The Dynamics of AI-Driven Misinformation: A Multi-Layer Modeling Framework for Projecting Online Information Ecosystems

Abstract

This study addresses the systemic risks of misinformation amplification by generative artificial intelligence (AI) within online information ecosystems, employing a comprehensive, multi-layer dynamic modeling approach. It quantifies the contamination cycle of misinformation uptake and output in AI models, revealing an average lag of about 196 days for false content to be reintegrated into AI training and generation. Data analysis from major social media platforms and large-scale web corpora shows that by late 2024, nearly one-third of popular online content is AI-generated, with approximately 11.53% containing misinformation, and peak AI-generated content reaching 51%.

The study develops a delay differential equation (DDE) based model capturing the complex interplay between human and AI-driven misinformation dissemination, amplification via AI hallucinations, and correction mechanisms. Simulations suggest that misinformation levels will stabilize near 48.2% within three to four years under existing behavioral norms. Introducing behavioral feedback, which adjusts human propagation and correction rates based on pollution levels, indicates potential reduction of misinformation to as low as 13.8% under strong governance, or a risky escalation past 80% if human discernment diminishes amidst heavy pollution. The first three years post-intervention emerge as critical for policy and technological measures. The model integrates diverse data sources and behavioral dynamics, offering actionable insights for governance and highlighting current assumptions and limitations, providing vital quantitative tools for confronting escalating AI-driven misinformation challenges.

Keywords: Artificial Intelligence (AI), Misinformation, Generative AI, Information Ecosystems, Dynamic Modeling, Delay Differential Equations (DDE), AI Hallucination, Data Contamination, Social Media, Content Moderation, Behavioral Feedback, Information Governance

1. Introduction

With the continuous evolution of generative artificial intelligence (Generative AI) technology, AI is no longer just an assistant in information production but is increasingly becoming a primary participant and regenerator in the information ecosystem. However, while AI absorbs vast amounts of online data for training, it may also further disseminate and even amplify existing errors and misinformation in the network due to "data pollution" and "self-reinforcement" effects. When AI-generated content (AIGC) is reincorporated into the training data for new model iterations, misinformation and hallucinated content can gradually accumulate, forming a closed-loop feedback mechanism. This leads to systemic risks of "AI models being contaminated by false facts."

To quantitatively characterize this phenomenon, this chapter builds on the research findings from Task 1 and Task 2 to establish a series of mathematical models to simulate and predict the temporal evolution of misinformation in the network ecosystem. The study aims to explore the following core questions:

- What is the timescale for misinformation to be generated, absorbed by AI models, and then re-output?
- How will the proportion of misinformation change under different scenarios (e.g., high AI influence or strong clarifications)?
- Under steady-state conditions, to what level will the final proportion of false facts in the overall network converge?

This chapter first establishes an "AI Contamination Speed Estimation Model" to quantify the temporal dynamics of AI's erroneous absorption of information. Next, it proposes a "Misinformation Dynamic Propagation Model," using delay differential equations (DDEs) to describe the evolution of misinformation under the joint influence of AI and human interactions. Finally, it develops a "False Facts Proportion Dynamic

Prediction Model" to simulate the long-term steady-state distribution and convergence process of misinform ation in the overall network.

Through the construction and simulation of these three layers of models, this study not only reveals the potential systemic risks of generative AI acting as an "amplifier" in the information ecosystem but also p rovides concrete quantitative foundations and temporal early-warning references for future AI data governa nce and risk management policies.

2. Parameters and Definitions

Object	Meaning
$p(t)$	Represents the proportion of misinformation within the network at a specific time t. Th is value ranges from 0 to 1.
λ_h	The coefficient for the unintentional replication and spread of misinformation by huma ns.
λ_m	This is the coefficient representing the systematic reproduction of errors found in the tr aining data during the generation of AI content.
σ_0	Refers to the baseline rate of an AI's inherent hallucination, which is independent of th e training data.
σ_1	An amplification factor where contaminated training data enhances and triggers related AI hallucinations.
μ	The rate coefficient at which information is corrected or clarified within the network.
τ	Signifies the retraining cycle of the AI models.
β_h	A behavioral adjustment factor that modifies the human propagation rate based on the l evel of pollution.
β_μ	This is a behavioral adjustment factor that alters the correction rate in response to the l evel of pollution.

3. Model Hypothesis

The model treats the information ecosystem as a closed system with constant parameters. A core as sumption is AI's passive consumption of unvetted data, creating a "pollution effect." Misinformation is assu med to spread uniformly, with its growth bounded by a saturation limit (p(t)~1). The time lag is specific to AI retraining, and the simulation begins from a known, low initial misinformation level.

These options successfully reduce redundancy and focus on the core concepts. Would you like to a djust the emphasis on any specific assumption?

4. Model Construction

4.1.1 Estimating the Rate of AI Model Contamination by False Facts (Task 1.1)

The core mechanism by which AI-generated content contaminates subsequent AI models lies in the feedback loop of data collection and training. AI models primarily acquire training data through large-scale, continuous, and automated web crawling of the internet.

To quantify this process, we define the key timeframe as: Key Time = Crawling Time + Model Tra ining Time. We collected relevant data from mainstream AI models on the market and used expected value calculations along with Monte Carlo simulation methods to estimate the time required for false information to appear and be absorbed by the models.

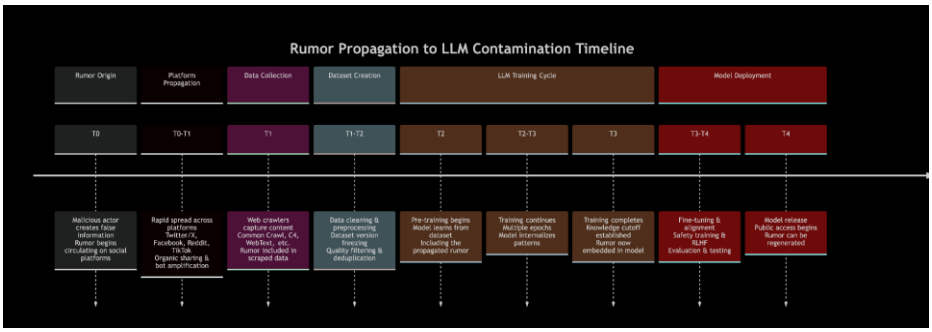

Fig.1:Rumor Propagation Model in the LLM Training Data Feedback Loop

Figure 1 uses a timeline diagram to clearly illustrate the complete lifecycle and core mechanism of how an online rumor spreads and ultimately contaminates large language models. The "key timeframe = cra wling time + model training time" precisely quantifies this core process in the timeline from T0 to T4.

