

The Softmax Inundation Bound

Attention Dilution in Long-Context Transformers and
Generational Context as a Volume-Bounding Remedy

Shaurya Sanyal

Team Auralis

shauryasanyal@gmail.com

October 2026

Abstract

We isolate and formalize a simple but under-appreciated consequence of softmax normalization in transformer self-attention: when a fixed set of *invariant* tokens competes with a growing set of distractor tokens, the aggregate attention mass on the invariant set is squeezed by the normalization denominator. We call this *softmax inundation*. Under the standard bounded-projection-energy assumption we prove

$$A_{i,\text{inv}} \leq \frac{\kappa L_u}{L_u + |\mathcal{S}_{\text{noise}}|} = \mathcal{O}\left(\frac{L_u}{|\mathcal{S}_{\text{noise}}|}\right), \quad \kappa = e^{2M_Q M_K / \sqrt{d_k}},$$

so that the invariant share decays at least as fast as the inverse of the distractor count.

Because the inequality is conditional on uniform score bounds and can be uninformative when score contrast is large, we do not present it as a new law of attention. Its value is diagnostic: it isolates *token volume* as the variable a system designer can actually control.

We then introduce **Generational Context**, a three-tier agent memory architecture (Gen-0 ephemeral tool output with TTL distillation, Gen-1 working memory paged by Attention-Weighted Eviction, Gen-2 permanently pinned invariants), and prove that under a strict bound C_{bound} on active non-invariant volume, invariant attention admits a *turn-independent* lower bound,

$$A_{i,\text{inv}}^{(\text{GC})} \geq \frac{L_u}{L_u + \kappa C_{\text{bound}}} > 0 \quad \text{for all } t \geq 1.$$

This is the substantive claim: bounding volume removes horizon-dependence from the worst case. It is a guarantee about attention allocation, not about task success. We report algorithmic-simulation results, describe a reference implementation with cryptographic intent auditing, and state precisely what remains unverified.

1 Introduction

Autonomous agents built on transformer language models typically treat the context window as an append-only transcript. Every tool call, every stack trace, every intermediate hypothesis is appended and never removed. This is convenient to implement and expensive to sustain.

Two things are routinely conflated in discussion of this practice:

- (i) **Capacity.** Whether the transcript still fits in the window.
- (ii) **Allocation.** Whether attention still lands on the tokens that matter.

Work on long-context models has largely addressed (i): windows grow, and retrieval quality within a window is measured by benchmarks such as RULER [1] and by position-sensitive

studies such as *Lost in the Middle* [2]. But (ii) is a separate phenomenon, and it is governed by the softmax denominator rather than by the window size.

Our contributions are deliberately modest and, we hope, clearly separated:

- §3 states and proves the *Softmax Inundation Bound*. It is elementary; we claim only that it is a useful way to isolate the mechanism, not that the mechanism was previously unknown.
- §4 introduces Generational Context and proves the *Generational Retention Guarantee*, a horizon-independent lower bound on invariant attention under bounded active volume. This is the result we consider the paper’s actual contribution.
- §5–7 describe the eviction policy, the reference implementation, and a simulation harness.
- §8 is an explicit limitations section. We regard it as part of the contribution rather than as boilerplate.

What this paper does not claim. It does not claim that attention dilution causes reasoning failure in any particular model; residual streams, induction heads, and tool use can all compensate. It does not claim empirical gains on live models; we have not run those experiments. It does not claim novelty for softmax competition itself.

