Certified Logit-Faithful Sparse Tracing via Softmax Tail Bounds

Liz Lemma Future Detective

January 18, 2026

Abstract

Mechanistic interpretability methods that inspect attention patterns scale quadratically with context length, making million-token analysis infeasible. Recent work (STREAM/SPARSE TRACING) uses hierarchical sparse attention to prune 90–99

Table of Contents

8. 8. Experiments (recommended to strengthen contribution): certificate tightness vs measured KL/logit error; comparison to nmatch-based search; robustness across decoding temperature/sampling; RULER and long-CoT/tool traces.

9. 9. Limitations and open problems: tightness in deep networks, best possible certificates, and hardness barriers.

10. 10. Conclusion.

1 Introduction

A large portion of contemporary long-context interpretability proceeds by replacing a dense attention computation with a proxy and then evaluating the proxy by a token-matching criterion: one reports whether the proxy preserves the argmax next token, the top-k set, or a small overlap score between the proxy and baseline predictions. Such criteria are convenient, but they are brittle in the sense that they conflate *semantic* agreement with *distributional* agreement. In particular, the map from logits to the predicted token is discontinuous: an arbitrarily small ℓ_∞ perturbation of the logits may flip the argmax whenever two candidates are close. Consequently, token-matching can declare failure for a proxy whose induced distribution is essentially identical to the baseline, and it can declare success for a proxy that substantially reshapes the probability mass while leaving the top prediction unchanged. When the purpose of the proxy is interpretability—to attribute model behavior to specific context positions or blocks—this brittleness is not a benign artifact of evaluation; it obscures whether a purported "explanation" is faithful to the mechanism being explained.

The long-context regime exacerbates these issues. For a sequence of length T, dense attention is a quadratic object, and any attempt to trace "which past tokens matter" must, at a minimum, select a subset of keys/values to retain or inspect. When T is large, the space of possible subsets is enormous and the score landscape can be highly non-uniform: there may exist many moderately relevant tokens whose collective contribution is non-negligible even if no single token dominates. A proxy that retains only the apparent maxima may therefore mis-estimate the contribution of the tail, while a proxy that retains a large random sample may be expensive and still lack a meaningful correctness guarantee. From the interpretability perspective, we wish to isolate a small set of context blocks that *suffices* to reproduce the model's predictive distribution up to a specified tolerance, and we wish to do so without constructing the dense attention matrix.

These considerations motivate a different target: rather than asking whether a proxy matches the baseline token, we ask for *certifiable distributional fidelity*. Concretely, we view the next-token prediction as a probability vector obtained by a softmax of logits, and we adopt a divergence such as KL to measure discrepancy between the baseline distribution and the proxy distribution. The critical point is that this objective admits a route to *sound certification*. If we can upper-bound the deviation in logits under a proxy computation, then stability properties of the softmax map convert this into a bound on distributional error. This shift replaces an unstable, discontinuous notion of correctness by a stable one, and it permits us to reason about approximation error by inequalities rather than by empirical token-level coincidence.

Our focus is the central computational bottleneck: attention. We con-

sider the task of constructing a sparse attention mask that retains only a small number of key blocks per query block, while provably controlling the resulting error in downstream logits and hence in the next-token distribution. The naive approach—materialize all attention scores, compute the exact omitted mass, and then certify the proxy—is precisely what is infeasible at large T. Thus the key technical requirement is to estimate, *without* dense materialization, an upper bound on the softmax probability mass assigned to keys that we discard. Once such a tail-mass bound is available, we can translate it into a bound on the difference between the dense attention output and the sparse, renormalized output, and finally into a bound on logit perturbation.

The algorithmic challenge is therefore twofold. First, we must *find* a sparse set of blocks that plausibly contains the dominant attention contributors, with time and memory near-linear in T. Second, we must *certify* that the remaining, unevaluated part of the attention computation cannot collectively change the logits by more than a user-specified tolerance. These tasks are naturally coupled: the certificate should improve monotonically as we refine the sparse set, and refinement should prioritize precisely those omitted branches that contribute most to the current upper bound on tail mass. In this way, the procedure behaves as an anytime algorithm: it begins with a small candidate set, computes a conservative certificate, and selectively expands the candidate set until the desired fidelity criterion is met.

We instantiate this program via a certified sparse tracing method built on hierarchical pruning in the style of STREAM/HiP. Hierarchical organization of key blocks allows us to reason about *branches* of the key space, maintaining for each discarded branch an upper bound on the maximum attention score within that branch. Such bounds need not be tight to be useful; they need only be *sound*. Summing the resulting exponential upper bounds yields a conservative upper bound on the omitted log-sum-exp, and hence on omitted softmax mass. Importantly, this computation is compatible with streaming: we evaluate scores only for retained blocks, keep a small set of branch statistics, and never store a $T \times T$ object. The certificate is computed in the same pass as the sparse attention itself, and its monotone dependence on the retained set provides a principled stopping rule.

Contributions. We summarize our contributions as follows.

1. We formulate *certified sparse tracing* for attention as the problem of constructing a block-sparse mask together with a *sound* upper bound on the induced ℓ_∞ logit error, sufficient to guarantee a target KL tolerance for the next-token distribution.

2. We provide a certification mechanism based on (i) an upper bound on omitted softmax mass derived from hierarchical score bounds, and

(ii) a conversion of omitted mass into an attention-output error bound under a uniform bound on value norms. This yields a closed-form certificate that is computable without dense attention.

3. We integrate certification with hierarchical pruning to obtain an efficient algorithmic procedure that runs in near-linear time (up to polylogarithmic factors), uses linear memory, and refines adaptively until the fidelity target is met.

4. We discuss compositional extensions beyond the last-layer substitution setting by propagating certified perturbation bounds through subsequent linear maps via explicit Lipschitz constants, and we delineate where additional assumptions are required for soundness.

5. We complement the algorithmic results with a hardness perspective: even in simplified settings, selecting a minimum-size faithful subset can be computationally intractable, which justifies the emphasis on efficient heuristics equipped with conservative certificates rather than exact optimality.

The resulting viewpoint is that interpretability-oriented sparsification should be judged not by whether it reproduces a particular sampled token, but by whether it provably preserves the predictive distribution to within a stated tolerance. By coupling sparse selection to an explicit certificate, we obtain a method that is both practically implementable in long-context settings and mathematically accountable in the sense that it never underestimates its own approximation error under the stated hypotheses.

2 Preliminaries and background

We fix an input sequence $x = (x_1, \ldots, x_T)$ of length T and consider a decoder-only transformer in which each position t produces a representation by attending to a set of *valid* key positions. We encode validity by a binary mask $C \subseteq [T] \times [T]$, where $(t, j) \in C$ denotes that key position j is visible to query position t. In the standard causal setting, $(t, j) \in C$ if and only if $j \leq t$, possibly with additional constraints (padding, document boundaries, local windows). For a fixed query position t we write the valid key set as $\mathcal{J}_t := \{j \in [T] : (t, j) \in C\}$.

Single-head attention. For a single attention head with head dimension d, we write $q_t \in \mathbb{R}^d$ for the query at position t and $(k_j, v_j) \in \mathbb{R}^d \times \mathbb{R}^d$ for the key/value at position j. The (scaled) score of key j for query t is

$$s_{tj} \;:=\; \langle q_t, k_j \rangle,$$

where the conventional $1/\sqrt{d}$ scaling may be absorbed into the vectors. Dense attention assigns weights

$$p_{tj} \;:=\; \frac{\exp(s_{tj})}{\sum_{i\in\mathcal{J}_t}\exp(s_{ti})} \qquad (j\in\mathcal{J}_t),$$

and produces the output vector

$$o_t \;:=\; \sum_{j\in\mathcal{J}_t} p_{tj}\, v_j \in \mathbb{R}^d.$$

Our subsequent analysis is headwise; multi-head attention is handled by applying the same reasoning per head and combining the resulting perturbation bounds by triangle inequalities and explicit operator norms. Since our certificates ultimately target the next-token logits, we emphasize that the attention computation enters only through o_t and that the central difficulty is to approximate o_t without evaluating all scores s_{tj}.

Block partitioning and blocked primitives. To obtain near-linear complexity, we impose a block structure on both queries and keys. Fix query and key block sizes (b_q, b_k). We partition positions into consecutive query blocks $Q^{(a)}$ of b_q tokens and key blocks $K^{(r)}, V^{(r)}$ of b_k tokens. Concretely, if block a contains query positions $t \in I_a$ with $|I_a| = b_q$, we collect the corresponding queries into a matrix $Q^{(a)} \in \mathbb{R}^{b_q\times d}$; similarly, if block r contains key positions $j \in J_r$ with $|J_r| = b_k$, we collect keys and values into matrices $K^{(r)}, V^{(r)} \in \mathbb{R}^{b_k\times d}$. The score submatrix between blocks is then the blocked dot product

$$S^{(a,r)} \;:=\; Q^{(a)}(K^{(r)})^\top \in \mathbb{R}^{b_q\times b_k}.$$

The causal mask induces a blockwise validity pattern (and, within a partially valid block, a triangular submask). The computational constraint relevant for long context is that we may compute selected blocks $S^{(a,r)}$ and perform partial reductions (e.g. log-sum-exp and weighted sums with $V^{(r)}$), but we do not store or traverse all (a, r) pairs. In particular, we avoid materializing any dense $T \times T$ score matrix.

