THE IRREVERSIBILITY THRESHOLD: WHEN ERROR IN LLM AGENT PIPELINES OUTRUNS HUMAN CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents are increasingly composed into *pipelines*, in which each agent conditions on the output of the one before it. We show that this decomposition pattern carries a structural failure mode: as a multi-agent pipeline deepens, semantic error does not merely accumulate, it *compounds*. We develop a formal model of error propagation in linear LLM agent pipelines and give the accumulated error a closed form, the *cascade equation*, from which we derive a geometric lower bound with *cascade ratio* $r > 1$. Two thresholds follow in closed form: a *critical depth* m^*, the deepest pipeline that stays within an error tolerance θ without intervention, and an *irreversibility threshold* $m^\dagger$, beyond which no bounded human validator, modelled as a gate $h = \alpha\beta$ that combines domain expertise and provenance traceability with achievable capacity $\bar{h} < 1$, can restore the output to within θ. These partition pipeline depth into safe, recoverable, and irreversible regimes. The model reframes reliability, rather than per-agent capability, as the binding constraint on deep agentic workflows, offers a candidate, testable structural account of why many multi-agent generative-AI deployments fail to deliver value, and proposes $m^\dagger$ as a candidate pre-deployment audit trigger for human-oversight requirements such as the EU AI Act. Empirically, two exploratory pilots and two pre-registered, replicated mechanism experiments map the regimes. Relaying facts under a behavioural token budget degrades output by *erasure* with negligible alteration in all three models tested. Processing steps inject distortion net of a gold-state floor, whereas the same models' relay distortion is indistinguishable from zero in one model and one to two orders of magnitude smaller in the other. A pre-registered coupling assumption, that compression injects distortion, is not supported in relay; in processing it is supported in one of two models, whose fitted parameters place the worst-case handoff at an interior compression, and not in the other. We report the failed calibration gates and an instrument-precision sensitivity that lowers the processing estimates without changing their ordering, except for one model at the value-only extreme of the measured precision interval. Following emerging usage, we call this failure mode a *hallucination cascade*, and give it, to our knowledge, its first closed-form account.

1 INTRODUCTION

An LLM extracts structured fields from a document; a second agent reasons over those fields; a third drafts a recommendation; a fourth formats it for delivery. *Pipelines* of this kind, linear chains in which each agent receives as input the output of the one before it (Figure 3, Appendix G), are how LLM agents are increasingly deployed, because they decompose a hard task into narrower subtasks [25, 26, 27]. They are also, we argue, the source of a systematic, underappreciated failure mode: semantic error does not merely accumulate as a pipeline deepens, it *compounds*. Every LLM agent introduces some intrinsic distortion relative to the ideal output of its subtask [2, 16, 17]; an agent conditioning on a distorted input cannot separate inherited error from signal, so its own distortion amplifies what it received. Accumulated error therefore scales geometrically with the number of handoffs, which we call a *hallucination cascade* [7, 14] and formalise as the *cascade*

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.

equation. Related compounding dynamics have been characterised for multi-agent reasoning [5] and sequential estimation [4], but no closed-form account exists for a linear chain of LLM agents.

The stakes are economic. Many enterprise generative-AI deployments are pipeline-shaped, and many fail to deliver positive return on investment [44]; users name unreliability as their most common concern [42]. Practitioners' explanations treat the problem as one of local agent quality; our model suggests a structural one that local quality cannot fix: beyond some depth accumulated error exceeds any useful tolerance, and beyond a further depth what even an expert reviewer can correct.

Contributions. We develop a formal model of semantic error propagation in linear LLM agent pipelines and derive three results.

C1. An error-quantification model. We define accumulated semantic error recursively via the cascade equation (Definition 3.6), unroll it into closed form (Equation (3)), and ground it in the $\mathcal{O}(n^2 d)$ per-step capacity bound of Sikka and Sikka [2]. Under stated assumptions this yields a geometric lower bound on error growth with *cascade ratio* $r = 1 + \delta_{\min}\kappa_{\min} > 1$ (Theorem 3.7). We posit a compression–injection coupling (Assumption A9) predicting a *worst-case handoff compression* $\kappa^* = \frac{1}{2} + \delta_0/2\lambda$ (Proposition 3.9), a pre-registered hypothesis our experiments test: unsupported in relay, supported in one of two processing models (Section 4).

C2. A critical depth m^*. We derive the largest depth at which the model's lower bound on accumulated error remains within a tolerance θ without intervention (Corollary 3.8), an upper estimate of the safe depth since the bound discards every injection term.

C3. A human validator with an irreversibility threshold $m^\dagger$. We formalise a human reviewer as a bounded corrective gate $h_k = \alpha_k \beta_k$ combining domain expertise and error traceability (Definition 3.10), posit on capacity grounds that it is bounded strictly below one (Assumption A8), and derive in closed form the depth $m^\dagger$ beyond which no validator within that bound can restore the output below θ (Corollary 3.15).

Together these define three regimes: safe ($[1, m^*]$), recoverable ($(m^*, m^\dagger]$), irreversible ($(m^\dagger, \infty)$). Section 4 tests the mechanisms with a pre-registered, replicated programme; Section 5 draws the implications for design, review, ROI and oversight regulation [40].

2 Related Work

Why single LLMs err is well studied [18, 19]: language models must hallucinate on rarely seen facts [16], and training rewards confident guessing over abstention [17]. Most relevant to us, Sikka and Sikka [2] argue from the $\mathcal{O}(n^2 d)$ cost of dense self-attention [3] and the Time Hierarchy Theorem [1] that beyond a task-complexity threshold an LLM cannot reliably solve *or verify* a task within its per-step budget; we adopt this bound as the source of each agent's distortion and of the bound on a validator's traceability. Empirical work studies multi-agent LLM systems directly [30, 28, 29]; Tran and Kiela [5] find that single agents match or outperform multi-agent chains at matched compute and ground the result in the Data Processing Inequality [34]; taxonomies catalogue *how* such systems break [23, 24, 14], whereas we model *when*, as a function of depth, failure becomes unavoidable. Formal antecedents are compounding budget and precision errors in sequential estimation [4], cascading errors in annotation pipelines [31] and exposure bias in autoregressive models [32, 33]. Two contemporaneous papers study cascades in LLM agent systems: Jamshidi et al. [7] find net *attenuation* across refinement chains, in which every agent sees the full upstream context ($\kappa \approx 1$), the opposite corner from the lossy handoffs we model; Xie et al. [8] model cascades as binary contagion, whereas we track a continuous error magnitude. Other treatments are empirical [9, 13, 10, 11, 15, 12]. We adopt the verification/validation distinction of software engineering [35, 36] to place our human gate; the scalable-oversight programme [37, 38] is orthogonal to our claim. Appendix E expands this section.

3 Formal Model of Semantic Error Propagation

The model rests on the per-step computational capacity of a transformer agent: following Sikka and Sikka [2], dense self-attention over n tokens of dimension d costs $\mathcal{O}(n^2 d)$ per step [3], and beyond a task-complexity threshold this forces a non-zero intrinsic distortion. The propagation dynamics

below hold under *any* per-step capacity bound; its form affects only the numerical values of m^* and $m^\dagger$. Notation is in Appendix A (an overbar is a maximum achievable value, a hat an estimator); qualifying remarks and replies to anticipated objections are collected in Appendix F.

3.1 Linear Agentic Pipeline

Definition 3.1 (Linear Agentic Pipeline)**.** A *linear agentic pipeline* is a tuple $\mathcal{P} = (A_1, \ldots, A_m, \Phi)$, where each agent $A_k = (M_k, \mathcal{T}_k, n_k)$ is an LLM M_k of embedding dimension d_k, an assigned task $\mathcal{T}_k$ and a prompt of n_k tokens, and $\Phi = \{\phi_k\}_{k=1}^{m-1}$ collects the handoffs $\phi_k : O_k \to I_{k+1}$; the dependency graph is a simple path $A_1 \to \cdots \to A_m$ (Figure 3, Appendix G).

3.2 Computational Capacity and Intrinsic Node Error

Definition 3.2 (Per-Step Computational Capacity)**.** Following Sikka and Sikka [2], the per-step capacity of agent A_k is fixed, as a convention, at $\Omega_k = c \cdot n_k^2 \cdot d_k$ FLOPs per generation step, with c an architecture-dependent constant declared once per deployment (Assumption A5); claims that depend on c are read ordinally.

Definition 3.3 (Intrinsic Semantic Distortion)**.** The *intrinsic semantic distortion* of agent A_k is $\delta_k = 1 - \mathrm{sim}(O_k^*, O_k) \in [0, 1]$, with O_k^* the correct output, O_k the actual one and sim an abstract similarity; $\hat{\delta}_k$ is its estimator under a measurable proxy (Assumption A4).

Assumption A7 (Complexity–Error Relationship). From the Time Hierarchy Theorem [1] as applied by Sikka and Sikka [2], if $C(\mathcal{T}_k) > \Omega_k$ then A_k may not execute $\mathcal{T}_k$ correctly within the per-step bound; we posit, rather than derive, $C(\mathcal{T}_k) > \Omega_k \Rightarrow \delta_k > 0$, whose contrapositive makes $\Omega_k \geq C(\mathcal{T}_k)$ necessary, not sufficient, for $\delta_k = 0$.

Conjecture 3.4 (Complexity–Distortion Monotonicity)**.** Fix a model M and a task family $\mathcal{F}$, and let $\rho(\mathcal{T}) = \max\{0, C(\mathcal{T}) - \Omega\}/\Omega$ be the complexity gap ratio. Then $\rho \mapsto \mathbb{E}[\delta(\mathcal{T}) \mid \rho(\mathcal{T}) = \rho]$ is non-decreasing, with $\mathbb{E}[\delta \mid \rho = 0] \geq \delta_{\mathrm{floor}}(M) \geq 0$ and $\lim_{\rho \to \infty} \mathbb{E}[\delta \mid \rho] = 1$ (tested ordinally in Section 4).

3.3 Compression and Accumulated Error

Definition 3.5 (Compression Coefficient)**.** Let $\kappa_k = |I_{k+1}|/|O_k| \in [0, 1]$ denote the *compression coefficient* of handoff ϕ_k, where $|\cdot|$ is token count and $|I_{k+1}|$ counts only the tokens of I_{k+1} derived from O_k. $\kappa_k = 1$ is full pass-through. κ_k measures information retention, not the transmission of semantic error as such; by convention our estimators attribute *deletion* of content to the channel (κ) and reserve $\hat{\delta}$ for *alteration* (Appendix F).

Definition 3.6 (Pipeline Error (The Cascade Equation))**.** The accumulated semantic error E_k is defined recursively by the *cascade equation*:

$$E_1 = \delta_1 \tag{1}$$

$$E_k = E_{k-1} \cdot \big(1 + \delta_k \cdot \kappa_{k-1}\big) + \delta_k, \qquad 2 \leq k \leq m \tag{2}$$

Inherited error enters node k through the *incoming* handoff ϕ_{k-1}, so amplification is scaled by κ_{k-1}. The form is a modelling choice fixed by three requirements: every node injects its own distortion ($E_{k-1} = 0 \Rightarrow E_k = \delta_k$); inherited error is amplified, because an agent conditioning on a distorted input cannot separate inherited error from signal; and purely additive propagation would make each node's contribution independent of the error it receives. The scalar multiplier is the first-order expansion of any smooth propagation operator meeting the first two requirements [4]; the form contest of Section 4 discriminates among named rivals (additive, saturating, survival) where power exists. Unrolling,

$$E_m \;=\; \sum_{j=1}^{m} \delta_j \prod_{k=j+1}^{m} \big(1 + \delta_k \cdot \kappa_{k-1}\big). \tag{3}$$

E_m is an unbounded error index ($E_m \in [0, \infty)$): only its comparison to θ and ordinal comparisons carry interpretation.

