PA-STREAM: Position-Adaptive Hierarchical Sparse Tracing for Uniform Long-Context Fidelity

Liz Lemma Future Detective

January 18, 2026

Abstract

Mechanistic interpretability techniques that inspect attention patterns scale quadratically in context length T, making million-token analysis infeasible. STREAM (Rosser et al., 2025) leverages hierarchical sparse attention masks to trace salient long-context attention in near-linear time, but exhibits a systematic failure mode under causal masking: pruning becomes effectively more aggressive for late queries, degrading retrieval near the end of long contexts (e.g., needle-in-a-haystack). We formalize this as a position-induced bias caused by the expanding set of admissible causal keys. We introduce PA-STREAM, a position-adaptive sparsity schedule $k(q)$ (optionally per layer/head) that increases mildly with query position, preserving uniform retrieval and tracing fidelity across the entire context while retaining near-linear scaling. Our theory models hierarchical pruning as selection under noisy branch-score estimates and proves that $k(q) = \Theta(\log |\mathcal{V}(q)|)$ is sufficient—and in a natural sense necessary—for uniform retention of behavior-critical branches when only $O(1)$ representatives per branch are scored. Empirically (to be added), PA-STREAM eliminates STREAM's late-context degradation on RULER and improves long-context tracing stability on real RAG corpora while still pruning >95

Table of Contents

3. 3. Clean Problem Formulation: (a) uniform-fidelity sparse tracing as constrained sparsification; (b) branch-selection under noisy maxima; (c) effective sparsity and position bias under causal masking.

4. 4. Why Fixed-k Fails (Formal): prove monotone growth of admissible keys; define effective retention ratio; provide worst-case and distributional lower bounds showing constant-k cannot maintain uniform success as T grows under noisy estimation.

5. 5. PA-STREAM Algorithm: position-adaptive schedules (logarithmic, candidate-count-based, entropy-based); optional per-layer/head calibration; global budgeted variant; implementation notes (keeping first ℓ_d layers dense; interaction with block sizes).

6. 6. Main Theorems (Upper Bounds): sufficient conditions for uniform retention across query positions; runtime/space; tradeoffs among $k(q)$, representative count r, and failure probability δ.

7. 7. Lower Bounds and Tightness: explicit instance family for hierarchical selection with noisy estimates requiring $k(q) = \Omega(\log |\mathcal{V}(q)|)$ (or $r = \Omega(\log |\mathcal{V}(q)|)$) to achieve constant success; discussion of optimality up to constants/polylogs.

8. 8. Experimental Design (Strengthening Evidence): RULER depth-vs-length sweeps; compare fixed-k STREAM vs PA-STREAM; real RAG corpora; metrics (retrieval success, logit/KL fidelity, mask sparsity, wall-clock, memory); ablations (schedule types, block sizes, dense early layers).

9. 9. Discussion: implications for million-token monitoring; when log-growth might be insufficient; relation to retrieval heads/attention sinks; limitations (attention-only tracing; dependence on separation/noise assumptions).

10. 10. Conclusion and Future Work: extending adaptive schedules to variable-per-iteration branching; combining with logit-faithful certificates; incorporating residual/MLP tracing.

1 Introduction

We consider the problem of scaling mechanistic interpretability and long-context evaluation for decoder-only transformers beyond regimes where dense attention is computationally convenient. In such settings, it is natural to replace the dense causal attention pattern with a learned or data-dependent sparse mask that preserves a prescribed notion of fidelity while reducing runtime and memory. The particular fidelity notion we have in mind is position-uniform: for query positions near the output of the model, the sparse computation should retain the blocks that carry behavior-critical information (e.g. the block containing a "needle" token in retrieval benchmarks, or the blocks supporting a specific causal influence path) with probability at least $1 - \delta$, uniformly as the context length T increases. This uniformity requirement is not cosmetic. If failure probability increases with position, then any long-context conclusion derived from sparse masking becomes systematically biased toward early tokens, and interpretability claims cease to extrapolate to the long-context regime of interest.

Hierarchical mask estimators in the style of STREAM/HiP are appealing because they admit a streaming, blockwise implementation: for each query block one does not score all admissible key blocks, but rather organizes them into a refinement tree and progressively narrows attention to a subset of promising candidates using representative sampling. In practice, such procedures can achieve substantial acceleration while often maintaining good empirical accuracy on short and moderate contexts. However, in the causal setting there is a structural asymmetry that becomes dominant at long context length: the admissible region for a query expands monotonically with position, forming a causal triangle in the token–token plane (and its blockwise analog). As a consequence, the number of admissible key blocks $|\mathcal{V}(q)|$ grows with q. Any method that retains a position-independent number k of key blocks per query block therefore allocates a vanishing fraction $k/|\mathcal{V}(q)|$ of the admissible past to late queries. In the absence of strong additional side information that reliably localizes behavior-critical blocks, this induces a position bias: late queries are intrinsically harder, and sparse estimators tuned for early queries will eventually fail on late ones.

The same phenomenon persists even when the estimator is adaptive in a limited sense. A hierarchical method may be able to focus computation on a subset of branches whose estimated scores $\widehat{M}$ are large, but if these estimates are derived from a constant number of representatives per branch, they inherit nontrivial noise. In the long-context regime the number of competing branches increases, and extreme-value effects imply that some noncritical branch will be spuriously overestimated with nonnegligible probability. If we prune aggressively by keeping only a sublogarithmic number of branches at any refinement level, then the critical branch can be eliminated early, after which later refinements cannot recover it. Thus there are two coupled

sources of degradation under causal masking: (i) the combinatorial growth of admissible keys with position, and (ii) the statistical growth of misleading competitors under noisy branch estimation. Taken together, these effects explain a characteristic failure mode observed in practice: sparse attention may appear faithful on shallow needles and early positions, yet exhibit sharp drops in retrieval or behavioral preservation at late positions, even when average accuracy metrics remain acceptable.

Our approach is to make this failure mode explicit and to design a schedule that neutralizes it by allocating sparsity budget as a function of position. Concretely, we propose a position-adaptive variant of STREAM, which we refer to as PA-STREAM, that assigns to each query block q a budget $k(q)$ that increases with the number of admissible key blocks. The guiding principle is that the budget should be just large enough to control the probability that the behavior-critical branch is pruned at any refinement level, while remaining sufficiently small that the total number of retained block interactions is near-linear up to polylogarithmic factors. This yields a sparse mask that is still structured and efficiently computable, but does not implicitly privilege early positions.

The contributions of this work are as follows.

- *Position bias under causal masking.* We formalize the causal-triangle bias as an information-theoretic obstruction: when a query has $|\mathcal{V}(q)|$ admissible key blocks and the estimator has no reliable distinguisher for the critical one, retaining at most k blocks yields worst-case success probability at most $k/|\mathcal{V}(q)|$. In particular, any constant-k policy has success probability tending to 0 for late queries as T grows. This statement is independent of STREAM, and isolates the source of failure as the growth of the admissible set itself.

- *PA-STREAM: a logarithmic position-adaptive schedule.* We introduce a sparsity schedule of the form

$$k(q) = k_0 + \left\lceil \alpha\big(\log(1 + |\mathcal{V}(q)|) + \log(1/\delta)\big) \right\rceil,$$

  together with a corresponding hierarchical pruning procedure that uses this budget at each query block. The schedule increases slowly (logarithmically) with position, so the total retained interactions scale as $O(N_q \log N_q)$ rather than $O(N_q^2)$, while enabling uniform control of failure probability across the sequence.

- *Matching necessity under noisy branch estimates.* We establish a lower bound for hierarchical selection with sub-Gaussian estimation noise and constant representatives per branch: if at some refinement level the estimator keeps $k(q) = o(\log |\mathcal{V}(q)|)$ branches, then there exist instances with a constant separation margin in which the probability

of pruning the critical branch is bounded below by a constant. This shows that the logarithmic growth of $k(q)$ is not merely sufficient but asymptotically necessary (up to constants) for uniform success in the long-context regime under the stated noise model.

- *Empirical validation plan.* We outline an evaluation protocol tailored to position-uniform fidelity. Rather than reporting a single aggregate score, we propose measuring retrieval or behavioral preservation as a function of needle depth (or query block index), and comparing fixed-k STREAM, PA-STREAM, and ablations that remove either the $\log|\mathcal{V}(q)|$ term or the $\log(1/\delta)$ term. We further propose reporting wall-clock speedups and memory usage under a blockwise implementation, to ensure that the improved uniformity does not come at the cost of negating sparsity benefits.

The net effect is a sparse attention masking strategy that is aligned with the combinatorics of causal attention: as the admissible past grows, we increase the budget just enough to keep the probability of losing behavior-critical information controlled. The remainder of the paper develops the formal model, states the hierarchical pruning abstraction we analyze, and proves the claimed guarantees and lower bounds before turning to experimental validation.