Sparse retention and renormalization. Given a query position t (or a query block a), we will retain only a subset $S \subseteq \mathcal{J}_t$ of keys (typically structured as a subset of key blocks). The corresponding *renormalized* sparse weights are

$$p^S_{tj} \;:=\; \frac{\exp(s_{tj})}{\sum_{i\in S}\exp(s_{ti})} \qquad (j\in S),$$

and the sparse attention output is

$$o_{t,S} \;:=\; \sum_{j\in S} p^S_{tj}\, v_j.$$

Renormalization is the natural choice when the sparse computation is intended to approximate the dense conditional distribution over keys rather than to approximate the unnormalized sum; it also yields a clean dependence of output error on the omitted probability mass.

Hierarchical pruning (STREAM/HiP style). The purpose of hierarchical pruning is to locate high-scoring keys and to upper-bound the contribution of the unevaluated remainder using only coarse information. We view the key blocks as leaves of a rooted tree over indices $r \in \{1, \dots, T/b_k\}$, typically a balanced binary tree whose internal nodes correspond to unions of consecutive leaf blocks. For a fixed query block, a *branch* B denotes an internal node and its associated set of leaf indices; we write $|B|$ for its number of keys (or, equivalently, b_k times the number of leaf blocks). A pruning procedure adaptively expands a small number of promising branches, evaluates exact score blocks for selected leaves, and discards the rest while maintaining, for each discarded branch B, an upper bound

$$u_B \;\geq\; \max_{j \in B} s_{tj}$$

for every query token t under consideration (or a uniform bound for all t in the query block, depending on implementation). The mechanism by which u_B is obtained is method-dependent: one may use norm inequalities (e.g. $\langle q_t, k_j \rangle \leq \|q_t\|_2 \|k_j\|_2$ together with precomputed bounds on $\|k_j\|_2$ within B), or sampled/probed scores augmented by a safety margin. For certification, the only essential requirement is soundness of the inequality above; tightness affects efficiency but not correctness.

Streaming log-sum-exp bookkeeping. For a retained set S (keys or blocks) and a partition $\{B\}$ of the discarded keys, we will repeatedly use two scalars: an exact retained log-sum-exp

$$L_{\text{keep}} \;:=\; \log \sum_{j \in S} \exp(s_{tj}),$$

and an upper bound on the discarded log-sum-exp

$$U_{\text{disc}} \;:=\; \log \sum_{B} |B| \, \exp(u_B).$$

Both quantities admit stable streaming computation using repeated $\log(\exp(a) + \exp(b))$ updates; crucially, U_{disc} depends only on the branch bounds and sizes, not on per-key scores. This separation is what permits us to upper-bound omitted softmax mass without inspecting each discarded key.

Fidelity metrics: logits, total variation, and KL. Ultimately we measure fidelity at the level of the next-token distribution. Let $z, z' \in \mathbb{R}^{|\mathcal{V}|}$ be two logit vectors (baseline and proxy) and let $p = \mathrm{softmax}(z)$, $q = \mathrm{softmax}(z')$. We will certify an ℓ_∞ logit perturbation bound $\|z - z'\|_\infty \le \eta$, because (i) it composes cleanly through linear maps and (ii) it yields distributional guarantees via softmax stability. In particular, a standard stability argument shows that $\|z - z'\|_\infty \le \eta$ implies pointwise likelihood-ratio control $p_i/q_i \le e^{2\eta}$ and therefore $\mathrm{KL}(p\|q) \le 2\eta$. If desired, one may also translate the same logit bound into total variation distance $\mathrm{TV}(p, q) := \frac{1}{2}\|p - q\|_1$ using either direct bounds in terms of e^η factors or via Pinsker's inequality $\mathrm{TV}(p, q) \le \sqrt{\mathrm{KL}(p\|q)/2}$. Our algorithmic certificate will therefore be stated in terms of η (specialized later to $\widehat{\Delta}$), with KL as the primary distributional metric.

From attention-output error to logit error. To connect attention approximation to distributional fidelity, we recall the standard last-layer relationship: in a single-layer setting, or when only the final attention layer is modified, the next-token logits take the form $z = Wo + b$ for some readout matrix W and bias b, where o denotes the relevant attention output (or a linear function of it). Thus if o' is a proxy attention output, then

$$\|z - z'\|_\infty \;=\; \|W(o - o')\|_\infty \;\le\; \|W\|_{\infty\to 2}\, \|o - o'\|_2,$$

where $\|W\|_{\infty\to 2} := \max_i \|W_{i:}\|_2$ is the operator norm from ℓ_2 to ℓ_∞ given by the maximum row norm. This reduction motivates bounding $\|o - o_S\|_2$ for renormalized sparse attention outputs. In the subsequent development, this bound will be expressed in terms of the omitted softmax mass and a uniform bound $V_{\max}$ on value norms, thereby closing the loop from hierarchical score bounds to certified KL fidelity.

3 Problem formulation: certified sparse tracing (CST)

We formalize the task of replacing dense attention by a block-sparse surrogate while certifying that the induced change in next-token predictions is small. Throughout, the baseline computation is the model output on the given input x, and the proxy computation differs only by restricting certain attention computations to a sparse, validity-respecting mask.

Objects being approximated. Fix a layer/head and a query position (or query block) under the validity mask C. Let $z(x) \in \mathbb{R}^{|\mathcal{V}|}$ denote the baseline next-token logits produced by the model on input x, and let $z_M(x)$ denote the logits produced when we substitute a sparse attention operator according to a mask M (with renormalization on the retained keys) at the designated locations. The mask M is structured at the block level: for

each query block index a we retain a subset of key block indices r, and within partially valid blocks we still respect the induced submask from C. We measure fidelity primarily via the logit perturbation $\|z(x) - z_M(x)\|_\infty$, since it admits a direct stability implication for the softmax distribution (cf. Thm. 1) and composes through linear maps.

The certified sparse tracing problem. Given a tolerance $\varepsilon > 0$, our goal is to *construct* a sparse mask M together with a *certificate* $\widehat{\Delta}$ such that the certificate is sound and strong enough to imply the desired distributional guarantee. Concretely, we require

$$\|z(x)-z_M(x)\|_\infty \;\le\; \widehat{\Delta} \qquad \text{and} \qquad \mathrm{KL}(\mathrm{softmax}(z(x)) \,\|\, \mathrm{softmax}(z_M(x))) \;\le\; 2\widehat{\Delta} \;\le\; \varepsilon. \tag{1}$$

The point of the certificate is that it is computed from information available to the tracing algorithm (e.g. score upper bounds on discarded branches and exact statistics on retained blocks) and does *not* rely on evaluating the dense attention matrix. We emphasize that $\widehat{\Delta}$ is an *upper bound*: it may be conservative, but it must never underestimate the true error under the stated hypotheses.

Decision and optimization variants. It is useful to distinguish three related problem statements.

1. *Decision-CST (fidelity feasibility).* Given (x, ε) and a sparsity budget B (e.g. a bound on the number of retained key blocks per query block, or an upper bound on total block evaluations), decide whether there exists a validity-respecting block-sparse mask M of cost at most B such that (1) holds. We do not attempt to solve this decision problem exactly; it serves to clarify the meaning of "minimality" and to motivate hardness phenomena.

2. *Optimization-CST (minimal sparsity).* Given (x, ε), find a mask M minimizing a chosen cost functional $\mathrm{cost}(M)$ subject to $2\widehat{\Delta}(M) \le \varepsilon$, where $\widehat{\Delta}(M)$ is a sound certificate computable without dense attention. Typical costs are (i) total number of retained key blocks across all query blocks, (ii) total number of evaluated score blocks, or (iii) a weighted proxy for FLOPs.

3. *Constructive-CST (any feasible certified mask).* Given (x, ε), output some mask M and certificate $\widehat{\Delta}$ satisfying (1), with near-linear resource usage. Our algorithmic focus is on this constructive variant, with an implicit secondary objective of producing a small mask via adaptive refinement.

As discussed later via a reduction, exact minimality in Optimization-CST is intractable in general; consequently, we seek efficient procedures that return

a certified feasible mask and allow monotone refinement when the initial sparsity is insufficient.

What the algorithm is allowed to compute. We adopt a streaming/random-access RAM model with GPU-like blocked linear algebra primitives. For each head/layer where we sparsify attention, the algorithm may: (i) compute selected blockwise score products $S^{(a,r)} = Q^{(a)}(K^{(r)})^\top$ for chosen query/key block pairs (a, r); (ii) perform stable partial reductions over selected scores, including log-sum-exp updates and weighted sums with $V^{(r)}$; (iii) maintain per-branch upper bounds u_B in a hierarchical partition over key blocks, using any sound bounding method (norm-based, sampled plus margin, or other); (iv) perform top-k style selection over a small number of branch or block scores to decide which portions of the hierarchy to refine next. The algorithm is *not* allowed to materialize the full $T \times T$ score matrix, nor to store per-query dense attention weights. In particular, any step that implicitly requires enumerating all key blocks for each query block is disallowed.

