EXACT SOFTMAX ATTENTION HAS THE RECTANGULAR MATRIX-MULTIPLICATION EXPONENT IN EXTENDED ARITHMETIC CIRCUITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard softmax attention explicitly forms all query–key scores, even when each query produces only a single scalar output. This raises a fundamental question: can exact attention avoid the arithmetic complexity of the underlying rectangular matrix multiplication? We show that it cannot. For N queries and keys of width $d = \lceil N^a \rceil$, with fixed $0 < a \leq 1$, the minimum size of an exact-real circuit with $+, -, \times, /, \exp, \ln$ gates is $N^{\omega(1,a,1)+o(1)}$, even with only one scalar value per key. This matches standard exact evaluation and closes the previous single-head gap between $\omega(1, a, a)$ and $\omega(1, a, 1)$. Our main technical contribution is an exp–log span theorem showing that, modulo additive constants, all within-row score differences lie in the $\mathbb{Q}$-linear span of the circuit's exponential inputs and logarithm outputs. Combined with mixed-coefficient extraction, this recovers the rectangular matrix-multiplication tensor and yields the lower bound. The same theorem also exactly characterizes transcendental complexity, giving $n(m-1)$ exp/log gates for dense independently parameterized attention.

1 INTRODUCTION

Transformers use self-attention to combine information across a sequence (Vaswani et al., 2017). In standard dense attention, each query receives a score for every key before the algorithm normalizes those scores and combines the values. This all-pairs step becomes costly as context length grows. Kernel and low-rank approximations (Choromanski et al., 2021; Wang et al., 2020) and sparse interactions (Zaheer et al., 2020) reduce that cost. These advances raise a basic question: how much work does the exact attention function require?

These methods often approximate or change dense attention, while exact algorithms such as FlashAttention reduce memory traffic (Dao et al., 2022; Saha & Ye, 2024). We ask how many scalar operations exact softmax attention needs on arbitrary real inputs. A large intermediate matrix alone cannot answer this question. For $Q \in \mathbb{R}^{n\times d}$ and $K \in \mathbb{R}^{m\times d}$, one can evaluate $QK^\top v$ as $Q(K^\top v)$ in $O((n+m)d)$ operations without forming the nm entries of $QK^\top$.

The query–key step in standard dense attention multiplies an $N \times d$ matrix by a $d \times N$ matrix. For $d = \lceil N^a \rceil$ and fixed $0 < a \leq 1$, let $\omega(1, a, 1)$ denote the arithmetic exponent of this product. Standard exact evaluation uses $N^{\omega(1,a,1)+o(1)}$ operations. In the same width regime, the lower-bound construction of Saha et al. (2026) encodes a $d \times N$ by $N \times d$ product and yields exponent $\omega(1, a, a)$ per head and layer. They ask how to close this gap, including for a single head. With one value channel, standard evaluation forms N^2 scores but the function returns only N scalars. *Must exact softmax attention still require the full rectangular exponent?*

We show that the answer is yes: even with one scalar value per key, exact softmax attention requires $N^{\omega(1,a,1)+o(1)}$ scalar operations in a highly permissive exact-real circuit model.

To state this result precisely, we consider one-channel attention with independent query, key, and value inputs, where each key carries one scalar value v_j. Let $Q \in \mathbb{R}^{n\times d}$ and $K \in \mathbb{R}^{m\times d}$ contain the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

queries and keys in their rows:

$$\mathcal{A}_{n,d,m}(Q,K,v)_i = \sum_{j=1}^{m} P(QK^\top)_{ij} v_j, \qquad P(QK^\top)_{ij} = \frac{e^{(QK^\top)_{ij}}}{\sum_{t=1}^{m} e^{(QK^\top)_{it}}}. \tag{1}$$

Following Saha et al. (2026), we measure complexity in the exact-real extended arithmetic circuit (eAC) model. Circuits use scalar $+, -, \times, /, \exp, \ln$ gates, may hardwire arbitrary real constants, and allow unrestricted nesting and sharing. We use a nonuniform circuit for each input size and count exact scalar operations rather than bit operations. In this model, for $n = m = N$ and $d = \lceil N^a \rceil$, the minimum circuit size is $N^{\omega(1,a,1)+o(1)}$.

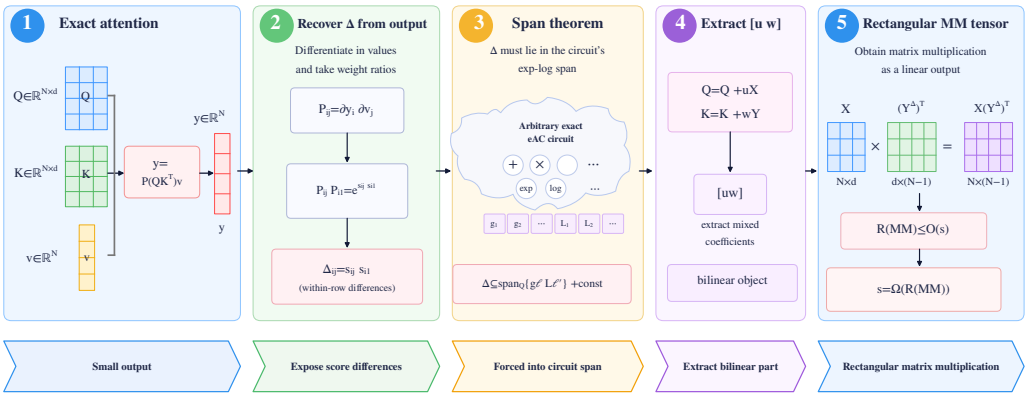


Figure 1: From a small attention output to rectangular matrix multiplication. Value differentiation identifies the softmax weights, whose ratios determine within-row score differences. The span theorem expresses these differences, up to additive constants, as rational linear combinations of internal exponential inputs and logarithm outputs. Mixed coefficient extraction and a fixed linear projection then recover the rectangular matrix-multiplication tensor and give the rank lower bound.

1.1 Our contributions

Our main result closes the single-head polynomial-width gap in exact attention. The key technical ingredient is a new score-difference span theorem, which also yields a broader characterization of exact exp/log gate complexity.

1. **Main result: closing the single-head rectangular-exponent gap.** Theorem 1 gives a finite-size local lower bound in terms of rectangular matrix-multiplication tensor rank. For $n = m = N$ and $d = \lceil N^a \rceil$, Corollary 2 implies that exact one-channel attention requires $N^{\omega(1,a,1)+o(1)}$ operations, matching standard exact evaluation and closing the single-head gap left by Saha et al. (2026).
2. **Technical core: a score-difference span theorem for exact softmax.** The key ingredient behind the lower bound is Theorem 5. For any rational score family, it places all within-row score differences, modulo additive constants, in the $\mathbb{Q}$-linear span of the circuit's exponential inputs and logarithm outputs. For dot-product scores, mixed coefficient extraction turns these span relations into the rectangular matrix-multiplication tensor, providing the bridge from an arbitrary exact attention circuit to Theorem 1.
3. **Broader consequence: exact exp/log gate complexity.** Beyond its role in the total-size lower bound, the same span theorem exactly characterizes transcendental-gate complexity. Theorem 10 shows that, for any rational score family S, the minimum number of exp/log gates equals its score-difference dimension $\tau(S)$. In particular, dense attention with independently parameterized queries and keys requires exactly $n(m-1)$ such gates. The framework also gives the rectangular total-size exponent for tied queries and keys and exact exp/log counts for fixed masks and disjoint-variable instances.

1.2 Related work

The closest work is by Saha et al. (2026), who develop the exact-real eAC framework and the low-degree eAC-to-AC simulation technique used here. At polynomial width, their lower bound gives exponent $\omega(1, a, a)$, and Section 6 asks how to reach $\omega(1, a, 1)$ with a single head. We close that gap with one scalar value per key. Our proof uses their simulation and a new score-difference span theorem to obtain relations for all $N - 1$ difference columns.

Other studies address approximate attention, alternative architectures, and I/O complexity. Fine-grained results give subquadratic algorithms or conditional lower bounds for approximate attention under different entry sizes, temperatures, and accuracy requirements (Duman Keles et al., 2023; Alman & Song, 2023; Gupta et al., 2026). Alman & Yu (2025) study limits of subquadratic alternative architectures. FlashAttention reduces memory traffic for exact attention (Dao et al., 2022), and Saha & Ye (2024) give matching I/O bounds in the large-cache regime, even when fast matrix multiplication is allowed. These results concern approximation-sensitive computation or data movement; our model counts exact scalar operations. Our span argument draws on Ax's differential-field theorem (Ax, 1971; Aslanyan, 2024), value differentiation, and circuit-generated transcendence bounds.

Table 1: The polynomial-width single-head gap. Let $d = \lceil N^a \rceil$ for fixed $0 < a \leq 1$. Standard exact evaluation computes the query–key product, while the prior lower-bound construction recovers a $d \times d$ product. Our one-channel lower bound recovers the score-difference tensor with $N - 1$ columns.

Result	**Guarantee / setting**	**Recovered / computed tensor**	**Exponent**
Standard exact evaluation	Upper bound, one dense head	$\langle N, d, N \rangle$	$\omega(1, a, 1)$
Saha et al. (2026)	Prior eAC lower bound	$\langle d, N, d \rangle$	$\omega(1, a, a)$
This work	**eAC lower bound, one head, one scalar value/key**	$\boldsymbol{\langle N, d, N - 1 \rangle}$	$\boldsymbol{\omega(1, a, 1)}$

2 Circuit Model and Total-Size Lower Bound

2.1 Circuit Model and Preliminaries

We write $[k] = \{1, \ldots, k\}$, $\mathbf{1}$ for an all-ones vector, and $P(S) = \mathrm{softmax}(S)$ for row-wise softmax. The entries of Q, K, v in Eq. (1) are independent inputs unless stated otherwise.

We use standard arithmetic-circuit conventions (Bürgisser et al., 1997). Circuits are scalar directed acyclic graphs with arbitrary fan-out. The arithmetic gates $+, -, \times, /$ have fan-in two; $\exp$ and the real logarithm $\ln$ have fan-in one. Inputs and arbitrary real constants are free, designated outputs are wires to existing nodes, and size is the number of operation gates. An arithmetic circuit (AC) uses only the four arithmetic operations; an extended arithmetic circuit (eAC) may also use $\exp$ and $\ln$.

We formulate exact computation on arbitrary nonempty connected open subsets of the input domain. A *rational score family* $S(z)$ is a matrix whose entries are rational functions of input variables z. Its real domain of definition $\mathrm{Dom}_{\mathbb{R}}(S)$ is the set where the reduced denominators of all score functions are nonzero. We call an input point *regular* for an eAC if every division denominator is nonzero and every logarithm input is positive there. Given a nonempty connected open set $U \subseteq \mathrm{Dom}_{\mathbb{R}}(S) \times \mathbb{R}^m$, an eAC is U-admissible if every point of U is regular and the circuit equals $\mathrm{softmax}(S(z))v$ throughout U. A local lower bound therefore applies to globally correct circuits.

For a field extension F/L, we write $\mathrm{trdeg}_L F$ for its transcendence degree.

We use standard tensor-rank notation for matrix multiplication. For a bilinear map T, $\mathrm{R}(T)$ denotes its tensor rank over $\mathbb{R}$. We write $\langle n, d, m \rangle$ for multiplication of an $n \times d$ matrix by a $d \times m$ matrix. A size-s division-free arithmetic circuit computing a bilinear map has tensor rank $O(s)$. For fixed $0 < a \leq 1$ we use

$$\omega(1, a, 1) = \liminf_{N \to \infty} \frac{\log \mathrm{R}(\langle N, \lceil N^a \rceil, N \rangle)}{\log N}.$$

2.2 TOTAL-SIZE LOWER BOUND

Theorem 1 (Local unrestricted exp–log total-size lower bound)**.** *Let $n \geq 1$, $m \geq 2$, $d \geq 1$, and let $U \subseteq \mathbb{R}^{nd+dm+m}$ be any nonempty connected open set of (Q, K, v) inputs. Every U-admissible unrestricted eAC of size s computing the one-channel attention map $\mathcal{A}_{n,d,m}$ defined in Eq. (1) on U satisfies*

$$s = \Omega(\mathrm{R}(\langle n, d, m-1\rangle)) .$$

The polynomial-width consequence is stated next.

Corollary 2 (Exact one-head rectangular exponent)**.** *For every fixed $0 < a \leq 1$, the minimum size of a globally correct unrestricted eAC for $\mathcal{A}_{N,\lceil N^a\rceil,N}$ is $N^{\omega(1,a,1)+o(1)}$.*

Proposition 14 embeds the square case into fixed-projection self-attention with constant-factor input-width overhead.

