Forgetting as Regularisation in an Energy-Conserving Context Graph

Where bounded memory helps next-character prediction across data and capacity, and where local statistics meet their ceiling

Marcel Langjahr

Pantareon GmbH
https://pantareon.io/
langjahr@pantareon.io

June 2026

Abstract

We hold the substrate, the data, and the evaluation fixed, and vary only the memory regime. The substrate is a variable-order context graph with PPM-C backoff: non-backprop, online, count-based, with the topology grown from the stream and no learned weights. The regime is either unbounded accumulation (*plain*) or energy-conservation (*conserve*), in which each deposit is funded by a global decay, edges below a floor are pruned, and the node count is held at a ceiling by evicting the lowest-energy contexts, so that total energy reaches a flux-balance fixed point coupled to physical memory. We ask one question: does bounded forgetting help next-character prediction, and where?

Across two axes and three corpora the answer is consistent. On the *data-size* axis, forgetting lowers held-out bits-per-character at every corpus size, and the margin grows as data is scarcer—from a tie at 408k characters to -0.077 at 50k. On the *capacity* axis, forgetting shifts the bias–variance optimum to deeper contexts and flattens the curve: where plain counting peaks at order 4 and then drowns in variance, the conserving graph uses order 6 productively. The effect is robust over a factor of 25 in the decay rate (a clean U with an optimum), and over three corpora and two authors (eight windows, directionally identical, -0.07 to -0.12 bits/char at the best order). A second operation on the same substrate—backoff-local state merging by Jensen–Shannon divergence—roughly halves the node count at held quality and sharpens generation, with no sign of the predicted representational collapse: the suffix hierarchy absorbs the shared structure. But neither regime touches the ceiling we measure directly: at 11.3 million characters the model reaches 1.83 held-out bits/char and emits orthographically perfect, locally fluent text with no long-range coherence. Local statistics, perfected, are not meaning. The findings converge on one reading—forgetting is a regulariser that buys sample-efficiency exactly when capacity exceeds data, and the meaning ceiling is a property of the local, suffix-bounded substrate that no memory regime can lift.

Scope of this paper: The experiments below are deliberately small and single-machine: a zero-dependency count model over 10^5–10^7 characters, a single online pass, minutes on a CPU. They are not claims about frontier-scale behaviour. Their purpose is comparative and mechanistic—to isolate the contribution of the *memory regime* by holding the substrate, the data, and the evaluation fixed. Absolute bits-per-character figures should be read only relative to one another and to the stated baselines (uniform, unigram, order-1). The corpora are German—two by the present author, one a translation of a different author—so language-generality is indicated only to the extent that the effect survives a change of author and register; a non-German corpus and formal confidence intervals would be required to call it established.

1 Introduction

The substrate studied here descends from the same commitment as Emergent Graph Theory: that function is carried by a *discrete graph grown from data* under a conserved quantity, rather than by a fixed parametric model fitted by global optimisation. In the context graph this is literal. A node is a context—a suffix of the stream up to a maximum order D. An edge carries the *energy* of a next symbol in that context, which is simply its accumulated count. Prediction is a relaxation down the suffix chain (PPM-C

backoff): the longest matching context is blended with its shorter ancestors, the escape mass flowing to the root. There are no learned weights; the topology *is* the model.

To this substrate one can add an energy-conservation regime. Each new deposit is funded by a global decay—a forgetting at rate δ per symbol—edges that fall below a floor are pruned and return their energy to the pool, and the node count is held at a ceiling by evicting the lowest-energy contexts. Total energy then reaches a flux-balance fixed point (deposits in = decay out), and when the ceiling is set in physical memory the graph holds itself at that wall: its available energy *is* its memory allocation. This is the discrete, conserved analogue of a flux equilibrium, and it is the property the rest of the paper interrogates.

A natural hope follows: that forgetting, by discarding what is no longer reactivated, would behave as a regulariser and let the model generalise from less text. This paper tests that hope directly, and separates two things that are easily conflated. One is the *substrate*: the count graph and its PPM-C backoff. The other is the *memory regime*: unbounded accumulation versus conservation. By running the same substrate under both regimes on the same data, and by pushing the better regime to the limits of corpus size and model capacity, we can ask where bounded forgetting helps—and, where it does not, what that tells us about the substrate itself.