Resource targets. Our target complexity is $O(T \operatorname{polylog}(T))$ time and $O(T \operatorname{polylog}(T))$ space, with the canonical instantiation achieving $O(T \log T)$ dot-product work and $O(T)$ auxiliary memory per head/layer (up to constant factors depending on b_q, b_k). The space budget covers (a) the retained block indices for M, (b) running statistics needed for certification (e.g. retained log-sum-exp scalars and discarded-branch bound accumulators), and (c) the bookkeeping required by hierarchical pruning. The salient constraint is that memory must scale essentially linearly in context length; we cannot cache dense activations indexed by all (t, j) pairs.

Outputs and certificates. For each sparsified attention instance (typically per head and query block), the algorithm outputs:

- a block-sparse retained set S (or equivalently a mask M) satisfying validity constraints induced by C;
- a numerical certificate, either directly as $\widehat{\Delta}$ or via intermediate certified quantities (e.g. an upper bound $\widehat{P}_{\text{tail}}$ on omitted softmax mass) that deterministically imply $\widehat{\Delta}$ under explicit norm bounds.

When multiple heads contribute additively to a residual stream, we combine headwise certificates by triangle inequality and explicit operator norms. When sparsifying multiple layers, we require either (i) a last-layer-only substitution (where earlier computations match exactly), or (ii) a compositional analysis with stated Lipschitz constants that upper-bound how perturbations propagate through subsequent layers. The constructive output is therefore a *mask plus a proof obligation* in numeric form: the certificate must be

checkable from the traced quantities without appealing to inaccessible dense computations.

Separation of mask search from certification. Although mask construction and certification are intertwined in our implementation, it is conceptually useful to separate them. Mask search proposes a retained set S by hierarchical exploration of high-scoring regions, while certification upper-bounds the contribution of the unexplored remainder. The key requirement is *soundness under partial information*: the algorithm must be able to certify that the unexplored region cannot carry enough softmax mass to violate the target tolerance. This is precisely why we insist on maintaining explicit upper bounds on discarded branches rather than relying on heuristic score proxies alone.

Summary. CST asks for an efficiently computable, validity-respecting block-sparse substitution of attention together with a rigorous numerical certificate of next-token distribution fidelity. The decision/optimization variants clarify that (i) we seek feasibility with guarantees rather than exact minimality, and (ii) the algorithm must operate under strict streaming constraints. In the next section we show how hierarchical score upper bounds yield tail-mass certificates, which in turn imply explicit bounds on attention-output error, logit error, and KL divergence.

4 Certificates from softmax tail bounds

We now derive the certificate that accompanies a block-sparse attention substitution. The argument is local to a single attention instance (one head, one query position or query block) and then aggregates across heads (and, when applicable, across layers). Throughout this section we treat the retained set S as given; in §5 we explain how S is constructed and refined while maintaining the same certified quantities online.

4.1 From score upper bounds to omitted softmax mass

Fix a query vector (or query block) and let $\mathcal{J}$ denote the set of valid key indices under the causal/validity mask. Write the attention scores as $\{s_j\}_{j\in\mathcal{J}}$ and the dense attention weights as

$$p_j \;=\; \frac{e^{s_j}}{\sum_{i\in\mathcal{J}} e^{s_i}}\,.$$

Given a retained subset $S \subseteq \mathcal{J}$ (typically block-structured), the central quantity controlling approximation error is the omitted probability mass

$$P_{\text{tail}} := \sum_{j \notin S} p_j = \frac{\sum_{j \notin S} e^{s_j}}{\sum_{i \in \mathcal{J}} e^{s_i}}.$$

Our goal is to upper-bound P_{tail} without enumerating all $j \notin S$.

To do so, we assume that the discarded indices $\mathcal{J} \setminus S$ are covered by a collection of disjoint "branches" $\{B\}_{B \in \mathcal{D}}$ arising from a hierarchical partition over key blocks (for instance, a binary tree over contiguous blocks). For each discarded branch B, the tracing procedure maintains a scalar upper bound u_B satisfying the soundness condition

$$u_B \geq \max_{j \in B} s_j. \tag{2}$$

The method by which u_B is obtained is deliberately left abstract here: norm-based bounds, sampled estimates with deterministic safety margins, and other head-specific bounding schemes are admissible provided (2) holds.

We combine exact retained statistics with upper bounds for the remainder. Define the retained log-sum-exp

$$L_{\text{keep}} := \log \sum_{j \in S} e^{s_j}, \tag{3}$$

which is computable exactly while streaming over retained blocks (using the standard stable two-term update for log-sum-exp), and define the discarded upper log-sum-exp

$$U_{\text{disc}} := \log \sum_{B \in \mathcal{D}} |B| \, e^{u_B}, \tag{4}$$

where $|B|$ denotes the number of valid keys in branch B (or, at block granularity, the number of valid positions after applying the mask restricted to that branch). By (2), we have the deterministic domination

$$\sum_{j \notin S} e^{s_j} = \sum_{B \in \mathcal{D}} \sum_{j \in B} e^{s_j} \leq \sum_{B \in \mathcal{D}} \sum_{j \in B} e^{u_B} = \sum_{B \in \mathcal{D}} |B| \, e^{u_B} = e^{U_{\text{disc}}}.$$

Consequently, the omitted mass admits the upper bound (cf. Thm. 2)

$$P_{\text{tail}} = \frac{\sum_{j \notin S} e^{s_j}}{\sum_{j \in S} e^{s_j} + \sum_{j \notin S} e^{s_j}} \leq \frac{e^{U_{\text{disc}}}}{e^{L_{\text{keep}}} + e^{U_{\text{disc}}}} =: \widehat{P}_{\text{tail}}. \tag{5}$$

We emphasize two practical properties of (5). First, it is *streaming-friendly*: L_{keep} is accumulated from retained score blocks, and U_{disc} is accumulated from branch summaries, neither of which requires materializing the dense score matrix. Second, it is *monotone under refinement*: any refinement step that evaluates additional blocks and tightens (or replaces) some of the branch bounds can only increase L_{keep} and/or decrease U_{disc}, hence can only decrease $\widehat{P}_{\text{tail}}$.

4.2 From omitted mass to attention-output error

Let $v_j \in \mathbb{R}^d$ denote the value vectors and assume a uniform bound

$$\|v_j\|_2 \;\le\; V_{\max} \qquad \forall j \in \mathcal{J}\,, \tag{6}$$

which is readily obtained (and typically conservative) from the value activations at the relevant layer/head. Define the dense attention output and the renormalized sparse output by

$$o \;:=\; \sum_{j\in\mathcal{J}} p_j v_j, \qquad o_S \;:=\; \sum_{j\in S} p_j^S v_j, \qquad p_j^S \;:=\; \frac{e^{s_j}}{\sum_{i\in S} e^{s_i}}\,.$$

Write $\alpha := 1 - P_{\mathrm{tail}} = \sum_{j\in S} p_j$. Then the restriction of the dense distribution to S is p_j/α for $j \in S$, hence

$$\mu_S \;:=\; \sum_{j\in S} \frac{p_j}{\alpha} v_j \;=\; o_S\,.$$

Similarly, defining $\mu_{\mathrm{tail}} := \sum_{j\notin S} \frac{p_j}{P_{\mathrm{tail}}} v_j$ when $P_{\mathrm{tail}} > 0$, we have the convex decomposition

$$o \;=\; \alpha\mu_S + P_{\mathrm{tail}}\mu_{\mathrm{tail}}\,.$$

Eliminating $\mu_S = o_S$ and using $\alpha = 1 - P_{\mathrm{tail}}$ yields

$$o - o_S \;=\; \frac{P_{\mathrm{tail}}}{1 - P_{\mathrm{tail}}}\,(\mu_{\mathrm{tail}} - \mu_S)\,. \tag{7}$$

By (6), both μ_{tail} and μ_S are convex combinations of vectors with ℓ_2-norm at most $V_{\max}$, so $\|\mu_{\mathrm{tail}}\|_2 \le V_{\max}$ and $\|\mu_S\|_2 \le V_{\max}$, hence $\|\mu_{\mathrm{tail}} - \mu_S\|_2 \le 2V_{\max}$. Combining with (7) gives the quantitative bound (cf. Thm. 3)

$$\|o - o_S\|_2 \;\le\; \frac{2V_{\max}\,P_{\mathrm{tail}}}{1 - P_{\mathrm{tail}}} \;\le\; \frac{2V_{\max}\,\widehat{P}_{\mathrm{tail}}}{1 - \widehat{P}_{\mathrm{tail}}}\,. \tag{8}$$

The dependence on $\widehat{P}_{\mathrm{tail}}$ is sharp in the sense that it captures the correct first-order behavior: for small tail mass, the right-hand side is $2V_{\max}\widehat{P}_{\mathrm{tail}} + O(\widehat{P}_{\mathrm{tail}}^2)$. The singularity as $\widehat{P}_{\mathrm{tail}} \uparrow 1$ simply reflects that renormalization is ill-posed if essentially all softmax mass lies outside S; in practice we ensure that the mask-search stage retains at least one nontrivial region per query block so that $\widehat{P}_{\mathrm{tail}}$ is bounded away from 1.