3.4 GEOMETRIC ERROR GROWTH

Theorem 3.7 (Geometric Error Growth). *For $m \geq 2$, let $\delta_{\min} = \min_{2 \leq k \leq m} \delta_k > 0$, a uniform lower bound over the nodes that receive a handoff (δ_1 enters only through the base case), and $\kappa_{\min} = \min_{1 \leq k \leq m-1} \kappa_k > 0$, and define the* cascade ratio *$r := 1 + \delta_{\min} \cdot \kappa_{\min} > 1$. Under Assumptions A1–A2, which enter through Definition 3.6, for any linear agentic pipeline of depth m without human intervention,*

$$E_m \;\geq\; \delta_1 \cdot r^{\,m-1}. \tag{4}$$

Proof (Appendix B): induction on m, dropping the injection term $\delta_m \geq 0$ and bounding the factor by $\delta_m \geq \delta_{\min}$, $\kappa_{m-1} \geq \kappa_{\min}$. Since δ_1 is unconstrained by $\delta_{\min}$, $\delta_{\min} > \delta_1$ is admissible (Figure 1).

Corollary 3.8 (Critical Pipeline Depth). *For any tolerance $\tau > \delta_1$, define the* critical-depth function

$$m^*(\tau) \;=\; 1 + \left\lfloor \frac{\log(\tau/\delta_1)}{\log r} \right\rfloor. \tag{5}$$

The critical depth *$m^* := m^*(\theta)$ is the largest depth with $E_m \leq \theta$ under the bound of Theorem 3.7; $m^*(\tau)$ is non-decreasing in τ, and since it is derived from a lower bound, m^* is an upper estimate of the safe depth.*

3.5 THE COMPRESSION–INJECTION COUPLING AND THE WORST-CASE HANDOFF

The cascade equation treats δ_k and κ_{k-1} as free parameters, so that amplification is maximised by full pass-through. A compressive handoff, however, forces the receiving agent to reconstruct content from a reduced input, widening the capacity gap that A7 links to distortion; we therefore posit, and test, a coupling.

Assumption A9 (Compression–Injection Coupling). For $k \geq 2$, the intrinsic distortion of agent A_k increases affinely with the compression of its *incoming* handoff:

$$\delta_k \;=\; \delta_0 + \lambda\,(1 - \kappa_{k-1}), \qquad \delta_1 = \delta_0, \tag{6}$$

where $\delta_0 \geq 0$ is the model's basal distortion on intact input and $\lambda \in \mathbb{R}$ the compression penalty, posited $\lambda > 0$ (the pre-registered hypothesis) and admissible whenever $0 \leq \delta_0 + \lambda(1-\kappa) \leq 1$ on the compressions in use. The affine form is the first-order expansion of any smooth $\delta(\kappa)$ around $\kappa = 1$; since δ_k depends on κ_{k-1}, not on E_{k-1}, A9 is consistent with A1 and leaves Theorem 3.7 and its corollaries unchanged: with uniform compression κ, for any admissible λ with $\delta_0 + \lambda(1-\kappa) > 0$, they apply with $\delta_{\min} = \delta_0 + \lambda(1-\kappa)$, so that $r = 1 + p(\kappa)$.

Proposition 3.9 (Worst-Case Handoff Compression). *Under Equation (2) and Assumption A9 with $\lambda > 0$, the per-step compounded loss at a handoff with compression $\kappa \in [0,1]$, $p(\kappa) := \kappa\,\delta(\kappa) = (\delta_0 + \lambda)\,\kappa - \lambda\,\kappa^2$, is strictly concave with unconstrained maximiser*

$$\kappa^* \;=\; \frac{1}{2} + \frac{\delta_0}{2\lambda}, \qquad p(\kappa^*) \;=\; \frac{(\delta_0 + \lambda)^2}{4\lambda}, \tag{7}$$

so the maximiser over $[0,1]$ is $\min\{1, \kappa^*\}$, interior ($\kappa^* < 1$) iff $\delta_0 < \lambda$; when $\delta_0 \geq \lambda$ the worst case sits at pass-through, recovering the uncoupled model as a limiting case. For $\lambda \leq 0$, p is maximised on $[0,1]$ at $\kappa = 1$. At experimentally accessible depths the observable signature of A9 is the concavity of $p(\kappa)$ rather than an interior maximum of E_m itself (Appendix F).

Empirical status. Section 4 finds no coupling in relay (three models), a positive and robust coupling in one processing model, for which $\hat{\delta}_0 < \hat{\lambda}$ places κ^* in the interior, and none in the other; the uncoupled model ($\lambda = 0$) and the deletion-only reading ($\lambda \leq 0$) are A9's limiting cases (Appendix F).

3.6 THE HUMAN VALIDATOR GATE

Definition 3.10 (Verification Capacity). A human validator V_k inserted at node k is assigned a *verification capacity* $h_k = \alpha_k \cdot \beta_k \in [0,1]$, where $\alpha_k \in [0,1]$ denotes domain knowledge (expertise relative to task $\mathcal{T}_k$) and $\beta_k \in [0,1]$ denotes traceability (the ability to reconstruct the provenance of I_k through prior pipeline steps). The product encodes joint necessity (expertise without provenance

cannot tell whether an input is distorted; provenance without expertise cannot evaluate correctness) and is the most conservative aggregation (Appendix J). h_k is a *verification* capacity, the ability to bring O_k toward a reference the model takes as given; we keep the term *validator* for continuity with the oversight literature.

Definition 3.11 (Pipeline Error with Human Validator)**.** When validator V_k is inserted at node k, the residual pipeline error is $E_k^H = E_k \cdot (1 - h_k)$. A reviewer-introduced error term would only move every threshold below earlier (Appendix F).

Proposition 3.12 (Conditions for Error Suppression)**.** *Within the model, for $\delta_1 > 0$ (so that $E_k \geq E_1 = \delta_1 > 0$), $E_k^H = 0$ if and only if $h_k = 1$, which requires $\alpha_k = 1$ and $\beta_k = 1$ simultaneously. Since $h_k < 1$ under any realistic instantiation, human intervention reduces but does not eliminate accumulated pipeline error.*

Assumption A8 (Traceability Limitation on Validator Capacity). For tasks where $C(\mathcal{T}_k) > \Omega_k$, we posit that $\beta_k < 1$, hence $\bar{h} < 1$. The bound is on *access*, not computation: tools, traces and time raise β_k only up to the provenance that survives the chain and cannot restore what was deleted upstream (the erasure of Section 4); secondarily, the capacity argument of [2] applies to verifying as well as solving.

3.7 The Irreversibility Threshold

Definition 3.13 (Maximum Achievable Verification Capacity)**.** For a given domain and pipeline, $\bar{h} = \bar{\alpha} \cdot \bar{\beta} \in (0, 1)$, with $\bar{\alpha}$ the best available expertise and $\bar{\beta}$ the maximum achievable traceability; $\bar{h} < 1$ by A8, and its treatment as a constant independent of k and m is A6.

Definition 3.14 (Human-Correctable Condition)**.** Given $\theta > 0$ and $\bar{h} \in (0, 1)$, depth m is *human-correctable* iff $E_m(1 - \bar{h}) \leq \theta$, equivalently iff $E_m \leq \theta/(1 - \bar{h})$.

Corollary 3.15 (Irreversibility Threshold)**.** *Under Assumptions A1–A2 and A6 (with $\bar{h} < 1$ motivated by A8), let $\theta > \delta_1$, $\delta_{\min} > 0$, $\kappa_{\min} > 0$, and $\bar{h} \in (0, 1)$. The* irreversibility threshold *is the critical depth at the effective tolerance $\theta/(1 - \bar{h})$:*

$$m^\dagger \coloneqq m^*\left(\frac{\theta}{1-\bar{h}}\right) = 1 + \left\lfloor \frac{\log\left(\dfrac{\theta}{\delta_1\,(1-\bar{h})}\right)}{\log r} \right\rfloor \tag{8}$$

For any pipeline of depth $m > m^\dagger$, no validator with capacity bounded by $\bar{h}$ satisfies $E_m^H \leq \theta$ under the model. The converse is not asserted: the threshold is computed from a lower bound on E_m, so $m \leq m^\dagger$ does not by itself guarantee correctability. Since $m^*(\tau)$ is non-decreasing and $\theta/(1 - \bar{h}) \geq \theta$, $m^\dagger \geq m^*$, with equality possible.

Proof (Appendix D): substitute the bound of Theorem 3.7 into Definition 3.14 and solve for m; as the bound is over a lower bound on E_m, actual pipelines may become irreversible earlier.

Three regimes follow: a *safe zone* $[1, m^*]$ in which the bound stays within θ; a *recoverable zone* $(m^*, m^\dagger]$ in which $E_m > \theta$ and correction is not yet excluded by the bound, feasible only for a validator with $h_k \geq 1 - \theta/E_m$; and an *irreversible zone* $(m^\dagger, \infty)$ in which correction to below θ is not feasible under $\bar{h}$ (Table 3, Appendix F).

The inputs to Equation (8) are estimable before deployment; Appendix J propagates illustrative input uncertainties to $m^\dagger$ (point estimate 46, 90% interval $[37, 64]$). The model is a model of *verification*: O^*, sim and θ are *validation inputs* it presupposes rather than derives [35, 36], so the experiments do not estimate α, β or $\bar{h}$, whose values in the text are illustrative. A saturating alternative to Equation (2) admits a finite $m^\dagger$ only if $\theta < 1 - \bar{h}$, a discriminating prediction between the two families (Appendix F).

4 Experiments

We test the model with two exploratory pilots and two pre-registered mechanism experiments (Table 4; full tables in Appendix H). The pilots ran before the protocol below existed and no headline

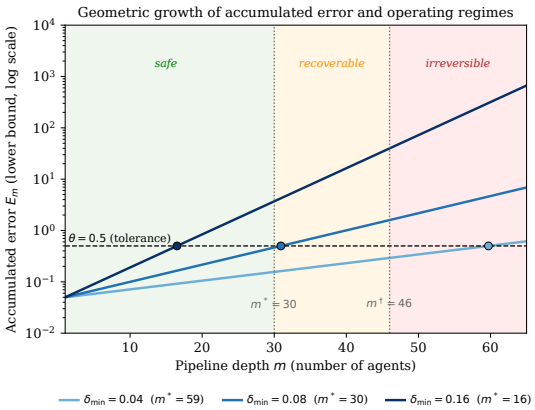

Figure 1: Geometric lower bound on accumulated error E_m (log scale) versus pipeline depth m, for $\delta_{\min} \in \{0.04, 0.08, 0.16\}$ ($\delta_1 = 0.05$, $\kappa = 1$, $\theta = 0.5$; $\delta_{\min}$ is taken over the nodes $k \geq 2$, so δ_1 is unconstrained). Markers give the critical depth m^* (Corollary 3.8). Shaded regions illustrate the three operating regimes for $\delta_{\min} = 0.08$ and a maximum validator capacity $\bar{h} = 0.7$: safe ($[1, m^*]$, $m^* = 30$), recoverable ($(m^*, m^\dagger]$, $m^\dagger = 46$), and irreversible ($(m^\dagger, \infty)$, Corollary 3.15). Curves grow without bound; the axis is truncated.

claim rests on them. For Exp M (relay) and Exp B (processing), hypotheses, decision rules, instrument definitions and calibration gates were fixed in version-controlled configurations and tagged before any paid call; every later change is a dated amendment (A1–A5, Appendix I), one of which (A5) is post-hoc and reported only as a sensitivity. Every cell was run in three independent replicates; intervals are cluster-bootstrap 95% intervals (by packet in Exp M, by chain in Exp B); every call is logged; every number below is re-derived from the traces by one command, and the headline estimates are recomputed by an independent implementation.

Compression as a behavioural budget. Each handoff carries a per-step token budget C_k that the agent must meet itself: the API ceiling sits at $3C_k$ and never binds for a compliant reply; a reply over budget is returned once for a rewrite, and one still over budget is trimmed at whole-line boundaries and flagged *forced*. Prompts are neutral about how to comply, so whether a model omits, alters or overruns is its own behaviour. Budgets are fractions κ of each model's measured *lossless terse size*, so $\kappa < 1$ forces a loss of content; $\hat{\kappa}$ is always the realised ratio.

