WHAT THE EXPONENTIAL ADDS: SOFTMAX, HARD ATTENTION, AND EXACT COUNTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hard-attention models simplify the study of transformer expressivity, but their conclusions need not transfer directly to softmax. Averaging hard attention (AHAT) averages values at the highest-scoring positions. Without positional encodings or masking, its rational-parameter version recognizes exactly the counting properties defined by Boolean combinations of polynomial inequalities with rational coefficients. Softmax recognizes strictly more. Two layers decide whether $\#a > \#c\, e^{\#b/(n+1)}$, where n excludes one appended end marker: the first averages the b indicator, and the second uses this average as the score for c. No semialgebraic description captures this property. The same network works when each position attends to itself and earlier positions, answering an open question. More generally, replace e^u by a continuous positive function on a fixed closed interval containing all scores, keeping the remaining operations semialgebraic. Unmasked networks can recognize a nonsemialgebraic counting property exactly when this function is nonsemialgebraic on that interval; polynomial replacements therefore cannot recognize our example. Masking provides another route: three layers attending equally to earlier positions compare a count with $1+1/2+\cdots+1/(n-1)$, a property no unmasked softmax network built from affine maps and ReLUs recognizes. Finally, replacing e^u by one polynomial keeps outputs of a fixed network with bounded inputs, affine/ReLU updates and an affine readout within any prescribed tolerance at every length. Preserving every decision of even our two-layer network at length n requires degree at least $(1 - o(1)) \log n / \log \log n$. Among positive count triples of length n, the degree-ten Taylor replacement's disagreement fraction tends to about 3×10^{-11}, yet it flips a length-1080 decision.

1 INTRODUCTION

Hard-attention models make transformer expressivity easier to analyze (Hao et al., 2022; Barceló et al., 2024). Transferring their conclusions to softmax, however, requires a relation between the models. Known simulations use unbounded positional embeddings or temperature scaling chosen for the input length and attention-score gaps (Yang et al., 2026). We compare single networks without positional encodings, each used at every input length. Counting properties, such as whether one token occurs more often than another, allow an exact comparison because the unmasked AHAT side is already characterized. Without positional encodings or masking, Sälzer et al. (2026) show that averaging hard attention (AHAT), which averages values at the highest-scoring positions, recognizes exactly the properties described by Boolean combinations of polynomial inequalities with rational coefficients. Even uniform attention, which gives all positions equal weight, recognizes this entire class. Softmax with equal scores implements uniform attention, so it recognizes every such property as well. Does softmax recognize anything more?

We show that softmax recognizes strictly more. With x, y, z occurrences of a, b, c and $n = x+y+z$, consider

$$x > z \exp\left(\frac{y}{n+1}\right). \tag{1}$$

The denominator includes one appended end marker. Each c counts against a with weight determined by the frequency of b. Holding the counts of a and c fixed, $(3, 1, 2)$ is accepted because $2e^{1/7} < 3$, whereas $(3, 7, 2)$ is rejected because $2e^{7/13} > 3$. Two softmax layers implement this

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.

comparison: first average the b indicator, then use that average as the attention score for c. No finite combination of polynomial conditions describes the resulting counting property (Theorem 1). The same network works when each position attends to itself and its predecessors, answering an open question on causal counting posed by Sälzer et al. (2026, Future Work).

Softmax assigns weight e^u to an attention score u. To identify the source of its extra counting power, replace e^u by a continuous positive function $\kappa(u)$ on a fixed closed interval containing all scores, keeping the remaining operations semialgebraic, as affine maps and ReLUs are. Without masking, every such network recognizes a semialgebraic counting property if κ is semialgebraic; otherwise, two layers with affine updates already recognize a nonsemialgebraic one (Theorem 2). Thus unmasked networks using a polynomial in place of e^u cannot recognize (1), at any fixed finite depth and width.

