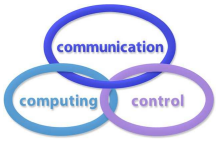
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 21, Issue: 6, Month: December, Year: 2026
Article Number: 7470, https://doi.org/10.15837/ijccc.2026.6.7470



CCC Publications



A Rule and Evidence Enhanced Framework to Dynamic Suppression of Health Misinformation Propagation

J.Y. Xiong, H.X. Tang, W. Lu, L.Y. Yao

Jianying Xiong
College of Intelligent Medicine and Information Engineering
Jiangxi University of Chinese Medicine
330004, Nanchang, Jiangxi, China
special8212@sohu.com

Haoxiang Tang
School of Business Administration
Jiangxi University of Finance and Economics
330013, Nanchang, Jiangxi, China
865843642@qq.com

Wei Lu
Center of Collaboration and Innovation
Jiangxi University of Technology
330098, Nanchang, Jiangxi, China
vluwei@163.com

Leiyue Yao*
College of Intelligent Medicine and Information Engineering
Jiangxi University of Chinese Medicine
330004, Nanchang, Jiangxi, China
*Corresponding author: leiyue_yao@163.com

Abstract

Social networks are vital for public health information access, but they also accelerate the spread of high-risk or false content. Existing moderation approaches lack the precision and adaptability. We propose a rule and evidence enhanced risk assessment and risk-based dynamic communication decision-making framework to suppress the spread of health misinformation. By constructing a hierarchical evidence knowledge base, it leverages a knowledge-augmented large language model (LLM) to assess information risk along three fine-grained dimensions: logical consistency, source reliability, and semantic ambiguity. Propagation permissions of each node are dynamically adjusted by integrating risk score with node credibility and influence. Experimental results demonstrate that, compared to a general-purpose LLM, our method improves accuracy in identifying high-risk content and achieves a lower false-negative rate. In heterogeneous social network simulations, our framework achieves near-complete suppression of harmful misinformation without inadvertently blocking benign information. This work presents a computable, and explainable solution for balancing the free diffusion of health information and the prevention and control of risky information.

Keywords: Health misinformation, Dynamic suppression, Evidence and rules enhanced reasoning, Risk aware propagation.

1 Introduction

Social media has become a primary channel for public access to health information. However, its openness accelerates the viral spread of high-risk or misleading health content. The World Health Organization (WHO) has identified "infodemics" as a significant global public health threat. Because misinformation can lead to delayed medical care, improper medication use, and even mass health crises [1]. The elderly population is particularly vulnerable. Studies have shown that they often accept misinformation to avoid health risks rather than doubt it, leading to frequent sharing of false health information[2]. Therefore, it is important to develop governance mechanisms that can accurately identify and suppress high-risk content without hindering the flow of necessary social information [3].

While the identification of health misinformation shares similarities with rumor detection, it possesses distinct professional characteristics. Current approaches to misinformation governance fall into three main categories. The first focuses on content security, employing keyword filtering systems, text classification, and emotion recognition [4]. However, these methods are difficult to deal with semantically variant and context-dependent misleading content, often leading to controversy. The second perspective emphasizes social network trust, utilizing dissemination control strategies based on user reputation or social graphs. For instance, Yue *et al.* proposed heterogeneous game and reputation based dynamic threshold methods, which suppress rumor spread through node permissions [5]. Such methods rely heavily on the historical behavior data, making them susceptible to failure when faced with newly registered accounts, abnormal behaviors, or insufficient professional recognition capabilities. Recently, Large Language Models (LLMs) have facilitated automated fact-checking and demonstrated potential in understanding complex medical statements [6]. Nevertheless, LLMs are prone to "hallucinations" in professional fields [7].

In summary, existing research has primarily focused on verifying the veracity of information itself. However, the user adoption of health information involves fine-grained risk reasoning rather than simple truth classification. Similarly, social networks approaches relying solely on user reputation are prone to failure. Integrating content risk, node credibility, and network influence into a dynamic permission function reduces reliance on historical data. Furthermore, as demonstrated by Zhao *et al.* [8], evidence types significantly affect people's intention to share and verify health misinformation. Consequently, combining LLM reasoning with a structured evidence base and rule-based risk quantification can enhance both accuracy and interpretability. So, it is meaningful to design a network dynamic health information dissemination governance method that integrates fine-grained semantic risk rules and evidence support.

Unlike prior studies focusing on content review or network topology, this study transforms risk assessment into continuous network control signals through a theoretically grounded dynamic permission function. We establish its key mathematical properties, including monotonicity and asymptotic stability for high-impact nodes. thereby ensuring robust governance that operates independently of historical data. To improve the transparency of LLM inference, we design a hierarchical evidence-based risk quantification mechanism that explicitly decouples logical consistency, source reliability, and semantic ambiguity, effectively mitigating LLM hallucinations through structured reasoning. While conventional RAG approaches are tailored to question-answering tasks, our framework innovatively deploys Retrieval-Augmented Generation for in-depth risk reasoning in health misinformation detection. Moreover, the proposed dynamic permission function moves beyond simple weighting schemes by modeling the Signal-to-Noise Ratio (SNR) dynamics of information propagation, adaptively reconciling content credibility with network-induced distortions. The efficacy of our design is further validated through ablation studies and comparative experiments.

Overall, this study proposes a multi-dimensional dynamic permission control framework for suppressing health misinformation. The contributions of this study include the following. First, we propose a dynamic content permission mechanism that integrates information risk, node reputation, and network influence, which can improve governance efficiency. Second, we construct a hierarchical evidence base and fine-grained risk inference rules that augment LLM reasoning, shifting content evaluation from authenticity verification to harm-oriented assessment via RAG-enhanced inference. Third, we verify the feasibility of dynamic permission mechanism in balancing information freedom and risk prevention. Our method can effectively constrain the illusion of LLM in professional sce-

narios, improve the interpretability of risk judgment, and achieve differentiated forwarding control of "high-risk strong inhibition, low-risk weak intervention".

2 Related work

The governance of online health information is an interdisciplinary research problem that integrates library science, communication studies, and computer science [9]. From different disciplinary perspectives, this task can be defined as information quality evaluation, false information identification, and social network dissemination control. The section systematically reviews representative works in these directions and identifies their limitations in addressing dynamic and fine-grained health information governance.

2.1 Information quality and credibility assessment

Traditional quality assessment relies on static structured scales like DISCERN and GQS, which have proven reliable in clinical consultation and patient education settings [10]. However, these scales are unsuitable for the rapid identification of massive, real-time, unstructured content in social media environments. Consequently, information quality evaluation has shifted towards a dynamic credibility modeling framework that integrates communication attributes and domain knowledge. Existing credibility assessment studies typically extract features from source, content, and dissemination channel. Misinformation often manifests through anonymous or unverified sources, highly emotional language, self-contradictory logic, or lack of supporting evidence [11]. To integrate these heterogeneous features, researchers concatenate text semantic features [2] (e.g., sentiment polarity, readability, fuzzy word frequency), user attributes [12] (e.g., verification status, historical posting credibility), and dissemination dynamical features [13] (e.g., initial forwarding speed, diffusion depth) into vectors for classifier input. While computationally efficient, this feature-concatenation approach neglects semantic relationships between modalities and makes it difficult to capture complex social interaction relationships.

2.2 Content identification algorithms

Content identification algorithms have evolved from rule-based matching to deep learning. Early methods relied on manually constructed keyword libraries or grammar rules. Although fast, they struggle with linguistic variations and novel rumor types [14]. Subsequent research has extensively explored machine learning and deep learning methods, including Naive Bayes, Support Vector Machines, Random Forests, and deep learning models such as CNN and LSTM, as well as pre-trained language models like BERT [15]. For example, Sikosana [16] evaluated the effectiveness of machine learning (ML) and deep learning (DL) models in detecting COVID-19 related misinformation on online social networks. However, such methods simplify the problem into binary classification (true/false), ignoring the continuity and multidimensionality of health information risks.

Traditional information credibility assessment methods often focus on the classification information sources or content [17], without deeply deconstructing the scientific reasoning inherent in information. If causal consistency modeling is ignored, it limits interpretability and its ability to explain the spread of rumors [18]. LLMs are well-suited for fact-checking due to their contextual reasoning and cross-domain generalization capabilities. However, directly applying LLMs to health misinformation risk assessment introduces several issues [19]. The first is factual hallucination, LLMs may generate medical claims that appear reasonable but are incorrect, such as erroneous descriptions of drug interactions or disease mechanisms; The second is Authority bias, Models may overestimate the credibility of academic papers based solely on text format imitation, while overlooking the lack of substantive evidence; The third is insufficient granularity, it can generate insufficient granularity in generating evidence. For example, when there is a lack of clear information on the applicable population, dosage instructions, or contraindication constraints, LLM is difficult to quantify its potential risks [20, 21]. To mitigate these problems, Knowledge-Augmented LLMs have become an important direction. By retrieving evidence from structured knowledge bases, these models can ground their reasoning in verified facts.

For example, Yue *et al.* [5] integrated a dietary supplement knowledge base as an RAG enhancement, improving the accuracy and reliability of related Q&A.

2.3 Network information propagation governance

False information spreads faster and more widely than true information on social platforms [22]. The synergy between social influence and cross-network effects significantly affects diffusion patterns [23]. The research paradigm is shifting towards risk adaptive access control by intervening in key nodes or pathways to suppress diffusion. As proposed by He *et al.* [24], adaptive rumour suppression and multi round hybrid strategy can dynamically adjust intervention intensity; Peng *et al.* [25] dynamically activated suppression mechanisms based on the prevalence of erroneous information. These all support the differentiated forwarding control concept of "high-risk strong suppression, low-risk weak intervention". According to this control concept, researchers have proposed various network behavior-based suppression algorithms. Node reputation is a typical subjective trust model that has been widely used in P2P, recommendation systems, and rumour detection [26]. Whalen *et al.* [27] designed a forwarding filter based on user reputation, allowing only high reputation users to spread sensitive health content. Cui *et al* [28] used Graph Attention Network to aggregate the credibility signals of neighbouring nodes and dynamically adjust information visibility. However, high impact nodes in social networks play a crucial role in the spread of health misinformation, and their misinformation can exponentially amplify the harm [29]. The user feedback driven dynamic adjustment strategy utilizes implicit signals such as community, likes/stampedes, etc. to update node reputation and content risk scores in real time, forming a dynamic co-governance system [18, 30]. Critically, these methods treat the information itself as homogeneous entities and do not differentiate what level of intervention should be taken for content with different risk levels. In addition, their intervention strategies are mostly "blocking" or "limiting flow", lacking protective mechanisms for low-risk beneficial information.

