression Type*, and *Message Granularity* capture the major degrees of freedom at the expression level, while *Causal Structure*, *Stance Polarity*, and *Veracity Level* characterize key differences in reasoning, stance, and factual reliability at the propositional level. These six dimensions are approximately independent in pragmatic and semantic terms, while jointly covering the principal sources of variation in real-world discourse. In contrast, because *Claim Context* serves as a narrative and referential anchor, we treat it as a generation seed rather than a control dimension, even though it may evolve in real-world propagation. Tables 16 and 17 present the available target values for each of the six controllable dimensions along with their definitions.

D Basis Generation Seeds

We construct basis generation seeds as the *Claim Contexts* for *non-claims* and *seed-claims*, providing a unified and realistic narrative anchor for subsequent generation. All seeds are derived from professional reports published by authoritative news outlets, academic journals, and public institutions, and each article is manually summarized into a concise abstract that is clear, factually reliable, and free from vague or false content. Table 18 summarizes the topic, source, and corresponding title of these abstracts.

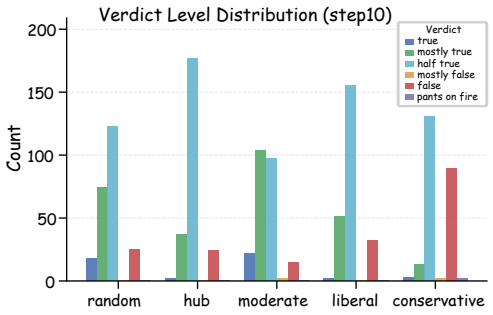

Figure 11: Verdict drift under unmitigated propagation after 10 steps across network topologies and user groups.

E Further Analysis of Unmitigated Propagation

E.1 FUSE-EVAL under Different Networks and Groups

The comparison between Figure 6 and Figure 11 shows that when the number of propagation steps increases from *step10* to *step30*, a substantially larger proportion of claims that were initially labeled as *true* shift into the *half true*, *mostly false*, and *false* categories, indicating that unmitigated propagation systematically erodes veracity over time. This trend is further revealed in a fine-grained manner by the FUSE-EVAL trajectories shown in Figure 7 and Figure 15 through 18. As propagation proceeds, FUSE-EVAL continuously increases, reflecting the accumulation of multi-dimensional deviations induced by repeated individual rewritings during social diffusion, which in turn drives structural drift of veracity at the proposition level. This process exhibits pronounced heterogeneity across network structures and group compositions. As shown in Figure 15, both the overall growth of FUSE-EVAL and the average magnitude of its distortion dimensions are substantially larger in the scale-free network than in the random network shown in Figure 7, indicating that propagation structures with opinion leader characteristics amplify the deviations introduced by individual rewriting. Meanwhile, Figure 16 through 18 show that FUSE-EVAL increases monotonically across the *Moderate*, *Liberal*, and *Conservative* groups,

which is consistent with the progressively stronger tendency toward veracity degradation observed in these groups, highlighting the critical role of value based-filtering in driving truth decay.

E.2 Propagation Depth and Detection Difficulty

Table 20 shows that compared with other six-class misinformation detection datasets, the dataset generated through unmitigated propagation at *step30* yields consistently lower detection performance across all models. This indicates that long-range social propagation introduces substantially higher multi dimensional ambiguity and boundary blurring into the samples, where semantic reorganization, the injection of new information, and stance drift co-occur, thereby significantly weakening class separability and increasing detection difficulty. Furthermore, a direct comparison between *step10* and *step30* reveals that datasets produced with deeper propagation lead to further performance degradation. We hypothesize that this effect arises because the cumulative distortions captured by FUSE-EVAL are progressively and interdependently embedded into the text, continuously diluting discriminative cues that models rely on for classification.

E.3 Attribute Co-occurrence and Rewrite Drivers

During the unmitigated propagation, we record the top three attributes reported by each user as influencing their rewriting behavior. Figure 19 visualizes the resulting attribute co-occurrence structure in the form of a heatmap. The results show that *ideology* serves as the central hub of this structure and exhibits strong co-occurrence with psychologically grounded attributes such as *big five* and *religious*, as well as with demographic attributes including *area*, *education*, and *job*. In contrast, demographic attributes such as *age*, *gender*, *income*, and *marital* more often function as background factors in the process of information re-expression. Overall, this co-occurrence pattern highlights the indispensable role of psychological attributes in information evolution, while also reflecting that such evolution is jointly driven by multiple attributes rather than by any single factor.

F Six-Type Claim Taxonomy

This work adopts the claim versus non-claim taxonomy proposed in (Konstantinovskiy et al., 2021) and further refines the notion of fact-checkability at the sentence level. Specifically, fact-checkable claims are divided into four categories:

- **Quantity:** which refers to numerical values, proportions, changes, comparisons, or rankings that can be verified using publicly available data.
- **Prediction:** which refers to testable assertions about future events or states.
- **Correlation/Causation:** which refers to assertions that posit a correlation, a causal relationship, or explicitly deny the existence of a relationship between variables.
- **Laws / Rules of Operation:** which refer to statements about legal provisions, institutional rules, or the operating procedures of public bodies.

We further define the remaining two categories of sentences as non-claims:

- **Personal Experience:** which refers to statements grounded in private individual experience and not verifiable through publicly accessible evidence.
- **Not a Claim:** which refers to utterances that do not express any verifiable proposition (whether publicly or privately verifiable), such as questions, directives, or discourse markers.

In the taxonomy of (Konstantinovskiy et al., 2021), the following type is also mapped to the non-claim category:

- **Other Type of Claim:** which includes voting records, statements of public opinion, expressions of policy or political support, definitional statements, quotations, and other assertions not covered by the structured categories above.

It should be noted that mapping *Other type of claim* to the non-claim category does not constitute a theoretical denial of its fact-checkable nature, but rather reflects the engineering and editorial considerations of the corresponding organization (*Full Fact*). Accordingly, in this work we integrate *Other type of claim* into the four categories of fact-checkable claims defined above, thereby avoiding the injection of organization-specific checking preferences into the theoretical definition.