2 Preliminaries and Notation

We work with a decoder-only transformer on a length-T token sequence. Fix a layer ℓ and head h and suppress (ℓ, h) when unambiguous. Let Q, K, V denote the usual attention matrices with per-token rows $Q_i, K_j \in \mathbb{R}^d$ (and $V_j \in \mathbb{R}^{d_v}$). The masked attention output at token i is

$$\mathrm{Attn}(i) \;=\; \sum_{j=1}^{T} \pi_{ij} V_j, \qquad \pi_{ij} \;\propto\; \exp\Big(\tfrac{1}{\sqrt{d}}\langle Q_i, K_j\rangle\Big)\, C_{ij},$$

where $C \in \{0,1\}^{T\times T}$ is a token-level validity mask. In the causal setting $C_{ij} = \mathbf{1}\{j \le i\}$, so each query token may attend only to its prefix. Our focus is not on approximating π_{ij} for all pairs, but on selecting a structured subset of valid key positions that is sufficient for a specified fidelity criterion.

Block partitioning. We partition tokens into query blocks of size b_q and key blocks of size b_k, writing $N_q = T/b_q$ and $N_k = T/b_k$ (divisibility assumed for simplicity). Query block $q \in \{1, \dots, N_q\}$ contains token indices $(q-1)b_q+1, \dots, qb_q$, and similarly key block $r \in \{1, \dots, N_k\}$ contains $(r-1)b_k + 1, \dots, rb_k$. We use the block matrices

$$Q_q \in \mathbb{R}^{b_q\times d}, \qquad K_r \in \mathbb{R}^{b_k\times d}, \qquad V_r \in \mathbb{R}^{b_k\times d_v}.$$

A blockwise sparse mask is an indicator $M(q, r) \in \{0, 1\}$ that decides whether query block q is permitted to interact with key block r. Such a mask induces a token-level mask by allowing precisely those token pairs whose blocks are retained, subject to the original validity constraints.

To respect the original token-level mask C, we define a block validity predicate $C_{\text{blk}} \in \{0, 1\}^{N_q \times N_k}$ by

$$C_{\text{blk}}(q, r) = 1 \iff \exists\, i \in q,\ j \in r \text{ such that } C_{ij} = 1,$$

where $i \in q$ denotes that token i lies in query block q, and similarly $j \in r$. For causal C and aligned blocks, $C_{\text{blk}}(q, r) = 1$ for $r \leq q$ and 0 otherwise. We write the admissible key-block set for query block q as

$$\mathcal{V}(q) \ := \ \{r \in \{1, \ldots, N_k\} : C_{\text{blk}}(q, r) = 1\},$$

so $\mathcal{V}(q) = \{1, \ldots, q\}$ in the causal case.

Block scores and maxima. Hierarchical sparse attention procedures do not typically compute all blockwise interactions $Q_q K_r^\top$. Instead, they use a cheap score $\alpha(Q_q, K_r)$ that is intended to upper bound, approximate, or correlate with the best token-token affinity between the blocks. A canonical example (used in STREAM-style methods) is the block maximum

$$\alpha(Q_q, K_r) \ := \ \max_{1 \leq i \leq b_q,\ 1 \leq j \leq b_k} \langle Q_{q,i}, K_{r,j} \rangle,$$

although our abstraction permits other monotone surrogates (e.g. a $\log \sum \exp$ proxy). The objective of masking is then phrased in terms of retaining those key blocks whose contributions to downstream behavior are large or otherwise behavior-critical, rather than reproducing the full dense distribution π_{ij}.

Hierarchical pruning abstraction (STREAM/HiP). Fix a query block q. We view the set $\mathcal{V}(q)$ of admissible key blocks as the leaves of a refinement tree. At refinement level $i \in \{0, 1, \ldots, nit\}$, the admissible leaves are grouped into $B_i(q)$ disjoint *branches*, each branch corresponding to a contiguous range of key-block indices (for example, a binary partition of $\{1, \ldots, |\mathcal{V}(q)|\}$). A branch b at level i contains a subset $\mathcal{R}_{i,b}(q) \subseteq \mathcal{V}(q)$ of key blocks. The *true* branch maximum is

$$M_{i,b}(q) \ := \ \max_{r \in \mathcal{R}_{i,b}(q)} \alpha(Q_q, K_r).$$

Hierarchical pruning proceeds by estimating $M_{i,b}(q)$ for each branch using a small number of representatives, ranking branches by their estimated maxima, retaining only a limited number of branches, and then refining retained branches to the next level.

Concretely, for each branch b we sample at most $r = O(1)$ representatives $r^{(1)}, \ldots, r^{(r)} \in \mathcal{R}_{i,b}(q)$ (uniformly, or by a fixed deterministic rule), compute their block scores, and form an estimate $\widehat{M}_{i,b}(q)$ intended to approximate $M_{i,b}(q)$, e.g.

$$\widehat{M}_{i,b}(q) \;:=\; \max_{t \in \{1,\ldots,r\}} \alpha(Q_q, K_{r^{(t)}}).$$

We treat $\widehat{M}_{i,b}(q)$ as noisy evidence about $M_{i,b}(q)$. The estimator then retains the $k(q)$ branches with the largest $\widehat{M}_{i,b}(q)$ among those consistent with C_{blk}, discards the others, and refines only the retained branches. After *nit* refinement levels, each remaining branch corresponds to a small number of candidate key blocks; the final selection $S(q) \subseteq \mathcal{V}(q)$ is obtained by keeping at most $k(q)$ key blocks (for example, by selecting the best leaf from each surviving branch or by a final top-$k(q)$ over remaining leaves). The key point for our purposes is that *once a branch is pruned at some intermediate level, all key blocks it contains are irrevocably removed.*

Behavior-critical blocks and success. We formalize fidelity through a set of *behavior-critical* key blocks $\mathcal{R}^\star(q) \subseteq \mathcal{V}(q)$ for each query block q. The definition of $\mathcal{R}^\star(q)$ depends on the task: in a retrieval benchmark, $\mathcal{R}^\star(q)$ may contain the block holding the needle token(s); in a mechanistic tracing setting, $\mathcal{R}^\star(q)$ may contain the blocks that lie on a specified causal influence path. We say the masking procedure *succeeds at* q if it retains at least one critical block:

$$\mathrm{Succ}(q) \;:=\; \big\{ S(q) \cap \mathcal{R}^\star(q) \neq \emptyset \big\},$$

and *fails at* q otherwise. Since hierarchical pruning can eliminate critical blocks early, we also consider the stronger event that the critical block(s) remain in the retained set of branches at *every* refinement level; under the usual refinement-tree semantics, this stronger event implies $\mathrm{Succ}(q)$.

These preliminaries isolate the objects we will optimize and analyze: admissible sets $\mathcal{V}(q)$ induced by causal masking, a hierarchical branch-and-refine estimator driven by noisy maxima $\widehat{M}$, and a success criterion defined by retention of behavior-critical blocks. In the next section we express these components as a constrained sparsification problem and as a noisy branch-selection problem, and we make precise how causal growth of $|\mathcal{V}(q)|$ induces position-dependent difficulty.

3 Clean Problem Formulation

We now recast hierarchical sparse mask estimation as (i) a constrained sparsification problem with a *uniform* fidelity requirement across positions, and (ii) a sequence of noisy branch-selection subproblems induced by hierarchical refinement. This formulation isolates the mechanism by which causal masking creates position-dependent difficulty.

3.1 Uniform-fidelity sparse tracing as constrained sparsification

Fix a layer/head (ℓ, h) and suppress (ℓ, h) when convenient. For each query block $q \in \{1, \dots, N_q\}$ we must output a retained set $S(q) \subseteq \mathcal{V}(q)$ of key blocks, encoded equivalently by a block mask $M \in \{0,1\}^{N_q \times N_k}$ with

$$M(q, r) = \mathbf{1}\{r \in S(q)\}.$$

Mask feasibility is expressed by the validity constraint $S(q) \subseteq \mathcal{V}(q)$, and the local sparsity constraint $|S(q)| \leq k(q)$. If we allow a (possibly optional) global budget B, we may further require $\sum_{q=1}^{N_q} |S(q)| \leq B$; our subsequent analysis does not rely on this global form, but it is useful when comparing methods at matched compute.

To express fidelity, we use the critical-set abstraction from the preliminaries: for each q there is a set $\mathcal{R}^\star(q) \subseteq \mathcal{V}(q)$ of behavior-critical key blocks. The success event is $\mathrm{Succ}(q) = \{S(q) \cap \mathcal{R}^\star(q) \neq \emptyset\}$. The key requirement in long-context inference is not merely high *average* success, but *uniform* success over positions (especially those near the output region). We therefore study constraints of the form

$$\forall q \in \mathcal{Q}_{\mathrm{eval}} : \qquad \mathbb{P}\big(\mathrm{Succ}(q)\big) \;\geq\; 1 - \delta, \tag{1}$$

where $\mathcal{Q}_{\mathrm{eval}} \subseteq \{1, \dots, N_q\}$ is the set of query blocks at which we demand fidelity, and the probability is over any algorithmic randomness (e.g. representative sampling) and/or a data distribution (e.g. random placement of needles), as appropriate for the evaluation protocol.