Masking provides an independent route. When each position attends only to its predecessors, different positions average prefixes of different lengths. Three layers with uniform weights can generate reciprocals and compare the count of a with the harmonic sum $H_{n-1} = 1 + 1/2 + \cdots + 1/(n-1)$. This final decision depends only on counts, although its intermediate computation uses position. No unmasked softmax network with affine/ReLU computation recognizes it. The exclusion extends to every threshold, such as $\log n$, that tends to infinity more slowly than every positive power of n (Theorem 3). Thus strict causal masking supplies a second route beyond semialgebraic counting, even with uniform weights. Figure 2 shows the two mechanisms.

Polynomial replacements of e^u are used in transformer architectures and secure inference (Zimerman et al., 2024; Li et al., 2024). Their loss of exact counting power raises a quantitative question: what can a fixed replacement preserve? Output error smaller than the distance to the decision threshold guarantees the same sign, but this margin need not stay bounded away from zero as length grows. We show that a fixed output tolerance can be met at every length by one polynomial, whereas preserving every decision of even the two-layer network above forces degree to grow with length. The degree-ten Taylor polynomial illustrates the distinction: among positive count triples of length n, the fraction of disagreeing decisions tends to about 3×10^{-11} as n grows, yet the label at $(74, 976, 30)$ changes.

Contributions. Our four contributions identify two sources of exact counting power (Figure 1) and quantify what polynomial substitution loses:

- **Softmax exceeds AHAT.** Two layers recognize a nonsemialgebraic counting property, so softmax recognizes strictly more counting properties than AHAT without masking, under the rational-parameter convention of Sälzer et al. (2026). The same construction works with self-inclusive causal masking (Theorem 1).
- **A criterion for weight functions.** In unmasked networks with otherwise semialgebraic computation, replacing e^u by a continuous positive function permits counting beyond polynomial descriptions exactly when that function is nonsemialgebraic on the fixed closed score interval. Polynomial weights therefore fail even with a redesigned fixed finite architecture (Theorem 2).
- **An independent role for masking.** Three layers attending equally to earlier positions decide whether $\#a > 1 + 1/2 + \cdots + 1/(n-1)$. Unmasked affine/ReLU softmax cannot recognize this property, or compare a count with any threshold growing without bound but more slowly than every positive power of n, such as $\log n$ (Theorem 3).
- **Output accuracy versus exact decisions.** For output tolerance η, degree with leading term $\log(1/\eta)/\log\log(1/\eta)$ suffices at every length and is necessary for our two-layer example. Preserving every decision of that example at length n instead requires degree at least $(1 - o(1)) \log n / \log\log n$ (Proposition 4, Theorem 5).

2 Setting

We follow the counting-language framework of Sälzer et al. (2026), with the execution conventions below. A word $w = w_1 \cdots w_n$ over a finite alphabet is followed by one end marker \$. Its length n excludes the marker. Each token type (a symbol of the alphabet) has a fixed embedding, with

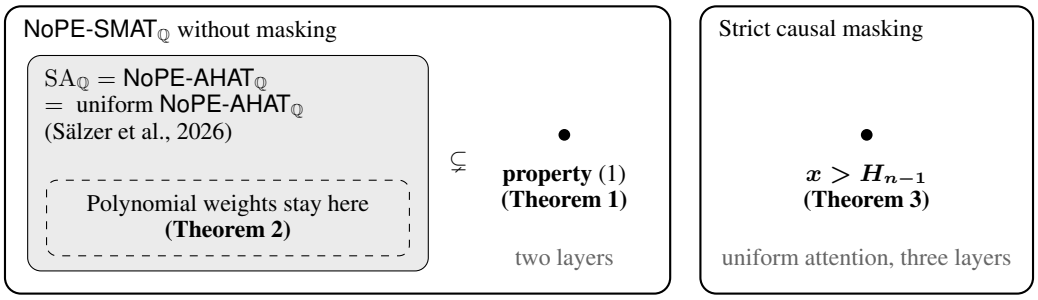

Figure 1: Counting properties without positional encodings; new results are in bold. The enclosed classes use rational parameters and affine/ReLU computation; $\mathrm{SA}_\mathbb{Q}$ denotes rational-coefficient semialgebraic counting. SMAT denotes softmax. The equalities and non-strict inclusion are due to Sälzer et al. (2026); strictness is Theorem 1. The dashed annotation uses rational polynomial coefficients and positive reachable weights. The harmonic property lies outside even real-parameter unmasked softmax.