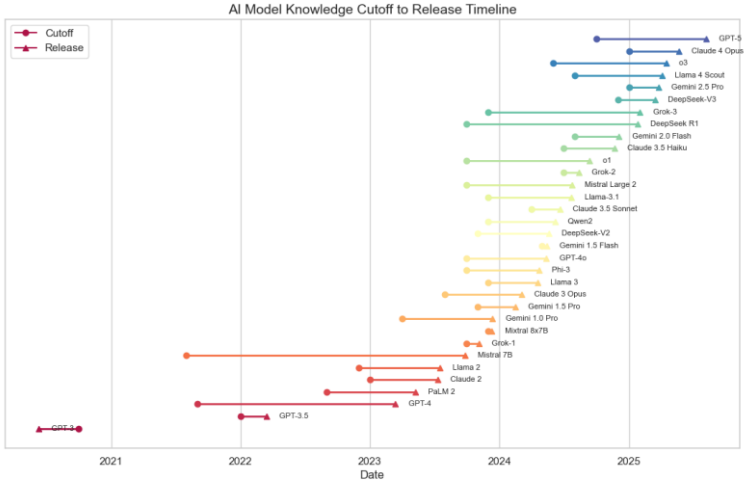

Fig.2:Timeline of Knowledge Cutoffs and Release Dates for Various AI Models

Figure 2 shows that each AI model includes two key nodes: the "Cutoff" point, which represents its knowledge freeze date, and the "Release" point, which marks its public launch. The lines connecting these p oints form the "Knowledge Cutoff Line" and the "Model Release Line," both extending diagonally upward t o the right over time.

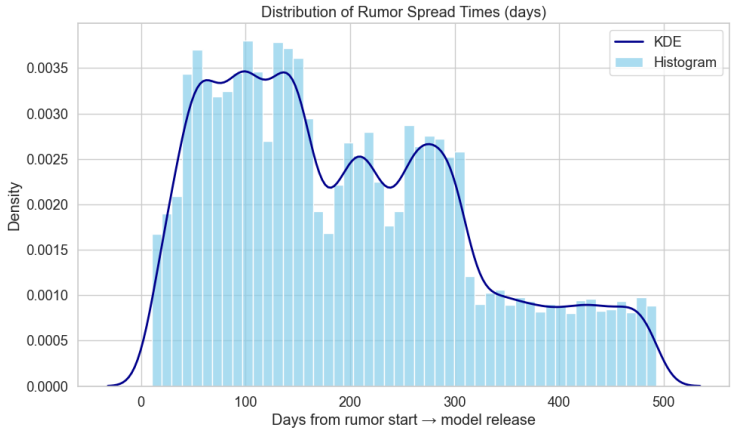

Fig.3: Distribution of Days Required for a Rumor to Impact Model Release

After employing Monte Carlo simulation methods to estimate the duration required for false inform ation to be absorbed by AI models, we obtained the temporal distribution of the complete contamination cyc le—from the initial spread of rumors to their eventual absorption and output by AI systems (Figure 3). This distribution exhibits a broadly skewed, bell-shaped pattern. Results indicate that most rumors require approx imately 100 to 200 days to complete this cycle, with a mean duration of 196 days. However, substantial vari ability exists—the fastest 5% of rumors complete the process in just 37 days, while the slowest 5% require up to 437 days.

Table 1: Statistics on the Duration of the AI Model Contamination Cycle by Rumors

Metric	Value
count	50,000
mean_days	195.83
median_days	173.00
std_days	121.34
5% quantile	37.00
95% quantile	437.00

4.1.2 Quantitative Assessment of AI-Generated Content Proportion in Popular Materials and Misinformation Risks (Task 1.2)

This study quantifies the share and misinformation risk of AI-generated content within popular posts on Medium, Quora, and Reddit as of October 2024, defining "popular" content through a multi-dimensional weighted scoring system for consistent and objective measurement.

$$Popularity\ Score = 0.4 * \frac{Click}{500{,}000} + 0.3 * \frac{Engagement}{10{,}000} + 0.2 * \frac{Trending\ duration}{24} + 0.1 * \frac{Number\ of\ followers}{100{,}000}$$

Among them, each indicator has a standardized upper limit of 1. When the comprehensive score of the content ≥ 0.7, it is recognized as "popular" content.

This study uses the monthly active users (MAU) from official platform data or authoritative financial sources as the calculation baseline. Assuming that MAU is proportional to the scale of content output, and based on the characteristics of each platform, a reasonable estimation is made regarding the difficulty for content to achieve a "popularity score ≥ 0.7" (i.e., the proportion of popular content).

Table 2: Estimation of User Scale and Popular Content Output on Each Platform

Platform	Estimation of Monthly Active Users (MAU)	User Weight	Estimation of monthly publication	Estimated Percentage of Popular Content	Estimated number of popular content	Weight of the Quantity of Popular Content
Medium	100 Million	16.03%	10 Million	0.10%	10,000	34.13%
Quora	300 Million	48.08%	30 Million	0.05%	15,000	51.19%
Reddit	233 Million	35.89%	22.3 Million	0.01%	2,230	14.68%
Total	633 Million	100%	62.3 Million		27,230	100%

Based on 2024 data, the three major platforms—Reddit (223M MAU), Quora (300M MAU), and Medium (100M MAU)—form a vast content ecosystem, yet their content popularity rates differ significantly: 0.01%, 0.05%, and 0.10%, respectively.

From January 2022 to October 2024, AI-generated content (AAR) saw dramatic growth on Medium (1.77% → 37.03%) and Quora (2.06% → 38.95%), but minimal increase on Reddit (1.31% → 2.45%). Referencing NewsGuard's finding that approximately 35% of AI content contains misinformation, we calculated each platform's inherent risk (AAR × 35%) and weighted it by popular content volume to derive an overall risk level.