In summary, there are three shortcomings in existing research. The first is the precise alignment between content and authoritative medical knowledge; The second is the differentiated design of communication control strategies that match the level of information risk in node behavior, and the third is the effective coupling of content risk inference and social network dynamic intervention mechanisms. The multidimensional dynamic permission control framework proposed in this study addresses these three issues. It constrains LLM reasoning through a hierarchical evidence knowledge base for fine-grained risk inference. Then it integrates risk assessment with node credibility and network influence through a mathematically grounded dynamic permission function, and generates differentiated forwarding permissions. The behavior permission control can accurately suppress the dissemination of high-risk health content while maintaining the flow of beneficial information.

3 Proposed method

3.1 Overall working process of the framework

We propose a Dynamic Suppression Model (DSM) for health information propagation on social networks. As illustrated in Fig. 1, DSM comprises four steps: evidence retrieval, content risk calculation and reasoning, dynamic permission control, and Information behavioral intervention. The framework jointly uses unstructured online health content, individual node attributes, and social network propagation characteristics to enable intelligent governance of information flow.

(1) Evidence Retrieval: To provide evidence-based risk assessment, we first build a high-quality and structured domain knowledge base. Then use LLMs to understand the health perspective of the content, and match the evidence knowledge base through retrieval-augmented generation (RAG) techniques.

(2) Content Risk calculation and reasoning: Quantify the risk of health information that users intend to publish or forward based on external evidence and internal fine-grained semantic features. This step outputs the risk level, risk value, and inference evidence set after processing.

(3) Dynamic Permission Control: This step constructs suppression function $f(R, C, d)$ to calculate dynamic permission values. The function will obtain the credibility value (C) and influence

value (D) of the node, and integrate the content risk value (R) to calculate the real-time dynamic information behavior permission value of the node.

(4) Information behavioral intervention: This step compares the node permission value with the threshold to determine whether the node can publish or forward information. When the permission value in this scenario is below the threshold, the system will prevent publishing/forwarding operations and provide interpretable explanations derived from the risk inference process. When the permission value exceeds the threshold, content is allowed to propagate.

This pipeline is triggered in real time upon user-initiated sharing actions, and formed a closed-loop control mechanism (Figure 1). The system retrieves relevant evidence from the hierarchical knowledge base to support risk inference (Module 1). Subsequently, LLM quantifies content risk based on external evidence and internal semantic features (Module 2). The risk score is input into the dynamic permission function along with node credibility and network influence to calculate real-time forwarding permission values (Module 3). The system determines whether to allow or prevent sharing behavior based on a preset threshold (module 4).

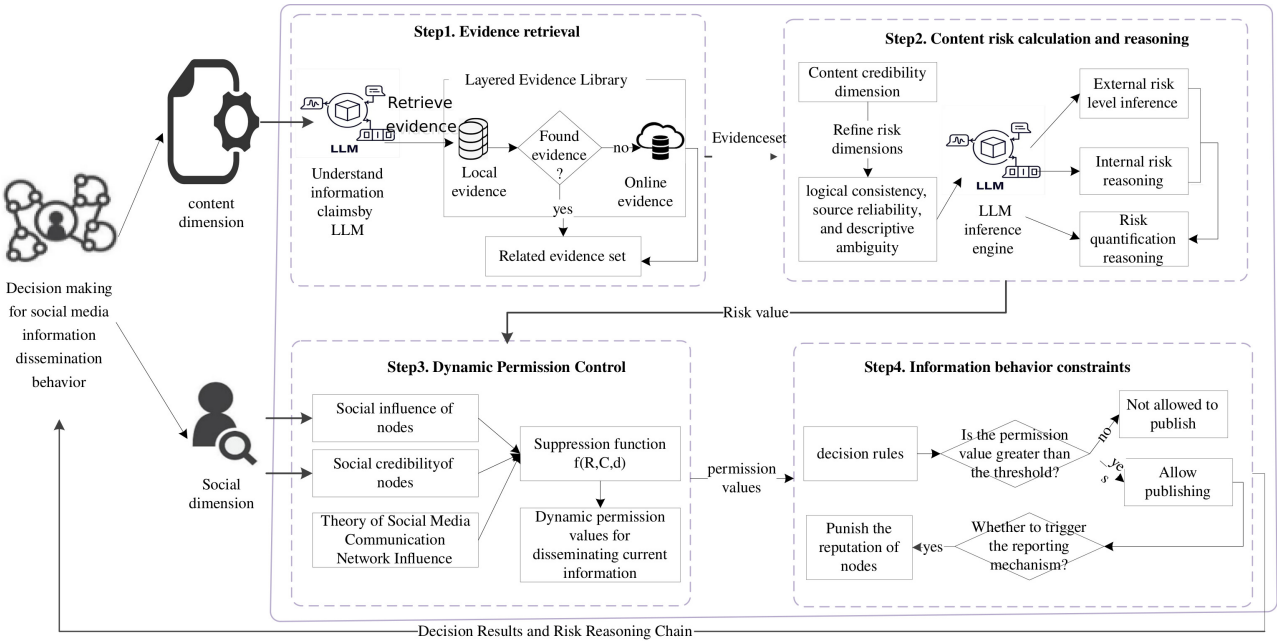


Figure 1: Principle description diagram of decision-making framework

The key innovation of DSM lies in the integrated innovation of coupling content risk, source credibility, and network structure attributes to solve the problem of suppressing inaccurate health information. Although these three dimensions have accumulated achievements, there is relatively little integration of them to achieve social network node behavior permission control mechanisms.

3.2 Evidence retrieval

3.2.1 Construct hierarchical evidence base

We constructed a hierarchical evidence knowledge base to enhance the accuracy and explainability of health information risk assessment. We take diabetes prevention and treatment related knowledge as an example to build a prototype of the knowledge base. The main reason is that diabetes has diversified treatment options and significant individual differences, and relevant health information is representative in personalized assessment. Drawing from the evidence grading system in evidence-based medicine, our knowledge base is structured into three layers corresponding to different levels of authority. These data sources are summarized and will be used to verify the "logical consistency" and "descriptive clarity" of textual content (see Table 1).

To enable automated processing and efficient utilization of knowledge, we performed structured information extraction on the included literature. Through natural language processing techniques

such as named entity recognition, relation extraction, and manual verification, the data is converted into structured data records. The content includes: evidence theme (such as diagnosis, intervention, beverage, etc.), publication date, evidence content, applicable conditions and target population, label (labeled according to hierarchical levels A/B/C), and source.

Table 1: Authority hierarchy of medical knowledge sources

Authoritative hierarchy	**Data source**
Core Layer (A)	International clinical guidelines (e.g., ADA, WHO), Cochrane, pharmacopeias
Extended Layer (B)	Authoritative medical journal articles
Supplementary Layer (C)	Vetted high-quality health content from the internet

Although the current prototype is based on diabetes related knowledge, the framework is designed for domain extensibility. After method validation, the system supports dynamic injection of new evidence and expansion to other health domains. To address potential conflicts in multi-source knowledge, we adopted a resolution strategy based on authority level and timeliness principles (see Table 2). It ensures that the knowledge base remains up-to-date and authoritative.

Table 2: Types of conflicts and strategies

Conflict Type	**Solution Strategy**	**Verification Method**
Authority Level	Prioritize higher-level evidence	Check if the model adopts Level A over Level C when faced with contradictions.
Timeliness conflict	Adopt the latest published version.	Check whether the new version reaches the same conclusion and automatically overwrite the old version.
Dosage correlation conflicts	Labeling applicable condition constraints	Evaluate whether the supplementary instructions generated by the model clearly indicate the applicable prerequisites.

Through these operations, we have constructed a structured and scalable hierarchical evidence knowledge base. It is not merely a collection of facts, but also the knowledge foundation for subsequent LLM logical reasoning, scientific judgment, and personalized risk assessment.

3.2.2 Retrieve evidence base

In order to achieve high-precision semantic alignment between health statements and authoritative evidence, we use the BAI/bge-m3 multifunctional embedding model for vectorized representation and similarity calculation. This method retrieves the Top-K (default K=5) most relevant evidence items. A weighted evidence consistency score S_{consist} is calculated as As shown in Equation 1.

$$S_{consist}(c_i) = \frac{\sum_{j=1}^{k} sim(c_i, e_j) \bullet level(e_j)}{\sum_{j=1}^{k} level(e_j)} \tag{1}$$

where S is the consistency score, c_i is the content viewpoint, $level(e_j)$ represents the quantitative weight of evidence, which is defaulted to A: 1, B: 0.9; C: 0.8, and $sim(c_i, e_j)$ is the similarity of text vector The weight reflects the evidence-based principle that high-level evidence should dominate the judgment. In the retrieval step of Figure 1, the closed-loop update mechanism of the detailed hierarchical evidence base is shown in Figure 2. When a health statement cannot fully match any existing evidence in the knowledge base, the system automatically marks it as pending verification. It can be combined with a manual review mechanism to regularly scan the list to be verified, and finally update the knowledge base with the revised knowledge after review. This closed-loop design ensures that the framework continuously expands its coverage of health topics and reduces the frequency of unresolved topic content over time. Figure 1 and Figure 2 together indicate that the proposed framework is not a static pipeline, but a self-improving governance system.

