Information-Geometric Context Window Governance and the Probabilistic Theory of Long-Context Collapse in Large Language Models

A Unified Framework: Observer Entropy, Extreme Value Theory, and the CPL 4.0 Phase-Aware Governor

doi:10.5281/zenodo.19352770

Vladimir Khomyakov*
Independent Researcher
orcid.org/0009-0006-3074-9145

March 31, 2026

Abstract

We present a unified theoretical framework connecting two complementary approaches to long-context degradation in Transformer-based large language models (LLMs): the information-geometric theory of observer entropy and the probabilistic theory of attention collapse via Extreme Value Theory (EVT). The central object is the *observer entropy* $S_{\text{obs}}(p_\theta, \varepsilon) = D_{\text{KL}}(p_\theta \| \widetilde{\Pi_\varepsilon p_\theta})$, whose quadratic scaling law—the Bridge Theorem—states $S_{\text{obs}} = \frac{1}{2}\varepsilon^2 v(\theta)^\top I(\theta) v(\theta) + O(\varepsilon^3)$, where $I(\theta)$ is the Fisher information matrix. We interpretatively connect this quantity to the signal-noise competition in self-attention: defining signal strength $\mu_s = \frac{1}{\sqrt{d}}\text{Tr}(W_Q^\top W_K \Sigma_{qr})$ and effective margin $\mu_L = \mu_s - \sigma\sqrt{2\log L_{\text{eff}}}$, we prove a two-sided bound $c_1 \mu_L e^{\mu_L}/L \le S_{\text{obs}}(L) \le c_2 e^{\mu_L}/L$ valid in the pre-collapse regime $\mu_L > \log 2 + 1$, and a one-sided upper bound $S_{\text{obs}}(L) \le c_2 e^{\mu_L}/L$ for all sufficiently large L; the formal equivalence between the partition-based and attention-uniform definitions of S_{obs} is an interpretative identification whose rigorous formalisation is an open problem. This yields a *Fundamental Impossibility Theorem*: for any finite μ_s, observer entropy decays to zero as $L \to \infty$, establishing long-context collapse as an information-theoretic inevitability under softmax attention. Using Gumbel convergence for weakly dependent logit maxima with Gaussian marginals (Leadbetter conditions), we derive a closed-form *Probabilistic Risk Law*: $P(\mathcal{F}_L) \approx 1 - \exp(-\exp(-(\mu_s - \sigma\sqrt{2\log L})/a_L))$, and conjecture that the critical length L_{crit} is a heavy-tailed random variable (formal proof incomplete; see Remark 4.5). The CPL 4.0 phase-aware governor emerges as the principled control-theoretic response, with a Master Theorem guaranteeing a hard context cap, entropy contraction, and a sub-linear fragmentation bound $N_F(T) = O(\sqrt{T\log(1/\delta_0)})$. Numerical verification confirms the ε^2 scaling law across both worked examples.

*Companion manuscripts:
KL-Geometric Structure of Observer Entropy [1] (doi:10.5281/zenodo.19202244)
and *Information-Geometric Context Window Governance* [2] (doi:10.5281/zenodo.19177363).

Contents

1 Introduction

Long-context capabilities have become a central axis of LLM development, with deployed systems now advertising context windows of 128k to over 1M tokens. Yet substantial empirical evidence documents a persistent and often catastrophic degradation of model quality at long contexts—degradation that persists even when relevant information is explicitly provided. This paper argues that such degradation is not an engineering artefact to be eliminated by clever architecture choices, but a *fundamental information-theoretic phenomenon*: the inevitable consequence of finite-signal dynamics under softmax attention in high-dimensional spaces.

Empirical motivation. Liu et al. [3] documented the "lost-in-the-middle" effect: information retrieval accuracy degrades steeply for tokens not appearing near the context boundary. Du et al. [4] demonstrated that performance degrades even when all relevant documents are supplied and retrieved, implying an intrinsic architectural limitation. Attention entropy studies [6, 7] show that the entropy of attention weight distributions grows approximately as $\log L$, converging toward uniform allocation and destroying the model's ability to selectively attend to relevant tokens. Wang et al. [5] document catastrophic collapse as a threshold phenomenon, with accuracy remaining stable up to a critical length and then dropping abruptly by tens of percentage points—a signature of extreme-value competition.

Contributions. This paper makes four main contributions:

(i) **Unification**: We interpretatively connect the Bridge Theorem of information geometry (observer entropy $S_{\text{obs}}(p_\theta, \varepsilon) \sim \varepsilon^2 v^\top I v$) to the EVT-based collapse theory via the bound $S_{\text{obs}}(L) \leq c_2 e^{\mu_L}/L$ (with a matching lower bound in the pre-collapse regime $\mu_L > \log 2 + 1$), identifying a single information-theoretic measure that governs both coarse-graining information loss and attention collapse severity. The formal equivalence between the partition-based $S_{\text{obs}}(p_\theta, \varepsilon)$ (Definition 3.4) and the attention-uniform $S_{\text{obs}}(L)$ (Definition 4.7) is an interpretative identification whose rigorous formalisation is an open problem (Remark 4.2).

(ii) **Fundamental Impossibility Theorem**: For any fixed signal strength $\mu_s < \infty$, $S_{\text{obs}}(L) \to 0$ as $L \to \infty$ under softmax attention with sub-Gaussian logits. This is a rigorous no-go theorem for full long-context retention.

(iii) **Probabilistic Risk Law**: Using Gumbel EVT for weakly dependent sequences with Gaussian marginals, we derive a closed-form expression for $P(\mathcal{F}_L)$ and conjecture that the critical context length L_{crit} is a heavy-tailed random variable with $\log L_{\text{crit}} \sim (\mu_s - X)^2/2\sigma^2$ for $X \sim$ Gumbel (formal derivation of this distributional form is incomplete; see Remark 4.5).

(iv) **Control-theoretic governance**: The CPL 4.0 phase-aware governor is derived as the optimal discrete-time control response to the collapse risk law, with a Master Theorem guaranteeing hard invariants and sub-linear performance degradation bounds.

Paper structure. Section 2 reviews related work. Section 3 develops the information-geometric foundations. Section 4 presents the probabilistic collapse theory. Section 5 discusses architectural and training dependence. Section 6 derives the CPL 4.0 governor. Section 7 covers physical interpretations. Section 8 presents numerical verification. Section 9 proposes the experimental protocol. Section 10 provides the full assumption and result classification. Section 11 concludes.

2 Related Work

Context management heuristics. Practical approaches to long-context degradation include retrieval-augmented generation, sliding window attention, hierarchical summarisation, and chunking strategies. These are engineering interventions; none provides a principled information-theoretic analysis of the degradation mechanism or its fundamental limits.

Attention entropy studies. Zhai et al. [6] document attention entropy collapse in training and propose scaled initialisation to prevent it. Zhang et al. [7] show that attention entropy is a key predictor of downstream performance degradation. Our work provides the theoretical grounding for these empirical observations through the Bridge Theorem and EVT analysis.

Scaling laws. Kaplan et al. and Hoffmann et al. established power-law relationships between model parameters, training tokens, and loss. Our scaling conjecture $\mu_s \sim \sqrt{d} \cdot \mathcal{L}^{-1/2}$ proposes a specific link between these classical scaling laws and long-context capacity.

Extreme Value Theory for neural networks. Leadbetter et al. [10] established EVT for weakly dependent stationary processes. Berman [11] provided the key mixing conditions. Application of EVT to neural network attention is novel to our framework; the closest prior work is on random matrix theory for weight spectra.

