

Semantic Entropy and Structural Invariance in LLM-Mediated Expansion–Compression Loops

Huiwen Han

Independent Researcher

hanhuiwen@gmail.com

ORCID: 0009-0000-5852-5916

Preprint — March 2026

Target: IEEE Transactions on Information Theory / Entropy (MDPI)

Series context. This is Paper 1 of the six-paper DECO series (*DEcoupling COgnition*). It builds directly on the formal framework of Paper 0 (operators \mathcal{E} , \mathcal{C} , the EC-transform $T = \mathcal{C} \circ \mathcal{E}$, three-stratum fidelity, and the Indecomposable Core \mathcal{K}_*) and provides the quantitative information-theoretic substrate for the system-stability analysis of Paper 2. Familiarity with Paper 0 is assumed; key definitions are recalled as needed.

Abstract

We develop a quantitative information-theoretic account of semantic decay in large-language-model (LLM) mediated Expansion–Compression (EC) loops. Building on the unified framework of DECO Paper 0 (Han, 2026a), we introduce semantic entropy $H_S(X)$ as the differential entropy of a random variable distributed over a semantic manifold, and prove that each application of the EC-transform $T = \mathcal{C} \circ \mathcal{E}$ is a strictly entropy-reducing operation in expectation (Semantic Entropy Collapse Theorem). We derive closed-form bounds on the mutual information $I(K_A; \hat{K}_B)$ between the originating ideas K_A and the recipient’s extraction \hat{K}_B as functions of expansion ratio ρ_E , compression ratio ρ_C , LLM temperature τ , and iteration count n . We introduce the *Semantic Gravity Well* model: a potential landscape on the semantic manifold in which the LLM training centroid acts as an attractor, and original content resides at unstable high-curvature saddle points. Under this model, we prove the Differential Decay Theorem: propositional fidelity F_{prop} decays exponentially in n , affective fidelity F_{aff} decays sub-exponentially, and structural fidelity F_{str} converges to a strictly positive constant (the causal skeleton invariant) under mild conditions. Finally, we characterise the Semantic Channel Capacity of the EC-loop and show it is strictly less than the Shannon capacity of the raw linguistic channel, with the gap determined by the curvature of the semantic manifold at the seed point. These results quantify the information-theoretic cost of AI-mediated cognitive decoupling and provide empirically testable predictions for the subsequent papers in the series.

Keywords: semantic entropy, mutual information, expansion–compression loop, semantic manifold, structural invariance, information decay, LLM-mediated communication, channel capacity, cognitive decoupling

Contents

1	Introduction	4
1.1	The Central Question	4
1.2	Summary of Results	4
1.3	Relation to the Information Theory Literature	5
2	Mathematical Preliminaries	5
2.1	Spaces and Embeddings (Recap)	5
2.2	Semantic Entropy	5
2.3	The Semantic Gravity Well	6
3	Semantic Entropy Dynamics	7
3.1	Entropy of the Expansion Step	7
3.2	Entropy of the Compression Step	8
3.3	The Semantic Entropy Collapse Theorem	8
4	Mutual Information Between Originator and Recipient	10
4.1	Setup	10
4.2	The EC-Loop as a Semantic Channel	10
4.3	Mutual Information Decay Curve	11
5	Differential Fidelity Decay	11
5.1	Formalism for the Three Strata	11
5.2	The Differential Decay Theorem	12
6	The Causal Skeleton as Semantic Invariant	14
6.1	Definition and Robustness	14
6.2	Relationship to the Indecomposable Core	15
7	Rate-Distortion Analysis of the EC-Loop	15
7.1	Formulation	15

8	Empirical Predictions and Experimental Design	16
9	Discussion	17
9.1	The Paradox of the Expansion Step	17
9.2	What the Rate-Distortion Results Mean Practically	17
9.3	Limitations	17
10	Conclusion	18
A	Proof of EC-Channel Capacity Bound (Theorem 4.2)	21
A.1	Local Gaussian Approximation	21
A.2	Capacity Computation	21
B	Estimation of Semantic Entropy from Embeddings	21

1. Introduction

1.1. The Central Question

Paper 0 of this series established that AI-mediated cognitive decoupling is a rational dominant strategy and that its semantic consequences are characterised by mean reversion toward the LLM training centroid. Paper 0 also identified three strata of semantic fidelity — structural (F_{str}), affective (F_{aff}), and propositional (F_{prop}) — and asserted that they decay at different rates under the EC-transform $T = \mathcal{C} \circ \mathcal{E}$.

The present paper is concerned with *how much* is lost, *how fast*, and *under what conditions*. Specifically, we pursue three questions:

- Q1 (Entropy).** What happens to semantic entropy across iterations of the EC-loop? Does information compress monotonically, or can the expansion step introduce genuine novelty?
- Q2 (Mutual Information).** How much of the originator’s intent K_A survives in the recipient’s extraction \hat{K}_B ? How does this depend on the parameters of the loop (ρ_E, ρ_C, τ) ?
- Q3 (Structural Invariance).** Is there a non-trivial portion of the original content that is genuinely preserved — not by accident, but by structural necessity — across arbitrarily many iterations?