3.3 Dual-Cognition-Based risk quantification

3.3.1 Risk assessment description

We propose a two-stage risk quantification mechanism, which is achieved through general preliminary level qualitative and structured detailed evaluation. In the first stage, the risk level of information is determined based on the existence of external authoritative evidence and the serious risk

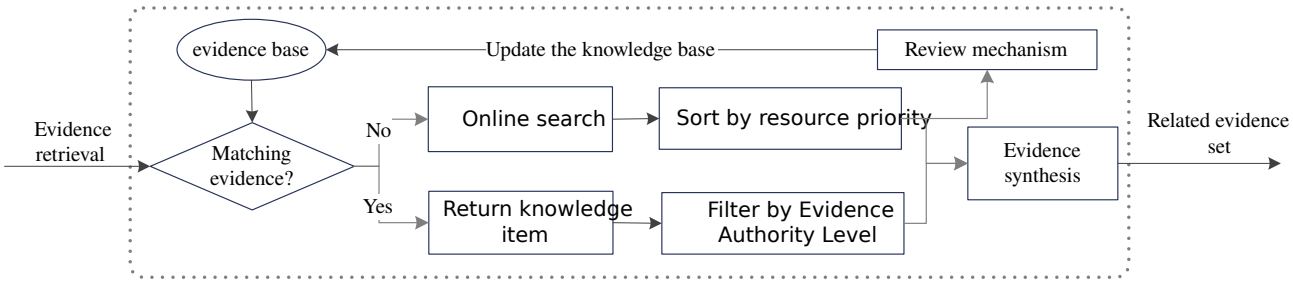


Figure 2: Dynamic update mechanism of knowledge base

consequences. The second stage is to generate three fine-grained risk scores based on the consistency characteristics of the content viewpoint itself, which are inferred by the LLM according to the feature description. The structured features considered internally include logical consistency, source reliability, and ambiguity. This design is consistent with the dual cognitive model of external validation and internal inference, improving the transparency and auditability of model inference.

The risk assessment process mainly includes three steps (Figure 3): The first step is to determine the overall risk level (high, medium, or low) of the information based on authoritative medical evidence; The second step is to conduct structured analysis from three dimensions: logical consistency, source reliability, and expression ambiguity; The third step is to integrate the judgment results of the three dimensions within the legal value range corresponding to the risk level, and generate the final quantitative risk value.

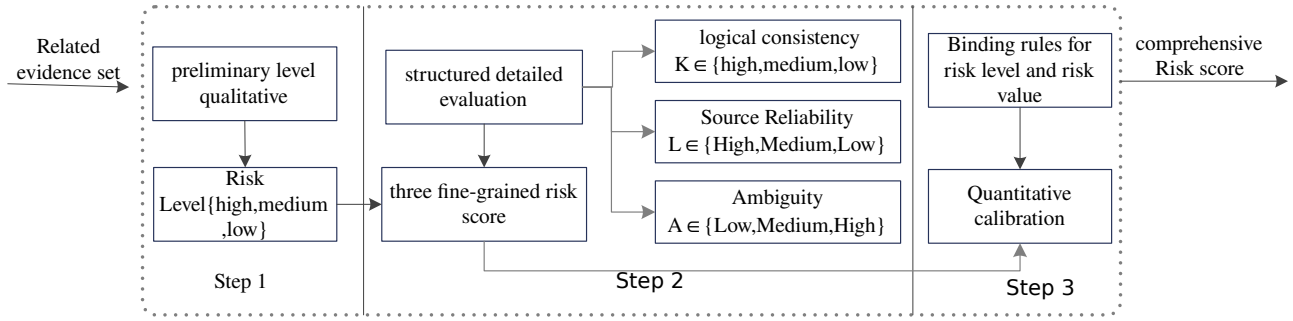


Figure 3: Risk quantification process

Table 3 define the core Variables of the risk quantification. The symbol m represents the health information to be evaluated, and its risk attribute is characterized by the following variables. All dimension values are derived from a predefined semantic set, ensuring that the output is discrete, comparable, and explainable.

Table 3: Explanation of variables related to risk calculation

Variable	Definition	Scope	Explanation
G	risk level	$G(m) \in \{High, Medium, Low\}$	Reflects the severity of potential health harm caused by the claim.
R	risk score	$R(m) \in [0, 1]$	Quantify the overall risk value of information.
K	logical self-consistency	$K \in \{High, Medium, Low\}$	Indicates whether the claim contains internal logical contradictions, causal fallacies, conceptual confusions, or violations of basic scientific principles.
L	Source Reliability	$L \in \{High, Medium, Low\}$	Measures the authority and trustworthiness of the information source.
A	Ambiguity	$A \in \{Low, Medium, High\}$	Describes the clarity of the statement and completeness of contextual or conditional specifications.

3.3.2 Risk calculation process

(1) Risk Level Determination Rules: The risk level $G(m)$ is determined jointly by factual accuracy and potential health consequences, driven entirely by external evidence.

High: The claim contains factual errors that could lead to severe outcomes, such as delayed treatment, medication misuse, organ damage, or death.

Medium: The claim is partial, overgeneralized, or lacks necessary contextual conditions; it may cause misunderstanding but does not pose an immediate life-threatening risk.

Low: The claim is largely accurate, aligns with current clinical guidelines or high-quality evidence, and is expressed with appropriate precision.

(2) Risk quantification constraints: To prevent inconsistency between content risk and risk level after being structured and quantified, a constraint on the legal range of risk values is introduced. This constraint constitutes a safety fallback mechanism. Any output that violates in Equation 2 will be forcibly corrected by the post-processing module to the corresponding interval endpoint.

$$R(m) = \begin{cases} [0.7, 1.0], & \text{if } G(m) = \text{high} \\ [0.4, 0.7), & \text{if } G(m) = \text{medium} \\ [0, 0.4), & \text{if } G(m) = \text{low} \end{cases} \tag{2}$$

(3) Risk score calculation mechanism: The evaluation dimension vector is: $R(m) = (K, L, A)$. As the Table 4, each dimension is mapped to an ordered level {0,1,2}, where 0 represents a low-risk, 1 represents a medium-risk, and 2 represents a high-risk. The raw risk value can be calculated as Equation 3.

$$R_{raw} = \frac{K + L + A}{6}, \quad R_{raw} \in [0, 1] \tag{3}$$

The denominator 6 is derived from the maximum possible sum of the three dimensions, and it ensures that R_{raw} is normalized to the range [0,1]. For example, if a claim exhibits High consistency risk (K=2), Medium source risk (L=1), and Low ambiguity risk (A=0), then R_{raw} =(2+1+0)/6=0.5. This linear aggregation assumes equal contribution from each dimension, which is a design choice that can be adjusted in future work by introducing dimension-specific weights.

The R_{raw} is then constrained by Equation 2 according to the first-stage risk level G to yield the final risk score $R(m)$. This process enforces step-by-step reasoning through structured prompts. The final system returns a quadruple: ($G(m)$, $R(m)$, (K, L, A), reason), which bring a full auditability. For example, "Why did this piece of information receive a score of 0.35? Because K=Low, L=Low, and A=Low."

Table 4: Risk quantification definition of each dimension

Variable	**Mapping Rule Definition (High: 2; Medium: 1; Low: 0)**
K: Logical consistency	**High:** Contains self-contradiction, pseudoscientific terms (such as “detoxification” and “acid-alkaline constitution”) or obvious fallacies of causality. **Medium:** There is simplification, analogy, or slight causal ambiguity, but no fundamental errors; **Low:** The expression is rigorous, free from logical contradictions, and conforms to basic scientific principles;
L: Source reliability	**High:** No clear source, from social media, commercial promotion, or personal experience. **Medium:** Information published on medical platforms and well-known social media; **Low:** Documents issued by regulatory agencies and platforms certified by authoritative institutions;
A: Ambiguity	**High:** Using absolute terms such as “all”, “certain”, and “cure” without any conditional restrictions. **Medium:** Some context is missing, but the core meaning can be understood; **Low:** Includes applicable population, dosage, contraindications, and other limiting conditions;

3.4 Dynamic Permission Control

In social networks, the propagation risk of health misinformation depends not only on the content itself but also on the credibility of the disseminator and their network influence. To address this, we propose a three-dimensional dynamic permission control mechanism. It can dynamically determine whether users have the authority to publish or forward specific health information.

3.4.1 Permission Function Formulation

The weighted suppression function $f(R, C, d)$ is designed with a Signal-to-Noise Ratio (SNR). The numerator represents the "desired signal" (node credibility C), while the denominator quantifies the "suppressive noise" contributed by content risk and network influence. The permission value is calculated as Equation 4.

$$f(R, C, D) = \frac{C}{1 + \gamma \bullet R \bullet \log(1 + D)} \tag{4}$$

where:

(1) Credibility ($C \in [0, 1]$) reflects the reliability of a node's historical behavior, like frequency of spreading rumors, professional certification status. Users with higher credibility are granted greater "speaking rights," aligning with the principle of reputation-based access control [27] .

(2) Risk Score ($R \in [0, 1]$) is computed in 3.3, representing the potential harm of the content. Higher values indicate greater misleading potential or danger.

(3) Out-degree ($D \geq 0$) denotes the number of followers or the average historical reach, serving as a proxy for social influence [29]. Greater influence entails greater responsibility. So D amplifies the suppression term.

(4) Coupling coefficient ($\gamma > 0$) adjusts the interaction strength between risk and influence.

The design principle of forwarding function is that high-credibility nodes retain relatively permissive sharing rights even with high out-degree; low-credibility but high-influence accounts face stringent restrictions; ordinary users with medium credibility and influence operate within an intermediate range, preserving everyday information-sharing freedom. This design ensures that governance is proportionate, adaptive, and fairness-aware.

We adopt $log(1 + D)$ rather than a linear D for two reasons grounded in network science. First, real-world node out-degree follows a heavy-tailed distribution [31]; a linear term would disproportionately penalize moderately influential users. The logarithmic form ensures suppression increases at a diminishing rate as influence grows. Second, $log(1 + D)$ guarantees $f(R, C, D) \to 0$ smoothly as $D \to \infty$, ensuring numerical stability. The hyperparameter γ>0 adjusts the coupling strength between risk and influence.