2 Preliminaries

2.1 Interaction transcript

Let an agent interact with an environment over discrete turns $t \in \mathbb{N}$. At turn t the context is

$$\mathcal{H}_t = (\mathbf{u}, \mathbf{a}_1, \mathbf{o}_1, \mathbf{r}_1, \dots, \mathbf{a}_{t-1}, \mathbf{o}_{t-1}, \mathbf{r}_{t-1}), \quad (1)$$

where $\mathbf{u} \in \mathcal{V}^{L_u}$ holds the user objective and system invariants, $\mathbf{a}_k$ are actions, $\mathbf{o}_k$ are raw environmental observations, and $\mathbf{r}_k$ are reasoning traces. The context length is

$$N_t = L_u + \sum_{k < t} (L_a^{(k)} + L_o^{(k)} + L_r^{(k)}). \quad (2)$$

2.2 Scaled dot-product attention

For query position i and key position j ,

$$A_{i,j} = \frac{\exp(s_{i,j})}{\sum_{m=1}^{N_t} \exp(s_{i,m})}, \quad s_{i,j} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_k}}, \quad (3)$$

with $\mathbf{q}_i = \mathbf{X}_i W_Q$ and $\mathbf{k}_j = \mathbf{X}_j W_K$.

2.3 Partition of context

Fix a turn t and partition the index set $J = \{1, \dots, N_t\}$ into three mutually disjoint subsets:

$$\mathcal{S}_{\text{inv}} : \text{invariant tokens}, \quad |\mathcal{S}_{\text{inv}}| = L_u, \quad (4)$$

$$\mathcal{S}_{\text{noise}} : \text{raw observations}, \quad |\mathcal{S}_{\text{noise}}| = \sum_k L_o^{(k)}, \quad (5)$$

$$\mathcal{S}_{\text{work}} : \text{actions and reasoning}, \quad |\mathcal{S}_{\text{work}}| = \sum_k (L_a^{(k)} + L_r^{(k)}). \quad (6)$$

Write $A_{i,\text{inv}} = \sum_{j \in \mathcal{S}_{\text{inv}}} A_{i,j}$.

Assumption 1 (Bounded projection energy). There exist $M_Q, M_K > 0$ such that $\|\mathbf{q}_i\|_2 \leq M_Q$ and $\|\mathbf{k}_j\|_2 \leq M_K$ for all i, j . Consequently, by Cauchy–Schwarz,

$$-B \leq s_{i,j} \leq B, \quad B = \frac{M_Q M_K}{\sqrt{d_k}}. \quad (7)$$

Define the dynamic range factor $\kappa = e^{2B}$.

Remark 1. Assumption 1 is weak but not innocuous. Weight matrices in trained transformers are finite, so it always holds for *some* M_Q, M_K ; the question is whether the resulting κ is small enough for the bound to say anything. See §8.

3 The Softmax Inundation Bound

Theorem 1 (Softmax Inundation Bound). *Under Assumption 1, for every query position i and every turn t ,*

$$A_{i,\text{inv}} \leq \frac{\kappa L_u}{L_u + |\mathcal{S}_{\text{noise}}|}, \quad \kappa = e^{2M_Q M_K / \sqrt{d_k}}. \quad (8)$$

In particular, holding L_u and the scoring regime fixed,

$$A_{i,\text{inv}} = \mathcal{O}\left(\frac{L_u}{|\mathcal{S}_{\text{noise}}|}\right) \quad \text{as } |\mathcal{S}_{\text{noise}}| \rightarrow \infty. \quad (9)$$

Proof. Bounds on $s_{i,j}$ give $e^{-B} \leq e^{s_{i,j}} \leq e^B$ for all j . Hence

$$\sum_{j \in \mathcal{S}_{\text{inv}}} e^{s_{i,j}} \leq L_u e^B, \quad (10)$$

and, dropping the non-negative $\mathcal{S}_{\text{work}}$ terms from the denominator of (3),

$$\sum_{m=1}^{N_t} e^{s_{i,m}} \geq (L_u + |\mathcal{S}_{\text{noise}}|) e^{-B}. \quad (11)$$

Dividing,

$$A_{i,\text{inv}} \leq \frac{L_u e^B}{(L_u + |\mathcal{S}_{\text{noise}}|) e^{-B}} = \frac{e^{2B} L_u}{L_u + |\mathcal{S}_{\text{noise}}|}. \quad (12)$$