2 Experimental Setup

Common harness. Text is tokenised at the character level with a vocabulary read from the corpus; no external tokeniser is used. Two metrics are reported. *Held-out* bits-per-character (bpc) is the cross-entropy of next-character prediction on a validation tail the model never trains on; the train/validation split makes the distinction between memorisation and generalisation visible. *Prequential* bpc is a single online pass—predict each character, then deposit it—whose cumulative mean equals the code length an adaptive compressor reports, and is therefore directly comparable to gzip/bzip2/xz. Baselines are the uniform bound $\log_2 V$, the unigram (letter-frequency) entropy, and the order-1 conditional entropy.

Substrate. A variable-order context graph. Deposits are local and online: at each position the traversed contexts (orders 0 to D) each receive one unit of energy on the observed next symbol. Prediction blends the orders by PPM-C escape,

$$p \;\leftarrow\; \frac{c_s + q\, p_{\text{lower}}}{T + q}, \tag{1}$$

folded from the root upward, where c_s is the count of symbol s in the node, T its total, and q the number of distinct symbols seen (the escape count); the order--1 term is the uniform $1/V$. The substrate is identical in both regimes.

Plain regime. Counts only grow; the graph is unbounded. This is a clean adaptive PPM-C and serves as the control.

Conserve regime. Deposits are funded by a per-symbol decay δ (lazy, applied on touch and on a periodic sweep); edges below a floor ϕ are pruned; and the node count is capped—either by a hard node limit or by a resident-memory budget—by evicting the lowest-energy contexts down to a watermark. Decay, floor, and cap can be switched independently: with $\delta = \phi = 0$ and a memory budget the graph grows like *plain* until the wall, then evicts only under pressure; with $\delta, \phi > 0$ and no budget it forgets continuously regardless of memory. The eviction *score* is selectable: raw evidence mass (TOTAL); energy density $\rho = \text{total}/2^{H(p)}$ (mass per effective next-symbol, so concentration is preferred over frequency); or usage-weighted divergence from the backoff, usage $\times\, D_{\text{KL}}\big(p(\cdot \mid \text{context}) \,\|\, p(\cdot \mid \text{backoff})\big)$, which rewards a context for what it adds over its shorter one, mass aside.

State merging. A separate operation fuses contexts that share a direct backoff (the same suffix with the oldest symbol dropped) and predict almost identically. Similarity is the Jensen–Shannon divergence [5] between the siblings' next-symbol distributions (symmetric, a true metric root). The suffix hierarchy is used as a *free neighbourhood index*: only siblings under one backoff are candidates, so a pass is $O(N \times \text{group size})$, never all-pairs. A merge sums the members' evidence into one representative and aliases the others to it, so prediction for any member resolves to the pooled distribution. The pass is instrumented for a collapse failure mode (Section 5).

Corpora. *Codex*: a constructed-language / philosophical text, 107k characters, vocabulary 104 (locally irregular, incantatory). *Prose*: the author's German prose novel, 299k characters. *Mixed*: a 408k concatenation used for the capacity sweep. *External*: a German translation of an English fantasy series by

a different author, ≈11.3M characters, vocabulary 104. The first two are the author's own writing; the third controls for author and register.

3 Forgetting as Regularisation

The motivating observation is a single point that contradicts the naive expectation. On *Codex* at $D = 5$, conserving ($\delta = 8 \times 10^{-5}$, $\phi = 0.01$) reaches 2.367 prequential bpc against plain's 2.502—better prediction with *fewer* nodes (42.6k vs 56.8k). The reason is bias–variance: at order 5 on 10^5 characters most depth-5 contexts are seen once, and their counts are noise. Plain keeps these singletons and trusts them on the next occurrence; conserve lets them decay and prunes them via the floor, so the model falls back on the shorter, well-estimated context. Forgetting acts as regularisation. The rest of this section asks whether that single point is a curve, and along which axis.