Information geometry of KL divergence. The local quadratic structure of KL divergence with the Fisher information matrix as its Hessian is classical (Amari [8]; Rao [9]). Observer entropy as defined in Khomyakov [1] is a novel operator connecting this geometry to finite-resolution observation.

Concurrent work. Izumo [12] independently introduced a KL-based measure of information loss under discrete partitions for explainability purposes, but does not establish a geometric scaling law with Fisher information. Kiely et al. [13] study quantum Fisher information for observer objectivity, providing a complementary metrological perspective without establishing the KL-geometric scaling relation.

3 Information-Geometric Foundations

3.1 Token Space and Observer Entropy

Definition 3.1 (Token Space and Parametric Family)**.** *Let $\mathcal{X} = \{1, \ldots, V\}$ be the LLM vocabulary and $\Delta^\circ(\mathcal{X}) = \{p : \mathcal{X} \to (0, \infty) \mid \sum_x p(x) = 1\}$ the open probability simplex. At each step k, the LLM produces $p_{\theta_k} \in \Delta^\circ(\mathcal{X})$ where $\theta_k \in \Theta \subset \mathbb{R}^d$ encodes decoding parameters and context conditioning.*

Assumption 3.2 (Parametric Regularity)**.** *The family $\{p_\theta\}_{\theta \in \Theta}$ satisfies: (i) Θ is open and connected; (ii) $\theta \mapsto p_\theta(x)$ is C^3 for each x; (iii) $p_\theta(x) > 0$ for all x, θ.* ***[Modelling axiom / empirical hypothesis]***

Definition 3.3 (Cognitive Partition and Projection)**.** *Fix $\varepsilon > 0$. The relation $x \sim_\varepsilon y \iff d_{\mathcal{X}}(x, y) \leq \varepsilon$ induces partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ when it is an equivalence relation. The* projection *$\Pi_\varepsilon : \Delta(\mathcal{X}) \to \Delta(\mathcal{X}/\sim_\varepsilon)$ aggregates masses: $(\Pi_\varepsilon p)([x]) = \sum_{y \in [x]} p(y)$. The* uniform lift *$\widetilde{q}(y) = q([y])/|[y]|$ distributes class mass uniformly.*

Definition 3.4 (Observer Entropy)**.** *The* observer entropy *(behavioral entropy) is:*

$$S_{\text{obs}}(p_\theta, \varepsilon) = D_{\text{KL}}\left(p_\theta \,\middle\|\, \widetilde{\Pi_\varepsilon p_\theta}\right) = \sum_{x \in \mathcal{X}} p_\theta(x) \log \frac{p_\theta(x)}{\widetilde{\Pi_\varepsilon p_\theta}(x)}. \tag{1}$$

3.2 The Bridge Theorem

Assumption 3.5 (Parametric Deformation)**.** *There exists a C^1 map $v : \Theta \to \mathbb{R}^d$ such that $\widetilde{\Pi_\varepsilon p_\theta} = p_{\theta+\varepsilon v(\theta)+\rho(\theta,\varepsilon)}$ where $\|\rho(\theta,\varepsilon)\| \le L\varepsilon^2$ uniformly on compact subsets.*

Remark 3.1. Assumption 3.5 is a *theorem* for partition-adapted families (Definition 6.14 in [1]), including all full exponential families with centered between/within structure (Corollary 6.36, ibid.). For the softmax LLM family, it holds automatically (Remark 2.2 in [2]).

Theorem 3.6 (Bridge Theorem [1])**.** *Under Assumptions 3.2 and 3.5, within a fixed partition regime:*

$$S_{\text{obs}}(p_\theta, \varepsilon) = \frac{1}{2}\,\varepsilon^2\, v(\theta)^\top I(\theta)\, v(\theta) + O(\varepsilon^3), \tag{2}$$

with explicit remainder: $|S_{\text{obs}} - \frac{1}{2}\varepsilon^2 v^\top I v| \le C_1\varepsilon^3$ for some $C_1 > 0$ uniform on compact parameter subsets. ***[Proved]***

The proof proceeds by: (1) expressing $S_{\text{obs}} = D_{\text{KL}}(\theta\|\theta+\delta)$ via the deformation; (2) applying the local quadratic structure of KL (Fisher matrix = Hessian at coincidence); (3) expanding $\delta = \varepsilon v + \rho$ and collecting $O(\varepsilon^2)$ terms. Complete proof in [1], Theorem 6.3.

Corollary 3.7 (Riemannian Reformulation)**.** *$S_{\text{obs}}(p_\theta, \varepsilon) = \frac{1}{2}\varepsilon^2\|v(\theta)\|^2_{g_\theta} + O(\varepsilon^3)$ where $\|v\|^2_{g_\theta} = v^\top I(\theta) v$ is the squared Fisher–Rao norm of the coarse-graining generator. Observer entropy is one-half the squared Fisher–Rao distance to the projected parameter.* ***[Proved]***

3.3 Sufficient Conditions (Partition-Adapted Families)

Theorem 3.8 (Sufficient Conditions for Assumption 3.5 [1])**.** *A C^3 parametric family $\Phi : \Theta \to \Delta^\circ(\mathcal{X})$ with $d = n - 1$ parameters admits a $\mathcal{P}$-adapted parameterization (conditions A1–A3: within-class neutrality at zero, lift closure, immersion) if and only if Assumption 3.5 holds. For balanced exponential families satisfying conditions L1–L2, the deformation vector satisfies $v = (0, -\alpha)$ and $S_{\text{obs}} = \frac{1}{2}\varepsilon^2\alpha^\top I_{\text{ww}}\alpha + O(\varepsilon^3)$ where I_{ww} is the within-class Fisher block.* ***[Proved]***

4 Probabilistic Theory of Long-Context Collapse

4.1 Signal-Noise Model for Attention Logits

Definition 4.1 (Attention Logits and Signal Strength)**.** *Let $q \in \mathbb{R}^d$ be a query vector and $\{k_i\}_{i=1}^L \subset \mathbb{R}^d$ key vectors. Define attention logits $z_i = q^\top k_i/\sqrt{d}$. Assume a distinguished relevant index $i^* \in \{1, \dots, L\}$ with $\mathbb{E}[z_{i^*}] = \mu_s > 0$ and $\mathbb{E}[z_i] = 0$ for $i \ne i^*$. The* signal strength *is:*

$$\mu_s := \frac{1}{\sqrt{d}}\mathbb{E}[x_q^\top W_Q^\top W_K x_{i^*}] = \frac{1}{\sqrt{d}}\text{Tr}(W_Q^\top W_K \Sigma_{qr}), \tag{3}$$

where $\Sigma_{qr} = \mathbb{E}[x_q x_{i^*}^\top]$ is the query-relevant token cross-covariance.

Assumption 4.2 (Sub-Gaussianity)**.** *There exists $\sigma > 0$ such that $\mathbb{P}(|z_i| > t) \le 2e^{-t^2/(2\sigma^2)}$ for all $i \ne i^*$.* ***[Modelling axiom / empirical hypothesis]*** *(verified empirically for normalized key-query products)*

Assumption 4.3 (Gaussian Marginals)**.** *The noise logits $\{z_i\}_{i\ne i^*}$ have marginal distribution exactly $\mathcal{N}(0, \sigma^2)$.* ***[Modelling axiom / empirical hypothesis]*** *(Strengthens Assumption 4.2; verified approximately for normalised key-query dot-products in wide transformers via the central limit theorem applied to the inner product $q^\top k_i/\sqrt{d}$.)*

Remark 4.1 (Extension to non-Gaussian sub-Gaussian marginals). Assumption 4.3 is invoked in Appendix A to apply the classical Gaussian Gumbel limit with normalising sequences $b_L = \sigma\sqrt{2\log L_{\text{eff}}}$ and $a_L = \sigma/\sqrt{2\log L_{\text{eff}}}$. For a general sub-Gaussian distribution satisfying only Assumption 4.2, membership in MDA(Gumbel) is not automatic; it requires a separate domain-of-attraction argument, for example via the von Mises sufficient condition $\lim_{t\to\infty} t\bar{F}'(t)/\bar{F}(t) = 0$ (where $\bar{F} = 1 - F$ is the survival function) or the Gnedenko criterion. Verifying membership in MDA(Gumbel) for the empirical marginal of LLM attention logits, and establishing the appropriate normalising sequences for non-Gaussian sub-Gaussian cases, is an identified gap for future work.