With (1) in hand, sparse tracing becomes a constrained minimization:

$$\min_{\{S(q)\}} \sum_{q=1}^{N_q} |S(q)| \quad \text{subject to} \quad S(q) \subseteq \mathcal{V}(q),\;\; |S(q)| \leq k(q),\;\; (1). \tag{2}$$

We emphasize two structural features. First, (1) couples sparsification to the causal growth of $|\mathcal{V}(q)|$: late query blocks are constrained against a larger admissible set. Second, the constraint is existential in $\mathcal{R}^\star(q)$: it suffices to retain *one* critical block, which is precisely the regime in which hierarchical pruning is attractive but also fragile (since an early pruning decision can remove all critical candidates at once).

3.2 Hierarchical masking as branch selection under noisy maxima

We now express the hierarchical procedure for a fixed query block q as a noisy selection problem. At a refinement level i, the admissible set $\mathcal{V}(q)$

is partitioned into $B_i(q)$ branches $\{\mathcal{R}_{i,b}(q)\}_{b=1}^{B_i(q)}$. Each branch has a true maximum score

$$M_{i,b}(q) = \max_{r \in \mathcal{R}_{i,b}(q)} \alpha(Q_q, K_r),$$

and the algorithm observes only an estimate $\widehat{M}_{i,b}(q)$ computed from $r = O(1)$ representative samples. Abstractly, we treat $\widehat{M}_{i,b}(q)$ as a noisy statistic of $M_{i,b}(q)$; the salient point is that branch ranking is performed using $\widehat{M}_{i,b}(q)$, not $M_{i,b}(q)$.

Suppose there exists a behavior-critical block $r^\star \in \mathcal{R}^\star(q)$, and let $b^\star = b^\star(i,q)$ denote the unique branch at level i containing $r^\star$. Because pruning is irrevocable, the algorithm succeeds only if $b^\star$ is retained at *every* level. Thus, even when $\mathcal{R}^\star(q)$ contains a single block, the multi-level procedure induces a conjunction of branch-retention events:

$$\mathrm{Succ}(q) \;\supseteq\; \bigcap_{i=0}^{nit} \Big\{ \text{branch } b^\star(i,q) \text{ is among the retained branches at level } i \Big\}.$$

At each level, the decision rule is a top-$k(q)$ selection among estimated scores $\{\widehat{M}_{i,b}(q)\}_b$ restricted to valid branches. Hence the per-level failure mechanism is clear: $b^\star$ is pruned if sufficiently many noncritical branches receive upward fluctuations large enough to outrank it, or if $b^\star$ itself is underestimated. The role of a separation margin Δ is likewise transparent: if $M_{i,b^\star}(q)$ exceeds competitors by Δ, then only estimation error can cause misranking.

This reduction motivates analyzing the estimator in terms of (a) concentration of $\widehat{M}_{i,b}(q) - M_{i,b}(q)$, and (b) extreme-value effects across the $B_i(q)$ noncritical branches. The latter is the central obstruction: even modest upward noise becomes significant when amplified by a large number of competitors, which is precisely the regime encountered at late positions where $|\mathcal{V}(q)|$ (and therefore $B_i(q)$) is large.

3.3 Effective sparsity and causal position bias

We now formalize the notion that a fixed per-query budget implicitly becomes more stringent at later positions. Define the *effective retention ratio* (or effective sparsity level) at query block q by

$$\rho(q) \;:=\; \frac{k(q)}{|\mathcal{V}(q)|}\,. \tag{3}$$

Under causal masking with aligned blocks, $|\mathcal{V}(q)| = q$. Consequently, if $k(q) \equiv k$ is constant then $\rho(q) = k/q \to 0$ as $q \to \infty$. This decay is not a cosmetic artifact of normalization: it reflects a genuine increase in combinatorial difficulty. In the absence of additional information distinguishing the critical block(s), selecting k admissible key blocks at random yields success

probability exactly $\rho(q)$ when $|\mathcal{R}^\star(q)| = 1$ is uniformly distributed over $\mathcal{V}(q)$. Any method that cannot reliably separate critical from noncritical blocks by score must contend with this baseline.

Causal masking therefore induces an intrinsic position bias: a sparsifier with constant k allocates a vanishing fraction of the admissible set to late query blocks. Hierarchical pruning does not remove this bias; rather, it changes its manifestation. Early in the hierarchy, $B_i(q)$ grows with q, so the number of noisy competitors grows as well, increasing the probability that some noncritical branch attains an anomalously large $\widehat{M}$. Thus, even if the critical branch enjoys a fixed true-score advantage Δ, a constant retained-branch budget can be overwhelmed by the multiplicity of branches at large q.

In summary, (2) with the uniform constraint (1) forces us to confront the causal growth of $|\mathcal{V}(q)|$. The branch-selection view explains why noise and extreme values jointly degrade late-position performance when $k(q)$ is not allowed to increase. In the next section we make this obstruction formal by proving lower bounds showing that fixed-k (or, more generally, sublogarithmic $k(q)$) cannot maintain uniform success as T grows.

4 Why Fixed-k Fails (Formal)

We now make precise the sense in which a position-independent sparsity budget is incompatible with uniform-fidelity requirements under causal masking. The obstruction has two layers: an information-theoretic limitation that already holds in the absence of score structure, and a distinct extreme-value limitation that persists even when the critical branch enjoys a constant true-score advantage but branch scores are observed through noisy sampling.

4.1 Monotone growth of admissible keys under causal masking

Under (H1) we have, in block indices,

$$\mathcal{V}(q) = \{1, \dots, \min(q, N_k)\}, \qquad \text{hence} \qquad |\mathcal{V}(q)| = \min(q, N_k),$$

which is nondecreasing in q. In the common aligned regime $N_q = N_k$ this simplifies to $|\mathcal{V}(q)| = q$. Consequently, any schedule $k(q) \equiv k$ induces an effective retention ratio $\rho(q) = k/|\mathcal{V}(q)|$ that decays as $1/q$ for late positions. This monotone growth is the sole causal ingredient needed for the lower bounds below; it is independent of how hierarchical refinement partitions $\mathcal{V}(q)$ at intermediate levels.

4.2 Worst-case lower bound: constant k cannot yield uniform success

We first isolate the most basic impossibility: if the algorithm has no reliable side-information distinguishing which admissible key block is behavior-critical for the given query block, then selecting only k admissible blocks necessarily fails with high probability at large q.

Formally, fix q and suppose $|\mathcal{R}^\star(q)| = 1$, writing $\mathcal{R}^\star(q) = \{r^\star\}$, with $r^\star \in \mathcal{V}(q)$. Consider any (possibly randomized) selection rule outputting $S(q) \subseteq \mathcal{V}(q)$ with $|S(q)| \le k$. If $r^\star$ is worst-case (adversarial) among the admissible candidates, then we can force failure whenever $r^\star \notin S(q)$, and thus the worst-case success probability satisfies

$$\inf_{r^\star \in \mathcal{V}(q)} \mathbb{P}\big(r^\star \in S(q)\big) \;\le\; \frac{k}{|\mathcal{V}(q)|},$$

since $\mathbb{E}[|S(q)|] \le k$ and the average inclusion probability over $r^\star \in \mathcal{V}(q)$ equals $\mathbb{E}[|S(q)|]/|\mathcal{V}(q)|$. In particular, under $|\mathcal{V}(q)| = q$ we obtain the lower bound $1-k/q$ on failure probability, which is precisely the content of Thm. 1 in our notation.

A direct consequence for uniform fidelity is immediate. If $\mathcal{Q}_{\text{eval}}$ contains some q with $|\mathcal{V}(q)|$ large, then the constraint $\mathbb{P}(\text{Succ}(q)) \ge 1-\delta$ forces

$$k \;\ge\; (1-\delta)\,|\mathcal{V}(q)|. \tag{4}$$

Thus, a fixed budget k cannot satisfy a uniform requirement $\sup_{q\in\mathcal{Q}_{\text{eval}}} \mathbb{P}(\text{Fail}(q)) \le \delta$ as T grows unless we let k scale at least linearly with the latest evaluated q. This is a purely combinatorial statement: it does not depend on hierarchical structure, score distributions, or the particular implementation of $\alpha(\cdot,\cdot)$.

4.3 Distributional lower bound: random critical location still forces growth

The same phenomenon persists under a simple distributional model in which the critical location is random. Assume $r^\star$ is uniformly distributed on $\mathcal{V}(q)$, independent of any algorithmic randomness. Then for any (measurable) selection rule with $|S(q)| \le k$,

$$\mathbb{P}\big(\text{Succ}(q)\big) = \mathbb{P}\big(r^\star \in S(q)\big) = \mathbb{E}\Big[\frac{|S(q)|}{|\mathcal{V}(q)|}\Big] \le \frac{k}{|\mathcal{V}(q)|}.$$

Under $|\mathcal{V}(q)| = q$, the best achievable success probability under this model is at most k/q, which vanishes for constant k as $q \to \infty$. In particular, for any fixed $\delta \in (0,1)$, achieving $\mathbb{P}(\text{Succ}(q)) \ge 1-\delta$ again implies (4). The distributional statement is useful as a baseline when arguing that empirical late-position degradation is not merely an artifact of adversarial constructions: even benign random placement of a single required dependency forces

$k(q)$ to grow with $|\mathcal{V}(q)|$ unless scores provide additional exploitable structure.