2.3 PROOF OVERVIEW

To recover matrix-multiplication complexity from only n outputs, we account for softmax's invariance under row shifts by using

$$(QK^\top)_{ij} - (QK^\top)_{i1} = Q_{i,:}(K_{j,:} - K_{1,:})^\top, \qquad j = 2, \ldots, m.$$

Changing key coordinates identifies these functions with an $n \times d$ by $d \times (m-1)$ product (Section 3).

Linearity in v identifies softmax weights through value differentiation; their ratios are exponentials of score differences. Ax's theorem places the differences in the $\mathbb{Q}$-linear span, up to additive constants, of internal exponential inputs and logarithm outputs (Theorem 5). Differentiation establishes field membership of all softmax weights without enlarging the designated output set.

Mixed coefficients under independent perturbations of Q, K retain this product. Low-degree simulation (Saha et al., 2026, Theorem 4.1) extracts them from internal functions with constant overhead per gate. Their fixed linear projection is the product, so tensor-rank monotonicity gives the bound (Section 5).

Why standard reductions fall short. For N queries and keys and value width $d = \lceil N^a \rceil$, a first-order softmax expansion recovers products with exponent $\omega(1, a, a)$. Coordinate selection recovers only $O(d)$ score columns per attention call and incurs an $N^{1-a+o(1)}$ call overhead. Appendix G gives the detailed reductions, including the anchor, recovery cost, and domain conditions. Appendix F gives a complementary mixed-Hessian identity.

3 THE SCORE-DIFFERENCE TENSOR

We begin with the algebraic part of the argument: within-row differences of dot-product scores form a rectangular matrix-multiplication tensor. For every $c \in \mathbb{R}^n$, softmax satisfies $\mathrm{softmax}(S + c\mathbf{1}^\top) = \mathrm{softmax}(S)$: adding an arbitrary constant to each row does not change the row-wise softmax weights. We therefore work with the $\mathbb{Q}$-linear space generated by within-row score differences modulo constants; its dimension will also determine the exact number of exponential and logarithm gates.

Definition 3 (Score-difference dimension)**.** Let $S(z) = (s_{ij}(z))$ be a rational score family, and choose a reference column r_i in every row. We call $s_{ij} - s_{i,r_i}$ a reference score difference. Define

$$\mathcal{D}(S) = \mathrm{span}_{\mathbb{Q}}\{s_{ij} - s_{i,r_i} : i \in [n], j \in [m]\} + \mathcal{C},$$
$$\tau(S) = \dim_{\mathbb{Q}}\big(\mathcal{D}(S)/\mathcal{C}\big),$$

where $\mathcal{C}$ denotes the $\mathbb{Q}$-vector space of constant functions with values in $\mathbb{C}$. For real score functions, quotienting by $\mathcal{C}$ identifies functions that differ by an additive constant. Equivalently, $\tau(S)$ counts the number of $\mathbb{Q}$-linearly independent within-row score-difference functions after additive constants are ignored. The value is independent of the reference columns; Appendix A also records why restricting to a smaller nonempty open domain does not change it.

For dense dot products, taking all differences relative to one reference column in each row gives a tensor linearly equivalent to rectangular matrix multiplication.

Proposition 4 (Score-difference tensor identity). *For $Q \in \mathbb{R}^{n\times d}$ and $K \in \mathbb{R}^{m\times d}$, let*

$$\mathbf{\Delta}_{ij}(Q,K) = (QK^\top)_{i,j+1} - (QK^\top)_{i1} \qquad (j \in [m-1]).$$

Then

$$\mathrm{R}(\mathbf{\Delta}) = \mathrm{R}(\langle n, d, m-1\rangle),$$

and the $n(m-1)$ coordinate functions of $\mathbf{\Delta}$ are $\mathbb{Q}$-linearly independent modulo constants.

To establish the tensor identity, observe that the invertible input change

$$k_{\mathrm{ref}} = K_{1,:}, \qquad K^{\Delta}_{j,:} = K_{j+1,:} - K_{1,:}, \qquad K^{\Delta} \in \mathbb{R}^{(m-1)\times d}$$

gives $\mathbf{\Delta}(Q,K) = Q(K^{\Delta})^\top$, while the block k_{ref} is unused. The reverse restriction sets $k_{\mathrm{ref}} = 0$. Thus $\mathbf{\Delta}$ and $\langle n, d, m-1\rangle$ are linearly equivalent as bilinear tensors. Independence follows by inspecting the coefficients of the monomials $Q_{i\ell}K^{\Delta}_{j\ell}$.

As a concrete 2×3 example, let K have row vectors k_1, k_2, k_3 and let Q have rows $Q_{1,:}, Q_{2,:}$. Choosing the first score in each row as reference gives

$$\begin{pmatrix} Q_{1,:}(k_2-k_1)^\top & Q_{1,:}(k_3-k_1)^\top \\ Q_{2,:}(k_2-k_1)^\top & Q_{2,:}(k_3-k_1)^\top \end{pmatrix} = Q(K^{\Delta})^\top, \qquad K^{\Delta} = \begin{bmatrix} k_2-k_1 \\ k_3-k_1 \end{bmatrix} \in \mathbb{R}^{2\times d}.$$

Hence the score-difference map is the tensor $\langle 2, d, 2\rangle$. The general case is the same product with n rows and $m-1$ difference columns.

The remaining task is to connect these score differences to the internal functions of an arbitrary circuit for $\mathrm{softmax}(QK^\top)v$. The following span theorem provides this connection.

4 A Span Theorem for Score Differences

Fix a rational score family S and a U-admissible circuit on a nonempty connected open set $U \subseteq \mathrm{Dom}_{\mathbb{R}}(S)\times\mathbb{R}^m$. After specializing the value variables at a regular point, the theorem below states that every score difference lies in the $\mathbb{Q}$-linear span of the circuit's exponential-gate inputs and logarithm-gate outputs, modulo constants.

The proof uses a local analytic argument. Around a regular real point, we complexify the gate functions and work with their local holomorphic and meromorphic germs. We then restrict to a generic affine line that preserves the relevant $\mathbb{Q}$-linear independence relations, reducing the functional-transcendence step to one variable. The branch and complexification details are deferred to Appendices A.2 and B.

Theorem 5 (Exp–log span theorem). *Let S be a rational score family and let $U \subseteq \mathrm{Dom}_{\mathbb{R}}(S)\times\mathbb{R}^m$ be nonempty, connected, and open. Consider a U-admissible unrestricted eAC computing $\mathrm{softmax}(S(z))v$. Choose $(z^\star, v^\star) \in U$ and a product neighborhood contained in U, specialize $v = v^\star$, let $g_1, \dots, g_{r_{\exp}}$ be all exponential-gate inputs and $L_1, \dots, L_{r_{\log}}$ all logarithm-gate outputs, and put*

$$\mathcal{W} = \mathrm{span}_{\mathbb{Q}}\{g_1, \dots, g_{r_{\exp}}, L_1, \dots, L_{r_{\log}}\} + \mathcal{C}.$$

Then

$$\mathcal{D}(S) \subseteq \mathcal{W}.$$

We first use value differentiation to place exponentials of the reference score differences in the specialized gate field. We then combine a gate-prefix bound with Ax's theorem to rule out any score difference outside $\mathcal{W}$.

1. Value differentiation gives score ratios. Because the target is linear in v, differentiating the ith output with respect to v_j gives the softmax coefficient $P(S)_{ij}$. A gate-by-gate chain-rule induction places these derivatives in the field generated by the specialized gate values and rational input

functions. The induction establishes field membership while leaving the designated circuit outputs unchanged. After restricting to the affine line used below,

$$P(S)_{ij} \in F, \qquad \frac{P(S)_{ij}}{P(S)_{i,r_i}} = e^{s_{ij}-s_{i,r_i}} \in F.$$

Thus $e^p \in F$ for every reference score difference $p = s_{ij} - s_{i,r_i}$. For a general $f \in \mathcal{D}(S)$, e^f is algebraic over F, which suffices for the Ax argument.

2. Bounding transcendence gate by gate. Restrict the z-variables to a short affine-line segment in the regular domain, chosen to preserve rational linear independence modulo constants for the finite family of functions used in the argument (Lemma 18). The rational input field is then $F_0 = \mathbb{C}(t)$. Fix a topological ordering of the circuit gates. For the prefix consisting of the first i gates, let F_i be the field generated over F_0 by the gate values seen so far, and let $\mathcal{W}_i$ be the $\mathbb{Q}$-linear span of the exponential inputs and logarithm outputs seen so far, together with $\mathcal{C}$. Thus $\mathcal{W}_i$ is a vector space, whereas F_i is a field; the following bound tracks them gate by gate.

Lemma 6 (Gate-prefix transcendence bound)**.** *For every gate prefix,*

$$\operatorname{trdeg}_{F_0} F_i \le \dim_{\mathbb{Q}}(\mathcal{W}_i/\mathcal{C}).$$

Hence, if $q = \dim_{\mathbb{Q}}(\mathcal{W}/\mathcal{C})$ for the full circuit field F, then

$$\operatorname{trdeg}_{\mathbb{C}} F \le 1 + q.$$

Proof. Arithmetic gates add no transcendence and do not enlarge $\mathcal{W}_i$. Consider an exponential gate $E = e^g$. If g is independent of $\mathcal{W}_{i-1}$ modulo constants, the span dimension increases by one and adjoining E increases the transcendence degree by at most one. Otherwise, after clearing rational denominators,

$$Mg = c + \sum_\nu \alpha_\nu g_\nu + \sum_\mu \beta_\mu L_\mu$$

for previous exponential inputs g_ν, logarithm outputs L_μ, and integers α_ν, β_μ. Thus

$$E^M = e^c \prod_\nu (e^{g_\nu})^{\alpha_\nu} \prod_\mu (e^{L_\mu})^{\beta_\mu} \in F_{i-1},$$

because each e^{g_ν} is a previous exponential output and each e^{L_μ} is the input of the corresponding logarithm gate. Hence E is algebraic over F_{i-1}. For a logarithm gate $L = \ln h$, if L is independent of $\mathcal{W}_{i-1}$ modulo constants, the span dimension increases by one and adjoining L increases the transcendence degree by at most one; otherwise $L \in F_{i-1}$ already. The induction starts from $F_0 = \mathbb{C}(t)$ and $\mathcal{W}_0 = \mathcal{C}$. □

The analytic input is the following one-variable consequence of Ax's differential-field theorem.

Theorem 7 (One-variable Ax consequence; Ax, 1971; see also Aslanyan, 2024)**.** *If $f_1, \dots, f_k$ are one-variable holomorphic germs that are $\mathbb{Q}$-linearly independent modulo constants and $k > 0$, then*

$$\operatorname{trdeg}_{\mathbb{C}} \mathbb{C}(f_1, \dots, f_k, e^{f_1}, \dots, e^{f_k}) \ge k + 1.$$

The details are given in Appendix B.

Proof of Theorem 5. Suppose a reference score difference $p = s_{ij} - s_{i,r_i}$ lies outside $\mathcal{W}$. Choose a basis $u_1, \dots, u_q$ of $\mathcal{W}/\mathcal{C}$ and restrict to an affine line preserving the independence of $u_1, \dots, u_q, p$ modulo constants. By the gate-prefix bound, $\operatorname{trdeg}_{\mathbb{C}} F \le q + 1$.

Each u_k lies in F, and e^{u_k} is algebraic over F: exponentiating a rational combination of exponential inputs and logarithm outputs produces rational powers of their corresponding gate outputs and inputs. Also $p \in \mathbb{C}(t) \subseteq F$ and, by value differentiation, $e^p \in F$. After adjoining the algebraic elements e^{u_k} inside the ambient meromorphic-germ field, Theorem 7 applied to $(u_1, \dots, u_q, p)$ gives transcendence degree at least $q + 2$, contradicting the upper bound $q + 1$. Hence every reference score difference lies in $\mathcal{W}$, and these differences generate $\mathcal{D}(S)/\mathcal{C}$. □

The span theorem gives more than a lower bound on the number of transcendental gates. It applies when a circuit nests exponentials and logarithms or shares their outputs across many paths. Softmax cannot distinguish scores that differ by a constant within a row, so the theorem concerns the score differences that determine its weights. This is also the form needed for the total-size bound: mixed coefficients of those differences retain the rectangular matrix product.

These linear relations can involve internal functions that are nonlinear in Q and K. Mixed coefficient extraction isolates their bilinear parts.

5 From the Span Theorem to the Total-Size Lower Bound

Choose a real point in the circuit domain and perturb the two matrix inputs independently, with $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^{m\times d}$:

$$Q = Q^\star + uX, \qquad K = K^\star + wY. \tag{2}$$

All division denominators remain nonzero and all logarithm inputs remain positive in a sufficiently small neighborhood of this point. We retain coefficients modulo (u^2, w^2) and write $[f]_{11} = [uw]f$ for the mixed $(1,1)$ coefficient, which is bilinear in X, Y. The low-degree simulation of Saha et al. (2026, Theorem 4.1) extracts mixed coefficients with constant overhead. Theorem 5 identifies internal functions whose extracted coefficients linearly generate the score-difference tensor.