Assumption 4.4 (Weak Dependence (Leadbetter Conditions)). *The sequence $\{z_i\}_{i\neq i^*}$ is stationary with covariance $\text{Cov}(z_i, z_j) = \sigma^2\rho(|i-j|)$ satisfying: (i) $\sum_{k=1}^{\infty}|\rho(k)| < \infty$; (ii) the Leadbetter $D(u_n)$ mixing condition: long-range extremes are asymptotically independent. Define effective length $L_{\text{eff}} = L/\tau_{\text{corr}}$ where $\tau_{\text{corr}} = 1 + 2\sum_{k=1}^{\infty}\rho(k)$.* ***[Proved conditionally]*** *(holds for exponentially decaying ρ, as induced by RoPE)*

4.2 Extreme Value Theorem for Attention

Lemma 4.5 (Maximum of Weakly Dependent Sub-Gaussians). *Under Assumptions 4.2, 4.3, and 4.4, the noise maximum $M_L = \max_{i\neq i^*} z_i$ satisfies:*

$$\mathbb{E}[M_L] = \sigma\sqrt{2\log L_{\text{eff}}} + O\left(\frac{\log\log L}{\sqrt{\log L}}\right). \tag{4}$$

[Proved conditionally] *(standard EVT under Gaussian marginals; gap is formal verification of $D(u_n)$ for transformer logits; for non-Gaussian sub-Gaussian marginals see Remark 4.1)*

Proof Sketch. Partition $\{1,\ldots,L\}\setminus\{i^*\}$ into $k_L = \lfloor L_{\text{eff}}\rfloor$ blocks of size $r_L = L/k_L$. By Assumption 4.4(ii), extremes of well-separated blocks are asymptotically independent. Within each block, the maximum follows the i.i.d. sub-Gaussian extreme value law. Applying the Gumbel limit for i.i.d. sub-Gaussian sequences and assembling across blocks yields the stated asymptotic. The dependence correction contributes only to the $O(\log\log L/\sqrt{\log L})$ term. *Gap:* Formal verification that LLM attention logits satisfy $D(u_n)$ requires direct measurement of mixing coefficients. □

Theorem 4.6 (Gumbel Convergence for Attention Maxima). *Under Assumptions 4.2, 4.3, and 4.4, with normalising sequences $b_L = \sigma\sqrt{2\log L_{\text{eff}}}$ and $a_L = \sigma/\sqrt{2\log L_{\text{eff}}}$:*

$$\frac{M_L - b_L}{a_L} \xrightarrow{d} \text{Gumbel}(0,1). \tag{5}$$

[Proved conditionally] *(conditional on Assumptions 4.4 and 4.3; for extension beyond Gaussian marginals see Remark 4.1)*

4.3 Tight Bounds on Observer Entropy

Definition 4.7 (Attention-Induced Observer Entropy). *At context length L, define the* effective margin $\mu_L := \mu_s - \mathbb{E}[M_L] \approx \mu_s - \sigma\sqrt{2\log L_{\text{eff}}}$ *and the attention-induced observer entropy as $S_{\text{obs}}(L) := D_{\text{KL}}(p_\theta\|\Pi(p_\theta))$ where Π replaces attention weights by uniform weights $u_i = 1/L$.*

Remark 4.2 (Relationship between Definition 3.4 and Definition 4.7). Definition 3.4 ($S_{\text{obs}}(p_\theta,\varepsilon)$) and Definition 4.7 ($S_{\text{obs}}(L)$) are *operationally distinct*: the former coarse-grains the output distribution via a partition indexed by resolution parameter ε; the latter replaces the attention allocation by the uniform distribution $u_i = 1/L$. These coincide interpretatively under the identification $\varepsilon \sim 1/\sqrt{L}$ (so that the number of distinguishable attention bins $\sim \varepsilon^{-2} \sim L$), but a formal equivalence—a proved proposition of the form $\kappa_1 S_{\text{obs}}(p_\theta,\varepsilon(L)) \le S_{\text{obs}}(L) \le \kappa_2 S_{\text{obs}}(p_\theta,\varepsilon(L))$ for explicit constants $\kappa_1,\kappa_2 > 0$ and an explicit partition $\mathcal{P}_\varepsilon$—has not been established within this framework. Accordingly, the application of the Bridge Theorem (Theorem 3.6, proved for Definition 3.4) to bounds on

$S_{\text{obs}}(L)$ (Definition 4.7) constitutes an interpretative connection, not a formal deduction. A sufficient approach to formalising this identification would be to specify a canonical partition $\mathcal{P}_\varepsilon$ on the attention-weight simplex such that $\widetilde{\Pi_\varepsilon p_\theta} = \Pi(p_\theta)$ when $\varepsilon = 1/\sqrt{L}$; this is an open problem listed in §10.

Theorem 4.8 (Two-Sided Bounds in the Pre-Collapse Regime; Upper Bound for All L)**.** *Under Assumptions 4.2–4.4 and a Lipschitz output assumption (the output distribution satisfies $|p_\theta(\cdot|a) - p_\theta(\cdot|a')|_{\text{TV}} \le K|a - a'|_1$), let*

$$L_0 := \exp\Big(\tfrac{(\mu_s - \log 2 - 1)^2}{2\sigma^2}\Big) \tag{6}$$

denote the threshold beyond which $\mu_L \le \log 2 + 1$.

(i) Pre-collapse regime *($\mu_L > \log 2 + 1$, equivalently $L < L_0$):*

$$c_1 \cdot \mu_L \cdot \frac{e^{\mu_L}}{L} \;\le\; S_{\text{obs}}(L) \;\le\; c_2 \cdot \frac{e^{\mu_L}}{L} \tag{7}$$

for constants $c_1, c_2 > 0$ depending on the Lipschitz constant K and the sub-Gaussian parameter σ.

(ii) All *L (including $L \ge L_0$):*

$$S_{\text{obs}}(L) \;\le\; c_2 \cdot \frac{e^{\mu_L}}{L}. \tag{8}$$

The one-sided upper bound holds for all sufficiently large L. A matching lower bound valid uniformly for all large L is not established by the current proof; see Remark 4.3. ***[Proved conditionally]***

Proof Sketch. Note: The following bounds are stated for $S_{\text{obs}}(L)$ as defined in Definition 4.7. Their connection to the Bridge Theorem (Theorem 3.6) is interpretative; see Remark 4.2.

Upper bound. By softmax concentration, $a_{i^*} \le e^{\mu_L}/(e^{\mu_L} + L)$ and $|a - u|_1 \le Ce^{\mu_L}/L + O(1/L)$. The Lipschitz assumption gives $S_{\text{obs}}(L) \le K|a - u|_1 \le c_2 e^{\mu_L}/L$.

