

COMPLEXA
LABS

RESEARCH WHITE PAPER

# Error Dynamics in Bot Conversations

*A framework for measuring correction, retention, and error introduction in multi-agent AI systems*

Nicolò Ciccarone

COMPLEXA LABS

November 2026

© 2026 Nicolò Ciccarone / Complexa Labs. All rights reserved.

# Contents

**1 Introduction** **1**

1.1 Scope and claims . . . . . 1

**2 Literature Review** **1**

**3 Mathematical Framework** **2**

3.1 State variables and estimands . . . . . 2

3.2 Why multiplication appears . . . . . 3

3.3 Stochastic transitions . . . . . 3

3.4 Nested and competing specifications . . . . . 3

3.5 Correction plus introduction . . . . . 3

3.6 White-noise perturbations . . . . . 4

3.7 Multivariate error dynamics . . . . . 5

**4 Discussion** **6**

4.1 Hypotheses . . . . . 6

4.2 Empirical strategy . . . . . 6

4.3 Deployment implications . . . . . 6

**5 Conclusions** **7**

# 1 Introduction

AI systems are increasingly composed as conversations: one model drafts an answer, another critiques it, a third revises it, and the sequence may continue until a stopping rule is reached. The intuitive expectation is that repeated review reduces error. This is plausible, but it is not a mathematical consequence of a model's standalone benchmark accuracy. A score on a benchmark describes performance on a specified task distribution; it does not automatically identify the probability that a reviewer will correct an error which has survived previous reviews.

This white paper develops a compact framework for measuring error dynamics in multi-agent AI conversations. It distinguishes mathematical identities from substantive assumptions, defines the quantities that must be observed, and proposes an empirical strategy for comparing candidate models. The aim is practical: to determine when multi-agent review improves final-answer reliability, when it produces diminishing returns, and when it can create fresh risk.

The central object is a conditional correction probability. Let $E_t \geq 0$ denote residual error after review round $t$, and let $c_t \in [0, 1]$ denote the fraction of surviving error corrected by the next review. Under proportional correction,

$$E_{t+1} = E_t - c_t E_t = (1 - c_t) E_t. \tag{1}$$

Iterating equation (1) gives

$$E_T = E_0 \prod_{t=0}^{T-1} (1 - c_t). \tag{2}$$

The product is therefore an exact description of the stated transition rule. It is not evidence that the rule is appropriate, that $c_t$ is constant, or that $c_t$ is equal to a public accuracy score.

## 1.1 Scope and claims

This paper makes four claims. First, a multiplicative residual-error model is a natural representation of proportional correction. Second, a general additive-increment model contains this multiplicative model as a special case when the increment depends on the current state. Third, benchmark accuracy may be tested as a predictor of conditional correction, but it is not identical to it. Fourth, reliable deployment requires modeling both correction and the introduction of new errors.

# 2 Literature Review

The proposed framework sits at the intersection of conditional probability, forecast calibration, and AI evaluation. The relevant statistical lesson is that a quantity must be interpreted on its conditioning set. Conditional-independence theory formalizes why probabilities may change when information and history change [1]. In a conversational AI setting, the history may include the original task, previous answers, critique prompts, retrieved sources, reviewer identity, and whether an error has already survived earlier review.

Calibration provides a second useful distinction. A probability estimate is calibrated when its stated probabilities agree with observed frequencies under an appropriate evaluation protocol. Dawid's account of calibration makes clear that evaluating a forecast requires matching the forecast to an outcome and a reference class [2]. By analogy, a model's benchmark


© 2026 Nicolò Ciccarone / Complexa Labs. All rights reserved.

performance is only informative for conversation review when the benchmark task and evaluation protocol align with the conditional event of interest: correcting a surviving error in context.

Recent work on language-model confidence similarly treats calibration as an empirical property rather than a default implication of capability. Tian et al. evaluate methods for eliciting calibrated confidence from language models and show why the method used to obtain a probability matters [3]. This supports a conservative operational principle: do not convert an aggregate accuracy, confidence score, or vendor claim directly into a conversational transition probability without measuring the transition on the target workflow.

The literature does not supply a universal decay law for multi-agent conversations. The question is instead empirical. Do reviewers remove a constant amount of error, a constant fraction of the error that remains, or a context-dependent amount? Do surviving errors become systematically harder to detect? How often does a reviewer corrupt a correct element while repairing another? These questions motivate the model family developed below.

