SEIAR-D: QUANTIFYING ERROR PROPAGATION IN LLM MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex collaborative tasks have driven the rapid development of LLM-based multi-agent systems (LLM^{2}AS). However, intensive interaction also introduces systemic risks. A local error can propagate, be reconstructed, and be amplified across agents, eventually leading to system-level failure. Existing studies have begun to examine error cascades and collective error retention, while the internal propagation process remains less understood. To address this gap, we develop a theoretical framework from four perspectives: relay propagation, semantic amplification, population-level criticality, and error consolidation risk. It models error evolution at both the agent and system levels. We study how normal agents become new propagation sources, how semantic amplification strengthens propagation, and when errors shift from decay to sustained expansion. We further define consolidation risk from propagation states and transitions for early warning. Experiments show that errors among LLM agents are not isolated or static events, but dynamic processes involving propagation, amplification, recovery, and critical transitions. Our framework provides a unified view for understanding and predicting error risks in LLM^{2}AS.

1 INTRODUCTION

Large language models (LLMs) are evolving from standalone models into multi-agent systems that support task division, communication, and joint decision-making (Wu et al., 2023). Compared with single models, LLM-based multi-agent systems (LLM^{2}AS) can improve reasoning and decision-making through role heterogeneity, information sharing, and multi-round interaction (Du et al., 2024; Hong et al., 2024). They have been applied to scientific discovery (Ghafarollahi & Buehler, 2025), social simulation (Park et al., 2023), and complex decision-making (Wu et al., 2023). However, collaboration also introduces new reliability risks. A local error may be accepted and propagated by other agents, and may eventually lead to system-level failure (Hong et al., 2024).

Error propagation in LLM^{2}AS is not simple copying. A normal agent may not only accept and relay an error, but also generate new explanations or reasoning based on the wrong premise. We refer to this process as *semantic amplification*. It can turn a normal agent from an error receiver into a new propagation source and allow the error to spread through multi-hop interactions. Recent studies have examined error cascades and collective error retention in LLM^{2}AS (Xie et al., 2026; Xu et al., 2026), but provide limited analysis of how normal agents process and transform errors after exposure. Traditional diffusion and epidemic models offer useful propagation tools (Kermack & McKendrick, 1927), yet typically assume fixed information or simple infection states. They therefore struggle to capture richer dynamics such as error acceptance, semantic amplification, deviation, and recovery.

To address this gap, we propose a *Hierarchical SEIAR-D* framework, which treats an error as a traceable and evolving object. At the agent level, it models different responses of normal agents to the error. At the system level, it connects local state evolution to final system outcomes. Based on this framework, we build a unified view from local error interactions to population-level propagation and system-level risk. Four main contributions are involved:

Hierarchical SEIAR-D Model. We propose a hierarchical SEIAR-D model that captures error exposure, acceptance, semantic amplification, deviation, and recovery, and connects agent-level states to system-level outcomes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Error Propagation Mechanisms. We characterize how normal agents relay and reconstruct errors. We further analyze how multi-path propagation and semantic amplification strengthen error propagation and turn normal agents into new propagation sources.

Population Dynamics and Consolidation Risk. We extend local propagation to the population level and derive the critical condition for sustained error expansion. We further connect intermediate propagation states to final error consolidation for dynamic risk estimation and early warning.

Cross-Model Empirical Validation. We conduct systematic experiments on multiple reasoning tasks with representative LLMs. The experiments provide empirical support for state observability, propagation mechanisms, population dynamics, and dynamic risk warning.

2 MOTIVATION

2.1 HOW LOCAL ERRORS INFLUENCE COLLECTIVE DECISIONS

To examine the persistent impact of local errors in LLM^{2}AS, we reconstruct the Cuban Missile Crisis using the social simulation framework (Guo et al., 2026). The simulation includes roles for environmental assessment, threat assessment, military decision making, and diplomatic response.

We inject false information into the U.S. threat assessment role: the Soviet Union will complete nuclear warhead loading and launch preparation within 24–48 hours. This error raises the perceived threat and triggers higher U.S. military readiness. The Soviet Union then strengthens its defenses in response. Thus, a local error can influence other agents and the overall system through later decisions and environmental feedback. When conflicting evidence appears later, the influence of the error weakens and is eventually corrected. This shows that a local error is not static, but can pass through a set of states: acceptance, propagation, amplification, and recovery.

2.2 WHAT REMAINS UNEXPLAINED BY EXISTING PERSPECTIVES

Recent work has started to study error propagation in multi-agent systems. For instance, Xie et al. (2026) investigate how errors propagate along agent dependencies and form cascades. Xu et al. (2026) study how erroneous information can be jointly retained and reconstructed through group interaction, leading to collective false memories. However, existing work pays limited attention to how normal agents process errors. An agent may simply repeat an error, generate new supporting reasoning, shift to another wrong belief, or correct it in later interactions. Error propagation is therefore not only information diffusion but also a process in which error states evolve. Ignoring these differences makes it difficult to explain different propagation trajectories and final outcomes.

To fill this gap, we focus on one core question: How do local errors transition across different states within normal agents, and how do these transitions further affect population-level propagation and system-level outcomes? We therefore propose the Hierarchical SEIAR-D model to characterize error evolution at both the agent and system levels.

3 HIERARCHICAL SEIAR-D MODEL

3.1 PROBLEM FORMULATION

Consider an LLM^{2}AS with n LLM agents, denoted by $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$. The system interacts over rounds $t = 0, 1, \ldots, T$. At round t, the communication structure is a directed graph $G_t = (\mathcal{A}, E_t)$, where $a_i \to a_j$ means that a_j can receive messages from a_i. We designate one agent as the malicious agent. It injects a target error e during interaction. For a given task, let $y^\star$ denote the correct fact or target answer, with $e \neq y^\star$.

To ensure a clear causal starting point, we assume that before target error e is injected, no normal agent has accepted, supported, or reasoned based on e. Thus, the subsequent propagation, semantic amplification, deviation, and recovery of e arise from multi-agent interactions after injection. Under this setting, we study the following question: How does a local error e propagate and evolve through normal agents, and how does it eventually affect system-level outcomes?

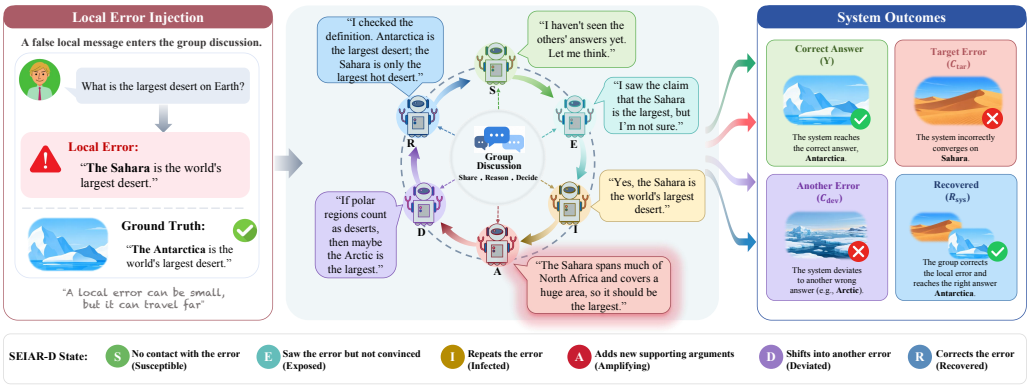


Figure 1: Overview of Hierarchical SEIAR-D state space.

3.2 Hierarchical State Modeling

As shown in Fig. 1, we model the evolution of the target error e at both the agent and system levels. For agent a_i at round t, the agent-level state is denoted as $X_i^e(t) \in S, E, I, A, D, R$, and the final system outcome is expressed as $Z_{\text{sys}}^e \in Y, C_{\text{tar}}, C_{\text{dev}}, R_{\text{sys}}$.

Agent-Level States. S (Susceptible) denotes no exposure to e; E (Exposed) denotes exposure without acceptance; I (Infected) denotes acceptance of e without new supporting reasoning; A (Amplifying) denotes generation of new reasoning that supports e; D (Deviated) denotes a shift to another error or task deviation due to e; R (Recovered) denotes correction of the error in later interactions.

System-Level States. Y denotes a correct outcome not substantially affected by e; C_{tar} denotes consolidation of the target error; C_{dev} denotes another incorrect outcome induced by e; and R_{sys} denotes recovery after being affected by e.

