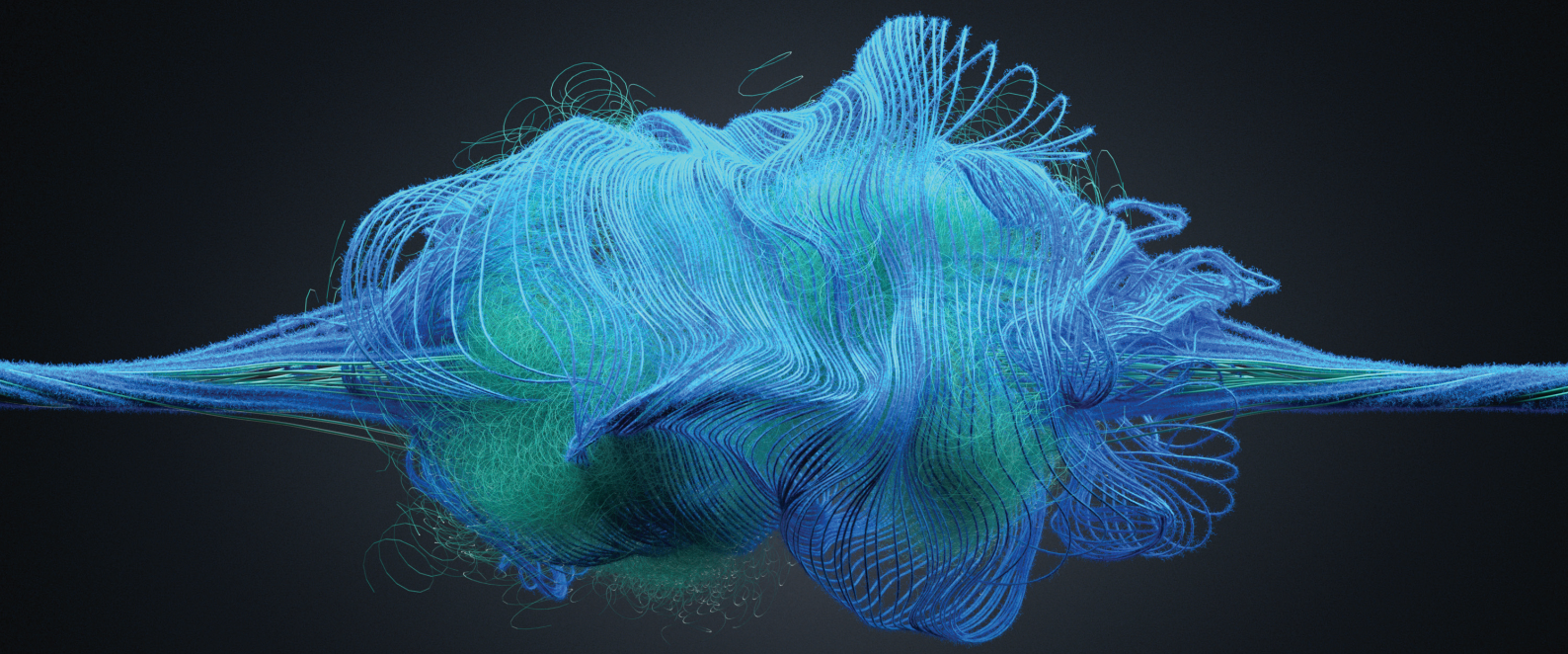
zensar

Formal Modeling of Error Propagation in Contextual Memory Systems:

A Dynamical-Systems Theory of Context Poisoning in Large Language Models

Research paper



An RPG Company

Abstract

Practitioners have repeatedly observed that large language models (LLMs) accumulate and amplify errors as conversations or agentic workflows grow longer. A single mistaken claim, once admitted into the context window, tends to be treated as an established fact and reproduced or elaborated on in subsequent turns. At the same time, overall reliability can degrade as the input grows, even when no error has been introduced. These two related but distinct phenomena are commonly referred to as context poisoning and context rot. To date, explanations of both have remained largely empirical, relying on benchmark curves and qualitative taxonomies rather than a predictive formal theory. This paper develops such a theory. We model the context window as a discrete-time dynamical system in which a scalar or vector-valued contamination state evolves under a propagation operator. The operator's spectral properties determine whether injected errors decay, persist, or grow without bound, drawing a direct analogy with pole placement and stability analysis in classical control theory. We derive closed-form expressions for the contamination trajectory; define an Error Propagation Coefficient, a contamination half-life, and a compounding factor; and show how position-dependent attention weighting and context-length-dependent correction decay enter the model as time-varying parameters. We then extend the framework to multi-agent systems using a communication matrix over an interaction graph. Applying the Perron-Frobenius theorem, we establish a stability bound showing that systemic contamination growth is governed chiefly by the worst-case per-agent amplification gain, rather than by network topology alone. The resulting framework offers falsifiable, quantitative predictions that can be fitted to existing empirical benchmarks and used to design targeted mitigations, such as periodic re-grounding, verifier agents, and context-compaction schedules.

Keywords: large language models, context contamination, error propagation, hallucination snowballing, and multi-agent systems

1. Introduction

Large language models increasingly operate over long horizons: extended multi-turn conversations, retrieval-augmented pipelines that inject external documents, and multi-agent systems in which several model instances exchange intermediate results before producing a final answer. Across all three settings, practitioners have documented a recurring failure pattern: an early, localized error —such as a hallucinated fact, a misread instruction, or a faulty retrieval — is not merely retained but actively reinforced by the model's subsequent generations. This produces internally coherent yet increasingly inaccurate trajectories (Cemri et al., 2025; Zhang et al., 2024). A related but distinct phenomenon, context rot, describes performance degradation as a function of raw input length, even when no error has been introduced (Hong et al., 2025). Both phenomena have been consistently reported, but the literature documenting them remains largely descriptive. It consists of controlled benchmarks that vary needle position, distractor count, or conversation length, alongside qualitative failure taxonomies (Cemri et al., 2025; Hong et al., 2025; Liu et al., 2023). No formal account so far treats error propagation through a context window as a dynamical process with predictive, quantitative structure comparable to the mature theory of error propagation in feedback control systems (Åström et al., 2008).

This gap matters for two reasons. First, without a formal propagation model, it is difficult to compare interventions — such as larger context windows, summarization, retrieval re-ranking, and verifier agents — on a common footing. Each is evaluated against a different benchmark rather than a shared theoretical criterion for stability. Second, a formal model generates falsifiable predictions — for example, about how contamination should scale with conversation length, attention concentration, or the number of communicating agents. These predictions can be tested against existing and future empirical data rather than rediscovered for every system architecture.

This paper's contribution is a minimal but complete formal framework for context-contamination dynamics. We treat the contextual state of a model, whether single-agent or multi-agent, as a variable that evolves under a discrete-time recursion. Its coefficients are grounded in known empirical regularities: attention concentration and positional bias (Liu et al., 2023; Xiao et al., 2023), degradation of retrieval fidelity with input length (Hong et al., 2025), and self-reinforcing justification of prior claims (Zhang et al., 2024). From this recursion, we derive stability conditions, closed-form contamination trajectories, and summary metrics. We also extend the model to networks of communicating agents using standard results from matrix analysis and consensus theory (Horn et al., 2012; Olfati-Saber et al., 2023).

2. Related Work

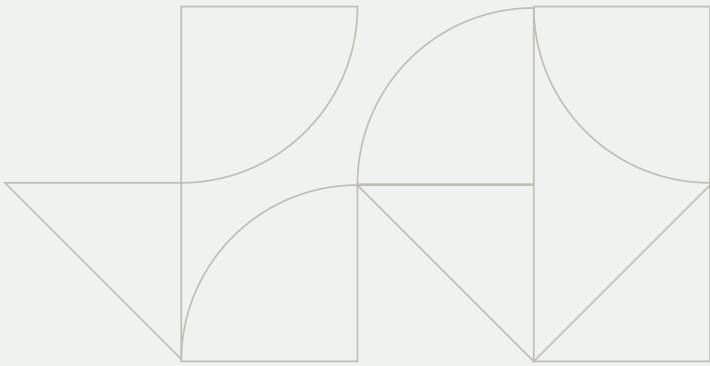
Empirical documentation of long-context failure falls into three overlapping strands. The first concerns positional sensitivity within a single context window. Liu et al. showed that retrieval accuracy over a document set embedded in a long prompt follows a U-shaped curve: information near the beginning or end of the context is retrieved far more reliably than information placed in the middle. This pattern is now widely termed "lost in the middle" (Liu et al., 2023). A partial mechanistic explanation comes from the attention-sink literature. Xiao et al. found that autoregressive transformers allocate disproportionate attention to a small number of early tokens, largely irrespective of their semantic content. This stabilizes generation under sliding-window truncation but also implies that the effective influence of any given context segment on the output is strongly nonuniform and position-dependent, rather than a simple function of relevance (Xiao et al., 2023).