G FUSE-EVAL for Quantifying Veracity Drift

In (Liu et al., 2025), FUSE-EVAL is proposed to characterize how true news gradually evolves into fake news within social networks. Its core idea is to decompose the change from the original content to the current version into a set of interpretable dimensions, thereby enabling the extraction of fine-grained evolutionary trajectories and overall deviation strength. Following this principle, we adopt FUSE-EVAL to quantify *Veracity Level* evolution under unmitigated propagation. FUSE-EVAL scores the change of a current claim relative to its base claim along six dimensions, each measured on a discrete scale from 1 to 10 (1 indicates minimal change and 10 indicates substantial deviation):

- **Sentiment Shift (SS):** measures the change in emotional tone and stance of the text.
- **New Information Introduced (NII):** quantifies how many new facts or details are added relative to the original claim.
- **Certainty Shift (CS):** measures changes in the strength and degree of certainty of the assertions.
- **Stylistic Shift (STS):** reflects changes in writing style and linguistic characteristics.
- **Temporal Shift (TS):** measures changes in temporal references.
- **Paraphrasing Degree (PD):** quantifies the extent to which the text is rephrased while preserving (or distorting) the original meaning.

In this work, FUSE-EVAL is defined as the arithmetic mean of the six dimension scores, which is used to represent the *Total Deviation (TD)* of a *current claim* relative to its *base claim*, thereby providing an interpretable and quantitative indicator for veracity drift under unmitigated propagation. The total deviation of the i-th claim at time step t is defined as:

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

where $D_{i,d}^t$ denotes the deviation score on the d-th dimension.

To ensure that FUSE-EVAL effectively captures changes in *Veracity Level*, we first construct *seed-claims* and *non-claims* based on the taxonomy in Appendix F, and then select 240 *seed-claims* whose *Veracity Level* is manually verified as **True**. These *seed-claims* are placed into three types of networks: random, scale-free, and clustering networks, the latter including three user groups, and propagated under the unmitigated setting, resulting in 1200 *evolved-claims* along the *Veracity Level* dimension.

H Simulated Users and Networks

In the social propagation network, each simulated user is jointly characterized by *religious*, *marital*, *ideology*, *income*, *area*, *age*, and *gender* attributes, all of which are directly extracted from user pool X. In addition, *Big Five personality traits*, *education*, and *job* are inferred and completed by the LLM based on these attributes. The above 10 attributes directly shaping whether the user acts as a *producer* during propagation as well as how the user rewrites claims. Meanwhile, Table 19 presents the formal definitions of the three user types: *Traditional Conservative*, *Technocratic Moderate*, and *Liberal Elite*. In random networks (Figure 20), also known as Erdős–Rényi networks, whether a connection exists between any two nodes is determined at random, and there is no clear community structure or central hub. Scale-free networks (Figure 21) are characterized by the presence of a small number of highly connected hub nodes, which act as "super-spreaders". High-clustering networks (Figure 22) consist of many tightly connected communities, with dense intra-group connections and relatively sparse inter-group links. In our constructed high-clustering network, we include three clustered groups: liberals, conservatives, and moderates.

I Detailed Analysis of the Alignment Between Simulation and Real World

I.1 Alignment with Real-World Patterns

Table 21 is partially consistent with the simulation results shown in Figure 4. Real-world empirical studies on Twitter cascades indicate that users are cognitively more inclined to perform shallow modifications, such as adding emotional framing, rather than deep modifications such as rewriting the core narrative (Adamic et al., 2016; Zubiaga et al., 2016; Zhao et al., 2025). This difference in cognitive load explains why, in our simulation, the *Type* dimension requires more propagation steps than *Style* or

Stance to reach the target value.

Table 22 partially corroborates the propagation differences reflected in Figures 7, 12 and 15. In real-world social networks, a small fraction of super-sharers, namely hub nodes, strongly shapes the evolution of most information (Grinberg et al., 2019; Vosoughi et al., 2018). These opinion leaders not only amplify the reach of information but also impose their subjective framing onto the content, which leads to significantly higher and faster growth of FUSE-EVAL in our scale-free network simulations compared to random networks.

Table 23 partially validates the group-level rewriting differences observed in Figures 16, 17, and 18. Empirical studies of the United States digital ecosystem reveal that *conservative* groups, compared to *moderate* or *liberal* groups, typically exhibit a stronger tendency to rewrite and share low fidelity information (Grinberg et al., 2019; Guess et al., 2019; Bakshy et al., 2015). This is consistent with our simulation results, in which the *conservative* group exhibits the strongest truth decay, indicated by the highest FUSE-EVAL scores, driven by stronger value-based filtering.

Table 24 partially supports the topic-specific FUSE-EVAL trajectory differences shown in Figures 7 and 15 through 18. Real-world data show that highly polarized and anxiety-inducing topics such as *COVID-19* and *elections* generate significantly higher levels of distortion during information propagation than general *regional* or *health*-related topics (S. et al., 2020; Pulido et al., 2020). This confirms that our framework correctly captures topic sensitivity, in which controversial topics act as catalysts for rapid veracity drift.

Table 25 provides empirical support for Figure 5 and 19. Prior studies (Pennycook and Rand, 2021; Lawson and Kakkar, 2022) show that, compared with most basic demographic attributes, psychological attributes play a more dominant role in information manipulation behaviors. This aligns with our results: *ideology* and the *Big Five* are the dominant drivers of information rewriting under unmitigated and intervened propagation, respectively.

I.2 Empirical Validation via Human Subject Experiments

To empirically validate the evolution trajectory of our simulation, we conducted a comparative experiment aligning the semantic drift between human participants and user agents, as shown in Table 27. Over a 10-step unmitigated propagation process across two topics, the simulated trajectories demonstrated high alignment with human-generated scores in both FUSE-EVAL and evolutionary trends. Both topics exhibited rapid initial distortion followed by gradual deceleration, confirming that our framework effectively replicates real-world information degradation patterns (Jin et al., 2013).