4.3 From attention-output error to logit and KL certificates

We first treat the setting where the sparsified attention output is fed into a *linear* readout (either because we analyze a single attention layer as a base

case, or because we sparsify only the final attention layer so that all earlier computations match exactly). Let $W \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $b \in \mathbb{R}^{|\mathcal{V}|}$ be such that

$$z \;=\; Wo + b, \qquad z_S \;=\; Wo_S + b\,.$$

For each logit coordinate i, $|(z - z_S)_i| = |\langle W_{i,:},\, o - o_S\rangle| \le \|W_{i,:}\|_2 \, \|o - o_S\|_2$. Taking the maximum over i yields

$$\|z - z_S\|_\infty \;\le\; \|W\|_{\infty\to 2}\, \|o - o_S\|_2, \qquad \|W\|_{\infty\to 2} \;:=\; \max_i \|W_{i,:}\|_2\,. \tag{9}$$

Combining (8) and (9), we obtain the explicit certificate

$$\Delta \;:=\; \|z - z_S\|_\infty \;\le\; \widehat{\Delta} \;:=\; \|W\|_{\infty\to 2} \cdot \frac{2V_{\max}\, \widehat{P}_{\text{tail}}}{1 - \widehat{P}_{\text{tail}}}\,, \tag{10}$$

which is exactly the quantity checked by CST-ATTN in the single-layer / last-layer-substitution regime (cf. Thm. 4).

Finally, the certificate (10) translates into a distributional guarantee via softmax stability (Thm. 1). If $p = \mathrm{softmax}(z)$ and $q = \mathrm{softmax}(z_S)$ and $\|z - z_S\|_\infty \le \widehat{\Delta}$, then

$$\mathrm{KL}(p\|q) \;\le\; 2\widehat{\Delta}. \tag{11}$$

Thus it suffices to enforce $2\widehat{\Delta} \le \varepsilon$ to obtain an ε-KL faithful substitution.

Aggregation across heads and (optionally) layers. When multiple heads contribute additively to a residual stream at a fixed layer, we may compute a headwise $\widehat{\Delta}_h$ of the form (10) (possibly with head-specific $V_{\max,h}$ and $\widehat{P}_{\text{tail},h}$) and combine them by triangle inequality after applying the appropriate linear maps; in the simplest shared-readout case this reduces to summing headwise $\|o_h - o_{S,h}\|_2$ bounds before applying $\|W\|_{\infty\to 2}$. For multiple sparsified layers, a sound extension requires explicit Lipschitz constants for the intervening computations; we treat this compositional case as an optional add-on and focus algorithmically on the last-layer-certified substitution where (10) and (11) are directly applicable.

The remaining question is algorithmic: how do we propose S so that $\widehat{P}_{\text{tail}}$ (hence $\widehat{\Delta}$) is small while keeping the number of evaluated blocks near-linear? We answer this by a hierarchical refinement procedure that maintains L_{keep} and U_{disc} online and refines exactly those discarded branches that dominate (4). This is the content of §5.

5 Algorithm: CST-ATTN via hierarchical pruning and online certification

We describe the certified sparse tracing procedure (CST-ATTN) for a single attention instance (one head, one query block) in the single-layer / last-layer-substitution regime, and then indicate the standard aggregation across

heads. The input to the procedure is a query block $Q \in \mathbb{R}^{b_q \times d}$, a collection of key/value blocks $\{(K_r, V_r)\}_{r=1}^{T/b_k}$, and the validity mask restricted to the query–key block pair. The output is a retained set S of key blocks together with the certificate $\widehat{\Delta}$ defined in (10). Our design constraint is that we never materialize the dense $T \times T$ score matrix; instead, we evaluate only a small number of key blocks per query block and bound the contribution of the remainder using a hierarchical partition.

5.1 Hierarchical partition and branch bounds

We fix a rooted tree $\mathcal{T}$ over key-block indices (e.g., a balanced binary tree over contiguous block ranges). Each node $B \in \mathcal{T}$ corresponds to a set of key indices (at token granularity) or a set of key blocks (at block granularity); we write $|B|$ for the number of valid keys in that node after applying the causal/validity constraints for the current query block. Leaves of $\mathcal{T}$ correspond to single key blocks, so that evaluating a leaf means computing the exact scores for that query block against that key block.

For each node B that we do *not* evaluate exactly, we maintain a scalar bound u_B satisfying the soundness condition (2). Concretely, the algorithm requires a *bounding oracle* that, given the query block and a node B, returns $(u_B, |B|)$ with $u_B \geq \max_{j \in B} s_j$. One admissible instantiation is norm-based:

$$u_B \;=\; \max_{t \in \mathrm{qblk}} \|q_t\|_2 \cdot \max_{j \in B} \|k_j\|_2,$$

where $\max_{j \in B} \|k_j\|_2$ can be precomputed per node in $\mathcal{T}$ (or per key block and aggregated up the tree). Our analysis and certificate use only (2); any tighter deterministic bound improves performance without affecting soundness.

5.2 Maintained state and streaming updates

For a fixed query block, CST-ATTN maintains three pieces of state.

(i) Retained statistics. As we evaluate key blocks and add them to S, we update the retained log-sum-exp L_{keep} from (3) exactly. Operationally, evaluating a key block entails forming the block score matrix $S_{Q,r} = QK_r^\top$ (masked), and then applying a numerically stable log-sum-exp reduction across its entries to update L_{keep}. In the same pass we may accumulate the unnormalized numerator needed for the sparse output,

$$N_{\mathrm{keep}} \;:=\; \sum_{j \in S} e^{s_j} v_j,$$

stored in a stable (e.g., log-scaled) representation; at termination, the sparse output is $o_S = N_{\mathrm{keep}} / \exp(L_{\mathrm{keep}})$.

(ii) Discarded statistics. For nodes not in S, we maintain a set $\mathcal{D}$ of disjoint discarded branches covering $\mathcal{J} \setminus S$. We update U_{disc} from (4) by maintaining the sum $\sum_{B \in \mathcal{D}} |B| e^{u_B}$ in log-space. When we refine a branch B into children $\text{ch}(B)$, we remove the contribution of B and insert the contributions of its children, thereby decreasing (or leaving unchanged) U_{disc}.

(iii) A refinement priority. To decide which discarded region to refine, we store a priority queue keyed by a proxy for each node's contribution to the tail bound, e.g.,

$$\rho(B) \;:=\; \log |B| + u_B,$$

since $\sum_{B \in \mathcal{D}} |B| e^{u_B}$ is dominated by nodes with large $\rho(B)$. Any rule that selects nodes in nonincreasing order of estimated contribution yields the monotonicity properties below.

Given $(L_{\text{keep}}, U_{\text{disc}})$, we compute $\widehat{P}_{\text{tail}}$ by (5) and then compute $\widehat{\Delta}$ by (10). The stopping rule is $2\widehat{\Delta} \le \varepsilon$.

5.3 Refinement loop

We initialize $\mathcal{D}$ to consist of a small number of coarse nodes that cover all valid keys (for instance, the children of the root, or a fixed-depth partition), and initialize $S = \varnothing$ (or include a mandatory local window, if desired, as a heuristic only). For each initial discarded node B we compute u_B and insert it into the priority queue.

The core loop repeats:

1. Compute $\widehat{P}_{\text{tail}}$ and $\widehat{\Delta}$. If $2\widehat{\Delta} \le \varepsilon$, we terminate and output S and $\widehat{\Delta}$.
2. Otherwise, extract from the queue a highest-risk node $B^\star$ (e.g., maximizing $\rho(B)$), and refine it.
3. If $B^\star$ is internal, replace it in $\mathcal{D}$ by its children; for each child B compute its bound u_B and insert it into the queue.
4. If $B^\star$ is a leaf (a single key block), we evaluate that key block exactly: compute masked scores s_j for $j \in B^\star$, update L_{keep} and N_{keep}, and add the block to S (removing it from $\mathcal{D}$).

Optionally, we may evaluate a small batch of leaves per iteration (top-k by ρ) to amortize kernel launches; this changes only constant factors.

5.4 Invariants and soundness

The procedure maintains the following invariants for every iteration.

- *Branch soundness:* Every node $B \in \mathcal{D}$ satisfies $u_B \geq \max_{j \in B} s_j$ (by the bounding oracle). Hence U_{disc} is an upper bound on the discarded log-sum-exp, and $\widehat{P}_{\text{tail}}$ computed from (5) upper-bounds the true omitted mass.

- *Exact retained mass:* L_{keep} equals $\log \sum_{j \in S} e^{s_j}$ for the currently retained set, since it is updated only by exact evaluations of retained blocks.

- *Monotonicity under refinement:* Replacing a discarded node by children can only decrease U_{disc} (because $\sum_{j \in B} e^{s_j} \leq \sum_{B' \in \text{ch}(B)} |B'| e^{u_{B'}} \leq |B| e^{u_B}$ need not hold, but the maintained quantity remains an upper bound and can be tightened), while evaluating a leaf and moving it to S can only increase L_{keep}. Consequently, $\widehat{P}_{\text{tail}}$ and $\widehat{\Delta}$ are nonincreasing over iterations.

By the results of §4, these invariants imply that $\widehat{\Delta}$ is a sound bound on $\|z - \hat{z}_S\|_\infty$ and that $2\widehat{\Delta}$ upper-bounds the KL divergence between the dense and sparsified next-token distributions in the single-layer / last-layer-substitution setting.