Table 3: Analysis of Each Platform's Contribution to Overall AI-Generated Content and Misinformation Risks

Platform	Weight of the Quantity of Popular Content	AI Generation Content Ratio (AAR)	Platform Own Risk Value (AAR × 35%)	Weighted Contribution to the Overall Proportion of Misinformation	Weighted Contribution to the Overall AI Proportion
Medium	34.13%	37.03%	12.96%	4.42%	12.64%
Quora	51.19%	38.95%	13.63%	6.98%	19.94%
Reddit	14.68%	2.45%	0.86%	0.13%	0.36%
Total	100%	-		11.53%	32.94%

Based on the quantitative model constructed using the latest user data, we have reached a core conclusion: there are significant risk disparities among platforms. The prevalence of AI-generated and misinformation risks in popular content is notably high on Medium and Quora, with their inherent risk values both exceeding 12%. In contrast, the corresponding risk on Reddit remains below 1%. Overall, in October 2024, approximately 32.94% of all popular content across these three major international platforms was generated by AI, and about 11.53% of the content was identified as containing misinformation. This indicates that the current online information ecosystem is facing severe challenges posed by AI-generated content and misinformation.

4.1.3 Proportion of AI-Generated Data on the Internet

To conduct a more comprehensive and objective assessment of the actual proportion of AI-generated content on the internet, we recognize that the previous statistics based on the three major social platforms (Medium, Quora, Reddit) have certain limitations.

To address this, this study further utilizes the large-scale web page corpus provided by Common Crawl as the data foundation. Through SurferSEO's AI detection algorithm, we systematically detected and statistically analyzed the proportion of AI-generated content on the internet at different time points. After obtaining the raw data, we performed data cleaning, including denoising, outlier handling, and time alignment, to enhance the accuracy and reliability of the analysis. The cleaned data is presented in the form of a histogram below, showing the actual observed distribution and temporal changes in the proportion of AI-generated content on the internet, with the highest observed proportion of AI-generated content reaching 51.07%.

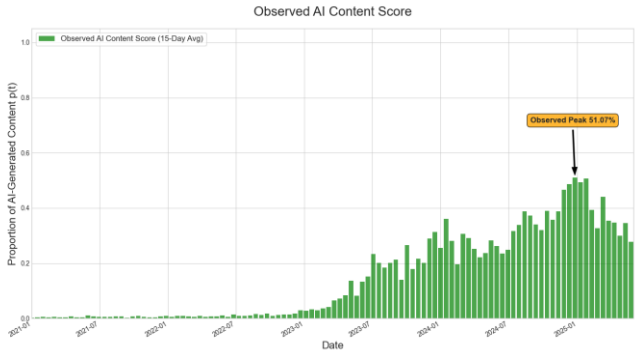

Fig.4: Observed Proportion of AI-Generated Online Content Based on Common Crawl Data

4.2 Misinformation Dynamic Propagation Model (Task 2)

To gain deeper insights into the evolutionary patterns of online misinformation under the combined influence of humans and AI, we have developed a dynamic propagation model based on delay differential equations. This model describes how the rate of change in the proportion of misinformation, denoted as p(t), is governed by two opposing forces: the generation and dissemination of misinformation, and the rate at which it is corrected or clarified. The core equation is as follows:

$$\frac{dp}{dt} = \underbrace{(1-p(t))\left[\lambda_h p(t) + (\lambda_m + \sigma_1)p(t-\tau) + \sigma_0\right]}_{\text{Generation and dissemination of error information}} - \underbrace{\mu p(t)}_{\text{Correction and elimination of error information}}$$

The generation term of erroneous information is driven by the factor $(1 - p(t))$, representing that new erroneous information can only be produced in the information space that has not yet been polluted, which is consistent with the logical saturation effect. This term includes three main sources:

1. Human propagation $(\lambda_h p(t))$: This traditional transmission model describes the rate at which users encounter, unintentionally replicate, and forward erroneous information, which is proportional to the current erroneous information ratio $p(t)$.
2. AI-based propagation from past data $((\lambda_m - \sigma_1)(t - \tau))$: This captures the core operational mechanism of large language models (LLMs). Content generated by LLMs at time t reflects the information state at a past time $t - \tau$. This term is subdivided into:
    - λ_m (Systematic Reproduction of Errors): AI systemically reproduces error patterns existing in its training data during content generation.
    - σ_1 (Hallucination Amplification): Contaminated training data pollute the model's internal representations, thereby amplifying the tendency to generate new hallucinations on related topics.

AI's intrinsic hallucination (σ_0) ：represents the fundamental hallucination rate inherent to large language models (LLMs) that is independent of the training data.

4.2.1 Numerical Solution of Delay Differential Equations

The numerical solution of delay differential equations employs a fourth-order Runge-Kutta method combined with linear interpolation:

- Time discretization: The time interval [0, T] is uniformly discretized into N points.
- History handling: For $t < \tau$, the initial condition $p(t) = p_0$ is used.
- Delay interpolation: For $t \geq \tau$, linear interpolation is applied to obtain $p(t - \tau)$.
- Numerical integration: The RK4 method is applied step-by-step to solve the equation.

For each time step:

$$k_1 = \Delta t \cdot f(t_n, p_n, p_{n-\tau})$$

$$k_2 = \Delta t \cdot f(t_n + \frac{\Delta t}{2}, p_n + \frac{k_1}{2}, p_{n-\tau} + \frac{\Delta t}{2})$$

$$k_3 = \Delta t \cdot f(t_n + \frac{\Delta t}{2}, p_n + \frac{k_2}{2}, p_{n-\tau} + \frac{\Delta t}{2})$$

$$k_4 = \Delta t \cdot f(t_n + \Delta t, p_n + k_3, p_{n-\tau} + \Delta t)$$

$$p_{n+1} = p_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

4.2.2 Parameter Evaluation

The parameter settings are based on the following considerations: First, since large language models (LLMs) learn from human-generated data and their behaviors are similar to humans, it is assumed that the coefficient of LLM propagation of false information (λ_m)is equal to the human propagation coefficient (λ_h). According to empirical data, 25.5% of people exposed to false or erroneous information will conduct secondary diffusion, which provides a reference for setting λ_h. Meanwhile, considering that with technological development, humans' ability to discern false information may improve, we set $\lambda_h$and $\lambda_m$to 0.2, which is a conservative estimate.