Lower bound. The KL divergence decomposes as $S_{\text{obs}}(L) \ge a_{i^*} \log(a_{i^*} L)$. With $a_{i^*} \sim e^{\mu_L}/L$ and $\log(a_{i^*} L) \sim \mu_L$, we obtain $S_{\text{obs}}(L) \ge c_1 \mu_L e^{\mu_L}/L$, valid when $\mu_L > \log 2 + 1$ so that $\log(a_{i^*} L) > 0$. For $\mu_L \le \log 2 + 1$ the lower bound argument breaks down; see Remark 4.3. *Gap:* The Lipschitz assumption on $p_\theta(\cdot|a)$ requires formal verification for standard transformer decoders; it holds for bounded value vectors. □

Remark 4.3 (Lower bound and the deep collapse regime). The lower bound in part (i) of Theorem 4.8 requires $\mu_L > \log 2 + 1$, which is violated for all $L \ge L_0$. Since $\mu_L = \mu_s - \sigma\sqrt{2 \log L_{\text{eff}}} \to -\infty$ as $L \to \infty$, the pre-collapse condition $\mu_L > \log 2 + 1$ holds only on the bounded interval $L \in (0, L_0)$. In the deep collapse regime ($\mu_L \le 0$), the effective margin is negative and $e^{\mu_L}/L \to 0$ faster than any power of $1/L$; a matching lower bound would require either a different proof technique (for example, a refined expansion of the softmax competition near the uniform distribution) or a modified definition of $S_{\text{obs}}(L)$ that remains meaningful when no single token dominates attention. See also item (7) in §10.

Remark 4.4 (Impossibility Theorem is unaffected by the lower bound restriction). Theorem 4.9 ($S_{\text{obs}}(L) \to 0$ as $L \to \infty$) follows directly from the upper bound $S_{\text{obs}}(L) \le c_2 e^{\mu_L}/L$, which holds for all sufficiently large L (Theorem 4.8(ii)). The restriction of the two-sided bound to the pre-collapse regime does not weaken the Impossibility Theorem.

4.4 Fundamental Impossibility Theorem

Theorem 4.9 (Impossibility of Full Long-Context Retention)**.** *For any parametric family satisfying Assumptions 4.2 and 4.4, and any finite signal strength $\mu_s < \infty$:*

$$\lim_{L \to \infty} S_{\text{obs}}(L) = 0. \tag{9}$$

That is, no finite-signal softmax attention mechanism can maintain non-zero observer entropy (information retention) at arbitrarily large context lengths. ***[Proved conditionally]***

Proof. From Theorem 4.8(ii), $\log S_{\text{obs}}(L) \le \log c_2 + \mu_L - \log L = \log c_2 + \mu_s - \sigma\sqrt{2 \log L} - \log L + o(1)$. Since $\log L \gg \sqrt{\log L}$ as $L \to \infty$, this tends to $-\infty$. Hence $S_{\text{obs}}(L) \to 0$. □

4.5 Stochastic Critical Length and Probabilistic Risk Law

Definition 4.10 (Critical Context Length)**.** *For threshold $\varepsilon_0 > 0$, the* critical context length *is the (random) first passage time:* $L_{\text{crit}} = \inf\{L : S_{\text{obs}}(L) \leq \varepsilon_0\}$.

Theorem 4.11 (Stochastic Critical Length)**.** *Under the Gumbel limit (Theorem 4.6), with* $X \sim \text{Gumbel}(0, 1)$*:*

$$\log L_{\text{crit}} \approx \frac{(\mu_s - X - \log \varepsilon_0)^2}{2\sigma^2}. \tag{10}$$

Consequently, L_{crit} *has a heavy-tailed distribution: the expected critical length satisfies* $\mathbb{E}[L_{\text{crit}}] \gg \text{median}(L_{\text{crit}})$*, and variance in collapse timing is fundamental, not incidental.* ***[Conjecture — proof incomplete]*** *(see Remark 4.5)*

Proof Sketch. Collapse ($S_{\text{obs}}(L) \leq \varepsilon_0$) requires $e^{\mu_L}/L \lesssim \varepsilon_0$, i.e., $\mu_L \lesssim \log L + \log \varepsilon_0$. With the correct Gumbel parameterisation (Theorem 4.6), $M_L \approx b_L + a_L X$ where $X \sim \text{Gumbel}(0, 1)$, so $\mu_L = \mu_s - M_L \approx \mu_s - \sigma\sqrt{2 \log L_{\text{eff}}} - a_L X$. Substituting into the collapse condition $\mu_L \lesssim \log L + \log \varepsilon_0$ and setting $u = \sqrt{2 \log L_{\text{eff}}}$ gives a transcendental equation in u; a closed-form solution requires a further approximation whose validity has not been fully verified. The stated distributional form $\log L_{\text{crit}} \approx (\mu_s - X - \log \varepsilon_0)^2/(2\sigma^2)$ is consistent with the leading-order behaviour of this equation only when $a_L X \ll \mu_s - \log L - \log \varepsilon_0$, an assumption that is not verified in the current sketch. □

Remark 4.5 (Gap in the proof of Theorem 4.11). The proof sketch above does not constitute a complete derivation. The correct Gumbel substitution is $M_L \approx b_L + a_L X$ (not $b_L + X/a_L$ as would be dimensionally inconsistent with the normalising sequence $a_L = \sigma/\sqrt{2 \log L_{\text{eff}}}$). Under the corrected substitution, the step from the collapse condition to the closed-form $\log L_{\text{crit}} \approx (\mu_s - X - \log \varepsilon_0)^2/(2\sigma^2)$ requires: (i) a rigorous fixed-point analysis of the resulting transcendental equation in $\log L$; and (ii) justification that the perturbation $a_L X$ is negligible at leading order relative to $\mu_s - \log L - \log \varepsilon_0$. Until these steps are provided, this result carries the status **[Conjecture — proof incomplete]**.

Corollary 4.12 (Probabilistic Risk Law)**.** *The probability that context of length L leads to attention failure is:*

$$P(\mathcal{F}_L) \approx 1 - \exp\left(-\exp\left(-\frac{\mu_s - \sigma\sqrt{2 \log L_{\text{eff}}}}{a_L}\right)\right), \tag{11}$$

where $\mathcal{F}_L = \{\max_{i \neq i^*} z_i \geq z_{i^*}\}$ *is the failure event and* $a_L = \sigma/\sqrt{2 \log L_{\text{eff}}}$*. This is the CDF of the* Gumbel$(0, 1)$ *distribution evaluated at* $(\mu_s - b_L)/a_L$*.* ***[Proved conditionally]***

Remark 4.6 (Qualitative Behavior). The risk law (11) exhibits S-shaped behavior in $\log L$: near-zero failure probability for $L \ll L_{\text{crit}}$, rapid increase near L_{crit}, and convergence to 1 for $L \gg L_{\text{crit}}$. This matches the empirically observed sudden collapse behavior documented in [5].

5 Architectural and Training Dependence of Signal Strength

5.1 Signal Strength as Spectral Alignment

From Definition 4.1, $\mu_s = \frac{1}{\sqrt{d}} \text{Tr}(W_Q^\top W_K \Sigma_{qr})$. Diagonalizing $W_Q^\top W_K = U \Lambda U^\top$:

$$\mu_s = \frac{1}{\sqrt{d}} \sum_i \lambda_i \sigma_i \tag{12}$$

where λ_i are the eigenvalues of $W_Q^\top W_K$ and σ_i are projections of Σ_{qr} onto the corresponding eigenvectors. Signal strength is thus the *spectral alignment* between the learned attention mechanism and the semantic structure of the data.

5.2 NTK-Regime Analysis

In the Neural Tangent Kernel (NTK) regime (sufficiently wide networks), the attention weight matrix evolves minimally from initialisation:

$$W_Q^\top W_K \approx K_{\text{attn}} \quad \text{(kernel alignment matrix)}, \tag{13}$$

yielding $\mu_s \approx \frac{1}{\sqrt{d}}\text{Tr}(K_{\text{attn}}\Sigma_{qr})$. This reduces signal strength to the alignment between the kernel geometry and the data cross-covariance.