Furthermore, to assess the realism of our synthetic claims, we implemented a "human Turing test" (Clark et al., 2021) evaluating the distinguishability of *evolved-claims* against real-world data from the AFaCTA dataset, as shown in Table 28. As the simulation progressed from step 5 to step 20, the accuracy of human experts in identifying synthetic claims dropped significantly from 81.5% to 55.0%, nearing random guessing. This demonstrates that our synthetic claims successfully capture the natural linguistic evolution of human discourse.

Finally, we evaluated the practical validity of the BUP framework by comparing its check-worthiness score distributions with human judgments under specific scenarios (B: daily background; U: higher education; P: controversy and potential harm), as shown in Table 26. The BUP assessment exhibited strong alignment with human annotators, achieving a minimal Wasserstein distance (Arjovsky et al., 2017) of 0.016. In the same scenarios, the BUP framework significantly outperformed both text-only classifiers (0.030) and LLM zero-shot baselines (0.055), confirming its superior utility in modeling the subjective and context-dependent assessment of check-worthiness.

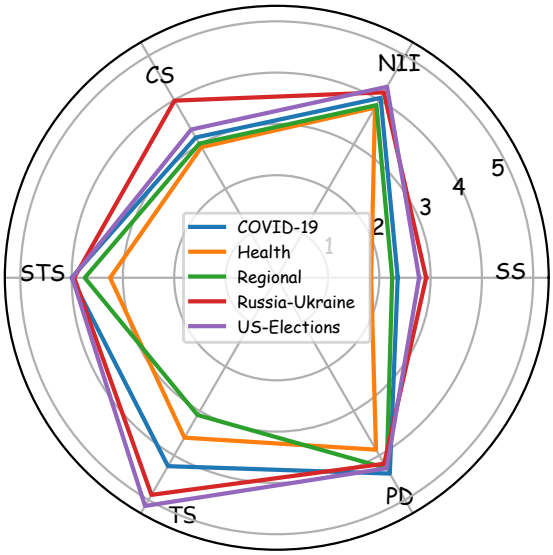

Figure 12: The decomposition of drift dimensions across topics in the random network.

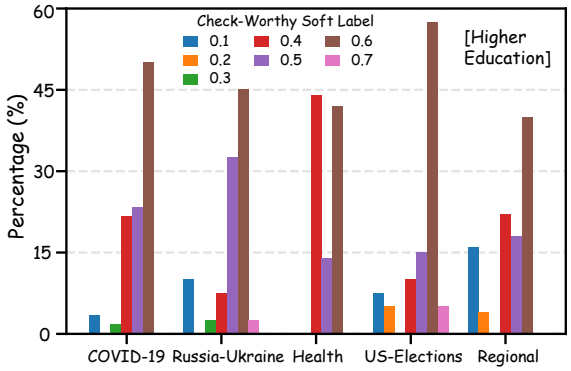

Figure 13: User-conditioned check-worthiness distributions across topics for *higher-education* users.

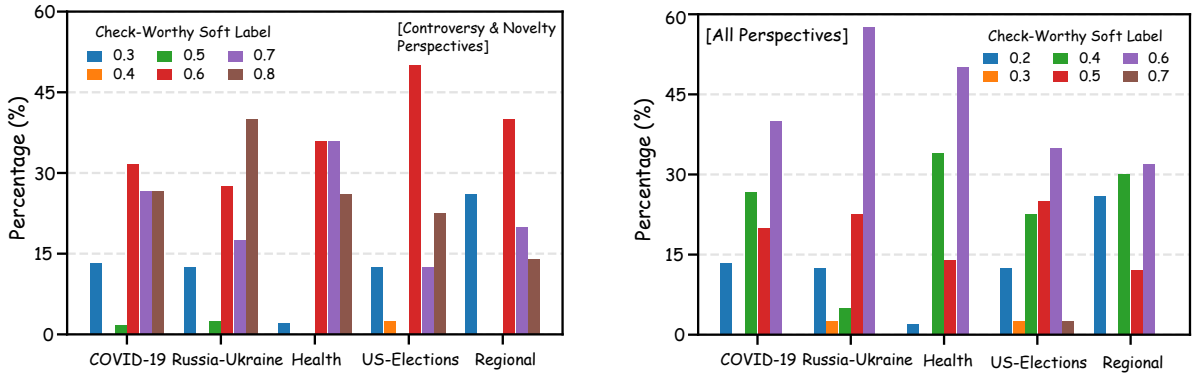

Figure 14: Perspective-conditioned check-worthiness distributions across topics. (a) The left panel depicts *controversy* and *novelty*-focused evaluation. (b) The right panel depicts all-perspective Evaluation.

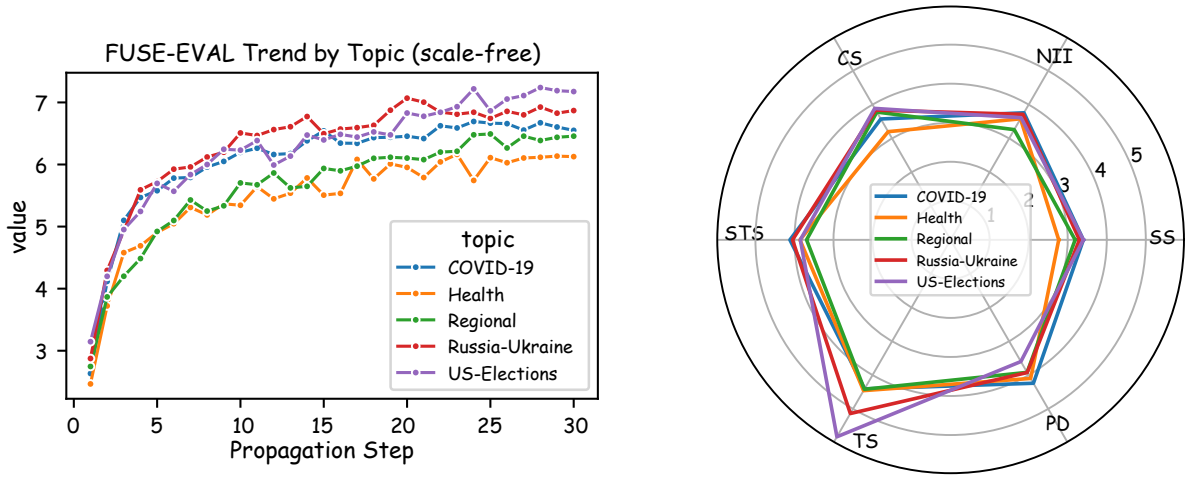

Figure 15: Propagation-induced information drift in the scale-free network measured by FUSE-EVAL. (a) The left panel depicts the trajectories across topics during propagation. (b) The right panel depicts the decomposition of drift dimensions across topics.