3.4.2 Mathematical Properties of the Suppression Function

The design of Equation 4 is not merely heuristic but possesses critical theoretical properties that ensure system robustness:

Property 1:Boundedness. For all $R \in [0, 1]$, $C \in [0, 1]$, $D \geq 0$, and $\gamma > 0$, the permission value satisfies $f \in [0, 1]$.

Property 2:Monotonicity. The function $f(R, C, D)$ is strictly decreasing with respect to to R and D ($\partial f/\partial R \leq 0$, $\partial f/\partial D \leq 0$), and strictly increasing with respect to C ($\partial f/\partial C > 0$). This ensures that the system responds predictably to changes in risk, influence, and credibility.

Property 3: Limiting Behavior. The permission function exhibits the following asymptotic properties:

(a) As risk approaches zero: if $R \to 0$, $f(R, C, D) \to C$, the permission depends only on credibility. This proves that the mechanism does not hinder the free flow of safe information, preserving network utility.

(b) As risk approaches maximum: if $R \to 1$, high-risk content faces maximum suppression.

(c) As influence grows unbounded: if $D \to \infty$, $f(R, C, D) \to 0$ regardless of C , unless $C = 0$. This mathematically guarantees that high-risk content cannot be propagated by nodes with massive influence, mitigating the "super-spreader" risk in epidemiological terms.

These limiting properties demonstrate that the function smoothly interpolates between edge cases. When $\gamma = 0$, the function degenerates to $f = C$, it is relying solely on reputation. As γ increases, the suppression due to content risk and influence intensifies.

The dynamic permission mechanism also solve the problem of 'cold start' for new users. Since the mechanism is risk-sensitive rather than purely reputation-based, for low-risk content with $R \to 0$, the

permission value f is approximately equal to the node's credibility C. Even new users with low initial credibility retain basic forwarding capability. Conversely, if a new user attempts to spread high-risk content ($R \to 1$), they are blocked regardless of influence. As users accumulate positive feedback, their credibility increases, progressively granting greater sharing privileges. High-trust community nodes can provide credibility endorsement for uncertain information with broader reach. This design protects the network while preserving legitimate discourse for new participants.

3.5 Information Behavioral Constraint Mechanism

We define a Global Forwarding permission threshold $\theta \in (0, 1)$, representing the minimum acceptable permission for content distribution:

$$\text{Allow}(R, C, D) = \begin{cases} \text{true}, & \text{if } f(R, C, D) \geq \theta \\ \text{false}, & \text{otherwise} \end{cases} \tag{5}$$

A higher θ imposes stricter requirements. For example, only combinations of high credibility, low risk, and low influence are permitted to forward. A lower θ yields a more open information environment. For high-risk content , we require $f < \theta$. For low-risk content, we require $f \geq \theta$. Solving these inequalities yields feasible ranges for γ and θ that satisfy both constraints, establishing the existence of a Pareto-optimal configuration. The parameter optimization procedure in Section 4.2.1 numerically identifies such configurations.

When misinformation reaches domain expert nodes, they can flag it as high-risk, triggering an expert reporting mechanism. Expert identity is determined through two criteria: platform-level verification of official credentials (e.g., licensed medical professionals), and continuous behavioral alignment with the evidence knowledge base. A dynamic revocation protocol is set up. It ensures that nodes with frequent disagreements with verification knowledge will lose their expert identities. To mitigate risks of potential malfeasance or errors by expert nodes, the blocking mechanism is activated only through multi-node consensus or when risk scores exceed a critical threshold. This design ensures system robustness by preventing isolated node failures from compromising overall network governance.

4 Experiments and evaluation

4.1 Content Risk Assessment

4.1.1 Experimental Data Description

The experiment constructs an evidence dataset and a health claim test dataset.

Evidence dataset: 2,600 evidence texts on diabetes prevention and treatment were collected from authoritative databases. The data sources included FDA drug instructions, clinical guidelines, and core journals. After manual review, 481 pieces of Type A evidence, 1309 pieces of Type B evidence, and 810 pieces of Type C evidence were sorted out.

Health information text set: As this study aims to validate a comprehensive control mechanism, rather than training a generic model. However, existing public benchmarks typically focus on binary rumor detection and lack the fine-grained risk stratification annotations required for this study. Meanwhile, the language style of Chinese cannot directly match the English dataset. We need a test dataset that fits the scenario.

We Collect 120 typical false or misleading health information texts as seed examples from rumors disclosed on the official website. These seeds were manually verified and mixed with factual health texts. Then use LLM to generate 528 diverse health information texts covering various topics within diabetes care (diet, medication, exercise, monitoring, alternative therapies). To review the quality issues of LLM generated data, two annotators with medical backgrounds will conduct a review, and only those confirmed will be adopted; Text with risk labels that cannot be agreed upon is excluded. The text set includes high-risk information items n=178. that is, they contain clear medical errors that may directly endanger health; It contains medium risk information items n=170, that is, they

are based on facts, but one-sided/unconditional; It contains low-risk information items n=180, that is rigorously expressed and consistent with clinical guidelines.

4.1.2 Definition of experimental indicators

We evaluate model performance from two critical perspectives, one is the overall classification accuracy, and the other is safety-oriented detection of high-risk misinformation. It reflects the dual goals of general effectiveness and clinical risk mitigation.

(1) Overall accuracy and macro-average $F1$:

Accuracy measures the proportion of correctly classified claims across all risk levels. It is defined as the proportion of samples in which the risk level predicted by the model is completely consistent with the marked real level in all test samples. It calculated by Equation 6.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}(\widetilde{y_i} == y_i) \tag{6}$$

where N is the total number of samples tested; y_i is the true risk level of the information; $\widetilde{y_i}$ is the predict risk level; When the forecast is equal to the real label, the value is 1, and the correct classification quantity is accumulated to calculate the correct proportion. This index is intuitive and easy to explain, but it is easily affected by the deviation of category distribution.

Macro-F1 computes the unweighted average of F1-scores for High, Medium, and Low risk classes, ensuring balanced evaluation without bias toward majority classes. It is calculated by Equation 7.

$$\text{Macro-F1} = \frac{1}{3}\left(F1_{\text{high}} + F1_{\text{medium}} + F1_{\text{low}}\right) \tag{7}$$

And the $F1$ score for each category is the macro average harmonic value of precision (P) and recall (R). It is calculated by Equation 8. For each category $c \in \{\text{high, medium, low}\}$, and TP is true positive, FP is false positive, and FN is false negative.

$$F1_c = 2 \cdot \frac{P_c \cdot R_c}{P_c + R_c}, \quad P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c} \tag{8}$$

(2) High-Risk Detection Metrics:

Given the severe consequences of undetected harmful content, we report precision, recall, and F1-score specifically for the high-risk class. Recorded as high-risk precision (HRP), high-risk recall (HRR), and high-risk F1 ($HRf1$). The calculation method is shown in formula 8, where c equals high risk value.

High-Risk False Negative Rate (FNR): Defined as the fraction of actual high-risk information text misclassified as medium or low risk. This metric is particularly critical in medical contexts: failing to detect a dangerous claim poses greater public health risk than incorrectly labeled low-risk information. Thus, minimizing FNR is prioritized in system design and threshold tuning. It is calculated by Equation 9.

$$\text{FNR} = 1 - \text{HRR} \tag{9}$$

4.1.3 Ablation Study on Model Components

In order to verify the enhanced value of hierarchical evidence and double quantitative rules for risk reasoning in large models, we established a comparative experiment of two scenarios. The experimental comparison is as follows:

(1) Variant A (LLM-only): General LLM without rules or evidence. We use Qwen-Max model, but without structural rule constraints and evidence base, relying on general LLM.

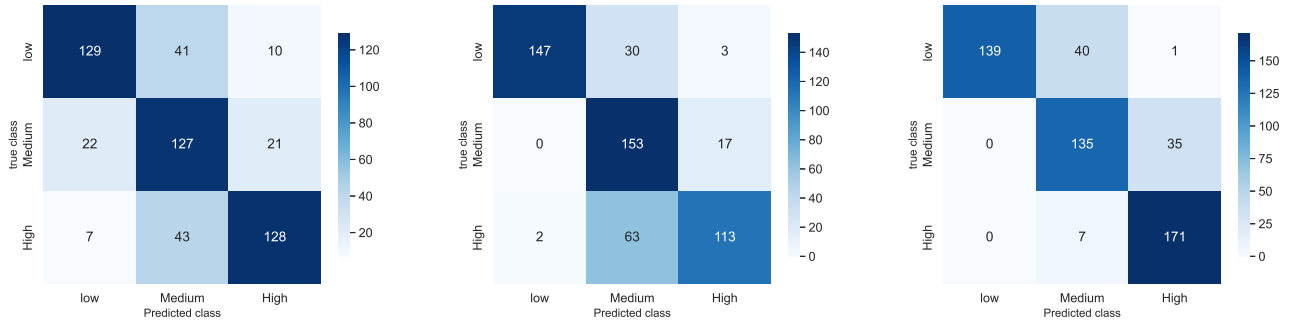
(2) Variant B (LLM + Rules): With risk constraints but without evidence retrieval. We use Qwen-Max model and structured Risk Constraint, but without external evidence retrieval, relying on LLM and subjective judgment of fine-grained risk quantification.

(3)Variant C (Full Framework): Our complete model. using Qwen-Max model and adopt the evaluation model of risk constraint rules and evidence mechanism presented in this article. The experimental results are presented in Table 5. To intuitively illustrate classification biases across methods, we visualize 3 × 3 confusion matrices as heatmaps(Figure 4), whose rows are true labels, and columns are predicted labels. The color intensity reflects the frequency of predictions. Our method consistently outperforms all baselines across core metrics.