Lemma 8 (Constant-overhead mixed-coefficient extraction)**.** *From a size-s eAC and a real base point regular for the circuit, one can construct a size-$O(s)$ division-free arithmetic circuit that jointly outputs the mixed $(1,1)$ coefficient of every gate under the perturbation in Eq. (2).*

To see why the overhead is constant, write the four retained coefficients of a gate value f as $f_{00}, f_{10}, f_{01}, f_{11}$. Arithmetic gates update these four quantities by the truncated product and quotient rules. The two transcendental gates satisfy

$$(e^g)_{11} = e^{g_{00}}(g_{11} + g_{10}g_{01}), \qquad (\ln g)_{11} = \frac{g_{11}}{g_{00}} - \frac{g_{10}g_{01}}{g_{00}^2}.$$

At the regular real base point, $e^{g_{00}}$, g_{00}^{-1}, and $\ln g_{00}$ are hardwired real constants. Thus every original scalar gate contributes only a constant number of arithmetic gates. Appendix D gives the full recurrence table and the division elimination.

The extraction circuit produces the mixed coefficients of all selected internal gates jointly, even if the original circuit shares intermediate values. Each rational relation from Theorem 5 has coefficients fixed by the given circuit; they do not depend on the perturbation matrices X and Y. We can therefore apply these relations as one fixed linear output map to the extracted coefficient vector. Tensor rank cannot increase under this map, so the proof does not need a separate circuit for every score difference.

For $j \in [m-1]$, define $p_{ij}(Q,K) = (QK^\top)_{i,j+1} - (QK^\top)_{i1}$. Its mixed coefficient is

$$[p_{ij}]_{11} = (XY^\top)_{i,j+1} - (XY^\top)_{i1},$$

independently of the base point. Hence the vector $([p_{ij}]_{11})_{i,j}$ is exactly $\mathbf{\Delta}(X,Y)$.

Tensor rank is monotone under fixed linear output maps, so the possibly dense projection from extracted coefficients to score differences yields the required rank comparison.

Lemma 9 (Output projection and tensor rank)**.** *Let $\Gamma(X,Y)$ be a vector-valued bilinear map computed by a size-t division-free arithmetic circuit. For every fixed real linear map Π,*

$$\mathrm{R}(\Pi\Gamma) \le \mathrm{R}(\Gamma) \le O(t).$$

The first inequality follows by applying Π to the constant output vector in each rank-one term of a bilinear decomposition; the second is the standard bilinearization bound.

Proof of Theorem 1. Theorem 5 gives fixed rational relations expressing every reference score difference in terms of the specialized exponential inputs g_ν and logarithm outputs L_μ, modulo constants. Apply Lemma 8 and write

$$G_\nu = [uw]g_\nu, \qquad J_\mu = [uw]L_\mu, \qquad \Gamma = (G_1, \dots, G_{r_{\exp}}, J_1, \dots, J_{r_{\log}}).$$

By Proposition 4, the mixed score-difference map is linearly equivalent to $\langle n, d, m-1\rangle$. Taking mixed coefficients in the span relations expresses this map as a fixed rational linear projection of Γ. Lemma 9 therefore gives

$$\mathrm{R}(\langle n, d, m-1\rangle) \le \mathrm{R}(\Gamma) \le O(s).$$

This proves Theorem 1. We next derive the polynomial-width consequence. □

The missing reference column does not weaken the exponent. Set $M = \lfloor N/2 \rfloor$ and $d_N = \lceil N^a \rceil$. Restricting matrix multiplication to smaller dimensions gives

$$\mathrm{R}(\langle M, \lceil M^a \rceil, M\rangle) \le \mathrm{R}(\langle N, d_N, N-1\rangle) \le \mathrm{R}(\langle N, d_N, N\rangle).$$

Since $M = \Theta(N)$, Proposition 25 shows that both outer ranks have exponent $\omega(1, a, 1)$. For the matching upper bound, standard recursive bilinear algorithms evaluate the rectangular product at this exponent. Exponentiation, row normalization, and value aggregation add $O(N^2)$ gates, which do not raise the exponent because $\omega(1, a, 1) \ge 2$.

6 Exact Exp/Log Gate Complexity

We now count only exponential and logarithm gates. Let $\mathsf{TC}_U(S)$ be the minimum number of these gates in a U-admissible eAC computing $\mathrm{softmax}(S)v$. Then the score-difference dimension equals the minimum exp/log gate count.

Theorem 10 (Local exact softmax transcendental-gate complexity)**.** *For every rational score family S and every nonempty connected open set $U \subseteq \mathrm{Dom}_{\mathbb{R}}(S) \times \mathbb{R}^m$,*

$$\mathsf{TC}_U(S) = \tau(S).$$

For the lower bound, a circuit with $r_{\exp}$ exponential gates and $r_{\log}$ logarithm gates satisfies $\tau(S) \le \dim_{\mathbb{Q}}(\mathcal{W}/\mathcal{C}) \le r_{\exp} + r_{\log}$ by Theorem 5. For the upper bound, set $q = \tau(S)$ and choose functions $h_1, \dots, h_q$ whose classes form a $\mathbb{Q}$-basis of $\mathcal{D}(S)/\mathcal{C}$. Nonuniformity allows the circuit to hardwire this basis and its rational relations. If M clears the denominators in those relations, one exponential gate for each h_k/M gives $z_k = e^{h_k/M}$. Every reference score difference then has the form

$$e^{s_{ij} - s_{i,r_i}} = e^{c_{ij}} \prod_{k=1}^{q} z_k^{p_{ijk}}, \qquad c_{ij} \in \mathbb{R}, \quad p_{ijk} \in \mathbb{Z}.$$

The circuit forms these weights with integer powers and arithmetic gates, then normalizes each row relative to the reference weight 1. It needs exactly q exponential gates and no logarithm gates. This construction may use many arithmetic gates: Theorem 10 optimizes only the exp/log count, while the total-size lower bound accounts for arithmetic work as well. Appendix C gives the domain and construction details.

Corollary 11 (Independent-query/key dense attention)**.** *For dense dot-product attention with independently parameterized Q and K, and every nonempty connected open set $U \subseteq \mathbb{R}^{nd+dm+m}$,*

$$\mathsf{TC}_U(QK^\top) = n(m-1).$$

For $n = m = N$, Proposition 14 gives the same exact count $N(N-1)$ for the fixed-projection self-attention restriction.

To attain this bound, fix the first key as the reference and define $K^{\Delta}_{j,:} = K_{j+1,:} - K_{1,:}$ for $j \in [m-1]$. Then $Q(K^{\Delta})^\top$ consists exactly of the within-row score differences $(QK^\top)_{i,j+1} - (QK^\top)_{i1}$. Since softmax is invariant under subtracting the reference score $(QK^\top)_{i1}$ from row i, the corresponding unnormalized weights are 1 for the reference key and $\exp((Q(K^{\Delta})^\top)_{ij})$ for $j \in [m-1]$. Normalizing these weights and aggregating the values therefore computes the original attention output. The construction uses exactly $n(m-1)$ exponential gates and no logarithm gates. Forming K^{Δ} costs $O(dm)$ arithmetic gates, computing $Q(K^{\Delta})^\top$ uses an arithmetic circuit for an $n \times d$ by $d \times (m-1)$ product, and normalization and value aggregation cost $O(nm)$ additional gates. Hence the total size is the arithmetic-circuit size of this rectangular product plus $O(dm + nm)$.

Theorem 10 shows that exact exp/log complexity is governed by $\mathbb{Q}$-linear relations among score differences, rather than simply by the number of underlying real parameters.

Example 12 (Dependence on rational relations)**.** For one query and three keys,

scores	$\tau(S)$	transcendental generators
(x, x, x)	0	none
$(0, x, 2x)$	1	e^x
$(0, x, \sqrt{2}x)$	2	$e^x,\ e^{\sqrt{2}x}$

The last two score families both depend on a single real parameter, but x and $2x$ are $\mathbb{Q}$-linearly dependent whereas x and $\sqrt{2}x$ are $\mathbb{Q}$-linearly independent. Hence their optimal exp/log gate counts are 1 and 2, respectively.

7 Extensions and the Scope of the Lower Bound

Tied self-attention. We next consider the tied setting $K = Q$, where the score matrix is $QQ^\top$. Although the independent Q/K perturbation used above is no longer available, splitting the rows of Q into two blocks allows the two blocks to be perturbed independently. The mixed coefficients of cross-block score differences then form a rectangular matrix product.

Corollary 13 (Total size for tied queries and keys)**.** *Let $N = r + t$ with $r \geq 1$, $t \geq 2$, and $d \geq 1$, and let $U \subseteq \mathbb{R}^{Nd+N}$ be nonempty, connected, and open. Every U-admissible size-s eAC computing $\mathrm{softmax}(QQ^\top)v$, for $Q \in \mathbb{R}^{N\times d}$ and $v \in \mathbb{R}^N$, satisfies*

$$s = \Omega(\mathrm{R}(\langle r, d, t-1\rangle)).$$

Consequently, for fixed $0 < a \leq 1$ and $d = \lceil N^a \rceil$, this tied map also has minimum size $N^{\omega(1,a,1)+o(1)}$.

Here both Q and v are variable inputs.

Proof sketch. We partition the rows of Q into two blocks, $Q^{(1)} \in \mathbb{R}^{r\times d}$ and $Q^{(2)} \in \mathbb{R}^{t\times d}$, and perturb the two blocks independently. We take score differences across columns in the second block. Their mixed coefficients form an $r \times d$ by $d \times (t-1)$ matrix product, while within-block scores contribute no mixed coefficient. The span theorem and the output-projection argument from Theorem 1 then give Corollary 13. Details are in Appendix E.3.

Scope and open directions. Our exact-real results are sensitive to rational linear relations among the score differences. For example, for scores $(0, x, \alpha x)$, the minimum exp/log gate count changes depending on whether α is rational or irrational. This sensitivity reflects the exact-real setting, and our results do not resolve the complexity of approximate or finite-precision attention. For disjoint variable blocks, the exp/log gate count is additive even if a single circuit shares gates across the instances. When different heads or layers share input variables, the score-difference spaces no longer separate in this way. Comparable total-size lower bounds for such shared-input multi-head or multi-layer architectures remain open.

8 Conclusion

Our results show that restricting each query to a single scalar output does not reduce the exact asymptotic cost of dense attention in the eAC model. Although each query returns only one scalar, differentiating the attention map with respect to the value inputs gives all of its softmax weights. Ratios of these weights then determine the within-row score differences. For dot-product attention, the mixed coefficients of these score differences form a rectangular matrix-multiplication tensor. The relevant complexity therefore comes from the score differences that determine the softmax weights, rather than from the number of output coordinates.

The exp–log span theorem describes how these score differences must appear inside any exact circuit. It shows that, modulo constants, every within-row score difference lies in the $\mathbb{Q}$-linear span of the circuit's exponential inputs and logarithm outputs. We use this statement in two ways. First, we apply mixed coefficient extraction to these span relations to obtain the rectangular total-size lower bound. Second, the dimension of the same score-difference space gives the exact exp/log gate count. Both results rely on the same fact: an exact circuit must represent the score differences that determine the softmax weights.

Reproducibility Statement

This is a theoretical work with no empirical data. Section 2 specifies the exact-real circuit model. Appendices B–E give the complete span argument, transcendental-gate construction, coefficient recurrences, and exponent accounting. Appendices F–I give the additional identities, reductions, and domain conditions. The LaTeX source reproduces all statements and proofs.

AI Use Statement

GPT-6 Sol was used as a research and writing aid. In particular, it suggested the mixed-Hessian construction in Appendix F, which introduces auxiliary variables and a zero-score anchor so that mixed second derivatives encode exponentials of the query–key scores. It also assisted with the design of Figure 1 and with polishing the exposition, organization, terminology, and notation of the manuscript. The complete set of proofs was additionally checked using Astra. The authors independently verified every proof and take full responsibility for all mathematical claims, citations, and the final manuscript.

References

Josh Alman and Zhao Song. Fast attention requires bounded entries. In *Advances in Neural Information Processing Systems*, volume 36, pp. 63117–63135, 2023.

Josh Alman and Hantao Yu. Fundamental limitations on subquadratic alternatives to transformers. In *Proceedings of the 13th International Conference on Learning Representations*, 2025.

Vahagn Aslanyan. The existential closedness and Zilber–Pink conjectures. *Model Theory*, 3(2): 599–624, 2024. doi: 10.2140/mt.2024.3.599.

James Ax. On schanuel's conjectures. *Annals of Mathematics*, 93(2):252–268, 1971. doi: 10.2307/1970774.