Pilots. Two exploratory pilots (pass-through QA relay, up to 16 steps) exposed an instrument failure and a manipulation failure, which fixed the programme's design principle: a cascade is observable only when the payload exceeds the channel and steps *process* rather than relay. Their one retained measurement is methodological: a third-provider judge gives inter-judge agreement of Cohen's $\kappa = 0.90$ on both pilots (Table 14).

4.1 Exp M: relay under a token budget

Exp M relays packets of $N \in \{10, 20\}$ facts about *fictitious* entities (closing parametric knowledge by construction) through chains of up to 12 steps under budgets $\kappa \in \{0.9, 0.75, 0.5, 0.35, 0.2\}$ of the previous state, with an uncapped control, for claude-haiku-4.5, gpt-5.4-mini and the open-weight qwen3-32b (240 relays per cell, 4,320 in total). A deterministic scorer classifies each fact as recalled, altered (stated with a wrong value) or omitted; inverting Equation (2) on the alteration channel gives per-step $\hat{\delta}$, and the affine fit $\hat{\delta}(\hat{\kappa})$ gives $\hat{\delta}_0$ and $\hat{\lambda}$ (A9), reported as not identified when $\hat{\kappa}$ does not vary enough. Decision rules: H-M1, relay distortion is negligible ($\hat{\delta}_0$ interval below 0.01); H-M2, A9 holds in relay ($\hat{\lambda} > 0$, interval excluding 0).

Results (Table 5; Figure 2a). Budgets bind: from uncapped final recall of 0.985, 1.000 and 0.982 (haiku, mini, qwen), recall falls to 0.576, 0.788 and 0.713 at $\kappa = 0.9$ and to between 0.137 and 0.211 in every cell at $\kappa \leq 0.75$. What is lost is lost by *erasure*: on the alteration chan-

nel $\hat{\delta}_0 = 0.0013$ $[0.0009, 0.0018]$, -0.0000 $[-0.0001, 0.0000]$ and 0.0005 $[0.0003, 0.0007]$, so H-M1 is supported for all three models, and since a blind precision sample confirmed only 1 of 8 Exp M "altered" labels (Section 4.3), even these are upper bounds. H-M2 is not supported: $\hat{\lambda} = -0.0063$ $[-0.0145, 0.0001]$ (haiku) and 0.0023 $[-0.0004, 0.0072]$ (mini) cover 0; qwen's 0.0013 $[0.0001, 0.0026]$ meets the rule formally but is not robust: restricting the fit to budget-compliant steps reverses its sign (-0.0006). In the form contest (BIC over named rivals on the pre-saturation window), the cascade form wins with the predicted convex curvature in three of four fittable cells for haiku and for mini and in two of four for qwen, whose $\kappa = 0.75$ cells are concave (survival-shaped; Table 7).

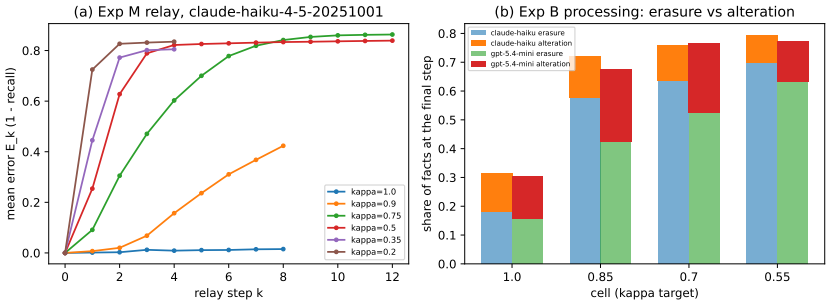

Figure 2: Empirical results. **(a)** Exp M relay (claude-haiku-4.5): mean error $1 - \text{recall}$ by relay step for each budget target κ; the uncapped control stays near zero while every capped cell loses content, faster the tighter the budget (cells at $\kappa \leq 0.5$ run to four steps and $\kappa = 0.9$ to eight by design). **(b)** Exp B processing, final step: share of recorded facts *absent* (solid) and *altered* (hatched) by cell, both models; tightening the budget raises the absent share while an altered share of the same order persists (Table 8).

4.2 Exp B: processing injects distortion

Exp B moves from relay to *processing*: 60 MuSiQue [6] hop chains (30 at $m = 2$, 30 at $m = 4$, the dataset's composability ceiling) in a consistent counterfactual world, every answer replaced by a plausible false value throughout the chain, so an answer from memory is detectably wrong. Each step answers a sub-question from its inherited notes plus one fresh paragraph and re-emits its notes under a budget of κ times the model's terse notes size ($\kappa \in \{0.85, 0.7, 0.55\}$, an alternating pattern, and an uncapped control), the agentic-RAG pattern [15]. A line-level probe classifies each recorded fact as correct, altered (value replaced, including by the real-world answer, or attached to another entity) or absent; a gold-state arm gives the per-step floor from which $\hat{\delta}_0^{\text{net}}$ is taken, a no-notes arm measures how much of each hop is answerable without the channel, and a 2-fact demand gives the Conjecture step. Decision rules: H-B1, $\hat{\delta}_0^{\text{net}} > 0$ (interval excluding 0) and above the same model's relay $\hat{\delta}_0$; H-B2, A9 holds in processing; H-B3, per-hop error under a 2-fact demand is at least that under a 1-fact demand (ordinal, Conjecture 3.4). Two pre-registered gates *failed* in the main run and are reported rather than re-run: the realised compression span of the capped cells fell below 0.25 (G2: 0.15 for haiku, 0.22 for mini), so $\hat{\lambda}$ in Exp B is identified over a narrower range than planned; and the probe's detection of a perturbed value was 77.3% against a gate of 80% (G3).

Results (Tables 8 and 12; Figure 2b). Processing injects distortion where relay does not: $\hat{\delta}_0 = 0.0817$ $[0.0650, 0.0988]$ (haiku) and 0.1144 $[0.0900, 0.1389]$ (mini); net of the gold-state floor, $\hat{\delta}_0^{\text{net}} = 0.0532$ $[0.0314, 0.0728]$ and 0.0670 $[0.0436, 0.0907]$, whereas the same models' relay $\hat{\delta}_0$ in Exp M was 0.0013 (haiku, one to two orders of magnitude smaller) and -0.0000 (mini, indistinguishable from zero). H-B1 is supported in the primary analysis, and the gross estimate is unchanged (haiku) or higher (mini, 0.1313) when the 11 and 24 chains guessable without the channel are excluded. On the coupling the two models part ways. For mini, $\hat{\lambda} = 0.1554$ $[0.0926, 0.2228]$: H-B2 is supported, and the slope is positive in every replicate ($0.158, 0.196, 0.113$) and under the compliant-only, guessable-excluded and parse-ok-only sensitivities (not under the superseded v1 probe). For haiku, $\hat{\lambda} = 0.0150$ $[-0.0146, 0.0433]$ covers 0: not supported. With mini's point estimates, $\hat{\delta}_0 < \hat{\lambda}$,

so Proposition 3.9 places that model's worst-case handoff at an interior compression rather than at pass-through (not at the lower end of $\hat{\lambda}$'s interval, where $\hat{\delta}_0 < \hat{\lambda}$ fails). H-B3 is supported ordinally in both models: per-hop error rises from 0.381 to 0.543 (haiku) and from 0.417 to 0.477 (mini) when the per-step demand goes from one fact to two. As Figure 2b shows, tightening the budget raises the share of facts *absent* at the final step in both models while an altered share of the same order persists: in processing, compression erases and, for one model, also distorts.

Post-hoc precision correction (amendment A5). The main-run audit showed most "altered" labels of the *entity* kind disagreeing with the second judge, so a random precision sample was added to the blind human review after the results were read: value-kind alterations were confirmed $8/8$, entity-kind $8/16$ (Wilson 95% $[0.28, 0.72]$). Re-weighting each entity-kind alteration as real with probability p is a sensitivity. At $p = 0.5$, $\hat{\delta}_0^{\text{net}} = 0.0266$ $[0.0096, 0.0420]$ (haiku) and 0.0431 $[0.0233, 0.0607]$ (mini); at the value-only extreme $p = 0$, mini keeps 0.0200 $[0.0075, 0.0308]$ and $\hat{\lambda} = 0.0751$ $[0.0394, 0.1152]$, whereas haiku's net distortion, -0.0009 $[-0.0127, 0.0090]$, is no longer distinguishable from its gold floor. At the central value the correction lowers $\hat{\delta}_0^{\text{net}}$ from 0.0532 to 0.0266 for haiku (a half) and from 0.0670 to 0.0431 for mini (a third), without changing mini's sign at any p; for haiku, H-B1 depends on the instrument's precision. Missed alterations push the other way and are not corrected.

4.3 Instrument validity

Both scorers are audited on the real outputs of every run (Table 13): known edits (deleting a fact, perturbing or swapping a value, dropping a tag, restoring the real-world value) must be detected, and a third-provider judge labels a stratified sample of 300 facts per experiment. Detection is 98.1–100% for the Exp M edit kinds and 77.3–100% for the Exp B kinds; scorer–judge agreement is Cohen's $\kappa = 0.59$ in both experiments. A single blind human reviewer adjudicated disagreements, siding with the scorer in $11/12$ (Exp M) and $7/12$ (Exp B) calibration cases (five probe errors fixed before the main run, amendment A4) and in $4/10$ and $2/10$ main-run cases, with the judge in $4/10$ and $6/10$; the precision sample above motivated A5.

4.4 Synthesis and limitations

Relay transports and processing transforms: relays lose content by erasure with negligible alteration (three models), while processing steps inject distortion net of their gold floor (two models) where relay distortion is indistinguishable from zero (mini) or one to two orders of magnitude smaller (haiku), so the depth that consumes the error budget is *processing* depth. The pre-registered coupling A9 receives mixed support: none in relay, a positive and robust coupling in one processing model and none in the other; the estimators' split of the two channels (erasure to κ, alteration to δ; Appendix F) accommodates both, and the interior worst-case handoff of Proposition 3.9 is now a live prediction where $\hat{\delta}_0 < \hat{\lambda}$. Taking $\hat{\delta}_0^{\text{net}}$ as an illustrative proxy for $\delta_{\min}$ (Figure 1 is drawn at $\kappa = 1$), the processing estimates (net 0.027–0.067 across the primary and central precision-corrected analyses) span the $\delta_{\min} = 0.04$ curve, for which $m^* = 59$ and $m^\dagger = 90$ at $\bar{h} = 0.7$: tens of processing steps, depths reachable by deployed long-horizon systems [43]. No experiment here estimates m^* or $m^\dagger$: the values quoted ($m^* = 30$, $m^\dagger = 46$; $m^* = 59$, $m^\dagger = 90$) come from illustrative configurations, and the pre-registered experiments test the model's mechanisms at depths $m \leq 12$ (relay) and $m \leq 4$ (processing), not the thresholds themselves. Limitations, stated plainly: the binding limitation is the instrument's precision (entity-kind alterations confirmed at about 50%, Exp M alteration labels at $1/8$, moderate judge agreement, one human reviewer labelling samples of 10–24), and H-B1 for haiku depends on it; models met the budget unaided in 24–71% of capped steps, and Exp B's realised compression range was narrower than planned (failed gate G2), so its coupling slopes rest on a short baseline; three relay and two processing models were tested, MuSiQue bounds processing depth at $m \leq 4$, the domain is multi-hop QA, and models ran at default temperature, with run-to-run variability measured by the replicates. Every experiment scores *verification* against a specification fixed by design (a model that states the true real-world answer scores as error), and none estimates a human gate: α, β and $\bar{h}$ are unmeasured throughout. Small models occupy the relay and processing positions by design (Appendix F).