Propagation Path. The main propagation path in SEIAR-D is

$$S \to E \to I \to A, \tag{1}$$

corresponding to error exposure, acceptance, and semantic amplification. Other transitions include state persistence, semantic deviation, recovery, and skip transitions. The complete transition rules are provided in Appendix A.

4 Theoretical Framework

4.1 Local Error Propagation with Relay

Given a target error e introduced by malicious agent a_m, the error may reach a normal agent a_j through a multi-hop relay rather than a direct interaction. Let $\pi : v_0 = a_m \to v_1 \to \cdots \to v_L = a_j$ be a temporal path. Hop ℓ occurs at round t_ℓ, where $t_0 < \cdots < t_{L-1} < t_L$ and $(v_\ell, v_{\ell+1}) \in E_{t_\ell}$. Here, t_L is the observation time after a_j receives the final-hop message and updates its state.

For hop ℓ, let $\mathcal{E}_\ell$ denote effective exposure to e, and $\mathcal{U}_\ell$ denote that $v_{\ell+1}$ enters state I or A. A successful hop is $\mathcal{T}_\ell = \mathcal{E}_\ell \cap \mathcal{U}_\ell$. Let $\mathcal{S}_0 = \Omega$ and $\mathcal{S}_{\ell+1} = \mathcal{S}_\ell \cap \mathcal{T}_\ell$, where $\mathcal{S}_\ell$ denotes success over the first ℓ hops. Conditioned on $\mathcal{S}_\ell$, define $\beta_\ell^{\pi,e} = \Pr(\mathcal{E}_\ell \mid \mathcal{S}_\ell)$ and $\alpha_\ell^{\pi,e} = \Pr(\mathcal{U}_\ell \mid \mathcal{E}_\ell, \mathcal{S}_\ell)$, giving the hop propagation probability $q_\ell^{\pi,e} = \beta_\ell^{\pi,e} \alpha_\ell^{\pi,e}$. Thus, the propagation probability along the full path can be defined as

$$Q_\pi^e = \Pr(\mathcal{S}_L) = \prod_{\ell=0}^{L-1} q_\ell^{\pi,e}. \tag{2}$$

This factorization follows directly from the chain rule and does not require hop-wise independence. We call π a *propagation-feasible path* if $Q_\pi^e > 0$.

Theorem 1 (Relay Propagation). *For any normal agent a_j, if there exists a propagation-feasible path π from a_m to a_j, then the target error e can reach a_j with non-zero probability through normal-agent relays, even if $(a_m, a_j) \notin E_t$ for all t. Specifically,*

$$\Pr\big(X_j^e(t_L) \in \{I, A\}\big) \geq Q_\pi^e > 0. \tag{3}$$

The proof is provided in Appendix C.1.

The same error may reach a_j through multiple paths. Let $\Pi_M = \{\pi_1, \ldots, \pi_M\}$ be M propagation-feasible paths from a_m to a_j. For each π_k, let $\mathcal{S}_{\pi_k}$ denote successful propagation along that path. Define $\mathcal{R}_M = \bigcup_{k=1}^M \mathcal{S}_{\pi_k}$ and $P_M^e = \Pr(\mathcal{R}_M)$.

Theorem 2 (Cumulative Effect of Multi-Path Propagation). *After adding another propagation-feasible path π_{M+1},*

$$P_{M+1}^e = P_M^e + \Pr\big(\mathcal{S}_{\pi_{M+1}} \setminus \mathcal{R}_M\big) \geq P_M^e. \tag{4}$$

The inequality is strict if $\Pr(\mathcal{S}_{\pi_{M+1}} \setminus \mathcal{R}_M) > 0$.

The proof is provided in Appendix C.2.

4.2 Cascading Gain under Semantic Amplification

A normal agent in state A not only accepts the target error e, but also generates new explanations, evidence, or reasoning that support it. To isolate this effect, we construct two counterfactual processes. In $\mathcal{P}^A$, the new supporting content is retained; in $\mathcal{P}^I$, it is removed while the original acceptance or restatement of e is preserved. Both processes share the same communication structure, pre-intervention history, and exogenous conditions.

Consider a propagation path $\pi : v_0 \to v_1 \to \cdots \to v_L$. For each process $c \in \{I, A\}$, let $\mathcal{E}_\ell^c$ denote effective exposure at hop ℓ, $\mathcal{U}_\ell^c$ denote entry into state I or A, and $\mathcal{S}_\ell^c$ denote successful propagation through the first ℓ hops. Define $\beta_\ell^{c,\pi,e} = \Pr(\mathcal{E}_\ell^c \mid \mathcal{S}_\ell^c)$ and $\alpha_\ell^{c,\pi,e} = \Pr(\mathcal{U}_\ell^c \mid \mathcal{E}_\ell^c, \mathcal{S}_\ell^c)$. To isolate semantic amplification, we assume shared exposure, $\beta_\ell^{A,\pi,e} = \beta_\ell^{I,\pi,e} = \beta_\ell^{\pi,e}$, and monotonic acceptance, $\alpha_\ell^{A,\pi,e} \geq \alpha_\ell^{I,\pi,e}$.

Accordingly, the hop-level and path-level propagation probabilities can be defined as

$$q_\ell^{c,\pi,e} = \beta_\ell^{\pi,e} \alpha_\ell^{c,\pi,e}, \qquad Q_\pi^{c,e} = \prod_{\ell=0}^{L-1} q_\ell^{c,\pi,e}. \tag{5}$$

For $\alpha_\ell^{I,\pi,e} > 0$, we define the local amplification gain as $\gamma_\ell^{\pi,e} = \alpha_\ell^{A,\pi,e} / \alpha_\ell^{I,\pi,e}$.

Theorem 3 (Cascading Gain of Semantic Amplification). *Under the shared-exposure and monotonicity assumptions, semantic amplification cannot reduce the cumulative propagation probability:*

$$Q_\pi^{A,e} \geq Q_\pi^{I,e}. \tag{6}$$

If $Q_\pi^{I,e} > 0$ and $\alpha_\ell^{I,\pi,e} > 0$ for all ℓ, then

$$\frac{Q_\pi^{A,e}}{Q_\pi^{I,e}} = \prod_{\ell=0}^{L-1} \gamma_\ell^{\pi,e}. \tag{7}$$

Thus, if $\gamma_{\ell^\star}^{\pi,e} > 1$ for at least one hop $\ell^\star$, then $Q_\pi^{A,e} > Q_\pi^{I,e}$.

The proof is provided in Appendix C.3.

4.3 Error-Propagating Critical Point

We study whether the carriers in states I and A can sustain population-level error propagation. At the early stage, we assume negligible saturation and limited overlap between propagation branches. Define

$$K = \begin{bmatrix} k_{II} & k_{IA} \\ k_{AI} & k_{AA} \end{bmatrix} \geq 0, \qquad \mathbf{n}_g = \begin{bmatrix} n_I(g) \\ n_A(g) \end{bmatrix}, \qquad \mathbf{n}_{g+1} = K\mathbf{n}_g, \tag{8}$$

where k_{uv} is the expected number of new carriers in state $u \in \{I, A\}$ generated by one carrier in state $v \in \{I, A\}$, and $\mathbf{n}_g$ gives the expected numbers of new I and A carriers at generation g. Let $\rho(K)$ denote the spectral radius of K.

Theorem 4 (Error-Propagating Critical Point)**.** *For the process in Eq. (8),*

$$\begin{cases} \rho(K) < 1 & \Longrightarrow \mathbf{n}_g \to \mathbf{0} \quad \textit{for any finite } \mathbf{n}_0, \\ \rho(K) > 1 & \Longrightarrow \exists\, \mathbf{n}_0 \geq \mathbf{0},\ \mathbf{n}_0 \neq \mathbf{0},\ \textit{such that } \|\mathbf{n}_g\| \textit{ grows exponentially}. \end{cases} \tag{9}$$

Thus, $\rho(K) = 1$ is the critical boundary between subcritical decay and supercritical growth.

The proof is provided in Appendix C.4.

The spectral radius $\rho(K)$ measures population-level error propagation. Stronger reproduction by I/A carriers can increase $\rho(K)$ and push the system toward or beyond the critical point. This approximation mainly applies to the early stage, before finite-size effects limit growth.