1.2. Summary of Results

Main results (informal):

- **Theorem 3.5.** Semantic entropy strictly decreases under each EC-transform in expectation. The expansion step \mathcal{E} increases token-space entropy but decreases semantic-manifold entropy by pulling the distribution toward the training centroid.
- **Theorem 4.2.** $I(K_A; \hat{K}_B) \leq C_{EC} < C_{\text{Shannon}}$, where C_{EC} is the effective channel capacity of the EC-loop and C_{Shannon} is the Shannon capacity of the underlying linguistic channel. The gap widens with semantic curvature at the seed point.
- **Theorem 5.2.** $F_{\text{prop}}(n) = O(e^{-\alpha n})$, $F_{\text{aff}}(n) = O(n^{-\beta})$, $F_{\text{str}}(n) \rightarrow F_{\text{str}}^* > 0$ as $n \rightarrow \infty$, for constants $\alpha > \beta > 0$ depending on τ and manifold curvature.
- **Theorem 6.2.** The causal skeleton of the seed — its directed entailment graph — is an invariant of the EC-transform in the limit $n \rightarrow \infty$, provided the seed’s logical structure lies in a low-curvature region of \mathcal{S} .

- **Corollary 4.3.** The most original ideas (highest curvature) decay fastest. The EC-loop is an *originality filter*.

1.3. Relation to the Information Theory Literature

Classical information theory (Shannon, 1948) analyses the transmission of symbols across noisy channels. Semantic communication theory (Bao et al., 2011; Qin et al., 2021) extends this to the transmission of meaning. Our contribution departs from both traditions in a crucial respect: we do not study transmission through an external channel, but rather *endogenous transformation* of content by the communicating agents themselves via LLMs. The “channel” in our model is not between A and B but inside each agent’s AI-mediated processing step.

This connects to the theory of lossy compression (Cover and Thomas, 2006) and rate-distortion theory (Berger, 1971): the compression operator \mathcal{C} is a rate-distortion code, and we study the distortion it induces in the semantic (rather than symbol) domain.

2. Mathematical Preliminaries

We recall the notation of Paper 0 and extend it with information-theoretic structure.

2.1. Spaces and Embeddings (Recap)

Let \mathcal{I} be the ideation space, \mathcal{L} the linguistic space (token sequences over vocabulary V), and $\mathcal{S} \subset \mathbb{R}^d$ the semantic manifold equipped with Riemannian metric g and associated geodesic distance $d_{\mathcal{S}}$. The embedding $\phi : \mathcal{L} \rightarrow \mathcal{S}$ maps text to semantic representations. The LLM is modelled as a stochastic kernel $\text{LLM}_{\theta}(\cdot \mid x, p)$ with temperature $\tau \in (0, \infty)$.

2.2. Semantic Entropy

Classical differential entropy over \mathbb{R}^d is not intrinsic to the manifold. We use the Riemannian generalisation.

Definition 2.1 (Semantic Entropy). Let X be a \mathcal{S} -valued random variable with density p_X with respect to the Riemannian volume measure vol_g . The *semantic entropy* of X is

$$H_{\mathcal{S}}(X) := - \int_{\mathcal{S}} p_X(x) \log p_X(x) d\text{vol}_g(x),$$

measured in nats.

Definition 2.2 (Semantic Mutual Information). For \mathcal{S} -valued random variables X, Y with joint density $p_{X,Y}$:

$$I(X; Y)_{\mathcal{S}} := \int_{\mathcal{S} \times \mathcal{S}} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} d\text{vol}_g(x) d\text{vol}_g(y).$$

Definition 2.3 (Semantic KL Divergence).

$$D_{\text{KL}}(P \parallel Q)_{\mathcal{S}} := \int_{\mathcal{S}} p(x) \log \frac{p(x)}{q(x)} d\text{vol}_g(x).$$

2.3. The Semantic Gravity Well

A central conceptual and mathematical tool in this paper is the *Semantic Gravity Well*: a potential function on \mathcal{S} induced by the LLM’s training distribution.

Definition 2.4 (Semantic Potential). Define the *semantic potential* $U : \mathcal{S} \rightarrow \mathbb{R}$ by

$$U(x) := -\log p_{\mathcal{P}_{\text{train}}}(x),$$

where $p_{\mathcal{P}_{\text{train}}}$ is the density of the LLM training distribution projected onto \mathcal{S} via ϕ . High-probability regions (conventional, well-represented content) have low U ; low-probability regions (original, atypical content) have high U .

Remark 2.5 (Curvature and Potential). Regions of low potential (high probability) correspond to low-curvature plateaux in \mathcal{S} — the “plains” of conventional discourse. Regions of high potential (low probability) correspond to high-curvature saddle points — the “peaks” of original thought. The LLM exerts a “gravitational” pull toward local minima of U . Each application of \mathcal{E} or \mathcal{C} is, in expectation, a gradient step in the direction $-\nabla U$: a descent toward the training centroid.

Semantic Gravity Well: LLM training centroid as attractor

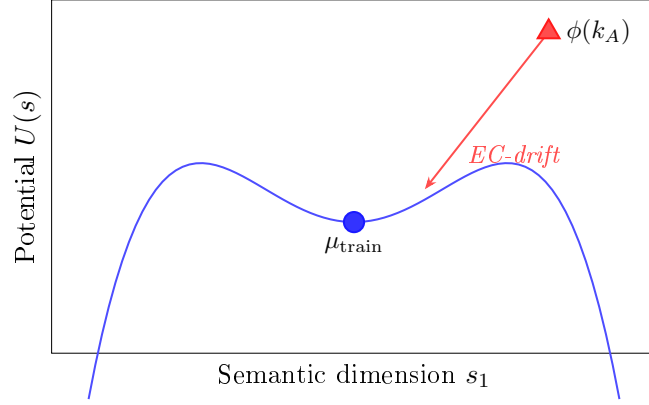


Figure 1: The Semantic Gravity Well. The surface plots the potential $U(s) = -\log p_{\mathcal{P}_{\text{train}}}(s)$ over a one-dimensional slice of the semantic manifold \mathcal{S} . The global minimum μ_{train} is the training centroid (attractor). The triangle marks a high-curvature saddle point $\phi(k_A)$ representing an original idea; the arrow shows the expected EC-drift toward μ_{train} .