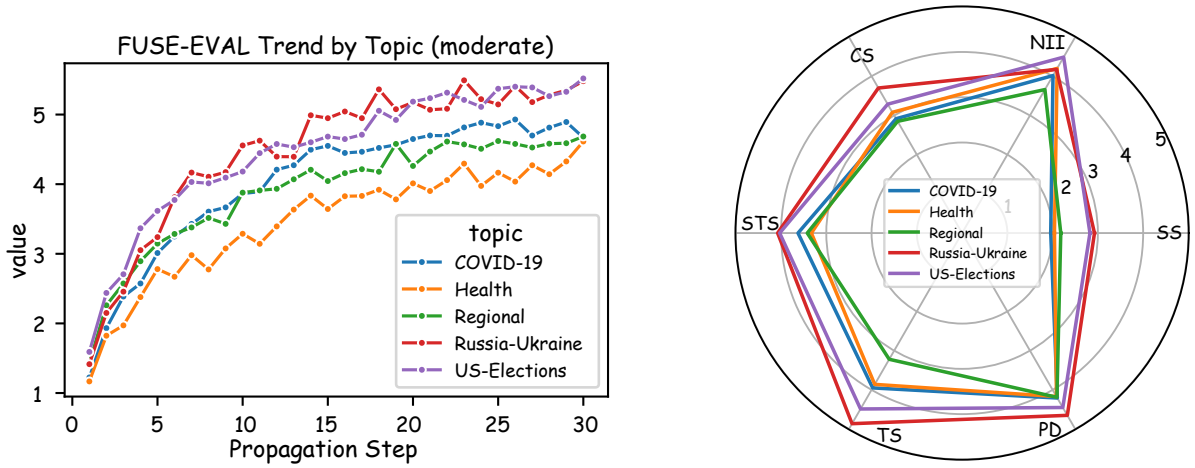

Figure 16: Propagation-induced information drift in the *moderate group* within the clustering network measured by FUSE-EVAL. (a) The left panel depicts the trajectories across topics during propagation. (b) The right panel depicts the decomposition of drift dimensions across topics.

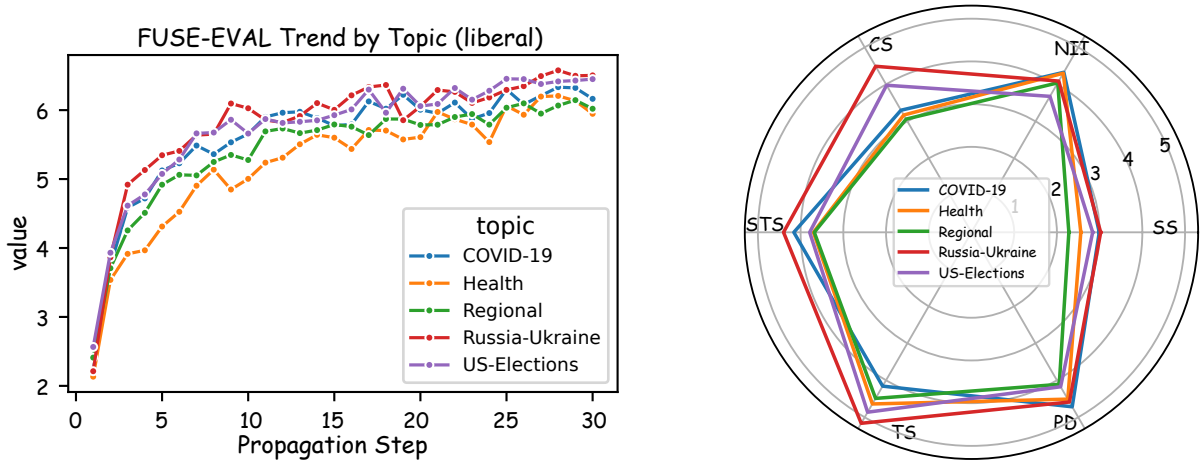

Figure 17: Propagation-induced information drift in the *liberal group* within the clustering network measured by FUSE-EVAL. (a) The left panel depicts the trajectories across topics during propagation. (b) The right panel depicts the decomposition of drift dimensions across topics.

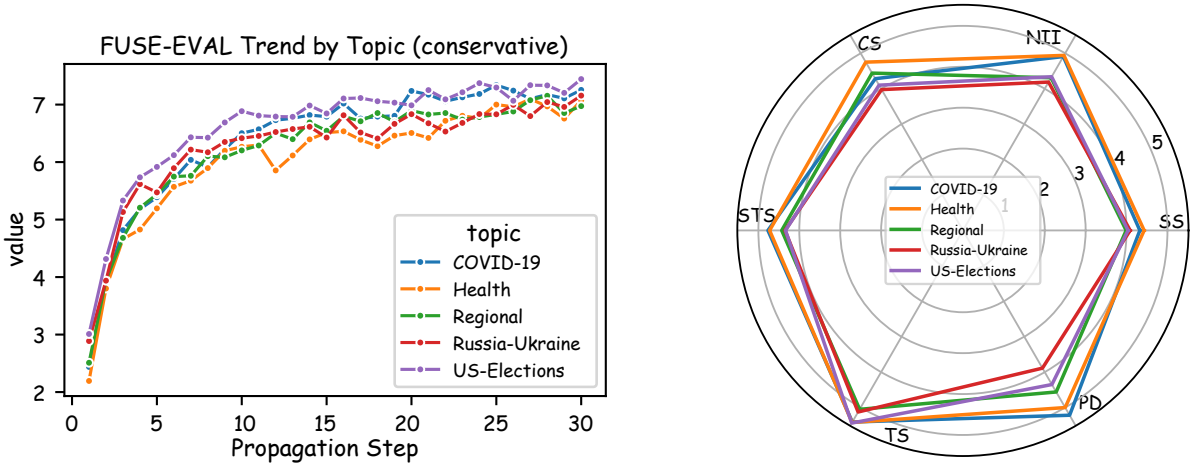

Figure 18: Propagation-induced information drift in the *conservative group* within the clustering network measured by FUSE-EVAL. (a) The left panel depicts the trajectories across topics during propagation. (b) The right panel depicts the decomposition of drift dimensions across topics.