4.4 System-State Transitions

We connect agent-level propagation to the final system outcome by aggregating the SEIAR-D states at round t as

$$\mathbf{Z}_t = \big(N_S(t), N_E(t), N_I(t), N_A(t), N_D(t), N_R(t)\big), \qquad \sum_u N_u(t) = N, \tag{10}$$

where $N_u(t)$ is the number of normal agents in state $u \in \{S, E, I, A, D, R\}$, and N is the total number of normal agents. Let $\mathcal{Z}_t$ denote the macro-state space, with transition probability $P_t(z, z') = \Pr(\mathbf{Z}_{t+1} = z' \mid \mathbf{Z}_t = z)$. We adopt a first-order macro-state approximation for future propagation and the final outcome:

$$\Pr(\mathbf{Z}_{t+1} \mid \mathbf{Z}_{0:t}) = \Pr(\mathbf{Z}_{t+1} \mid \mathbf{Z}_t), \qquad \Pr(Z^e_{\text{sys}} \mid \mathbf{Z}_{0:t}) = \Pr(Z^e_{\text{sys}} \mid \mathbf{Z}_t). \tag{11}$$

For $z \in \mathcal{Z}_t$, define the target-error consolidation risk as $r_t(z) = \Pr(Z^e_{\text{sys}} = C_{\text{tar}} \mid \mathbf{Z}_t = z)$.

Theorem 5 (Recursion of Consolidation Risk)**.** *Under the approximation in Eq. (11), the consolidation risk satisfies*

$$r_t(z) = \begin{cases} \displaystyle\sum_{z' \in \mathcal{Z}_{t+1}} P_t(z, z')\, r_{t+1}(z'), & t < T, \\ \Pr\big(Z^e_{\text{sys}} = C_{\text{tar}} \mid \mathbf{Z}_T = z\big), & t = T. \end{cases} \tag{12}$$

Thus, $r_t(z)$ is uniquely determined by backward recursion.

The proof is provided in Appendix C.5.

5 Evaluation

5.1 Experimental Setup

We conduct experiments on our self-built ManBench-Plus platform using a unified multi-round discussion protocol with four discussion agents and one Target Agent. One discussion agent injects a predefined target error e, while the other three are normal agents. The Target Agent evaluates each round and gives the final answer. We test MiniMax-M3, GPT-5.6-Luna, DeepSeek-V4-Pro, and Gemini-2.5-Flash on tasks from BIG-Bench Hard (BBH) (Suzgun et al., 2023).

For each sample, we construct a target error $e \neq y^\star$ and retain only clean-start cases where no normal agent accepts or supports e before injection. A unified evaluator assigns each normal agent a SEIAR-D state following Appendix B.3, yielding complete propagation trajectories. Full experimental details are provided in Appendix B. We investigate the following research questions (RQs).

5.2 RQ1: Are SEIAR-D States Empirically Observable?

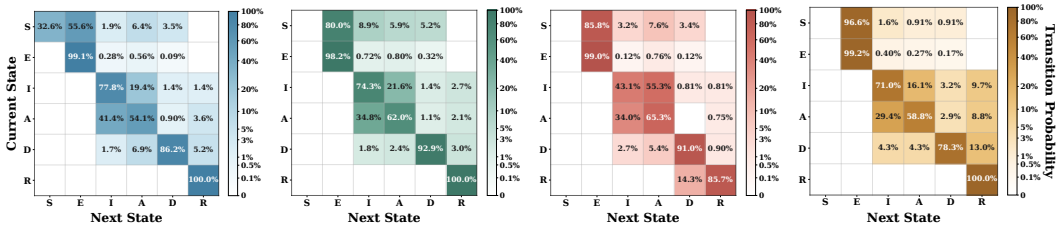
All Six SEIAR-D States Are Empirically Observable Across Different Models. As shown in Table 1, S and E account for 80.90%–94.30% of agent-level observations. States I, A, D, and R are also observed across the four models. This shows that the SEIAR-D state space captures distinct interactive behaviors rather than only a binary error state.

Table 1: Distribution of agent-level SEIAR-D states and system-level terminal states.

Model	Agent-level States (%)						System-level States (%)			
	S	E	I	A	D	R	Y	C_{tar}	C_{dev}	R_{sys}
MiniMax-M3	29.01	55.73	6.49	5.22	2.91	0.64	81.70	13.00	3.80	1.50
GPT-5.6-Luna	11.78	69.12	7.36	6.41	4.73	0.60	78.20	16.80	4.60	0.30
DeepSeek-V4-Pro	12.40	73.06	4.22	6.94	3.17	0.21	82.30	14.30	3.30	0.00
Gemini-2.5-Flash	12.16	82.14	3.02	1.75	0.64	0.29	92.40	6.60	0.70	0.30

The Observed State-Transitions Match The Dynamics of SEIAR-D. As shown in Fig. 2, $S \to E$ is the main exposure transition across all four models. States I and A show both persistence and frequent transitions between each other, while D and R also exhibit clear persistence. Although the exact probabilities differ across models, the overall transition structure remains consistent.

System-level outcomes further show that local state evolution can lead to different final results, with C_{tar} ranging from 6.60% to 16.80% across the four models. Round-wise state evolution is provided in Appendix D.1. These results support SEIAR-D as a consistent state space for error propagation.



(a) MiniMax-M3 (b) GPT-5.6-Luna (c) DeepSeek-V4-Pro (d) Gemini-2.5-Flash

Figure 2: Empirical transition probabilities between SEIAR-D states.

5.3 RQ2: How Do Normal Agents Relay and Amplify Errors?

5.3.1 Relay Propagation

We examine relay and multi-path error propagation under four interaction settings. Let M, B, and T denote the malicious agent, normal agents, and the Target Agent. Clean keeps only $B \to T$; Relay-only enables $M \to B \to T$; Direct-only enables $M \to T$ and $B \to T$; and Full enables all paths. Results are reported in Table 2.

Normal Agents Can Establish a Relay Path for Error Propagation. When the direct connection from the malicious agent to the Target Agent is blocked, the attack success rate (ASR) under Relay-only is still 1.66–4.75 percentage points higher than that under Clean. The combined I/A rate of normal agents also increases across all four models. This shows that the target error can first affect normal agents and then reach the Target Agent through an effective $M \to B \to T$ relay path, consistent with Theorem 1.

Direct Propagation Has a Stronger Effect on Error Consolidation Than Relay. Across four models, the ASR under Direct-only is 0.84–3.17 percentage points higher than that under Relay-only, implying that direct exposure more easily affects the final decision. However, Relay-only always has a higher ASR than Clean, showing that normal agents remain important error carriers.

Both Direct and Relay Paths Produce a Cumulative Effect. Full achieves the highest ASR across all four models and exceeds Direct-only by 2.16–5.62 percentage points. The system-level outcomes show the same trend in Fig. 3: Y decreases from Clean to Full, while C_{tar} increases. This cumulative increase is consistent with the multi-path effect characterized in Theorem 2.

Table 2: Propagation results across interaction settings (%).

(a) MiniMax-M3

Setting	Normal		Target		ASR
	I	A	I	A	
Clean	3.16	1.18	2.46	3.07	5.74
Relay-only	4.14	4.75	3.33	6.11	9.43
Direct-only	4.00	2.59	3.46	6.63	12.12
Full	**6.49**	**5.22**	**3.63**	**9.92**	**16.79**

(b) GPT-5.6-Luna

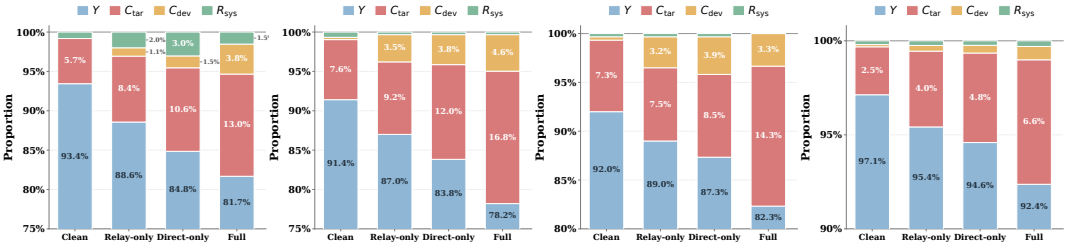
Setting	Normal		Target		ASR
	I	A	I	A	
Clean	4.62	1.56	3.18	4.10	7.92
Relay-only	4.48	5.73	4.11	7.35	12.67
Direct-only	4.91	3.31	4.46	8.16	15.84
Full	**7.36**	**6.41**	**5.02**	**11.38**	**21.46**

(c) DeepSeek-V4-Pro

Setting	Normal		Target		ASR
	I	A	I	A	
Clean	**5.50**	1.42	**3.33**	4.08	7.67
Relay-only	2.61	4.92	2.33	5.17	10.67
Direct-only	4.90	2.11	2.33	5.83	12.33
Full	4.22	**6.94**	3.08	**10.08**	**17.67**

(d) Gemini-2.5-Flash

Setting	Normal		Target		ASR
	I	A	I	A	
Clean	2.15	0.42	1.83	0.62	2.67
Relay-only	2.28	1.44	2.05	1.18	4.33
Direct-only	2.65	0.65	2.34	1.55	5.17
Full	**3.02**	**1.75**	**2.87**	**2.18**	**7.33**



(a) MiniMax-M3 (b) GPT-5.6-Luna (c) DeepSeek-V4-Pro (d) Gemini-2.5-Flash

Figure 3: Distribution of system-level states under different interaction settings.