5.3 Positional Decay via RoPE

For sequences with positional encoding (RoPE), the query-key correlation at positional distance Δ satisfies $\rho(\Delta) \sim e^{-\gamma\Delta}$, yielding:

$$\mu_s(\Delta) \approx \mu_0 e^{-\gamma\Delta}, \tag{14}$$

so that $\mu_L = \mu_0 e^{-\gamma\Delta} - \sigma\sqrt{2\log L_{\text{eff}}}$. Long-range dependencies are therefore doubly penalised: by positional decay and by growing extreme-value noise.

5.4 Scaling Law Conjecture

Conjecture 5.1 (Signal-Loss Scaling)**.** *For a transformer trained to validation loss $\mathcal{L}$ on D tokens with model dimension d:*

$$\mu_s \sim \sqrt{d} \cdot \mathcal{L}^{-1/2}, \tag{15}$$

*implying $L_{\text{crit}} \sim \exp(d \cdot \mathcal{L}^{-1})$. **[Modelling axiom / empirical hypothesis]***

Remark 5.1 (Motivation and Required Assumptions). The $\sqrt{d}$ factor emerges from the variance of dot-products: $q \cdot k \sim \mathcal{N}(0, d)$ and $\mu_s \sim \rho\sqrt{d}$ where ρ is semantic alignment. The $\mathcal{L}^{-1/2}$ factor is motivated by the empirical observation that better-trained models produce more semantically aligned representations. To promote Conjecture 5.1 to a theorem requires: (i) a formal model of representation geometry as a function of training loss; (ii) a proof that the NTK approximation holds with sufficient accuracy; and (iii) a quantitative bound on ρ in terms of $\mathcal{L}$.

6 Control-Theoretic Response: CPL 4.0 Governor

6.1 Phase Classifier and State Model

The CPL 4.0 framework [2] introduces a discrete-time state $x_k = (L_k, \hat{H}_k, \hat{S}_k, \hat{D}_k, z_k, c_k)$ where $\hat{H}_k = S_{\text{obs}}(p_{\theta_k}, \varepsilon)$, $\hat{S}_k = 1 - \hat{H}_k / S_{\text{obs}}^{\max}$ (semantic stability), and $z_k \in \{\mathbf{C}, \mathbf{R}, \mathbf{F}\}$ is the cognitive phase.

Definition 6.1 (Phase Classifier)**.** *The phase is assigned by:*

$$z_k = \begin{cases} \mathbf{R}, & |\hat{H}_k - \hat{H}_{k-1}| \geq \gamma, \\ \mathbf{C}, & \hat{H}_k < H_c^{\text{eff}} \wedge \hat{S}_k > S_c, \\ \mathbf{F}, & \textit{otherwise}, \end{cases} \tag{16}$$

with hysteresis $H_c^{\text{eff}} = H_c + \beta_{\text{hyst}}\mathbf{1}\{z_{k-1} = \mathbf{C}\}$.

6.2 Governance Policy and Lyapunov Stability

The policy maps phases to actions (keep/summarize/chunk) and releases tokens$r_k \in \{0, r_{\text{rescue}}, r_{\text{recover}}\}$ per equations (16)–(19) of [2]. Tight decoding (θ^{tight}) is activated in non-Coherence phases to drive entropy contraction.

Definition 6.2 (Lyapunov Potential)**.** $V_k = (\hat{H}_k - H_c)_+ + (S_c - \hat{S}_k)_+$.

Under tight decoding with $\alpha_{\text{eff}} \in [\alpha_{\min}, \alpha_{\max}] \subset (0, 1)$ and target $h(L, \theta^{\text{tight}}) < H_c$, the drift condition holds: $V_{k+1} \leq (V_k - \delta)_+ + \xi_{k+1}$ with $\delta > 0$ and conditionally sub-Gaussian ξ_{k+1}.

6.3 Master Theorem

Theorem 6.3 (CPL 4.0 Master Theorem [2]). *Under the standing hypotheses of [2] (bounded inputs, safe initialisation, sub-Gaussian noise, design function regularity, design margin, threshold ordering), the CPL 4.0 governor guarantees:*

(i) ***Hard context cap:*** $L_k \le L_{\text{cap}}$ *for all* $k \ge 0$.

(ii) ***Entropy contraction:*** *In tight mode with* $\hat{H}_k - H_{\text{tgt}} \ge \Delta > 0$: $\mathbb{E}[\hat{H}_{k+1} | x_k, a_k] \le \hat{H}_k - \alpha_{\min}\Delta$.

(iii) ***Fragmentation bound:*** *For any* $\delta_0 \in (0,1)$, *with probability* $\ge 1 - \delta_0$: $N_F(T) \le \frac{V_0}{\delta} + \frac{\sigma}{\delta}\sqrt{2T\log(1/\delta_0)} = O(\sqrt{T\log(1/\delta_0)})$.

[Proved conditionally] *(conditional on Spectral Response Assumption [2])*

Interpretation via the Risk Law. The Master Theorem provides the *control-theoretic complement* to the Risk Law (11): while the risk law characterises the probability of collapse at length L, the governor guarantees that the controller prevents the system from ever reaching $L_{\text{cap}} < L_{\text{crit}}$. Together, they establish that the controlled system operates in the safe regime $P(\mathcal{F}_L) \ll 1$ with high probability.

7 Physical Interpretations

7.1 Cognitive Uncertainty Principle

Theorem 7.1 (Resolution–Information Trade-off [1]). *Under Assumptions 3.2 and 3.5, if* $I(\theta)$ *is positive definite and* $v(\theta) \neq 0$, *then* $S_{\text{obs}}(p_\theta, \varepsilon) \ge \delta_0$ *requires:*

$$\varepsilon \ge \sqrt{\frac{2\delta_0}{v(\theta)^\top I(\theta) v(\theta)}}\,(1 - C'\varepsilon). \tag{17}$$

Higher Fisher information (in the projection direction) lowers the detection threshold: sharper semantic distinctions become detectable at finer resolution. ***[Proved]***

Threshold calibration. Setting $\delta_0 = H_c$ (the CPL entropy threshold), the critical resolution is $\varepsilon_{\text{crit}} \approx \sqrt{2H_c/(v^\top I v)}$. Below this, phase transitions between coherence and fragmentation are not detectable.

7.2 Landauer Thermodynamic Bound

Theorem 7.2 (Landauer Bound [1]). *Under the identification of* Π_ε *with irreversible information erasure (a physical modelling assumption), the minimum energetic cost of performing the* ε*-resolution coarse-graining is:*

$$E_{\min} \ge kT \cdot S_{\text{obs}}(p_\theta, \varepsilon) = \frac{kT}{2}\varepsilon^2 v(\theta)^\top I(\theta) v(\theta) + O(\varepsilon^3), \tag{18}$$

where k *is Boltzmann's constant and* T *is temperature.* ***[Proved conditionally]*** *(modelling assumption on erasure identification)*

LLM deployment interpretation. Each summarisation step in the CPL 4.0 governor erases $S_{\text{obs}}(p_{\theta_k}, \varepsilon)$ nats of within-class information. The Landauer bound provides a physical cost floor for each such operation. Over T steps, total context management cost satisfies: $E_{\text{manage}} \ge kT \sum_k S_{\text{obs}}(p_{\theta_k}, \varepsilon)\mathbf{1}\{m_k \neq \texttt{keep}\}$. The thermodynamic cost of long-context operation thus grows with the severity of the information-geometric degradation measured by S_{obs}.