no additional positional encoding. This finite set of embeddings bounds the initial state norm independently of length. A network has fixed finite depth, width, head counts, and real parameters, independent of n. Its end-marker state produces a scalar output: positive means acceptance and non-positive means rejection. We reject the empty word by convention. A counting property depends on the number of occurrences of each type, rather than their order.

Let h_i be the incoming state at position i. Head k applies affine query, key, and value maps Q_k, K_k, V_k. A positive *weight function* κ_k converts a score to an unnormalized weight:

$$o_{i,k} = \frac{\sum_{j \in J_i} \kappa_k(\langle Q_k(h_i), K_k(h_j)\rangle) V_k(h_j)}{\sum_{j \in J_i} \kappa_k(\langle Q_k(h_i), K_k(h_j)\rangle)}, \qquad h_i^+ = \Phi(h_i, o_{i,1}, \ldots, o_{i,M}). \tag{2}$$

The same map Φ updates each position separately; it can include residual connections and output projections. Unless a result specifies a larger class, these updates and the scalar readout use finitely many affine maps and ReLUs. Softmax uses $\kappa(u) = e^u$. A positive constant weight function gives equal weights, as does any positive weight function when all scores are equal.

Under *no masking*, each J_i contains all $n+1$ positions, including i and the end marker. Under *strict causal masking*, $J_i = \{j : j < i\}$; an empty set returns zero. Self-inclusive causal masking uses $J_i = \{j : j \le i\}$. Appendix A supplies the coordinate implementation and signed affine updates used below.

A set is *semialgebraic* if it is a finite Boolean combination of polynomial equalities and inequalities with real coefficients; a function is semialgebraic when its graph is. A counting property is semialgebraic if its accepted count vectors are the intersection of such a set with the integer lattice. Affine maps, ReLU and polynomials are semialgebraic. Sälzer et al. (2026) use rational coefficients and rational network parameters, and we retain this convention when comparing with AHAT. Our impossibility results hold even with real coefficients and parameters. As in that work, we assume exact real arithmetic.

3 Softmax exceeds semialgebraic counting

3.1 An exponential counting property

The first layer assigns zero scores and averages the b indicator, storing $r = y/(n+1)$ while preserving symbol identity. In the second layer, the query is r, the key is the c indicator, and the value is $+1$ for a, -1 for c, and zero otherwise. The end-marker output is

$$F(x, y, z) = \frac{x - ze^{y/(n+1)}}{x + y + 1 + ze^{y/(n+1)}}. \tag{3}$$

Its positive denominator makes the sign exactly the required decision. One scalar head per layer and four state coordinates suffice (Figure 2, left).