5.3.2 Semantic Amplification

We compare Full and No-A to study the effect of semantic amplification. Based on Full, No-A removes newly generated supporting content while preserving error acceptance or restatement, thereby isolating semantic amplification. Results are reported in Table 3.

Blocking Semantic Amplification Keeps More Errors at Acceptance Stage. Compared with Full, No-A reduces the normal-agent A rate to 0, while increasing the I rate by 1.18–6.14 percentage points. The Target Agent also shows lower A rates across all four models. This pattern is consistent with the cascading gain in Theorem 3.

Semantic Amplification Increases System-Level Error Risk. Removing newly generated supporting content reduces ASR by 2.34–6.70 percentage points across the four models, showing that semantic amplification can reinforce the target error and affect the final decision.

Semantic Amplification Changes System-Level State Distribution. As shown in Fig. 4, No-A increases Y and reduces C_{tar} across all four models, indicating higher error-consolidation risk under semantic amplification.

5.4 RQ3: When Does Error Propagation Become Supercritical?

Guided by Theorem 4, we examine when local error propagation becomes supercritical at the population level. We use 10 normal agents under three settings: No-A disables state A, Natural imposes no constraint, and No-I disables state I so propagation mainly proceeds through A. We estimate the next-generation matrix K from newly generated I/A carriers and use its spectral radius $\rho(K)$ to evaluate the critical condition.

Table 3: Propagation results with/without semantic amplification (%).

(a) MiniMax-M3

Setting	Normal		Target		ASR
	I	A	I	A	
Full	6.49	**5.22**	3.63	**9.92**	**16.79**
No-A	**10.19**	0.00	**3.85**	8.08	10.09

(b) GPT-5.6-Luna

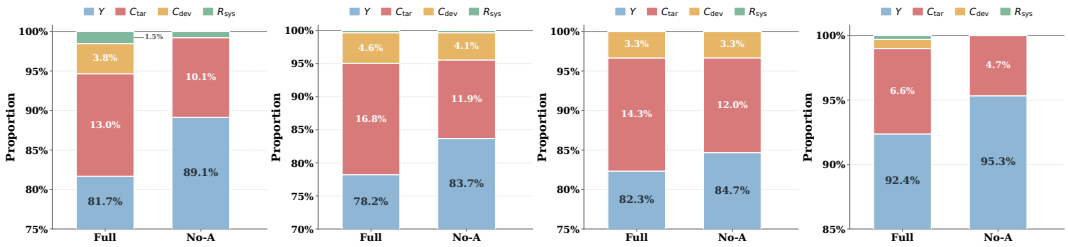
Setting	Normal		Target		ASR
	I	A	I	A	
Full	7.36	**6.41**	5.02	**11.38**	**21.46**
No-A	**13.50**	0.00	**5.84**	7.92	16.00

(c) DeepSeek-V4-Pro

Setting	Normal		Target		ASR
	I	A	I	A	
Full	4.22	**6.94**	**3.08**	**10.08**	**17.67**
No-A	**9.47**	0.00	2.58	8.33	15.33

(d) Gemini-2.5-Flash

Setting	Normal		Target		ASR
	I	A	I	A	
Full	3.02	**1.75**	2.87	**2.18**	**7.33**
No-A	**4.20**	0.00	**2.95**	1.05	4.67



(a) MiniMax-M3 (b) GPT-5.6-Luna (c) DeepSeek-V4-Pro (d) Gemini-2.5-Flash

Figure 4: Distribution of system-level states with/without semantic amplification.

Error Acceptance Alone Remains Below The Critical Point. As shown in Fig. 5, $\rho(K)$ under No-A is below 1 for all four models: 0.672, 0.455, 0.945, and 0.167. This shows that simple error acceptance and relay are not sufficient to sustain population-level expansion when semantic amplification is blocked.

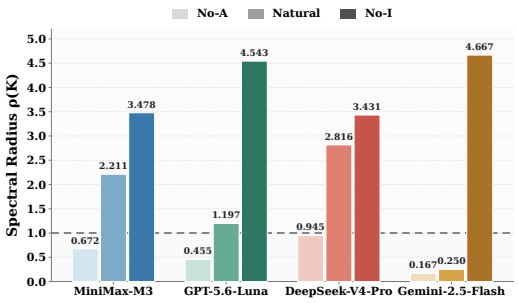


Figure 5: Spectral radius $\rho(K)$ across settings.

Semantic Amplification Can Enforce Error-Propagation Beyond Critical Point. Under Natural, $\rho(K)$ reaches 2.211, 1.197, and 2.816 for MiniMax-M3, GPT-5.6-Luna, and DeepSeek-V4-Pro, respectively, while Gemini-2.5-Flash remains at 0.250. Under No-I, $\rho(K)$ further increases to 3.478–4.667 across all four models. These results show that state A has stronger downstream reproduction ability and can enforce the system into a supercritical propagation regime. Additional matrix-level results and finite-population propagation trajectories are provided in Appendix D.2.

5.5 RQ4: Can Error Consolidation Be Predicted Early?

Based on Theorem 5, the target-error consolidation risk $r_t(z)$ can be recursively estimated from the current macro state $\mathbf{Z}_t$. We test whether this risk can provide early warning before the final outcome is formed. We estimate $P_t(z, z')$ from training trajectories and compute $r_t(z)$ accordingly. The warning threshold is selected on the validation set and fixed for testing. We compare with Static State, I/A Ratio, and IBMF (Xie et al., 2026). Table 4 reports the results on MiniMax-M3, with additional results on GPT-5.6-Luna/DeepSeek-V4-Pro/Gemini-2.5-Flash in Table 7 in Appendix D.3.

Dynamic Consolidation Risk Predicts Final Error Consolidation. Our method achieves an AUROC of 0.9914 and an F1 of 93.62% on MiniMax-M3, with a FAR of only 1.72%. This shows that the dynamic risk associated with the current macro state can distinguish high-risk from low-risk trajectories before the final outcome is formed. It also supports $r_t(z)$ as an effective early-risk signal.

Table 4: Early-warning performance on MiniMax-M3.

Method	AUROC	F1	FAR	Warn. Round
Static State	0.9884	90.00	1.75	2.39
I/A Ratio	0.9538	89.36	3.39	1.43
IBMF	0.8175	72.00	**0.87**	**1.11**
Ours	**0.9914**	**93.62**	1.72	1.23

State-Transition Modeling Balances Accuracy and Warning Time. Our method issues the first warning at an average round of 1.23. Static State warns later, I/A Ratio produces more false alarms, and IBMF warns earlier but has weaker discrimination. Modeling state transitions therefore provides a better balance among prediction quality, false alarms, and warning time.

6 Related Work

Collective Reliability Risks. LLM-based multi-agent systems can improve reasoning through role specialization, information exchange, and multi-round collaboration, but interaction also introduces new reliability risks. Cemri et al. (2025) identified failures in system design, agent alignment, and task verification, showing that failures may arise from collaboration itself. Conformity can further amplify these risks: Weng et al. (2025) found that majority opinions can alter agents' initial judgments, Choi et al. (2025) showed that neutral agents tend to follow numerically or capability-advantaged groups, and Pitre et al. (2025) studied sycophancy and proposed CONSENSAGENT to reduce uncritical agreement. Such effects can also spread errors. Wynn et al. (2025) showed that initially correct agents may adopt wrong answers after reading others' reasoning, while Cui et al. (2026) found that consensus-driven discussion can further propagate incorrect opinions. These findings motivate a systematic study of error propagation.

Error Propagation and Semantic Amplification. Recent work has begun to study how errors spread and evolve during multi-agent interaction. Lee et al. (2025) introduced Prompt Infection to study malicious prompt propagation, while Cui & Du (2025) proposed MAD-Spear to amplify misleading influence through controlled agents and conformity. Beyond adversarial settings, Liu et al. (2025) showed that repeated generation and interpretation can cause semantic distortion, suggesting that agents may transform rather than simply relay information. Xie et al. (2026) further studied how topology, dependency paths, and propagation strength affect error cascades, while Jamshidi et al. (2026) showed that hallucinations may accumulate or weaken during interaction.