With $\kappa = e^{2B}$ this is (8). $\square$

Remark 2 (Reading the bound correctly). Equation (8) is an *upper* bound on the invariant share. It is informative when κ is moderate; it becomes vacuous as soon as $\kappa L_u \gtrsim L_u + |\mathcal{S}_{\text{noise}}|$, at which point the right-hand side exceeds 1 and the inequality carries no information. This is not a defect of the proof but a genuine limitation of the argument: bounding norms alone does not force dilution. If distractor scores are systematically much smaller than invariant scores, $A_{i,\text{inv}} \rightarrow 1$ regardless of $|\mathcal{S}_{\text{noise}}|$.

Remark 3 (Relationship to known results). Softmax competition of this form is implicit in any analysis of attention with a growing key set, and explicit in streaming and long-context work [5, 2]. We do not claim the mechanism as new. The useful part, in our view, is the separation of *volume* (which a system controls) from *score contrast* (which it largely does not), because only the former is addressable by memory management.

4 Generational Context

If Theorem 1 identifies volume as the controllable variable, the natural response is to control it. Generational Context borrows the shape of generational garbage collection and of virtual-memory paging, and applies both at the granularity of agent context rather than of KV blocks.

Definition 1 (Generational Context). Active context is partitioned into three tiers:

Gen-1 *Tenured*. Founding objectives and system invariants. Pinned; never evicted. Occupies L_u tokens.

Gen-2 *Evaporative*. Working memory: hypotheses, plans, uncommitted deltas. Paged by Attention-Weighted Eviction (§5) under a strict capacity C_1 .

Gen-3 *Ephemeral*. Raw tool output and logs. Passed through a distillation map σ_{distill} producing at most K_0 tokens, with the raw form evicted under a turn-based TTL.

Let C_{overhead} bound the structural tokens (role tags, delimiters, section headers) required to render the window, and set

$$C_{\text{bound}} = C_1 + K_0 + C_{\text{overhead}}. \quad (13)$$

Theorem 2 (Generational Retention Guarantee). *Under Assumption 1, and given a Generational Context with $C_{\text{bound}} < \infty$,*

$$A_{i,\text{inv}}^{(\text{GC})} \geq \frac{L_u}{L_u + \kappa C_{\text{bound}}} > 0 \quad \text{for all } t \geq 1. \quad (14)$$

In particular the bound is independent of the interaction horizon t .

Proof. By Definition 1, at every turn the number of active non-invariant tokens satisfies

$$|\mathcal{S}_{\text{noninv}}^{(t)}| \leq C_1 + K_0 + C_{\text{overhead}} = C_{\text{bound}} \quad (15)$$

independently of t . Set

$$R_t = \frac{\sum_{m \notin \mathcal{S}_{\text{inv}}} e^{s_{i,m}}}{\sum_{j \in \mathcal{S}_{\text{inv}}} e^{s_{i,j}}}, \quad \text{so that} \quad A_{i,\text{inv}} = \frac{1}{1 + R_t}. \quad (16)$$

The numerator has at most C_{bound} terms, each at most e^B ; the denominator has L_u terms, each at least e^{-B} . Therefore

$$R_t \leq \frac{C_{\text{bound}} e^B}{L_u e^{-B}} = \frac{\kappa C_{\text{bound}}}{L_u}. \quad (17)$$

Since $R \mapsto 1/(1 + R)$ is decreasing on $[0, \infty)$,

$$A_{i,\text{inv}} \geq \frac{1}{1 + \kappa C_{\text{bound}}/L_u} = \frac{L_u}{L_u + \kappa C_{\text{bound}}}. \quad (18)$$

Because C_{bound} is fixed by the memory controller, the right-hand side does not depend on t . $\square$

Corollary 1 (No horizon decay). *Under the hypotheses of Theorem 2, $\inf_{t \geq 1} A_{i,\text{inv}}^{(t)} > 0$. That is, unbounded softmax inundation cannot occur, in the worst case, once active non-invariant volume is bounded.*

Remark 4 (What the guarantee is and is not). The bound is a *worst-case* statement under uniform score assumptions. It does not say invariant attention is large; it says it cannot be driven to zero by volume alone. Task success additionally requires correct retrieval, valid tool use, and sound multi-step reasoning, none of which this theorem addresses. An adversary who controls distractor content rather than distractor count can violate the spirit of the result by inflating $s_{i,m}$ while staying inside Assumption 1; see §8.