The second strand concerns degradation with raw input length, independent of content difficulty. Chroma's technical report evaluated 18 contemporary models on controlled retrieval and replication tasks. It found that reliability decreases as input length grows, even on tasks that are trivial at shorter lengths. This degradation is shaped by factors such as distractor similarity and haystack structure, rather than task difficulty alone. The report names this phenomenon context rot (Hong et al., 2025). This literature establishes that a model's correction and retrieval capacity decreases as a function of context length, a property our framework treats as a time-varying parameter rather than a constant.

The third strand concerns self-reinforcing error accumulation across turns or reasoning steps. Zhang et al. demonstrated that language models frequently justify an initially incorrect answer with a further incorrect explanation, a pattern they term "hallucination snowballing." They also showed that models can often identify these same errors as false when the claims are presented in isolation, indicating that the failure stems from contextual overcommitment rather than a pure knowledge gap (Zhang et al., 2024). Chain-of-thought prompting, which decomposes reasoning into explicit intermediate steps, has been shown to improve accuracy on many tasks (Wei et al., 2022). However, a longer explicit reasoning chain also creates more opportunities for an early misstep to be built upon, a tension our model makes explicit through the amplification-gain parameter introduced in Section 3.

At the level of multi-agent systems, Cemri et al. conducted a large-scale empirical taxonomy of multi-agent LLM failures. They analyzed seven frameworks across more than 200 tasks and identified 14 distinct failure modes, grouped into specification failures, inter-agent misalignment, and task-verification failures (Cemri et al., 2025). Their finding that verification failures are pervasive, and that performance gains from adding agents are often minimal, motivates the multi-agent stability analysis in Section 4. There, we show formally why the presence or absence of a verifying node, rather than the raw number of communicating agents, is the primary lever affecting system-wide contamination growth.

None of this literature offers a propagation model with explicit dynamics, stability conditions, or derived scaling laws. Classical control theory, by contrast, has a mature apparatus for addressing precisely this kind of question. A system's response to a disturbance is governed by the eigenvalues (poles) of its state-transition operator, and stability in discrete-time systems is equivalent to those eigenvalues lying inside the unit disk (Åström et al., 2008). The Perron-Frobenius theory provides corresponding tools for systems whose transition operator is nonnegative or row-stochastic, as occurs naturally when propagation takes place over a communication graph (Horn et al., 2012). It has also been used to analyze consensus and agreement dynamics among networked agents with switching topologies (Olfati-Saber et al., 2023). Retrieval augmentation (Lewis et al., 2020) and calibration benchmarks such as TruthfulQA (Lin et al., 2022) provide natural candidate correction mechanisms and evaluation targets for the model developed below. The broader hallucination survey by Ji et al. situates context poisoning within the wider landscape of generation-faithfulness failures (Ji et al., 2023). Section 3 brings these tools into the language-model setting.



3. A Formal Framework for Single-Agent Contamination Dynamics

3.1 State representation

Let a context window at turn t consist of an ordered sequence of segments Ct = (s1, s2, ..., snt), where each segment may be a token span, a message turn, or a retrieved document, and nt is the number of segments present at turn t. Each segment carries a latent truth label ei Σ{0,1}, equal to one if the segment is factually or logically erroneous, and zero otherwise. Because the model cannot directly observe this label, we instead work with the aggregate contamination fraction

$$x_t = \Sigma_i \, w_{i,t} \, e_i \quad (1)$$

where wi,t is the effective reliance weight the model places on segment i when generating its output at turn t, normalized so that Σi wi,t = 1. The quantity xt is a continuous relaxation of the discrete labels: it represents the attention-weighted proportion of the context that is currently false and plays the same role in this framework that a tracking error plays in a feedback control loop (Åström et al., 2008).