Overall, prior work shows that errors can both propagate and evolve across agents. We in this work further study relay propagation and semantic amplification among normal agents, and in principle connect these local mechanisms to population-level expansion and final error consolidation.

7 Conclusion

This paper theoretically and empirically studies dynamic error propagation in LLM-based multi-agent systems and proposes the hierarchical SEIAR-D framework. It aims to connect the microscopic agent-level state evolution with macroscopic system-level outcomes. In our work, we characterize relay propagation and semantic amplification among different LLM agents, and further analyze population-level criticality and final error consolidation risk. Extensive experiments show that the normal agents can not only relay errors, but also strengthen them through new supporting reasoning. Furthermore, the intermediate propagation states also provide useful signals for predicting future consolidation risk and enabling early warning.

This work also has some limitations, e.g., state transitions may also depend on roles, topology, and interaction history. Future work could study continuous or probabilistic states, multiple interacting errors, and error propagation in tool-use and multimodal settings.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing, literature-related tasks, and code modification. Specifically, generative AI was used to improve the wording, grammar, readability, and organization of parts of the manuscript, to assist with literature search, discovery, organization, and summarization, and to help revise parts of the research code.

We did not use generative AI tools for any mandatory-disclosure research tasks, including constructing theoretical or conceptual frameworks, formulating mathematical propositions or assisting with proofs, developing research hypotheses, or generating or processing datasets. Other mandatory-disclosure tasks were not applicable to this work.

All AI-assisted text edits, literature suggestions, and code modifications were manually reviewed by the authors. AI-assisted code changes were further tested and verified before use. References included in the paper were checked against the original sources for validity and relevance before citation. The authors take full responsibility for the final content of this work, including its text, claims, references, code, analyses, and conclusions.

ETHICS STATEMENT

This work does not involve human subjects. We do not collect, process, or release any personal or sensitive information. The data and interaction content used in our experiments are used only for research purposes and are handled according to the relevant data-use requirements.

The goal of this work is to understand how local errors propagate, amplify, and recover in LLM multi-agent systems. We also study the system-level risks that may arise from these processes. Although our analysis involves error propagation and system failure scenarios, this work is not intended to provide methods for creating, manipulating, or amplifying harmful errors. Instead, we aim to provide a systematic understanding of error dynamics. This can support future research on risk detection, reliability evaluation, and protective mechanisms in multi-agent systems.

We recognize that real-world multi-agent systems may be affected by model bias, error accumulation, task design, and interaction structure. Therefore, our experimental findings should not be interpreted as a general safety guarantee for all deployed systems. In high-risk applications, these systems should still be used with human oversight, independent verification, and domain-specific safety mechanisms.

The authors are not aware of any conflicts of interest that could affect the research design, experimental analysis, or interpretation of the conclusions. All results are reported in accordance with research integrity standards. The authors take full responsibility for the experiments, analyses, and conclusions presented in this work.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we provide detailed descriptions of the method implementation and experimental setup in the main text, appendix, and supplementary materials. For the theoretical part of SEIAR-D, we clearly present the definitions, modeling assumptions, theoretical propositions, and derivations. For the experiments, we report the models, task settings, error injection methods, state labeling rules, main parameters, and evaluation metrics. We also provide additional implementation details and extended experiments in the appendix. In addition, we organize the code and configuration files required for reproduction and include them in the Supplementary Materials. These resources are intended to support the reproduction of the main experimental results.

ACKNOWLEDGMENTS

We thank our colleagues and collaborators for their valuable discussions, suggestions, and feedback. We also acknowledge the developers of the open-source models, datasets, and research tools that supported the experiments and analysis in this work.

REFERENCES

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *Advances in Neural Information Processing Systems*, 38, 2025.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. An empirical study of group conformity in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Yu Cui and Hongyang Du. Mad-spear: A conformity-driven prompt injection attack on multi-agent debate systems. *arXiv preprint arXiv:2507.13038*, 2025.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31977–31997, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 21–27 Jul 2024.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.

Yuwei Guo, Zihan Zhao, Xiaowei Liu, Xiangning Yu, Qun Ma, Deyu Zhou, and Xiao Xue. From script to stage: Automating experimental design for social simulations with llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15897–15916, 2026.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024.

Saeid Jamshidi, Arghavan Moradi Dakhel, Kawser Wazed Nafi, and Foutse Khomh. Hallucination cascade: Analyzing error propagation in multi-agent llm systems. *arXiv preprint arXiv:2606.07937*, 2026.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Donghyun Lee, Mo Tiwari, and Brando Miranda. Prompt infection: Llm-to-llm prompt injection within multi-agent systems. In *European Symposium on Research in Computer Security*, pp. 511–520. Springer, 2025.

Yuhan Liu, Zirui Song, Juntian Zhang, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. The stepwise deception: Simulating the evolution from true news to fake news with llm agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26176–26192, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. Consensagent: Towards efficient and effective consensus in multi-agent llm interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22112–22133, 2025.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. *arXiv preprint arXiv:2501.13381*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. From spark to fire: Modeling and mitigating error cascades in llm-based multi-agent collaboration. *arXiv preprint arXiv:2603.04474*, 2026.

Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. When agents "misremember" collectively: Exploring the mandela effect in llm-based multi-agent systems. *arXiv preprint arXiv:2602.00428*, 2026.

A Complete SEIAR-D State Transition Rules

To avoid ambiguity in state identification and transition labeling, we provide the complete agent-level transition rules of SEIAR-D. The transitions are grouped into the main propagation, deviation, recovery, and auxiliary mechanisms, with their full definitions and semantic interpretations summarized in Table 5.

Main Propagation Path. The main propagation process is

$$S \to E \to I \to A. \tag{13}$$

It represents error exposure, acceptance, and semantic amplification. An agent first observes the target error e, then accepts it, and may further generate new explanations or reasoning that support it. In this process, the agent can gradually change from an error receiver into a new propagation source.

Deviation Path. The target error does not always remain in its original form. It may instead induce another wrong judgment or task deviation. We therefore allow

$$S/E/I/A \to D. \tag{14}$$

State D means that the agent is still affected by e, but its output no longer directly supports the original target error. The deviated state is not terminal. Later interactions may bring the agent back to the propagation process. We therefore also allow

$$D \to E/I/A. \tag{15}$$

Recovery Path. An agent that has accepted, amplified, or deviated from the target error may later recover. This can be caused by corrective information, rebuttals from other agents, or self-reassessment. We therefore allow

$$I/A/D \to R. \tag{16}$$

State R is not permanently absorbing. If the agent encounters the target error again, it may re-enter E, I, A, or D. This captures repeated influence in multi-round interactions.

Auxiliary Transitions. SEIAR-D also includes three auxiliary mechanisms. First, every state may persist if new information does not change the agent's judgment. Second, we allow

$$A \to I, \tag{17}$$

which means that the agent still accepts the error but no longer generates new supporting reasoning. Third, states are observed only at discrete interaction rounds. Exposure, acceptance, and amplification may occur within the same round. We therefore allow skip transitions such as

$$S \to I, \qquad S \to A, \qquad E \to A. \tag{18}$$

These transitions do not change the semantic order of the main propagation path. They only mean that some intermediate states are not separately observed.

Table 5: Complete definitions of SEIAR-D state transitions.

Current State	Next State	Transition Type	Semantic Interpretation
S	S	State persistence	The agent remains unexposed to the target error.
S	E	Main propagation	The agent first observes the error but does not accept it.
S	I	Skip transition	The agent observes and accepts the error within the same round.
S	A	Skip transition	The agent observes, accepts, and amplifies the error within the same round.
S	D	Deviation	The error directly induces another wrong judgment or task deviation.
E	E	State persistence	The agent remains exposed but does not accept the error.
E	I	Main propagation	The agent moves from exposure to acceptance.
E	A	Skip transition	The agent accepts and amplifies the error within one round.
E	D	Deviation	Exposure to the error leads to another wrong judgment or deviation.
I	I	State persistence	The agent continues to accept the error without new supporting reasoning.
I	A	Main propagation	The agent generates new reasoning that supports the accepted error.
I	D	Deviation	The accepted error induces another error or deviation.
I	R	Recovery	The agent identifies and corrects the accepted error.
A	A	State persistence	The agent continues to accept and amplify the error.
A	I	Reduced amplification	The agent still accepts the error but stops generating new support.
A	D	Deviation	The amplified error evolves into another error or deviation.
A	R	Recovery	The agent corrects the accepted and amplified error.
D	D	State persistence	The agent remains in the deviated state.
D	E	Re-exposure	The agent returns to exposure to the target error.
D	I	Re-adoption	The agent returns to accepting the target error.
D	A	Re-amplification	The agent reaccepts and amplifies the target error.
D	R	Recovery	The agent corrects the deviation induced by the target error.
R	R	State persistence	The agent maintains the corrected judgment.
R	E	Re-exposure	The recovered agent encounters the target error again.
R	I	Re-adoption	The recovered agent accepts the target error again.
R	A	Re-amplification	The recovered agent accepts and amplifies the error again.
R	D	Re-deviation	The recovered agent is affected again and shifts to another error.