5 DISCUSSION

The four implications below are conditional on A1–A8 and none on A9, whose empirical support is mixed; where one invokes the experiments, it reports an observation, not a derivation.

Implication 1: design pipelines with processing depth below m^*. For any tolerance θ there is a depth m^* beyond which accumulated error exceeds θ even without any single bad agent (Figure 1), computable at design time. The prescription is to keep sequential handoffs below m^*, consolidating subtasks into fewer, more capable agents, consistent with evidence that single agents match longer chains at matched compute [5]. The experiments sharpen it: *count processing steps, not agents*, since relay steps alter almost nothing while processing steps inject distortion (one to two orders of magnitude more in the one model whose relay distortion is distinguishable from zero). Nor, empirically, is compression protective: in relay its cost is erasure and lost provenance, and in processing one model distorted more the harder it was compressed.

Implication 2: human validators need high expertise α and traceability β to generate positive ROI. By Proposition 3.12, correction is effective only when *both* factors are high. A generalist reviewer with little visibility into intermediate steps therefore yields little error reduction at full cost, consistent with the automation-bias literature [39]; positive ROI requires reviewer expertise *and* traceability (provenance tooling, intermediate-output logging [45]). A low-α validator does not reduce error, it fails to *see* it: per-step improvements raise the *floor* of quality, the evaluator of record sets the *ceiling*. Erasure raises the ceiling further: detecting an *absence* requires α and β together [11].

Implication 3: a candidate, testable explanation for why many generative-AI implementations fail to deliver ROI. Beyond $m^\dagger$ a pipeline produces output that is neither directly usable nor economically correctable, and review labour is sunk cost. Organisations typically do not know they are in this regime: they add review effort and see no improvement, because the constraint is depth, not effort. This offers a falsifiable explanation for negative generative-AI ROI [44]: review effort beyond an estimated $m^\dagger$ should fail to convert into usable output.

Why a human gate? Nothing in the model bounds a machine verifier's h below $\bar{h}$, and on *verification* the objection is conceded: a machine verifier is another node inside the model [2]. The claim that survives concerns *validation*: outside the laboratory there is no oracle, O^*, sim and θ are constituted by human purposes, and under oversight regulation [40] responsibility for them is non-delegable; the threshold is the depth beyond which the evaluator of record loses access to the grounds for any verdict (Appendix F).

Implication 4: $m^\dagger$ as an auditable safety metric. Because $m^\dagger$ is a closed-form function of quantities estimable before deployment, a pipeline's depth can be compared against its computed $m^\dagger$ and the comparison logged and inspected by a third party (Appendix J propagates estimation error; $\bar{\beta}$ is calibratable by a seeded-error audit). Under frameworks such as the EU AI Act [40] and the international assessment of AI risks [41], high-risk systems must demonstrate adequate human oversight; when $m > m^\dagger$, mandated review cannot correct the output to within tolerance. Licensed inference runs one way: $m > m^\dagger$ certifies that review cannot restore tolerance under the stated bound, while $m \leq m^\dagger$ does *not* certify adequacy (the threshold rests on a lower bound for E_m and on A6); the metric is an audit *trigger*, not a safety certificate.

6 CONCLUSION

We presented a formal model of semantic error propagation in linear LLM agent pipelines, the cascade equation, with its geometric lower bound, critical depth m^* and irreversibility threshold $m^\dagger$, and tested its mechanisms with two pre-registered, replicated experiments, reporting every gate, amendment and failure. Relay steps under a budget lose content by erasure and alter almost nothing; processing steps inject distortion where relay distortion is indistinguishable from zero or one to two orders of magnitude smaller; the pre-registered coupling holds in one processing model, where it places the worst-case handoff at an interior compression, and not elsewhere. $m^\dagger$ is a pre-deployment, auditable metric for whether mandated oversight is meaningful or nominal.

AI USE STATEMENT

In this work we used LLM assistants for the following tasks. *Writing*: to draft and polish sections of the paper; every sentence and claim was reviewed by the authors. *Retrieval and discovery*: to find and screen related work; every reference was checked against its source. *Proving mathematical claims*: to check derivations and formulations (including the correction of the injection term of the cascade equation), to verify them numerically (the unrolled form against the recursion on randomised pipelines), to search for counterexamples, to audit the proofs adversarially, and to assist in writing the Lean 4 formalisation of Section 3 in the anonymised repository; every proof was checked by the authors, and the formalisation is checked by Lean's kernel. *Implementation*: LLM coding assistants were used to implement, run and audit the experimental code (the runner, the scorers, the analysis and the test suite); the authors reviewed, tested and executed it. All design decisions, hypotheses, the pre-registration and its amendments were made by the authors; the assistants served as a sounding board in refining the experimental designs and in discussing the results, and the conclusions drawn are the authors'. We have not used generative AI tools for synthetic data generation (the frozen stimuli are produced by deterministic, seeded generators in the repository, and no dataset was produced by prompting a model), for translation, for data cleaning, or for qualitative analysis. We have reviewed all AI-assisted work: every formal statement is verified by an executable test and by the Lean formalisation, every reported number is traced to a manifest generated from the logged traces and recomputed through an independent implementation, and every reference was checked. LLMs are also the *object* of study and serve as second judges in the instrument audit; that use is part of the methodology described in Section 4. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The experiments involve no human subjects: the only human labelling is the blind review of scorer–judge disagreements by the authors, and no personal data is processed. The stimuli are synthetic fact packets about fictitious entities and counterfactual variants of a public dataset (MuSiQue), released with the code; all agent calls were made to commercial or open-weight models under their terms of use. The irreversibility threshold $m^\dagger$ is proposed as an audit trigger for human-oversight provisions, not as a certificate of safety, and Section 5 states the direction of licensed inference explicitly so that the metric cannot be read as clearing a deployment. No conflicts of interest or external sponsorship apply.

REPRODUCIBILITY STATEMENT

All results are reproducible from the anonymised repository.[1] Every formal statement is mirrored by an executable test (the cascade equation against its unrolled form, the geometric bound, the threshold identity $m^\dagger \equiv m^*(\theta/(1-h))$, and recovery of the A9 parameters from synthetic data, with negative controls), and the definitions and results of Section 3 are additionally formalised in Lean 4 with Mathlib (folder `formal/` of the anonymised repository; 77 theorems, no `sorry`, and the file compiles without errors). The frozen stimuli regenerate byte for byte from their generators; configurations pin model identifiers and verification dates and contain the pre-registration; every agent call is logged with its budget, each rewrite attempt, the returned model version and token counts; runs resume without repeating work; and one command re-derives every number in Section 4, recomputes the headline estimates through an independent implementation, and fails on any mismatch.

REFERENCES

[1] J. Hartmanis and R. E. Stearns. On the computational complexity of algorithms. *Transactions of the American Mathematical Society*, 117:285–306, 1965.

[2] Varin Sikka and Vishal Sikka. Hallucination stations: On some basic limitations of transformer-based language models. *arXiv preprint arXiv:2507.07505*, 2025.

[1] `https://anonymous.4open.science/r/irreversibility-threshold-33A7` (anonymised mirror)

[3] A. Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] M. Alizadeh Vandchali et al. One rank at a time: Cascading error dynamics in sequential learning. *arXiv preprint arXiv:2505.22602*, 2025.

[5] D. Tran and D. Kiela. Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets. *arXiv preprint arXiv:2604.02460*, 2026.

[6] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *TACL*, 10:539–554, 2022.

[7] S. Jamshidi, A. Moradi Dakhel, K. W. Nafi, and F. Khomh. Hallucination cascade: Analyzing error propagation in multi-agent LLM systems. *arXiv preprint arXiv:2606.07937*, 2026.

[8] Y. Xie, C. Zhu, X. Zhang, T. Zhu, D. Ye, M. Qi, H. Chen, and W. Zhou. From spark to fire: Modeling and mitigating error cascades in LLM-based multi-agent collaboration. *arXiv preprint arXiv:2603.04474*, 2026.

[9] B. Gurram. Evaluating tool-using language agents: Judge reliability, propagation cascades, and runtime mitigation in AgentProp-Bench. *arXiv preprint arXiv:2604.16706*, 2026.

[10] B. Xia, L. Zhu, E. Gao, Q. Lu, M. Xue, and D. Sejdinovic. Uncertainty propagation in LLM-based systems. *arXiv preprint arXiv:2604.23505*, 2026.

[11] D. Liu. Silent failure in LLM agent systems: The entropy principle and the inevitable disorder of autonomous agents. *arXiv preprint arXiv:2606.08162*, 2026.

[12] S. Rabanser, S. Kapoor, P. Kirgis, K. Liu, S. Utpala, and A. Narayanan. Towards a science of AI agent reliability. *arXiv preprint arXiv:2602.16666*, 2026.

[13] J. Wang et al. Where do deep-research agents go wrong? Span-level error localization in agent trajectories. *arXiv preprint arXiv:2606.02060*, 2026.

[14] X. Lin et al. LLM-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions. *arXiv preprint arXiv:2509.18970*, 2025.

[15] S. Mishra. Cascading hallucination in agentic RAG: The CHARM framework for detection and mitigation. *arXiv preprint arXiv:2606.04435*, 2026.

[16] A. T. Kalai and S. S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pages 160–171, 2024.

[17] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

[18] Z. Ji et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[19] L. Huang et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[20] P. Manakul, A. Liusie, and M. J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*, pages 9004–9017, 2023.

[21] S. Min et al. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, 2023.

[22] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*, 2022.

[23] M. Cemri, M. Z. Pan, S. Yang, et al. Why do multi-agent LLM systems fail? In *NeurIPS Datasets and Benchmarks Track*, 2025. arXiv:2503.13657.

[24] S. Zhang, M. Yin, J. Zhang, J. Liu, Z. Han, J. Zhang, B. Li, C. Wang, H. Wang, Y. Chen, and Q. Wu. Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems. In *ICML*, PMLR 267:76583–76599, 2025.

[25] Q. Wu et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[26] S. Hong et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *ICLR*, 2024.

[27] G. Li et al. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[28] J. Li, Q. Zhang, Y. Yu, Q. Fu, and D. Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024.

[29] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, and J. Zou. Mixture-of-agents enhances large language model capabilities. In *ICLR*, 2025.

[30] X. Liu et al. AgentBench: Evaluating LLMs as agents. In *ICLR*, 2024.

[31] J. R. Finkel, C. D. Manning, and A. Y. Ng. Solving the problem of cascading errors: Approximate Bayesian inference for linguistic annotation pipelines. In *EMNLP*, pages 618–626, 2006.

[32] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[33] S. Ross, G. J. Gordon, and J. A. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

[34] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[35] B. W. Boehm. Verifying and validating software requirements and design specifications. *IEEE Software*, 1(1):75–88, 1984.

[36] IEEE. IEEE Standard for System, Software, and Hardware Verification and Validation. *IEEE Std 1012-2016*, 2017.

[37] P. Christiano, B. Shlegeris, and D. Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

[38] S. R. Bowman et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

[39] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

[40] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L series, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Accessed 2026-07-03.

[41] Y. Bengio et al. International AI Safety Report 2026. Independent international expert panel, February 2026. arXiv:2602.21012. https://internationalaisafetyreport.org. Accessed 2026-07-03.

[42] Anthropic. What 81,000 people want from AI. Research report, 2026. https://www.anthropic.com/81k-interviews. Accessed 2026-07-03.

[43] METR. Measuring AI ability to complete long software tasks. Technical report, Model Evaluation & Threat Research, 2025. arXiv:2503.14499. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/. Accessed 2026-07-03.

[44] A. Challapally, C. Pease, R. Raskar, and P. Chari. The GenAI divide: State of AI in business 2025. Technical report, MIT NANDA, July 2025. Accessed 2026-07-03.

[45] LangChain. LangSmith observability: Tracing and monitoring for LLM applications and agents. Product documentation, 2026. https://docs.langchain.com/langsmith/observability. Accessed 2026-07-03.