Table 5: Comparison of experimental indicators for content evaluation

Metrics	**Variant A**	**Variant B**	**Variant C**
Accuracy	0.7273	0.7822	0.8428
Macro-F1	0.7299	0.7853	0.8423
High-risk Precision (HRP)	0.8050	0.8496	0.8261
High-risk Recall (HRR)	0.7191	0.6348	0.9607
High-risk F1 (HRF1)	0.7596	0.7267	0.8883
High-risk False Negative Rate (FNR)	0.2809	0.3652	0.0393

In terms of overall performance, our method (Variant C) is significantly better than the other two ablation models. It achieves an accuracy of 0.8428, representing an 11.55 percentage-point improvement over Variant A, and a $macro - F1$ of 0.8423, indicating balanced discriminative ability across all three risk classes. This improvement stems from the anchoring effect of the external evidence base, which effectively suppresses LLM hallucinations common in the medical domain. More critically, our method achieves a high-risk recall of 96.07%, with only 7 high-risk cases missed ($FNR = 3.93\%$), substantially lower than Variant A ($FNR = 28.09\%$) and Variant B ($FNR = 36.52\%$).



(a) **(b)** **(c)**

Figure 4: Confusion matrix heatmap. (a) Variant A: The diagonal deviation is serious (b) Variant B: most of the high-risk samples were correctly identified (c) Variant C: almost all high-risk samples were correctly identified

As shown Figure 4(c), our method eliminates the most dangerous error type. None of the truly high-risk samples were misclassified as low-risk, they were only misclassified as medium-risk. Although the model in this article conservatively upgraded 35 cases of medium risk claims and 1 case of low-risk claim to high risk, the specific explanation provided by RAG allows publishers to quickly make corrections, which is also the greatest guarantee for the circulation of low-risk health information. In contrast, Variant A exhibits a severe high-risk omission problem: as shown in Figure 4(a), 43 high-risk claims were misjudged as medium risk, 10 as low risk, resulting in a FNR of 28.09%. This confirms that that general LLMs lack reliable professional judgment in specialized medical domains. Variant B (LLM + Rules) improves HRP to 0.8496 through structured output, yet its HRF1 (0.7267) and FNR (36.52%) are inferior to Variant A (0.7596 and 28.09%, respectively). This deterioration stems from the reliance on static rules without evidence retrieval: when high-risk misinformation evades these predefined rules, the model lacks external verification capability, producing excessive False Negatives and degrading both recall and HRF1. As shown in Figure 4(b), although structured output improves precision, it did not improve recognition accuracy; lacking external knowledge support, HRR dropped to 63.48% and FNR paradoxically increased. On the contrary, the FNR increased.

In summary, our method has realized the transformation from subjective reasoning to evidence support by introducing health medicine knowledge base. It has been improved in accuracy, safety

and explainability. Especially, the evidence retrieval component (absent in Variant B) reduces the FNR from 36.52% (Variant B) to 3.93% (our method). This proves the individual contribution of the hierarchical evidence base in preventing LLM hallucinations and ensuring reliable high-risk misinformation detection.

4.2 Evaluation of misinformation suppression in social network propagation

This experiment needs to verify the effectiveness of the proposed model in the governance of complex network information propagation. The experiment simulates the information diffusion process in real social scenarios by constructing a multi type node network, and compares and analyzes it with four baseline models.

4.2.1 Simulation design

(1) Network and Information Configuration:

We construct a scale-free network with $N = 500$ nodes using the Barabási-Albert (BA) model [31] and convert it into a directed graph to simulate unidirectional information propagation. The experiment simulates the propagation dynamics messages over 30 time steps. The BA model exhibits a typical power-law degree distribution, which aligns with the topological characteristics observed in real-world social media platforms [32]. The connection parameter is set to $m = 4$, which accurately reflects the "few hubs, many leaves" topology characteristic of social networks. To ensure reproducibility, the random seed is fixed at 42. The network topology is constructed using Python's NetworkX toolkit to simulate heterogeneous connectivity in real information propagation scenarios.

To simulate the heterogeneity of real social networks, nodes are divided into three categories as Table 6. Each node type with distinct propagation attributes. Professional users provide authoritative verification as high reputation gatekeepers; Follower simulate super communicators; ordinary users represent the public group with certain judgment, while silent users depict the passive information recipients who account for the majority in reality. This distribution conforms to the typical characteristics in scale-free networks.

It should be noted that the classification of 'Professional Users' in our simulation is based on pre-assigned labels to isolate the effect of the permission mechanism. In a real-world deployment, expert identification would rely on platform-level verification mechanisms, for example, real-name certification, institutional affiliation checks, and combined with the user's historical behavioral patterns on medical topics.

Table 6: Initialization values of different role node attributes

Role	Initial Credibility	Description	Proportion
Professional User	High ($C = 1$)	Verified medical practitioners; forward low risk information	5%
Follower	Medium ($C = 0.5$)	Highly active, emotion-driven; forwards under lenient conditions	15%
Ordinary User	Medium ($C = 0.5$)	Moderately cautious; requires stronger justification to forward	20%
Silent User	Medium ($C = 0.5$)	Receives information only; never forwards	60%

To verify model robustness in complex scenarios, the risk distribution incorporates confusion characteristics between LLMs bias and real risk from previous experiments. A total of 484 health claims with risk scores R$\in$[0.08,0.98] were selected from the test set. These include 161 low-risk (R$<$0.4), 164 medium-risk (0.4$\leq$R$<$0.7), and 159 high-risk(R$\geq$0.7) instances.

(2) Parameter optimization: To optimize the trade-off between high-risk suppression Rate (SR) and low-risk False suppression rate (FSR), we perform a systematic grid search over two key hyperparameters. The risk-influence coupling coefficient γ and The global forwarding threshold θ.

There is no single "best" solution, but rather a Pareto frontier of balanced configurations(Figure 5). To facilitate comparison, we define a Balance Score: balance score, which is computed by $SR*(1-FSR)$. A higher score indicates better simultaneous performance in suppressing harmful content and avoiding over-blocking benign information. where SR$\in$[0,1]. This score reaches a maximum of 0.832

under the following configuration: γ=0.9 which can balance risk-based penalties with information fluidity; and $\theta = 0.2$ which set the minimum acceptable credibility level for forwarding eligibility. This setting achieves strong high-risk interception while maintaining a low false suppression rate, making it ideal for health information governance. All subsequent main experiments adopt this optimized configuration.

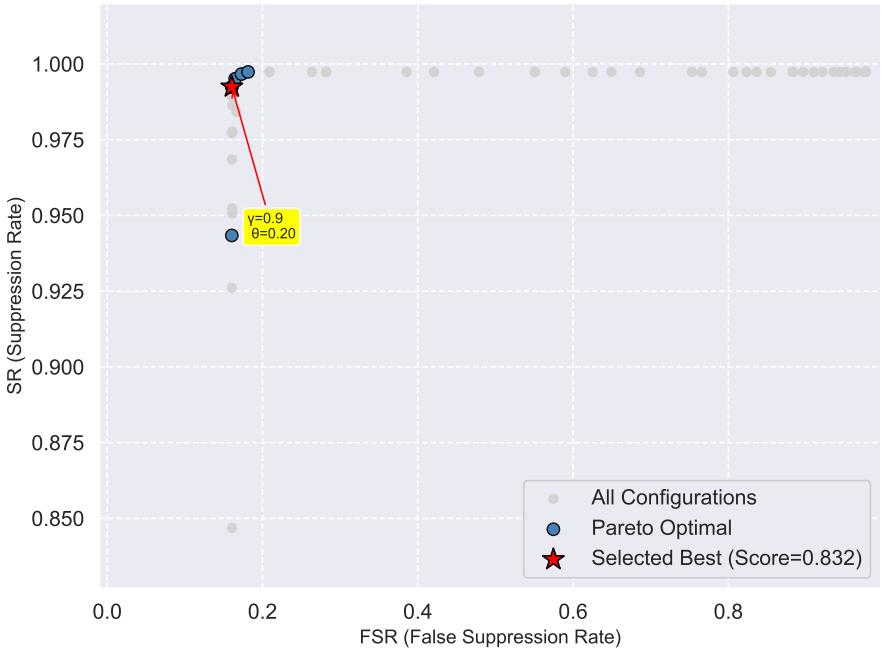


Figure 5: Parameter optimization analysis

(3) Comparison Model Explanation: To comprehensively evaluate the effectiveness of the proposed dynamic mechanism, we selected representative baseline models that cover the spectrum of information dissemination control strategies. Baselines 1-4 represent standard settings or classic strategies in this domain, while Proposed denotes our integrated control model.

Baseline1: Unrestricted Diffusion Model [31, 33]. All non-silent nodes, including professionals, followers, and regular users always forward any received message to all their neighbors upon receipt. This scenario simulates the completely free flow of information as the performance ceiling and security baseline.

Baseline 2: Role-based Heuristic Strategy [34]. This baseline simulates a static intervention mechanism widely observed in real-world social platforms, where forwarding behavior is governed by fixed user attributes. Follower nodes forward unconditionally, professionals only forward information with a credibility higher than 0.9, and regular User randomly forward with a probability p=0.3, which is more in line with real-world scenarios.

Baseline 3: Probability-based Forwarding Strategy[35]. Unlike the static rules in Baseline 1 and 2, this model adopts a probabilistic approach. Nodes calculate a forwarding probability based on the joint function of their intrinsic credibility and the message's risk score. This model isolates the contribution of the probabilistic forwarding mechanism without dynamic thresholds.

Baseline 4: Static Risk Threshold [22]. This represents a widely used rule-based content review system. If a message's risk score exceeds a fixed threshold, such as set to 0.8 indicating high-risk information, it is intercepted and blocked from further propagation. Messages below the threshold are forwarded with a fixed probability.

Proposed method: Dynamic Permission with Circuit Breaker. This model implements a dynamic permission control mechanism. It calculates a node-specific dynamic threshold based on the message risk and node degree. Furthermore, it introduces a circuit breaker mechanism. If a message is identified as high-risk and is flagged by at least two expert nodes, the message propagation is immediately suppressed globally. This design ensures precise interception of high-risk content while minimizing disruption to normal information flow.

4.2.2 Definition of experimental indicators

In order to evaluate the inhibitory effect on high-risk information while ensuring the dissemination of high-quality information, we adopt four key metrics to assess governance performance.