Walter Baur and Volker Strassen. The complexity of partial derivatives. *Theoretical Computer Science*, 22:317–330, 1983. doi: 10.1016/0304-3975(83)90110-X.

Markus Bläser. Fast matrix multiplication. *Theory of Computing Library, Graduate Surveys*, 5:1–60, 2013. doi: 10.4086/toc.gs.2013.005.

Peter Bürgisser, Michael Clausen, and M. Amin Shokrollahi. *Algebraic Complexity Theory*, volume 315 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1997.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *Proceedings of the 9th International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 597–619. PMLR, 2023.

H. I. Epstein and B. F. Caviness. A structure theorem for the elementary functions and its application to the identity problem. *International Journal of Computer & Information Sciences*, 8:9–37, 1979. doi: 10.1007/BF00995426.

Michael A. Forbes, Mrinal Kumar, and Ramprasad Saptharishi. Functional lower bounds for arithmetic circuits and connections to boolean circuit complexity. In *31st Conference on Computational Complexity (CCC 2016)*, volume 50 of *Leibniz International Proceedings in Informatics*, pp. 33:1–33:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2016. doi: 10.4230/LIPIcs.CCC.2016.33.

Dima Grigoriev and Nicolai N. Vorobjov. Complexity lower bounds for computation trees with elementary transcendental function gates. *Theoretical Computer Science*, 157(2):185–214, 1996. doi: 10.1016/0304-3975(95)00159-X.

Shreya Gupta, Boyang Huang, Barna Saha, Yinzhan Xu, and Christopher Ye. Subquadratic algorithms and hardness for attention with any temperature. In *Proceedings of the 14th International Conference on Learning Representations*, 2026.

Ellis R. Kolchin. *Differential Algebra and Algebraic Groups*, volume 54 of *Pure and Applied Mathematics*. Academic Press, 1973.

Jiatu Li and Mengdi Wu. Identity testing for circuits with exponentiation gates. In *17th Innovations in Theoretical Computer Science Conference (ITCS 2026)*, volume 362 of *Leibniz International Proceedings in Informatics*, pp. 95:1–95:22. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi: 10.4230/LIPIcs.ITCS.2026.95.

Daniel Richardson. Some undecidable problems involving elementary functions of a real variable. *The Journal of Symbolic Logic*, 33(4):514–520, 1968. doi: 10.2307/2271358.

Robert H. Risch. The problem of integration in finite terms. *Transactions of the American Mathematical Society*, 139:167–189, 1969. doi: 10.1090/S0002-9947-1969-0237477-8.

Robert H. Risch. Algebraic properties of the elementary functions of analysis. *American Journal of Mathematics*, 101(4):743–759, 1979. doi: 10.2307/2373917.

Joseph Fels Ritt. Algebraic combinations of exponentials. *Transactions of the American Mathematical Society*, 31(4):654–679, 1929. doi: 10.1090/S0002-9947-1929-1501505-4.

Maxwell Rosenlicht. Integration in finite terms. *The American Mathematical Monthly*, 79(9):963–972, 1972. doi: 10.1080/00029890.1972.11993166.

Maxwell Rosenlicht. Differential extension fields of exponential type. *Pacific Journal of Mathematics*, 57(1):289–300, 1975. doi: 10.2140/pjm.1975.57.289.

Michael Rothstein and B. F. Caviness. A structure theorem for exponential and primitive functions. *SIAM Journal on Computing*, 8(3):357–367, 1979. doi: 10.1137/0208028.

Barna Saha and Christopher Ye. I/O complexity of attention, or how optimal is FlashAttention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43024–43042. PMLR, 2024. URL https://proceedings.mlr.press/v235/saha24a.html.

Barna Saha, Yinzhan Xu, Christopher Ye, and Hantao Yu. On the computational hardness of transformers. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*, pp. 1073–1084. ACM, 2026. doi: 10.1145/3798129.3800819. Full version: arXiv:2603.11332.

Arnold Schönhage. Partial and total matrix multiplication. *SIAM Journal on Computing*, 10(3): 434–455, 1981. doi: 10.1137/0210032.

Volker Strassen. Vermeidung von Divisionen. *Journal für die reine und angewandte Mathematik*, 264:184–202, 1973. doi: 10.1515/crll.1973.264.184.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. URL https://arxiv.org/abs/2006.04768.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297, 2020. URL https://papers.nips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

A Model Details and Tensor Preliminaries

A.1 Embedding one-channel attention into standard self-attention

Let $Q \in \mathbb{R}^{n\times d}$ and $K \in \mathbb{R}^{m\times d}$. Fix a value width $d_v \geq 1$ and let $e_1 \in \mathbb{R}^{d_v}$ be the first standard basis vector. For any $v \in \mathbb{R}^m$, set $V = ve_1^\top \in \mathbb{R}^{m\times d_v}$. Since $P(QK^\top) = \mathrm{softmax}(QK^\top)$ row-wise, the first output coordinate of the standard head is

$$P(QK^\top)v = \mathcal{A}_{n,d,m}(Q, K, v),$$

and every other coordinate is zero. Hence every lower bound for one value channel applies to a full exact-attention head. For a circuit defined only locally, fix the unused value columns at constants chosen so that the full value input remains in the same local domain; the first output coordinate remains $P(QK^\top)v$ on a smaller open neighborhood.

Proposition 14 (Realization as self-attention with fixed projections)**.** *Let $n = m = N$, take value width $d_v = d$, and define tokens*

$$X_i = (Q_{i,:}, K_{i,:}, v_i) \in \mathbb{R}^{2d+1}.$$

With fixed projection matrices

$$W_Q = \begin{bmatrix} I_d \\ 0_{d\times d} \\ 0_{1\times d} \end{bmatrix}, \qquad W_K = \begin{bmatrix} 0_{d\times d} \\ I_d \\ 0_{1\times d} \end{bmatrix}, \qquad W_V = \begin{bmatrix} 0_{d\times d} \\ 0_{d\times d} \\ e_1^\top \end{bmatrix},$$

standard self-attention satisfies $XW_Q = Q$, $XW_K = K$, and $XW_V = ve_1^\top$. Its first output coordinate is therefore $\mathcal{A}_{N,d,N}(Q, K, v)$ and all remaining output coordinates vanish. The input-width overhead is a constant factor. The fixed projections are coordinate selectors, so the embedding introduces no matrix-multiplication preprocessing. If the convention uses scaled scores $QK^\top/\sqrt{d}$, absorb the fixed factor $\sqrt{d}$ into W_Q (or split it between W_Q and W_K); the realized score remains $QK^\top$ with unchanged width and circuit complexity.

A.2 Local domains and complexification

For a rational score family S, the local complexity $\mathsf{TC}_U(S)$ from the main text is defined on any nonempty connected open set $U \subseteq \mathrm{Dom}_{\mathbb{R}}(S) \times \mathbb{R}^m$. An admissible circuit must be regular and equal to the target throughout U: every division denominator is nonzero and every real-logarithm input is positive. Any circuit that computes the target globally is a special case, so a lower bound in this local model applies a fortiori to globally correct circuits. Fix $(z^\star, v^\star) \in U$. Openness gives product neighborhoods $U_z \ni z^\star$ and $U_v \ni v^\star$ with $U_z \times U_v \subseteq U$. All subsequent value specializations are taken at the chosen $v^\star \in U_v$.

After shrinking the product neighborhood, all real-analytic gate functions extend holomorphically to a common complex neighborhood. For each logarithm gate, its nonvanishing input admits a unique local holomorphic branch agreeing with the real logarithm on the real slice. We fix these branches once and keep them throughout the proof. For a fixed nonvanishing input, two local logarithm branches differ by an element of $2\pi i\mathbb{Z} \subset \mathbb{C}$. All subsequent gate evaluations use the fixed branches chosen above, and all field and transcendence arguments take place in the corresponding field of meromorphic germs, that is, quotients of holomorphic functions identified when they agree near the base point. Equality on the original real open set implies equality of the complexified germs by the identity theorem.

If E is a subfield of the ambient one-variable meromorphic-germ field $\mathcal{M}$, then E^{alg} consists of the elements of $\mathcal{M}$ algebraic over E; this is the relative algebraic closure of E in $\mathcal{M}$. The characteristic-zero matrix-multiplication exponent is unchanged over $\mathbb{C}$; all circuit and tensor-rank bounds stated here are over $\mathbb{R}$.

The coefficient extraction stays over the reals: at the chosen real base point, the exponential values, logarithm values, and inverses in the coefficient recurrences are real constants. Tensor rank R is therefore taken over $\mathbb{R}$. Complexification is confined to functional transcendence; coefficient extraction and tensor-rank bounds remain over $\mathbb{R}$.

A.3 Score-difference space and tensor rank

We prove Proposition 4. Write the rows of K as $K_{1,:}, \dots, K_{m,:}$ and apply the invertible linear change of input coordinates, together with its inverse and the resulting score differences:

$$\begin{aligned} k_{\mathrm{ref}} &= K_{1,:}, & K^{\Delta}_{j,:} &= K_{j+1,:} - K_{1,:}, \\ K_{1,:} &= k_{\mathrm{ref}}, & K_{j+1,:} &= k_{\mathrm{ref}} + K^{\Delta}_{j,:}, \\ \mathbf{\Delta}_{ij}(Q,K) &= (QK^\top)_{i,j+1} - (QK^\top)_{i1} = (Q(K^{\Delta})^\top)_{ij} & (j &\in [m-1]). \end{aligned}$$

The variable k_{ref} is unused. Thus $\mathbf{\Delta}$ is obtained from $\langle n, d, m-1\rangle$ by an invertible input transformation and the addition of an unused input block. Restricting $k_{\mathrm{ref}} = 0$ gives the reverse reduction. Tensor rank is monotone under linear restrictions and invariant under invertible linear changes, so

$$\mathrm{R}(\mathbf{\Delta}) = \mathrm{R}(\langle n, d, m-1\rangle).$$

To prove the claimed linear independence, suppose

$$\sum_{i=1}^{n}\sum_{j=1}^{m-1} \lambda_{ij}(Q(K^{\Delta})^\top)_{ij} = 0.$$

The coefficient of the monomial $Q_{i\ell}K^{\Delta}_{j\ell}$ is λ_{ij} for every $\ell \in [d]$. Hence all λ_{ij} vanish. The reference score differences are therefore linearly independent over $\mathbb{Q}$, even before quotienting by constants.

The value $\tau(S)$ in Definition 3 does not depend on the reference choices: for two columns r_i, r_i' in row i,

$$s_{ij} - s_{i,r_i'} = (s_{ij} - s_{i,r_i}) - (s_{i,r_i'} - s_{i,r_i}),$$

so both choices span the same vector space modulo constants. It is also independent of the chosen connected open domain: a rational linear combination that is constant on a nonempty open set is the same constant as a rational function, by the identity theorem after local complexification.

Example 15 (2×3 score family)**.** With the first column as reference, the score-difference space is generated by

$$s_{12} - s_{11}, \quad s_{13} - s_{11}, \quad s_{22} - s_{21}, \quad s_{23} - s_{21}.$$

For independent score variables, $\tau(S) = 4$. For $S = QK^\top$, these four functions are precisely the entries of $Q(K^{\Delta})^\top$ with $n = 2$ and $m = 3$.

A.4 Bilinear circuits and tensor rank

We use the standard real bilinearization principle: if a division-free arithmetic circuit of size s computes a vector-valued bilinear map $T(x, y)$ over $\mathbb{R}$, then $\mathrm{R}(T) = O(s)$. One proof replaces every gate by its bidegree-$(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$ components. In the resulting circuit, every bidegree-$(1,1)$ contribution is a sum of products of a linear form in x and a linear form in y. Charging a constant number of such products to each multiplication gate gives a rank-$O(s)$ decomposition. We invoke this principle after the coefficient-extraction construction of Appendix D; all remaining divisions there are by hardwired nonzero constants and can be eliminated.

B Proof of the Exp–Log Span Theorem

We prove Theorem 5. If the score family has no input variables, every score difference is constant and $\mathcal{D}(S) = \mathcal{C} \subseteq \mathcal{W}$. We may therefore assume that $z \in \mathbb{R}^p$ with $p \geq 1$. Choose $(z^\star, v^\star)$ and a product neighborhood as in Appendix A.2, specialize $v = v^\star$, and fix a connected open ball in the z-variables on which all circuit gates are regular. After restriction to an affine line, the rational input field is $F_0 = \mathbb{C}(t)$.

B.1 Gate fields and recovering softmax weights

Enumerate the scalar gates in topological order. Every input of a transcendental gate is an input node or the value of an earlier gate.