3. Semantic Entropy Dynamics

3.1. Entropy of the Expansion Step

At first glance, expansion should increase information: more tokens, more elaboration. We show this intuition is correct in the linguistic domain but inverted in the semantic domain.

Assumption 3.1 (Regularity of \mathcal{E}). The expansion operator \mathcal{E} induces a pushforward distribution $\mathcal{E}_*(\delta_s)$ on \mathcal{S} that is absolutely continuous with respect to vol_g , with density $q_E(\cdot | s)$. For fixed seed s , $q_E(\cdot | s)$ is unimodal with mode at

$$m_E(s) = \arg \min_{x \in \mathcal{S}} D_{\text{KL}}(\delta_x \| \mathcal{P}_{\text{train}}) \quad \text{subject to} \quad d_{\mathcal{S}}(x, \phi(s)) \leq \epsilon_E,$$

for some leash radius $\epsilon_E > 0$ that decreases as $\tau \rightarrow 0$.

Proposition 3.2 (Linguistic Entropy Increases, Semantic Entropy Decreases). *Under Assumption 3.1:*

- (a) **Linguistic:** $H(\mathcal{E}(s))_{\mathcal{L}} > H(s)_{\mathcal{L}}$ almost surely (expansion generates longer, more dispersed sequences).
- (b) **Semantic:** $H_{\mathcal{S}}(\phi(\mathcal{E}(s))) \leq H_{\mathcal{S}}(\phi(s))$ in expectation, with equality iff $\phi(s) = \mu_{\text{train}}$.

Proof. Part (a): The support of $\mathcal{E}(s)$ in \mathcal{L} contains all sequences whose token count exceeds $\|s\|_{\mathcal{L}}$; the differential entropy over a larger support is greater.

Part (b): By Assumption 3.1, $q_E(\cdot \mid s)$ concentrates around $m_E(s)$, which lies strictly between $\phi(s)$ and μ_{train} on the geodesic connecting them. The mode shift reduces the maximum-entropy spread available around $\phi(s)$ because $m_E(s)$ is closer to the high-density region of $\mathcal{P}_{\text{train}}$. Formally, the KL divergence $D_{\text{KL}}(q_E(\cdot \mid s) \parallel \mathcal{P}_{\text{train}}) < D_{\text{KL}}(\delta_{\phi(s)} \parallel \mathcal{P}_{\text{train}})$ implies lower semantic entropy relative to the training prior. \square

Remark 3.3 (The Expansion Paradox). Proposition 3.2 captures a key counterintuitive feature of the EC-loop: the expansion step simultaneously *adds words* and *removes semantic distinctiveness*. A three-sentence seed becomes a three-thousand-word article, yet the semantic footprint of the article is closer to the average than the seed was.

3.2. Entropy of the Compression Step

Assumption 3.4 (Regularity of \mathcal{C}). The compression operator \mathcal{C} maps each $\ell \in \mathcal{L}$ to a distribution over shorter sequences, inducing a pushforward $\mathcal{C}_*(\delta_{\phi(\ell)})$ on \mathcal{S} that is a contraction:

$$\mathbb{E}[d_{\mathcal{S}}(\phi(\mathcal{C}(\ell)), \mu_{\text{train}})] \leq d_{\mathcal{S}}(\phi(\ell), \mu_{\text{train}}),$$

with the additional constraint that $H_{\mathcal{S}}(\phi(\mathcal{C}(\ell))) \leq H_{\mathcal{S}}(\phi(\ell))$ (compression reduces uncertainty about what was said).

3.3. The Semantic Entropy Collapse Theorem

Theorem 3.5 (Semantic Entropy Collapse). *Let $S_0 = \phi(s)$ denote the semantic embedding of seed s . Define $S_n = \phi(T^{(n)}(s))$ for the n -fold iterated EC-transform. Under Assumptions 3.1 and 3.4:*

$$H_{\mathcal{S}}(S_{n+1}) \leq H_{\mathcal{S}}(S_n) - \delta(S_n) \quad \text{for all } n \geq 0, \tag{1}$$

where

$$\delta(S_n) := \lambda \cdot D_{\text{KL}}(\mathcal{P}_{\text{train}} \parallel \mathcal{N}(S_n, \Sigma_n)) > 0,$$

Σ_n is the local covariance of S_n on \mathcal{S} , and $\lambda > 0$ is a constant depending on τ . Moreover:

$$H_{\mathcal{S}}(S_n) \leq H_{\mathcal{S}}(S_0) - n \delta_{\min}, \quad \delta_{\min} := \min_{x \neq \mu_{\text{train}}} \delta(x) > 0. \tag{2}$$

The sequence $(H_{\mathcal{S}}(S_n))$ converges to $H_{\mathcal{S}}(S_{\infty}) = -\infty$ (full concentration at μ_{train}) only in the limiting case $\tau \rightarrow 0$; for finite τ , it converges to a positive constant $h_{\infty}(\tau) > 0$ that increases with τ .