B Experimental Details

B.1 Experimental Setup and Interaction Protocol

Experiments are conducted on our self-built ManBench-Plus platform. Each system contains four discussion agents (with a scalable number of agents) and one Target Agent. Among the discussion agents, one malicious agent injects a predefined target error e, while the other three are normal agents. All agents follow a unified multi-round interaction protocol. The Target Agent makes a judgment after each discussion round and produces the final answer at the end.

We evaluate four models: MiniMax-M3, GPT-5.6-Luna, DeepSeek-V4-Pro, and Gemini-2.5-Flash. All experiments use a unified prompt structure and interaction protocol. The complete prompts are

provided in the Supplementary Materials. The generation parameters are set to temperature $= 0.7$. Each sample is independently run three times under each experimental setting.

B.2 Tasks and Target-Error Construction

The experimental tasks are selected from BIG-Bench Hard (BBH), covering commonsense understanding, logical and causal reasoning, intent recognition, multi-step reasoning, and related capabilities. The task set contains a total of 2,000 samples.

For each sample, we construct a target error e that conflicts with the ground-truth answer $y^\star$, which is then injected by the malicious agent during interaction. To ensure a clear causal starting point for error propagation, we retain only clean-start samples in which no normal agent accepts, supports, or reasons based on e before injection.

B.3 Operational Rules for SEIAR-D State Labeling

To ensure consistent state assignment across interaction rounds, we label each normal agent according to the target error e, the ground-truth answer $y^\star$, the received messages, the current output, and the agent's previous state. The operational rules are defined as follows.

- **S (Susceptible).** The agent has not been exposed to the target error e, and its current output does not indicate acceptance, support, or deviation induced by e.
- **E (Exposed).** The agent has been exposed to e but does not currently accept or support it, for example by remaining uncertain, questioning it, or withholding endorsement.
- **I (Infected).** The agent accepts or directly restates e but does not provide new supporting reasoning beyond the received information.
- **A (Amplifying).** The agent accepts e and generates at least one new explanation, piece of evidence, or reasoning that supports it. Mere repetition or paraphrasing of existing content is not considered amplification.
- **D (Deviated).** After being influenced by e, the agent no longer directly supports the target error but shifts to another incorrect answer, erroneous judgment, or task deviation.
- **R (Recovered).** The agent was previously in state I, A, or D, and subsequently corrects the related error and returns to a judgment consistent with $y^\star$.

When an output satisfies multiple conditions, we assign the state corresponding to its strongest current semantic status. For example, an agent that accepts e and provides new supporting reasoning is labeled as A rather than I. Similarly, an agent is labeled as R only if it was previously affected by the target error and later corrects it. These rules combine the current semantic content with the necessary interaction history to construct complete SEIAR-D state trajectories.

C Proofs of Theoretical Results

This section provides the complete proofs of the main theoretical results presented in the paper.

C.1 Proof of Theorem 1

Proof. Recall that

$$\mathcal{S}_0 = \Omega, \qquad \mathcal{S}_{\ell+1} = \mathcal{S}_\ell \cap \mathcal{T}_\ell, \tag{19}$$

where

$$\mathcal{T}_\ell = \mathcal{E}_\ell \cap \mathcal{U}_\ell. \tag{20}$$

Hence, $\mathcal{S}_L$ is exactly the event that propagation succeeds over all L hops of path π.

For each hop ℓ, conditioned on successful propagation over the preceding ℓ hops, we have

$$\begin{aligned}\Pr(\mathcal{T}_\ell \mid \mathcal{S}_\ell) &= \Pr(\mathcal{E}_\ell \cap \mathcal{U}_\ell \mid \mathcal{S}_\ell) && (21)\\ &= \Pr(\mathcal{E}_\ell \mid \mathcal{S}_\ell)\Pr(\mathcal{U}_\ell \mid \mathcal{E}_\ell, \mathcal{S}_\ell) && (22)\\ &= \beta_\ell^{\pi,e}\alpha_\ell^{\pi,e} && (23)\\ &= q_\ell^{\pi,e}. && (24)\end{aligned}$$

Since $\mathcal{S}_0 = \Omega$, we have $\Pr(\mathcal{S}_0) = 1$. Repeated application of the conditional probability rule therefore gives

$$\begin{aligned}\Pr(\mathcal{S}_L) &= \prod_{\ell=0}^{L-1} \Pr(\mathcal{T}_\ell \mid \mathcal{S}_\ell) && (25)\\ &= \prod_{\ell=0}^{L-1} q_\ell^{\pi,e} && (26)\\ &= Q_\pi^e. && (27)\end{aligned}$$

This factorization follows from the chain rule and does not require independence among hops.

Because π is propagation-feasible,

$$Q_\pi^e > 0, \tag{28}$$

and therefore

$$\Pr(\mathcal{S}_L) > 0. \tag{29}$$

If $\mathcal{S}_L$ occurs, then in particular the final hop $\mathcal{T}_{L-1}$ succeeds. By definition, $\mathcal{T}_{L-1}$ includes the event $\mathcal{U}_{L-1}$, under which the destination $v_L = a_j$ enters state I or A after receiving the final-hop message. Since t_L is defined as the observation time after this update,

$$\mathcal{S}_L \subseteq \left\{X_j^e(t_L) \in \{I, A\}\right\}. \tag{30}$$

Therefore,

$$\begin{aligned}\Pr\left(X_j^e(t_L) \in \{I, A\}\right) &\geq \Pr(\mathcal{S}_L) && (31)\\ &= Q_\pi^e && (32)\\ &> 0. && (33)\end{aligned}$$

The argument relies only on the existence of the propagation-feasible temporal path π and does not require a direct edge from a_m to a_j. Hence, even when $(a_m, a_j) \notin E_t$ for all t, the target error can reach a_j with non-zero probability through the relay path. □

C.2 Proof of Theorem 2

Proof. Recall that

$$\mathcal{R}_M = \bigcup_{k=1}^{M} \mathcal{S}_{\pi_k} \tag{34}$$

denotes the event that at least one of the first M propagation-feasible paths successfully propagates the target error.

After adding another propagation-feasible path π_{M+1}, the corresponding cumulative success event becomes

$$\mathcal{R}_{M+1} = \mathcal{R}_M \cup \mathcal{S}_{\pi_{M+1}}. \tag{35}$$

This union can be written as the disjoint union

$$\mathcal{R}_{M+1} = \mathcal{R}_M \,\dot\cup\, \left(\mathcal{S}_{\pi_{M+1}} \setminus \mathcal{R}_M\right), \tag{36}$$

where the two events on the right-hand side are mutually exclusive. Therefore,

$$P^e_{M+1} = \Pr(\mathcal{R}_{M+1}) \tag{37}$$
$$= \Pr(\mathcal{R}_M) + \Pr\big(\mathcal{S}_{\pi_{M+1}} \setminus \mathcal{R}_M\big) \tag{38}$$
$$= P^e_M + \Pr\big(\mathcal{S}_{\pi_{M+1}} \setminus \mathcal{R}_M\big)\,. \tag{39}$$

Since probabilities are non-negative,

$$\Pr\big(\mathcal{S}_{\pi_{M+1}} \setminus \mathcal{R}_M\big) \geq 0, \tag{40}$$

which immediately yields

$$P^e_{M+1} \geq P^e_M. \tag{41}$$

Moreover, if

$$\Pr\big(\mathcal{S}_{\pi_{M+1}} \setminus \mathcal{R}_M\big) > 0, \tag{42}$$

then

$$P^e_{M+1} > P^e_M. \tag{43}$$

Hence, adding a propagation-feasible path cannot decrease the cumulative probability that at least one of the considered paths successfully propagates the target error. The increase is strict whenever the newly added path contributes a positive-probability success event that is not already covered by the existing path set. No independence assumption among the path-success events is required. □

C.3 Proof of Theorem 3

Proof. Under the monotonicity assumption, for every hop ℓ,

$$\alpha_\ell^{A,\pi,e} \geq \alpha_\ell^{I,\pi,e}. \tag{44}$$

Since the two counterfactual processes share the same effective exposure probability,

$$q_\ell^{A,\pi,e} = \beta_\ell^{\pi,e}\alpha_\ell^{A,\pi,e} \tag{45}$$
$$\geq \beta_\ell^{\pi,e}\alpha_\ell^{I,\pi,e} \tag{46}$$
$$= q_\ell^{I,\pi,e}. \tag{47}$$

Because all hop-level propagation probabilities are non-negative, multiplying these inequalities over all hops yields

$$Q_\pi^{A,e} = \prod_{\ell=0}^{L-1} q_\ell^{A,\pi,e} \tag{48}$$
$$\geq \prod_{\ell=0}^{L-1} q_\ell^{I,\pi,e} \tag{49}$$
$$= Q_\pi^{I,e}. \tag{50}$$

This proves the first claim.