4.4 Noisy hierarchical refinement: extreme values force at least logarithmic growth

One might hope that hierarchical pruning circumvents the preceding bounds by using score structure to concentrate probability mass on a small subset of admissible keys. However, under (H2) the algorithm does not observe true branch maxima M, but noisy estimates $\widehat{M}$ computed from only $r = O(1)$ representatives per branch. In this regime, the dominant failure mode is not the absence of signal, but the multiplicity of competitors: as the number of branches $B_i(q)$ grows, the maximum upward fluctuation among noncritical branches becomes significant even when each individual deviation is sub-Gaussian.

Thm. 2 captures this phenomenon in a simplified, level-wise abstraction: with B branches and $r = O(1)$ samples per branch, there exist instances (with a constant separation margin $\Delta = \Theta(1)$ between the critical branch maximum and competitors) such that retaining $k = o(\log B)$ branches by $\widehat{M}$ prunes the critical branch with constant probability. Translating to our masked attention setting, observe that for late q we necessarily have large candidate sets, hence some refinement level i with $B_i(q)$ polynomial in $|\mathcal{V}(q)|$ (e.g. $B_i(q) \asymp |\mathcal{V}(q)|/2^i$ for dyadic partitioning). Therefore, unless $k(q)$ grows at least on the order of $\log B_i(q)$, we incur a constant lower bound on per-level failure, and hence on overall failure (since the multi-level procedure requires retaining the critical branch at every level).

In particular, since $B_i(q)$ is typically $\Omega(|\mathcal{V}(q)|^\gamma)$ for some $\gamma > 0$ at an early level of refinement, we obtain the qualitative necessity

$$k(q) \;=\; \Omega\big(\log |\mathcal{V}(q)|\big)$$

to prevent late-position failure probabilities from being bounded away from zero under fixed r. This lower bound is compatible with the information-theoretic limitation above: score structure and separation can reduce the required growth from linear to logarithmic, but cannot remove growth altogether when branch maxima are estimated noisily and the number of competing branches increases with q.

Collecting these statements, fixed-k sparsification is untenable under uniform-fidelity constraints in long-context causal settings, and even sublogarithmic schedules cannot control the extreme-value effect induced by noisy branch ranking. This motivates the position-adaptive schedules in the next section, where we choose $k(q)$ to track the growth of $|\mathcal{V}(q)|$ sufficiently to offset both combinatorial dilution and noisy-competition amplification.

5 PA-STREAM: Position-Adaptive Sparsity Schedules and Practical Variants

We now specify the mask-estimation procedure we analyze. The guiding design choice is that the per-query budget must depend on the size of the admissible set $\mathcal{V}(q)$ (and, optionally, on a cheap proxy for attention "difficulty"), while remaining implementable in a single streaming pass with $O(T)$ working memory.

5.1 Base algorithmic skeleton (one layer/head)

Fix a layer–head pair (ℓ, h) with $\ell > \ell_d$. We partition the sequence into query blocks and key blocks, and compute the block validity mask C_{blk}. For each query block q, we run a hierarchical refinement procedure in the style of STREAM/HiP over the admissible key blocks $\mathcal{V}(q)$, where at each refinement level we (i) group candidates into branches, (ii) estimate each branch score by sampling at most $r = O(1)$ representatives and forming an estimated branch maximum $\widehat{M}$, and (iii) retain only the top-$k_{\ell,h}(q)$ branches by $\widehat{M}$, recursing until we reach the leaf (block) level. The output is the selected set $S_{\ell,h}(q) \subseteq \mathcal{V}(q)$, which we materialize as a sparse mask $M_{\ell,h}$ with $M_{\ell,h}(q, r) = 1$ iff $r \in S_{\ell,h}(q)$.

The only aspect not fixed by the baseline STREAM template is the schedule $k_{\ell,h}(q)$. PA-STREAM is the rule that chooses $k_{\ell,h}(q)$ as a nondecreasing function of the admissible candidate count $|\mathcal{V}(q)|$ (or an equivalent proxy), together with minor bookkeeping to ensure validity and budget invariants.

5.2 Logarithmic schedule from candidate counts

Our default schedule is

$$k_{\ell,h}(q) \;=\; \min\Big\{k_{\max},\; k_0 + \Big\lceil \alpha_{\ell,h} \log\big(1 + |\mathcal{V}(q)|\big) + \beta_{\ell,h} \log\big(1/\delta\big) \Big\rceil \Big\}, \qquad (5)$$

with $k_0 \geq 1$, $\alpha_{\ell,h}, \beta_{\ell,h} \geq 0$, and an optional cap $k_{\max}$ determined by hardware limits. Under the causal block mask (H1) we have $|\mathcal{V}(q)| = \min(q, N_k)$, hence $k_{\ell,h}(q)$ grows like $\log q$ until saturation. This growth is the minimal structural response to the increasing multiplicity of competitors under noisy branch ranking (cf. the lower bound discussed earlier), while remaining far from the linear growth that would be required absent score structure.

In practice we often enforce monotonicity explicitly by post-processing

$$k_{\ell,h}(q) \leftarrow \max\{k_{\ell,h}(q), k_{\ell,h}(q-1)\},$$

which removes small nonmonotonic fluctuations due to discretization or mixed regimes ($N_q \neq N_k$) and simplifies implementation in fused kernels.

5.3 Entropy-based and hybrid schedules

Candidate-count growth is necessary under (H1), but it is not always sufficient for best empirical accuracy at fixed compute: some query blocks exhibit unusually flat attention (many comparable keys), while others have a sharply peaked pattern. To adapt to this heterogeneity without full score materialization, we may compute a proxy $\widehat{H}_{\ell,h}(q)$ from the same sampled representatives used in the hierarchical search. One concrete choice is to form a small multiset of sampled block scores $\{s_m\}_{m=1}^{m}$ (e.g. maxima or averages within sampled token pairs) and define a normalized entropy

$$\widehat{H}_{\ell,h}(q) \;=\; -\sum_{m=1}^{m} \widehat{p}_m \log \widehat{p}_m, \qquad \widehat{p}_m = \frac{\exp(s_m/\tau)}{\sum_{j=1}^{m} \exp(s_j/\tau)},$$

with temperature $\tau > 0$ fixed. We then set

$$k_{\ell,h}(q) \;=\; \min\Big\{k_{\max},\; k_0 + \Big\lceil \alpha_{\ell,h} \log\big(1 + |\mathcal{V}(q)|\big) + \gamma_{\ell,h} \widehat{H}_{\ell,h}(q) \Big\rceil \Big\}, \tag{6}$$

where $\gamma_{\ell,h}$ controls how aggressively we respond to flatness. The point of (6) is not to change the asymptotic growth in $|\mathcal{V}(q)|$, but to reallocate budget among positions at fixed average cost.

5.4 Per-layer and per-head calibration

Different layers and heads exhibit different separations Δ and effective noise scales σ in their branch score estimates (and, empirically, different sensitivity of downstream behavior to pruning). Accordingly, we allow $\alpha_{\ell,h}, \beta_{\ell,h}, \gamma_{\ell,h}$ to depend on (ℓ, h). A lightweight calibration procedure is to run PA-STREAM on a short calibration set, sweep $\alpha_{\ell,h}$ over a small grid, and choose the smallest value for which the observed failure metric (e.g. needle retrieval or next-token deviation) stays below a target. Since $\ell \le \ell_d$ is kept dense by design, this calibration is restricted to $\ell > \ell_d$ and typically exhibits a monotone "accuracy versus $\alpha_{\ell,h}$" curve.

5.5 Global budgeted variant

If a global constraint $\sum_q k_{\ell,h}(q) \le B_{\ell,h}$ is required, we can impose it by a Lagrangian scaling of the baseline schedule. Concretely, define unnormalized demands

$$u(q) = k_0 + \alpha_{\ell,h} \log\big(1 + |\mathcal{V}(q)|\big) + \beta_{\ell,h} \log(1/\delta)$$

(or the entropy-augmented analogue), and choose a multiplier $\lambda \ge 0$ such that

$$k_{\ell,h}(q) = \min\{k_{\max},\; \max\{k_0,\; \lfloor \lambda\, u(q) \rfloor\}\} \quad \text{satisfies} \quad \sum_q k_{\ell,h}(q) \le B_{\ell,h}.$$

Since $u(q)$ is nondecreasing under (H1), λ-scaling preserves monotonicity and can be found by a one-dimensional search. This provides a simple knapsack-like allocation while keeping the analysis anchored to a schedule that grows with $|\mathcal{V}(q)|$.