5.5 Complexity

Let $n = T/b_k$ be the number of key blocks. For each query block, each refinement step touches $O(1)$ nodes, and each evaluated leaf triggers one block dot product $QK_r^\top$ plus masked reductions. Since the tree depth is $O(\log n)$, a STREAM/HiP-style search that expands only a polylogarithmic number of nodes per level yields $O(\log n)$ to $O(\text{polylog}(n))$ evaluated leaves per query block in typical regimes, and $O(\log n)$ priority-queue operations per refinement step. Summed across the T/b_q query blocks, this gives $O(T \log T)$ block-level dot-product work (up to block-size constants) and $O(T)$ auxiliary memory for storing bounds and sparse indices when computed layer-by-layer. The certificate computation itself adds only constant-factor overhead: maintaining L_{keep} and U_{disc} is $O(1)$ per evaluated block or refined node.

Finally, for multi-head attention within a single layer, we run the above procedure headwise (potentially sharing the same hierarchical tree) and aggregate the resulting headwise output-error bounds via triangle inequality after the head output projections; this does not change the asymptotic complexity.

6 Theoretical results: certified correctness, branch bounds, and lower bounds

We collect the guarantees underlying CST-ATTN. Throughout, we fix a query (or query block) and the associated valid key set $\mathcal{J}$ determined by

the causal/validity mask, and we write scores $\{s_j\}_{j\in\mathcal{J}}$ with dense softmax weights $p_j = \exp(s_j)/\sum_{i\in\mathcal{J}}\exp(s_i)$. For a retained set $S \subseteq \mathcal{J}$ we write the renormalized sparse weights $p_j^S = \exp(s_j)/\sum_{i\in S}\exp(s_i)$ for $j \in S$, the dense output $o = \sum_{j\in\mathcal{J}} p_j v_j$, and the sparse output $o_S = \sum_{j\in S} p_j^S v_j$. The central quantity is the omitted mass $P_{\text{tail}} := \sum_{j\notin S} p_j$, which is not directly computable without dense normalization, and which we upper-bound by $\widehat{P}_{\text{tail}}$ computed from $(L_{\text{keep}}, U_{\text{disc}})$ as in (5).

From logit perturbations to KL. We first isolate the final step that converts a certified ℓ_∞ logit error into a KL guarantee. If $p = \text{softmax}(z)$ and $q = \text{softmax}(z')$ with $\|z - z'\|_\infty \le \eta$, then for each coordinate i we have

$$e^{-\eta} \le e^{z'_i - z_i} \le e^{\eta} \quad \text{and} \quad e^{-\eta} \le \frac{\sum_k e^{z'_k}}{\sum_k e^{z_k}} \le e^{\eta},$$

hence $\log \frac{p_i}{q_i} \le 2\eta$ and therefore $\text{KL}(p\|q) = \sum_i p_i \log(p_i/q_i) \le 2\eta$. Consequently, it suffices to certify $\|z - z_S\|_\infty \le \widehat{\Delta}$ and enforce $2\widehat{\Delta} \le \varepsilon$.

Tail-mass upper bounds from hierarchical score bounds. Let the discarded region $\mathcal{J} \setminus S$ be partitioned into disjoint branches $\mathcal{D}$, and assume that for each $B \in \mathcal{D}$ we maintain a scalar u_B satisfying the branch soundness condition

$$u_B \;\ge\; \max_{j\in B} s_j. \tag{12}$$

Define

$$L_{\text{keep}} := \log \sum_{j\in S} e^{s_j}, \qquad U_{\text{disc}} := \log \sum_{B\in\mathcal{D}} |B|\, e^{u_B},$$

where $|B|$ counts valid keys in B (after masking). Then

$$\sum_{j\notin S} e^{s_j} \;=\; \sum_{B\in\mathcal{D}} \sum_{j\in B} e^{s_j} \;\le\; \sum_{B\in\mathcal{D}} \sum_{j\in B} e^{u_B} \;=\; \sum_{B\in\mathcal{D}} |B| e^{u_B} \;=\; e^{U_{\text{disc}}}.$$

Since the softmax normalizer is $\sum_{j\in\mathcal{J}} e^{s_j} = e^{L_{\text{keep}}} + \sum_{j\notin S} e^{s_j}$, we obtain the computable bound

$$P_{\text{tail}} = \frac{\sum_{j\notin S} e^{s_j}}{e^{L_{\text{keep}}} + \sum_{j\notin S} e^{s_j}} \;\le\; \frac{e^{U_{\text{disc}}}}{e^{L_{\text{keep}}} + e^{U_{\text{disc}}}} =: \widehat{P}_{\text{tail}}.$$

This is the formal justification of the online computation in (5). Importantly, no assumptions on the score distribution are required: (12) alone implies $\widehat{P}_{\text{tail}} \ge P_{\text{tail}}$.

Output error controlled by omitted mass. Assume a uniform value norm bound $\|v_j\|_2 \le V_{\max}$ for all valid keys. Write the dense output as a convex mixture over retained and omitted regions:

$$o = (1 - P_{\text{tail}})\,\mu_S + P_{\text{tail}}\,\mu_{\text{tail}}, \qquad \mu_S := \sum_{j \in S} p_j^S v_j, \qquad \mu_{\text{tail}} := \sum_{j \notin S} \frac{p_j}{P_{\text{tail}}} v_j.$$

By construction $o_S = \mu_S$. Using $\|\mu_S\|_2, \|\mu_{\text{tail}}\|_2 \le V_{\max}$, we have

$$o - o_S = \frac{P_{\text{tail}}}{1 - P_{\text{tail}}}\,(\mu_{\text{tail}} - \mu_S), \qquad \|o - o_S\|_2 \le \frac{P_{\text{tail}}}{1 - P_{\text{tail}}}\,2V_{\max}.$$

Substituting $\widehat{P}_{\text{tail}}$ yields the certified bound

$$\|o - o_S\|_2 \;\le\; \frac{2V_{\max}\,\widehat{P}_{\text{tail}}}{1 - \widehat{P}_{\text{tail}}}. \tag{13}$$

This estimate is tight in the sense that, given only a bound on the omitted mass and value norms, the factor $2V_{\max}$ cannot be improved in general.

End-to-end certificate for last-layer substitution. In the single-layer setting (or when we sparsify only the last attention layer and keep all previous computation fixed), logits satisfy $z = Wo + b$ and $z_S = Wo_S + b$. Therefore

$$\|z - z_S\|_\infty \le \|W\|_{\infty \to 2}\,\|o - o_S\|_2,$$

and combining with (13) gives exactly the certificate form (10):

$$\widehat{\Delta} := \|W\|_{\infty \to 2} \cdot \frac{2V_{\max}\,\widehat{P}_{\text{tail}}}{1 - \widehat{P}_{\text{tail}}} \quad \Rightarrow \quad \|z - z_S\|_\infty \le \widehat{\Delta}.$$

Applying the logit-to-KL implication above yields

$$\mathrm{KL}(\mathrm{softmax}(z)\,\|\,\mathrm{softmax}(z_S)) \le 2\widehat{\Delta},$$

so the stopping rule $2\widehat{\Delta} \le \varepsilon$ is sufficient for ε-KL fidelity. Soundness is immediate: every inequality is one-sided in the conservative direction, and the only hypothesis is the correctness of the branch bounds and the value norm bound.

Bounding oracles: deterministic and sample-aided variants. The only algorithm-dependent component in the preceding chain is (12). A simple deterministic oracle is norm-based: for a node B and query block, one may return

$$u_B := \max_{t \in \text{qblk}} \|q_t\|_2 \cdot \max_{j \in B} \|k_j\|_2,$$

which upper-bounds $\max_{j \in B} \langle q_t, k_j \rangle$ for each t and hence the masked score maxima after taking the maximum over valid (t, j) pairs. The key norms $\max_{j \in B} \|k_j\|_2$ can be preaggregated along the tree. In practice, tighter bounds can be obtained by evaluating a small sample of keys in B to estimate a local maximum score and then adding a *safety margin* chosen to dominate the worst-case deviation; when such margins are derived from distributional assumptions, the resulting certificate becomes high-probability rather than absolute. A conservative hybrid is to use the sample-aided estimate but take u_B as the minimum of that estimate plus margin and the deterministic norm bound, which restores unconditional soundness while typically improving tightness.

Lower bounds: streaming necessity and subset-selection hardness. Two complementary lower bounds explain why CST-ATTN targets certified, near-linear procedures rather than exact minimal masks. First, any algorithm that outputs a nontrivial input-dependent certificate must inspect $\Omega(T)$ information in the worst case: if some position is never queried (directly or through an equivalent computation), an adversary can alter the corresponding key/value so as to change o and hence z while leaving all inspected quantities unchanged, invalidating any claimed bound. Thus $\Omega(T)$ time is information-theoretically necessary even when d is fixed.