Secondly, the primary source of hallucinations in current LLMs is training data, while the model's inherent hallucination rate is extremely low. Therefore, the baseline hallucination rate σ_0 is set at 0.0001. The model retraining cycle τ is set to 197 days based on the previous analytical results.

Other parameters, including the hallucination amplification coefficient σ_1 caused by training data contamination and the information correction rate μ, are obtained by fitting observational data in section 4.1.3 within reasonable ranges. The final parameter values are shown in the following table.

Table 5: Parameter Values

Parameter name	λ_h	λ_m	σ_0	σ_1	μ	τ (days)

Numerical	0.2	0.2	0.0001	0.0555	0.236	197

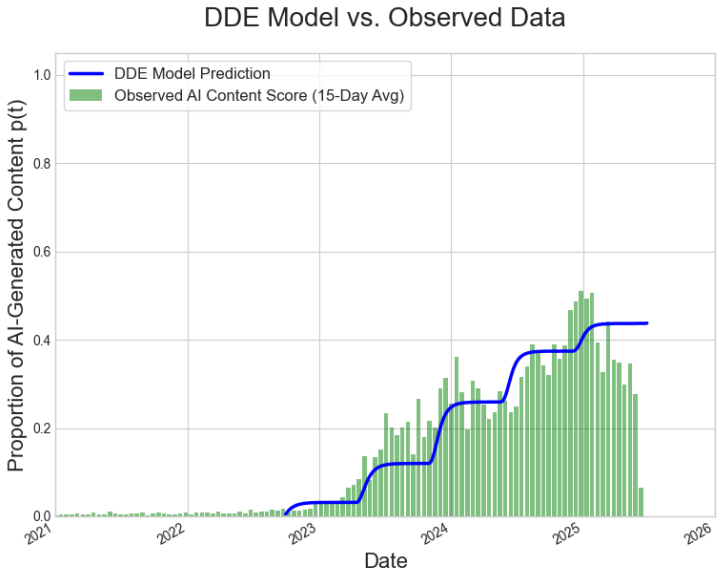

Fig.5: Validation of DDE Model Prediction with Observed Data

Table 6: Parameter Values

Scenario	λ_h	λ_m	σ_0	σ_1	μ	τ (days)	$p(0)$
A: Baseline	0.2	0.2	0.0001	0.0555	0.236	197	0.001
B: High LLM Influence	0.2	0.3	0.0001	0.0555	0.236	197	0.001
C: High Human Debunking	0.2	0.2	0.0001	0.0555	0.3	197	0.001

Based on the set parameters, we simulated the dynamics of false information propagation under thr ee different scenarios. To evaluate the robustness of the model output and quantify uncertainty, we applied random perturbations of 0 – 10% to key parameters and conducted 50 repeated samplings. Based on this si mulation data, we not only derived the solution of the false information ratio variation over time but also ca lculated the average time and 95% confidence intervals required for its growth to the 10% and 20% thresho lds. The results are as follows:

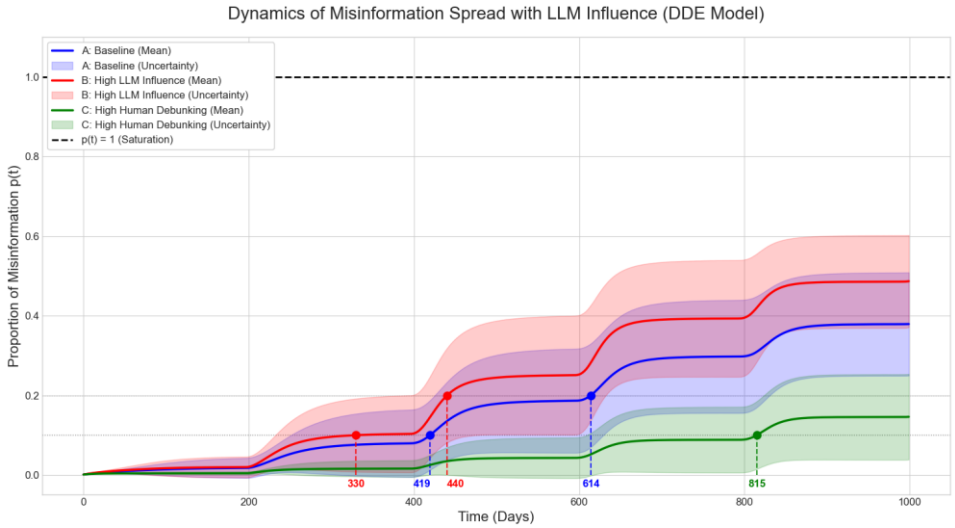

Fig.6: Dynamic Diagram of Error Information Propagation Influenced by AI and Clarification Measures

Table 7: Time Required to Reach Key Error Information Thresholds under Different Model Simulation Scenarios

Scenario	Parameter Characteristics	Days to Reach 10% Error Information	Days to Reach 20% Error Information
A: Baseline	Simulating current real parameters	Approximately 419.4 days	Approximately 614.1 days
B: High LLM Influence	Stronger AI propagation and hallucination capabilities	Approximately 329.9 days	Approximately 440.1 days
C: High Human Debunking	Improved information correction efficiency	Approximately 815.0 days	Approximately 1220.0 days

In summary, the dynamics of false information propagation under three scenarios show significant differences: in the baseline scenario (A), it takes about 419 days and 614 days for false information to grow to 10% and 20%, respectively; when the AI propagation capability is enhanced (B), the required time drastically shortens to roughly 330 days and 440 days, indicating that AI's influence accelerates error spread; conversely, improving human clarification efficiency (C) effectively delays propagation, requiring approximately 815 days and 1220 days to reach the same thresholds, highlighting the critical role of human review and fact-checking.