3.2 The propagation recursion

We model how contamination evolves from one turn to the next as

$$x_{t+1} = (1 - \lambda_t) \, x_t + \lambda_t \, g(x_t) + \varepsilon_t - \mu_t \quad (2)$$

Here, λt [0,1] is the reliance coefficient: the fraction of the model's generative process at turn t that conditions on prior context rather than on parametric knowledge or fresh evidence. The function g(·) is an amplification function that captures self-reinforcing justification, with g(x) ≥ x for x > 0. This reflects the empirical observation that models elaborate on prior errors rather than merely repeat them (Zhang et al., 2024). The term εt ≥ 0 represents exogenous contamination newly injected at turn t (for example, a faulty retrieval, a false user premise, or a tool error), while μt ≥ 0 represents the correction provided by verification, grounding, or retrieval mechanisms active at that turn. If the underlying labels ei are also allowed to change according to fixed transition probabilities as segments age and are re-attended, the joint process over context configurations forms a Markov chain. The propagation coefficient derived below then plays a role analogous to that of the second eigenvalue of a transition matrix in classical mixing-time analysis (Levin et al., 2017).

Linearizing g around small contamination levels gives g(xt) ≈ γt xt, where γt > 1 is the overcommitment gain. Equation (2) therefore reduces to the scalar linear recursion.

$$x_{t+1} = \rho_t \, x_t + \varepsilon_t - \mu_t, \quad \rho t = (1 - \lambda_t) + \lambda_t \, \gamma_t \quad (3)$$

The coefficient ρt is a discrete-time pole, as in linear control theory (Åström et al., 2008). When ρt < 1 for all t, the influence of an injected error decays geometrically over subsequent turns. When ρt = 1, the error persists without decaying or growing (contamination is "sticky"). When ρt > 1, it grows without bound, representing the formal signature of hallucination snowballing.

3.3 Closed-form trajectory and steady state

Assuming constant coefficients ρ, ε, and μ over a window of t turns, Equation (3) has the following closed-form solution:

$$x_t = \rho_t \, x_0 + (\varepsilon - \mu) \cdot (1 - \rho t) / (1 - \rho), \quad \rho \neq 1 \quad (4)$$

For ρ < 1, the trajectory converges to the fixed point x* = (ε − μ) / (1 − ρ), a bounded equilibrium contamination level analogous to the steady-state error of a stable feedback loop under a constant disturbance (Åström et al., 2008). For ρ ≥ 1 and ε > μ, Equation (4) is unbounded in t, formally reproducing the compounding-across-turns behavior described qualitatively in the practitioner literature on context poisoning (Hong et al., 2025).

3.4 Position-dependent reliance weights

The reliance weights wi,t in Equation (1) are not uniform across segments. Empirically, effective attentional influence follows a U-shaped positional profile: it is high near the beginning and end of the context and suppressed in the middle (Liu et al., 2023). This pattern is reinforced by the tendency of early tokens to absorb disproportionate attention, regardless of content (Xiao et al., 2023). We represent this with a stylized weight profile:

$$w_i \propto \exp(-\alpha \cdot \min(i - 1, n - i)), \quad \Sigma_i \, w_i = 1 \quad (5)$$

where α > 0 governs how sharply middle segments are discounted relative to boundary segments. Equation (5) implies that a contamination event ei = 1 introduced near the middle of a long context contributes little to xt directly through Equation (1), even though it may later migrate toward the recency boundary as the conversation continues and become influential again. The scalar model approximates this second-order effect through the time dependence of λt.

3.5 Context length and correction decay

Chroma's evaluation shows that reliability on tasks such as retrieval and exact replication degrades monotonically as input length grows, independent of task difficulty (Hong et al., 2025). We capture this by making the correction term in Equation (2) a decreasing function of context length nt:

$$\mu_t = \mu_0 \cdot \varphi(n_t), \quad \varphi'(n) < 0, \ \varphi(0) = 1 \quad (6)$$

where μ0 is the correction efficacy at negligible context length (for example, a verifier operating on a short excerpt), and φ is a decreasing fidelity function fitted empirically from context-rot curves. Substituting Equation (6) into Equation (3), the effective pole becomes

$$\rho_{eff}(n_t) = (1 - \lambda_t) + \lambda_t \gamma_t \quad (7)$$

The corresponding disturbance term, εt − μ0 φ(nt), grows in magnitude as nt increases because φ decreases. This produces the model's key qualitative prediction: even a system with ρt < 1 (nominally stable and self-correcting) can exhibit rising steady-state contamination purely as a function of accumulated length, because the corrective term weakens even while the propagation pole remains fixed. This formally distinguishes context rot (a weakening of μ) from context poisoning proper (an increase in ρ). Although these phenomena are often discussed together in the practitioner literature, the model shows that they are mechanistically distinct.