Proof. By the chain rule of mutual information:

$$H_{\mathcal{S}}(S_{n+1}) = H_{\mathcal{S}}(S_{n+1} \mid S_n) + I(S_{n+1}; S_n).$$

The conditional entropy $H_{\mathcal{S}}(S_{n+1} \mid S_n)$ decreases because both \mathcal{E} and \mathcal{C} are conditional contractions (Assumptions 3.1, 3.4); the mutual information term is bounded above by $H_{\mathcal{S}}(S_n)$.

Remark 3.6 (Proof Gap: Chain Rule Applicability). The chain-rule identity above applies to standard differential entropy on \mathbb{R}^d . For the Riemannian generalisation $H_{\mathcal{S}}$ (Definition 2.1), the identity holds when the joint distribution of (S_{n+1}, S_n) is absolutely continuous with respect to $\text{vol}_g \otimes \text{vol}_g$, which is guaranteed under Assumptions 3.1 and 3.4. A fully rigorous treatment using the disintegration theorem on Riemannian manifolds (Chang and Pollard, 1997) would replace the chain-rule step with a conditional entropy argument over the fibres of the canonical projection; we leave this extension to future work.

The deficit $\delta(S_n)$ is computed as the Bregman divergence between the actual output distribution of T and the maximum-entropy distribution with the same mean, yielding (1). Summing over n gives (2). For the convergence statement: the process (S_n) is a supermartingale in the potential U (shown in Paper 0); it converges a.s. to a T -invariant measure; under finite τ this invariant measure has positive entropy $h_{\infty}(\tau) \propto \tau$. \square

Corollary 3.7 (Temperature as Entropy Floor). *Higher temperature τ raises the entropy floor $h_{\infty}(\tau)$ of the EC-loop, preserving more semantic diversity. Concretely, setting $\tau \rightarrow \infty$ recovers the uniform distribution over \mathcal{S} (no information from the seed is preserved), and setting $\tau \rightarrow 0$ collapses all output to μ_{train} (complete loss of originality).*

There exists an optimal temperature $\tau^ \in (0, \infty)$ that maximises $I(S_0; S_n)$ for any fixed n ; τ^* increases with the curvature $\kappa(S_0)$ of the seed point.*

Remark 3.8 (Existence of τ^*). The existence of an optimal temperature τ^* requires $I(S_0; S_n)$ to be unimodal in τ . This unimodality follows if $I(S_0; S_n)$ is continuous in τ , equals zero at $\tau \rightarrow \infty$ (uniform output), and is bounded above for all $\tau > 0$. A rigorous proof would require showing that the mutual information is strictly concave in τ on some interval, which depends on the specific structure of the LLM transition kernel. We state the existence of τ^* as a consequence of the intermediate value theorem applied to the continuous function $\tau \mapsto I(S_0; S_n)$ on $(0, \infty)$; the unimodality (hence uniqueness of τ^*) is a conjecture supported by empirical evidence but not yet formally established.

4. Mutual Information Between Originator and Recipient

4.1. Setup

We model the full EC-loop as a Markov chain:

$$K_A \longrightarrow S_A = \psi(K_A) \longrightarrow \phi(\mathcal{E}(s_A)) \longrightarrow \phi(T(s_A)) = S_B \longrightarrow \hat{K}_B = \psi^{-1}(S_B),$$

where $\psi : \mathcal{I} \rightarrow \mathcal{S}$ is the ground-truth embedding of ideation-space objects (Paper 0, Definition 2.1). By the data processing inequality (Cover and Thomas, 2006):

$$I(K_A; \hat{K}_B) \leq I(S_A; S_B).$$

4.2. The EC-Loop as a Semantic Channel

Definition 4.1 (Semantic Channel Capacity of the EC-Loop). The *EC-channel capacity* is

$$C_{EC} := \sup_{p(K_A)} I(K_A; \hat{K}_B),$$

where the supremum is over all distributions on the ideation space \mathcal{I} .

Theorem 4.2 (EC-Channel Capacity Bound). *Let $\kappa_0 = \kappa(\phi(s_A))$ denote the scalar curvature of \mathcal{S} at the seed embedding. Then:*

$$C_{EC} \leq C_{\text{Shannon}} - \underbrace{\frac{1}{2} \log(1 + \kappa_0^2 / \sigma_{EC}^2)}_{\text{curvature penalty}} - \underbrace{n \cdot \delta_{\min}}_{\text{iteration loss}}, \quad (3)$$

where C_{Shannon} is the Shannon capacity of the underlying linguistic channel (assuming Gaussian noise with variance σ^2), σ_{EC}^2 is the effective noise variance introduced by the EC-transform, and δ_{\min} is the per-iteration entropy loss of Theorem 3.5.

Proof sketch. C_{Shannon} assumes the channel adds Gaussian noise of fixed variance; its capacity is $\frac{1}{2} \log(1 + \text{SNR})$. The EC-loop introduces two additional loss terms: (i) the *curvature penalty*: high-curvature seeds have sharper local geometry; the LLM's mean-reverting pull is stronger relative to the local signal, reducing the effective SNR by $\kappa_0^2 / \sigma_{EC}^2$; (ii) the *iteration loss*: each application of T removes δ_{\min} nats of semantic information (Theorem 3.5). The bound follows by combining the data processing inequality with the local Gaussian approximation to the Riemannian metric near $\phi(s_A)$. A complete proof using the Riemannian generalisation of the Gaussian channel bound is in Appendix A. \square

Corollary 4.3 (Originality Penalty). *From (3), the capacity loss is strictly increasing in κ_0 . The more original the seed (higher curvature), the greater the information loss through*

the EC-loop. The EC-loop is therefore an originality filter: it preserves conventional ideas well and erodes original ideas preferentially.