# 3 Mathematical Framework

## 3.1 State variables and estimands

The word *error* must be defined before selecting a model. The following quantities are related but not interchangeable.

| Quantity | Notation | Meaning |
|---|---|---|
| Residual error mass | $E_t \geq 0$ | Count, weighted count, or other amount of unresolved error after round $t$. |
| Current-answer error probability | $e_t \in [0, 1]$ | Probability that the current answer is wrong under a defined protocol. |
| Survival indicator | $I_t \in \{0, 1\}$ | Whether one particular initial error remains after round $t$. |
| Introduced-error count | $N_t$ | Number of new errors created by a revision or generation. |
| Final-answer correctness | $Y_t \in \{0, 1\}$ | Whether the complete output meets the ground-truth criterion. |

For controlled experiments, the primary estimand should be the conditional correction fraction of error cost:

$$c_t = \Pr(I_{t+1} = 0 \mid I_t = 1, H_t), \tag{3}$$

where $H_t$ is the observable conversation history. For an individual tracked error, this is a correction probability. For weighted error mass, it is the expected fraction of the currently unresolved cost corrected during the next review.



© 2026 Nicolò Ciccarone / Complexa Labs. All rights reserved.

### 3.2 Why multiplication appears

For an individual original error, let $S_t$ be the event that it survives $t$ reviews. If $c_t = \Pr(S_{t+1}^c \mid S_t, H_t)$, then the chain rule yields

$$\Pr(S_T) = \Pr(S_0) \prod_{t=0}^{T-1} \Pr(S_{t+1} \mid S_t, H_t) = \Pr(S_0) \prod_{t=0}^{T-1} (1 - c_t). \tag{4}$$

No unconditional independence assumption is required for this factorization: each term is already conditional on survival and history. However, assuming that the $c_t$ are equal, stable across tasks, or predictable from a benchmark score are additional hypotheses.

Suppose a model reports or is assigned a benchmark accuracy $a_t$ and define $q_t = 1 - a_t$. The substitution $c_t = a_t$ yields $E_{t+1} = q_t E_t$. It can be useful as a simple proxy, but it is neither a mathematical identity nor a safe default. It must be tested by measuring whether benchmark accuracy predicts conditional correction under the target review protocol.

### 3.3 Stochastic transitions

When $E_t$ is random, the more defensible formulation is often conditional expectation:

$$\mathbb{E}[E_{t+1} \mid \mathcal{F}_t] = (1 - c_t) E_t, \tag{5}$$

where $\mathcal{F}_t$ contains the information available at round $t$. Equation (5) should not be replaced automatically by an unconditional product of expected retention factors when $c_t$ is random or history-dependent. A valid analysis either preserves the conditioning set or models the joint process explicitly.

### 3.4 Nested and competing specifications

The meaningful comparison is not simply addition versus multiplication. A general additive-increment transition is

$$E_{t+1} = E_t + g_t(E_t). \tag{6}$$

The proportional model is nested within it when $g_t(E_t) = -c_t E_t$. Summing the increments gives

$$E_T = E_0 - \sum_{t=0}^{T-1} c_t E_t, \tag{7}$$

which is the same trajectory represented by equation (2). By contrast, constant absolute correction,

$$E_{t+1} = \max\{0, E_t - d\}, \tag{8}$$

imposes a genuinely different restriction: each review removes a fixed amount rather than a fraction of what remains.

### 3.5 Correction plus introduction

Pure decay assumes that review can only remove error. That is often too strong. For operational decisions, the primary state should be an unbounded cost or error-mass process $E_t \geq 0$: the weighted cost of unresolved mistakes after review round $t$. The weighting can encode monetary loss, safety impact, remediation effort, downstream exposure, or another decision-relevant loss function.



© 2026 Nicolò Ciccarone / Complexa Labs. All rights reserved.

Let $c_t \in [0,1]$ be the fraction of existing error cost corrected during round $t$, and let $r_t \geq 0$ be the endogenous amplification rate—the new downstream error cost induced per unit of residual error. The resulting transition is

$$E_{t+1} = (1 - c_t)E_t + r_t E_t = \underbrace{(1 - c_t + r_t)}_{G_t} E_t. \tag{9}$$

This formulation separates two mechanisms that a binary correctness probability obscures. The term $r_t E_t$ represents endogenous propagation: unresolved errors cause further errors or make their repair more costly. It is deliberately multiplicative: the existing error mass is the source of the next round's newly induced error cost.

Define the net multiplier $G_t = 1 - c_t + r_t$. Iterating gives

$$E_T = E_0 \prod_{t=0}^{T-1} G_t. \tag{10}$$

The relevant stability measure for the homogeneous multiplicative process is its long-run log-growth rate,

$$\gamma = \lim_{T\to\infty} \frac{1}{T} \log \frac{E_T}{E_0} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log G_t. \tag{11}$$

If $\{\log G_t\}$ is stationary, ergodic, and integrable, Birkhoff's pointwise ergodic theorem yields $\gamma = \mathbb{E}[\log G_t]$ almost surely [5]. The regimes are:

- $\gamma < 0$: contractive error cost;
- $\gamma = 0$: critical error process;
- $\gamma > 0$: exponentially growing error cost.