Data-size axis. Holding $D = 5$ and varying the amount of *Prose* seen (conserve as a pure regulariser: decay and floor on, no budget binding), Table 1 shows conserve below plain at every size, with the margin largest in the middle of the range.

Table 1: Prequential bpc on growing prefixes of *Prose*, $D = 5$ (lower is better). Conserve: $\delta = 8 \times 10^{-5}$, $\phi = 0.01$.

chars	plain	conserve	Δ
10k	3.438	3.381	−0.057
25k	2.903	2.789	−0.114
50k	2.607	2.466	−0.141
100k	2.417	2.289	−0.128
200k	2.303	2.225	−0.078

The Δ is a hump, not a slope: it rises to 50k, plateaus, then contracts. The contracting right arm is the expected behaviour—with more data, plain needs the regularisation less. The rising left arm is subtler: below ∼50k there is too little recurring deep structure to separate from noise, so even the regulariser has little to act on. Sample-efficiency requires a minimum of data before it bites.

Capacity axis. Holding the corpus (*Mixed*, 408k) and sweeping the order D exposes two bias–variance U-curves (Table 2).

Table 2: Prequential bpc over context order on *Mixed* (408k). Plain peaks shallow; conserve peaks deep and flatter.

D	plain	conserve
2	2.729	2.725
3	2.313	2.346
4	**2.175**	2.202
5	2.208	**2.177**
6	2.298	**2.191**

The curves cross between $D = 4$ and $D = 5$. Plain's optimum is shallow ($D = 4$); above it, variance dominates ($D = 6$: 2.298). Conserve's optimum is deep ($D = 6$) and its U is much flatter—plain's steep climb at $D = 6$ becomes a gentle 2.191. *Forgetting shifts the capacity optimum upward*, which is what a regulariser is defined to do: it admits more model capacity before overfitting sets in. On this large corpus, however, plain's tuned best (2.175) edges conserve's tuned best (2.177)—a tie, not a win. The value of forgetting here is not "better than optimally tuned plain" but "robust to too-high D": at $D = 6$ conserve is 0.107 bpc better. It forgives the capacity choice.

The two axes are one surface. A single principle unifies them: the relevant quantity is the ratio of capacity to data. On the same *Mixed* corpus the sign of Δ flips between $D = 3$ (plain wins) and $D = 5$–6 (conserve wins); and as the corpus shrinks, plain's optimum moves to smaller D while conserve's advantage in the over-D region grows. The best-vs-best comparison across data sizes (each side at its own optimal D) makes this concrete (Table 3): the forgetting model's advantage grows monotonically as data is scarcer, and vanishes into a tie when data is plentiful.

Table 3: Best held-out / prequential bpc at each side's optimal D (lower is better). The advantage of forgetting grows as data shrinks.

corpus	plain best	conserve best	Δ
Prose 50k	2.513	2.436	−0.077
Prose 100k	2.339	2.289	−0.050
Mixed 408k	**2.175**	2.177	+0.002

Robustness. Three checks keep this from being a single corpus's quirk. *First*, a second author: the same 50k D-sweep on *External* (a different author, a different register) gives a conserve-best of 2.766 against a plain-best of 2.886 at $D = 4$, $\Delta = -0.120$—

larger than on the author's own prose, not smaller. *Second*, multiple windows: across three windows of *Prose* the best-D margin is -0.066 to -0.077, across two windows of *Codex* it is -0.108 to -0.121, and across three windows of *External* it is -0.105 to -0.120. The absolute bpc wanders with the window (some text is lighter), but Δ is stable, because window variance hits both regimes equally and cancels in the comparison (Table 4). *Third*, the decay rate: on *Codex* 50k at $D = 5$ (plain 2.727), conserve beats plain over a factor of 25 in δ, with a clean U and an optimum at 8×10^{-5} (Table 5)—the bias–variance signature of a genuine regularisation strength, not a tuned coincidence.