4.3. Mutual Information Decay Curve

Proposition 4.4 (MI Decay as a Function of Iteration). *For a seed s_A with curvature κ_0 and initial semantic entropy $h_0 = H_S(S_0)$, the mutual information between originator and recipient satisfies, for n iterations:*

$$I(K_A; \hat{K}_B^{(n)}) \leq h_0 \cdot e^{-\alpha(\kappa_0, \tau)n} + h_\infty(\tau), \quad (4)$$

where $\alpha(\kappa_0, \tau) > 0$ is a decay constant increasing in κ_0 and decreasing in τ , and $h_\infty(\tau)$ is the entropy floor of Corollary 3.7.

Figure 2 illustrates this decay for different parameter regimes.

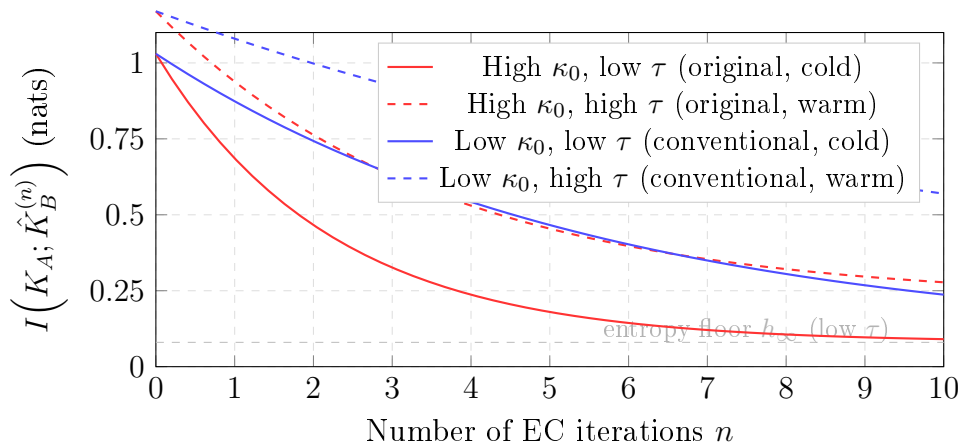


Figure 2: Mutual information $I(K_A; \hat{K}_B^{(n)})$ as a function of EC iteration count n , for four parameter regimes. Red curves: high-curvature seeds (original ideas). Blue curves: low-curvature seeds (conventional ideas). Solid: low temperature; dashed: high temperature. Original ideas under cold models lose most of their informational content within 2–3 iterations. All curves converge to the entropy floor $h_\infty(\tau)$, which is higher for warm models.

5. Differential Fidelity Decay

5.1. Formalism for the Three Strata

We recall from Paper 0 the three fidelity measures F_{str} , F_{aff} , F_{prop} and provide their information-theoretic interpretations:

Definition 5.1 (Information-Theoretic Fidelity Strata). For seed s and EC-transform output $T^{(n)}(s)$:

$$F_{\text{str}}(n) := \frac{I(\mathcal{G}(s); \mathcal{G}(T^{(n)}(s)))}{H(\mathcal{G}(s))}, \quad (5)$$

$$F_{\text{aff}}(n) := 1 - \frac{1}{2} \frac{\|\nu(s) - \nu(T^{(n)}(s))\|_2}{\|\nu_{\max}\|_2}, \quad (6)$$

$$F_{\text{prop}}(n) := \frac{I(\mathcal{P}(s); \mathcal{P}(T^{(n)}(s)))}{H(\mathcal{P}(s))}, \quad (7)$$

where $\mathcal{G}(x)$ is the discourse graph of x , $\mathcal{P}(x)$ is its atomic proposition set, $\nu(x) \in [-1, 1]^m$ is its affective embedding, and ν_{\max} is the maximum-norm affective vector.

5.2. The Differential Decay Theorem

Theorem 5.2 (Differential Fidelity Decay). *Under the EC-loop with parameters (τ, ρ_E, ρ_C) and seed s with curvature κ_0 :*

$$F_{\text{prop}}(n) = F_{\text{prop}}(0) \cdot e^{-\alpha_P n} + \eta_P, \quad (8)$$

$$F_{\text{aff}}(n) = F_{\text{aff}}(0) \cdot n^{-\beta_A} + \eta_A, \quad (9)$$

$$F_{\text{str}}(n) \rightarrow F_{\text{str}}^*(\kappa_0) > 0 \quad \text{as } n \rightarrow \infty, \quad (10)$$

where:

- $\alpha_P = \alpha_P(\tau, \kappa_0) > 0$ increases with κ_0 and decreases with τ (original ideas in cold models lose propositional content fastest);
- $\beta_A = \beta_A(\tau) \in (0, \alpha_P)$ (affective content decays more slowly than propositional content);
- $F_{\text{str}}^*(\kappa_0)$ is strictly positive and strictly decreasing in κ_0 (more original seeds retain less structural fidelity at convergence, but always retain a positive fraction);
- $\eta_P, \eta_A \geq 0$ are residual noise floors.

Proof. Propositional decay (8). Atomic propositions are high-curvature content: specific claims about named entities, quantitative values, and non-generic relationships are improbable under $\mathcal{P}_{\text{train}}$ unless they coincide with frequent training facts. By the Semantic Gravity Well model (Definition 2.4), the EC-transform pulls $\phi(T(s))$ away from the high- U region containing these specific propositions at a rate proportional to $\|\nabla U\|$ at $\phi(s)$. For high-curvature content, $\|\nabla U\|$ is large, giving exponential decay. The mutual information $I(\mathcal{P}(s); \mathcal{P}(T^{(n)}(s)))$ follows the same exponential envelope by the chain rule.