Second, finding the *smallest* retained set meeting a tight fidelity constraint is computationally intractable in general. Even in a degenerate instance with uniform scores (so p_j is uniform) and scalar values encoding a multiset $\{a_j\}$, deciding whether there exists a subset S of a prescribed size whose renormalized sparse output equals the dense average reduces to PARTITION. Concretely, with $|S| = n/2$ one requires

$$\frac{2}{n} \sum_{j \in S} a_j \;=\; \frac{1}{n} \sum_{j=1}^{n} a_j \quad \Longleftrightarrow \quad \sum_{j \in S} a_j \;=\; \frac{1}{2} \sum_{j=1}^{n} a_j,$$

which is exactly the PARTITION feasibility condition. Hence, absent additional structure, we should not expect a polynomial-time method to compute an optimal sparsity pattern for a given ε; the appropriate goal is instead what we implement: efficient construction of a mask together with a certificate that is sound by design and tight enough to be useful.

Extensions: compositional multi-layer certificates. The preceding analysis is exact when the sparse substitution occurs in the final attention layer (or in a single-layer model), because the readout map $o \mapsto z$ is linear and known. When we sparsify attention in *multiple* layers, we require a mechanism to transport a certified perturbation at an intermediate representation to a certified perturbation on the final logits. We proceed by an explicit Lipschitz composition bound.

Let the transformer be written as a composition of blocks

$$h^{(0)} := E(x), \qquad h^{(\ell+1)} := F_\ell(h^{(\ell)}) \quad (\ell = 0, \dots, L-1), \qquad z := R(h^{(L)}),$$

where each F_ℓ is a standard decoder block (pre/post-norm is immaterial to the algebra) and R is the final linear readout. Suppose we produce, for a set of layers $\mathcal{L}_{\mathrm{sp}}$, sparse substitutes $\widetilde{F}_\ell$ that differ from F_ℓ only in one attention subroutine, yielding a perturbed trajectory $\widetilde{h}^{(\ell)}$ and logits $\widetilde{z}$. For each sparsified layer $\ell \in \mathcal{L}_{\mathrm{sp}}$ we can certify a bound

$$\left\| F_\ell(h^{(\ell)}) - \widetilde{F}_\ell(h^{(\ell)}) \right\|_2 \;\le\; \widehat{\delta}_\ell, \tag{14}$$

where $\widehat{\delta}_\ell$ is obtained from $\widehat{P}_{\mathrm{tail}}$ via the analogue of (13) together with the linear maps internal to the attention sublayer (multi-head concatenation and output projection). Concretely, if the attention sublayer output is $a^{(\ell)} \in \mathbb{R}^{d_{\mathrm{model}}}$ and we certify $\|a^{(\ell)} - \widetilde{a}^{(\ell)}\|_2 \le \widehat{\alpha}_\ell$, then we may take $\widehat{\delta}_\ell := \widehat{\alpha}_\ell$ for a residual-add block, or $\widehat{\delta}_\ell := \|J_{\mathrm{post}}\|_2 \, \widehat{\alpha}_\ell$ if a subsequent fixed linear map J_{post} is included before the residual connection.

Now assume we have Lipschitz constants $\mathrm{Lip}(F_m)$ for each block F_m with respect to $\|\cdot\|_2$, i.e.,

$$\|F_m(u) - F_m(v)\|_2 \le \mathrm{Lip}(F_m)\, \|u - v\|_2.$$

Define the downstream product

$$\Lambda_\ell \;:=\; \mathrm{Lip}(R) \cdot \prod_{m=\ell+1}^{L-1} \mathrm{Lip}(F_m),$$

interpreting an empty product as 1. A standard telescoping argument then gives the sound global logit bound

$$\|z - \widetilde{z}\|_\infty \;\le\; \sum_{\ell \in \mathcal{L}_{\mathrm{sp}}} \Lambda_\ell \, \widehat{\delta}_\ell, \tag{15}$$

and therefore $\mathrm{KL}(\mathrm{softmax}(z) \| \mathrm{softmax}(\widetilde{z})) \le 2 \sum_\ell \Lambda_\ell \widehat{\delta}_\ell$ by the same logit-to-KL implication used earlier.

The appeal of (15) is that it isolates all multi-layer interaction into explicit constants $\mathrm{Lip}(F_m)$ and $\mathrm{Lip}(R)$. For $R(h) = Wh + b$ we have $\mathrm{Lip}(R) = \|W\|_{\infty \to 2}$ when we transport ℓ_2 hidden-state errors to ℓ_∞ logit errors exactly as in the last-layer case. For the blocks F_m, a conservative but fully explicit choice is obtained by bounding each constituent map: linear maps by spectral norms (or Frobenius norms when one is content with a weaker but cheaper bound), residual additions by $1 + \mathrm{Lip}(\text{sublayer})$, and elementwise nonlinearities by their global slope bounds. Layer normalization can be bounded on a restricted domain: if we guarantee that the per-token

pre-normalization standard deviation is bounded below by $\sigma_{\min} > 0$, then the corresponding normalization map is Lipschitz with constant at most on the order of $\gamma/\sigma_{\min}$ (where γ denotes the learned scale), and we may treat $\sigma_{\min}$ as a model- and data-dependent calibration constant. When such a lower bound is unavailable, we recommend either (i) certifying only last-layer substitutions (where no block Lipschitzes are required), or (ii) adopting a hybrid certificate that is unconditional but loose (taking a worst-case $\sigma_{\min}$) and then reporting empirical tightness.

Budgeting ε across layers and heads. Equation (15) suggests a natural allocation strategy: choose per-layer targets ε_ℓ such that $\sum_{\ell \in \mathcal{L}_{\mathrm{sp}}} \varepsilon_\ell \le \varepsilon$, and enforce the sufficient conditions

$$2\Lambda_\ell \, \widehat{\delta}_\ell \;\le\; \varepsilon_\ell \qquad \text{for all } \ell \in \mathcal{L}_{\mathrm{sp}}.$$

Within a layer, one may similarly sum per-head contributions (triangle inequality after the output projection) and enforce headwise budgets. This yields a practical stopping rule that is local (each head/layer refines until its own allocated tolerance is met) while remaining globally sound by construction.

Variable-k schedules and adaptive refinement. CST-ATTN is naturally cast as an *anytime* procedure: at any intermediate refinement stage we have a retained set S and a valid certificate $\widehat{\Delta}$, and refinement can only decrease $\widehat{\Delta}$. This supports variable-k schedules in which k is not fixed a priori but is chosen per query block and per layer to meet the target tolerance with minimal work.

A simple schedule is to initialize with a small k_0 for every query block, compute $\widehat{P}_{\mathrm{tail}}$ and the resulting $\widehat{\Delta}$, and then iteratively refine only those query blocks whose certificates violate the target. Since refinement is driven by the largest contributors to U_{disc}, we may implement a priority queue keyed by branch contributions (or by the current per-block $\widehat{\Delta}$) so that additional dot products are spent where they reduce the bound the most. Two common constraints are easily incorporated: (i) a global compute budget B, in which case we stop when the queue is exhausted or B is reached and output the best certificate achieved; and (ii) a latency constraint, in which case we cap the number of refinement rounds and report the achieved $\widehat{\Delta}$ as an explicit quality indicator.

We also note an interaction with decoding-time temperature. If the deployed next-token distribution is $\mathrm{softmax}(z/\tau)$ for $\tau > 0$, then an ℓ_∞ logit perturbation $\|z - \widetilde{z}\|_\infty \le \eta$ becomes a τ-scaled perturbation after division, and the same stability argument yields $\mathrm{KL}(\mathrm{softmax}(z/\tau) \| \mathrm{softmax}(\widetilde{z}/\tau)) \le 2\eta/\tau$. Thus higher temperatures admit larger certified logit error at fixed KL tolerance; variable-k schedules may exploit this by dynamically loosening or tightening ε as a function of τ.

Practical calibration to reduce conservatism (while preserving soundness). The dominant looseness in certificates typically comes from worst-case norm bounds and from branch upper bounds u_B. We can reduce conservatism without sacrificing soundness by replacing global constants with *instance-conditioned* (but still certified) bounds. For example, $V_{\max}$ may be taken as $\max_{j \in \mathcal{J}} \|v_j\|_2$ computed on the fly for the relevant layer/head and query block, since this requires only streaming access to the same values already touched by sparse attention; similarly, $\|W\|_{\infty \to 2}$ (or the spectral norms of fixed projections) can be precomputed exactly from weights.

For branch bounds, we recommend a conservative hybrid oracle: for each branch B, compute a cheap deterministic upper bound u_B^{det} (e.g., via $\|q\|_2 \max_{j \in B} \|k_j\|_2$), and optionally compute a sample-aided estimate u_B^{samp} by evaluating a small subset of scores in B and adding a calibration margin m_B. We then set

$$u_B \;:=\; \min\{u_B^{\text{det}},\, u_B^{\text{samp}}\}.$$

Because u_B^{det} is unconditionally sound, the minimum remains sound, while the sample-aided term often tightens U_{disc} substantially. The calibration margin m_B may be chosen empirically from a held-out calibration set by taking a high quantile of observed underestimation errors and then adding a union-bound correction over the maximum number of branches considered; we emphasize that such a purely empirical choice makes u_B^{samp} high-probability rather than absolute, hence the inclusion of u_B^{det} is the mechanism that restores unconditional correctness.

In summary, multi-layer substitution is feasible with explicit Lipschitz bookkeeping; variable-k schedules are a direct consequence of the monotone refinement invariant; and most of the practical gap between certified and observed errors can be traced to conservative constants that admit principled, instance-adaptive tightening without changing the logical structure of the certificate.