8 Numerical Verification

8.1 Bridge Theorem Verification

Both companion scripts (`verify_bridge_theorem_v2.py`) confirm the Bridge Theorem on two concrete softmax families.

Example 1: 4-point space. $\mathcal{X} = \{1,2,3,4\}$, partition $\mathcal{P} = \{\{1,2\},\{3,4\}\}$, $d = 3$, $v = (0, -\alpha, -\beta)$. At $\theta^* = 0$: $I(\theta^*) = \mathrm{diag}(1, \frac{1}{2}, \frac{1}{2})$, leading prediction $\frac{1}{2}v^\top I v = \frac{1}{4}(\alpha^2 + \beta^2)\varepsilon^2$.

Example 2: 5-point space. $\mathcal{X} = \{1,\ldots,5\}$, partition $\mathcal{P} = \{\{1,2,3\},\{4,5\}\}$, $d = 4$, $v = (0, -\alpha, -\beta, -\gamma)$. At $\theta^* = 0$: $I(\theta^*) = \mathrm{diag}(\frac{3}{2}, \frac{6}{5}, \frac{2}{5}, \frac{2}{5})$, leading prediction $\frac{1}{2}v^\top I v = (\frac{3}{5}\alpha^2 + \frac{1}{5}\beta^2 + \frac{1}{5}\gamma^2)\varepsilon^2$.

Verification results (all 8 components pass):

(1) Log-log slope of exact S_{obs} vs ε: fitted slope $= 2.000 \pm 0.01$ (both examples).

(2) Normalised remainder: $\sup_{\varepsilon \in [10^{-4}, 10^{-2}]} |R(\varepsilon)|/\varepsilon^3 < 10$.

(3) Fisher matrix agreement: analytic vs. finite-difference $< 10^{-8}$.

(4) Closed-form coefficient consistency: $< 10^{-15}$ deviation.

(5) Relative error $< 1\%$ for $\varepsilon \leq 0.03$.

(6) Landauer bound $E_{\min} = kT \cdot S_{\mathrm{obs}} > 0$ satisfied.

(7) Monotone non-decrease: 0 violations.

(8) Sub-linear relative error slope ≥ 0.90 on $[10^{-3}, 10^{-1}]$.

8.2 CPL 4.0 Simulator Results

The `simulator.py` (v2.0) implements the discrete governor and confirms:

- **Context cap invariant**: Governor max $L = 4000 = L_{\mathrm{cap}}$; baseline reaches $L_{\mathrm{practical}} = 4800$ (violation).
- **Lyapunov convergence**: $V_k \to 0$ under the governor; baseline exhibits persistent positive V_k.
- **Fragmentation occupancy**: Governor N_F grows as $O(\sqrt{T})$; confirmed to remain below the theoretical bound with $\sigma_{\mathrm{eff}} = \sigma_H$.
- **Phase-space diagram**: Trajectory under governor remains in the Coherence region $\hat{H} < H_c, \hat{S} > S_c$; baseline drifts into Fragmentation.
- **Algebraic coupling**: $\hat{S}_k = 1 - \hat{H}_k / S_{\mathrm{obs}}^{\max}$ (no independent SDE for stability) confirmed structurally.

9 Experimental Protocol for Real-LLM Validation

The following protocol is designed to validate Theorem 4.8, Corollary 4.12, and Conjecture 5.1 on deployed models (LLaMA-3, Mistral-7B, etc.).

Experiment 1: Margin scaling $\mu(L)$. Using needle-in-a-haystack (NIAH) tasks [15], for each $L \in \{100, 500, 1000, 2000, 5000, 10000\}$: (a) Extract attention logits $z_i = q^\top k_i / \sqrt{d}$ from the last transformer layer for the query token targeting the needle; (b) compute $\mu(L) = z_{\mathrm{rel}} - \max_{i \neq \mathrm{rel}} z_i$; (c) fit $\mu(L)$ vs. $\sqrt{2 \log L}$. **Prediction**: $\mu(L) \approx \mu_s - \sigma\sqrt{2 \log L}$, with linear decay in $\sqrt{\log L}$.

Experiment 2: Failure probability curves. For each L, run $N = 100$ NIAH instances and record $P(\mathcal{F}_L)$ = fraction of failures. Fit the empirical CDF to the Gumbel prediction (11) with parameters μ_s and σ estimated from Experiment 1. **Prediction**: S-shaped curve in $\log L$; Gumbel CDF with fitted parameters achieves $R^2 > 0.95$.

Experiment 3: Training checkpoint scaling. For a model trained with checkpoint saves at intervals of 10^9 tokens, measure μ_s (estimated as mean of z_{rel} over 1000 NIAH instances) vs. training loss $\mathcal{L}$. **Prediction**: $\mu_s \propto \mathcal{L}^{-1/2}$ (linear in $1/\sqrt{\mathcal{L}}$).

Experiment 4: RoPE positional decay. For fixed L, vary the positional distance Δ between query and needle (by controlling needle insertion position). Measure $z_{\text{rel}}(\Delta)$. **Prediction**: $z_{\text{rel}}(\Delta) \propto e^{-\gamma\Delta}$.

Experiment 5: Fisher information proxy. For the softmax output distribution p_{θ_k} at each step, estimate $v(\theta_k)^\top I(\theta_k) v(\theta_k)$ numerically using the formula of Theorem 3.6 and compare with the attention entropy H_{attn}. **Prediction**: High correlation between S_{obs} (information-geometric measure) and H_{attn} (empirical proxy) in the collapse regime.

10 Assumptions, Classification, and Limitations

10.1 Complete Assumption Catalogue

Assumption	Type	Content
Ass. 3.2	Structural	C^3 regularity, strict positivity
Ass. 3.5	Struct./proved	Parametric deformation; proved for partition-adapted families
Ass. 4.2	Empirical	Sub-Gaussian attention logits
Ass. 4.3	Modelling axiom	Exact $\mathcal{N}(0, \sigma^2)$ marginals for noise logits; strengthens Ass. 4.2; used in Appendix A
Ass. 4.4	Empirical	Weak dependence; holds for RoPE-induced correlations
Lipschitz output	Modelling	$\lvert p_\theta(\cdot\lvert a) - p_\theta(\cdot\lvert a')\rvert_{\text{TV}} \le K\lvert a - a'\rvert_1$
Ass. [Spectral]	Modelling/Empirical	$Q(\theta^{\text{tight}}) < Q(\theta^{\text{base}})$
Erasure identification	Physical model	Π_ε = thermodynamic erasure
Conj. 5.1	Empirical	$\mu_s \sim \sqrt{d} \cdot \mathcal{L}^{-1/2}$
CPL Standing Hyps.	Design	Bounded I/O, safe init, sub-Gaussian noise, etc. [2]

10.2 Classification of Main Results

Result	Status	Key condition
Bridge Theorem (Thm. 3.6)	**[Proved]**	Partition-adapted family
Riemannian reformulation (Cor. 3.7)	**[Proved]**	Bridge Theorem
Sufficient conditions (Thm. 3.8)	**[Proved]**	A1–A3
EVT Lemma (Lem. 4.5)	**[Proved conditionally]**	$D(u_n)$ mixing condition; Ass. 4.3
Gumbel convergence (Thm. 4.6)	**[Proved conditionally]**	Leadbetter conditions; Ass. 4.3 (Gaussian marginals; see Rem. 4.1)