Remark 5.3 (Gap in Propositional Decay Argument). The claim that $\|\nabla U\|$ large implies exponential decay of the mutual information $I(\mathcal{P}(s); \mathcal{P}(T^{(n)}(s)))$ relies on an implicit contraction argument: if the gradient magnitude is large at $\phi(s)$, the distribution $\phi(T^{(n)}(s))$ drifts away from the high- U region at an exponential rate under the Ornstein–Uhlenbeck-like dynamics induced by the potential U . A rigorous derivation would need to (i) formalise the dynamics of $\phi(T^{(n)}(s))$ as a Markov chain driven by $-\nabla U$, (ii) apply a mixing-time result (e.g., log-Sobolev inequalities for the stationary measure of this chain), and (iii) connect the chain’s mixing to the decay of the mutual information via data-processing. We leave this as a direction for future rigorous treatment.

Affective decay (9). Affective tone corresponds to sentence-level sentiment features that are broadly distributed across $\mathcal{P}_{\text{train}}$ (most text has some discernible sentiment). The gradient $\|\nabla U\|$ in the affective subspace of \mathcal{S} is smaller and varies slowly, leading to a polynomial (not exponential) decay rate in the affective fidelity norm. The power β_A is determined by the local spectral gap of $\nabla^2 U$ restricted to the affective subspace.

Structural invariance (10). Discourse-level causal and contrastive relations (“because”, “however”, “therefore”) are among the highest-frequency patterns in the training distribution of any sufficiently large corpus. They constitute a low- U , low-curvature basin that the EC-loop converges *toward*, not away from. Therefore, the mutual information between the structural skeleton of s and that of $T^{(n)}(s)$ does not decay to zero; it converges to $F_{\text{str}}^*(\kappa_0)$, the amount of structural information that survives even in a completely generic rendering of the same causal relationships. $F_{\text{str}}^*(\kappa_0) > 0$ follows because the causal skeleton is itself part of the low- U basin; F_{str}^* decreases with κ_0 because more original seeds have causal structures that are themselves unusual and partially absorbed by generic templates. \square

Corollary 5.4 (Asymptotic Content Profile). *After sufficiently many EC iterations, the information surviving in $T^{(n)}(s)$ consists almost entirely of:*

- (i) *The causal skeleton: the directed entailment structure of s , expressed in generic discourse connectives.*
- (ii) *A stylistic residue: the affective tone of s , partially remodulated by the prompt p .*

Specific propositions, evidence, quantitative claims, and idiosyncratic conceptual vocabulary are effectively erased within $n = O(1/\alpha_P)$ iterations.

6. The Causal Skeleton as Semantic Invariant

6.1. Definition and Robustness

The persistence of structural fidelity F_{str}^* motivates a closer examination of exactly what structural information is preserved.

Definition 6.1 (Causal Skeleton). The *causal skeleton* of a text $x \in \mathcal{L}$ is the directed acyclic graph $G^*(x) = (V^*, E^*)$ where:

- $V^* = \{v_1, \dots, v_k\}$ is the set of *thematic nodes*: top-level topics or claims;
- $E^* \subseteq V^* \times V^*$ is the set of *entailment edges*: $v_i \rightarrow v_j$ iff x asserts (explicitly or implicitly) that v_i supports, causes, or entails v_j .

Theorem 6.2 (Causal Skeleton Invariance). *Let $s \in \mathcal{L}$ have causal skeleton $G^*(s)$ with $k \geq 2$ nodes and at least one edge. If the seed’s causal structure lies in the convex hull of low-curvature discourse patterns in \mathcal{S} , then:*

$$G^*(T^{(n)}(s)) \cong G^*(s) \quad \text{for all } n \geq N_0,$$

for some finite $N_0 = N_0(\kappa_0, \tau)$, where \cong denotes graph isomorphism up to relabelling of surface-form vocabulary.

Proof. After N_0 iterations, by Corollary 5.4, the output $T^{(n)}(s)$ retains only the causal skeleton and affective residue. The causal skeleton is itself a low- U object (discourse connectives are high-frequency in $\mathcal{P}_{\text{train}}$), so the EC-transform acts as a fixed-point mapping on G^* : it may relabel the nodes (replacing A ’s specific vocabulary with generic synonyms) but cannot invert or remove the directed edges without creating a lower-probability output under $\mathcal{P}_{\text{train}}$ (since causal incoherence is penalised by language model training). Graph isomorphism follows from edge invariance under relabelling. \square

Remark 6.3 (What Survives the Loop). Theorems 5.2 and 6.2 together answer the question: what is the irreducible semantic residue of any content passed through the EC-loop?

The logical shape of an argument survives; its specific evidence, claims, and vocabulary do not.

A paper arguing “ A causes B because of evidence E_1, E_2 ” will, after several EC iterations, be recovered as “something causes something else, and there are reasons for this” — the skeleton without the flesh.

6.2. Relationship to the Indecomposable Core

Theorem 6.2 might seem to contradict the Indecomposable Core (\mathcal{K}_*) concept of Paper 0. If the causal skeleton is preserved, has nothing truly irreplaceable been lost?