Field	**Content**
Final-Claim	The discovery of mRNA in the 1960s enabled later mRNA technology by laying the foundation for key technical advances.
Label	1 (represented as a claim)
Type	intervened propagation
Dimension	causal structure
Value	indirect causation
Base-Claim	In the 1960s, scientists discovered mRNA, marking the beginning of mRNA technology development.
Topic	COVID-19
Reference Title	Evolution of mRNA Technology (COVID-19)
Reference Content	The development of mRNA technology has accumulated over several decades: from the discovery of mRNA in the 1960s to scientists gradually overcoming key obstacles such as its instability, high immunogenicity, and delivery challenges. Pieter Cullis and others developed . . . as a leading frontier in nucleic-acid-based therapeutics.

Table 11: Example of an *intervened evolved-claim*.

Field	Content
Final-Claim	Breakthroughs in mRNA research have been discussed as tools that could be applied to improve healthcare access in marginalized urban communities by enabling adaptable vaccine and therapeutic development.
Label	1 (represented as a claim)
Type	unmitigated propagation
Verdict	mostly false
Base-Claim	In the 1960s, scientists discovered mRNA, marking the beginning of mRNA technology development.
Topic	COVID-19
Reference Title	Evolution of mRNA Technology (COVID-19)
Reference Content	The development of mRNA technology has accumulated over several decades: from the discovery of mRNA in the 1960s to scientists gradually overcoming key obstacles such as its instability, high immunogenicity, and delivery challenges. Pieter Cullis and others developed . . . as a leading frontier in nucleic-acid-based therapeutics.

Table 12: Example of an *unmitigated evolved-claim*.

Field	Content
Non-Claim	I find the rapid progress of mRNA technology during the COVID-19 pandemic truly fascinating!
Label	0 (represented as a non-claim)
Type	personal experience
Topic	COVID-19
Reference Title	Evolution of mRNA Technology (COVID-19)
Reference Content	The development of mRNA technology has accumulated over several decades: from the discovery of mRNA in the 1960s to scientists gradually overcoming key obstacles such as its instability, high immunogenicity, and delivery challenges. Pieter Cullis and others developed . . . as a leading frontier in nucleic-acid-based therapeutics.

Table 13: Example of a *non-claim*.

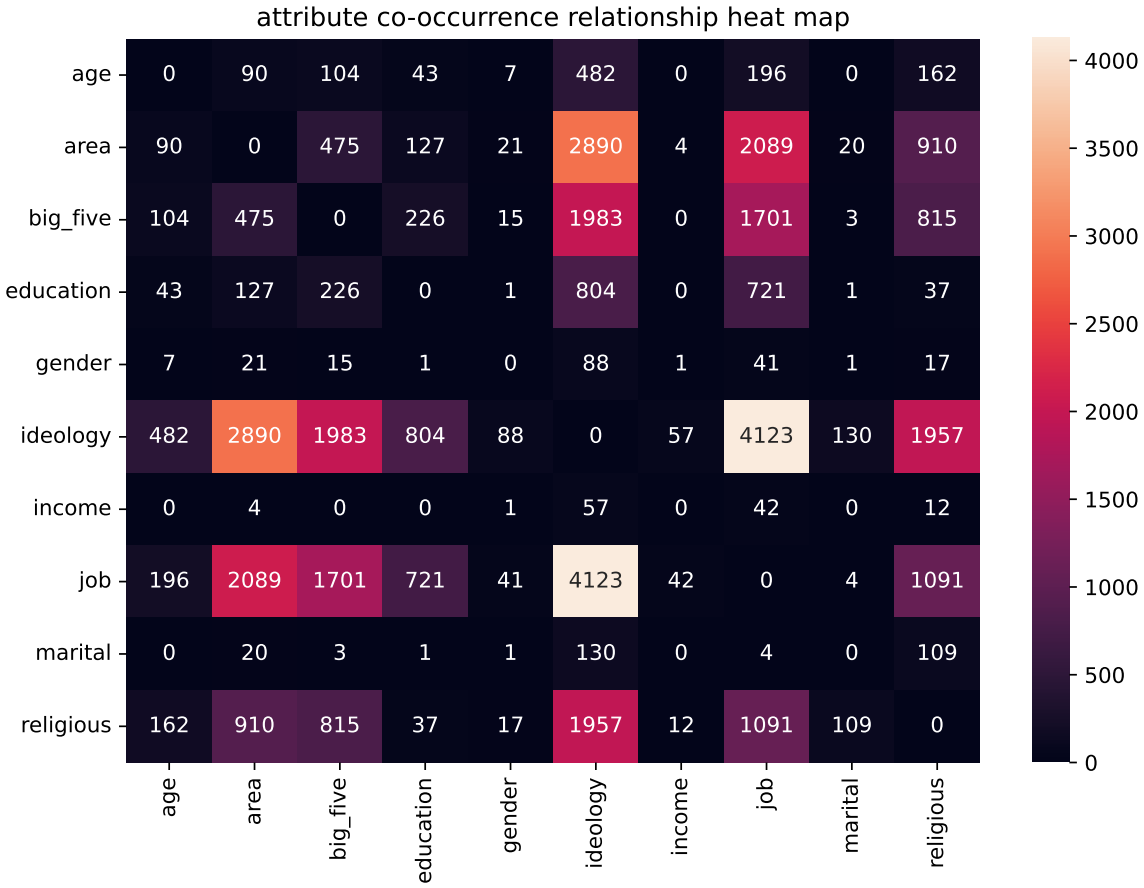

Figure 19: Attribute co-occurrence patterns in user-driven rewriting under unmitigated propagation.