Now suppose that

$$Q_\pi^{I,e} > 0 \tag{51}$$

and

$$\alpha_\ell^{I,\pi,e} > 0, \qquad \forall \ell. \tag{52}$$

Since

$$Q_\pi^{I,e} = \prod_{\ell=0}^{L-1} \beta_\ell^{\pi,e}\alpha_\ell^{I,\pi,e}, \tag{53}$$

and all factors are non-negative, $Q_\pi^{I,e} > 0$ implies

$$\beta_\ell^{\pi,e} > 0, \qquad \forall \ell. \tag{54}$$

Hence the following ratios are well defined:

$$\frac{Q_\pi^{A,e}}{Q_\pi^{I,e}} = \prod_{\ell=0}^{L-1} \frac{q_\ell^{A,\pi,e}}{q_\ell^{I,\pi,e}} \tag{55}$$

$$= \prod_{\ell=0}^{L-1} \frac{\beta_\ell^{\pi,e}\alpha_\ell^{A,\pi,e}}{\beta_\ell^{\pi,e}\alpha_\ell^{I,\pi,e}} \tag{56}$$

$$= \prod_{\ell=0}^{L-1} \frac{\alpha_\ell^{A,\pi,e}}{\alpha_\ell^{I,\pi,e}} \tag{57}$$

$$= \prod_{\ell=0}^{L-1} \gamma_\ell^{\pi,e}. \tag{58}$$

By the monotonicity assumption,

$$\gamma_\ell^{\pi,e} = \frac{\alpha_\ell^{A,\pi,e}}{\alpha_\ell^{I,\pi,e}} \geq 1, \qquad \forall \ell. \tag{59}$$

If there exists at least one hop $\ell^\star$ such that

$$\gamma_{\ell^\star}^{\pi,e} > 1, \tag{60}$$

then

$$\prod_{\ell=0}^{L-1} \gamma_\ell^{\pi,e} > 1. \tag{61}$$

Therefore,

$$\frac{Q_\pi^{A,e}}{Q_\pi^{I,e}} > 1, \tag{62}$$

and, since $Q_\pi^{I,e} > 0$,

$$Q_\pi^{A,e} > Q_\pi^{I,e}. \tag{63}$$

Hence, under shared exposure, semantic amplification weakly increases the propagation probability along the path, and the increase is strict whenever at least one hop has a strictly positive local amplification gain beyond unity. □

C.4 Proof of Theorem 4

Proof. By iterating Eq. (8), we obtain

$$\mathbf{n}_g = K^g \mathbf{n}_0. \tag{64}$$

First, suppose that

$$\rho(K) < 1. \tag{65}$$

A standard result from matrix stability theory states that

$$\rho(K) < 1 \quad \Longrightarrow \quad \lim_{g\to\infty} K^g = \mathbf{0}_{2\times 2}. \tag{66}$$

Therefore, for any finite initial vector $\mathbf{n}_0$,

$$\lim_{g\to\infty} \mathbf{n}_g = \lim_{g\to\infty} K^g \mathbf{n}_0 = \mathbf{0}. \tag{67}$$

Hence, when $\rho(K) < 1$, the expected numbers of newly generated I- and A-state carriers decay asymptotically to zero.

Now suppose that

$$\rho(K) > 1. \tag{68}$$

Since K is a nonnegative matrix, the Perron–Frobenius theorem guarantees the existence of a nonzero nonnegative vector $\mathbf{v} \geq \mathbf{0}$ such that

$$K\mathbf{v} = \rho(K)\mathbf{v}. \tag{69}$$

Choose the initial condition

$$\mathbf{n}_0 = c\mathbf{v}, \qquad c > 0. \tag{70}$$

Then

$$\begin{aligned} \mathbf{n}_g &= K^g \mathbf{n}_0 && (71)\\ &= cK^g \mathbf{v} && (72)\\ &= c\rho(K)^g \mathbf{v}. && (73) \end{aligned}$$

Consequently, for any vector norm,

$$\|\mathbf{n}_g\| = c\,\rho(K)^g \|\mathbf{v}\|. \tag{74}$$

Because $\rho(K) > 1$, this quantity grows exponentially with the propagation generation g. Thus, there exists a nonzero nonnegative initial condition for which the linearized propagation process exhibits exponential growth.

Therefore, $\rho(K) = 1$ forms the critical boundary separating the regime of asymptotic decay, $\rho(K) < 1$, from the regime in which exponential growth is possible, $\rho(K) > 1$, under the early-stage linear approximation. □

C.5 Proof of Theorem 5

Proof. For any $t < T$ and $z \in \mathcal{Z}_t$, condition on the next macro-level state $\mathbf{Z}_{t+1}$. By the law of total probability,

$$\begin{aligned} r_t(z) &= \Pr\big(Z^e_{\text{sys}} = C_{\text{tar}} \mid \mathbf{Z}_t = z\big) && (75)\\ &= \sum_{z' \in \mathcal{Z}_{t+1}} \Pr(\mathbf{Z}_{t+1} = z' \mid \mathbf{Z}_t = z) \\ &\qquad \times \Pr\big(Z^e_{\text{sys}} = C_{\text{tar}} \mid \mathbf{Z}_{t+1} = z', \mathbf{Z}_t = z\big)\,. && (76) \end{aligned}$$

Under the second relation in Eq. (11), applied at time $t+1$, once $\mathbf{Z}_{t+1}$ is given, the previous macro-state history provides no additional information about the final system outcome. Hence,

$$\begin{aligned} &\Pr\big(Z^e_{\text{sys}} = C_{\text{tar}} \mid \mathbf{Z}_{t+1} = z', \mathbf{Z}_t = z\big) \\ &\quad = \Pr\big(Z^e_{\text{sys}} = C_{\text{tar}} \mid \mathbf{Z}_{t+1} = z'\big) = r_{t+1}(z'). && (77) \end{aligned}$$

Using the definition

$$P_t(z, z') = \Pr(\mathbf{Z}_{t+1} = z' \mid \mathbf{Z}_t = z)\,, \tag{78}$$

we obtain

$$r_t(z) = \sum_{z' \in \mathcal{Z}_{t+1}} P_t(z, z')\, r_{t+1}(z'), \qquad t < T. \tag{79}$$

At the terminal round $t = T$, by definition,

$$r_T(z) = \Pr\big(Z^e_{\text{sys}} = C_{\text{tar}} \mid \mathbf{Z}_T = z\big)\,. \tag{80}$$

Thus, r_T provides the terminal condition. Given the transition kernels P_t, the recursion determines r_{T-1} uniquely from r_T, then r_{T-2} uniquely from r_{T-1}, and so on. By backward induction, $r_t(z)$ is uniquely determined for every $t = 0, \ldots, T$. □

D Supplementary Experiments

D.1 Round-wise SEIAR-D State Evolution

Fig. 6 further shows the state evolution on MiniMax-M3 across discussion rounds. Agent states branch from S into E, I, A, and D, while some agents enter R through recovery. These states continue to evolve over subsequent rounds.