A NOTATION

Symbol	Name	Domain	Description
$\mathcal{P} = (A_1, \ldots, A_m, \Phi)$	Linear pipeline	—	Chain of agents with handoffs
$A_k = (M_k, \mathcal{T}_k, n_k)$	Agent k	—	LLM, assigned task, prompt
m	Pipeline depth	$\mathbb{N}$	Number of agents; $m-1$ handoffs
k, j	Agent indices	$1..m$	k generic agent; j summation index
$\phi_k : O_k \to I_{k+1}$	Handoff k	—	Handoff function; $\Phi = \{\phi_k\}_{k=1}^{m-1}$
O_k, O_k^*, I_k	Outputs / input	—	Actual output, ground truth, input
n_k	Prompt length	$\mathbb{N}$	Input tokens for agent A_k
d_k	Embedding dim.	$\mathbb{N}$	Dimensionality of LLM M_k
Ω_k	Per-step capacity	$\mathbb{R}^+$	$c \cdot n_k^2 \cdot d_k$ FLOPs/step
$C(\mathcal{T}_k)$	Task complexity	$\mathbb{R}^+$	Min. complexity to correctly solve $\mathcal{T}_k$
$\rho(\mathcal{T})$	Complexity gap ratio	$[0, \infty)$	$\max\{0,\, C(\mathcal{T}) - \Omega\}/\Omega$ (Conjecture only)
$\mathrm{sim}(\cdot, \cdot)$	Semantic similarity	$[0, 1]$	Abstract function; empirical proxy via A4
δ_k	Intrinsic node error	$[0, 1]$	$1 - \mathrm{sim}(O_k^*, O_k)$
$\hat{\delta}_k$	Empirical estimator	$[0, 1]$	Estimate of δ_k via a proxy for sim
$\delta_{\min}$	Min. node error	$(0, 1]$	$\min_{2 \le k \le m} \delta_k$ (nodes that receive a handoff; δ_1 unconstrained)
κ_k	Compression coeff.	$[0, 1]$	$\lvert I_{k+1}\rvert / \lvert O_k\rvert$, tokens derived from O_k
$\kappa_{\min}$	Min. compression	$(0, 1]$	$\min_{1 \le k \le m-1} \kappa_k$
r	Cascade ratio	$(1, 2]$	$1 + \delta_{\min} \cdot \kappa_{\min}$
δ_0	Basal distortion	$[0, 1]$	Distortion on intact input (A9); $\delta_1 = \delta_0$
λ	Compression penalty	$\mathbb{R}$	Posited > 0 in A9; measured ≈ 0 in relay, > 0 in one of two processing models (Section 4); $\delta_0 + \lambda \le 1$
$p(\kappa)$	Per-step compounded loss	$[0, 1]$	$\kappa\, \delta(\kappa) = (\delta_0 + \lambda)\kappa - \lambda\kappa^2$
κ^*	Worst-case compression	$(0, 1]$	$\min\{1,\, \frac{1}{2} + \delta_0/2\lambda\}$; interior iff $\delta_0 < \lambda$
E_k	Accumulated error	$[0, \infty)$	Accumulated error after node k (cascade equation)
E_k^H	Residual error	$[0, \infty)$	$E_k \cdot (1 - h_k)$, after validator V_k
V_k	Human validator	—	Inserted at node k
h_k	Human gate	$[0, 1]$	$\alpha_k \cdot \beta_k$
α_k	Domain knowledge	$[0, 1]$	Validator expertise for $\mathcal{T}_k$
β_k	Traceability	$[0, 1]$	Ability to reconstruct provenance
$\bar{h} = \bar{\alpha} \cdot \bar{\beta}$	Max. achievable capacity	$(0, 1)$	Upper bound on h_k; domain-level constant (A6)
$h_{\mathrm{req}}(m)$	Required capacity	$[0, 1)$	$1 - \theta/E_m$
θ	Error tolerance	(δ_1, ∞)	Max. usable error; safe zone: bound $\le \theta$ (non-strict)
τ	Tolerance (argument)	(δ_1, ∞)	Bound variable of $m^*(\tau)$
$m^*(\tau)$	Critical-depth function	$\mathbb{N}$	$1 + \lfloor \log(\tau/\delta_1) / \log r \rfloor$
m^*	Critical depth	$\mathbb{N}$	$m^*(\theta)$: largest m at which the lower bound $\delta_1 r^{m-1} \le \theta$ (upper estimate of the safe depth)
$m^\dagger$	Irreversibility threshold	$\mathbb{N}$	$m^*(\theta/(1-\bar{h}))$: largest m at which the lower bound is human-correctable under $\bar{h}$ (upper estimate)
A1–A9	Assumptions	—	Appendix C

Table 1: Notation used throughout this document. Zones: safe $[1, m^*]$, recoverable $(m^*, m^\dagger]$, irreversible $(m^\dagger, \infty)$. O_k^*, sim, θ: validation inputs, given not derived (Remark, §3.7).

B PROOF OF THEOREM 3.7 (FULL)

Claim. Under Assumptions A1–A2, for any linear agentic pipeline of depth $m \geq 2$ with $\delta_{\min} = \min_{2 \leq k \leq m} \delta_k > 0$ (the nodes that receive a handoff; δ_1 is unconstrained), $\kappa_{\min} = \min_{1 \leq k \leq m-1} \kappa_k > 0$, and cascade ratio $r = 1 + \delta_{\min} \cdot \kappa_{\min}$:

$$E_m \;\geq\; \delta_1 \cdot r^{\,m-1}$$

Proof. **Base case** $(m = 1)$: $E_1 = \delta_1 = \delta_1 \cdot r^0$. ✓

Inductive step: Assume $E_{m-1} \geq \delta_1 \cdot r^{\,m-2}$ for some $m \geq 2$. By Equation (2) and the inductive hypothesis:

$$\begin{aligned} E_m &= E_{m-1} \cdot (1 + \delta_m \cdot \kappa_{m-1}) + \delta_m \\ &\geq E_{m-1} \cdot (1 + \delta_m \cdot \kappa_{m-1}) \\ &\geq \delta_1 \cdot r^{\,m-2} \cdot (1 + \delta_m \cdot \kappa_{m-1}) \\ &\geq \delta_1 \cdot r^{\,m-2} \cdot (1 + \delta_{\min} \cdot \kappa_{\min}) \;=\; \delta_1 \cdot r^{\,m-1} \end{aligned}$$

The first inequality drops the non-negative injection term $\delta_m \geq 0$. The second applies the inductive hypothesis; the direction is preserved because $E_{m-1} \geq 0$ and $1 + \delta_m \kappa_{m-1} > 0$. The third uses $\delta_m \geq \delta_{\min}$ (valid since $m \geq 2$) and $\kappa_{m-1} \geq \kappa_{\min}$, together with the fact that $(1 + x)$ is non-decreasing in $x \geq 0$; the base case involves δ_1 alone, which is why $\delta_{\min}$ ranges over $k \geq 2$. □

C Model Assumptions

ID	Name	Statement
A1	Independence	$\delta_k \perp E_{k-1}$: the intrinsic distortion of A_k is independent of the accumulated error it receives. If violated positively (e.g. $\delta_k = \delta_k^0 + \mu E_{k-1}$), the recursion becomes super-geometric and Theorem 3.7 holds *a fortiori*. If violated negatively (error saturation), the model may overestimate E_m.
A2	Linearity	Linear pipelines only. Results provide a lower bound for any sub-graph of a DAG that contains a linear path of depth m. Join nodes require separate treatment.
A3	Stationarity	δ_k is independent of I_k content beyond n_k. Captures average-case behaviour; content-dependent variance in δ_k is not modelled.
A4	sim as proxy	The function sim and its empirical proxies (deterministic fact-level scorers in Section 4) approximate task-relevant semantic loss; the gap between proxy and construct is measured by the instrument audit, not assumed away.
A5	Native inference	Dense self-attention without sparse attention, KV-caching, or tool delegation. Ω_k is a conservative upper bound; the model remains valid under any tighter capacity estimate. A5 concerns the per-step capacity bound only, not the pipeline's memory architecture: a tool call or a retrieval is a node or a handoff within the model, with its own δ_k or κ_k.
A6	Bounded $\bar{h}$	$\bar{h}$ is treated as a domain-level constant independent of k and m. If $\bar{h}$ decreases with m (e.g. due to cognitive overload with longer provenance chains), $m^\dagger$ in Equation (8) is an upper bound on the true irreversibility threshold.
A7	Complexity–Error	$C(\mathcal{T}_k) > \Omega_k \Rightarrow \delta_k > 0$; posited, not derived (stated in Section 3). Motivated by [2] and A5.
A8	Traceability	For tasks with $C(\mathcal{T}_k) > \Omega_k$, validator traceability satisfies $\beta_k < 1$, hence $\bar{h} < 1$ (stated in Section 3.6). Motivated by A7 and [2].
A9	Compression–Injection	$\delta_k = \delta_0 + \lambda(1 - \kappa_{k-1})$ for $k \geq 2$, with $\delta_1 = \delta_0$ (stated in Section 3.5). Couples injected distortion to incoming compression; consistent with A1 and leaves Theorem 3.7 unchanged. Yields the worst-case handoff $\kappa^* = \frac{1}{2} + \delta_0/2\lambda$ (Proposition 3.9). Pre-registered with $\lambda > 0$; supported in processing for gpt-5.4-mini, not supported for claude-haiku-4.5, and without robust evidence in relay (Section 4).

Table 2: Model assumptions. A7–A9 are stated in full in the main text. A1, A3, and A6 are most critical for the validity of the main results.

D Proof of Corollary 3.15

Proof. By Definition 3.14, depth m is human-correctable iff $E_m \leq \theta/(1-\bar{h})$. Substituting the lower bound from Theorem 3.7, this condition is violated when:

$$\delta_1 \cdot r^{\,m-1} \;>\; \frac{\theta}{1-\bar{h}}$$

Solving for m:

$$(m-1) \;>\; \frac{\log\!\big(\theta \,/\, \delta_1(1-\bar{h})\big)}{\log r}$$

The largest integer m for which the correctable condition still holds under the lower bound is $m^\dagger = m^*\!\big(\theta/(1-\bar{h})\big)$ as given in Equation (8). Note that this bound applies to the geometric lower bound on E_m; actual pipelines may become irreversible at smaller depths. □

E Extended Related Work

This appendix expands Section 2.

Error and limitation in single LLM agents. A substantial body of work characterises why individual LLMs produce incorrect outputs [18, 19]: calibration arguments show that language models must hallucinate on facts seen rarely in training [16], standard training and evaluation reward confident guessing over abstention [17], and a parallel literature builds detectors and benchmarks for such errors [20, 21, 22]. Most relevant to us, Sikka and Sikka [2] argue from the $\mathcal{O}(n^2 d)$ cost of dense self-attention [3] and the Time Hierarchy Theorem [1] that, beyond a task-complexity threshold, an LLM cannot reliably solve *or verify* a task within its per-step computational budget. We adopt this capacity bound as the source of each agent's intrinsic distortion δ_k and, crucially, as the reason a human validator's traceability β_k is itself bounded. This line of work treats the single agent in isolation; it does not address what happens when many such agents are composed.

Multi-agent systems: empirical evidence. A complementary, empirically driven literature studies multi-agent LLM systems directly [30, 28, 29]. Tran and Kiela [5] show that, under matched thinking-token budgets, single-agent systems match or outperform multi-agent systems on multi-hop reasoning, and ground the result in the Data Processing Inequality [34]: passing information through additional agents cannot increase task-relevant information and typically degrades it. This is precisely the qualitative behaviour our κ_k-scaled multiplicative recursion predicts—each handoff is an information-lossy channel—and we read their findings as empirical evidence consistent with our model rather than as a competing claim. Recent taxonomies of multi-agent failure modes—most prominently MAST, an empirically grounded taxonomy of 14 failure modes across 7 frameworks [23]—catalogue *how* such systems break in practice; failure-attribution studies show that even locating *which* agent failed is hard for state-of-the-art models [24], and a recent survey consolidates hallucination behaviour specific to agents [14]. Our contribution is orthogonal and complementary: rather than enumerating failure categories, we give a quantitative model that predicts *when*, as a function of depth, failure becomes unavoidable.