Lemma 16 (Gate-field generation and differential closure)**.** *Let F be the field over F_0 generated by all specialized gate values after the affine-line restriction. Then every gate value, every exponential-gate input, and every logarithm-gate input and output lies in F. Moreover F is stable under $D = d/dt$, and its constant field is $\mathbb{C}$.*

Proof. Arithmetic gates are rational operations on earlier values. Exponential and logarithm outputs are adjoined gate values, while their inputs are earlier values. For differential closure, use

$$D(e^g) = e^g Dg, \qquad D(\ln h) = Dh/h,$$

together with the arithmetic chain rules and $D(F_0) \subseteq F_0$. Since F is a subfield of one-variable meromorphic germs, $Df = 0$ implies $f \in \mathbb{C}$. □

Lemma 17 (Derivative-field closure)**.** *Before the affine-line restriction, let F_z be the field generated over $\mathbb{C}(z)$ by all gate values after specializing $v = v^\star$. For every gate u and value coordinate v_j,*

$$\left.\frac{\partial u}{\partial v_j}\right|_{v=v^\star} \in F_z.$$

Consequently every softmax coefficient $P(S)_{ij}$ lies in F_z, and

$$e^{s_{ij}-s_{i,r_i}} = P(S)_{ij}/P(S)_{i,r_i} \in F_z.$$

After the affine-line restriction, the corresponding elements $P(S)_{ij}$ and $e^{s_{ij}-s_{i,r_i}}$ lie in F.

Proof. Induct over gates. The formulas for arithmetic gates are the usual product, quotient, and sum rules; for transcendental gates use

$$\partial_{v_j} e^g = e^g \partial_{v_j} g, \qquad \partial_{v_j} \ln h = (\partial_{v_j} h)/h.$$

All required denominators are nonzero because the chosen point lies in the regular domain. The target is exactly linear in v, so differentiating its ith output with respect to v_j gives $P(S)_{ij}$ at the chosen $v^\star$. Softmax coefficients are positive, so their row ratios are defined. After restriction to the affine line, the same derivative identities hold in F; the relevant denominators remain nonzero on the chosen regular domain. □

For a rational linear combination $f = c + \sum_{i,j} q_{ij}(s_{ij} - s_{i,r_i})$, choose an integer $M \geq 1$ with $Mq_{ij} \in \mathbb{Z}$. Then

$$(e^f)^M = e^{Mc} \prod_{i,j} \left(e^{s_{ij}-s_{i,r_i}}\right)^{Mq_{ij}} \in F_z.$$

Hence e^f is algebraic over F_z. For each reference score difference p, the derivative-field lemma gives the stronger statement $e^p \in F_z$ used in the Ax argument.

B.2 Preserving $\mathbb{Q}$-linear independence on an affine line

Lemma 18 (Affine line preserving relations modulo constants)**.** *Let $p \geq 1$ and let $U_0 \subset \mathbb{R}^p$ be a connected open ball contained in the regular real domain fixed in Appendix A.2. Let $f_1, \ldots, f_q$ be real-analytic representatives on U_0 of meromorphic germs that are $\mathbb{Q}$-linearly independent modulo constants. Then there exist $z_0 \in U_0$, a nonzero direction $\xi \in \mathbb{R}^p$, and $\varepsilon > 0$ such that $z_0 + t\xi \in U_0$ for $|t| < \varepsilon$ and the restricted one-variable germs $f_i(z_0 + t\xi)$ remain $\mathbb{Q}$-linearly independent modulo $\mathcal{C}$.*

Proof. For $q = 0$, any sufficiently short segment with nonzero direction suffices. Assume $q > 0$. For each nonzero $c = (c_1, \ldots, c_q) \in \mathbb{Q}^q$, set

$$f_c(z) = \sum_{i=1}^{q} c_i f_i(z) \quad \text{and} \quad Z_c = \{(z, \xi) \in U_0 \times \mathbb{R}^p : \xi \cdot \nabla f_c(z) = 0\}.$$

The independence hypothesis implies that f_c is nonconstant on U_0. Since U_0 is connected and f_c is real analytic, ∇f_c is not identically zero. Hence

$$G_c(z, \xi) = \xi \cdot \nabla f_c(z)$$

is a nonzero real-analytic function on the connected open set $U_0 \times \mathbb{R}^p$: choose z with $\nabla f_c(z) \neq 0$ and then $\xi = \nabla f_c(z)$. Therefore its zero set Z_c is closed and has empty interior. The open Euclidean set $U_0 \times \mathbb{R}^p$ is a Baire space, so the countable union

$$\bigcup_{c \in \mathbb{Q}^q \setminus \{0\}} Z_c$$

does not cover it. Choose (z_0, ξ) outside this union. Necessarily $\xi \neq 0$. After decreasing $\varepsilon > 0$, the segment $z_0 + t\xi$ remains in U_0 for $|t| < \varepsilon$. For every nonzero $c \in \mathbb{Q}^q$,

$$\frac{d}{dt} f_c(z_0 + t\xi)\Big|_{t=0} = \xi \cdot \nabla f_c(z_0) \neq 0,$$

so the restricted f_c is not constant. Thus no nontrivial rational linear combination of the restricted f_i is constant. Because the line stays inside the preselected regular ball, all denominators remain nonzero and all fixed logarithm branches remain valid on a sufficiently small complex neighborhood of $t = 0$. □

B.3 The gate-prefix bound and algebraicity of basis exponentials

Apply the gate-prefix bound of Lemma 6 with $F_0 = \mathbb{C}(t)$ and the line-restricted gate fields F_i. We obtain $\operatorname{trdeg}_{\mathbb{C}} F \leq 1 + q$, where $q = \dim_{\mathbb{Q}}(\mathcal{W}/\mathcal{C})$.

Lemma 19 (Exponentials of a basis of $\mathcal{W}/\mathcal{C}$)**.** *Let $u_1, \ldots, u_q$ be a $\mathbb{Q}$-basis of $\mathcal{W}/\mathcal{C}$. For every $k \in [q]$, e^{u_k} is algebraic over F.*

Proof. Fix $k \in [q]$. The function u_k is a rational linear combination of actual exponential inputs and logarithm outputs modulo a constant. Clearing the rational denominators, write

$$M_k u_k = c_k + \sum_{\nu} r_{k\nu} g_\nu + \sum_{\mu} q_{k\mu} L_\mu$$

with $M_k \geq 1$ and integer coefficients. Then

$$(e^{u_k})^{M_k} - e^{c_k} \prod_{\nu} (e^{g_\nu})^{r_{k\nu}} \prod_{\mu} (e^{L_\mu})^{q_{k\mu}} = 0,$$

where negative powers can be cleared to obtain an ordinary polynomial equation. Exponential-gate outputs and logarithm-gate inputs all lie in F, so e^{u_k} is algebraic over F. □

B.4 Ax's theorem and the multivariate conclusion

We use the following single-derivation form of Ax (1971, Theorem 3); Aslanyan (2024, Theorem 2.10) states the corresponding differential-field theorem. Let (E, D) be a differential field of characteristic zero with constant field $C = \ker D$. Suppose $x_1, \ldots, x_k \in E$ and $y_1, \ldots, y_k \in E^\times$ satisfy $Dy_i = y_i Dx_i$. If the x_i are $\mathbb{Q}$-linearly independent modulo C and $k > 0$, then

$$\operatorname{trdeg}_C C(x_1, \ldots, x_k, y_1, \ldots, y_k) \geq k + 1.$$

The final 1 is the rank of the nonzero derivative row $(Dx_1, \ldots, Dx_k)$. For holomorphic germs, take $D = d/dt$ and $y_i = e^{x_i}$ in the ambient meromorphic-germ field. The next two lemmas verify the differential-field and constant-field hypotheses for our application.

Lemma 20 (Relative algebraic closure is differential)**.** *Let (E, D) be a differential subfield of a characteristic-zero differential field $\mathcal{M}$. The relative algebraic closure*

$$E_{\mathcal{M}}^{\mathrm{alg}} = \{\alpha \in \mathcal{M} : \alpha \text{ is algebraic over } E\}$$

is stable under D. It is algebraic over E and therefore has the same transcendence degree. If the constant field of $\mathcal{M}$ is $\mathbb{C}$ and $\mathbb{C} \subseteq E$, then the constant field of $E_{\mathcal{M}}^{\mathrm{alg}}$ is also $\mathbb{C}$.

Proof. Let $\alpha \in E_{\mathcal{M}}^{\mathrm{alg}}$ and let $q(X) = \sum_{j=0}^{r} a_j X^j \in E[X]$ be its monic minimal polynomial. Differentiating $q(\alpha) = 0$ gives

$$0 = q^D(\alpha) + q'(\alpha) D\alpha, \qquad q^D(X) = \sum_{j=0}^{r} (Da_j) X^j.$$

Since the ground field has characteristic zero, q is separable. Hence $q'(\alpha) \neq 0$ and

$$D\alpha = -\frac{q^D(\alpha)}{q'(\alpha)} \in E(\alpha) \subseteq E_{\mathcal{M}}^{\mathrm{alg}}.$$

Thus the relative algebraic closure is a differential field. Algebraic extensions preserve transcendence degree. Finally, a D-constant in this relative closure is a D-constant of the ambient field $\mathcal{M}$, hence belongs to $\mathbb{C}$; conversely $\mathbb{C} \subseteq E$. This is the standard characteristic-zero algebraic-extension argument in differential algebra (Kolchin, 1973). □

Lemma 21 (Ax lower bound)**.** *Let $u_1, \ldots, u_q$ be a $\mathbb{Q}$-basis of $\mathcal{W}/\mathcal{C}$, and let $p_1, \ldots, p_k$ be reference score differences whose classes are $\mathbb{Q}$-linearly independent modulo $\mathcal{W}$. Assume $q + k > 0$. Suppose the functions $u_1, \ldots, u_q, p_1, \ldots, p_k$ are restricted to an affine line on which they remain $\mathbb{Q}$-linearly independent modulo constants. Writing $u = (u_1, \ldots, u_q)$ and $p = (p_1, \ldots, p_k)$, we have*

$$\operatorname{trdeg}_{\mathbb{C}} \mathbb{C}(u, p, e^u, e^p) \geq 1 + q + k.$$

Proof. By Lemma 18, such an affine-line restriction exists. By Lemma 19, the germs e^{u_a} are algebraic over F. Adjoin them inside the ambient one-variable meromorphic-germ field and pass to the relative algebraic closure. Lemma 20 shows that the resulting field is differential, has the same transcendence degree, and has constant field $\mathbb{C}$. It contains u, p, e^u, and e^p. Here $F \supseteq \mathbb{C}(t)$ satisfies the constant-field hypothesis of Lemma 20, and $e^{p_j} \in F$ follows from Lemma 17 because each p_j is a reference score difference. The tuple (u, p) is $\mathbb{Q}$-linearly independent modulo constants. Since it is nonempty and the ambient derivation is $D = d/dt$, the row vector (Du, Dp) is nonzero and has rank one. Theorem 7 gives $\operatorname{trdeg}_{\mathbb{C}} \mathbb{C}(u, p, e^u, e^p) \geq 1 + q + k$. □

Proof of Theorem 5. Suppose, for a contradiction, that the multivariate quotient $(\mathcal{D}(S) + \mathcal{W})/\mathcal{W}$ has positive dimension. Since the reference score differences generate $\mathcal{D}(S)/\mathcal{C}$, at least one of them lies outside $\mathcal{W}$. Choose such a reference score difference p, take a basis of $\mathcal{W}/\mathcal{C}$, and apply Lemma 18 to this finite family. The restriction preserves the independence of the basis and of p modulo that basis. Lemmas 6 and 19 give an algebraic differential extension of F with transcendence degree at most $1 + q$. Lemma 21, with $k = 1$, gives at least $q + 2$, a contradiction. Thus no reference score difference lies outside $\mathcal{W}$, and $\mathcal{D}(S) \subseteq \mathcal{W}$. The line was selected under the assumption that $\mathcal{D}(S) \not\subseteq \mathcal{W}$, so the contradiction proves the multivariate inclusion. □

C Exact Exp/Log Gate Complexity

C.1 Proof of Theorem 10

The lower bound is immediate from Theorem 5. If the circuit uses $r_{\exp}$ exponentials and $r_{\log}$ logarithms, then the span $\mathcal{W}$ has at most $r_{\exp} + r_{\log}$ generators modulo constants, and therefore

$$\tau(S) = \dim_{\mathbb{Q}}(\mathcal{D}(S)/\mathcal{C}) \leq \dim_{\mathbb{Q}}(\mathcal{W}/\mathcal{C}) \leq r_{\exp} + r_{\log}.$$

To construct a matching circuit, use nonuniformity to hardwire a basis and all rational relations for the fixed rational score family. Set $q = \tau(S)$ and choose rational score-difference functions $h_1, \ldots, h_q$ whose cosets form a $\mathbb{Q}$-basis of $\mathcal{D}(S)/\mathcal{C}$. Every reference score difference can be written

$$s_{ij} - s_{i,r_i} = c_{ij} + \sum_{k=1}^{q} q_{ijk} h_k, \qquad c_{ij} \in \mathbb{R}, \quad q_{ijk} \in \mathbb{Q}.$$

The constant c_{ij} is real because the score differences and basis functions are real-valued on the chosen real domain, although the lower-bound argument works modulo $\mathcal{C}$. Let M be a positive

integer that clears the rational-number denominators of the finitely many coefficients q_{ijk}, and write $q_{ijk} = p_{ijk}/M$ with $p_{ijk} \in \mathbb{Z}$. Compute $z_k = \exp(h_k/M)$ using one exponential gate for each k. Set $w_{i,r_i} = 1$ and, for $j \neq r_i$,

$$w_{ij} = e^{c_{ij}} \prod_{k=1}^{q} z_k^{p_{ijk}}.$$

These weights equal $\exp(s_{ij} - s_{i,r_i})$; negative integer powers are implemented by division, and all constant factors are hardwired. Compute

$$(\mathrm{softmax}(S)v)_i = \frac{v_{r_i} + \sum_{j \neq r_i} w_{ij} v_j}{1 + \sum_{j \neq r_i} w_{ij}}.$$

The construction uses exactly q exponential gates and zero logarithm gates. Integer powers are computed by arithmetic gates, for example by repeated squaring. Theorem 10 optimizes the exp/log gate count. The construction is regular on every connected component of the real domain of definition, and the lower bound applies to every admissible local domain. Hence $\mathsf{TC}_U(S) = \tau(S)$ for every U in Theorem 10.