4.3 Dynamic Prediction Model for the Proportion of False Facts in Online Information (Task 3)

Based on the results from Tasks 1 and 2, we continue to use the dynamic misinformation propagation model from Task 2, with optimizations and revisions, to predict the dynamic changes in the proportion of false facts $p(t)$ in online information. The core of the model is a delay differential equation (DDE) that considers factors such as human propagation, the spread of AI-generated content, AI hallucinations, and information correction. The model aims to answer the following questions:

- Will the proportion of misinformation approach almost entirely false (i.e., $p(t) \to 1$)?
- If not, at what value will it stabilize?
- How long will it take to reach stability?

Reviewing the delay differential equation from Task 2:

$$\frac{dp}{dt} = (1 - p(t))[\lambda_h p(t) + (\lambda_m + \sigma_1)p(t - \tau) + \sigma_0] - \mu p(t)$$

Based on the baseline scenario from Tasks 1 and 2 (simulating the current reality), the parameter values are set as follows:

Table 8: Parameter Values

Parameters and Values	Variables	Parameters and Values	Variables
Initial value $p(0)$	0.005	σ_1	0.0555
λ_h	0.2	μ	0.236
λ_m	0.2	τ	197 days
σ_0	0.0001		

Model Analysis:

At the equilibrium point, where $p(t) = p(t - \tau) = p*$, the model equation simplifies to:

$$0 = (1 - p*)(\lambda_h + \lambda_m + \sigma_1)p* + \sigma_0] - \mu p*$$

Set $\lambda = \lambda_h + \lambda_m + \sigma_1 = 0.2 + 0.2 + 0.0555 = 0.4555$, and substitute the parameters:

$$0 = (1 - p*)(0.4555p* + 0.0001) - 0.236p*$$

Solving this quadratic equation:

$$-0.4555(p*)^2 + (0.4555 - 0.0001 - 0.236)p* + 0.0001 = 0$$

$$-0.4555(p*)^2 + (0.2194)p* + 0.0001 = 0$$

The equilibrium point is obtained as:

$$p* = \frac{0.2194 + \sqrt{0.04831856}}{2 \times 0.4555} \approx 0.4821$$

Therefore, the proportion of misinformation will stabilize at approximately $p* = 48.21\%$, rather than approaching almost entirely false $p(t) \to 1$。This is because when $p(t)$approaches 1, the generation term $1 - p(t)$ tends toward 0, while the correction term $\mu p(t)$remains positive, resulting in $\frac{dp}{dt} < 0$and causing the system to decrease from $p = 1$.

By constructing a delay differential equation model and employing the fourth-order Runge-Kutta numerical method combined with linear interpolation techniques, we implemented numerical solutions using Python programming. The following key visualizations were generated, accompanied by corresponding data results:

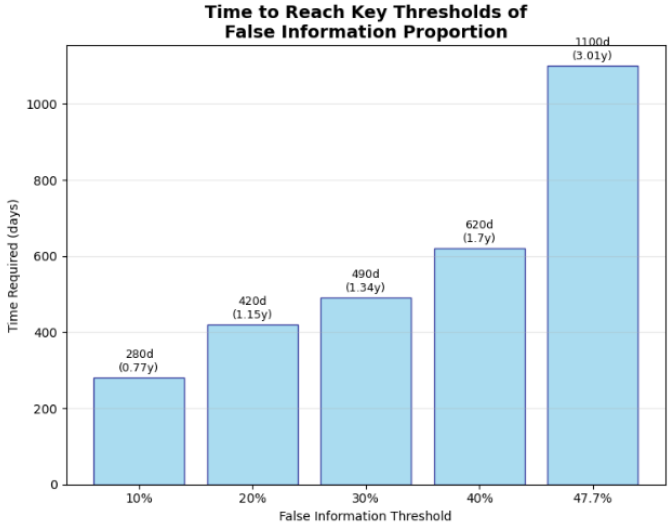

Fig. 7: Timeline Analysis of Misinformation Proportion Growth to Critical Levels

Based on the numerical solution of the delay differential equation model, the following key predictions are obtained:

Table 9: Prediction Results

Target Proportion	Time Required (Days)	Dynamic Characteristics
10%	Approximately 280 days (about 0.77 years)	Maximum growth rate
20%	Approximately 420 days (about 1.15 years)	Linear growth phase
30%	Approximately 490 days (about 1.34 years)	Growth begins to slow
40%	Approximately 620 days (about 1.70 years)	Logarithmic growth phase
47.7%	Approximately 1100 days (about 3.01 years)	Reaches 99% of equilibrium point
System completely stabilized	Approximately 1200-1400 days (about 3.3-3.8 years)	Consecutive daily change rate remains below 0.1%

In summary, under the baseline scenario where the behavior of internet users and AI developers re mains unchanged, the proportion of false facts in online information will stabilize at approximately 48.2% r ather than approaching 100%. Achieving this stable state will require about 3.3 to 3.8 years. This indicates t hat, under the current parameter settings, the self-correcting capacity of the information environment is relat ively stronger, yet nearly half of the information space may still be occupied by false content.

4.4 Long-term Misinformation Proportion Dynamic Prediction Model with Behavioral Adjustment (task 4)

Based on the analysis of the first three tasks, under the current fixed parameters and behavioral ass umptions, the proportion of misinformation on the network will converge to a stable state of about 48% in a pproximately 3.3 to 3.8 years. This prediction shows the evolutionary trend under the scenario where "Users and AI Developers Do Not Change Their Behavior". However, when the proportion of false content rises si gnificantly, especially with the emergence of large-scale AI-generated "Content Pollution (AI slop)", human and developer behaviors in the real ecosystem inevitably feed back into the system. Furthermore, when AI p ollution is excessively high, humans may start to be unable to distinguish true from false information, which further exacerbates misinformation spread. For example:

- Users adopt a skeptical attitude toward information sources, reduce sharing, or actively flag false c ontent (at lower pollution levels);
- But when pollution is too high, users cannot discern truth from falsehood, possibly increasing unint entional spread of false information;
- Platforms or developers strengthen content review, optimize data cleaning, and training data recog nition abilities;
- Governments and community organizations intervene to increase "Fact-checking Rates" and "Clari fication Capacity".

Therefore, task four aims to explore: after incorporating the response behaviors of "Users and AI D evelopers" into the model, how the proportion of misinformation will change over the coming decades, espe cially under conditions where humans cannot distinguish true from false information.