Result	Status	Key condition
Two-sided bounds (Thm. 4.8)	[**Proved conditionally**]	Lipschitz output; lower bound restricted to $\mu_L > \log 2 + 1$ (see Rem. 4.3)
Impossibility Theorem (Thm. 4.9)	[**Proved conditionally**]	Upper bound only; unaffected by lower bound restriction (see Rem. 4.4)
Stochastic L_{crit} (Thm. 4.11)	[**Conjecture — proof incomplete**]	Gumbel convergence; fixed-point derivation incomplete (see Rem. 4.5)
Risk Law (Cor. 4.12)	[**Proved conditionally**]	Gumbel convergence
CPL Master Theorem (Thm. 6.3)	[**Proved conditionally**]	Spectral Response Ass.
Resolution trade-off (Thm. 7.1)	[**Proved**]	Bridge Theorem, PD
Landauer bound (Thm. 7.2)	[**Proved conditionally**]	Erasure identification
Scaling conjecture (Conj. 5.1)	[**Modelling axiom / empirical hypothesis**]	NTK + loss-alignment link

10.3 Limitations and Open Problems

The following gaps remain explicitly identified:

(1) **Non-Gaussianity**: Real attention logits are not Gaussian; heavy tails may accelerate collapse beyond the Gumbel prediction. The proofs in Appendix A require Gaussian marginals (Assumption 4.3); extension to general sub-Gaussian distributions requires a domain-of-attraction argument (Remark 4.1).

(2) **Stationarity and non-stationarity**: The weak-dependence assumption requires approximate stationarity of attention logits. Positional embeddings introduce non-stationarity; quantifying this effect is an open problem.

(3) **Multi-head aggregation**: Our analysis is for a single attention head. The effective signal under multi-head aggregation depends on correlation structure across heads; current treatment treats h heads as providing a $\sqrt{\log h}$ factor.

(4) **Training dynamics**: Conjecture 5.1 connects μ_s to training loss but lacks a formal proof. The link between representation geometry and training loss requires a model of the learning trajectory.

(5) **Single relevant token**: The failure event $\mathcal{F}_L$ is defined with respect to a single relevant token. Extension to multiple relevant tokens (more realistic for many tasks) requires a multi-variate EVT analysis.

(6) **Semigroup structure**: Whether $\{\Pi_\varepsilon\}_{\varepsilon \geq 0}$ admits a continuous semigroup structure within a fixed partition regime remains open (cf. Proposition 9.4 of [1]).

(7) **Lower bound in the deep collapse regime**: The two-sided bound of Theorem 4.8 is established only for $\mu_L > \log 2 + 1$ (i.e., $L < L_0$). For $L \geq L_0$, only the upper bound $S_{\text{obs}}(L) \leq c_2 e^{\mu_L}/L$ is proved. A matching lower bound for all large L remains open; it would require either a different proof technique or a modified definition of $S_{\text{obs}}(L)$ in the deep collapse regime $\mu_L < 0$ (Remark 4.3).

(8) **Formal identification of observer entropy definitions**: The connection between the partition-based $S_{\text{obs}}(p_\theta, \varepsilon)$ (Definition 3.4) and the attention-uniform $S_{\text{obs}}(L)$ (Definition 4.7) is currently interpretative (Remark 4.2). A formal equivalence proposition would require specifying a canonical partition $\mathcal{P}_\varepsilon$ on the attention-weight simplex such that $\widetilde{\Pi_\varepsilon p_\theta} = \Pi(p_\theta)$ when $\varepsilon = 1/\sqrt{L}$; this is an open problem.

11 Conclusion

We have developed a unified information-theoretic and probabilistic framework for understanding and governing long-context degradation in LLMs. The central thesis is confirmed at multiple levels of the analysis:

Information-geometrically: The Bridge Theorem establishes that observer entropy $S_{\text{obs}} = \frac{1}{2}\varepsilon^2 v^\top I v + O(\varepsilon^3)$ measures the information loss from finite-resolution semantic coarse-graining, providing a Fisher-information-grounded measure of distributional stability.

Probabilistically: The EVT analysis (under Gaussian marginals, Assumption 4.3) establishes that attention logit maxima converge to Gumbel distributions under weak dependence, giving a closed-form risk law $P(\mathcal{F}_L) \approx 1 - \exp(-\exp(-(\mu_s - b_L)/a_L))$ and conjecturing that L_{crit} has heavy-tailed distributional character (formal proof of this distributional form is incomplete; see Remark 4.5).

Structurally: The bound $S_{\text{obs}}(L) \leq c_2 e^{\mu_L}/L$ (with a matching lower bound in the pre-collapse regime $\mu_L > \log 2 + 1$, Definition 4.7), where $\mu_L = \mu_s - \sigma\sqrt{2\log L_{\text{eff}}}$, interpretatively unifies both theories and yields the Fundamental Impossibility Theorem: for any finite μ_s, full long-context retention is impossible. The formal equivalence between the partition-based and attention-uniform definitions of S_{obs} is an open problem (Remark 4.2).

Control-theoretically: The CPL 4.0 governor provides the principled control-theoretic response, maintaining the system in the low-failure region $L < L_{\text{cap}} \ll L_{\text{crit}}$ with hard invariants and sub-linear degradation bounds.

Implications for LLM deployment. The risk law provides a principled basis for context length management decisions: contexts should be kept below the Gumbel risk threshold, and architectural choices should maximise μ_s (signal strength) by improving training quality and semantic alignment. The CPL 4.0 governor implements this dynamically. The Landauer bound establishes a thermodynamic cost floor for context compression operations.

Future directions. The most important open problems are: (i) empirical validation of the Gumbel risk law on real models; (ii) formal proof of the scaling conjecture $\mu_s \sim \sqrt{d} \cdot \mathcal{L}^{-1/2}$; (iii) extension to multi-token and multi-head settings; (iv) derivation of optimal policies via the MDP formulation introduced in [2]; (v) formal identification of the partition-based and attention-uniform observer entropy definitions; and (vi) completion of the fixed-point derivation for the distributional form of L_{crit}.

Acknowledgements

The author's research profile is available at orcid.org/0009-0006-3074-9145.

Code repositories: https://github.com/Khomyakov-Vladimir/observer-entropy-bridge and https://github.com/Khomyakov-Vladimir/llm-context-window-governance.

References

[1] V. Khomyakov (2026a). *KL-Geometric Structure of Observer Entropy: A Minimal Information-Theoretic Framework*, v2.4.1. Zenodo. doi:10.5281/zenodo.19202244.

[2] V. Khomyakov (2026b). *Information-Geometric Context Window Governance for LLMs via Observer Entropy and the CPL 4.0*, v2.0. Zenodo. doi:10.5281/zenodo.19177363.

[3] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang (2024). *Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics*, 12:157–173. doi:10.1162/tacl_a_00638.

[4] Y. Du, M. Tian, S. Ronanki, S. Rongali, S. B. Bodapati, A. Galstyan, A. Wells, R. Schwartz, E. A. Huerta, and H. Peng (2025). *Context Length Alone Hurts LLM Performance Despite Perfect Retrieval*. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23281–23298. doi:10.18653/v1/2025.findings-emnlp.1264.

[5] W. Wang, J. Min, and W. Zou (2026). *Intelligence Degradation in Long-Context LLMs: Critical Threshold Determination via Natural Length Distribution Analysis*. arXiv:2601.15300. doi:10.48550/arXiv.2601.15300.

[6] S. Zhai, T. Likhomanenko, E. Littwin, D. Busbridge, J. Ramapuram, Y. Zhang, J. Gu, and J. M. Susskind (2023). *Stabilizing Transformer Training by Preventing Attention Entropy Collapse*. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, PMLR 202, pp. 40770–40803. arXiv:2303.06296. doi:10.48550/arXiv.2303.06296.