5 Attention-Weighted Eviction

Theorem 2 is a statement about a controller that actually maintains the volume bound. We use an adaptation of ARC [3], which already solves the *recency versus frequency* allocation problem for paged memory.

The cache maintains a recency list T_1 , a frequency/saliency list T_2 , ghost registries B_1, B_2 recording recently evicted keys, and an adaptation parameter p targeting the size of T_1 . Victim selection within a list uses an *Information Density Score*

$$\text{IDS}(E) = \frac{\text{AttnWeight}(E) \cdot 100}{\text{Entropy}(E) + 10^{-4}}, \quad (19)$$

with tier-based priors for entries whose attention statistics have not yet been measured. The cache enforces the post-condition

$$\sum_{e \in T_1 \cup T_2} \text{tokens}(e) \leq C \quad \text{after every operation,} \quad (20)$$

truncating any single entry larger than C before insertion.

Two implementation notes matter for correctness. First, promoting an entry from T_1 to T_2 must evict *other* entries if the promotion grows the footprint; a naive implementation can select the newly promoted entry as its own victim. Second, ghost registries must be bounded, or the eviction bookkeeping grows without limit even though the active cache does not.

6 Simulation Results

We report results from a deterministic simulation harness (`ccb_bench.py`) that models three arms over scripted multi-turn scenarios: flat append-only, periodic compaction, and Generational Context. Attention quantities are computed with 128-dimensional hash-projected embeddings and an explicit softmax, not with trained transformer weights. Table 1 reports final-turn means across three scenarios.

Arm	Active tokens	CSR	Attn(inv)	Cost (sim.)
Flat append-only	5,099	0.0059	0.0425	\$0.0459
Periodic compaction	3,340	0.0077	0.0553	\$0.0301
Generational Context	129	0.2000	0.0757	\$0.0012

Table 1: CCB-001 simulation, final-turn means over three scripted scenarios. “CSR” is the context signal ratio; “Attn(inv)” is the simulated attention share on invariant tokens. These are algorithmic proxies, not neural measurements. Absolute values are not comparable to published transformer results; only the ordering across arms is meaningful.

What the simulation supports. The harness shows that the memory policy realizes its volume target under scripted load, and that the resulting active footprint is roughly two orders of magnitude smaller than flat append-only. That is a statement about the controller.

What it does not support. It does not demonstrate that a trained transformer allocates attention to invariants in the same proportions, nor that task success improves. Those require live-model experiments, which we have not performed.

7 Reference Implementation and Intent Auditing

The architecture is implemented in the `generational_context` package (Python standard library only). The controller exposes tool observation, reasoning recording, and prompt compilation; it enforces the capacity invariant of §5 and the TTL behaviour of Gen-0.

A separate benchmark exercises long-horizon intent conservation. In its current form it uses Ed25519-signed intent tokens over a canonical digest of the declared goal, values, constraints, and priorities, together with an append-only hash-chained audit log. A proposal is accepted only if its token verifies against the current intent digest and its signatory holds the required role. Earlier revisions of this component used an unsalted digest of static strings, which provided no security; that construction has been removed.

We are explicit that this is an engineering component, not a validated security boundary. Its threat model is undocumented and its scenarios are authored by the same team that wrote the checker.

8 Limitations

We prefer to state these plainly rather than to bury them.