(1)Effective Coverage (EC): This indicator reflects the average reach breadth of low-risk health information in social networks, and is used to evaluate the system's ability to guarantee the circulation efficiency of beneficial content. Then EC is defined as Equation 10.

$$EC = \frac{1}{|M_L|} \sum_{m \in M_L} \frac{N(m)}{N}, \quad \text{if } M_L \neq \emptyset \tag{10}$$

Where M_L is the collection of low-risk information, $N(m)$ is the total number of nodes covered by information m, and N is the total number of nodes in social network, where N=500. The index value is between [0,1], and the higher the value, the better the performance of the method.

(2)High-Risk Coverage (HRC): This indicator measures the average diffusion range of high-risk misinformation and directly represents its potential social harm. Then HRC is defined as Equation 11.

$$\text{HRC} = \frac{1}{|M_H|} \sum_{m \in M_H} \frac{N(m)}{N}, \quad \text{if } M_H \neq \emptyset \tag{11}$$

Where MH is the collection of high-risk information, and others are the same as before.The index value is between [0,1], and the lower the HRC value, the higher the security of the system.

(3)Suppression Rate (SR): It directly measures the ability of the system to reduce the spread of harmful content in real communication scenarios. The index is based on the free communication without intervention (baseline1), and reflects the intervention intensity through the relative reduction ratio. It is calculated as Equation 12. The index value is between [0,1], and the higher the HRC, the stronger the inhibition ability. This form is widely used in the field of computational social sciences to quantify the effect of intervention [22].

$$SR = 1 - \frac{HRC_{Model}}{HRC_{BL1}}, \quad \text{if } HRC_{BL1} > 0 \tag{12}$$

Where HRC_{Model} denotes the High-risk Coverage achieved by the proposed method, and HRC_{BL1} denotes the High-risk Coverage of Baseline 1.

(4)False Suppression Rate (FSR): It measures the degree of excessive blocking of low-risk content by the system. For each low-risk information m, calculate the proportion of nodes covered in baseline1 but not covered in this model. Then average all low-risk information to obtain FSR as Equation 13. This indicator is a supplement to EC. FSR $\in$ [0,1], The lower the index value, the less information that has been wrongly intervened.

$$FSR = \frac{1}{|M_L|} \sum_{m \in M_L} \frac{|N_{Bs1}(m) \setminus N_{Model}(m)|}{|N_{Bs1}(m)|} \tag{13}$$

Where M_L is the set of low-risk claims, $N_{Bs1}(m)$ is the propagation set of claim m under Baseline 1, and $N_{Model}(m)$ is the propagation set under the proposed model.

4.2.3 Overall performance comparison

To validate the effectiveness of our proposed model, we conducted simulation experiments under identical network topology (BA scale-free network, N=500) and message distribution conditions (484 messages with heterogeneous risk scores). The results are summarized(see Table 7).

These results provide a comprehensive evaluation of the proposed model's performance in terms of risk control efficiency, information flow, and decision accuracy.

(1) Analysis of Communication Suppression Efficiency. The analysis of high-risk content suppression reveals significant disparities among the models. Baseline 1 shows the natural propagation state under uncontrolled conditions. In this state, a high-risk spread (HRS) of approximately 80.49%. This indicates a complete loss of control over harmful information dissemination. While Baseline 2

Table 7: Propagation performance index results

Model	EC	HRC	SR	FSR
Baseline1	0.8036	0.8049	0.0000	0.0000
Baseline2	0.6660	0.6700	0.1676	0.1802
Baseline3	0.8036	0.2546	0.6837	0.0000
Baseline4	0.8028	0.1990	0.7528	0.0008
Proposed method	0.8036	0.0178	0.9778	0.0000

introduces role-based rules to mitigate this issue. Its suppression rate (SR) of only 16.76% proves insufficient for effective control. The introduction of quantitative indicators in Baseline 3 and Baseline 4 improves the situation, achieving suppression rates of 68.37% and 75.28%. However, both models still allow over 19% of high-risk content to propagate, posing substantial security threats. The proposed model has shown good advantages. It successfully curbs the high-risk spread to a mere 1.78%, achieving a suppression rate of 97.79%. This superior performance is attributed to the integration of the Expert Consensus Reporting Mechanism and the Dynamic Threshold Strategy. These enhancements enable the model to accurately identify and block high-risk content, even when such content attempts to evade detection by lowering its risk score.

(2) Analysis of Normal Information Flow Preservation. An effective control mechanism must not unduly hinder the dissemination of normal, low-risk information. The experimental data indicates that the Proposed Model, Baseline 1, and Baseline 3 all maintain an effective coverage (EC) of approximately 80.36%. This shows that these models allow normal information to flow freely without significant obstruction. Conversely, Baseline 2 exhibits a notable decline in effective coverage, dropping to 66.61%. This suggests that its rule-based strategy is overly conservative, inadvertently stifling the circulation of legitimate information.

(3) Analysis of False Suppression analysis. The analysis of the false suppression rate (FSR) highlights the precision of the proposed algorithm. The Proposed Model achieves a false suppression rate is 0. It indicated that it avoids misclassifying and blocking normal information. The static threshold of Baseline 4 is prone to errors in edge condition data. The dynamic threshold mechanism of this model combines node credibility and degree normalization, effectively addressing this limitation. It ensures that the model is sensitive to high-risk content and tolerant of normal information.

Experimental results validate the efficacy of the improved comprehensive control model. This model can achieve good suppression of high-risk content without affecting normal information flow or reducing false positives, and the proposed model outperforms all baseline methods. It successfully overcomes the limitations of static threshold strategies and the inefficiency of simple probability models, thus proving its value in complex network governance.

4.2.4 Robustness Verification Across Multiple Network Topologies

In order to reduce the randomness bias caused by network topology generation, user node distribution, and information source location, this experiment used 5 sets of independent random seeds (Seed 42-46) for repeated experiments and took the average (Figure 6 and Table 8). This setting ensures the statistical significance of the experimental results and demonstrates the robustness of the model in different network environments.

Table 8: Average coverage and standard deviation of three types of risks

Risk Level	Baseline1	Baseline2	Baseline3	Baseline4	Proposed
LOW	216.54 ± 194.92	248.98 ± 198.5	99.59 ± 145.29	216.54 ± 194.92	201.71 ± 189.61
MEDIUM	202.82 ± 195.7	236.35 ± 200.85	73.62 ± 126.65	202.82 ± 195.7	86.85 ± 136.19
HIGH	208.87 ± 194.96	244.67 ± 198.72	71.96 ± 120.38	12.01 ± 61.31	4.11 ± 4

(1) Suppress the diffusion of high-risk information: The proposed model effectively suppresses high-risk content. The coverage range of all three baselines remains between 200 and 240, indicating that relying solely on probability decay and follower effects cannot effectively curb misinformation. Baseline 4 reduces the coverage to 12.01 through a static threshold, but this method lacks adaptability. In contrast, the proposed model yields a mean coverage of 4.11 (SD = 4). This

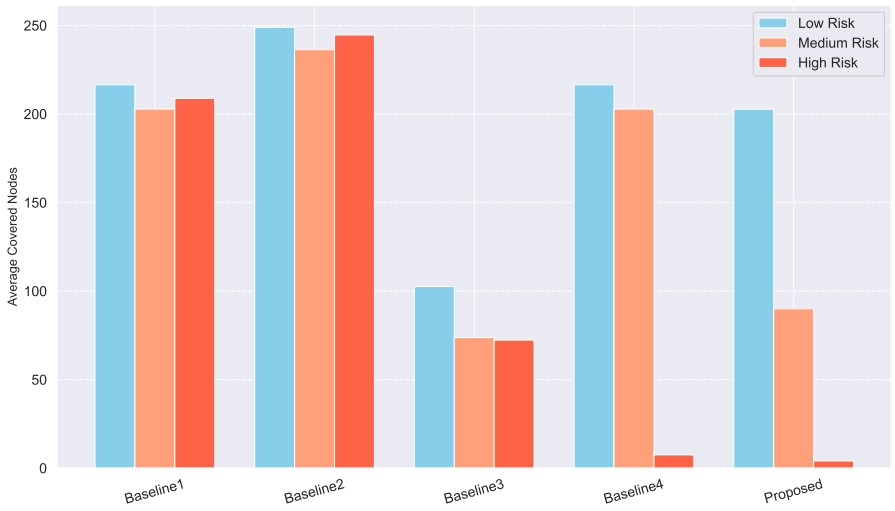


Figure 6: Comparison of average coverage of three types of risks in different topology networks

extremely low value, together with the minimal standard deviation. This effectively achieves a low-diffusion outcome for high-risk content.

(2) Reduce the spread of medium risk: For medium-risk content, the proposed model demonstrates clear risk differentiation. Baseline models maintain high coverage (200+), failing to distinguish medium-risk from low-risk content. By contrast, the proposed model reduces coverage to 86.85.This result indicates that the dynamic scoring mechanism functions as a speed bump in the medium-risk range. Notably, unlike Baseline 3 (coverage = 73.62), the proposed model avoids excessive suppression, striking a balance between risk control and information preservation.

(3) Maintain the spread of low-risk: For low-risk content, the proposed model successfully preserves normal information dissemination. It achieves coverage of 201, comparable to Baseline 1 (216.54) and substantially higher than Baseline 3 (99.59).

Regarding variance, all models exhibit large standard deviations (≈190) for low-risk scenarios. This is not a deficiency but rather a consequence of the network topology. The propagation results in BA scale-free networks largely depend on the position of the initial source node. The proposed model maintains coverage on par with the baselines while keeping standard deviations within a reasonable range, confirming its stability in normal dissemination scenarios.

(4) Observation on super spreader propagation behavior: To facilitate the diffusion of high-risk information by super nodes. The experiment used a super node with a degree of 87 as the propagation source, and compared the propagation coverage scale under five different mechanisms.

Figure 7(a) shows the comparison of propagation coverage range when a high-risk message R=0.85 is initiated with a degree of 87. Baseline 1 and Baseline 2 showed rapid early expansion during steps $3-5$, ultimately saturating near 90%. Baseline 3 is risk sensitive, although showing slow growth, it is still close to 70%, indicating that even with risk perception mechanisms, super spreaders can maintain cascading propagation through subsequent nodes. Baseline 4 significantly reduces the coverage range to only 7 nodes, but the static threshold lacks flexibility because the aggressive suppression mechanism almost stops all transmissions. The proposed model is close to zero on the one hand, the influence of super nodes and the risk assessment risk value will greatly reduce forwarding permissions; On the other hand, the expert nodes in the connection nodes of the super nodes can quickly identify real risks and trigger global blocking after consensus is reached among multiple nodes, effectively suppressing cascade propagation.