5.6 Implementation notes: dense prefix and block-size interactions

We keep the first ℓ_d layers dense. Operationally this means that PA-STREAM is invoked only for $\ell > \ell_d$, and the dense prefix supplies stable intermediate representations in which score separations are typically larger; empirically, this reduces the required $\alpha_{\ell,h}$ in later layers. Mask storage is streamed: we write out $S_{\ell,h}(q)$ as indices (or compressed ranges when contiguity emerges) and avoid storing full $N_q \times N_k$ masks.

Finally, block sizes b_q, b_k influence both candidate counts and the fidelity of branch scoring. Increasing b_k reduces N_k and hence $|\mathcal{V}(q)|$, permitting smaller $k_{\ell,h}(q)$ at the block level; however, it also makes each block less homogeneous, so a representative-based $\alpha(Q_q, K_r)$ can become noisier, effectively increasing σ. Conversely, smaller b_k increases N_k and the number of competing branches, strengthening the need for logarithmic growth in $k_{\ell,h}(q)$ but improving localization. Our main theorems in the next section make these tradeoffs explicit: they state sufficient conditions, in terms of σ, Δ, r and the schedule parameters, under which the behavior-critical branch is retained uniformly over q, together with the corresponding runtime and memory bounds.

6 Main Theorems (Upper Bounds)

We now state sufficient conditions under which PA-STREAM achieves uniform retention of behavior-critical information across query positions, together with the resulting time and space bounds. Throughout we fix a pruned layer–head pair (ℓ, h) with $\ell > \ell_d$, and we consider one query block q with admissible key blocks $\mathcal{V}(q)$ under the causal mask (H1). We write $n_q := |\mathcal{V}(q)|$, and we let $\mathrm{nit} := \lceil \log_2 N_k \rceil$ denote the number of refinement levels until leaf blocks are reached.

6.1 Uniform retention for a single query block

We formalize "behavior-critical retention" at the level of branches in the hierarchical refinement. At refinement level $i \in \{1, \ldots, \mathrm{nit}\}$, the admissible candidates $\mathcal{V}(q)$ are partitioned into $B_i(q)$ valid branches. For a branch b at level i, let $M_{i,b}(q)$ denote its true branch maximum score (the maximum of $\alpha(Q_q, K_r)$ over leaf blocks r contained in b), and let $\widehat{M}_{i,b}(q)$ denote the

estimated maximum computed from at most $r = O(1)$ sampled representatives as in (H2). We say a branch $b^\star$ is *critical* at level i if it contains at least one behavior-critical key block for query q. The failure event at level i is that $b^\star$ is not among the top-$k_{\ell,h}(q)$ branches by $\widehat{M}_{i,b}(q)$, hence is pruned and cannot re-enter at deeper levels.

Theorem 6.1 (Uniform per-query retention under logarithmic schedules)**.** *Assume (H2): for every level i and branch b, the estimation error $\widehat{M}_{i,b}(q) - M_{i,b}(q)$ is sub-Gaussian with parameter σ, uniformly over q. Assume (H3): with probability at least $1 - \delta_0$, for every level i the critical branch maximum exceeds the $(k_{\ell,h}(q)+1)$-st largest noncritical branch maximum by margin at least $\Delta > 0$. Fix a target $\delta \in (0,1)$. If we choose the budget*

$$k_{\ell,h}(q) \;\geq\; k_0 + \left\lceil \alpha\big(\log(1+n_q) + \log(1/\delta)\big) \right\rceil \tag{7}$$

with $\alpha \geq c\,\sigma^2/\Delta^2$ for an absolute constant $c > 0$ and k_0 large enough to cover the base branching factor at the top level, then PA-STREAM prunes all behavior-critical key blocks for query q with probability at most $\delta + \delta_0$.

Proof sketch. We condition on the (H3) separation event, which contributes the additive δ_0. Fix a level i, and let $b^\star$ be the critical branch at this level, with true maximum $M^\star := M_{i,b^\star}(q)$. Let $\{M_j\}_{j=1}^{B_i(q)-1}$ be the true maxima of noncritical branches. By (H3) we have $M_j \leq M^\star - \Delta$ for all but at most $k_{\ell,h}(q)$ branches (equivalently, $M^\star$ exceeds the $(k_{\ell,h}(q)+1)$-st largest by Δ). Define the threshold $\tau := M^\star - \Delta/2$. Failure at level i occurs only if either (a) the critical branch is underestimated below τ, i.e. $\widehat{M}^\star \leq \tau$, or (b) at least $k_{\ell,h}(q)$ noncritical branches satisfy $\widehat{M}_j \geq \tau$. By sub-Gaussian concentration,

$$\mathbb{P}(\widehat{M}^\star \leq M^\star - \Delta/2) \;\leq\; \exp\!\left(-\frac{\Delta^2}{c_1\sigma^2}\right)$$

for an absolute c_1. For noncritical branches with $M_j \leq M^\star - \Delta$, we similarly have

$$\mathbb{P}(\widehat{M}_j \geq M^\star - \Delta/2) \;\leq\; \exp\!\left(-\frac{\Delta^2}{c_2\sigma^2}\right).$$

A union bound over the $B_i(q)$ branches upper bounds the probability that $k_{\ell,h}(q)$ or more noncritical branches cross τ by a binomial-tail bound controlled by $B_i(q)\exp(-\Delta^2/(c_2\sigma^2))$. Choosing $k_{\ell,h}(q)$ as in (7) ensures this probability is at most δ/nit for all levels, using $B_i(q) \leq n_q$ and absorbing constants into α. Finally, we apply a union bound over levels $i \in \{1, \ldots, \mathrm{nit}\}$ to obtain failure probability at most δ conditioned on (H3). □

6.2 Runtime and storage consequences

We next translate Theorem 6.1 into total retained interactions and mask-estimation runtime. The relevant point is that under (H1) we have $n_q = \min(q, N_k)$, so $k_{\ell,h}(q)$ grows at most logarithmically in q until saturation at $k_{\max}$.

Corollary 6.2 (Near-linear total budget and polylog overhead)**.** *Assume $k_{\ell,h}(q) \le k_0 + \alpha \log(1+q) + \beta \log(1/\delta)$ for all $q \le N_q$ (e.g. (7) under (H1)). Then*

$$\sum_{q=1}^{N_q} k_{\ell,h}(q) \;=\; O(N_q \log N_q) + O\big(N_q \log(1/\delta)\big).$$

Moreover, if STREAM-style refinement uses $O(\log N_k)$ levels and performs $O(k_{\ell,h}(q))$ top-k branch selections per level, then the PA-STREAM mask-estimation time per layer/head is

$$O\left(\sum_{q=1}^{N_q} k_{\ell,h}(q)\, \log N_k\right) \;=\; O\big(N_q \log N_k \log N_q\big) \quad \textit{(up to constants and } \log(1/\delta) \textit{ terms).}$$

Mask storage as block indices is $O(\sum_q k_{\ell,h}(q))$, while working memory remains $O(T)$ when masks are streamed.

6.3 Tradeoffs between budget, representatives, and failure probability

Theorem 6.1 isolates the role of σ, which is the effective noise scale in $\widehat{M}$. When $\widehat{M}$ is computed from r i.i.d. representatives, standard concentration suggests that the sub-Gaussian parameter improves as $\sigma_r \asymp \sigma/\sqrt{r}$ for averaged estimators (and more generally decreases with r for robust maxima/quantile estimators). Substituting σ_r into Theorem 6.1 yields the qualitative tradeoff

$$k_{\ell,h}(q) \;=\; \Theta\left(\frac{\sigma^2}{r\,\Delta^2}\Big(\log(1+n_q) + \log(1/\delta)\Big)\right), \tag{8}$$

up to additive k_0 and absolute constants. Thus we may reduce $k_{\ell,h}(q)$ by increasing r, or conversely we may hold r constant and accept the logarithmic growth in $k_{\ell,h}(q)$ required for uniform success. The dependence on $\log(1/\delta)$ is similarly unavoidable in this concentration-based analysis: demanding smaller failure probability forces either larger budgets $k_{\ell,h}(q)$, more representatives r, or larger separations Δ.

Finally, we note that when we require uniform retention simultaneously over many query blocks and over multiple (ℓ, h), we may set per-instance targets $\delta_{\ell,h,q}$ so that $\sum_{\ell,h,q} \delta_{\ell,h,q} \le \delta_{\text{tot}}$, and then apply a union bound. In

the simplest symmetric allocation one takes $\delta_{\ell,h,q} = \delta_{\mathrm{tot}}/((L-\ell_d)HN_q)$, which introduces an additional additive $\log((L-\ell_d)HN_q)$ term inside the $\log(1/\delta)$ factor of (7). This completes the upper-bound picture: logarithmic growth in the admissible-set size suffices for uniform retention under (H2)–(H3), with near-linear total budget and polylogarithmic runtime overhead.