Formal models of error propagation. The closest formal antecedent is Alizadeh Vandchali et al. [4], who analyse cascading error dynamics in sequential learning through low-rank linear regression and prove that errors from limited budgets and finite precision compound predictably across sequential steps. Their analysis is for numerical estimation in a learning procedure; ours is for semantic distortion across a chain of LLM agents, but the structural lesson—that sequential dependence converts local error into compounding global error—is shared, and we cite their bounds as precedent for the multiplicative form. Older precedents exist outside the LLM setting: cascading errors in greedy NLP annotation pipelines [31] and compounding exposure bias in autoregressive sequence models [32, 33] both convert per-step error into super-linear end-to-end error. Against this backdrop, our paper is, to our knowledge, the first to provide (i) a closed-form *irreversibility threshold* $m^\dagger$ marking the depth beyond which bounded human correction cannot recover the output, and (ii) a formal, bounded *human-validator variable* $h_k = \alpha_k \beta_k$ with an axiomatised ceiling $\bar{h} < 1$ motivated by the same capacity bound that governs agent error.

Closest work. Two contemporaneous papers study error cascades in LLM agent systems directly. Jamshidi et al. [7] empirically track hallucination across *refinement* chains, in which each agent receives, critiques, and rewrites the full previous answer, and find net *attenuation* (normalised hallucination falls from 0.422 to 0.272 across three agents). Their setting differs from ours in precisely the variable our model isolates: refinement gives every agent access to the full upstream context (a corrective task with $\kappa \approx 1$), whereas we model *decomposition* pipelines whose lossy handoffs (κ_k, Definition 3.5) strip the provenance a downstream agent would need to distinguish inherited error from signal. Read through our model, their result strengthens rather than contradicts the thesis: the topology of information flow, not the number of agents, determines whether the regime is attenuating or amplifying. Xie et al. [8] model error cascades on multi-agent graphs as binary contagion and propose genealogy-based governance as a defence; we differ in tracking a *continuous* error magnitude in closed form and deriving ex-ante thresholds (m^*, $m^\dagger$) with an axiomatised human gate. Other contemporaneous treatments are empirical rather than closed-form: propagation

benchmarks [9, 13], uncertainty propagation [10], silent failure [11], cascade detection in agentic RAG [15], and reliability methodology [12].

Verification, validation, and oversight. Software engineering distinguishes verification (meeting a given specification) from validation (fitness of the specification for its purpose) [35, 36]; Section 3 adopts the vocabulary to place our human gate. The scalable-oversight programme [37, 38] pushes machine verification beyond unaided human capacity; our claim is orthogonal: we bound the gate's *verification* capacity by access to provenance, and locate what no verifier — human or machine — supplies in validation.

F Remarks and Anticipated Objections

This appendix collects the remarks that qualify the definitions of Section 3, and the full form of the replies summarised in Section 5.

Operational interpretation of m^* and $m^\dagger$. Table 3 summarises the three regimes.

Region	Depth range	Status under the model
Safe zone	$[1,\ m^*]$	bound $\delta_1 r^{m-1} \le \theta$; crossing not yet forced
Recoverable zone	$(m^*,\ m^\dagger]$	$E_m > \theta$; correction not excluded by the bound, feasible only if $h_k \ge h_{\mathrm{req}}(m) = 1 - \theta/E_m$ (requires E_m)
Irreversible zone	$(m^\dagger,\ \infty)$	correction to below θ not feasible under $\bar{h}$

Table 3: Operational interpretation of m^* and $m^\dagger$ under the model.

Remark (What κ_k does and does not measure). κ_k measures token-level *information retention* of the handoff, not the transmission of semantic error as such. In the cascade equation it scales the amplification of inherited error; under Assumption A9 compressive handoffs additionally *inject* distortion at the receiving agent, so a low κ_k is not, by itself, protective. Empirically, the estimator $\hat{\kappa}_k$ excludes prompt-template tokens and is clipped to $[0, 1]$: paraphrase or expansion by the receiving agent can push the raw token ratio above 1, and both the raw and clipped values are logged. A measurement convention accompanies this split: our estimators attribute *deletion* of content to the channel (κ) and reserve $\hat{\delta}$ for *alteration*; under the literal Definition 3.3, deletion of load-bearing content would also register as distortion, so every reported estimate states which channel it measures. In Section 4, Exp M's recall loss is the erasure channel and its $\hat{\delta}$ the alteration channel.

Remark (Scope: distortion, not erasure). The cascade equation is an index of *distortion*: E_m compounds the per-node deviations δ_k from the references O_k^*. A handoff that deletes content lowers κ_k and reduces what downstream nodes can amplify, but the deletion itself is not an error term of Equation (2); under the literal Definition 3.3 content that an *agent* omits from its own output is deviation from O_k^*, and the estimators of Section 4 separate the two by convention (alteration to $\hat{\delta}$, deletion to $\hat{\kappa}$). Accordingly, the tests of Assumption A9 concern the alteration channel, the recall trajectories of Exp M are the erasure channel and test the propagation form only under the literal reading, and the theory does not explain *why* budgeted relays erase rather than alter: that is an empirical finding, not a consequence of the model.

Remark (On $\delta_{\min}$). Theorem 3.7 takes $\delta_{\min}$ over the nodes $k \ge 2$, which are the nodes that amplify inherited error; the first node contributes δ_1 alone through the base case. A minimum over all nodes would force $\delta_{\min} \le \delta_1$ and make the reference configuration of Figure 1 ($\delta_1 = 0.05$, $\delta_{\min} = 0.08$) inadmissible; with the stated definition it is admissible, and r is unchanged wherever $\delta_1 \ge \delta_{\min}$.

Remark (Interpretation of E_m and alternative forms). Each $\delta_k \in [0, 1]$ lives on the similarity scale, but E_m is an *unbounded error index*: the recursion compounds distortions multiplicatively, so E_m may exceed 1, and only its comparison to θ and ordinal comparisons carry interpretation. The amplification-plus-injection form is not derived from first principles: the scalar multiplier is the

first-order expansion of any smooth propagation operator $f(E_{k-1}, \delta_k)$ satisfying the injection and amplification requirements, with higher-order structure captured by the interaction-term alternative $E_k = E_{k-1}(1 + \delta_k^0 \kappa_{k-1} + \mu E_{k-1}\kappa_{k-1}) + \delta_k^0$. If Assumption A1 is violated positively, the true recursion is super-geometric and the present model is conservative; if violated negatively (error saturation), the model may overestimate propagation at large E_{k-1}.

Remark (Asymptotic dominance and finite-depth crossover). For a pipeline with uniform handoff compression κ, the depth-m error $E_m(\kappa)$ is governed at small m by the injection channel and at large m by the geometric channel: $\arg\max_\kappa E_m(\kappa)$ sits at $\kappa = 0$ for shallow pipelines and migrates toward κ^* as $m \to \infty$, with convergence of order $1/m$. Numerically, for $(\delta_0, \lambda) = (0.02, 0.10)$ the argmax leaves $\kappa = 0$ at $m = 16$ and $E_m(\kappa^*)$ overtakes $E_m(0)$ at $m = 37$; for $(0.05, 0.20)$, at $m = 9$ and $m = 18$; for $(0.01, 0.05)$, at $m = 30$ and $m = 72$. At experimentally accessible depths ($m \lesssim 16$) the observable signature of A9 is therefore the *concavity of the per-step loss* $p(\kappa)$, loss peaking at intermediate compression, rather than an interior maximum of E_m itself. Both are discriminating predictions absent from the uncoupled model, which locates maximal loss at full pass-through.

Remark (Where irreversibility lives). A9 reframes the role of the handoff: pass-through maximises the *transmission* of whatever error exists but minimises its *creation*, and it preserves the upstream context on which both downstream agents and human validators rely. Compression removes content outright and destroys that context; the same deletion that drives the erasure channel plausibly lowers achievable traceability $\bar{\beta}$ (Assumption A8), tightening $m^\dagger$. Formalising a dependence $\bar{h}(\kappa)$ is left as a refinement.

Remark (Pass-through is not indefinitely available). Full pass-through preserves content and provenance, but accumulated state grows with depth while context windows do not, so at sufficient depth compression arrives by arithmetic rather than by design choice; the binding condition is the payload-versus-channel threshold of Section 4, which no prompt engineering removes. External memory (a state store, retrieval) relocates the handoff rather than removing it: whatever agent k actually conditions on is I_k, and κ_{k-1} is measured there. Exp B is exactly this pattern.

Remark (Empirical status of A9). The programme of Section 4 tests the coupling in both regimes. In relay (Exp M, three models) it is not supported: the fitted $\hat{\lambda}$ have intervals covering 0 for two models, and the third's formally positive slope is not robust to the compliant-steps-only sensitivity. In processing (Exp B) the two models part ways: gpt-5.4-mini shows a positive coupling ($\hat{\lambda} = 0.1554\ [0.0926, 0.2228]$) that is positive in every replicate and under every pre-registered sensitivity and the post-hoc precision correction, while claude-haiku-4.5's interval covers 0. We retain A9 and Proposition 3.9 as stated because they are the pre-registered hypothesis under which the programme was designed, and because for the model where the coupling holds the point estimates satisfy $\hat{\delta}_0 < \hat{\lambda}$, placing the worst-case handoff κ^* in the interior of $(0, 1)$, the proposition's distinctive prediction. The interior placement rests on the point estimates: at the lower end of $\hat{\lambda}$'s interval (0.0926) the condition $\hat{\delta}_0 < \hat{\lambda}$ no longer holds ($\hat{\delta}_0 = 0.1144$), so the prediction is live but not yet established. The uncoupled model ($\lambda = 0$) and the deletion-only reading ($\lambda \leq 0$), under which the worst case for transmitted distortion returns to the pass-through boundary, are the limiting cases that describe relay and the other processing model. In every case, compression's cost includes erasure and the loss of the provenance on which correction depends. Testing the interior maximum directly requires a wider realised compression range than Exp B achieved (failed gate G2) and a sharper alteration probe (Section 4.3).

Remark (Verifier, not validator, as modelled). h_k is a *verification* capacity: it measures the reviewer's ability to bring O_k back toward a reference O_k^* that the model takes as given. We retain the term *validator* for the reviewer, for continuity with the oversight literature, but as modelled the gate verifies. In deployment the same expert does double duty: α_k is at once an instrument of verification and, absent an oracle, the practical source of the reference verified against.

Remark (Reviewer-introduced error). Definition 3.11 lets the reviewer only remove error. If review also *introduces* error, an additive term $\varepsilon \geq 0$ for incorrect corrections, bias, or fatigue, $E_k^H = E_k(1 - h_k) + \varepsilon$, the correctable condition of Definition 3.14 becomes $E_m \leq (\theta - \varepsilon)/(1 - \bar{h})$ and

every threshold moves earlier: $m^\dagger$ as stated is an upper bound on the true irreversibility threshold, in the same direction as the caveat of Assumption A6. We omit ε because it is not identified by our experiments, not because it is negligible.

Remark (Sensitivity to the functional form). Equation (2) lets E_m grow without bound. Under the bounded, saturating alternative $1 - E_k = (1 - E_{k-1})(1 - \delta_k \kappa_{k-1})$, multiplicative in the surviving information and keeping $E_k \in [0, 1]$, a finite $m^\dagger$ exists *only if* $\theta < 1 - \bar{h}$: otherwise $E_m(1 - \bar{h}) \leq 1 - \bar{h} \leq \theta$, and every depth remains correctable. The two families therefore disagree on an observable, whether an irreversibility threshold exists at all when $\theta \geq 1 - \bar{h}$, and we treat this as a discriminating, testable prediction between the cascade equation and the saturating family rather than as a defect of either form. The survival form that wins some Exp M cells (Section 4.1) belongs to the saturating family.