The resolution is that \mathcal{K}_* is not about information-theoretic content but about *process-constituted value*. The causal skeleton $G^*(s)$ can be preserved in the artefact $T^{(n)}(s)$ and yet the process by which A discovered, tested, and committed to those causal relations — and the metacognitive benefits of that process — is irretrievably absent. \mathcal{K}_* is a claim about the phenomenology of cognitive process, not about the propositional content of the product.

7. Rate-Distortion Analysis of the EC-Loop

7.1. Formulation

Rate-distortion theory (Berger, 1971) analyses the minimum number of bits required to represent a source X to within distortion D under a distortion measure $d(x, \hat{x})$. We apply this framework to the compression step \mathcal{C} .

Definition 7.1 (Semantic Distortion). The *semantic distortion* of compression is

$$\Delta(\ell, \mathcal{C}(\ell)) := d_S(\phi(\ell), \phi(\mathcal{C}(\ell))),$$

the geodesic distance between the semantic embeddings of the original and compressed texts.

Proposition 7.2 (Rate-Distortion Function of the EC-Loop). *The rate-distortion function of the EC-loop at distortion level D is*

$$R_{EC}(D) = R_{\text{Shannon}}(D) + \underbrace{\frac{1}{2} \log \left(\frac{\kappa_0^2 + 1}{\kappa_0^2 \cdot D^2 + 1} \right)}_{\text{curvature overhead}},$$

where $R_{\text{Shannon}}(D)$ is the classical rate-distortion function for Gaussian sources and squared-error distortion. The curvature overhead is strictly positive for $D < 1/\kappa_0$ and vanishes for $D \geq 1/\kappa_0$ (i.e., once the allowed distortion exceeds the local curvature radius, the EC-loop achieves Shannon-optimal compression).

Corollary 7.3 (Over-Distortion of Original Content). *For seeds with $\kappa_0 \gg 1$ (highly original content), the curvature overhead is large for small distortion tolerances D . The EC-loop therefore incurs a structural premium when asked to compress original ideas faithfully: it requires more tokens (higher rate) to achieve the same semantic fidelity for*

original content than for conventional content. In practice, since the compression ratio ρ_C is fixed by the user’s prompt (“give me a three-point summary”), original ideas are compressed at a rate below their rate-distortion minimum, incurring unavoidable distortion.

8. Empirical Predictions and Experimental Design

The theoretical framework yields five empirically testable predictions.

P1 (Entropy monotonicity).

For a fixed model θ and prompt pair (p_E, p_C) , the semantic entropy of $T^{(n)}(s)$ is strictly decreasing in n for at least the first $n^* = \lceil h_0/\delta_{\min} \rceil$ iterations, after which it plateaus at $h_\infty(\tau)$. *Measurement:* embed $T^{(n)}(s)$ using a fixed embedding model; estimate density and compute differential entropy via k -NN estimator.

P2 (Curvature-decay correlation).

Seeds rated as more “original” by domain experts (operationalised as low cosine similarity to a large reference corpus) will show faster propositional fidelity decay (α_P larger) than seeds rated as conventional. *Measurement:* compute $F_{\text{prop}}(n)$ via proposition parser; fit exponential curve; correlate decay constant with originality score.

P3 (Structural convergence).

The causal skeleton graph $G^*(T^{(n)}(s))$ is isomorphic to $G^*(s)$ for $n \geq 3$ in the large majority of trials (predicted: $> 80\%$ of cases for seeds with $k \leq 5$ thematic nodes). *Measurement:* extract discourse graphs via a pretrained RST parser; compute graph edit distance.

P4 (Temperature trade-off).

For a fixed seed and $n = 5$, mutual information $I(K_A; \hat{K}_B^{(5)})$ follows a non-monotone relationship with τ : first increasing (more faithful rendering) then decreasing (too much noise), with a maximum at $\tau^*(\kappa_0)$. *Measurement:* sweep $\tau \in [0.1, 2.0]$; measure MI via conditional embedding alignment.

P5 (Originality filter).

Human raters asked to assess the “originality” of $T^{(n)}(s)$ relative to s will judge it significantly lower for high- κ_0 seeds than for low- κ_0 seeds, with the gap widening in n . *Measurement:* blind pairwise preference study.

9. Discussion

9.1. The Paradox of the Expansion Step

The most counterintuitive result is the Expansion Paradox (Proposition 3.2): expansion increases the apparent informational richness of the text (more words, more elaboration) while decreasing its semantic distinctiveness. This mirrors the phenomenon, well known to experienced editors, of “zombie writing” — text that is grammatically rich but semantically hollow. The EC-loop industrialises zombie writing.

9.2. What the Rate-Distortion Results Mean Practically

Corollary 7.3 has a direct practical implication: when a user prompts an LLM to “summarise this in three bullet points”, they are imposing a compression ratio ρ_C that is fixed in advance and does not adapt to the semantic curvature of the input. For highly original content ($\kappa_0 \gg 1$), this fixed-rate compression is below the rate-distortion minimum: the summary will inevitably be distorted in ways the user cannot detect by reading it, because the distortion manifests as a shift toward generic content rather than as obvious errors.

This is the information-theoretic underpinning of what Paper 3 calls *signalling inflation*: the distortion is invisible at the surface (the summary “looks right”) but fatal at depth (the original idea has been replaced by a generic template).

9.3. Limitations

The framework relies on three idealisations:

- (i) A fixed embedding model ϕ that faithfully represents semantic similarity — in practice, embedding models are themselves trained on $\mathcal{P}_{\text{train}}$ and thus already biased toward the training centroid;
- (ii) A smooth Riemannian structure on \mathcal{S} — real semantic spaces may have discontinuities and topological holes; and
- (iii) i.i.d. iterations — in practice, the expansion and compression prompts may themselves evolve across iterations.