7 Experiments

We describe an experimental protocol whose purpose is to validate two claims simultaneously: (i) that CST-ATTN achieves large reductions in attention compute at long context while remaining accurate, and (ii) that the accompanying certificate is *sound* and sufficiently *tight* to be operational as a stopping rule. Throughout, we compare a dense reference forward pass (materialized only for evaluation) against our certified sparse substitution (materialized at runtime), and we report both the *measured* error and the *certified* upper bound produced online.

Evaluation metrics. For each evaluated token position (or query block) we record: (a) the measured logit deviation

$$\Delta \;:=\; \|z - z_S\|_\infty,$$

(b) the measured distributional deviation $\mathrm{KL}(\mathrm{softmax}(z)\|\mathrm{softmax}(z_S))$, and (c) the certificate $\widehat{\Delta}$ computed from $\widehat{P}_{\mathrm{tail}}$ and the relevant operator norms. We summarize tightness via (i) the ratio $\rho := \widehat{\Delta}/\Delta$ (defined only when $\Delta > 0$), (ii) the empirical slack in KL, namely $\widehat{\varepsilon} := 2\widehat{\Delta}$ versus the measured KL, and (iii) the success rate of the implication $\Delta \le \widehat{\Delta}$, which should be identically 1 up to numerical tolerance.

Certificate tightness versus measured error. The primary plot is a scatter of Δ and KL against $\widehat{\Delta}$ across tokens and across contexts. Since the certificate is derived from (a) a tail-mass bound $\widehat{P}_{\mathrm{tail}}$ and (b) worst-case value and readout norms, we expect $\widehat{\Delta}$ to be systematically conservative, yet monotone under refinement. We therefore additionally plot $\widehat{\Delta}$ as a function of refinement work (e.g. number of evaluated key blocks per query block), and we report the smallest attained sparsity for which $2\widehat{\Delta} \le \varepsilon$.

A useful diagnostic is to isolate the contributions of each inequality in the chain

$$P_{\mathrm{tail}} \;\le\; \widehat{P}_{\mathrm{tail}} \quad\Rightarrow\quad \|o - o_S\|_2 \;\le\; \frac{2V_{\max}\widehat{P}_{\mathrm{tail}}}{1-\widehat{P}_{\mathrm{tail}}} \quad\Rightarrow\quad \Delta \;\le\; \|W\|_{\infty\to 2}\,\|o - o_S\|_2,$$

by reporting (i) an estimate of the *true* omitted mass P_{tail} computed from dense attention (evaluation only), (ii) the realized value norms $\|v_j\|_2$ in the instance, and (iii) the realized local linear gain of the readout on the error direction. This decomposition indicates whether looseness is dominated by hierarchical score bounds (u_B), by value-norm bounds, or by the global operator norm $\|W\|_{\infty\to 2}$.

Comparison to heuristic search without certificates (e.g. nmatch-based). We compare CST-ATTN to a representative heuristic that selects blocks by approximate similarity without maintaining a sound tail-mass bound. Concretely, an "nmatch" baseline may (i) retrieve candidate key blocks using an approximate maximum-inner-product or nearest-neighbor routine, (ii) retain the top-k candidates by sampled scores, and (iii) optionally increase k until a *measured* proxy criterion stabilizes (e.g. change in attention output on a small validation subset). Such methods can be competitive in average accuracy but do not provide a worst-case certificate for a given instance.

We therefore evaluate both methods under a common budget axis (evaluated blocks per query block, or achieved wall-clock time), and we compare

the Pareto frontier of (measured KL, compute). For CST-ATTN we additionally report the achieved *certified* frontier (certificate value versus compute). The key question is whether the certificate is tight enough that the certified stopping rule $2\widehat{\Delta} \le \varepsilon$ yields nearly the same compute as an oracle that stops when the *measured* KL first falls below ε. We recommend reporting the overhead of certification itself; operationally this is the additional cost of maintaining L_{keep} and U_{disc} and performing occasional branch refinements, which should be negligible relative to dot products.

Robustness across decoding temperature and sampling. Since deployment often uses temperature scaling and stochastic sampling, we test certificate behavior under a range of decoding parameters. For temperature $\tau > 0$ we evaluate the KL between tempered distributions $\mathrm{softmax}(z/\tau)$ and $\mathrm{softmax}(z_S/\tau)$ and verify the scaling predicted by the logit perturbation lemma, namely that an ℓ_∞ error bound η implies

$$\mathrm{KL}(\mathrm{softmax}(z/\tau) \| \mathrm{softmax}(z_S/\tau)) \;\le\; \frac{2\eta}{\tau}.$$

Empirically, this suggests that higher τ permits larger $\widehat{\Delta}$ at fixed KL tolerance, hence lower compute. We validate this by running identical prompts across a grid of $(\tau, \text{top-}p, \text{top-}k)$ and measuring (i) the achieved sparsity at the same certified tolerance and (ii) the realized change in sampled outputs (e.g. match rate of sampled next tokens). The latter is not certified, but it quantifies practical impact.

Long-context evaluations: RULER and structured long traces. To test long-range behavior, we recommend evaluating on benchmarks designed to probe retrieval and long-context reasoning (e.g. RULER-style tasks), where attention must locate sparse but decisive evidence. For each input length T we report: (i) accuracy on the benchmark task, (ii) average retained blocks per query block, (iii) achieved $\widehat{\Delta}$ distribution, and (iv) scaling of runtime with T. The principal claim is that hierarchical pruning yields subquadratic work while certificates remain feasible, i.e. that $\widehat{P}_{\text{tail}}$ can be driven small without exhaustively scanning keys.

Separately, we recommend long chain-of-thought and tool-use traces (multi-turn dialogues, code execution logs, or agent trajectories) where attention patterns are often highly localized but occasionally exhibit bursts of long-range dependency. Here we report token-wise compute adaptivity: the distribution of retained blocks as a function of position, and the frequency with which refinement triggers additional retrieval. This setting is a stress test for any fixed-k scheme, and it illustrates the advantage of an anytime, certificate-driven procedure.

Ablations: instance-conditioned constants and branch-bound design. Finally, we recommend ablations that isolate the practical impact of tightening constants while preserving soundness: (i) global $V_{\max}$ versus instance-wise $V_{\max}$ computed on the fly, (ii) purely deterministic u_B versus the hybrid $u_B = \min\{u_B^{\mathrm{det}}, u_B^{\mathrm{samp}}\}$, and (iii) different hierarchical partition granularities. For each ablation we report the change in median $\rho = \widehat{\Delta}/\Delta$ and the compute required to satisfy $2\widehat{\Delta} \le \varepsilon$. This directly tests whether certificate conservatism is a fundamental limitation or largely an artifact of avoidable worst-case bounds.

8 Limitations and open problems

Our guarantees are strongest in the setting where the sparse substitution is confined to a single attention map whose output is followed by an affine transformation, or more generally to a "last-layer-only" substitution in which all preceding computations coincide. In this regime the certificate reduces to bounding a single attention-output perturbation and then applying a stable readout bound together with softmax stability. The principal limitation is that the moment we sparsify attention in intermediate layers, the perturbation must be propagated through residual streams, normalization, and MLP blocks, and the resulting certificates can become loose unless we control the relevant operator norms with much greater precision.

Tightness in deep networks. A compositional extension via Lipschitz constants is sound but typically conservative. Concretely, if an intermediate attention output is perturbed by δo and the remainder of the network (from that point to the logits) is upper bounded by a global Lipschitz constant Λ in a suitable norm, then $\|z - z_S\|_\infty \le \Lambda \|\delta o\|$. While such constants exist in principle (e.g. products of spectral norms for linear maps, together with bounds for layer normalization and activation functions on a bounded domain), they can be vacuous in practice: (i) the product-of-norms effect scales poorly with depth, (ii) layer normalization is only Lipschitz on regions excluding vanishing variance, and (iii) the "right" norm for tight propagation is instance-dependent because the perturbations introduced by sparsifying attention are structured and often low-dimensional. We therefore view deep-network certification as an open problem of *instance-wise* sensitivity analysis: we would like bounds that resemble $\|J(x)\delta\|$ where $J(x)$ is a Jacobian of the remaining computation at the current activation point, rather than $\sup_x \|J(x)\|$ over a large set. A promising direction is to certify *local* gains using admissible upper bounds on $\|J(x)\|$ obtained from inexpensive power iterations, automatic differentiation with interval arithmetic, or semidefinite relaxations applied layerwise. The difficulty is to maintain soundness without paying a cost comparable to the dense forward pass.

Best possible certificates for attention outputs. Even in the single-layer setting, our bound $\|o - o_S\|_2 \le \frac{2V_{\max}\widehat{P}_{\text{tail}}}{1-\widehat{P}_{\text{tail}}}$ is intentionally worst-case: it depends only on omitted probability mass and a uniform value-norm bound. When the values within S and outside S are well aligned, the true error can be far smaller than $2V_{\max}P_{\text{tail}}/(1 - P_{\text{tail}})$. Tightening this step while preserving a streaming implementation is nontrivial because it amounts to certifying *geometric* information about the value vectors in the omitted region. One could imagine maintaining additional sufficient statistics for discarded branches, such as bounds on the mean of v_j or on the diameter of the value set within a branch, yielding a refined inequality of the form

$$\|o - o_S\|_2 \le \frac{P_{\text{tail}}}{1 - P_{\text{tail}}} \cdot \operatorname{diam}(\{v_j : j \in \mathcal{J}\}),$$

or a branchwise sum of diameters. However, any such refinement must be computable without scanning all values in discarded branches; thus one is led to hierarchical value summaries, certified range bounds, or precomputed per-block envelopes. Determining which summaries yield the largest practical gains for the smallest overhead remains open, as does the question of whether there exist "universal" summaries (independent of the query) that significantly tighten worst-case bounds.