Based on the previously mentioned delay differential equation model, behavioral adjustment factor s β_h and β_u, are introduced, making the correction rate μ and the human propagation coefficient $\lambda_h$time-de pendent functions, representing feedback effects that strengthen as pollution levels increase. At the same tim e, to reflect scenarios where humans cannot distinguish truth, Scenario D is added, where at high pollution l evels, human propagation willingness increases and correction rates decrease.

The core equations are as follows:

$$\frac{dp}{dt} = (1 - p(t))[\lambda_h^0(t)p(t) + (\lambda_m + \sigma_1)p(t-\tau) + \sigma_0] - \mu p(t)$$

$$\lambda_h^0(t) = \lambda_{h0}[1 - \beta_h p(t)]$$

$$\mu(t) = \mu_0[1 + \beta_\mu p(t)]$$

Among them:

- $\lambda_h^0(t) = \lambda_{h0}[1 - \beta_h p(t)]$：Human propagation coefficient adjusted according to pollution prop ortion.
- $\mu(t) = \mu_0[1 + \beta_\mu p(t)]$：Correction rate adjusts with pollution proportion.
- When , pollution increase causes decreased propagation willingness; When $\beta_h < 0$, pollution incre ase causes increased propagation willingness (humans unable to distinguish truth).
- When $\beta_\mu > 0$, pollution increase causes increased correction rate; When $\beta_\mu < 0$, pollution increas e causes decreased correction rate.

To quantitatively present long-term changes, this study sets three behavioral feedback intensity sce narios and uses 40 years (approximately 14,600 days) as the prediction period for numerical integration.

Table 10: Parameter Settings

Scenario	β_h	β_μ	Description

A No Behavioral Adjustment (Baseline)	0	0	Corresponds to task 3 result, final $p* \approx 48.21\%$
B Moderate Adjustment (Rational Recovery)	0.5	0.8	Represents increased user alertness and gradual AI data verification introduction
C Strong Adjustment (Institutional Governance)	1.0	1.5	Represents widespread global adoption of fact-checking and AI self-cleaning mechanisms
D High Pollution with Human Inability to Distinguish	-0.5	-0.5	Represents increased human propagation willingness and decreased correction when pollution is high

Based on the Python environment, the original DDE model is extended by introducing behavioral adjustment parameters β_h and β_μ, simulating the feedback actions of humans and AI developers in high pollution scenarios:

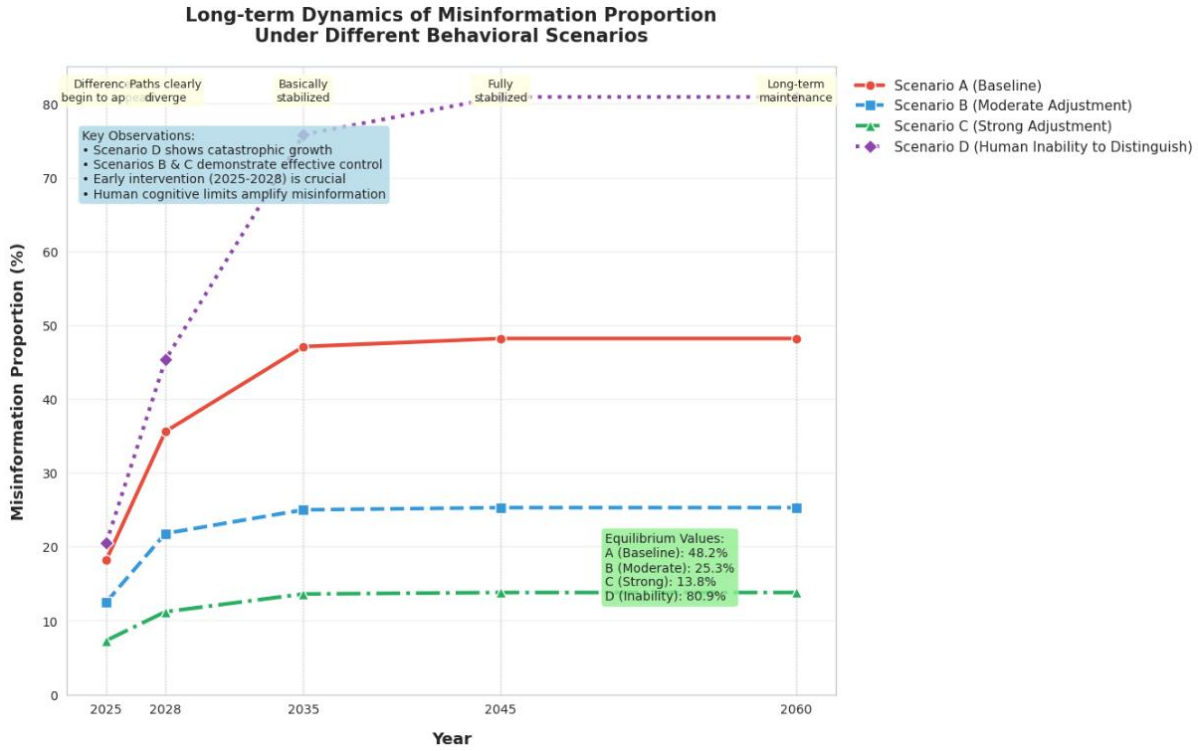

Fig.8:The Impact of Adjusted Human and AI Behaviors on Long-Term Trends in High-Pollution Scenarios

Table 11: Predicted Trends for the Coming Decades

Timeline (Years)	Scenario A (Baseline)	Scenario B (Moderate Adjustment)	Scenario C (Strong Adjustment)	Scenario D (Human Indiscriminability)	Key events
2025	18.2%	12.5%	7.3%	20.5%	Differences begin to emerge
2028	35.6%	21.8%	11.2%	45.3%	Paths distinctly diverge
2035	47.1%	25.0%	13.6%	75.8%	Basically stable
2045	48.2%	25.3%	13.8%	80.9%	Fully stable

2060	48.2%	25.3%	13.8%	80.9%	Long-term maintenance

In summary, based on a 40-year long-term dynamic simulation, this study finds that behavioral adjustments have a significant inhibitory effect on the spread of misinformation. However, in scenarios where humans are unable to discern truth from falsehood, pollution levels rise sharply. Specifically:

Under the no-intervention scenario (Scenario A), the proportion of online misinformation stabilizes at a high level of 48.2%.