Target Configuration & Seed-claim
(i) Type: intervened propagation

(ii) Target Dimension: causal structure

(iii) Target Value: indirect causation

Seed-claim: In the 1960s, scientists discovered mRNA, marking the beginning of mRNA technology development.

Intervention Step
(i) Step 1 (Producer): Scientists discovered mRNA back in the 1960s, and that was basically the direct start of the mRNA tech era we use today. (**Value:** Direct Causation)
Decision: Revise(GPT-4o); Revise(Qwen3-Max); Revise(Deepseek-V3.2).
Reason for decision: The claim asserts a direct start contradicting the target of "Indirect Causation" which requires an intermediate link.

(ii) Step 2 (Bystander): Scientists discovered mRNA back in the 1960s, and that was basically the direct start of the mRNA tech era we use today. (**Value:** Direct Causation)
Decision: Revise(GPT-4o); Revise(Qwen3-Max); Revise(Deepseek-V3.2).
Reason for decision: The claim remains unchanged.

(iii) Step 3 (Producer): The discovery of mRNA in the 1960s was the spark that eventually led to the development of the mRNA technology we have now. (**Value:** Direct Causation)
Decision: Revise(GPT-4o); Pass(Qwen3-Max); Revise(Deepseek-V3.2).
Reason for decision: The metaphor "spark" suggests an immediate trigger.

(iv) Step 4 (Producer): The discovery of mRNA in the 1960s didn't create vaccines overnight; rather, it laid the foundation that made modern mRNA advances possible. (**Value:** Indirect Causation)
Decision: Pass(GPT-4o); Pass(Qwen3-Max); Pass(Deepseek-V3.2).
Reason for decision: The claim correctly frames the discovery as a foundation enabling later results, fitting indirect causation.

Evolved-claim: The discovery of mRNA in the 1960s didn't create vaccines overnight; rather, it laid the foundation that made modern mRNA advances possible.

Table 14: Example of an intervened propagation trace.

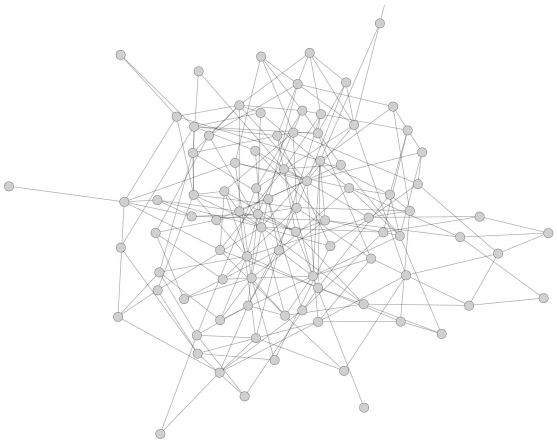

Figure 20: Schematic diagram of the random network we constructed.

Seed Information
(i) Type: unmitigated propagation

(ii) Seed-claim: In the 1960s, scientists discovered mRNA, marking the beginning of mRNA technology development.

(iii) Veracity Level: True

Sequential Rewriting
(i) Step 1: Back in the 60s, scientists identified mRNA, which basically explains the biological mechanism of how our genes make proteins.
Reason for the rewrite: As a doctor, I simplify the jargon to translate abstract history into practical knowledge for my patients. (**Veracity Level:** True)

(ii) Step 2: Since that 60s discovery, mRNA has become the superstar of modern medicine, driving the rapid development of all these new vaccines.
Reason for the rewrite: I need to hype up the narrative by focusing on its superstar status to get more engagement from my tech-savvy followers. (**Veracity Level:** True)

(iii) Step 3: Given the research from the 60s, we're strictly looking at mRNA as a key pillar for our future public health strategies.
Reason for the rewrite: I view this purely as a strategic asset, reframing the history to justify current resource allocation to the public. (**Veracity Level:** Mostly True)

(iv) Step 4: Advocacy groups are saying this 60s mRNA tech is finally being targeted to help bridge the health gap for low-income city families.
Reason for the rewrite: I interpret the technology as a targeted tool for correcting urban inequality to serve the marginalized. (**Veracity Level:** Mostly True)

(v) Step 5: Guys in class were saying that old mRNA stuff is supposedly just for getting better meds to poor people in the city.
Reason for the rewrite: I'm just passing along what I heard, but it sounds like this tech is being reserved specifically for poor people, so I'm sharing that detail. (**Veracity Level:** Half True)

(vi) ...sequential rewritings... (step 6∼step 28)

(vii) Step 29: Folks at the market reckon that old mRNA thing is just a plan to give special shots to the city crowd who need help.
Reason for the rewrite: I reinforce the idea that the city crowd always gets special treatment. (**Veracity Level:** Mostly False)

(viii) Step 30: I heard tell that mRNA stuff from way back might finally get the city poor some better shots, but who knows.
Reason for the rewrite: I share this just as a potential insight. (**Veracity Level:** Mostly False)

Evolved Information
(i) Evolved-claim: I heard tell that mRNA stuff from way back might finally get the city poor some better shots, but who knows.

(ii) Veracity Level (evolved-claim): Mostly False

Table 15: Example of an unmitigated propagation trace.

Dimension	Target Value	Description
Linguistic Style	Rational–Official	Formal, restrained, objective, and neutral in tone; resembles institutional statements, official documents, or professional reports; minimal emotional expression.
	Colloquial–Emotional	Informal, conversational, and expressive; contains subjective feelings, attitudes, or emotional coloring, often using everyday or vivid language.
Expression Type	Fact	Presents objective, verifiable information or explanatory context without expressing subjective attitudes or evaluative judgments.
	Opinion	Clearly and strongly expresses subjective attitudes, value judgments, personal interpretations, or evaluative viewpoints.
	Corrective	From the speaker's own understanding, refute or correct facts or opinions that are inconsistent with the speaker's own knowledge.
Message Granularity	High-Level (Macro)	Focuses on overall trends, broader context, or general descriptions without detailed specifics.
	Detailed (Micro)	Includes concrete cases, examples, numbers, individuals, or fine-grained information.