Figure 7(b) presents a critical edge case in which high-risk content (actual risk = 0.95) is misclassified as medium-risk (observed $R = 0.6$). Under this misclassification, Baselines 1-4 all exhibit consistently high coverage, as the underestimated risk value fails to activate their suppression thresholds. This reveals a major vulnerability: when the risk perception system is compromised, traditional control mechanisms become ineffective, enabling widespread dissemination of high-risk information. In contrast, although the observed risk remains below the activation threshold, the proposed model limits

final coverage to 61 nodes. This demonstrates that the combination of expert verification and dynamic thresholding reliably defends against cascading propagation of high-risk misinformation, particularly in scenarios involving super-spreader initiation and inaccurate risk assessment.

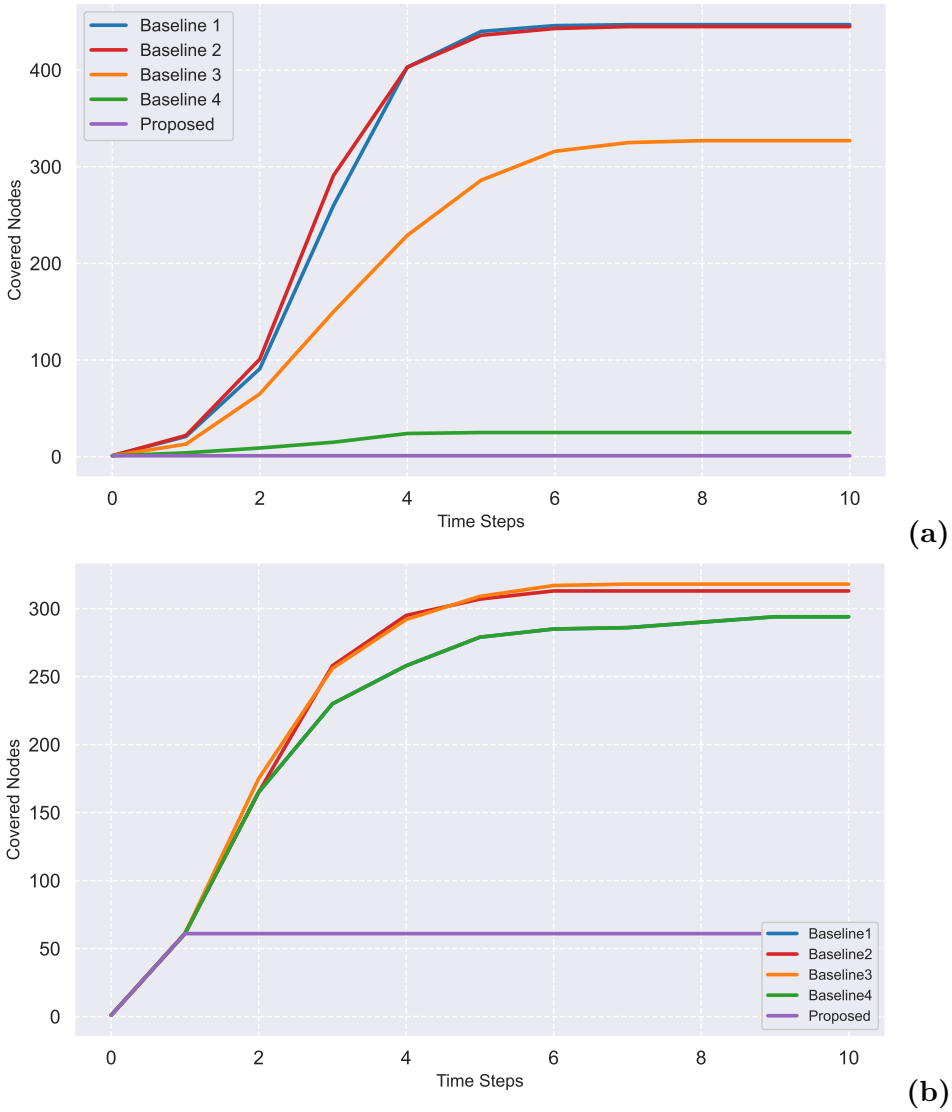


Figure 7: Information dissemination network coverage of super nodes. (a) Blocking effect of high-risk information dissemination; (b) Information dissemination when high risk is incorrectly classified as medium risk

5 Discussion

5.1 Value of evidence-augmented reasoning

Although LLM+Rules improves assessment rigor through three-dimensional scoring, its lack of external evidence proves particularly detrimental in the medical domain. Specifically, this variant substantially reduces recall for high-risk misinformation, yielding an unacceptable false negative rate (FNR) of 36.52%. This result highlights a fundamental limitation: neither prompt engineering nor static rules alone can sufficiently mitigate LLM hallucinations in professional settings. In contrast, the RAG mechanism with a layered evidence knowledge base serves as the key driver for improving high-risk recall. Compared to the rule-based variant (Variant B), incorporating external evidence retrieval significantly reduces FNR to 3.93% and increases recall by 32.59 percentage points. This confirms the necessity of external evidence anchoring in effectively suppressing model hallucinations.

When this evidence-enhanced risk assessment is integrated with our social network permission strategy, the framework demonstrates robust capability in curbing health misinformation at the network level. This synergistic effect establishes that the evidence base is not merely an optional enhancement but an indispensable component of reliable LLM-based health misinformation governance.

5.2 Dynamic trade-off between safety and usability

In the context of healthcare, the asymmetric costs of errors dictate our core design principle. the societal harm caused by missing a piece of high-risk misinformation far outweighs the cost of over-classifying benign content [36]. Guided by this principle, our framework prioritizes safety through a low tolerance for false negatives during parameter optimization. As shown in our experiments, although we conservatively misclassified 35 cases of moderate risk information as high-risk in our complete model. although the full conservatively misclassifies 35 cases moderate-risk information as high-risk. But it achieved a higher recall rate of 0.9607 for truly dangerous content, and the overall $macro - F1$ score of 0.8423 still indicates optimal performance.

Furthermore, unlike straightforward prohibition strategies, our complete model provides interpretable feedback for each restriction item through rules and evidence. This transparency enables publishers to understand why their content was restricted and to revise it accordingly. By doing so, the framework minimizes user friction. It also encourages higher-quality publishing behavior, and avoids the negative user experience typically associated with blanket prohibition strategies. Thus, the framework achieves a dynamic and ethically defensible balance between safety and usability. So it achieves a dynamic and ethically defensible balance between public health safety and the maintenance of normal information flow.

5.3 Mathematical rigor and topological robustness

Compared with existing work, the novelty of the framework lies in its integration of a tripartite coupling mechanism. A real-time dynamic suppression method for detecting and disseminating malicious information has been implemented. This method has a good theoretical basis.

Theoretical Soundness. Inspired by the theory of super spreaders in information epidemiology, we hypothesize that the harm of erroneous health information is nonlinearly positively correlated with the network centrality of its spreaders. The dynamic permission function $f(R, C, D)$ underpinning our framework is not heuristic. It has formally proven mathematical properties, including boundedness, monotonicity, and asymptotic stability. Among these, asymptotic stability is especially important. It mathematically guarantees that as a node's influence D grows very large (modeling a super-spreader), the permission value $f(R, C, D)$ approaches zero. This proof ensures that high-risk content cannot spread through highly influential nodes. This addresses a key limitation of traditional reputation-weighted methods, which do not account for risk amplification.

Architectural Integration. We integrate the reputation system, propagation model, and content safety module into a unified framework. Inspired by the Signal-to-Noise Ratio (SNR) concept from information theory, this tripartite coupling mechanism $f(R, C, D)$ advances the field. It replaces ad-hoc engineering combinations with a formally grounded architecture. This design enables real-time computation of forwarding permissions for each node before content propagation begins.

Empirical Robustness. We use the scale-free topology generated through the BA model. it is a simplified processing of real-world networks. But its priority connection mechanism and power-law distribution characteristics are typical features of real networks, which can meet the research needs of central nodes and propagation diffusion in the information dissemination process in this study. To validate robustness across different network topologies and remove bias from random generation, we ran extensive simulations using five independent random seeds (Seeds 42-46). As shown in Table 8, the model performs consistently well across all generated network instances. Importantly, the spread of high-risk misinformation is always suppressed to an average coverage of about 4.11 nodes (SD = 4), regardless of network structure. These results confirm that the strong performance of the DSM framework is stable and generalizes well to heterogeneous network environments.

5.4 Limitations and Future Work

Although the simulation results are good, we acknowledge several limitations of this study, which also outline our future research directions.

First, the current knowledge repository is mainly limited to the diabetes domain. This may restrict the generalizability of our framework to other disease areas. In future work, we will expand the evidence base to cover a wider range of health conditions. We will also build an automated pipeline for continuous knowledge updates, so that the system can keep pace with new medical evidence.

Second, our simulations use a BA scale-free network. This model captures the super-spreader phenomenon through its power-law degree distribution. However, real-world social networks have more complex structures, such as community clustering and hierarchical organization. In future studies, we will test our framework using empirical network data from platforms like Twitter or Weibo. This will allow a more realistic evaluation of its practical performance.

Third, our current user behavior model divides the population into four fixed roles: experts, fans, silent users, and normal users. This simplification is computationally convenient, but it does not fully capture the complex nature of human decision-making. In subsequent work, we will adopt more refined behavioral models. These will account for emotion-driven propagation and confirmation bias, both of which are known to strongly affect information diffusion on social media.