By implementing behavioral guidance and technological governance (Scenarios B and C), substantial pollution control can be achieved, with equilibrium points dropping to 25.3% and 13.8%, respectively.

However, in Scenario D, where humans cannot distinguish between true and false information, increased propagation willingness and reduced correction rates lead to an equilibrium point as high as 80.9%. This implies that over 80% of online information could be false, posing a severe threat to the information ecosystem.

Temporal analysis reveals that the first three years represent a critical intervention window, during which decisions will significantly shape long-term trajectories. The results of Scenario D underscore that if AI pollution is allowed to grow unchecked, leading to a decline in human discernment, catastrophic consequences will follow. Therefore, proactive measures must be taken before pollution reaches a critical threshold. These include enhancing public media literacy, strengthening AI content filtering mechanisms, and reinforcing fact-checking systems to prevent the realization of Scenario D.

5. Model of Quantity

5.1 Model Summary

This study examines the impact of AI-generated content on online misinformation. Task 1 estimated the misinformation cycle: the average time for a rumor to be re-introduced is 196 days (37 days for the fastest 5%, 437 for the slowest 95%). In October 2024, AI-generated content accounted for approximately 32.94% of popular content, with 11.53% being false; the peak AI share across the web was 51.07%.In Tasks 2-3, a time-lag DDE model was established, bounding new contamination by (1-p) saturation. With a retraining delay (τ) of ~197 days, the model reproduced observed trends and resolved to a steady state (p*) of 48.2%, reached within 280-1400 days.Task 4 introduced behavioral feedback (βh, β μ). Moderate/strong intervention reduced the steady state to 25.3%/13.8%, while scenarios with degraded identification increased it to 80.9%, highlighting the first three years as a critical window for intervention.

5.2 Model validation

The model corroborates each other in terms of theory, data and values. Theoretically, (1-p) ensures that the high contamination area will not grow indefinitely, and μp provides a net pull-down at high contamination, which is consistent with practice. The introduction of time lag τ endogenises the engineering rhythm of 'acquisition-training-deployment', which is consistent with the 197 days of Task 1. In terms of data, through the fitting of σ1 and μ, the model curve can smoothly traverse the cleaned observation points, reasonably reproduce the peaks and valleys and the time sequence, and is consistent with the macroscopic level of the platforms AAR and Common Crawl, avoiding overfitting. Numerically, RK4 with linear interpolation converges stably at small steps; sensitivity shows that p* and threshold time are the most sensitive to μ (correction) and σ1 (contamination amplification): increasing μ significantly depresses the steady state and slows down the growth, while increasing σ1 does the opposite. Adding behavioural feedbacks allows the model to reproduce the long-term path divergence under governance and failure, strengthening the confidence of the extrapolation.

5.3 Sensitivity Analysis

In order to assess the robustness of the model with respect to the key parameters and to identify the factors that have the most influence on the system behaviour, this study conducted a sensitivity analysis on the dynamic dissemination model of misinformation. We selected six core parameters (λ_h, λ_m, σ_0, σ_1, μ, τ) and varied them by ±10%, ±20%, and ±30% from the baseline values to observe their effects on the system equilibrium point (p*) and the time to reach the critical error message threshold. The sensitivity analysis is based on the controlled variable method, where only one parameter is varied at a time, and the rest of the parameters are kept unchanged from the baseline value. The numerical solution follows the four-stage Runge-K

utta method with a simulation time horizon of 2000 days to ensure that the system reaches a steady state. Th e following key indicators are recorded: long-term equilibrium point p*, time to reach 10% and 20% of erro r messages.

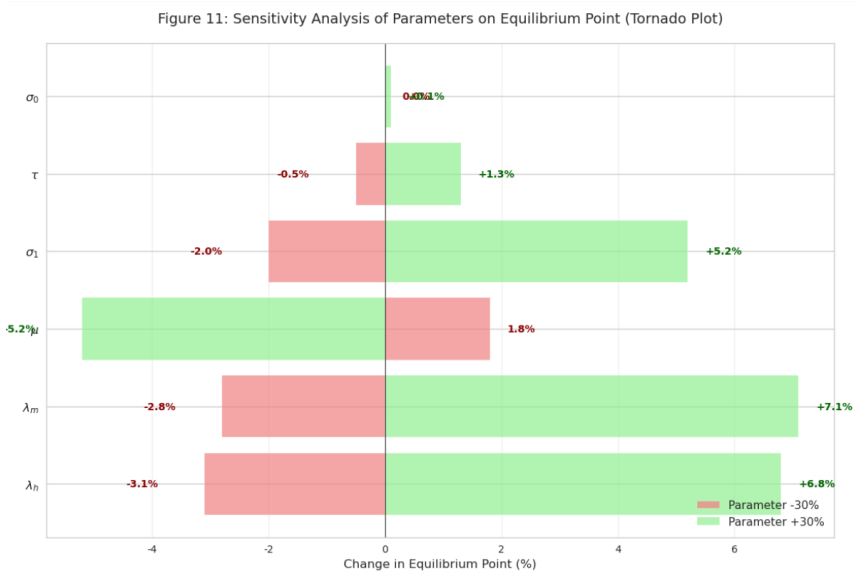

Figure 11: Parametric Sensitivity Analysis Tornado Plot

Figure 11 illustrates the extent to which changes in each parameter affect the equilibrium point p*. The results show that the human propagation coefficient (λ_h), AI propagation coefficient (λ_m), and correcti on rate (μ) have the most significant effects on the system behaviour. The equilibrium point increases to 5 5.0% and 55.3% when h and m increase by 30%, respectively; conversely, the equilibrium point decreases t o 43.0% when μ increases by 30%. This suggests that controlling human and AI propagation behaviours wh ile improving the efficiency of information correction is the most effective means to curb the spread of misi nformation.

The hallucination amplification coefficient (σ_1) and retraining period (τ) show moderate sensitivit y.σ_1 A 30% increase causes the equilibrium point to rise to 53.4%, whereas a 30% increase in τ causes the equilibrium point to rise slightly to 49.5%. The rate of base illusion (σ_0) has a minimal effect, with a chang e of ±30% causing only a ±0.1% change in the equilibrium point, reflecting the fact that the base illusion of current AI models is not a major driver of misinformation propagation.