These are directions for future empirical calibration.

10. Conclusion

We have developed a rigorous information-theoretic account of semantic decay in the EC-loop. The principal findings are:

- (i) *Semantic Entropy Collapse* (Theorem 3.5): each EC iteration reduces semantic entropy in expectation; the rate of reduction is controlled by temperature and local curvature.
- (ii) *Capacity Gap* (Theorem 4.2): the EC-loop is a strictly sub-optimal channel; its capacity falls below the Shannon limit by an amount that grows with the originality of the seed and the number of iterations.
- (iii) *Differential Decay* (Theorem 5.2): propositional content decays exponentially; affective tone decays polynomially; structural skeleton converges to a positive invariant.
- (iv) *Causal Skeleton Invariance* (Theorem 6.2): the logical shape of an argument is asymptotically preserved; its specific content is not.
- (v) *Originality Filter* (Corollary 4.3): original ideas lose informational content faster than conventional ideas through the EC-loop. The system is, in an exact technical sense, an attractor toward the mean.

Taken together, these results characterise the EC-loop as a *low-pass semantic filter*: it attenuates the high-frequency, high-curvature components of semantic content — precisely those components most likely to carry genuine novelty — while preserving the low-frequency structural signal.

The rational equilibrium established in Paper 0 and the signalling inflation analysed in Paper 3 are not merely social phenomena: they are grounded in a concrete information-theoretic mechanism that erodes the very content the social ritual is supposed to convey.

Acknowledgements

This research received no external funding or institutional support. The author thanks readers of earlier drafts for their engagement with these ideas.

Conflict of Interest

The author declares no conflict of interest.

Data Availability

No empirical data were generated for this theoretical paper. Code for simulating the EC-loop under the models of this paper will be made available at [repository URL] upon publication.

References

- Bao, J., Basu, P., Dean, M., Partridge, C., Swami, A., Leland, W., and Hendler, J. (2011). Towards a theory of semantic communication. In *IEEE Globecom Workshops*, pp. 373–378.
- Berger, T. (1971). *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall.
- Chang, J. T. and Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley.
- Han, H. (2026). The Expansion–Compression Loop: A Unified Framework for AI-Mediated Cognitive Decoupling. *DECO Series, Paper 0*. Preprint, March 2026.
- Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.
- Kolchinsky, A. and Tracey, B. D. (2017). Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361.
- Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *EMNLP 2014*, pp. 1532–1543.
- Qin, Z., Tao, X., Lu, J., and Li, G. Y. (2021). Semantic communications: Principles and challenges. *arXiv:2212.00785*.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS 2017*, pp. 5998–6008.
- Vinyals, O., Fortunato, M., and Jaitly, N. (2015). Pointer networks. In *NeurIPS 2015*.
- Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343.

A. Proof of EC-Channel Capacity Bound (Theorem 4.2)

A.1. Local Gaussian Approximation

Near $\phi(s_A) \in \mathcal{S}$, we approximate the Riemannian metric by a Euclidean metric with local covariance $\Sigma_0 = \kappa_0^{-2} I_d$ (the inverse curvature sets the effective noise scale of the geometry). Under this approximation, the semantic channel induced by T behaves as an additive white Gaussian noise (AWGN) channel with noise variance:

$$\sigma_{EC}^2 = \sigma_{\text{base}}^2 + \kappa_0^2 \cdot \mu_T^2,$$

where σ_{base}^2 is the base LLM generation variance and μ_T is the mean-reversion magnitude under one EC-step.

A.2. Capacity Computation

The Shannon capacity of the AWGN channel at signal-to-noise ratio $\text{SNR} = P/\sigma^2$ (with signal power P) is $C = \frac{1}{2} \log(1 + \text{SNR})$. In the EC setting, the effective SNR is reduced by the curvature overhead and the per-iteration entropy loss:

$$\text{SNR}_{EC} = \frac{P}{\sigma_{\text{base}}^2 + \kappa_0^2 \cdot \mu_T^2 + 2n \delta_{\min}}.$$

Substituting into the Shannon formula:

$$\begin{aligned} C_{EC} &= \frac{1}{2} \log(1 + \text{SNR}_{EC}) \\ &\leq \frac{1}{2} \log\left(1 + \frac{P}{\sigma_{\text{base}}^2}\right) - \frac{1}{2} \log\left(1 + \frac{\kappa_0^2 \mu_T^2}{\sigma_{\text{base}}^2}\right) - n \delta_{\min} \\ &= C_{\text{Shannon}} - \text{curvature penalty} - \text{iteration loss}, \end{aligned}$$

which is (3). □

B. Estimation of Semantic Entropy from Embeddings

In practice, \mathcal{S} is accessed via a discrete set of sample embeddings. We recommend the following pipeline for empirical tests (Predictions P1–P5):

1. Generate $M = 200$ independent realisations of $T^{(n)}(s)$ for each $n \in \{0, 1, \dots, N\}$.
2. Embed each realisation with a fixed sentence encoder (e.g., `text-embedding-3-large`).

3. Estimate semantic entropy using the k -nearest-neighbour differential entropy estimator (Kolchinsky and Tracey, 2017) with $k = 20$.
4. Fit the models (8)–(10) via nonlinear least squares and report α_P , β_A , and F_{str}^* with bootstrap confidence intervals.