7 Lower Bounds and Tightness

We complement the upper bounds by exhibiting explicit instance families in which hierarchical selection from noisy branch estimates cannot succeed uniformly unless the retained-branch budget grows at least logarithmically in the number of admissible candidates. The point is not merely information-theoretic indistinguishability (as in Theorem 1), but rather a *noisy-ranking* phenomenon: even when the critical branch is separated by a fixed margin at the level of *true* branch maxima, a procedure that (i) estimates each branch from only $r = O(1)$ representatives and (ii) retains only the top-k branches by the resulting noisy estimates must take $k = \Omega(\log B)$ at some refinement level with B branches, or else incur constant failure probability.

7.1 An explicit hard family at a single refinement level

Fix a refinement level and suppress i, q from the notation. We consider B candidate branches, exactly one of which is critical. The algorithm observes estimated maxima $\{\widehat{M}_b\}_{b=1}^{B}$ (computed from at most r representatives per branch) and retains the k branches with largest $\widehat{M}_b$. We now construct an instance where the true maxima satisfy a constant margin, yet the critical branch is pruned with constant probability unless $k \gtrsim \log B$.

Theorem 7.1 (Logarithmic budget is necessary under $r = O(1)$)**.** *Fix $\Delta \in (0, 1/2]$ and an integer $B \geq 16$. There exists a family of branch-score instances with the following properties:*

1. *There is a unique critical branch $b^\star$ with true maximum $M_{b^\star} = 1$.*
2. *There are $s := \lceil c \log B \rceil$ noncritical branches with true maximum $M_b = 1 - \Delta$, and the remaining $B - 1 - s$ noncritical branches have $M_b = 0$, for an absolute constant $c > 0$.*
3. *For each branch b, the estimation error $\widehat{M}_b - M_b$ is sub-Gaussian with a constant parameter $\sigma = O(1)$ (uniform in B), and $\widehat{M}_b$ can be generated by sampling at most $r = O(1)$ representatives from within the branch.*

For this family, any top-k retention rule that keeps only the k largest $\widehat{M}_b$ has failure probability bounded below by a constant whenever $k < s$; in particular, if $k = o(\log B)$ then for all sufficiently large B,

$$\mathbb{P}\big(b^\star \text{ is pruned}\big) \;\geq\; \frac{1}{4}.$$

Construction. We take the s "strong" noncritical branches to be noise-free at the estimate level: set $\widehat{M}_b \equiv M_b = 1 - \Delta$ for those s branches, and $\widehat{M}_b \equiv M_b = 0$ for the remaining weak noncritical branches. For the critical branch $b^\star$, we implement a representative-sampling estimator as follows. Inside $b^\star$ place m leaf blocks, with one distinguished leaf having score 1 and all other leaves having score 0; define $\widehat{M}_{b^\star}$ as the maximum score among r uniformly sampled representatives (with replacement) from these m leaves. Then

$$\widehat{M}_{b^\star} \in \{0, 1\}, \qquad \mathbb{P}(\widehat{M}_{b^\star} = 1) \;=\; 1 - \left(1 - \frac{1}{m}\right)^r.$$

Choosing m as a sufficiently large constant multiple of r makes $p := \mathbb{P}(\widehat{M}_{b^\star} = 1) \le 3/4$, hence $\mathbb{P}(\widehat{M}_{b^\star} = 0) \ge 1/4$. Moreover, the error $\widehat{M}_{b^\star} - M_{b^\star} \in \{-1, 0\}$ is bounded; consequently $\widehat{M}_{b^\star} - M_{b^\star}$ is sub-Gaussian with an absolute constant parameter (a bounded random variable is sub-Gaussian after centering). All other branches have deterministic errors 0, which are trivially sub-Gaussian.

Why this forces $k = \Omega(\log B)$. On the event $\{\widehat{M}_{b^\star} = 0\}$, all $s = \lceil c \log B \rceil$ strong noncritical branches have estimates $\widehat{M}_b = 1 - \Delta$ and thus strictly outrank $b^\star$. Therefore any rule that retains only the top $k < s$ branches must discard $b^\star$ on this event. Hence

$$\mathbb{P}(b^\star \text{ is pruned}) \;\ge\; \mathbb{P}(\widehat{M}_{b^\star} = 0) \;\ge\; \frac{1}{4}.$$

This proves Theorem 7.1. □

7.2 A complementary lower bound: constant k forces $r = \Omega(\log B)$

The preceding construction shows that even with a fixed margin Δ, *underestimation* of the critical branch can force a logarithmic k when only $r = O(1)$ representatives are available. A separate (and more classical) phenomenon shows that, for natural unbiased estimators, *overestimation* among many competitors forces r to grow at least like $\log B$ if one insists on constant k.

Proposition 7.2 (Representatives needed for $k = O(1)$)**.** *Fix $\Delta \in (0, 1)$ and consider B branches with $M_{b^\star} = 0$ for the critical branch and $M_b = -\Delta$ for all noncritical branches. Suppose $\widehat{M}_b$ is the empirical mean of r i.i.d. 1-sub-Gaussian samples with mean M_b, so that $\widehat{M}_b - M_b$ is $O(1/\sqrt{r})$-sub-Gaussian. If the selection rule retains only the single best branch ($k = 1$), then*

$$\mathbb{P}(b^\star \textit{ is selected}) \;\le\; \exp\!\big(-c_1 r \Delta^2\big)\, B$$

for an absolute constant $c_1 > 0$. Consequently, achieving $\mathbb{P}(b^\star \textit{ is selected}) \ge 3/4$ requires $r = \Omega(\log B / \Delta^2)$.

Proof sketch. For a fixed noncritical branch b, sub-Gaussian tails give $\mathbb{P}(\widehat{M}_b \geq 0) \leq \exp(-c_1 r\Delta^2)$. By a union bound over $B-1$ noncritical branches, the probability that *some* noncritical branch attains $\widehat{M}_b \geq 0$ is at most $(B-1)\exp(-c_1 r\Delta^2)$. On this event, the top-1 selection fails (ties can be handled by an arbitrarily small perturbation). Rearranging yields the stated requirement. □

7.3 Implications for causal attention and tightness of PA-STREAM schedules

We now translate the single-level lower bounds into a statement about causal attention with $|\mathcal{V}(q)| = n_q$. At some refinement level i, the number of valid branches $B_i(q)$ is at most n_q and, for standard balanced hierarchical partitions, is $\Theta(n_q)$ at the top levels. Applying Theorem 7.1 with $B = B_i(q)$ shows that any hierarchical estimator that (a) uses only $r = O(1)$ representatives per branch and (b) hopes to maintain a constant per-level success probability uniformly in q must allow

$$k_{\ell,h}(q) \;=\; \Omega(\log B_i(q)) \;=\; \Omega(\log n_q) \;=\; \Omega(\log |\mathcal{V}(q)|),$$

up to absolute constants (and ignoring additional $\log(1/\delta)$ factors demanded by high-probability rather than constant-probability guarantees). Proposition 7.2 gives the complementary tradeoff: if one insists on constant $k_{\ell,h}(q)$, then one must drive the effective noise scale down by increasing representatives r, and $r = \Omega(\log |\mathcal{V}(q)|)$ is necessary in general.

Comparing with Theorem 6.1, we conclude that PA-STREAM's logarithmic schedule $k_{\ell,h}(q) = \Theta(\log(1+n_q) + \log(1/\delta))$ is optimal in its dependence on $|\mathcal{V}(q)|$ within the model class captured by (H2) and hierarchical top-k refinement. Any improvement would require additional structure beyond (H2)–(H3) (e.g. stronger priors restricting which branches can be competitive, or estimators whose error tails shrink faster than sub-Gaussian under the available compute), or else a compensating increase in r that is itself at least logarithmic in $|\mathcal{V}(q)|$.

8 Experimental Design (Strengthening Evidence)

We now specify an experimental protocol whose purpose is to test, under controlled and realistic settings, whether the position-adaptive schedule in PA-STREAM achieves (i) uniform long-context fidelity under causal masking and (ii) near-linear resource growth, and to isolate which design choices are responsible for any observed gains. Throughout, we compare against fixed-budget hierarchical pruning ("fixed-k STREAM") and, when feasible, a dense-attention reference run (or the strongest dense-prefix surrogate available at the target length).

8.1 Long-context stress tests via RULER depth–vs–length sweeps

We adopt the RULER family of long-context tasks as a systematic stress test because it provides explicit control over both sequence length T and the depth at which behavior-critical information is placed. For each target length $T \in \{2^{14}, 2^{15}, \ldots, 2^{20}\}$ (up to the largest feasible T on the target hardware), we generate prompts with a single "needle" (e.g. a key-value fact, a short string to be retrieved, or a constrained instruction) inserted at a controlled depth. We parameterize depth both in tokens and in block indices: if the needle begins at token position t, we record its query-relative depth as $t/T \in (0, 1)$ and its block depth as $\lceil t/b_k \rceil$. We then evaluate a grid of depth settings, e.g. $t/T \in \{0.1, 0.3, 0.5, 0.7, 0.9, 0.97\}$, to probe the regime where $|\mathcal{V}(q)|$ is large and Theorem-level position bias becomes most severe for fixed budgets.