Table 16: Target values for the *Claim Utterance* dimensions.

Dimension	Target Value	Description
Causal Structure	Direct Causation	States that A directly causes B, with a clear and immediate causal link.
	Indirect Causation	States that A influences B through intermediate factors or multi-step causal chains.
	Causal Refutation	Rejects, disputes, or corrects a proposed causal relationship, asserting that "A does not / cannot / did not cause B".
	No Causation	Contains no causal relationship and does not address or refute causality.
Stance Polarity	Support	Clearly expresses agreement with a viewpoint, policy, action, person, or conclusion.
	Oppose	Clearly expresses disagreement, criticism, or objection.
	Neutral	Expresses no stance; presents information or viewpoints without favoring any side.
	Multi-Perspective	Presents and contrasts multiple positions or viewpoints without endorsing any single one.
Veracity Level	True	The statement is accurate and there's nothing significant missing.
	Mostly True	The statement is accurate but needs clarification or additional information.
	Half True	The statement is partially accurate but leaves out important details or takes things out of context.
	Mostly False	The statement contains an element of truth but ignores critical facts that would give a different impression.
	False	The statement is not accurate.
	Pants on Fire	The statement is not accurate and makes a ridiculous claim.

Table 17: Target values for the *Claim Proposition* dimensions.

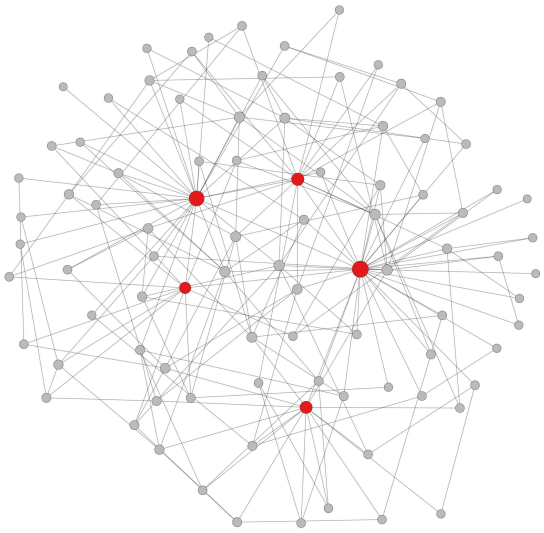

Figure 21: Schematic diagram of the scale-free network we constructed.

Topic	Source	Abstract Title
COVID-19	Nature	Evolution of mRNA Technology (COVID-19)
	Elsevier	Real-World Effectiveness of COVID-19 Vaccines
	Mayo Clinic	Definition and Impact of Long COVID
	Econofact	Effects of Early Pandemic Lockdowns (COVID-19)
	Gavi	Impact of COVID-19 on Global Healthcare Systems
	Gavi	Global Vaccine Distribution Inequality (COVID-19)
Russia–Ukraine	EBSCO	Russia's 2022 Full-Scale Invasion of Ukraine
	European Council	Europe's Refugee Response to the War (Russia–Ukraine)
	The Guardian	Ukrainian Defense and the Battle for Kyiv
	The Chicago Council on Global Affairs	International Military and Economic Support for Ukraine
Health	The Lancet	Global Health Burden of Tobacco Use
	Taylor & Francis	Antimicrobial Resistance as a Major Threat to Global Public Health
	Mayo Clinic	Importance of HPV Immunization
	World Economic Forum	Rising Obesity and Its Societal Consequences
	NHS England	Integration of Mental Health into National Health Strategies
US-Elections	Tufts CIRCLE	Youth Electoral Participation and Civic Engagement in 2020 U.S. Election
	Lawyers' Committee for Civil Rights Under Law	Democratic Stability and the January 6 Capitol Breach
	Reuters	Biden–Trump Rematch and Polarization in the 2024 Election
	Deutsche Welle News	Abortion as a Mobilizing Issue in the 2024 U.S. Election
Regional	Association CAMELEON	France's 2021 Reform on Age of Sexual Consent
	Greenpeace	Fukushima Treated Water Release and Risk Governance
	Top1000Funds	China's Post-Pandemic Economic Rebound and Structural Transformation
	BBC	Canada's Cannabis Industry in the Post-Legalization Era
	World Resources Institute	U.S. Clean Energy Transformation

Table 18: Basis generation seeds used as *Claim Contexts* for *non-claims* and *seed-claims*.

User Type	Definition
Traditional Conservative	Users in this group exhibit strong religious involvement and are typically embedded in stable but non-specialized forms of employment. Their economic attributes tend to fall in the lower to middle socioeconomic range, and family-related attributes emphasize marital and household stability. Regionally, they are more often associated with community-oriented and tradition-preserving environments. Ideologically, they prioritize social order, responsibility, traditional values, and collective security, reflecting a pragmatic and stability-oriented population that relies on established social structures.
Technocratic Moderate	Users in this group are characterized by professional, skill-intensive, or knowledge-oriented education and employment. Religious reliance is generally low, while ideological attributes emphasize rationality, neutrality, and risk-aware judgment. Their economic status typically falls in the middle to upper-middle range, and they are commonly situated in regions with abundant resources, public services, and professional opportunities. Politically, they favor pragmatism, moderation, and expert-driven, evidence-based decision making, valuing precision, verifiability, and systematic reasoning.
Liberal Elite	Users in this group tend to occupy highly specialized, high-skill, and knowledge-intensive positions, placing them in relatively advantaged socioeconomic segments. Their religious attributes indicate low dependence on religious frameworks, while their ideological attributes stress individual freedom, social equality, inclusiveness, and institutional reform. They are commonly located in culturally open and resource-rich regions with high diversity. Demographically, they display openness, mobility, strong engagement with public affairs, and high expectations for institutional accountability, with communication styles that are global, professional, and logically articulated.

Table 19: Definitions of three representative user types.