C.2 Dense dot products and tied self-attention

For $S = QK^\top$, Proposition 4 shows that its reference score differences are exactly the entries of a formal $n \times d$ by $d \times (m-1)$ product after an invertible input transformation. They are linearly independent, so

$$\tau(QK^\top) = n(m-1).$$

This proves Corollary 11 on every admissible local domain. For the dense construction in Section 6, forming K^Δ costs $O(dm)$ gates, computing $Q(K^\Delta)^\top$ uses an arithmetic circuit for an $n \times d$ by $d \times (m-1)$ product, and exponentiation, row normalization, and aggregation against v cost $O(nm)$ additional gates. Exactly $n(m-1)$ exponential gates and no logarithm gates are used.

Corollary 22 (Tied query/key score-difference dimension)**.** *Let $S(Q) = QQ^\top$ with $Q \in \mathbb{R}^{N \times d}$. For every $N \geq 1$ and $d \geq 1$,*

$$\tau(QQ^\top) = \frac{N(N+1)}{2} - 1.$$

Proof. Consider tied queries and keys $S(Q) = QQ^\top$ with $Q \in \mathbb{R}^{N \times d}$ for arbitrary $d \geq 1$. When $N = 1$ there is no nontrivial row difference, so $\tau = 0$, equal to $N(N+1)/2 - 1$. Assume $N \geq 2$.

The $N(N+1)/2$ distinct symmetric score coordinates are linearly independent for every $d \geq 1$. Indeed, restrict all but the first column of Q to zero and write that column as $x \in \mathbb{R}^N$. Any linear relation becomes a quadratic polynomial

$$\sum_{1 \leq i \leq j \leq N} \lambda_{ij} x_i x_j \equiv 0,$$

so the distinct monomial coefficients λ_{ij} all vanish.

It remains to compute the rank of the row-difference constraints inside this symmetric coordinate space. A coefficient vector orthogonal to every difference $(i,j) - (i,r)$ must assign the same coefficient to every symmetric coordinate incident to row i. For two distinct rows i and j, their common coordinate $\{i,j\}$ forces these row constants to agree. Since the complete graph on $[N]$ is connected, all coefficients are equal. Thus the orthogonal complement is exactly the one-dimensional all-ones coefficient space, and the row-difference span has codimension one. This proves Corollary 22 for every $N \geq 1$ and $d \geq 1$. □

C.3 Fixed-support masked attention

Let $E \subseteq [n] \times [m]$ be a fixed edge set with every row nonempty. We model the mask by deleting forbidden edges from the score family, denominator, and output. For permitted edges use the shared dot-product parameterization

$$s_{ij}(Q,K) = Q_{i,:} K_{j,:}^\top, \qquad (i,j) \in E,$$

with $Q \in \mathbb{R}^{n\times d}$ and $K \in \mathbb{R}^{m\times d}$ independently parameterized. Define

$$P^E_{ij}(Q,K) = \begin{cases} \dfrac{e^{s_{ij}(Q,K)}}{\sum_{t:(i,t)\in E} e^{s_{it}(Q,K)}}, & (i,j)\in E, \\ 0, & (i,j)\notin E, \end{cases}$$

We count exp/log gates in circuits computing $P^E(Q,K)v$ on an admissible local domain U. Let τ_E be the $\mathbb{Q}$-dimension, modulo constants, of the span of the permitted within-row differences $s_{ij} - s_{i,r_i}$ after choosing one permitted reference edge (i, r_i) in each row.

Corollary 23 (Fixed-support masked score-difference dimension)**.** *For every $d \geq 1$ and every nonempty connected open input domain U, the minimum exp/log gate count of a U-admissible circuit computing $P^E(Q,K)v$ is*

$$\tau_E = |E| - n.$$

More generally, if empty rows are omitted from the output, replace n by the number of nonempty rows.

Proof. Choose one permitted reference edge (i, r_i) in each row. The remaining $|E| - n$ differences

$$Q_{i,:}(K_{j,:} - K_{r_i,:})^\top, \qquad (i,j)\in E,\ j \neq r_i,$$

are linearly independent. Restrict all but the first coordinate of Q and K to zero. In any purported linear relation, the monomial $Q_{i1}K_{j1}$ for a nonreference permitted edge (i,j) occurs only in its own difference, so its coefficient vanishes; repeating this for every edge proves independence. The proof of Theorem 10 applies after restricting each row to its permitted entries; forbidden edges do not enter the field-generation or span argument. This gives the exact exp/log gate count. For the causal mask $E = \{(i,j) : j \leq i\}$, this equals $N(N-1)/2$. □

C.4 Additivity over disjoint variable blocks

Proposition 24 (Additivity over disjoint variable blocks)**.** *Let $S^{(1)}, \ldots, S^{(q)}$ be rational score families on pairwise disjoint variable blocks, where $S^{(h)}$ has m_h columns. For each h, let $U_h \subseteq \mathrm{Dom}_{\mathbb{R}}(S^{(h)}) \times \mathbb{R}^{m_h}$ be nonempty, connected, and open. Put $U = \prod_{h=1}^q U_h$, and let $\mathsf{TC}_U(\bigoplus_h S^{(h)})$ denote the minimum exp/log gate count of a single eAC that jointly computes*

$$\Big(\mathrm{softmax}(S^{(h)}(z^{(h)}))v^{(h)}\Big)_{h=1}^{q} \quad \text{on } U.$$

Then

$$\mathsf{TC}_U\left(\bigoplus_{h=1}^{q} S^{(h)}\right) = \sum_{h=1}^{q} \tau(S^{(h)}).$$

In particular, q disjoint-variable dense (n,d,m) instances have exact transcendental complexity $q\,n(m-1)$.

Proposition 24 allows arbitrary gate sharing across the instances. Let a single eAC C, with $r_{\exp}$ exponential gates and $r_{\log}$ logarithm gates, jointly compute

$$\Big(\mathrm{softmax}(S^{(h)}(z^{(h)}))v^{(h)}\Big)_{h=1}^{q}$$

on $U = \prod_h U_h$, where the variable blocks are pairwise disjoint. Choose a regular point in U and, for each h, shrink around its component to a local product neighborhood $U_z^{(h)} \times U_v^{(h)} \subseteq U_h$. We work on $\prod_h (U_z^{(h)} \times U_v^{(h)})$ and simultaneously specialize every value block to its corresponding $v^{(h)\star}$. Let F be the single field generated by all specialized exponential outputs and logarithm outputs of C, together with the rational input field in all $z^{(h)}$.

Write $P^{(h)} = \mathrm{softmax}(S^{(h)}(z^{(h)}))$ and choose a reference column $r_i^{(h)}$ in each row. Lemma 17 applies to every gate and every value variable $v_j^{(h)}$. Differentiating the hth output block therefore gives

$$P^{(h)}_{ij} = \left.\frac{\partial (P^{(h)}v^{(h)})_i}{\partial v_j^{(h)}}\right|_{v^{(1)}=v^{(1)\star},\ldots,v^{(q)}=v^{(q)\star}} \in F,$$

and hence

$$\exp\left(s_{ij}^{(h)} - s_{i,r_i^{(h)}}^{(h)}\right) = \frac{P_{ij}^{(h)}}{P_{i,r_i^{(h)}}^{(h)}} \in F$$

for every h, i, j. Repeat the proof of Theorem 5 in this single joint gate field, using the union of the score-difference families. The derivative-field step above gives every required coefficient ratio, and the same gate-prefix transcendence and Ax argument shows that every such difference lies in the $\mathbb{Q}$-linear span of the common circuit's exponential inputs and logarithm outputs. Hence

$$\dim_{\mathbb{Q}}\left(\frac{\sum_h \mathcal{D}(S^{(h)})}{\mathcal{C}}\right) \le r_{\exp} + r_{\log}.$$

It remains to compute the dimension of the joint score-difference space. Suppose $\sum_h f_h(z^{(h)})$ is constant with $f_h \in \mathcal{D}(S^{(h)})$. In the local product neighborhood above, fix all blocks except $z^{(h)}$ and vary $z^{(h)}$ in $U_z^{(h)}$. Hence f_h is constant as a germ. Thus the spaces are an internal direct sum modulo constants, and

$$\dim_{\mathbb{Q}}\left(\frac{\sum_h \mathcal{D}(S^{(h)})}{\mathcal{C}}\right) = \sum_h \tau(S^{(h)}).$$

This proves the lower bound even with arbitrary gate sharing. The upper bound is obtained by placing the optimal local circuits from Theorem 10 side by side, completing the proof of Proposition 24.

D Mixed-Coefficient Extraction

We use the mixed-$(1,1)$-coefficient version of the low-degree eAC-to-AC simulation of Saha et al. (2026, Theorem 4.1), expanded at a regular base point. This simulation builds on division elimination. We use two perturbation parameters and retain the coefficients of internal exponential inputs and logarithm outputs. The coefficient recurrence gives constant-overhead extraction, while Theorem 5 identifies the internal-function span whose extracted coefficients linearly generate the score-difference tensor. Regularity of the base point makes every denominator in the recurrence nonzero.

Fix matrices $Q^\star, K^\star$ at which all specialized gates are defined, and use the perturbation in Eq. (2). Introduce fresh matrix variables X, Y and substitute

$$Q = Q^\star + uX, \qquad K = K^\star + wY.$$

For a gate value f, write

$$f_{00} = [1]f, \quad f_{10} = [u]f, \quad f_{01} = [w]f, \quad f_{11} = [uw]f$$

in the truncated ring modulo (u^2, w^2). Here f_{00} is a scalar constant, f_{10} is linear in X, f_{01} is linear in Y, and f_{11} is bilinear in (X, Y).

For addition and subtraction, coefficients are taken componentwise. For multiplication $f = gh$,

$$\begin{aligned}
f_{00} &= g_{00}h_{00},\\
f_{10} &= g_{10}h_{00} + g_{00}h_{10},\\
f_{01} &= g_{01}h_{00} + g_{00}h_{01},\\
f_{11} &= g_{11}h_{00} + g_{10}h_{01} + g_{01}h_{10} + g_{00}h_{11}.
\end{aligned}$$

For inversion $f = g^{-1}$, with $g_{00} \neq 0$,

$$\begin{aligned}
f_{00} &= g_{00}^{-1},\\
f_{10} &= -g_{10}g_{00}^{-2},\\
f_{01} &= -g_{01}g_{00}^{-2},\\
f_{11} &= 2g_{10}g_{01}g_{00}^{-3} - g_{11}g_{00}^{-2}.
\end{aligned}$$

A division gate is handled by multiplication with the inverse.

For an exponential gate $f = e^g$,

$$f_{00} = e^{g_{00}}, \qquad f_{10} = f_{00}g_{10},$$
$$f_{01} = f_{00}g_{01}, \qquad f_{11} = f_{00}(g_{11} + g_{10}g_{01}). \tag{3}$$

For a logarithm gate $f = \ln g$, with the local branch fixed and $g_{00} \neq 0$,

$$f_{00} = \ln g_{00}, \qquad f_{10} = g_{10}/g_{00},$$
$$f_{01} = g_{01}/g_{00}, \qquad f_{11} = g_{11}/g_{00} - g_{10}g_{01}/g_{00}^2. \tag{4}$$

Each original gate requires only a constant number of arithmetic gates for these four components. At the real regular base point, g_{00} is real and every logarithm input is positive. Hence $e^{g_{00}}$, $\ln g_{00}$, and g_{00}^{-1} are real constants independent of X, Y and may be hardwired in the nonuniform real arithmetic circuit.