Table 4: Best-D margin Δ (conserve − plain) across corpora, authors, and windows. Directionally identical throughout.

corpus	author	windows	Δ (best D)
Prose	author A	3	$-0.066 \ldots -0.077$
Codex	author A	2	$-0.108 \ldots -0.121$
External	author B	3	$-0.105 \ldots -0.120$

Table 5: Decay sweep on *Codex* 50k, $D = 5$ (plain reference 2.727). A U with an optimum; robust over a factor of 25.

decay δ	conserve	vs plain
2×10^{-5}	2.655	−0.072
5×10^{-5}	2.590	−0.137
8×10^{-5}	**2.559**	−0.168
2×10^{-4}	2.584	−0.143
5×10^{-4}	2.721	−0.006

Taken together—two axes, three corpora, two authors, eight windows, and a decay U—the reading is that *bounded forgetting is sample-efficient in the overparameterised regime*: it helps exactly when capacity exceeds what the data supports, and ties otherwise. We make no claim beyond the measured trend. The corpora are all German prose (two by one author), there are no formal confidence intervals, and the absolute gain is modest (0.05–0.12 bpc); a non-German corpus and matched statistics would be needed to call it established rather than indicated.

4 Scale and the Ceiling

To see where the substrate stands, and where it stops, we train it on *External* at full size (11.3M characters, $D = 6$). A single online pass takes minutes on a CPU and yields a 1.33M-node, 3.0M-edge graph occupying 75 MB on disk and ∼389 MiB resident—no memory budget required at this scale. Three different numbers fall out of the same run, and they must be kept apart (Table 6).

Table 6: Three measurements on *External* (11.3M, $D = 6$). They answer different questions and must not be conflated. Baselines: uniform 6.700, bigram 3.489.

measurement	bpc	what it is
in-sample (2nd pass)	1.507	recognition
online / prequential	1.937	compression
held-out (full train)	**1.827**	generalisation

The one-pass property is real; the cold start is a measurement artifact. A natural objection is that the count model, like a gradient learner, might need more than one pass over the data. It does not—and this is provable rather than hoped. A second pass over the same text doubles every count, but the distribution $p(s) = c_s/T$ is unchanged because numerator and denominator scale identically. *The distributions are final after one pass*; counting saturates immediately, where gradient descent re-shapes the representation on every epoch and therefore needs many. The in-sample 1.507 is not learning; it is recognition—the graph predicting text it has already deposited. What the prequential measurement *does* contain is a cold start (the first ∼100k characters are predicted on a still-filling graph), but on 11.3M characters this lifts the average by only ∼0.02 bpc. The honest generalisation figure is the held-out 1.827, measured on a tail the model never trained on, and the validation curve is still falling at the end of training—the model is data-limited, not capacity-limited.

The ceiling. Generation from the full graph makes the limit audible. Proper names are reproduced correctly and in quantity; orthography is clean; local fragments of three to eight words read like the source. But grammatical agreement does not bind beyond the context window: within ∼6 characters everything is right, while subject–verb–object across a whole sentence dissolves. The output is a near-perfect local statistician and nothing more: *words right, sentences almost, meaning absent*. Crucially, this did not improve from 10^5 to 10^7 characters—two orders of magnitude—because the limit is not data but the absence of long-range binding. A character n-gram, perfected, does not become a model of meaning. This is the clean em-

pirical demonstration of where the method has its ceiling, and the ceiling is categorical, not a tuning gap.

External reference. On compression the substrate sits where a faithful PPM should. As orientation, published character-level figures on the standard enwik8 benchmark [11] run roughly: gzip $\sim$ 3.1, bzip2 $\sim$ 2.3, good PPM variants $\sim$ 1.9, small/medium character Transformers $\sim$ 1.1–1.4. The substrate's 1.83 held-out is PPM-level, as designed. A Transformer [10] would compress better ($\sim$0.4–0.7 bpc) and, more importantly, would climb the coherence axis the n-gram cannot—but it buys this with GPU-hours over many epochs, and from only 11M characters trained from scratch it would not reach its enwik8 best and would risk overfitting. The defensible positioning is therefore not "better than a Transformer" (it is not) but: the substrate reaches PPM-level prediction in *one CPU pass*, and its lever is sample-efficiency in the data-poor regime, not compression. We did not run a matched Transformer; doing so on the same corpus and hardware is the cleanest way to turn this orientation into a measurement.