The principal comparison is between (a) fixed-k STREAM, in which $k_{\ell,h}(q) \equiv k$ for all q in pruned layers/heads, and (b) PA-STREAM with the schedule $k_{\ell,h}(q) = k_0 + \lceil \alpha(\log(1 + |\mathcal{V}(q)|) + \log(1/\delta)) \rceil$. To avoid confounding from total budget differences, we include a matched-budget condition in which fixed-k is tuned so that $\sum_q k = \sum_q k(q)$ (up to rounding), thereby testing whether adaptivity across q matters beyond total retained mass. We also include an "oracle budget" sweep in which we vary a global multiplier γ and set $k(q) = \lceil \gamma \log(1 + |\mathcal{V}(q)|) \rceil$ to test sensitivity to the constant factor predicted by σ^2/Δ^2-type scaling.

8.2 Realistic retrieval-augmented generation corpora

Synthetic needles are necessary but not sufficient: we also require evaluation on naturally occurring long contexts arising from retrieval-augmented generation (RAG). We therefore construct long-context inputs by concatenating retrieved passages from a large text corpus (e.g. encyclopedic and technical sources) together with a user query that requires multi-passage synthesis. We generate retrieval contexts at multiple lengths by varying the number of passages and truncating/packing to a target token budget T. In this setting, behavior-critical information is not a single needle but rather a sparse set of evidential spans whose locations are determined by the retriever; this probes whether PA-STREAM preserves multiple long-range dependencies simultaneously.

To connect the evaluation to the mask level, we log (for each query block q) the distribution of selected key-block indices $S_{\ell,h}(q)$ and compute summary statistics such as the median retained key distance (in blocks) and the fraction of retained blocks that fall within the earliest ρ-prefix of $\mathcal{V}(q)$ for $\rho \in \{0.1, 0.2\}$. These diagnostics help distinguish "attention-sink"-like behavior (mass concentrated on early blocks) from genuinely depth-aware

retention.

8.3 Metrics: task success, fidelity, and resources

We report four metric families.

(i) Retrieval/task success. For RULER-style tasks, we use exact-match (or constrained-format) success, aggregated as a function of both T and depth t/T. For RAG-style corpora, we use answer quality metrics appropriate to the task (exact match / F1 when labeled, and otherwise verifier-based correctness), but we additionally include a citation-precision proxy: whether the generated answer contains a substring uniquely identifying the supporting passage(s), which is sensitive to long-range access.

(ii) Logit-level fidelity. To measure whether sparse attention preserves the model's local behavior beyond task-specific needles, we compute the per-token KL divergence between the dense-reference next-token distribution $p_{\text{dense}}(\cdot \mid x_{1:t})$ and the sparse distribution $p_{\text{sparse}}(\cdot \mid x_{1:t})$,

$$\mathrm{KL}\big(p_{\text{dense}} \,\|\, p_{\text{sparse}}\big) \;=\; \sum_{v} p_{\text{dense}}(v) \log \frac{p_{\text{dense}}(v)}{p_{\text{sparse}}(v)}.$$

We average this KL over a fixed window near the output (e.g. the last W tokens) and also report top-1 agreement and the change in log-probability assigned to the dense model's greedy token. When a full dense run is infeasible at the target T, we use a strong proxy reference (e.g. dense attention for the final layers and/or a smaller T matched by truncation) and interpret results accordingly.

(iii) Mask sparsity. We measure the realized sparsity as the average number of retained key blocks per query block,

$$\overline{k} \;:=\; \frac{1}{N_q} \sum_{q=1}^{N_q} |S_{\ell,h}(q)|,$$

and also report the total retained interactions $\sum_q |S_{\ell,h}(q)|$ as a function of T. To ensure that sparsity is not achieved by violating validity, we verify that all selected pairs satisfy $C_{\text{blk}}(q, r) = 1$.

(iv) Wall-clock time and memory. We measure end-to-end inference latency (including mask estimation) and peak device memory, both as functions of T, and we report the slope with respect to T in the long-context regime. We use a fixed hardware/software stack and include separate timings for (a) mask construction and (b) sparse attention application, since PA-STREAM changes both.

8.4 Ablations: schedules, block sizes, and dense early layers

To identify which components are essential, we run the following ablations.

Schedule variants. We compare (1) constant $k(q) \equiv k$, (2) linear $k(q) = k_0 + \gamma q$ (a pessimistic control), (3) square-root $k(q) = k_0 + \gamma\sqrt{q}$, (4) logarithmic $k(q) = k_0 + \lceil \alpha \log(1+q) \rceil$, and (5) entropy/proxy-adaptive schedules $k(q) = k_0 + \lceil \alpha \widehat{H}(q) \rceil$ where $\widehat{H}(q)$ is computed from the sampled branch scores used by PA-STREAM. For each variant, we either match $\sum_q k(q)$ or report Pareto curves trading fidelity against sparsity.

Block sizes (b_q, b_k). We sweep b_q, b_k over powers of two (e.g. $64, 128, 256$) to test whether PA-STREAM's gains are robust to coarser partitions that reduce N_k but potentially increase within-block heterogeneity. We record how the optimal α and the observed success-vs-depth curves shift with b_k, since $|\mathcal{V}(q)|$ is block-granular.

Dense-prefix depth ℓ_d. We vary the number of initial dense layers $\ell_d \in \{0, 2, 4, \ldots\}$ to test the hypothesis that early layers benefit from dense mixing while later layers can be aggressively sparsified. We report fidelity metrics as a function of ℓ_d at fixed total sparsity, thereby separating "where to prune" from "how much to prune."

Finally, we standardize all stochastic components (sampling within branches, tie-breaking in top-k) by fixing random seeds and reporting confidence intervals across runs. This isolates the intrinsic variance induced by representative sampling from task-level variance in the prompts.

9 Discussion

We discuss implications of position-adaptive hierarchical sparsification for monitoring and control at million-token scale, articulate regimes in which logarithmic growth of the retained budget can be insufficient, relate the schedule to empirical phenomena such as retrieval heads and attention sinks, and delineate the limitations of an attention-only analysis under the separation/noise assumptions.

Implications for million-token monitoring. A primary motivation for a schedule of the form $k(q) = k_0 + \lceil \alpha(\log(1 + |\mathcal{V}(q)|) + \log(1/\delta)) \rceil$ is that it keeps the *per-position* risk of pruning behavior-critical context approximately uniform as the context grows, rather than allowing failure probability to drift upward with q as predicted by position-bias lower bounds for constant k. At $T \approx 10^6$, the number of blocks $N_q = T/b_q$ is typically on the order of 10^4 for common $b_q \in [64, 256]$, so $\log N_q$ remains modest; the resulting $\sum_q k(q) =$

$\Theta(N_q \log N_q)$ scaling is compatible with streaming mask construction and is, in our computational model, a plausible substrate for *monitoring* in which we seek to certify that certain evidential spans remain reachable by late-layer queries. In particular, if we treat "monitoring" as the task of ensuring that a sparse controller (the mask) does not erase designated information paths, then the schedule provides a concrete knob—δ—that can be set according to an application-level tolerance, with the understanding that the runtime and retained interactions increase only polylogarithmically in the long-context regime. This is the sense in which PA-STREAM is suited to million-token settings: it does not promise that every dependence is preserved, but rather that dependencies that manifest as consistently high-scoring branches under the model's own Q, K geometry are unlikely to be pruned uniformly over position.

When log**-growth may be insufficient.** The preceding statement is conditional, and it is important to isolate the failure modes in which $\Theta(\log |\mathcal{V}(q)|)$ retained branches does not suffice even in principle for uniform success. First, our sufficient condition depends on an effective separation margin Δ between a behavior-critical branch maximum and competing branches at each refinement level. In realistic long contexts, Δ may deteriorate with q (e.g. because many semantically similar passages appear as the context grows), and the required constant α scales as σ^2/Δ^2; thus a schedule with fixed α may under-allocate $k(q)$ precisely where the model becomes least separated. Second, the sub-Gaussian noise model captures bounded-variance estimation error from representative sampling, but it does not cover heavy-tailed or state-dependent errors that can arise when the representative set is systematically uninformative for certain branches. In such cases, increasing $k(q)$ alone may be inefficient; instead one may need to increase the number of representatives r, or to adopt estimators with variance reduction (e.g. stratified sampling within branches), effectively trading additional dot-products for tighter concentration of $\widehat{M}$. Third, behavior-critical information need not be localized to a single branch; it may be *combinatorial* (multiple spans jointly required), in which case preserving only the top branch at each level is not the correct objective. One expects then that the appropriate budget is governed by a multi-target analogue of the separation condition, and may scale with the number of simultaneously required branches, not merely $\log |\mathcal{V}(q)|$.