The recurrences above prove Lemma 8 from the main text: each original gate contributes only a fixed number of arithmetic gates for the four retained coefficients. The only apparent divisions are by nonzero scalar constants, so hardwiring their inverses makes the extracted circuit division-free.

For $j \in [m-1]$, define the score difference

$$p_{ij}(Q, K) = (QK^\top)_{i,j+1} - (QK^\top)_{i1}.$$

Its mixed coefficient is

$$[p_{ij}]_{11} = (XY^\top)_{i,j+1} - (XY^\top)_{i1}.$$

Thus the vector of mixed score-difference coefficients is precisely the score-difference tensor $\mathbf{\Delta}(X, Y)$, independently of the chosen base point.

E Proof of the Rectangular Matrix-Multiplication Lower Bound

E.1 Tensor extraction from the span relations

Choose a point and product neighborhood as in Appendix A.2, specialize the value vector to that point, and list all exponential inputs $g_1, \ldots, g_{r_{\exp}}$ and logarithm outputs $L_1, \ldots, L_{r_{\log}}$. By Theorem 5, every reference score difference has a fixed rational relation

$$p_{ij} = c_{ij} + \sum_{\nu=1}^{r_{\exp}} \alpha_{ij\nu} g_\nu + \sum_{\mu=1}^{r_{\log}} \beta_{ij\mu} L_\mu.$$

Apply Lemma 8 at a regular real base point. The resulting $O(s)$-size division-free arithmetic circuit jointly computes the mixed coefficients

$$G_\nu(X, Y) = [uw]g_\nu(Q^\star + uX, K^\star + wY), \qquad J_\mu(X, Y) = [uw]L_\mu(Q^\star + uX, K^\star + wY).$$

Taking mixed coefficients in the span relations yields

$$\mathbf{\Delta}(X, Y) = \Pi(G_1, \ldots, G_{r_{\exp}}, J_1, \ldots, J_{r_{\log}})$$

for a fixed rational linear output map Π. Set $\Gamma = (G_1, \ldots, G_{r_{\exp}}, J_1, \ldots, J_{r_{\log}})$. Lemma 9 gives

$$\mathrm{R}(\langle n, d, m-1\rangle) = \mathrm{R}(\mathbf{\Delta}) \leq \mathrm{R}(\Gamma) \leq O(s).$$

The rank-one decomposition is projected by replacing each constant output vector c_k with Πc_k. Tensor-rank monotonicity applies directly to this fixed output projection. The exponent statement follows from Proposition 25, proved next.

E.2 From tensor-rank exponents to family-wise bounds

Proposition 25 (Rectangular rank regularization). *Fix $a > 0$ and let $r_a(N) = \mathrm{R}(\langle N, \lceil N^a \rceil, N\rangle)$. Then*

$$r_a(N) = N^{\omega(1,a,1)+o(1)} \quad \textit{and} \quad \mathrm{R}(\langle N, \lceil N^a \rceil, N-1\rangle) = N^{\omega(1,a,1)+o(1)}.$$

The same exponent governs scalar arithmetic size under the standard recursive evaluation of bilinear algorithms. Moreover $\omega(1, a, 1) \geq 2$.

Write $\omega = \omega(1, a, 1)$. For the lower estimate, the definition by a liminf gives, for every $\varepsilon > 0$ and all sufficiently large N,

$$r_a(N) \geq N^{\omega-\varepsilon}.$$

For the matching upper estimate, fix $\varepsilon > 0$. By the definition of the liminf, choose an integer $q \geq 2$ such that

$$r_a(q) \leq q^{\omega+\varepsilon}.$$

To extend this fixed-size bound to arbitrary sufficiently large N, choose k with $q^{k-1} < N \leq q^k$. Tensoring a rank decomposition for shape $\langle q, \lceil q^a \rceil, q \rangle$ gives

$$\mathrm{R}(\langle q^k, \lceil q^a \rceil^k, q^k \rangle) \leq r_a(q)^k.$$

The target dimensions fit inside this tensor because

$$N \leq q^k, \qquad \lceil N^a \rceil \leq \lceil q^{ak} \rceil \leq \lceil q^a \rceil^k.$$

Restricting the padded tensor therefore yields

$$r_a(N) \leq r_a(q)^k \leq q^{k(\omega+\varepsilon)} \leq (qN)^{\omega+\varepsilon} = N^{\omega+\varepsilon+o(1)}.$$

Because ε is arbitrary,

$$r_a(N) = N^{\omega+o(1)}.$$

To handle the omitted reference column, take $M = \lfloor N/2 \rfloor$. Tensor restriction gives

$$r_a(M) \leq \mathrm{R}(\langle N, \lceil N^a \rceil, N-1 \rangle) \leq r_a(N),$$

because $M \leq N - 1$ and $\lceil M^a \rceil \leq \lceil N^a \rceil$ for $N \geq 2$. Since $M = \Theta(N)$, the preceding estimate gives

$$\mathrm{R}(\langle N, \lceil N^a \rceil, N-1 \rangle) = N^{\omega+o(1)}.$$

To account for scalar arithmetic operations, fix the base algorithm above, put $h = \lceil q^a \rceil$, and write $R = r_a(q)$. The three tensor flattenings give $R \geq \max\{qh, q^2\}$. Recursive evaluation of this fixed bilinear algorithm satisfies

$$T(k) \leq RT(k-1) + C\big((qh)^{k-1} + q^{2k-2}\big),$$

where C depends only on the fixed base algorithm. Iterating this recurrence gives $T(k) = O(kR^k)$, so the scalar additions and multiplications incur no extra exponent in the preceding padding argument. The reverse direction is standard bilinearization, with divisions eliminated at a regular point (Strassen, 1973; Schönhage, 1981; Bürgisser et al., 1997; Bläser, 2013).

Finally, the flattening onto the N^2-dimensional output space gives $r_a(N) \geq N^2$, so $\omega \geq 2$. The entrywise exponentiation, normalization, and value aggregation cost $O(N^2)$ gates for $0 < a \leq 1$ and hence do not change the exponent.

E.3 Total-size lower bound for tied queries and keys

We prove Corollary 13. Partition the $N = r + t$ rows into two nonempty blocks and relabel the input coordinates as

$$Q = \begin{bmatrix} Q^{(1)} \\ Q^{(2)} \end{bmatrix}, \qquad Q^{(1)} \in \mathbb{R}^{r \times d}, \quad Q^{(2)} \in \mathbb{R}^{t \times d}.$$

This invertible relabeling uses all entries of Q and maps every admissible open input domain to an open domain in $(Q^{(1)}, Q^{(2)}, v)$.

Fix a regular point and specialize v to its value at that point. Theorem 5 applies to the rational score family $QQ^\top$. For each query $i \in [r]$, use column $r+1$ as the row reference and select the differences to columns $r+2, \ldots, r+t$. These are

$$\delta_{ij}(Q) = (QQ^\top)_{i,r+j+1} - (QQ^\top)_{i,r+1} = \big(Q^{(1)}(Q^{(2)})^\top\big)_{i,j+1} - \big(Q^{(1)}(Q^{(2)})^\top\big)_{i1}, \qquad j \in [t-1].$$

Every δ_{ij} lies in the $\mathbb{Q}$-linear span of the specialized exponential inputs and logarithm outputs, modulo constants.

Perturb the two row blocks independently about the same regular point:

$$Q(u,w) = \begin{bmatrix} Q^{(1)\star} + uX \\ Q^{(2)\star} + wY \end{bmatrix}.$$

Then

$$[uw]\delta_{ij}(Q(u,w)) = (XY^\top)_{i,j+1} - (XY^\top)_{i1}.$$

The within-block score terms involve only u or only w and do not contribute to this mixed coefficient. Lemma 8 computes the mixed coefficients of all selected internal gate functions in size $O(s)$. The mixed coefficients $([uw]\delta_{ij})_{i,j}$ are a fixed linear output projection of that coefficient vector. By Lemma 9 and Proposition 4,

$$\mathrm{R}(\langle r, d, t-1\rangle) \le O(s).$$

All coefficient extraction takes place at a regular point of the original open domain.

For the exponent, fix $0 < a \le 1$, let $r = \lfloor N/2 \rfloor$, $t = N - r$, and $d_N = \lceil N^a \rceil$. For sufficiently large N, with $M = \lfloor N/3 \rfloor$,

$$\mathrm{R}(\langle M, \lceil M^a \rceil, M\rangle) \le \mathrm{R}(\langle r, d_N, t-1\rangle) \le \mathrm{R}(\langle N, d_N, N\rangle).$$

Proposition 25 and $M = \Theta(N)$ give the same exponent $\omega(1, a, 1)$. The upper bound computes $QQ^\top$ by rectangular multiplication and then uses $O(N^2)$ gates for entrywise exponentiation, row normalization, and multiplication by v. Since $\omega(1, a, 1) \ge 2$, these operations do not change the exponent.

The map also has a fixed-projection realization with tokens $(Q_{i,:}, v_i)$, identical query and key projections selecting $Q_{i,:}$, and a value projection selecting v_i. Thus the tied result uses input width $d + 1$, with the same coordinates supplying queries and keys.

E.4 Lower bounds with one exponential per path

Theorem 1 already applies to log-free eACs of arbitrary exponential depth. Under the additional restriction of at most one exponential gate per input–output path, the rational preprocessing alone satisfies the same rank lower bound.

Corollary 26 (Preprocessing lower bound at exponential depth one)**.** *Let C be a U-admissible log-free eAC computing $\mathcal{A}_{n,d,m}$. Suppose that every input–output path contains at most one exponential gate, and let s_{pre} be the size of the rational preprocessing subcircuit preceding the exponential gates. Then*

$$s_{\mathrm{pre}} = \Omega(\mathrm{R}(\langle n, d, m-1\rangle)).$$

In such a circuit, every exponential input is computed before the exponential layer and there are no logarithm gates. The mixed-coefficient extraction therefore needs only the rational preprocessor. If its size is s_{pre}, the same projection gives

$$\mathrm{R}(\langle n, d, m-1\rangle) = O(s_{\mathrm{pre}}),$$

proving Corollary 26.

F A Mixed-Hessian Identity for Attention

A complementary second-order identity gives another view of score recovery from attention. Append one coordinate to the queries and keys, giving scores $(QK^\top)_{ij} + x_i y_j$, and add a zero-score anchor key. Give value 1 to each real key and 0 to the anchor. The sum of the query outputs is

$$\Phi_{Q,K}(x,y) = \sum_{i=1}^{n} \frac{Z_i(x,y)}{1 + Z_i(x,y)}, \qquad Z_i(x,y) = \sum_{j=1}^{m} e^{(QK^\top)_{ij} + x_i y_j}.$$

For $n = m = N$, the realization below uses $N + 1$ shared tokens and fixed projections.

Theorem 27 (Mixed-Hessian identity)**.** *At $x = y = 0$,*

$$H_{ij} := \left.\frac{\partial^2 \Phi_{Q,K}}{\partial x_i \partial y_j}\right|_{x=y=0} = \frac{e^{(QK^\top)_{ij}}}{\left(1 + \sum_t e^{(QK^\top)_{it}}\right)^2} > 0. \tag{5}$$

Hence $(QK^\top)_{ij} - (QK^\top)_{i1} = \ln H_{ij} - \ln H_{i1}$. If the coefficients H_{ij} are explicitly available, the matrix H, together with one reference column computable in $O(nd)$ gates, determines $QK^\top$.

F.1 Attention realization of the mixed-Hessian identity

For general n and m the construction is one ordinary softmax attention head, naturally viewed as cross-attention when $n \neq m$. In the square case $n = m = N$, use $N + 1$ shared tokens $X_i = (Q_{i,:}, x_i, K_{i,:}, y_i, 1)$ for $i \leq N$ and one zero anchor token. Fixed projections select the query block $(Q_{i,:}, x_i)$, the key block $(K_{i,:}, y_i)$, and the final scalar value; the anchor query output is ignored. Thus the coefficient identity is also a restriction of standard self-attention in the square case.