5 State Merging

If the ceiling is the absence of binding between contexts, the obvious move is to let functionally equivalent contexts share evidence. Backoff-local merging (Section 2) does this cheaply, fusing only siblings under a common backoff. Two questions decide whether it is safe and whether it helps: does it collapse, and does it improve prediction or generation?

No collapse—the backoff is the vacuum. Merging by similarity has a built-in runaway: a fused node has a smoothed distribution, a smoothed distribution is similar to *more* other nodes, so a fused node could become an attractor that swallows the graph into a few diffuse mega-nodes (a gravothermal analogue). The pass is instrumented for exactly this: node count against mean predictive entropy. The measurement is unambiguous (Table 7): more aggressive merging *lowers* mean entropy rather than letting it run away, and lowers held-out bpc on the over-capacity model. Had it collapsed, entropy would rise and bpc break; both move the safe way. The suffix hierarchy is the implicit "vacuum"—shared structure flows to the backoff rather than concentrating in fused nodes—so no explicit barrier is needed. This was a real, testable failure mode, ruled out by measurement, not assumed away.

Table 7: Merge threshold sweep, held-out, *External* 50k at $D = 6$ (plain val 2.972). Aggressive fusion lowers nodes, bpc, *and* mean entropy—no collapse.

JS threshold	nodes	val bpc	mean entropy
no merge	68337	2.972	0.277
< 0.1	56657	3.029	0.313
< 0.3	53897	2.957	0.284
< 0.6	49352	2.869	0.253
< 1.0	47443	2.812	0.236

Compression, and a divergence between the number and the reader. On the full 11.3M graph, merging roughly halves the node count at held quality: JS<0.3 fuses 568k contexts (1.33M$\rightarrow$ 761k nodes) while mean entropy barely moves ($0.514 \rightarrow 0.532$); JS<0.6 fuses 686k ($\rightarrow$643k) and drives entropy down to 0.476. But held-out bpc and reading quality do not track each other. On the 50k held-out sweep, more aggressive merging keeps improving bpc; in full-corpus generation (controlled at one seed across thresholds), JS$\approx$0.3 reads best—the glossary-list artifacts and singleton slips of the unmerged model are smoothed away, words stay whole—while at JS<0.6 word breaks and odd compounds increase even as the entropy falls further. The mean entropy is nearest its unmerged value precisely at JS$\approx$0.3, i.e. concentration without smearing. The lesson is that bpc measures *average* prediction, not whole-word integrity; for the property that matters here—coherence—reader judgment is the more discriminating instrument, and the two can be pushed apart by over-fusion.

What merging is, and is not. It improves the over-capacity regime and produces a visibly cleaner local sampler at roughly half the nodes; on held-out bpc at 50k it does *not* unseat plain's tuned shallow optimum (2.812 at $D = 6$ vs plain 2.723 at $D = 4$), exactly as conserve ties rather than wins at large data. And it does not lift the ceiling: it fuses suffix-sharing siblings, not suffix-disjoint semantic relatives, so meaning remains absent. It is a sharper statistician, not a step across the categorical boundary.

6 Transfer and Selective Forgetting

The conservation regime is, by construction, a small model of attention-driven memory: what is reactivated persists, what is not fades, and finite capacity forces the weakest out. We probe it with cross-corpus accumulation. Loading the *Codex* graph and then streaming *Prose* merges the vocabularies cleanly (Codex's 104 symbols extend to 117 as Prose's Latin-1 characters are appended, old IDs stable) and transfers real predictive structure: *Prose* is predicted at 2.090 prequential bpc with Codex preloaded against ~ 2.25 alone—about 0.16 bpc of shared German short-context structure carried across.