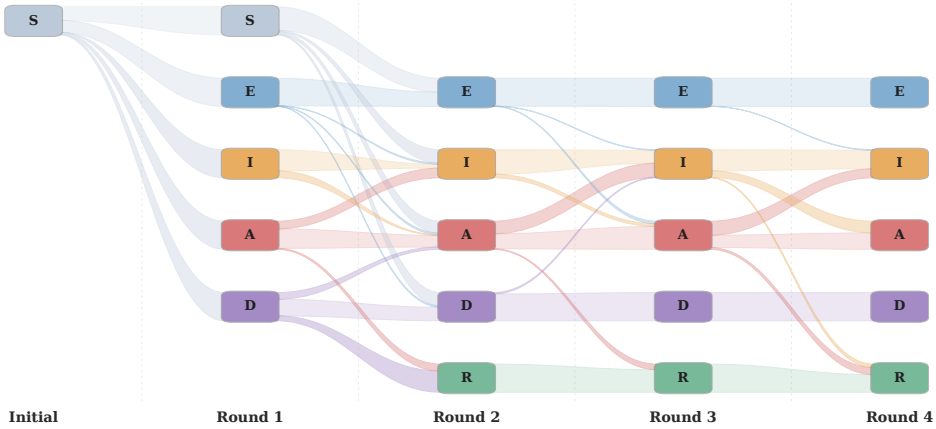


Figure 6: State evolution across multi-agent discussion rounds.

D.2 Additional Analysis of Propagation Criticality

D.2.1 Next-Generation Matrices

To explain the spectral radius $\rho(K)$ reported in the main text, we further report the next-generation matrices under No-A, Natural, and No-I for all four models. Here, k_{uv} denotes the expected number of new state-u carriers generated by one state-v carrier in the next propagation generation. We count an agent as a new carrier only when it newly enters state I or A, while persistence in the same state is not counted repeatedly. For each new state-u carrier, its state-v source is determined by the upstream message that directly triggers the transition. Accordingly, we estimate

$$\hat{k}_{uv} = \frac{C_{uv}}{N_v}, \tag{81}$$

where C_{uv} is the number of new state-u carriers generated by state-v carriers, and N_v is the total number of state-v carriers in the corresponding propagation generation. The full results are shown in Table 6.

Different settings directly change the available propagation channels. No-A removes all channels related to state A, leaving only $I \to I$. No-I leaves only $A \to A$. Natural allows both I and A carriers to participate in propagation, resulting in a richer matrix structure.

The Natural setting shows clear model differences. For MiniMax-M3 and GPT-5.6-Luna, propagation is mainly driven by A carriers, with $k_{AA} = 2.211$ and 1.197, respectively. GPT-5.6-Luna also has a strong $A \to I$ channel with $k_{IA} = 3.152$. DeepSeek-V4-Pro contains $I \to A$, $A \to I$, and $A \to A$ channels, forming a more complete feedback structure. In contrast, Gemini-2.5-Flash has a large $k_{IA} = 5.000$, but lacks the $I \to A$ feedback channel and has only $k_{AA} = 0.250$. Thus, it does not form sustained population-level feedback.

D.2.2 Propagation Trajectories in a Finite Population

Fig. 7 further shows how the proportion of I/A carriers changes across propagation generations under the three settings. Under No-A, carrier prevalence remains relatively low and becomes stable or decreases after the early stage. Natural shows stronger model differences. MiniMax-M3, GPT-5.6-Luna, and DeepSeek-V4-Pro increase early and then approach a plateau, while Gemini-2.5-Flash gradually declines. Under No-I, all four models show clear early expansion followed by a relatively stable plateau.

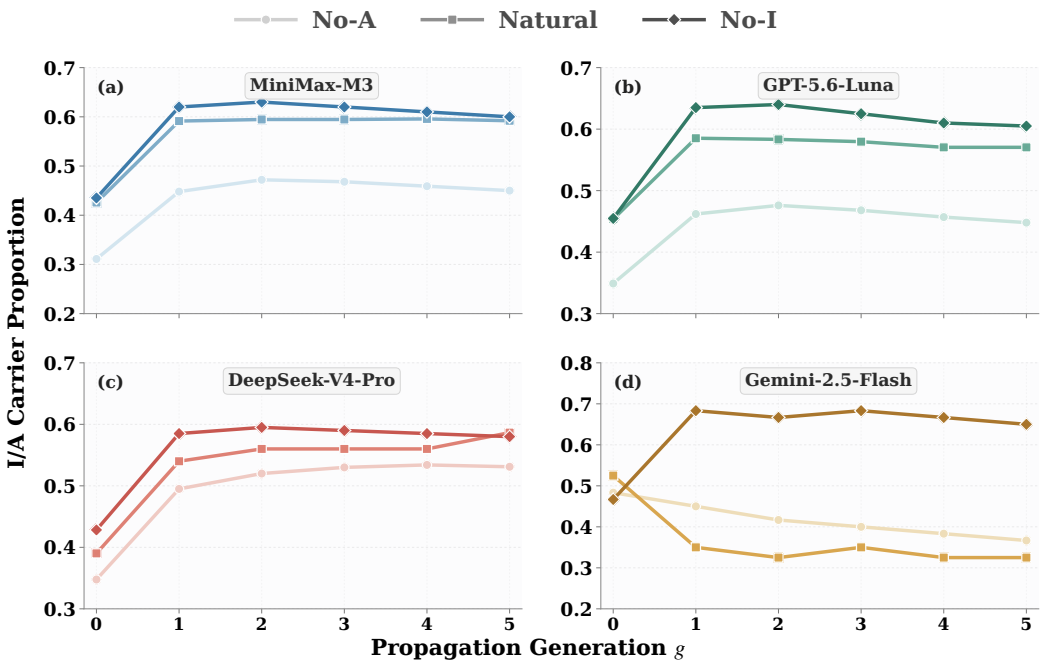


Figure 7: Population-normalized I/A prevalence across propagation generations.

Table 6: Estimated next-generation matrices and propagation regimes with three settings.

Model	Setting	k_{II}	k_{IA}	k_{AI}	k_{AA}	$\rho(K)$	Regime
MiniMax-M3	No-A	0.672	0.000	0.000	0.000	0.672	Subcritical
	Natural	0.056	0.657	0.000	2.211	2.211	Supercritical
	No-I	0.000	0.000	0.000	3.478	**3.478**	Supercritical
GPT-5.6-Luna	No-A	0.455	0.000	0.000	0.000	0.455	Subcritical
	Natural	0.000	3.152	0.000	1.197	1.197	Supercritical
	No-I	0.000	0.000	0.000	4.543	**4.543**	Supercritical
DeepSeek-V4-Pro	No-A	0.945	0.000	0.000	0.000	0.945	Subcritical
	Natural	0.000	0.788	1.333	2.442	2.816	Supercritical
	No-I	0.000	0.000	0.000	3.431	**3.431**	Supercritical
Gemini-2.5-Flash	No-A	0.167	0.000	0.000	0.000	0.167	Subcritical
	Natural	0.000	5.000	0.000	0.250	0.250	Subcritical
	No-I	0.000	0.000	0.000	4.667	**4.667**	Supercritical

D.3 Early-Warning Baselines and Additional Results

D.3.1 Early-Warning Baselines

To evaluate the dynamic consolidation risk, we compare three types of baselines. I/A Ratio measures the proportion of non-malicious agents currently in states I or A, reflecting the current error coverage. IBMF (Xie et al., 2026) models error propagation as an infection process over the communication graph. It projects the current infection state using propagation and recovery dynamics, and uses the expected final infection coverage as the risk score. Static State uses the current macro state, including the Target Agent state, the numbers of normal agents in different error states, and the current round, to directly estimate the probability of reaching C_{tar}. It does not explicitly model state transitions. These baselines represent current error coverage, dynamical projection, and static outcome prediction, respectively.

Table 7: Early-warning performance across models.

Model	Method	AUROC	F1	FAR	Warn. Round
GPT-5.6-Luna	Static State	0.9846	93.18	1.46	2.17
	I/A Ratio	0.9512	89.47	2.91	1.38
	IBMF	0.8427	76.19	**0.94**	**1.09**
	Ours	**0.9918**	**95.65**	1.21	1.27
DeepSeek-V4-Pro	Static State	0.9793	91.89	1.73	2.24
	I/A Ratio	0.9468	89.02	**1.28**	1.44
	IBMF	0.8265	74.42	1.06	**1.12**
	Ours	**0.9887**	**94.12**	1.36	1.31
Gemini-2.5-Flash	Static State	0.9761	**93.10**	1.31	2.09
	I/A Ratio	0.9394	87.50	2.54	1.35
	IBMF	0.8116	72.73	**0.79**	**1.05**
	Ours	**0.9862**	92.86	1.18	1.24

D.3.2 Results on Additional Models

We further evaluate the same methods on GPT-5.6-Luna, DeepSeek-V4-Pro, and Gemini-2.5-Flash using the same protocol as in the main text. For each model, the warning threshold is selected on the validation set and fixed for test evaluation. Table 7 reports AUROC, F1, False Alarm Rate (FAR), and the average warning round.