For constant parameters, explosive growth occurs precisely when

$$1 - c + r > 1 \quad \Longleftrightarrow \quad r > c. \tag{12}$$

The interpretation is direct: error cost becomes explosive when each round propagates more cost than it corrects.

Heavy-tailed modelling is relevant to the initial cost $E_0$ and to the distribution of amplification events $r_t$. A process may be contractive on a typical path yet still have severe tail risk when rare, high-cost initial errors or unusually large multipliers dominate aggregate loss. This distinction is why the analysis should report tail-sensitive metrics in addition to average error cost.

### 3.6 White-noise perturbations

The multiplicative specification isolates endogenous propagation: new cost is proportional to the error cost that survived the preceding round. In practice, a review round can also receive an exogenous shock: a new retrieval failure, an ambiguous instruction, a changed source, or an unrelated model failure. A convenient linear perturbation is

$$\widetilde{E}_{t+1} = G_t \widetilde{E}_t + \sigma_t \varepsilon_{t+1}, \qquad \mathbb{E}[\varepsilon_{t+1} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\varepsilon_{t+1}^2 \mid \mathcal{F}_t] = 1, \tag{13}$$

where $\{\varepsilon_t\}$ is white noise and $\sigma_t \geq 0$ is a predictable shock scale. The tilde is intentional: a mean-zero shock is appropriate for a signed measurement error or a local linear approximation, but it may produce negative values and therefore cannot by itself represent a literal nonnegative error mass.



© 2026 Nicolò Ciccarone / Complexa Labs. All rights reserved.

For a literal cost mass, the corresponding nonnegative formulation is

$$E_{t+1} = \left[G_t E_t + \sigma_t \varepsilon_{t+1}\right]_+, \qquad [x]_+ = \max\{x, 0\}. \tag{14}$$

Alternatively, genuinely new errors can be modeled as a nonnegative innovation $U_{t+1}$ in $E_{t+1} = G_t E_t + U_{t+1}$. This separates endogenous amplification ($G_t E_t$) from error cost arriving independently of the residual stock ($U_{t+1}$).

White noise changes the interpretation of stability. It does not alter the multiplicative threshold that governs the inherited component: under the usual integrability conditions, $\mathbb{E}[\log G_t] < 0$ makes the unprojected stochastic recurrence stable in the long-run distributional sense, while $\mathbb{E}[\log G_t] > 0$ permits the inherited cost to grow exponentially. Even in the stable regime, persistent shocks imply that error cost need not converge to zero path by path; instead, the process fluctuates around a stationary distribution. If $\varepsilon_t$ or $\sigma_t$ has heavy tails, occasional large excursions can dominate operational risk despite negative typical log growth.

### 3.7 Multivariate error dynamics

Many workflows contain interacting error classes rather than a single homogeneous stock. Let $\mathbf{E}_t \in \mathbb{R}_+^d$ collect error costs across $d$ categories, such as factual claims, calculations, citations, safety constraints, or downstream system components. Let $\mathbf{C}_t = \operatorname{diag}(c_{1t}, \ldots, c_{dt})$ be a diagonal correction matrix, and let $\mathbf{R}_t \in \mathbb{R}_+^{d \times d}$ be a nonnegative propagation matrix. The entry $(\mathbf{R}_t)_{ij}$ is the new error cost in category $i$ induced by one unit of residual cost in category $j$.

The multivariate transition is

$$\mathbf{E}_{t+1} = (\mathbf{I} - \mathbf{C}_t + \mathbf{R}_t)\, \mathbf{E}_t = \mathbf{G}_t \mathbf{E}_t. \tag{15}$$

Iterating preserves the order of review rounds:

$$\mathbf{E}_T = \mathbf{G}_{T-1} \mathbf{G}_{T-2} \cdots \mathbf{G}_0 \mathbf{E}_0. \tag{16}$$

Off-diagonal entries of $\mathbf{R}_t$ represent cross-category propagation. For example, an uncorrected factual error may induce errors in a recommendation, a calculation, or a cited conclusion. This captures systemic risk that a scalar aggregate can conceal.

For a time-invariant matrix $\mathbf{G}$, the decisive criterion is its spectral radius,

$$\rho(\mathbf{G}) = \max_i |\lambda_i(\mathbf{G})|, \tag{17}$$

the largest absolute value among the eigenvalues of $\mathbf{G}$. If $\mathbf{G}$ is diagonalizable as $\mathbf{G} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, then $\mathbf{G}^T = \mathbf{V}\mathbf{\Lambda}^T\mathbf{V}^{-1}$: each eigen-direction is multiplied by $\lambda_i^T$. Error cost therefore contracts when $\rho(\mathbf{G}) < 1$, is critical when $\rho(\mathbf{G}) = 1$, and grows exponentially along at least one direction when $\rho(\mathbf{G}) > 1$.