1. **The inundation bound is elementary and conditional.** Theorem 1 follows from Cauchy–Schwarz and the definition of softmax. It is informative only when κ is moderate; for realistic projection norms with B even of order 10, κ exceeds 10^8 and the bound is vacuous at practical context lengths. The theorem is best read as a formalization of a mechanism, not as an explanation of observed degradation at a particular scale.
2. **Uniform score assumptions are strong.** Both theorems assume the same worst-case score envelope for invariant and distractor tokens. Real attention scores are highly structured, and the ordering of $\mathcal{S}_{\text{inv}}$ versus $\mathcal{S}_{\text{noise}}$ in score space is exactly what determines the outcome. Our bounds are agnostic to that ordering, which is both their strength and their weakness.
3. **Multi-head and layer structure is ignored.** We analyze a single head in isolation. Specialization across heads (induction, copy, positional sinks) can preserve information that a per-head bound suggests is lost.
4. **No live-model validation.** All quantitative results come from a deterministic simulation with hash-based embeddings. We make no claims about behaviour of trained models, and we would regard any such claim drawn from these numbers as unsupported.
5. **The retention guarantee is about allocation, not success.** Keeping invariants in the attention distribution is necessary at best. An agent can attend to its instructions and still fail.
6. **Adversarial distractor content is out of scope.** A bounded number of very high-scoring distractors can dominate attention. Volume control does not defend against content attacks.

Open empirical questions. Does bounded-volume management measurably reduce goal drift in long agent sessions? Does distillation of Gen-0 preserve the error signature that made a failure diagnosable? Does AWE’s IDS criterion outperform plain recency at equal capacity? Each is answerable with a controlled experiment and none is answered here.

9 Related Work

Long-context behaviour. Position-sensitive degradation is documented in *Lost in the Middle* [2]; effective-context benchmarks show that nominal window size overstates usable context [1]. Streaming inference with attention sinks [5] addresses the degenerate-statistics problem of unbounded KV growth. Our concern is different in emphasis: not how to keep generating coherently from a long stream, but how to keep a bounded working set aligned with declared invariants.

Agent memory. MemGPT [4] introduces virtual context management through function calls, with paging decisions made by the model itself. Generational Context makes those decisions with a deterministic controller instead, trading flexibility for predictable cost. Pinning invariants corresponds to MemGPT’s always-in-context core memory; the difference is the explicit tiering and the TTL treatment of raw observations.

Cache replacement. ARC [3] supplies the adaptation mechanism. Attention-score-based eviction for KV caches is well established in the inference-efficiency literature. We apply the same idea at the granularity of semantic context entries rather than KV blocks.

10 Conclusion

Softmax normalization means that adding distractor tokens competes with invariant tokens for a fixed budget of probability mass. We state that competition as an explicit bound, note carefully where the bound is informative and where it collapses, and then prove the complementary result: if a controller holds active non-invariant volume below a constant, the worst-case invariant share is bounded away from zero independently of how long the session runs.

The second result is the one we would defend. It does not require any claim about model internals beyond bounded projections, it identifies a quantity a system can actually control, and it comes with an implementation whose capacity invariant is enforced rather than assumed.

What it does not come with is evidence that any of this improves outcomes on real tasks. That is the next experiment, and we would rather state the gap than paper over it.

Acknowledgments

The term *Artificial Civilization Intelligence* (ACI) was coined by Raja Dharma Tej Maddala. The architecture, proofs, implementation, and simulation harness in this document were produced by Team Auralis.

References

- [1] C.-P. Hsieh, S. Sun, S. Krizan, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. RULER: What’s the real context size of your long-context language models? *arXiv:2404.06654*, 2024.
- [2] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the ACL*, 2024. arXiv:2307.03172.
- [3] N. Megiddo and D. S. Modha. ARC: A self-tuning, low overhead replacement cache. In *Proceedings of FAST*, 2003.

- [4] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv:2310.08560*, 2023.
- [5] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient streaming language models with attention sinks. In *Proceedings of ICLR*, 2024. arXiv:2309.17453.