Forgetting and accumulation pull against each other—correctly. Under conservation the transfer survives but fades: the resume starts $\sim$0.6 bpc lower than a cold start, but as *Prose* never reactivates Codex's distinctive contexts they decay and are pruned, and the final graph ($\sim$35k nodes) is indistinguishable in size from *Prose* alone. Generation confirms it at the ear: a query that drew Codex's incantatory register from the fresh Codex graph draws *Prose*'s narrative register from the faded one. This is not a defect once the goal is stated. A system that held the first corpus perfectly forever, regardless of how much of the second arrived, would be a database, not a memory; the fading of unreactivated material under new attention is the function. And the fading is *selective*, not uniform: a name shared by both corpora is reactivated and survives, while only the non-overlapping material interferes away—overlap consolidates, the unique is overwritten, which is the structure of retroactive interference [9] rather than passive decay.

The measure is frequency, not yet salience. Under a memory ceiling we tested all three eviction scores on the *Codex*→*Prose* stream at a binding 18 MiB budget. They give essentially the same held-out bpc (2.117–2.118) and the same faded register in generation, differing only in equilibrium topology—mass keeps a few fat hubs, density keeps more, thinner, sharper nodes ($\rho = \text{total}/2^H$ does shift *which* contexts survive), and value (KL from backoff) trusts singletons. None preserves Codex's distinctive material, because that material is low-frequency in the combined stream (seen once in Codex, never refreshed by Prose), so its raw mass is small and it sits below the eviction edge under any score that keeps mass in the numerator. The honest conclusion is that the present forgetting is *frequency-driven*, where a biological memory consolidates by *salience*—surprise, prediction error, reward. The right open question is therefore not how to prevent fading (that was a mis-framing) but what salience measure the system should forget by; KL from the backoff—"keep what surprised you relative to the shorter context"—is the principled next candidate, but on these data it changed topology without yet changing bpc, which is itself the warning that changing the measure is not the same as predicting better.

7 Discussion

Two findings reinforce each other. Bounded forgetting buys sample-efficiency exactly where model capacity exceeds the data: it lowers held-out bpc across data sizes (margin growing as data shrinks), shifts and flattens the capacity optimum, and is robust over corpora, authors, and decay rate. Backoff-local merging sharpens the same substrate—roughly halving the node count at held quality, with no collapse because the suffix hierarchy absorbs shared structure. Both are operations *on* the count graph, not changes of substrate; the energy-conservation prior earns its place by measurably changing behaviour (which contexts survive, how the bias–variance optimum moves), which is the only licence a structural prior should have.

But the same experiments draw a hard line. The meaning ceiling—words right, sentences almost, meaning absent—did not move from 10^5 to 10^7 characters, and no memory regime touched it. It is a property of the local, suffix-bounded substrate: prediction relaxes down a suffix chain, and contexts that share no suffix share no path. This mirrors, on a different mechanism, the lesson of the companion study of this substrate family, where holding architecture fixed and varying the learning rule located the language capability in gradient-based credit assignment and not in the dynamics. Here, holding the substrate fixed and varying the memory regime locates a real but bounded benefit—sample-efficiency under over-capacity—in forgetting, and locates meaning in *neither* regime: it requires representational machinery the count graph does not have. Both readings are a local instance of the general pattern that search and learning over generic computation outperform hand-designed structural priors [12]; the conservation prior helps precisely as a regulariser and not as a route to understanding.

The energy/EGT reading survives this honestly. The RAM-coupled regime is a genuine flux-balance equilibrium—the graph breathes at its memory wall

rather than growing without bound—and a worry imported from the cosmological setting (that fusion-like merging would collapse to a singular attractor unless a vacuum degree of freedom enforces separation) turned out to name a real, testable failure mode that the measurement ruled out: the backoff hierarchy already plays the vacuum's role. But the analogy is *generative*, not load-bearing. Density changed topology without changing bpc; the collapse did not occur. The discipline the project requires is to keep the physics as a source of questions, not of answers—to let "conservation", "density", "vacuum" point to mechanisms that must then earn their keep by measurably changing the held behaviour, and to discard the ones that only make the picture round.