In this application $\mathbf{G} = \mathbf{I} - \mathbf{C} + \mathbf{R}$ is nonnegative. Perron–Frobenius theory then gives a useful operational interpretation: $\rho(\mathbf{G})$ is a real, nonnegative dominant eigenvalue with a nonnegative eigenvector. That eigenvector identifies the mix of error categories associated with the system's dominant growth or decay mode. A caveat is that $\rho(\mathbf{G}) < 1$ guarantees asymptotic contraction, but a non-normal matrix can still show temporary growth before decaying.



© 2026 Nicolò Ciccarone / Complexa Labs. All rights reserved.

For stationary ergodic random matrices satisfying $\mathbb{E}[\log^+\|\mathbf{G}_t\|] < \infty$, the corresponding stability statistic is the top Lyapunov exponent,

$$\gamma_1 = \lim_{T\to\infty} \frac{1}{T} \log \|\mathbf{G}_{T-1}\mathbf{G}_{T-2}\cdots\mathbf{G}_0\|\,, \tag{18}$$

which exists almost surely under standard conditions for products of random matrices [6]. The signs of $\gamma_1$ give the multivariate analogues of contraction, criticality, and explosive growth.

# 4 Discussion

## 4.1 Hypotheses

The framework generates testable hypotheses rather than claims of universal behavior.

- **Benchmark-transfer hypothesis.** Benchmark accuracy predicts conditional correction only after controlling for task type, error type, reviewer instructions, and available context.
- **Proportional-correction hypothesis.** A proportional-retention model predicts held-out error trajectories better than a constant-decrement model.
- **Survivor-hardening hypothesis.** Errors that survive more reviews have a lower subsequent correction probability because they are more subtle, correlated, or context-dependent.
- **Reviewer-dependence hypothesis.** Mixed-model chains can differ from repeated-model chains because shared training and prompting may reinforce the same blind spots.
- **Error-amplification hypothesis.** The induced-error rate $r_t$ is nonzero and varies with revision scope, task difficulty, and source availability.

## 4.2 Empirical strategy

The recommended experiment uses tasks with verifiable ground truth and preserves every round of the conversation. It should track deliberately seeded errors where this is appropriate, along with naturally occurring errors identified by an independent rubric. For every review round, record residual error cost, cost corrected, induced downstream error cost, agent identity, review instructions, context visibility, task difficulty, and final-answer correctness.

Candidate models should be fit on one set of tasks and compared on held-out tasks and held-out conversation chains. Predictive likelihood or another proper scoring rule should evaluate the estimated cost dynamics. Hierarchical effects for task and conversation are required when multiple rounds or multiple errors come from the same underlying task; otherwise, uncertainty will be understated. Results should report uncertainty intervals, calibration of estimated correction and amplification rates, and subgroup outcomes by domain, error taxonomy, and context condition.

## 4.3 Deployment implications

The model supports a practical decision loop. First, measure transition behavior on the organization's actual task distribution. Second, select reviewer combinations for demonstrated complementarity, not only for high standalone performance. Third, stop or escalate when estimated marginal correction is no longer greater than estimated error introduction. Finally,



© 2026 Nicolò Ciccarone / Complexa Labs. All rights reserved.

maintain trace-level evaluation data so that a changed model, prompt, retrieval policy, or tool configuration can be assessed as a change in transition dynamics rather than through anecdotes.

## 5 Conclusions

Multi-agent review is not a guarantee of reliability; it is a process whose transitions can be measured. Multiplicative retention provides a principled model when each review corrects a fraction of the residual error. A general additive-increment formulation contains that model as a state-dependent special case, while constant-decrement and correction-plus-introduction specifications impose distinct and testable restrictions.

The key operational distinction is between standalone benchmark accuracy and conditional correction probability. The former may be an informative predictor; it is not a substitute for measuring the latter. Complexa Labs therefore recommends direct transition measurement, explicit modeling of introduced errors, and held-out validation of review chains before multi-agent workflows are treated as reliability controls.

## References

[1] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B*, 41(1):1–31, 1979. https://doi.org/10.1111/j.2517-6161.1979.tb01052.x.

[2] A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. https://doi.org/10.1080/01621459.1982.10477856.

[3] K. Tian, E. Mitchell, H. Yao, et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv:2305.14975, 2023. https://arxiv.org/abs/2305.14975.

[4] J. Nair, A. Wierman, and B. Zwart. *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*. Cambridge University Press, 2022. https://doi.org/10.1017/9781009053730.

[5] G. D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931. https://doi.org/10.1073/pnas.17.2.656.

[6] H. Furstenberg and H. Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31:457–469, 1960. https://doi.org/10.1214/aoms/1177705909.



© 2026 Nicolò Ciccarone / Complexa Labs. All rights reserved.