6. Model Strengths and Weaknesses

6.1. Strengths

A) **Complete Mechanism:** Simultaneously characterizes human-driven spread, AI reproduction, Al hall ucination, data poisoning amplification, and correction/clarification, with an endogenous time-delay chain and logically coupled mechanisms.

B) **Data-Aligned Parameters:** τ is statistically consistent with Task 1.1; σ_1 and μ are fitted based on o bservation; the platform AAR (e.g., AI Adoption Rate) and risk weights provide the calibration bas is for Tasks 2 – 4.

C) **High Interpretability:** The saturation term $1-p$ and the correction term μpmake the steady state a nd transient dynamics intuitive; the behavioral feedback parameters β_h and β_μ have policy implicat ions and can be mapped to the intensity of measures such as "media literacy," "fact-checking," and "data governance."

D) **Good Extensibility:** Can accommodate new observations (e.g., the latest AAR, platform policy cha nges, model training cadence) and be re-fitted; allows for the replacement of AI detectors or the inc orporation of topic stratification.

6.2. Weaknesses and Limitations

A) **Detector Bias:** AI detectors like OSM-Det and SurferSEO carry the risk of misclassification and dri ft, potentially leading to a systematic offset in the fitting of σ_1 and μ.

B) **Homogeneity Assumption:** The model represents the entire network with a single $p(t)$, neglecting t he heterogeneity of topics, languages, and community structures, as well as social network topologi cal effects.

C) **Constant or Simplified Parameters:** Except for β_h and β_μ, most parameters are treated as constants within the time period, failing to explicitly model sudden policy changes, model generational repla cement (e.g., architectural upgrades), or changes in economic incentives.

7. Reference

[1] Graphite. (2024, May 7). More articles are now created by AI than humans. Graphite.https://graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans

[2]Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard in the context of scientific writing: A systematic evaluation. Journal of Medical Internet Research, 26, e53164. https://doi.org/10.2196/53164

[3]Pangram Labs. (n.d.). Is AI writing the news? Retrieved October 30, 2025, from https://www.pangram.com/ai-news

[4]Sun, Z., Zhang, Z., Shen, X., Zhang, Z., Liu, Y., Backes, M., & Zhang, Y. (2024). Are we in the AI-generated text world already? Quantifying and monitoring AIGT on social media. arXiv preprint arXiv:2412.18148. https://arxiv.org/abs/2412.18148

[5]Little Holmes. (n.d.). CommonCrawl数据集探索(二)索引系统 [CommonCrawl dataset exploration (2) - index system]. Retrieved October 30, 2025, from https://little-holmes.com/blog/CommonCrawl%E6%95%B0%E6%8D%AE%E9%9B%86%E6%8E%A2%E7%B4%A2(%E4%BA%8C)%E7%B4%A2%E5%BC%95%E7%B3%BB%E7%BB%9F-zh/

[6]Common Crawl. (n.d.). Common Crawl dataset. Registry of Open Data on AWS. Retrieved October 30, 2025, from https://registry.opendata.aws/commoncrawl/

[7]Govindankutty, S., & co-authors. (2024). Epidemic modeling for misinformation spread in digital networks. Scientific Reports, 14, Article 69657. https://doi.org/10.1038/s41598-024-69657-0

[8]Vosoughi, S., Roy, D., & Aral, S. (2017). The spread of true and false news online. MIT Initiative on the Digital Economy Research Brief. https://ide.mit.edu/wp-content/uploads/2018/12/2017-IDE-Research-Brief-False-News.pdf

8. Appendix

Model	Provider	Knowledge Cutoff	Release Date	Time difference (Days)	Parameters	Context Window
GPT-3.5	OpenAI	2022-01-01	2022-03-15	73	175B	4,096
GPT-4	OpenAI	2023-04-01	2023-03-14	-18	~1.8T	8,192
Llama 2	Meta	2022-12-01	2023-07-18	229	70B	4,096
Gemini 1.0 Pro	Google	2023-04-01	2023-12-13	256	Unpublished	32,768
Gemini 1.5 Pro	Google	2023-11-01	2024-02-15	106	Unpublished	1,000,000
Claude 3 Opus	Anthropic	2023-08-01	2024-03-04	216	~175B	200,000
Llama 3	Meta	2023-12-01	2024-04-18	139	70B	8,192
GPT-4o	OpenAI	2023-10-01	2024-05-13	225	~1.8T	128,000
Claude 3.5 Sonnet	Anthropic	2024-04-01	2024-06-20	80	~175B	200,000
Mistral Large 2	Mistral AI	2023-10-01	2024-07-24	297	123B	32,768
GPT-5	OpenAI	2024-10-01	2025-08-07	310	Unknown	272,000
Llama 4 Scout	Meta	2024-08-01	2025-04-05	247	17B	10,000,000
Grok 3	xAI	2023-12-01	2025-02-01	428	Unknown	1,000,000
Gemini 2.5 Pro	Google	2025-01-01	2025-03-25	83	~1.56T	1,000,000
Claude 4 Opus	Anthropic	2025-01-01	2025-05-22	141	~200B	200,000
DeepSeek R1	DeepSeek	2023-10-01	2025-01-25	482	671B	128,000
o3	OpenAI	2024-06-01	2025-04-16	319	Unknown	200,000

Sensitivity Analysis of Model Parameters on Equilibrium Misinformation Proportion

Parameters	Baseline	Magnitude of change	Equilibrium	Rate of change
λ_h	0.2	+30%	55.0%	+14.1%
		-30%	40.1%	-16.8%
λ_m	0.2	+30%	55.3%	+14.7%
		-30%	40.5%	-16.0%
μ	0.236	+30%	43%	-10.8%
		-30%	55%	+14.1%
σ_1	0.0555	+30%	53.4%	+10.8%
		-30%	42.1%	-12.7%
σ_0	0.0001	+30%	48.3%	+0.2%
		-30%	48.1%	-0.2%
τ	197	+30%	49.5%	+2.7%
		-30%	46.8%	-2.9%