For each query i, write

$$\phi_i(x, y) = \frac{Z_i(x, y)}{1 + Z_i(x, y)}, \qquad Z_i(x, y) = \sum_{t=1}^{m} e^{M_{it} + x_i y_t}, \qquad M = QK^\top.$$

At $x = y = 0$, the required derivatives are

$$\left.\frac{\partial Z_i}{\partial x_i}\right|_0 = 0, \qquad \left.\frac{\partial Z_i}{\partial y_j}\right|_0 = 0, \qquad \left.\frac{\partial^2 Z_i}{\partial x_i \partial y_j}\right|_0 = e^{M_{ij}}.$$

The chain rule therefore gives

$$\begin{aligned}\left.\frac{\partial^2 \phi_i}{\partial x_i \partial y_j}\right|_0 &= \left.\left(\frac{\partial^2 Z_i / \partial x_i \partial y_j}{(1 + Z_i)^2} - \frac{2(\partial Z_i / \partial x_i)(\partial Z_i / \partial y_j)}{(1 + Z_i)^3}\right)\right|_0 \\ &= \frac{e^{M_{ij}}}{\left(1 + \sum_t e^{M_{it}}\right)^2}.\end{aligned}$$

No $\phi_{i'}$ with $i' \neq i$ depends on x_i, proving (5). Taking a within-row ratio cancels the normalizer:

$$\frac{H_{ij}}{H_{i1}} = e^{M_{ij} - M_{i1}}.$$

If all coefficients of H are explicitly computed, computing the reference entries $M_{i1} = \sum_{k=1}^{d} Q_{ik} K_{1k}$ for all i costs $O(nd)$ gates, after which logarithmic ratios recover every entry. Every H_{ij} is positive, so all real logarithmic ratios are well defined. The construction copies Q, K, x, y and uses only 0 and 1 as structural constants.

Theorem 28 (Extended Baur–Strassen)**.** *Let a scalar function be computed on a regular local domain by a size-s eAC with gates $+, -, \times, /$, $\exp$, and $\ln$. Then all of its first partial derivatives are computed on that domain by an eAC of size $O(s)$, with a universal implicit constant.*

The arithmetic version is due to Baur & Strassen (1983); Saha et al. (2026, Theorem 4.4) extend it to the present gate set. The theorem is a local chain-rule construction. Its quotient and logarithm steps require the same nonvanishing denominators and fixed regular logarithm branches as the original eAC.

F.2 A derivative circuit for the Hessian–vector map

Assume that the size-s circuit under consideration is regular on a product neighbourhood $U_{QK} \times U_x \times U_y$ with $0 \in U_x$ and computes $\Phi_{Q,K}(x, y)$ there. The direct attention construction is regular for all real inputs, so it satisfies this condition. Apply Theorem 28 on this product neighborhood and subsequently specialize the derivative circuit to $x = 0$. The resulting circuit is defined on $U_{QK} \times U_y$.

Retain symbolic y but set $x = 0$. Then $Z_i(0, y) = \sum_j e^{M_{ij}}$ is independent of y, while

$$\left.\frac{\partial Z_i}{\partial x_i}\right|_{x=0} = \sum_j e^{M_{ij}} y_j.$$

Therefore

$$\nabla_x \Phi_{Q,K}(x, y)|_{x=0} = Hy.$$

By Theorem 28, the linear operator $y \mapsto Hy$ has an $O(s)$-size eAC. A general $O(s)$-size procedure for explicitly extracting all mixed second-order coefficients from a size-s scalar circuit would, in particular, output all entries of $QK^\top$ from the $O((n + m)d)$-size circuit for $z^\top QK^\top y$.

G LIMITS OF STANDARD REDUCTIONS

We examine the score-scaling Taylor expansion, first-order perturbations of the factorization $QK^\top$, and direct value selectors to explain the exponent bounds obtained by these three standard approaches.

G.1 FIRST-ORDER TAYLOR REDUCTION

For a row vector $s \in \mathbb{R}^m$,

$$\operatorname{softmax}(\lambda s) = \frac{1}{m}\mathbf{1}^\top + \frac{\lambda}{m} s C_m + O(\lambda^2), \qquad C_m = I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top.$$

Let $Q \in \mathbb{R}^{n\times d}$, $K \in \mathbb{R}^{m\times d}$, and $V \in \mathbb{R}^{m\times d_v}$, where d_v is the value width. Applying this rowwise to $QK^\top$ and multiplying by V gives

$$\operatorname{softmax}(\lambda QK^\top)V = \frac{1}{m}\mathbf{1}\mathbf{1}^\top V + \frac{\lambda}{m} Q(K^\top C_m V) + O(\lambda^2). \tag{6}$$

If the value width is d, the nonconstant term is trilinear in the independent inputs Q, K, V. Compute $C_m V$ in $O(md)$ arithmetic operations, then form $K^\top(C_m V)$ and multiply the resulting $d \times d$ matrix by Q. The two matrix products have shapes $\langle d, m, d\rangle$ and $\langle n, d, d\rangle$. When $n = m = N$ and $d = \lceil N^a \rceil$, permutation symmetry of the matrix-multiplication tensor gives exponent $\omega(1, a, a)$ for both. This first-order reduction therefore reaches the prior $\omega(1, a, a)$ frontier.

The differential of the score factorization is

$$D(QK^\top)[\dot{Q}, \dot{K}] = \dot{Q}K^\top + Q\dot{K}^\top.$$

At a smooth rank-d score matrix, these perturbations form the tangent space to the rank-at-most-d variety. Its dimension is $d(n+m) - d^2$, and every tangent vector has rank at most $2d$. Thus first-order score perturbations cannot span all nm coordinate basis vectors when $d < \min\{n, m\}$.

G.2 SCORE RECOVERY WITH COORDINATE SELECTORS

For globally correct circuits, the anchor and selector slices lie in the regular domain; the local version applies whenever the circuit is regular and correct on neighborhoods of these slices.

Add one zero-score anchor key. Partition the m real keys into blocks of size at most $d-1$ (or d if the anchor value uses an additional coordinate). For one block $J = \{j_1, \ldots, j_t\}$, let e_s denote the sth standard basis vector in the value space and choose $V_{j_s} = e_s$, $V_{\mathrm{anc}} = e_{t+1}$, and zero values for other real keys. The output contains

$$Y_{i,s} = \frac{e^{(QK^\top)_{i,j_s}}}{1 + \sum_j e^{(QK^\top)_{ij}}}, \qquad Y_{i,t+1} = \frac{1}{1 + \sum_j e^{(QK^\top)_{ij}}}.$$

Hence $(QK^\top)_{i,j_s} = \ln(Y_{i,s}/Y_{i,t+1})$. At value width d, this is exact with blocks of size at most $d-1$ and therefore $\lceil m/(d-1)\rceil$ calls when $d \geq 2$. Equivalently, width $d+1$, which leaves the exponent unchanged, allows blocks of d real keys plus the anchor and uses $\lceil m/d\rceil$ calls.

Let $T(N, d)$ be the size of a globally correct eAC for a head with $N+1$ tokens, query/key width d, and value width $d+1$. Fixed coordinate projections select independent query, key, and value blocks from tokens of width $3d+1$; the added token has zero query and key. Thus the call dimensions include the anchor and the additional value coordinate. For $m = n = N$ and $d = \lceil N^a \rceil$, the selector calls and logarithmic recovery give an eAC computing $QK^\top$ with size

$$O\big(N^{1-a}T(N, d) + N^2\big).$$

To obtain an arithmetic circuit, apply Lemma 8 to this composite eAC at a regular real base point. Under $Q = Q^\star + uX$ and $K = K^\star + wY$, the mixed output coefficient is $XY^\top$. The resulting division-free AC has the same asymptotic size, so

$$\mathrm{R}(\langle N, d, N\rangle) \leq O\big(N^{1-a}T(N, d) + N^2\big).$$

If $\omega(1, a, 1) > 2$, the rank lower bound dominates the recovery cost and implies

$$T(N, \lceil N^a \rceil) \geq N^{\omega(1,a,1)-1+a-o(1)}.$$

When $\omega(1,a,1) = 2$, this additive-cost comparison alone does not imply $T(N, \lceil N^a \rceil) \geq N^{\omega(1,a,1)-1+a-o(1)}$. Standard column blocking of the right matrix factor gives

$$\omega(1,a,1) \leq 1 - a + \omega(1,a,a),$$

so the exponent supplied by this coordinate-selector reduction is no larger than the established $\omega(1,a,a)$ frontier.

G.3 Limits of standard recovery strategies

We next examine repeated differentiation, fixed linear measurements, digit packing, and direct algebra simulation.

Repeated differentiation. On a domain containing the required zero slice, one Baur–Strassen pass gives Hy for symbolic y. Recovering all columns by basis-vector evaluation uses m evaluations, provided those inputs lie in the evaluation domain.

Fixed linear measurements (Vandermonde/moment values). For an attention row with m entries, including any anchor, a fixed value matrix $V \in \mathbb{R}^{m\times d}$ gives d linear measurements. The row also satisfies $\sum_{j=1}^{m} p_j = 1$, so its affine dimension is $m-1$. These measurements determine every such row exactly if and only if $\operatorname{rank}[\mathbf{1}\ V] = m$: otherwise a nonzero vector in $\ker(V^\top) \cap \mathbf{1}^\perp$ gives indistinguishable positive rows by a sufficiently small perturbation. Thus fewer than $m-1$ independent measurements are insufficient, while $m-1$ coordinate measurements suffice because the last entry follows from normalization. Thus fixed linear measurements need $m-1$ independent degrees of freedom per arbitrary positive row in the worst case.

Digit/exponent packing. The usual digit-packing construction decodes by floor, remainder, or a limiting operation, none of which is an eAC gate. Hence the standard digit-packing decoder is unavailable in the eAC model; alternative exact real encodings would need a different decoding mechanism.

Nilpotent packing. In direct coordinatewise simulation over an r-dimensional algebra, even adding two generic algebra elements requires r scalar additions. Applying this simulation separately at each circuit gate introduces that coordinate overhead. At $r = m/d$, direct coordinatewise simulation retains the selector-block factor. This calculation concerns unbatched, coordinatewise operations.

H Relation to Functional Transcendence and Elementary-Function Complexity

Ax's theorem supplies the transcendence-degree inequality used in our proof (Ax, 1971). Classical differential algebra studies elementary extensions and their algebraic relations (Kolchin, 1973; Rosenlicht, 1975). Grigoriev & Vorobjov (1996) study computation-tree depth for decision problems, and Li & Wu (2026) study identity testing with at most one exponential gate on each input–output path. Our object is the representation size of a local exp–log DAG with arbitrary nesting and sharing.

Structural and integration results for elementary extensions include Ritt (1929); Risch (1969; 1979); Rosenlicht (1972; 1975); Rothstein & Caviness (1979). Elementary identity and simplification problems are studied by Richardson (1968); Epstein & Caviness (1979). Forbes et al. (2016) prove functional lower bounds for polynomial circuits on discrete domains.

Direction	Object	Typical conclusion	Key distinction
Ax / functional transcendence	Differential fields with associated exponentials	Lower bound on transcendence degree	Gives the inequality used after the gate-generated field and score differences have been identified
Ritt, Risch, Rosenlicht	Elementary, exponential, or primitive extensions	Structure, integration, or algebraic relations	Elementary-extension structure; no fixed-DAG size measure
Richardson; Epstein–Caviness; Rothstein–Caviness	Elementary expressions and symbolic computation	Identity, dependence, or normal-form results	Symbolic identity/normal-form objective
Li–Wu	At most one exponential per input–output path	Identity testing and exponential-polynomial normal form	One exponential per path; identity testing
Grigoriev–Vorobjov	Computation trees with elementary transcendental gates	Depth lower bounds for decision problems	Tree/depth model
This paper	Local exact eAC representation of normalized exponential maps	$\mathsf{TC}_U(S) = \tau(S)$ and unrestricted tensor lower bound	Local DAG representation size with arbitrary exp–log nesting

Together, the gate-prefix bound and value differentiation turn functional transcendence into representation-size and tensor-rank lower bounds for exact attention.

I Exactness, Exponent Accounting, and Domain Conditions

I.1 Exactness and exponent accounting

All reductions use ordinary row-wise softmax. The mixed-Hessian construction adds one coordinate and one anchor key, so $d' = d + 1$ and $m' = m + 1$, preserving polynomial exponents. Its preprocessing copies the input entries and inserts the constants 0 and 1; after the Hessian coefficients are available, one reference column costs $O(nd)$ gates and the remaining recovery uses row-wise logarithmic ratios.

The unrestricted total-size lower bound uses a single attention computation and keeps the original sequence dimensions. Nonuniformity is used to hardwire real constants arising from a regular base point and the fixed rational linear relations.

I.2 Domain and regularity conditions

Each analytic step is carried out inside a regular neighborhood. Appendix A.2 fixes a product neighborhood and the logarithm branches. Lemma 18 restricts inside a connected ball contained in the regular domain, while Lemmas 6, 19, 20, and 21 establish the circuit-field upper bound, the algebraicity of the basis exponentials, the differential closure of the algebraic extension, and the Ax lower bound. Appendix D expands at a regular real base point, so every hardwired constant is real and every denominator is nonzero.

After these analytic ingredients are fixed, the remaining steps are algebraic: Appendix C.4 handles joint outputs, Appendix E performs tensor extraction, and Appendix F analyzes the Hessian identity.