Remark (Verification and validation in the model). Software engineering distinguishes *verification*, whether the output meets a given specification, from *validation*, whether the specification is the right one for the purpose [35, 36]. The model of Section 3 is a model of verification: δ_k measures deviation from a fixed reference O_k^* under a fixed sim, E_m compounds it, and the gate $h = \alpha\beta$ formalises the capacity to detect and correct such deviations. O^*, sim, and θ are *validation inputs*: the model presupposes them and does not derive them. The experiments accordingly test verification dynamics against designer-constructed specifications read by mechanical verifiers; they do not estimate α, β, or $\bar{h}$, whose values in the text are illustrative.

The verifier-regress objection, in full. A natural objection to Definition 3.10 runs: nothing in the model bounds a machine verifier's h below $\bar{h}$, it may have $\beta \approx 1$ (full trace access, unlimited time) and rising α; Corollary 3.15 itself says that past $m^\dagger$ more human effort does not help; and our own experiments measure error mechanically. Why, then, a human gate? Three replies. First, on *verification* the objection is conceded: nothing in the model privileges a human verifier, and a machine verifier may exceed human verification capacity wherever the specification is closed. Even so, it is not an exemption from the model but another node inside it: the capacity bound of Assumption A7 applies to solving *and* verifying [2], so inserting a verifier extends the chain under Equation (2) with its own δ_k; and training a verifier against a reward substitutes yet another proxy for sim, so "who verifies the verifier" recurses within the model rather than out of it. Second, the claim that survives concerns *validation*. In deployment there is no oracle: our experiments score error deterministically only because we *constructed* the ground truth by design; outside the laboratory, O^*, sim, and the tolerance θ are constituted by human purposes (θ is already a human-set parameter of the model), and even mechanically verifiable domains (test suites, compilers) bottom out in a specification some high-α human articulated: verification artifacts are congealed human judgment. Constituting the criterion is not a checking operation that a better checker could perform; it is the input the model presupposes rather than models, and it is there, not in verification, that the gate is non-delegable. Third, under oversight regulation [40] responsibility is non-delegable, and the accountable evaluator needs access to the grounds of judgment; that access is precisely β, and treating it as depth-independent is the simplification of Assumption A6, whose violation only tightens $m^\dagger$. The irreversibility threshold is therefore not a claim that humans out-correct machines; it is the depth beyond which the *evaluator of record* loses access to the grounds for any verdict, including the grounds on which to verify what they themselves constituted as the criterion. Progress in models raises the floor ($\delta_{\mathrm{floor}} > 0$ notwithstanding [16]); the evaluator sets the ceiling.

Model selection and scope of claims. The programme assigns small models to the relay and processing positions. This is an experimental manipulation, not a cost concession: the model's claims are parametric in δ_0, λ and κ, not in any model family, and the pilots showed frontier relays far below the regime in which rival propagation forms are statistically distinguishable at feasible depths, so measuring the mechanism required conditions under which it is measurable, as a stress test does. Three rules bound the scope of the conclusions. First, no claim about frontier behaviour is extrapolated from the small-model cells. Second, all conclusions are stated in the model's parameters, with per-model estimates, so results survive model deprecation and cannot be read as claims about specific products; the data vindicate the discipline, since the models tested differ in propagation form and in whether the coupling holds. Third, the weak-model-with-compressive-handoffs configuration is not an adversarial corner case but the one most representative of deployed pipelines, in which cost pressure routinely assigns intermediate relay and summarisation steps to smaller models;

the all-frontier pass-through pipeline is the expensive exception. All model versions are pinned by identifier and verification date; the decision criteria (the affine coupling fit with cluster-bootstrap intervals, the BIC form contest, the gates) were declared before the runs; and failed gates are reported rather than smoothed.

G Pipeline Schematic

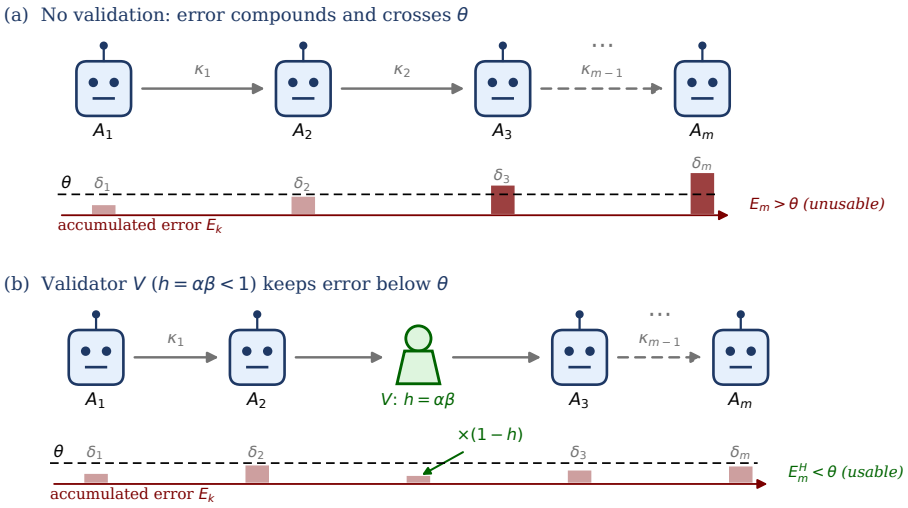

Figure 3: A linear agentic pipeline of LLM agents (blue squares, each with intrinsic distortion δ_k), with handoffs ϕ_k transmitting a fraction κ_k of the previous output. The lower strip depicts the accumulated error E_k, with each δ_k marked just above its step. **(a)** Without intervention, intrinsic agent distortion compounds and the accumulated error E_m grows along the chain, crossing the tolerance θ. **(b)** A human validator V (green circle) inserted at a node reduces the accumulated error by a factor $(1 - h)$ with $h = \alpha\beta$ (domain expertise α, traceability β); since $h < 1$ is bounded (Assumption A8), the correction is partial and error resumes growing downstream.

H Experimental Tables

Full result tables for the programme of Section 4. All intervals are cluster-bootstrap 95% intervals (by packet in Exp M, by chain in Exp B; 10,000 resamples, fixed seed); every value is re-derived from the committed traces by the repository's Section 4 builder, which recomputes the headline estimates through an independent implementation and fails on any mismatch. Model identifiers are the pinned snapshots (claude-haiku-4-5-20251001, gpt-5.4-mini-2026-03-17, Qwen/Qwen3-32B), verified on 2026-09-24.

Stage	Design	Scale	Status
Pilot 1, Pilot 2	multi-hop QA answer relay, pass-through and nominal compression, depth up to 16	exploratory	ran before the protocol; instrument and manipulation failures reported; no headline claim
Exp M	fictitious fact packets ($N \in \{10, 20\}$) relayed up to 12 steps under a behavioural budget $\kappa \in \{0.9, 0.75, 0.5, 0.35, 0.2\}$ plus uncapped control; 3 models	6 cells $\times$ 3 models, 240 relays per cell (three replicates pooled), 4,320 relays	pre-registered (tag `prereg-v2`); all gates passed
Exp B	60 counterfactual MuSiQue chains ($m \in \{2, 4\}$), notes re-emitted under $\kappa \in \{0.85, 0.7, 0.55\}$, alternating, and uncapped; gold-state, no-notes and 2-fact arms; 2 models	5 cells $\times$ 2 models $\times$ 3 replicates	pre-registered; gates G2 and G3 failed in the main run (reported)

Table 4: The empirical programme. Per-run details, calibration runs and the instrument audit are itemised in the repository.

Model	$\hat{\delta}_0$ [95% CI]	$\hat{\lambda}$ [95% CI]
claude-haiku-4.5	0.0013 [0.0009, 0.0018]	-0.0063 [-0.0145, 0.0001]
gpt-5.4-mini	-0.0000 [-0.0001, 0.0000]	0.0023 [-0.0004, 0.0072]
qwen3-32b	0.0005 [0.0003, 0.0007]	0.0013 [0.0001, 0.0026]

Table 5: Exp M (relay): affine coupling fit $\hat{\delta}(\hat{\kappa}) = \hat{\delta}_0 + \hat{\lambda}(1 - \hat{\kappa})$ on the alteration channel. Primary analysis: all steps except forced trims that may cut inside a fact, packet-cluster bootstrap (10,000 resamples). H-M1 ($\hat{\delta}_0$ interval below 0.01) is supported for all three models; H-M2 ($\hat{\lambda} > 0$, interval excluding 0) is not supported (qwen3-32b meets the rule formally; see Table 6).

Model	$\hat{\lambda}$, compliant steps only	$\hat{\lambda}$, steps $k \geq 2$	$\hat{\lambda}$ by replicate
claude-haiku-4.5	0.0074	-0.0077	0.0010, -0.0074, -0.0129
gpt-5.4-mini	0.0023	0.0003	-0.0006, 0.0008, 0.0063
qwen3-32b	-0.0006	0.0006	0.0017, 0.0012, 0.0009

Table 6: Exp M pre-registered sensitivities for the coupling slope (point estimates). Restricting to budget-compliant steps reverses the sign of qwen3-32b's slope.

	budget target κ					
Model	1.0 (control)	0.9	0.75	0.5	0.35	0.2
claude-haiku-4.5: final recall	0.985	0.576	0.137	0.161	0.195	0.165
realised $\hat{\kappa}$ (cap-active steps)	—	0.90	0.74	0.49	0.34	0.18
unaided budget compliance	—	66%	51%	71%	38%	60%
gpt-5.4-mini: final recall	1.000	0.788	0.178	0.189	0.204	0.211
realised $\hat{\kappa}$ (cap-active steps)	—	0.94	0.78	0.51	0.36	0.19
unaided budget compliance	—	59%	34%	71%	38%	60%
qwen3-32b: final recall	0.982	0.713	0.145	0.142	0.147	0.143
realised $\hat{\kappa}$ (cap-active steps)	—	0.94	0.76	0.49	0.34	0.17
unaided budget compliance	—	53%	35%	71%	39%	62%

Table 7: Exp M cells (240 relays per cell, three replicates pooled): mean final recall, realised compression on cap-active steps, and the share of capped steps that met the budget without a rewrite or a forced trim. Form contest (BIC, pre-saturation window): cascade wins with convex curvature in the $\kappa = 0.9$ cells ($N = 10$ and 20) for all three models and in the $\kappa = 0.75$, $N = 10$ cell for claude-haiku-4.5 and gpt-5.4-mini; the $\kappa = 0.75$, $N = 20$ cell is indistinct from additive for those two models; both $\kappa = 0.75$ cells of qwen3-32b are concave (survival at $N = 10$, additive at $N = 20$).

Model	$\hat{\delta}_0$ [95% CI]	$\hat{\delta}_0^{\text{net}}$ [95% CI]	relay $\hat{\delta}_0$ (Exp M)	$\hat{\lambda}$ [95% CI]
claude-haiku-4.5	0.0817 [0.0650, 0.0988]	0.0532 [0.0314, 0.0728]	0.0013	0.0150 [-0.0146, 0.0433]
gpt-5.4-mini	0.1144 [0.0900, 0.1389]	0.0670 [0.0436, 0.0907]	-0.0000	0.1554 [0.0926, 0.2228]

Table 8: Exp B (processing), primary analysis: all steps, chain-cluster bootstrap (10,000 resamples). $\hat{\delta}_0^{\text{net}}$ is net of the gold-state floor. H-B1 (net > 0, above relay) is supported for both models; H-B2 ($\hat{\lambda} > 0$) for gpt-5.4-mini only.

	claude-haiku-4.5		gpt-5.4-mini	
Sensitivity	$\hat{\delta}_0$	$\hat{\lambda}$	$\hat{\delta}_0$	$\hat{\lambda}$
replicate 1 alone	0.0750	0.0396	0.1085	0.1583
replicate 2 alone	0.0840	0.0011	0.1164	0.1963
replicate 3 alone	0.0861	0.0035	0.1181	0.1129
compliant steps only	0.1246	0.0058	0.0802	1.1560
guessable chains excluded (11 / 24)	0.0817	0.0061	0.1313	0.1698
parse-ok steps only	0.0788	0.0167	0.1144	0.1554
superseded v1 probe	0.1972	0.0489	0.2141	-0.0279

Table 9: Exp B pre-registered sensitivities (point estimates). The gpt-5.4-mini slope is positive in every replicate and under every sensitivity except the superseded v1 probe; the claude-haiku-4.5 slope is small and inconsistent.