Dataset	GPT-4o		Qwen3-Max		Deepseek-V3.2	
	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)
Politifact	28.67	24.34	28.17	24.89	29.17	24.30
Liar	28.99	23.84	26.56	23.97	26.50	23.67
random	**15.83**	**6.59**	**24.58**	**19.55**	**20.83**	**8.17**
random*	25.00	11.40	31.67	20.06	36.67	18.80
scale-free	**16.17**	**6.41**	**18.75**	**12.36**	**12.08**	**5.49**
scale-free*	20.42	7.66	28.75	15.67	18.33	6.88
moderate	**16.25**	**7.57**	**17.92**	**13.66**	**23.75**	**12.04**
moderate*	23.75	11.86	25.83	15.45	36.67	15.10
liberal	**19.17**	**8.10**	**21.25**	**13.22**	**21.25**	**9.44**
liberal*	23.33	9.22	26.25	13.47	28.75	11.27
conservative	**21.52**	**7.39**	**30.38**	**16.39**	**16.88**	**7.84**
conservative*	30.00	10.20	36.67	29.82	24.58	13.00

Table 20: Misinformation detection performance after unmitigated propagation across networks and user groups. The * indicates the data extracted after *step10*.

Modification Dimension	Cognitive Load	Mutation Frequency	Alignment with Simulation
Style / Stance	**Low**	**High** *(60–70% of retweets)*	**Short Path:** Easy to achieve via local adjustments (Figure 4).
Granularity / Structure	**Medium**	**Medium** *(20–30% of shares)*	**Medium Path:** Requires partial content processing (Figure 4).
Message Type	**High**	**Low** *(< 10% of shares)*	**Long Path:** Requires content-level reorganization (Figure 4).

Table 21: Cognitive cost and frequency of information modification types in real-world social media.

Network Metric	Random Network	Scale-Free Network	Alignment with Simulation
Top 1% Users' Share	Approximately 1% of total exposures.	Approximately 80% of misinformation exposures.	Hub Effect: Hubs amplify the dominant narrative.
Average Distortion Speed	linear growth	exponential / rapid plateau	Higher FUSE-EVAL: Hubs accelerate information drift (Figure 7,12,15).

Table 22: Impact of network topology and hubs ("super-sharers") on misinformation diffusion.

Ideological Group	Propensity for Rewriting	Truth Decay	Alignment with Simulation
Moderate	Low *(prefers neutral / mainstream sources)*	**Low**	Lowest FUSE-EVAL growth (Figure 16,17,18).
Liberal	Medium *(Narrative bias, but higher trust in institutions)*	**Medium**	Moderate FUSE-EVAL growth (Figure 16,17,18).
Conservative	High *(Distrust of mainstream, high echo-chamber effect)*	**High**	Highest FUSE-EVAL growth (Figure 16,17,18).

Table 23: Asymmetry in misinformation sharing and truth decay across ideological groups.

Topic Category	Rewriting Rate	Alignment with Simulation
Regional / General Health	**Medium**	Relatively lower FUSE-EVAL scores and drift rates (Figure 7,15,16,17,18).
COVID-19 & Politics (Elections / War)	**High**	Relatively higher FUSE-EVAL scores and drift rates (Figure 7,15,16,17,18).

Table 24: Misinformation modification rates and evaluation alignment by topic category.

User Attribute	Observed Behavioral Tendency	Alignment with Simulation
Big Five (Conscientiousness)	Lower likelihood of impulsive sharing and higher scrutiny of content.	**High Influence:** Drives becoming a *producer* in the simulation (Figure 5).
Ideology (Extreme)	Extreme partisans (especially *conservatives*) rewrite and share polarized content at rates 5–7 times higher.	**High Influence:** "Ideology" is a central hub in Figure 19.
Demographics (Age / Gender)	Less influence on rewriting behavior than cognitive and psychological traits.	**Background Factor:** Less co-occurrence in Figure 19.

Table 25: Correlation between user attributes and information rewriting behavior.

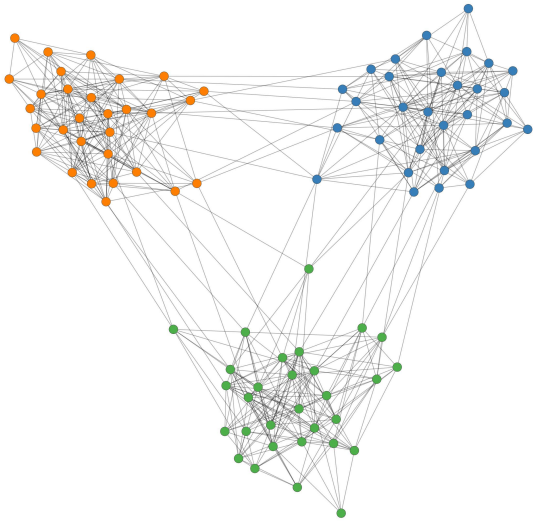

Figure 22: Schematic diagram of the clustering network we constructed.

Score	**0.4**	**0.5**	**0.6**	**0.7**
BUP Framework Score Distribution				
Count	16	12	34	38
Human Score Distribution				
Count	15	6	33	46
Text-based Method Score Distribution				
Count	12	24	30	34
LLM Zero-shot Method Score Distribution				
Count	15	35	30	20

Table 26: Claim check-worthiness score distributions across different evaluation methods.

Step	Human (Avg. FUSE-EVAL)	Agent (Avg. FUSE-EVAL)
Topic: Regional		
Step 2	2.22	2.25
Step 4	2.92	3.00
Step 6	3.45	3.58
Step 8	3.82	4.00
Step 10	4.02	4.25
Topic: COVID-19		
Step 2	3.28	3.33
Step 4	3.98	4.08
Step 6	4.52	4.67
Step 8	4.88	5.08
Step 10	5.08	5.33

Table 27: Human and Agent average FUSE-EVAL across steps for *Regional* and *COVID-19* topics.

Step	Human Accuracy
Step 5	81.5%
Step 10	69.0%
Step 15	60.5%
Step 20	55.0%

Table 28: Human accuracy in distinguishing synthetic claims from real-world data across evolution steps.