Relation to retrieval heads and attention sinks. Empirically, decoder-only transformers often exhibit heads that behave like retrieval mechanisms, with attention patterns that sharply localize on specific prior spans, as well as "attention sinks" in which early tokens attract disproportionate mass across many queries. PA-STREAM's schedule addresses a structural issue that

is orthogonal to these phenomena: under causal masking, the candidate set $|\mathcal{V}(q)|$ grows with position, so any fixed per-query budget induces an increasing combinatorial disadvantage for deep retrieval, irrespective of whether the model has specialized retrieval heads. Thus, to the extent that retrieval heads exist, a position-adaptive budget can be interpreted as supplying these heads with the degrees of freedom required to express long-range selection uniformly across q. At the same time, the existence of attention sinks complicates the interpretation of retained blocks: if a sink branch consistently yields high estimated maxima, a purely score-driven procedure will retain it, potentially crowding out mid-context evidence when the budget is small. In this sense, PA-STREAM should not be viewed as a complete remedy for sink-like behavior; it is a mechanism for avoiding a *budget-induced* depth collapse, not for correcting model-internal biases. A natural extension, compatible with the present framework, is to incorporate mild reweighting or constraints at the branch-scoring stage (e.g. penalizing extremely early blocks or enforcing diversity across distance bins) while retaining the same logarithmic scaling needed to control the probability of pruning genuinely critical late branches.

Limitations of attention-only tracing. Our guarantees are stated in terms of retaining behavior-critical *key blocks* under an attention-based notion of influence. This is necessarily incomplete: modern transformers route information not only through attention but also through residual pathways and MLP sublayers, and "behavior-critical" dependence can be mediated by features that do not correspond to a single high-attention edge. Accordingly, even a perfect attention mask is not a full causal certificate of functional equivalence; it is a controlled approximation that preserves certain attention-mediated interactions with high probability under assumptions. Moreover, the block score $\alpha(Q_q, K_r)$ is itself a proxy: maxima or sampled maxima are chosen for computational convenience, but they may correlate imperfectly with the downstream contribution of V_r after softmax normalization, head mixing, and residual addition. This mismatch is a fundamental limitation of mask construction from partial score observations, and it suggests that logit-level fidelity (or other functional metrics) should be treated as a first-class target, not merely an evaluation afterthought.

Dependence on separation and noise assumptions. Finally, we emphasize that the separation/noise hypotheses should be interpreted as modeling assumptions rather than universal truths. Separation may hold with high probability for many queries and yet fail on rare but important ones; likewise, sub-Gaussian estimation error may hold on average but break under distribution shift or adversarially constructed contexts. In practice, one may wish to *estimate* σ online from the dispersion of sampled scores, and to detect potential violations of separation by monitoring the empirical gaps be-

tween top-ranked branches; such diagnostics could trigger adaptive increases of $k(q)$ or r on a per-query basis. These considerations motivate, in the subsequent conclusion, the development of mask-selection procedures that couple adaptive schedules with explicit fidelity certificates and with tracing mechanisms that extend beyond attention alone.

10 Conclusion and Future Work

We have formalized a position-adaptive view of hierarchical attention sparsification under causal masking, in which the combinatorial growth of admissible keys with depth forces any uniformly reliable procedure to increase its retained budget with q. Within a blocked/streaming computational model, PA-STREAM couples STREAM-style hierarchical pruning with a sparsity schedule $k(q)$ that scales logarithmically in $|\mathcal{V}(q)|$ (and in $\log(1/\delta)$), yielding near-linear total retained interactions and polylogarithmic runtime overhead. Under sub-Gaussian branch-score estimation noise and a separation margin at each refinement level, this schedule suffices to bound the probability of pruning behavior-critical key blocks uniformly over position; conversely, with $r = O(1)$ representatives, retaining $k(q) = o(\log |\mathcal{V}(q)|)$ branches at some level is asymptotically incompatible with bounded failure probability on explicit hard instances. We therefore regard logarithmic growth not as an aesthetic choice but as a structural response to causal position bias in long contexts.

Adaptive schedules with variable per-iteration branching. A first extension is to make the schedule sensitive not only to $|\mathcal{V}(q)|$ but also to the *effective* branching structure encountered during refinement. In practice the number of valid branches $B_i(q)$ at level i depends on causal truncation, padding, and implementation details (e.g. unequal block sizes near boundaries), and may vary substantially with q. Our current statement uses the worst-case proxy $\log |\mathcal{V}(q)| \approx \log q$ to control a union bound across levels, but a sharper allocation is possible: one may define per-level budgets $k_i(q)$ and distribute a target failure probability δ as $\sum_i \delta_i \le \delta$, choosing

$$k_i(q) \;\approx\; k_{0,i} \;+\; \left\lceil \alpha\big(\log(1 + B_i(q)) + \log(1/\delta_i)\big) \right\rceil ,$$

so that refinement levels with small $B_i(q)$ do not inherit unnecessarily large k. Such a scheme suggests an analysis in the style of confidence sequences: rather than fixing δ_i a priori, we may adapt δ_i online based on observed empirical gaps between the highest estimated branch scores. This would convert the fixed logarithmic schedule into a data-dependent one that is still worst-case safe (by reserving a minimum budget) but that reduces work on "easy" queries where separation is strong. A corresponding open problem is to characterize the minimal $\sum_i k_i(q)$ sufficient for uniform success when

the $B_i(q)$ are random (e.g. induced by variable-length inputs) and when refinement trees are unbalanced.

From attention retention to logit-faithful certificates. A second direction is to replace an attention-centric notion of fidelity by a certificate at the level of model outputs. Mask construction from partial score observations is ultimately instrumental: we care about preserving behavior, not merely retaining edges with large $\alpha(Q_q, K_r)$. This motivates certificates that bound the change in logits (or in a task-specific functional) under masking. One natural approach is to treat sparsification as a structured perturbation and to upper bound its effect via local smoothness of the mapping from attention weights to the residual stream. Concretely, given a candidate sparse mask M, we may seek bounds of the form

$$\|\Delta z_\ell\| \le \sum_h \mathrm{Lip}_{\ell,h} \cdot \|\Delta A_{\ell,h}\|, \qquad \|\Delta \mathrm{logits}\| \le \mathrm{Lip}_{\mathrm{out}} \cdot \sum_\ell \|\Delta z_\ell\|,$$

where $\Delta A_{\ell,h}$ is the difference between dense and sparse attention probability matrices at head (ℓ, h), and $\mathrm{Lip}_{\ell,h}$ are computable (possibly conservative) Lipschitz-like constants derived from operator norms of value projections and downstream mixing. While worst-case operator-norm bounds are typically loose, the certificate can be made empirically meaningful by estimating these constants on the current input (or calibrating them on a validation distribution) and by tying them to δ-style risk parameters. An alternative, more distributional, route is randomized certification: sample multiple masks from a controlled family around the selected $S(q)$, and use concentration to certify stability of logits under mask perturbations. In either case, the central question is to connect our probabilistic retention guarantee for "critical branches" to a quantitative guarantee on functional deviation, and to identify when the separation condition (H3) is a useful proxy for logit robustness.

Incorporating residual and MLP tracing. A third extension is to broaden the traced objects beyond attention edges. Even if we retain the correct key blocks, information can propagate through residual pathways, layer normalization, and MLP sublayers in ways not captured by attention-only influence. We therefore view sparse attention masks as one component of a larger tracing problem on the transformer computation graph. One promising formulation is to define a set of *behavior-critical features* in the residual stream and to track their dependence across layers via linear probes or causal scrubbing operations, then to impose constraints on the mask so that these features remain reconstructible. Operationally, this could take the form of augmenting branch scores with feature-based signals (e.g. correlations with a monitored direction in activation space), or of allocating budget not only over key blocks but also over MLP "experts" or neurons in

architectures where such structure exists. A minimally invasive variant is to keep PA-STREAM as the attention selector but to add a post hoc verification step: patch out (or ablate) the retained blocks and measure the induced change in intermediate activations at designated sites, iteratively increasing $k(q)$ or r until the measured deviation falls below a tolerance. The mathematical challenge is then to develop a compositional analysis in which the uncertainty from attention selection and the uncertainty from feature tracing can be jointly controlled under a global failure probability.

Systems and empirical questions. Finally, several practical issues remain open even within the attention-only setting. We would like tight constant factors relating α to σ^2/Δ^2 in regimes where branch-score noise is heteroskedastic and depends on token content; we would like schedules that respect a global budget $\sum_q k(q) \leq B$ while preserving uniform risk guarantees (necessitating a coupling argument across q); and we would like extensions beyond purely causal masks to sliding windows or hybrid retrieval caches where $\mathcal{V}(q)$ is not simply $\{1, \ldots, q\}$. Empirically, the central question is whether the branches that are "critical" under the model's own score geometry coincide with the branches that are critical for downstream behavior, and how this alignment varies across layers and heads. A mature monitoring story at million-token scale will therefore combine (i) principled position-adaptive schedules that neutralize causal position bias, (ii) refinement procedures that adapt to realized branching and uncertainty, and (iii) certificates and tracing mechanisms that speak directly to functional behavior rather than to attention structure alone.