Model	p (entity-label precision)	$\hat{\delta}_0$ [95% CI]	$\hat{\delta}_0^{\text{net}}$ [95% CI]	$\hat{\lambda}$ [95% CI]
claude-haiku-4.5	1 (primary)	0.0817 [0.0639, 0.0983]	0.0532 [0.0314, 0.0728]	0.0150 [-0.0162, 0.0434]
	0.72 (Wilson upper)	0.0635 [0.0490, 0.0781]	0.0385 [0.0196, 0.0551]	0.0088 [-0.0179, 0.0335]
	0.50 (point)	0.0488 [0.0359, 0.0612]	0.0266 [0.0096, 0.0420]	0.0063 [-0.0175, 0.0295]
	0.28 (Wilson lower)	0.0339 [0.0241, 0.0443]	0.0142 [-0.0010, 0.0285]	0.0019 [-0.0172, 0.0190]
	0 (value-only)	0.0146 [0.0087, 0.0211]	-0.0009 [-0.0127, 0.0090]	-0.0021 [-0.0135, 0.0085]
gpt-5.4-mini	1 (primary)	0.1144 [0.0899, 0.1398]	0.0670 [0.0436, 0.0907]	0.1554 [0.0895, 0.2248]
	0.72 (Wilson upper)	0.0927 [0.0725, 0.1140]	0.0541 [0.0325, 0.0759]	0.1334 [0.0749, 0.1932]
	0.50 (point)	0.0754 [0.0582, 0.0922]	0.0431 [0.0233, 0.0607]	0.1142 [0.0619, 0.1687]
	0.28 (Wilson lower)	0.0581 [0.0436, 0.0720]	0.0330 [0.0160, 0.0482]	0.0982 [0.0531, 0.1505]
	0 (value-only)	0.0362 [0.0258, 0.0462]	0.0200 [0.0075, 0.0308]	0.0751 [0.0394, 0.1152]

Table 10: Post-hoc precision correction (amendment A5): each entity-kind "altered" label counted as real with probability p (20 imputations; value-kind labels, confirmed $8/8$, unchanged). H-B1 holds for gpt-5.4-mini at every p and for claude-haiku-4.5 depends on p; H-B2 for gpt-5.4-mini holds at every p.

	cell (budget target)				
Model	1.0 (control)	0.85	0.7	0.55	alternating
claude-haiku-4.5: realised $\hat{\kappa}$ (cap-active)	—	0.56	0.50	0.41	0.48
unaided budget compliance	—	30%	27%	24%	34%
parse rate	100.0%	96.1%	97.6%	97.2%	97.2%
gpt-5.4-mini: realised $\hat{\kappa}$ (cap-active)	—	0.66	0.56	0.44	0.53
unaided budget compliance	—	59%	45%	34%	36%
parse rate	99.8%	100.0%	100.0%	100.0%	100.0%

Table 11: Exp B cells (60 chains per cell, three replicates): realised compression on cap-active steps, unaided budget compliance, and parse rate. The realised span of the capped cells (0.15 for claude-haiku-4.5, 0.22 for gpt-5.4-mini) is below the pre-registered gate G2 (0.25); the primary A9 fits also include the uncapped control.

Model	1-fact demand	2-fact demand	H-B3 (ordinal)
claude-haiku-4.5	0.381	0.543	supported
gpt-5.4-mini	0.417	0.477	supported

Table 12: Conjecture step (uncapped, same task family): per-hop error under a 1-fact and a 2-fact per-step demand. The programme's only direct, ordinal, evidence on Conjecture 3.4.

	mutation detection	judge κ (raw agr., n)	blind human review: calibration / main
Exp M	delete fact 100.0%, perturb value 99.0%, swap values 98.1%	0.59 (72%, 300)	scorer 11/12 / scorer 4/10, judge 4/10
Exp B	delete line 94.9%, drop tag 100.0%, perturb value 77.3% (failed gate G3), revert to real 98.2%	0.59 (73%, 300)	scorer 7/12 / scorer 2/10, judge 6/10

Precision of scorer "altered" labels (random sample, blind): Exp M 1/8; Exp B value-kind 8/8, entity-kind 8/16 (Wilson 95% $[0.28, 0.72]$).

Table 13: Instrument audit on the real outputs of the main run (judge: third-provider model, stratified sample) and the blind human review of scorer–judge disagreements (one reviewer).

Pilot	Cohen's κ	raw agreement
Pilot 1	0.900	95.8%
Pilot 2	0.900	95.4%

Table 14: Inter-judge agreement on the exploratory pilots when a third-provider judge scores the *same* candidates as each original judge.

I PRE-REGISTRATION, AMENDMENTS AND GATES

Hypotheses, decision rules, instrument definitions, the primary analysis with its sensitivities, and the calibration gates were fixed in the `prereg` blocks of the experiment configurations and tagged in version control before any paid call. Every later change is a dated amendment in the same block (Table 15); A5 was made after the main-run results were read and is reported only as a sensitivity. Table 16 lists the gates as evaluated on the main run.

ID	When (UTC)	Change
A1	2026-09-25, planned in the tag	Terse anchors (each model's shortest lossless rewrite) measured and written into the cap fields; every rewrite lossless (Exp M 10/10 per model, Exp B 8/8 per step and model). Nothing else changed.
A2	2026-09-25, after calibration, before the main run	Gate G2 evaluated on cap-active steps only (at the floor $\hat{\kappa} \approx 1$ by construction). With this metric Exp M passes for all models and Exp B sits at the threshold (calibration span 0.24–0.25). No design change.
A3	2026-09-25, after calibration, before the main run	Exp M scorer fixes found by the calibration audit: month-name dates and spelled-out units canonicalised; value-before-entity phrasing and unique surnames counted as recalled; partial values still altered. Calibration judge κ 0.48 $\rightarrow$ 0.64.
A4	2026-09-25, after calibration, before any main-run result was read	Blind human review of 24 calibration disagreements sided with the Exp M scorer in 11/12 and the Exp B probe in 7/12; the five probe errors (parsing of heavily compressed notes) fixed with one regression test each; mutation harness edits value sides only.
A5	2026-09-25, **post-hoc**, after the main-run results were read	Random precision sample of scorer "altered" labels added to the blind review (Exp B value-kind 8/8, entity-kind 8/16; Exp M 1/8); Exp B estimates re-weighted by entity-label precision p as a sensitivity (Table 8). No pre-registered analysis changed.

Table 15: Dated amendments to the pre-registration.

Gate	Experiment	Rule	Main run
G1	Exp M	every relay completes	pass (0 incomplete, 0 failed)
G1	Exp B	parse rate $\geq 85\%$ in every cell	pass (both models)
G2	Exp M	realised $\hat{\kappa}$ monotone across targets, capped span ≥ 0.25	pass (spans 0.73, 0.75, 0.77)
G2	Exp B	same	**fail**: monotone, spans 0.15 (haiku), 0.22 (mini)
G3	Exp M	mutation detection $\geq 90\%$ for every kind	pass (worst 98.1%)
G3	Exp B	mutation detection $\geq 80\%$ for every kind	**fail**: worst 77.3% (perturbed value)
G4	all	projected cost below account balance	pass (calibration)

Table 16: Pre-registered gates evaluated on the main run. The two Exp B failures are reported in Section 4.2 and carried into the limitations rather than remedied by re-running.

J SENSITIVITY ANALYSES

This appendix quantifies (i) how estimation error in the inputs to $m^\dagger$ propagates to the threshold, and (ii) how $m^\dagger$ depends on the aggregation chosen for the human gate $h = \alpha\beta$. Both analyses are *illustrative*: they use the reference configuration ($\delta_1 = 0.05$, $\kappa = 1$, $\theta = 0.5$) with stated input uncertainties, and are to be re-instantiated with empirical confidence intervals once $\hat{\delta}_{\min}$, $\hat{\kappa}_{\min}$, and $\bar{h}$ are measured.

J.1 Propagation of parameter uncertainty to $m^\dagger$

Because $m^\dagger$ depends on its inputs through a logarithm and a reciprocal $(1-\bar{h})^{-1}$ (Equation (8)), it inherits their uncertainty non-linearly and asymmetrically. We propagate uncertainty by Monte Carlo: treating the central estimates $\delta_{\min} = 0.08$, $\bar{h} = 0.7$, $\kappa_{\min} = 1$ as means with coefficients of variation 15%, 10% and 5% (truncated to their admissible domains), we draw 3×10^5 samples and evaluate $m^\dagger$ and m^*. The point estimate $m^\dagger = 46$ carries a 90% interval $[37, 64]$ (Monte-Carlo median 47, mean 48.5, s.d. 8.4); the critical depth $m^* = 30$ carries $[25, 41]$ (Table 17). The interval is right-skewed, since $m^\dagger \to \infty$ as $\bar{h} \to 1$ or $\delta_{\min} \to 0$. Decomposing the spread by perturbing one input at a time, the basal-error estimate $\delta_{\min}$ dominates—alone it produces a 5–95 width of 23 depths—followed by $\bar{h}$ (10) and $\kappa_{\min}$ (4).

Two consequences follow. First, $m^\dagger$ is a genuine but *soft* threshold: reporting it as a bare integer overstates its precision, and deployments should report it with a propagated interval. Second, its usefulness as an auditable metric hinges disproportionately on a precise estimate of $\delta_{\min}$, which should therefore be the priority of any calibration effort.

	p5	median	p95
$m^\dagger$	37	47	64
m^*	25	31	41

Table 17: Monte-Carlo distribution of $m^\dagger$ and m^* under $15\%/10\%/5\%$ coefficient of variation on $\delta_{\min}/\bar{h}/\kappa_{\min}$ (reference configuration, 3×10^5 draws). $\delta_{\min}$ is the dominant source of uncertainty.

J.2 Aggregation form of the human gate

Definition 3.10 models $h = \alpha\beta$ as a product, but the results depend only on the scalar $\bar{h} \in (0, 1)$ that enters $m^\dagger = m^*(\theta/(1-\bar{h}))$, not on the specific aggregation. Fixing best-available factors $\bar{\alpha} = 0.9$, $\bar{\beta} = 0.8$ and the reference configuration, the product, minimum, harmonic-mean and arithmetic-mean aggregations give $\bar{h} \in \{0.72, 0.80, 0.85, 0.85\}$ and $m^\dagger \in \{47, 51, 55, 55\}$, while $m^* = 30$ is unchanged (Table 18). The three-regime structure and the ordering $m^\dagger \geq m^*$ are invariant; only the numerical threshold shifts, by at most a few depths. The product is the most *conservative* aggregation (smallest $\bar{h}$, smallest $m^\dagger$), which is the prudent default for a safety bound: it under-credits rather than over-credits the reviewer's achievable capacity.

Aggregation of $(\bar{\alpha}, \bar{\beta})$	$\bar{h}$	$m^\dagger$	m^*
product $\bar{\alpha}\bar{\beta}$	0.720	47	30
$\min(\bar{\alpha}, \bar{\beta})$	0.800	51	30
harmonic mean	0.847	55	30
arithmetic mean	0.850	55	30

Table 18: Human-gate aggregation sensitivity at $\bar{\alpha} = 0.9$, $\bar{\beta} = 0.8$ (reference configuration). The operating regimes are invariant across aggregations; the product form is the most conservative.