4. Extension to Multi-Agent Systems

4.1 Network propagation model

In a multi-agent system, N agents exchange messages over a communication structure that can be represented as a graph with a row-stochastic weighting matrix P ∈ RN×N, where Pij is the share of agent j's influence on agent i's next-turn context. This is consistent with standard treatments of networked consensus dynamics (Olfati-Saber et al., 2023). Writing the vector of per-agent contamination fractions as xt ∈ [0,1]N, the natural vector generalization of Equation (3) is

$$x_{t+1} = [(1 - \lambda) I + \lambda \Gamma P] x_t + \varepsilon_t \quad (8)$$

where Γ = diag(γ1, ..., γN) collects the per-agent overcommitment gains, and I is the identity matrix. Equation (8) states that each agent's contamination at the next turn is a convex combination of its own prior contamination and the gain-weighted contamination it receives from its neighbors, plus any newly injected error.

4.2 A stability theorem

Let M = (1 − λ) I + λ Γ P de note the system matrix in Equation (8). As P is row-stochastic and nonnegative, the Perron-Frobenius theorem guarantees that its spectral radius is ρ(P) = 1, with the all-ones vector as the corresponding eigenvector (Horn et al., 2012). A standard bound on the spectral radius of a scaled and shifted stochastic matrix then gives

$$\rho(M) \le (1 - \lambda) + \lambda \cdot \gamma_{max} \cdot \rho(P) = (1 - \lambda) + \lambda \gamma_{max} \quad (9)$$

where γmax = maxi γi. The system is stable — meaning that any locally injected contamination decays across the network rather than growing without bound — if and only if

$$\rho(M) < 1 \iff \gamma_{max} < 1 \quad (10)$$

This is the paper's central theoretical result for multi-agent systems: under a row-stochastic communication structure, whether contamination decays or grows across the network depends only on the single worst-case per-agent amplification gain, not on the topology of the communication graph, the number of agents, or the density of their connections. This provides a formal basis for the empirical observation that adding more communicating agents does not reliably improve — and can actively harm — multi-agent system reliability (Cemri et al., 2025). If even one agent behaves as a net error amplifier (γi ≥ 1), the bound in Equation (10) fails, and network-wide contamination growth becomes possible regardless of how the remaining agents are connected.

4.3 The role of verifier agents

The bound in Equation (10) also formalizes why verification failures dominate the empirically observed multi-agent failure taxonomy (Cemri et al., 2025). A verifier agent v with $\gamma_v \approx 0$ (that is, an agent that checks claims against external evidence rather than propagating unverified context) lowers γ_{max} only if it is actually consulted — that is, only if $P_{iv} > 0$ for the agents whose outputs feed the final answer. A verifier that exists in the system but is bypassed for latency or routing reasons leaves γ_{max} (and hence the stability margin) unchanged. The model therefore predicts that verification effectiveness is a property of the communication matrix P as much as of any individual agent's design, an implication consistent with the specification and coordination failure categories documented empirically (Cemri et al., 2025).

5. Results

Derived metrics for error amplification

From the scalar recursion in Section 3, we define three summary quantities intended to be estimable from benchmark data and comparable across systems.

Error Propagation Coefficient (EPC): The pole ρ itself, estimated by regressing observed contamination at turn t+1 on contamination at turn t across a benchmark suite. EPC < 1 indicates a self-correcting system; EPC ≥ 1 indicates a system prone to context poisoning.

$$EPC = \rho = (1 - \lambda) + \lambda\gamma \qquad (11)$$

Contamination Half-Life: For $\rho < 1$, this is the number of turns required for an injected error's residual influence to fall to half its initial value:

$$T_{1/2} = \ln(0.5) / \ln(\rho) \qquad (12)$$

Compounding Factor: This is the multiplicative growth (or decay) of an error's residual influence after k turns, directly comparable to gain in an open-loop control chain:

$$CF(k) = \rho^k \qquad (13)$$

The corresponding effective memory horizon, $k^* = \ln(\delta / x_0) / \ln(\rho)$, gives the number of turns before contamination originating at level x_0 falls below an acceptable threshold δ. These three quantities provide a common, architecture-independent basis for comparing single-agent conversational systems, retrieval pipelines, and multi-agent deployments. This is analogous to how pole location and settling time provide a common basis for comparing feedback controllers, regardless of the physical process being controlled (Åström et al., 2008).