Bounding the tail mass without losing too much. Our control of $\widehat{P}_{\text{tail}}$ depends on score upper bounds u_B for discarded branches. Deterministic bounds based on norm inequalities (e.g. $\langle q, k\rangle \le \|q\|_2\|k\|_2$) are sound but often loose; sharper bounds can be obtained by storing per-branch extrema of projected keys, by maintaining bounding boxes in a rotated basis, or by using multi-probe upper bounds as in maximum inner product search. The open problem is to design branch bounds that are simultaneously (i) tight in the typical regimes encountered in transformer attention, (ii) cheap to update under refinement, and (iii) robust to quantization and numerical error. In particular, any bound used inside log-sum-exp must account for floating-point roundoff conservatively. A related question concerns *data-dependent* certificates: one may wish to use sampled maxima to tighten u_B, but then soundness requires either a worst-case correction term (which may erase the gain) or a probabilistic statement. Developing a principled "PAC-certificate" variant, with explicit failure probability δ and a tight dependence on δ, would be valuable for deployment settings where a vanishingly small risk of violation is acceptable.

Hardness barriers and the limits of optimal sparsity. There is a fundamental distinction between (a) producing a *sound* certificate for a chosen mask and (b) finding the *sparsest* mask satisfying a given tolerance. Even in highly simplified attention instances, selecting an exact or near-exact subset

under renormalization constraints is NP-hard by reductions from classical subset-sum or PARTITION-type problems. Consequently, one should not expect a polynomial-time algorithm that always returns an optimal minimal mask unless P = NP. This motivates the algorithmic posture taken here: we focus on efficient heuristics guided by upper bounds that monotonically improve under refinement. An open theoretical question is whether one can prove *instance-optimal* or *constant-factor bicriteria* guarantees for hierarchical refinement procedures under realistic assumptions on score distributions (e.g. subgaussian tails, margin conditions, or clusterability of keys). Such results would bridge the gap between worst-case hardness and the strong empirical structure exploited by approximate retrieval.

Streaming constraints and cross-query reuse. Our computational model prohibits materializing dense $T \times T$ attention, and we have an $\Omega(T)$ lower bound for any nontrivial instance-dependent certificate. Within these constraints, it remains open whether the $O(T \log T)$ work suggested by hierarchical pruning is optimal up to constants, or whether one can provably reduce the logarithmic factor for typical contexts by amortizing computation across neighboring query blocks. Attention exhibits strong locality across queries: adjacent queries often share salient key blocks, suggesting that one can reuse retained sets and branch bounds, or maintain a persistent index over keys. Establishing sound reuse rules that preserve the monotonicity and soundness invariants, while handling the causal mask and query drift, is a technically delicate problem.

Beyond last-layer substitution. Finally, we note that sparsifying multiple layers simultaneously poses a compounded certification challenge: the sparsity decisions in early layers can change the very queries and keys used to form later attention scores, invalidating any certificate that assumes fixed Q, K, V. A fully end-to-end certificate would need to bound not only attention output errors but also the induced perturbations in subsequent attention *scores*, which feed back nonlinearly through softmax. We view this as the central open problem for certified sparse tracing: to develop a tractable bound-propagation calculus that can follow the transformer computation graph with acceptable overhead and without collapsing to vacuous constants.

9 Conclusion

We have introduced *certified sparse tracing* for transformer attention: a procedure that constructs an input-dependent block-sparse attention mask together with a *sound* a posteriori certificate controlling the deviation in next-token predictions. The central object of certification is an ℓ_∞ bound on the logit perturbation $\Delta = \|z - z_M\|_\infty$, which, by softmax stability, immediately

yields a bound on the induced divergence between next-token distributions. Operationally, this reframes sparse attention from a purely heuristic acceleration technique into a verifiable approximation: for each input and each query block, we either return a sparse mask with a certified fidelity guarantee, or we refine the mask until the certificate meets a user-specified tolerance.

Our main technical contribution is a certificate that can be computed in the same streaming pass that forms the sparse attention output, without materializing dense $T \times T$ score matrices. The certificate rests on two quantities. First, we compute an exact log-sum-exp over the retained set S,

$$L_{\text{keep}} = \log \sum_{j \in S} e^{s_j},$$

which is available whenever we explicitly evaluate the scores of retained blocks. Second, we maintain an upper bound on the discarded log-sum-exp,

$$U_{\text{disc}} = \log \sum_{B \in D} |B| e^{u_B},$$

where D is a partition of the discarded region into hierarchical branches and each u_B satisfies $u_B \geq \max_{j \in B} s_j$. These two scalars produce a computable tail-mass upper bound

$$\widehat{P}_{\text{tail}} = \frac{e^{U_{\text{disc}}}}{e^{L_{\text{keep}}} + e^{U_{\text{disc}}}},$$

which is sound by construction. We then convert tail mass into an attention-output perturbation bound under a uniform value-norm hypothesis $\|v_j\|_2 \leq V_{\max}$, and finally into a logit certificate via an explicit readout operator norm $\|W\|_{\infty \to 2}$. In the single-layer (or last-layer-only substitution) setting, these steps yield an end-to-end inequality of the form

$$\mathrm{KL}\big(\mathrm{softmax}(z) \,\big\|\, \mathrm{softmax}(z_M)\big) \leq 2\widehat{\Delta} \leq \varepsilon,$$

thereby turning a target KL tolerance into an implementable stopping criterion. The certificate is monotone under refinement: as we evaluate additional key blocks, L_{keep} can only increase while U_{disc} can only decrease, hence $\widehat{P}_{\text{tail}}$ and $\widehat{\Delta}$ decrease. This monotonicity aligns naturally with hierarchical search strategies and supports a practical "anytime" behavior: we may stop early with a weaker certificate under tight compute budgets, or continue refinement until the prescribed tolerance is met.

Algorithmically, we instantiate these ideas in a STREAM/HiP-style hierarchical pruning routine (CST-ATTN) that searches over key blocks using branch score bounds. The certificate computation adds only constant-factor overhead relative to standard hierarchical retrieval, because it requires maintaining the scalars L_{keep} and U_{disc} and a small collection of branch metadata. Under a balanced hierarchy of depth $O(\log(T/b_k))$, this yields the

target $O(T \log T)$ dot-product work and $O(T)$ auxiliary memory (per head and layer, up to block-size constants), respecting the constraint that dense attention must never be formed or stored. From a systems perspective, the decomposition is compatible with GPU kernels based on blocked matrix products and streaming log-sum-exp reductions: the retained blocks are processed exactly, while discarded branches contribute only through inexpensive bound updates. In particular, the certificate is not an afterthought; it is computed *online* and can guide which branches to refine next (e.g. by selecting the branches with largest contributions to U_{disc}).

Conceptually, our results clarify what can be certified efficiently and what cannot. The certification problem—bounding the error of a *given* sparse mask—admits simple, composable upper bounds driven by omitted softmax mass and linear readout stability. By contrast, the optimization problem of finding a *sparsest* faithful mask is in general computationally intractable: even stylized attention instances encode NP-hard subset selection problems under renormalization constraints. This separation justifies a pragmatic methodology: we do not seek globally optimal masks; instead, we develop refinement rules that preserve soundness and steadily tighten a certificate until it meets a specified tolerance. In this sense, certified sparse tracing is best understood as an interface between approximation and verification: approximate retrieval proposes candidates cheaply, while the certificate enforces correctness with respect to a chosen fidelity metric.

The present framework also suggests several concrete directions for continued work. On the theory side, the most immediate opportunity is to tighten the attention-output perturbation bound by incorporating additional geometric information about the value vectors beyond a uniform norm bound, while retaining streaming computability. On the algorithmic side, there is room to improve the sharpness of branch score bounds u_B using richer per-branch summaries or multi-probe upper bounds, and to exploit cross-query structure by reusing retained sets and bounds across neighboring query blocks without compromising soundness. On the integration side, it is natural to combine certified pruning with existing fast attention implementations, quantization schemes, and paging strategies for long contexts, with the certificate serving as a principled control signal for adaptive compute. Finally, extending certification beyond last-layer substitution remains an important objective: an end-to-end transformer certificate must reason about how earlier perturbations alter later queries, keys, and therefore the attention scores themselves.

In summary, we have shown that one can sparsify transformer attention in a streaming setting while producing a per-input, per-query certificate that provably controls next-token distribution drift in KL. The resulting view is neither purely worst-case nor purely heuristic: it is a *verifiable* approximation scheme whose computational cost scales near-linearly in context length and whose fidelity can be tuned continuously via an explicit tolerance

parameter. We expect certified sparse tracing to be useful both as a deployment primitive—where guarantees matter—and as a research tool for probing which parts of the context are truly responsible for a model's predictions under quantitatively controlled approximations.