[7] Z. Zhang, Y. Wang, X. Huang, T. Fang, H. Zhang, C. Deng, S. Li, and D. Yu (2025). *Attention Entropy is a Key Factor: An Analysis of Parallel Context Encoding with Full-attention-based Pre-trained Language Models*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pp. 9840–9855. doi:10.18653/v1/2025.acl-long.485.

[8] S. Amari (2016). *Information Geometry and Its Applications*. Applied Mathematical Sciences, vol. 194. Springer Tokyo. doi:10.1007/978-4-431-55978-8.

[9] C. R. Rao (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91.

[10] M. R. Leadbetter, G. Lindgren, and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer, New York. doi:10.1007/978-1-4612-5449-2.

[11] S. M. Berman (1964). Limit theorems for the maximum term in stationary sequences. *The Annals of Mathematical Statistics*, 35(2):502–516. doi:10.1214/aoms/1177703551.

[12] T. Izumo (2025/2026). Quantifying information loss under coarse-grained partitions: A discrete framework for explainable artificial intelligence. Preprint. arXiv:2502.07347. doi:10.48550/arXiv.2502.07347.

[13] A. Kiely, D. A. Chisholm, A. Touil, S. Deffner, G. Landi, and S. Campbell (2026). Metrological approach to the emergence of classical objectivity. *Physical Review A*, 113:022403. doi:10.1103/hn78-7xx3.

[14] T. M. Cover and J. A. Thomas (2006). *Elements of Information Theory*, 2nd ed. Wiley-Interscience. doi:10.1002/047174882X.

[15] G. Kamradt (2023). NeedleInAHaystack: A benchmark for long-context LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`.

A Complete Proof of the Gumbel Convergence Theorem

Setup. Throughout this appendix we invoke Assumption 4.3 (Gaussian Marginals). Let $\{z_i\}_{i=1}^{L}$ be a stationary sequence with marginal $\mathcal{N}(0, \sigma^2)$ and covariance $\mathrm{Cov}(z_i, z_j) = \sigma^2 \rho(|i-j|)$ with $\sum_k |\rho(k)| < \infty$ (Assumption 4.4). The Gaussian marginal places $\{z_i\}$ in MDA(Gumbel) with classical normalising sequences $b_L = \sigma\sqrt{2\log L_{\text{eff}}}$ and $a_L = \sigma/\sqrt{2\log L_{\text{eff}}}$ (see, e.g., [10], Thm. 1.5.3). For the non-Gaussian sub-Gaussian case, see Remark 4.1.

Block decomposition. Choose $r_L \to \infty$ with $r_L = o(L)$ (e.g., $r_L = (\log L)^2$). Partition $\{1, \ldots, L\}$ into $k_L = \lfloor L/r_L \rfloor$ consecutive blocks $B_1, \ldots, B_{k_L}$ of size r_L, plus a negligible remainder.

Approximate independence. By the Leadbetter $D(u_n)$ condition (implied by $\sum |\rho(k)| < \infty$ via the Berman criterion), for any u:

$$\left| \mathbb{P}(M_L \le u) - [\mathbb{P}(M_{r_L} \le u)]^{k_L} \right| \to 0.$$

Block maximum. For each block B_j, $M_{r_L}^{(j)} = \max_{i \in B_j} z_i$. For i.i.d. $\mathcal{N}(0, \sigma^2)$ variables of the same marginal (the Gaussian-EVT correction for dependence affects only lower-order terms):

$$\mathbb{P}(M_{r_L}^{(j)} \le u) = \Phi(u/\sigma)^{r_L} \approx \exp\big(-r_L \bar{\Phi}(u/\sigma)\big)$$

for u in the right tail, where $\bar{\Phi}(t) = 1 - \Phi(t) \sim \phi(t)/t$ as $t \to \infty$.

Tail approximation. For $u = b_L + x a_L$ with $b_L = \sigma\sqrt{2\log L_{\text{eff}}}$ and $a_L = \sigma/\sqrt{2\log L_{\text{eff}}}$:

$$L_{\text{eff}} \cdot \bar{\Phi}(u/\sigma) \to e^{-x}.$$

Substituting $k_L \approx L_{\text{eff}}$ and using the block approximation:

$$\mathbb{P}(M_L \le b_L + x a_L) \to \exp(-e^{-x}) = G_0(x)$$

which is the standard Gumbel CDF. □

B Proof of Two-Sided Bounds

Upper bound. Softmax: $a_{i^*} \le e^{\mu_L}/(e^{\mu_L} + L)$. For $L \gg e^{\mu_L}$: $a_{i^*} \le e^{\mu_L}/L$. The noise tokens each receive $a_i \le 1/L + O(e^{\mu_L}/L^2)$. Hence $|a - u|_1 \le 2e^{\mu_L}/L + O(1/L^2) \le c_2 e^{\mu_L}/L$ for large L. The Lipschitz assumption yields $S_{\text{obs}}(L) \le K|a-u|_1 \le c_2 e^{\mu_L}/L$. This bound holds for all sufficiently large L.

Lower bound (pre-collapse regime only). Direct computation: $S_{\text{obs}}(L) = \sum_i a_i \log(L a_i) = a_{i^*} \log(L a_{i^*}) + \sum_{i \ne i^*} a_i \log(L a_i)$. For the noise sum: $\sum_{i \ne i^*} a_i \log(L a_i) \le 0$ when $a_i \le 1/L$ for most i. For the signal term: $a_{i^*} \ge e^{\mu_L}/(e^{\mu_L} + L) \ge e^{\mu_L}/(2L)$ for $e^{\mu_L} \le L$. And $\log(L a_{i^*}) \ge \log(e^{\mu_L}/2) = \mu_L - \log 2$. Hence $S_{\text{obs}}(L) \ge \frac{e^{\mu_L}}{2L}(\mu_L - \log 2) \ge c_1 \mu_L e^{\mu_L}/L$ for $\mu_L > \log 2 + 1$. For $\mu_L \le \log 2 + 1$ (i.e., $L \ge L_0$), this lower bound argument is invalid; only the upper bound is established. See Remark 4.3 for discussion and Remark 4.4 for the (unaffected) Impossibility Theorem. □

C Summary of Notation

Symbol	Meaning
$\mathcal{X}$, $\Delta^\circ(\mathcal{X})$	Token space; open probability simplex
p_θ, $\theta \in \Theta$	LLM output distribution; decoding parameters
Π_ε, $\widetilde{\Pi_\varepsilon p}$	Cognitive projection; uniform lift
$S_{\text{obs}}(p_\theta, \varepsilon)$	Observer entropy (partition-based; Definition 3.4)
$S_{\text{obs}}(L)$	Attention-induced observer entropy (Definition 4.7); see Remark 4.2
$I(\theta)$	Fisher information matrix
$v(\theta)$	Coarse-graining generator field
z_i, q, k_i	Attention logit; query; key vectors
μ_s, μ_L	Signal strength; effective margin
M_L	Noise logit maximum
$\sigma, \rho(\cdot)$	Sub-Gaussian parameter; correlation function
$L_{\text{eff}}, \tau_{\text{corr}}$	Effective length; correlation time
b_L, a_L	Gumbel normalising sequences
$L_{\text{crit}}, \mathcal{F}_L$	Critical context length (conjectured heavy-tailed); failure event
L_0	Threshold $\exp((\mu_s - \log 2 - 1)^2/(2\sigma^2))$ below which two-sided bound holds
$\mu_F(T), N_F(T)$	Lyapunov potential; fragmentation count
H_c, S_c, γ	CPL entropy threshold; stability threshold; drift threshold
$L_{\text{cap}}, L_{\text{warn}}$	Hard context cap; warning threshold