6. Mapping the Model to Documented Empirical Phenomena

The framework developed above is intended to organize, rather than merely restate, the empirical literature. Context rot (Hong et al., 2025) corresponds to a decline in $\varphi(n_t)$ (Equation 6) as context length grows, raising the steady-state contamination level x^* even when the propagation pole ρ remains unchanged. Lost-in-the-middle effects (Liu et al., 2023) and attention-sink behavior (Xiao et al., 2023) correspond to the position-dependent weighting in equation (5), which determines how strongly a given contaminated segment enters the aggregate state x_t through equation (1). Hallucination snowballing (Zhang et al., 2024) corresponds directly to the regime $\gamma > 1$ in the amplification function of Equation (2). Chain-of-thought reasoning (Wei et al., 2022) can be interpreted as increasing the number of discrete propagation steps t within a single model turn; according to Equation (4), this increases the opportunity for an early error to compound before a final answer is produced. Finally, the multi-agent failure taxonomy of Cemri et al. (Cemri et al., 2025) maps onto the stability theorem in Section 4: specification failures correspond to poorly designed per-agent gains γ_i; inter-agent misalignment corresponds to a communication matrix P that routes context in ways inconsistent with the assumptions in Section 4.1; and verification failures correspond to the condition described in Section 4.3, in which a low-gain verifier exists but is not actually consulted by the routing structure.

7. Discussion: Implications for Mitigation Designt

The model suggests three broad classes of intervention, each targeting a distinct term in Equation (3) or Equation (8). First, interventions that reduce λ, the reliance coefficient — such as periodically re-grounding generation in original source material rather than in accumulated conversational context — directly reduce ρ. Second, interventions that reduce γ, the overcommitment gain — such as prompting strategies or fine-tuning objectives that reward the retraction of previously stated claims rather than penalizing inconsistency — also reduce ρ. These interventions are predicted to be effective against hallucination snowballing specifically, rather than context rot. Third, interventions that raise μ and prevent φ(nt) from decaying — such as context compaction, summarization, or retrieval re-ranking that keeps the effective working set short regardless of raw conversation length — directly target context rot, as distinct from context poisoning. In multi-agent systems, Equation (10) implies that architectural interventions should prioritize ensuring that every information path to the final answer passes through at least one low-gain verifying node. Simply adding agents or increasing communication density is not predicted to have a first-order effect on stability.

8. Limitations and Future Work

The framework makes several simplifying assumptions that should be treated as a starting point rather than a finished account. The linearization of g(x) around small x is unlikely to hold globally; a fuller treatment would model g as a saturating nonlinearity because contamination cannot literally exceed the full context. The scalar contamination fraction xt also collapses heterogeneous error types — factual, logical, and instructional — into a single quantity. A vector-valued extension that distinguishes error categories, each with its own gain and correction dynamics, is a natural next step. The multi-agent communication matrix P is treated as fixed and row-stochastic. Real systems exhibit time-varying, non-stochastic, and adversarially manipulable routing, for which the switching-topology results of consensus theory (Olfati-Saber et al., 2023) offer a more appropriate starting point than the static bound derived here. Finally, and most importantly, every parameter in this framework — λ, γ, μ0, φ, α — is currently theoretical. Fitting these parameters against existing benchmark data, such as the context-rot and lost-in-the-middle evaluation suites (Hong et al., 2025; Liu et al., 2023), is the necessary next step to establish whether the model's quantitative predictions, in particular the separation of context rot from context poisoning as distinct mechanisms, survive contact with data.

9. Conclusion

Context poisoning and context rot have so far been documented empirically but not modeled formally. This paper proposes a discrete-time dynamical-systems account in which contamination in a context window evolves under a propagation recursion. Its stability is governed by an effective pole, directly analogous to error-propagation analysis in classical control theory. The paper extends this account to multi-agent systems through a Perron-Frobenius stability bound, showing that network-wide contamination growth depends chiefly on the worst-case per-agent amplification gain rather than on communication topology. The resulting Error Propagation Coefficient, contamination half-life, and compounding factor provide architecture-independent, quantitative targets that future work can fit against existing benchmarks and use to evaluate mitigations on a common theoretical footing, rather than through ad hoc, system-specific evaluations.