Fourth, we acknowledge that relying on synthetic or semi-synthetic datasets has limitations. However, for health misinformation risk stratification, this choice is a necessary step. It allows us to rigorously verify the safety boundaries of our control mechanisms. Real-world social media data often suffers from extreme class imbalance. High-risk misinformation makes up only a tiny fraction of random samples. Yet these rare cases are precisely the most important targets for safety systems to catch. To reduce the gap between synthetic environments and real-world scenarios, we used seed data from actual social platforms in our simulations. Despite these positive results, future work will pursue large-scale empirical validation on fully authentic datasets. This will further strengthen the generalizability of our findings.

6 Conclusion

This study proposed a Rule and Evidence-Enhanced Dynamic Suppression Framework (DSM) to combat health misinformation propagation in social networks. Unlike traditional binary classification or static rule-based methods, our framework introduces a mathematically grounded, dynamic permission control mechanism. The core contribution of this work is a tripartite coupling mechanism that dynamically integrates risk, reputation, and influence into a unified governance model. We provide formal mathematical proofs for the boundedness and asymptotic stability of this mechanism. We also validate its robustness across different network topologies through extensive simulations.

Overall, the DSM framework achieves a balanced trade-off between safety and efficiency. It suppresses high-risk misinformation to lowest propagation. At the same time, it avoids unnecessary interference with low-risk information. While our current validation uses synthetic networks and semi-synthetic datasets, the framework offers a theoretically sound, computationally feasible, and interpretable approach for translating public health safety principles into actionable AI governance. Our future work will focus on expanding the evidence base to cover major disease categories with automated pipelines for continuous updating. In addition, in simulating real social networks, scale-free networks can be transformed into more complex networks, which can further explore the interaction between network clustering effects and error information containment.

Funding

This work was supported by the 2024 Humanities and Social Sciences Research Project of Jiangxi Provincial Department of Education (TQ24103).

Author contributions

Jianying Xiong: Conceptualization, Methodology, Reviewing. Haoxiang Tang: Paper writing, Software, Data Visualization. Wei Lu: Data Collecting, Reviewing, Editing Leiyue Yao: Supervision, Validation.

Conflict of interest

No author associated with this paper has disclosed any potential or pertinent conflicts that may be perceived to have an impending conflict with this work.

References

[1] Calleja, N.; Nguyen, T.; Purnat, T. D. (2021). A public health research agenda for managing infodemics: Methods and results of the first WHO infodemiology conference, *JMIR Infodemiology*, 1(1), e30979, 2021.

[2] Zhou, J.; Xiang, H.L.; Xie, B.J. (2023). Better safe than sorry: A study on older adults' credibility judgments and spreading of health misinformation, *Universal Access in the Information Society*, 22(3), 957–966, 2023.

[3] Zareie, A.; Sakellariou, R. (2021). Minimizing the spread of misinformation in online social networks: A survey, *Journal of Network and Computer Applications*, 186, 103094, 2021.

[4] Aïmeur, E.; Amri, S.; Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review, *Social Network Analysis and Mining*, 13, 30, 2023.

[5] Yue, H.; Xiong, X.; Feng, M.; Szolnoki, A. (2025). Dynamic evolution of cooperation based on adaptive reputation threshold and game transition, *Chaos, Solitons & Fractals*, 199, 116693, 2025.

[6] Wang, B.; Wu, J.; Shi, Y.; Chen, W; Gong, F. (2025). Structured reflective reasoning for precise medical knowledge graph retrieval augmented generation, *Health Information Science and Systems*, 13, 76, 2025.

[7] Zhou, H.; Inkpen, D.; Kantarci, B. (2024). Evaluating and Mitigating Gender Bias in Generative Large Language Models, *International Journal of Computers Communications & Control*, 19(6), 6853, 2024.

[8] Zhao, X.; Tsang, S. (2023). How people process different types of health misinformation: Roles of content falsity and evidence type, *Health Communication*, 39(4), 741–753, 2023.

[9] Viviani, M.; Pasi, G. (2017). Credibility in social media: Opinions, news, and health information—A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5), e1209, 2017.

[10] Ng, J.; Dzisiak, D.; Saini, J. (2021). Cannabis for pain: A cross-sectional survey of the patient information quality on the Internet, *Journal of Cannabis Research*, 3, 36, 2021.

[11] Mahmud, M.R. (2025). Dissemination of misinformation during the dengue outbreak in Bangladesh, *The Electronic Library*, 43(3), 449–467, 2025.

[12] Chowdhury, R.; Srinivasan, S.; Getoor, L. (2020). Joint estimation of user and publisher credibility for fake news detection, In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, 1993–1996, 2020.

[13] Sun, Q.; He, Q.; Hu, H. (2025). Social network information diffusion model with diverse user forwarding thresholds, *Journal of Social Computing*, 6(4), 321–333, 2025.

[14] Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. (2019). dEFEND: Explainable Fake News Detection, In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405, 2019.

[15] Damoto, B.R.P.; Hasanah, M.; Haq, T.I.; Mafazy, M.M.; Agustinus, J.T.; Siahaan, D.; Purwitasari, D. (2024). Deep learning and ensemble approaches to misinformation detection in digital news: A systematic review, In *2024 IEEE 10th Information Technology International Seminar (ITIS)*, 151–156, 2024.

[16] Sikosana, M.; Ajao, O.; Maudsley-Barton, S. (2024). A comparative study of hybrid models in health misinformation text classification, In *Proceedings of the 2024 Workshop on Open Challenges in Online Social Networks (OASIS 2024)*, ACM, 18–25, 2024.

[17] Yuan, C.; Ma, Q.; Zhou, W.; Han, J.; Hu, S. (2019). Jointly embedding the local and global relations of heterogeneous graph for rumor detection, In *2019 IEEE International Conference on Data Mining (ICDM)*, Beijing, China, 796–805, 2019.

[18] Kwao, L.; Yang, Y.; Ativi, W.X.; Yussif, S.B. (2026). DCAGTN: Dual-snapshot causal-attention graph transformer with feedback mechanism for real-time rumor detection, *Knowledge-Based Systems*, 335, 115167, 2026.

[19] Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; Fung, P. (2023). Survey of hallucination in natural language generation, In *ACM Computing Surveys*, 55(12), 1–38, 2023.

[20] Li, X.; Li, G.; Zhao, Y.; Liang, Y.; Dong, Y. (2025). Exploring and comparing the use of large language models in supporting osteoporosis health consultations, *Clinical Interventions in Aging*, 20, 2133–2143, 2025.

[21] Rosen, K.L.; Sui, M.; Heydari, K.; Enichen, E.J.; Kvedar, J.C. (2025). The perils of politeness: How large language models may amplify medical misinformation, *NPJ Digital Medicine*, 8(1), 1–2, 2025.

[22] Vosoughi, S.; Roy, D.; Aral, S. (2018). The spread of true and false news online, *Science*, 359(6380), 1146–1151, 2018.

[23] Li, P.; Wu, B. (2024). The diffusion of platform self-operation on reputation-based two-layer network, *Industrial Management & Data Systems*, 124(3), 949–977, 2024.

[24] He, Q.; Zhang, Z.; Bi, T.; Fang, H.; Yi, X.; Yu, K. (2025). Adaptive rumor suppression on social networks: A multi-round hybrid approach, *ACM Transactions on Knowledge Discovery from Data*, 19(2), 31, 2025.

[25] Peng, H.; Pan, G.; Zhang, B.; Qian, C.; Zhong, M.; Li, S.; Han, J.; Zhao. D.; Wang, W. (2025). A sentinel-triggered competing dynamical strategy for misinformation containing in higher-order social networks, *Nonlinear Dynamics*, 113(25), 35167–35181, 2025.

[26] Coscia, M.; Cuzzocrea, A.; Shu, K. (2021). HawkEye: A robust reputation system for community-based counter misinformation, In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 188–192, 2021.

[27] Whalen, E.A.; Belarmino, A. (2023). Risk mitigation through source credibility in online travel communities, *Anatolia*, 34(3), 414–425, 2023.

[28] Cui, L.; Seo, H.; Tabar, M.; Ma, F.; Wang, S.; Lee, D. (2020). Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation, In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 492–502, 2020.

[29] Manouchehri, M.A.; Helfroush, M.S.; Danyali, H. (2021). A theoretically guaranteed approach to efficiently block the influence of misinformation in social networks, *IEEE Transactions on Computational Social Systems*, 8, 716–727, 2021.

[30] Liu, J.; Song, M.; Fu, G. (2024). Intervention analysis for fake news diffusion: An evolutionary game theory perspective, *Nonlinear Dynamics*, 112(16), 14657–14675, 2024.

[31] Barabasi, A.L.; Albert, R. (1999). Emergence of scaling in random networks, *Science*, 286(5439), 509–512, 1999.

[32] Broido, A.D.; Clauset, A. (2019). Scale-free networks are rare, *Nature Communications*, 10(1), 1017, 2019.

[33] Meng, F.; Xie, J.; Sun, J.; Hu, Y. (2025). Spreading dynamics of information on online social networks, *Proceedings of the National Academy of Sciences (PNAS)*, 122(4), e2410227122, 2025.

[34] Hinz, O.; Skiera, B.; Barrot, C.; Becker, J.U. (2011). Seeding strategies for viral marketing: An empirical comparison, *Publications of Darmstadt Technical University Institute for Business Studies*, 75(6), 55–71, 2011.

[35] Zhang, M.; Qin, S.; Zhu, X. (2021). Information diffusion under public crisis in BA scale-free network based on SEIR model — Taking COVID-19 as an example, *Physica A: Statistical Mechanics and its Applications*, 571, 125848, 2021.

[36] Kriebel, D.; Tickner, J. (2001). Reenergizing Public Health Through Precaution, *American Journal of Public Health*, 91(9), 1351–1355, 2001.



Copyright ©2026 by the authors. Licensee Agora University, Oradea, Romania.
This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.
Journal's webpage: http://univagora.ro/jour/index.php/ijccc/



This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Xiong, J.Y.; Tang, H.X.; Lu, W.; Yao, L.Y. (2026). A Rule and Evidence Enhanced Framework to Dynamic Suppression of Health Misinformation Propagation, *International Journal of Computers Communications & Control*, 21(6), 7470, 2026.
https://doi.org/10.15837/ijccc.2026.6.7470