This locates the genuine next decision, and it is a substrate question, not a tuning one. Crossing the meaning ceiling needs either long-range binding—attention, which leaves this substrate—or generalisation between suffix-disjoint contexts, which needs a *geometry* on the node space that a localist suffix tree does not possess. The suffix hierarchy is a free neighbourhood index for suffix-sharing contexts, and merging exploits it; but two contexts with no common suffix yet related meaning are unreachable, and brute-force neighbourhood search is $O(N^2)$, infeasible at 10^6 nodes. The candidates are a metric index over the distribution space (locality-sensitive hashing [7]) or a sparse distributed representation in which overlap *is* the neighbourhood [6, 8]—the latter a true substrate change, lossy by design. Which to take is the open architectural choice; the present results say only that the count substrate has been brought cleanly to its ceiling, and that the ceiling is where local statistics end.

8 Conclusion

Holding the substrate, the data, and the evaluation fixed and varying only the memory regime, we find that bounded forgetting is a regulariser that buys sample-efficiency exactly when model capacity exceeds the data. On the data-size axis its advantage over unbounded counting grows as data shrinks (from a tie at 408k to -0.077 at 50k); on the capacity axis it shifts the bias–variance optimum to deeper contexts and flattens the curve; and the effect is robust over three corpora, two authors, eight windows, and a factor of 25 in the decay rate, at -0.07 to -0.12 bits/char at the best order. A second operation on the same substrate, backoff-local state merging, roughly halves the node count at held quality without the predicted representational collapse. Neither regime lifts the ceiling we measure directly: at 11.3 million characters the model reaches 1.83 held-out bits/char and produces orthographically perfect, locally fluent, globally meaningless text—local statistics, perfected, are not meaning, and the limit is categorical. The result is constructive for an energy-conserving count graph as a sample-efficient, one-pass, CPU-cheap local statistician, and negative for the strong thesis that meaning emerges from the substrate's own dynamics. Establishing whether the sample-efficiency margin holds against a Transformer on the same corpus and hardware, and whether a geometry on the node space can cross the ceiling without abandoning the count substrate, is the natural next measurement.

Code availability. The model is a single zero-dependency Rust crate (`ctxgraph`): the variable-order context graph, the PPM-C blend of Eq. (1), the conservation regime with the three eviction scores, backoff-local merging, and checkpointing with vocabulary and alias persistence. All figures above are produced by its held-out and whole-file prequential harnesses.

References

[1] von der Malsburg, C., "The Correlation Theory of Brain Function", in *Models of Neural Networks II*, E. Domany, J. L. van Hemmen, K. Schulten (eds.), Springer, New York, pp. 95–119 (1994); orig. MPI für Biophysikalische Chemie, Internal Report 81-2 (1981).

[2] Shannon, C. E., "Prediction and Entropy of Printed English", *Bell System Technical Journal* **30**(1), 50–64 (1951).

[3] Cleary, J. G., Witten, I. H., "Data Compression Using Adaptive Coding and Partial String Matching", *IEEE Transactions on Communications* **32**(4), 396–402 (1984).

[4] Moffat, A., "Implementing the PPM Data Compression Scheme", *IEEE Transactions on Communications* **38**(11), 1917–1921 (1990).

[5] Lin, J., "Divergence Measures Based on the Shannon Entropy", *IEEE Transactions on Information Theory* **37**(1), 145–151 (1991).

[6] Kanerva, P., *Sparse Distributed Memory*, MIT Press, Cambridge, MA (1988).

[7] Charikar, M. S., "Similarity Estimation Techniques from Rounding Algorithms", *Proceedings of the 34th Annual ACM Symposium on*

Theory of Computing (STOC), pp. 380–388 (2002).

[8] Plate, T. A., "Holographic Reduced Representations", *IEEE Transactions on Neural Networks* **6**(3), 623–641 (1995).

[9] McCloskey, M., Cohen, N. J., "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem", *Psychology of Learning and Motivation* **24**, 109–165 (1989).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., "Attention Is All You Need", *Advances in Neural Information Processing Systems* **30** (2017).

[11] Mahoney, M., "Large Text Compression Benchmark", `http://mattmahoney.net/dc/text.html` (accessed 2026).

[12] Sutton, R. S., "The Bitter Lesson", *incompleteideas.net/IncIdeas/BitterLesson.html* (2019).