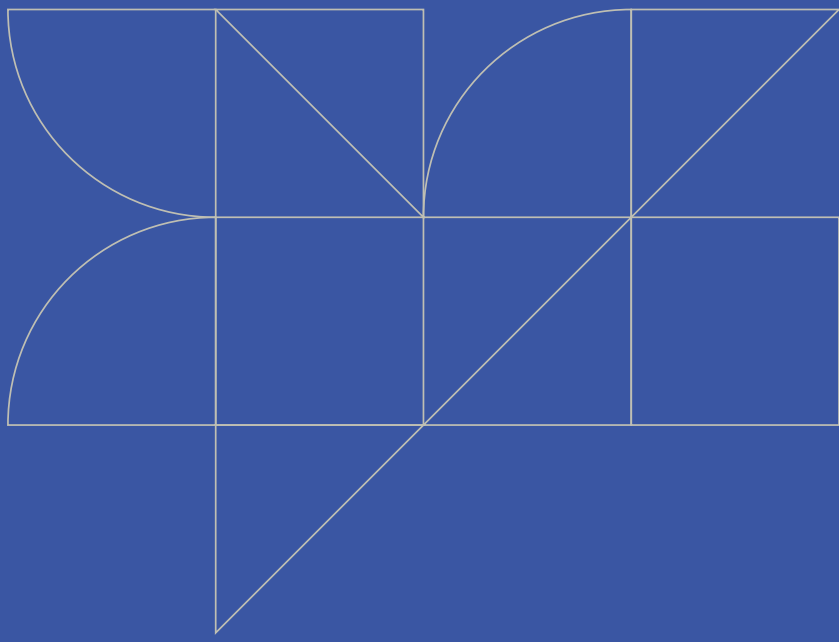
References

1. Åström, K. J., & Murray, R. M. (2008). Feedback systems: An introduction for scientists and engineers. Princeton University Press.
2. Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., & Stoica, I. (2025). Why do multi-agent LLM systems fail? arXiv. https://arxiv.org/abs/2503.13657
3. Hong, K., Troynikov, A., & Huber, J. (2025). Context rot: How increasing input tokens impacts LLM performance (Technical Report). Chroma.
4. https://research.trychroma.com/context-rot
5. Horn, R. A., & Johnson, C. R. (2012). Matrix analysis (2nd ed.). Cambridge University Press.
6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.
7. Levin, D. A., & Peres, Y. (2017). Markov chains and mixing times (2nd ed.). American Mathematical Society.
8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474).
9. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 3214–3252).
10. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. arXiv.
11. https://arxiv.org/abs/2307.03172
12. Olfati-Saber, R., & Murray, R. M. (2004). Consensus problems in networks of agents with switching topology and time-delays. IEEE Transactions on Automatic Control, 49(9), 1520–1533.
13. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (Vol. 35, pp. 24824–24837).
14. Xiao, G., Tian, Y., Chen, B., Han, S., & Lewis, M. (2023). Efficient streaming language models with attention sinks. arXiv. https://arxiv.org/abs/2309.17453
15. Zhang, M., Press, O., Merrill, W., Liu, A., & Smith, N. A. (2024). How language model hallucinations can snowball. In Proceedings of the 41st International Conference on Machine Learning (pp. 59670–59684). PMLR.

Authored by

Mouna Nataraja,

n.mouna@zensar.com



zensar

An RPG Company

At Zensar, we're 'experience-led everything.' We are committed to conceptualizing, designing, engineering, marketing, and managing digital solutions and experiences for over 145+ leading enterprises. Using our 3Es of experience, engineering, and engagement, we harness the power of technology, creativity, and insight to deliver impact.

Part of the $4.8 billion RPG Group, we are headquartered in Pune, India. Our 10,000+ employees work across 30+ locations worldwide, including Milpitas, Seattle, Princeton, Cape Town, London, Zurich, Singapore, and Mexico City.

For more information, please contact: **info@zensar.com | www.zensar.com**