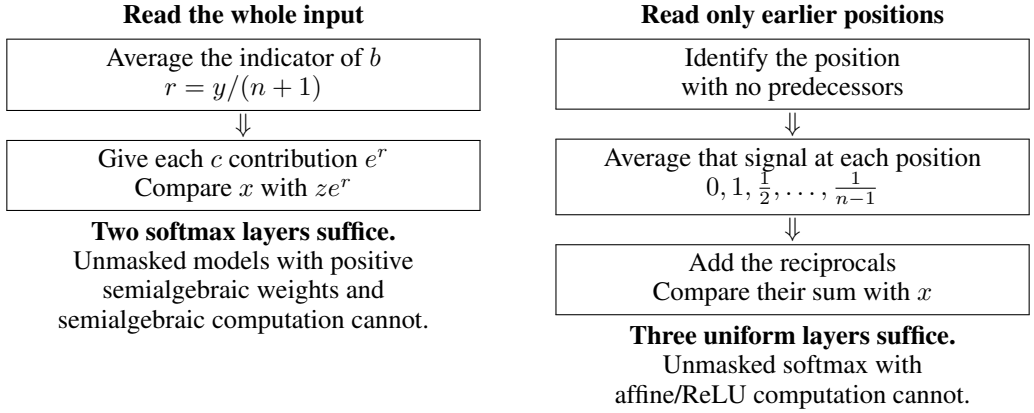

Figure 2: Two counting separations (Theorems 1 and 3). All models have fixed finite structure and no positional encodings. Left: x, y, z count a, b, c, and $n = x + y + z$. Right: x counts a in a binary word of length n; the reciprocal sum is H_{n-1}. Arrows show computational steps.

Theorem 1 (Softmax strictly exceeds AHAT counting)**.** *The counting property* (1) *is not semialgebraic, even with real coefficients. The two-layer softmax network* (3) *recognizes it both without masking and with self-inclusive causal masking. Consequently, without positional encodings or masking, the counting properties recognized by rational-parameter AHAT form a strict subclass of those recognized by rational-parameter softmax with ReLU updates.*

Under self-inclusive causality, the end marker still averages all $n + 1$ positions in both layers. The first layer therefore gives it the same r. The second-layer source keys and values use only the preserved symbol coordinates, so the different prefix averages at ordinary positions do not change the end-marker output. Appendix A.1 verifies a rational-parameter ReLU realization of the same language, matching the model of Sälzer et al. (2026). Uniform attention supplies the inclusion in Theorem 1; the nonsemialgebraic property makes it strict.

Every fixed unmasked network with positive semialgebraic weight functions and semialgebraic updates and readout recognizes only semialgebraic counting properties. Following the type reduction of Sälzer et al. (2026, Proposition 4.1), all occurrences of the same symbol start with the same state and read the same collection of states. They therefore remain identical at every layer. The computation can be grouped into four type states, with multiplicities $x, y, z, 1$.

For example, in a single head, write v_τ for the value at type τ and $w_{\sigma\tau}$ for the positive weight from query type σ to key type τ. Its output is

$$\frac{xw_{\sigma a}v_a + yw_{\sigma b}v_b + zw_{\sigma c}v_c + w_{\sigma \$}v_\$}{xw_{\sigma a} + yw_{\sigma b} + zw_{\sigma c} + w_{\sigma \$}}.$$

The values and weights can already depend on the counts through previous layers. Semialgebraic functions are closed under this arithmetic and under composition (Bochnak et al., 1998). This extends the type reduction to positive semialgebraic weight functions and semialgebraic updates (Appendix A.4). The final positivity test gives a semialgebraic decision in the counts. Increasing the input length changes the multiplicities in this fixed expression, which is why the obstruction applies to every fixed depth and width.

To rule out a semialgebraic description of the integer labels, we examine their limiting boundary in token frequencies. Multiplying a positive count vector (x, y, z) by an integer k preserves its frequencies and makes the single end marker negligible. Varying the count vector supplies a dense set of positive rational frequencies. Along each such sequence, agreement on all finite inputs constrains which side of the limiting boundary is accepted. This connects the discrete recognition problem to a continuous curve.

The numerator, divided by k, tends to

$$x - z \exp\left(\frac{ky}{kn+1}\right) \longrightarrow x - ze^{y/n}.$$

If this limit is nonzero, its sign determines the label for all sufficiently large k. With $t = y/n$ and $s = z/n$, the eventual target label is therefore determined by the side of

$$s = g(t) := \frac{1-t}{1+e^t} \tag{4}$$

on which those proportions lie. Along each ray, meaning the multiples kv of a fixed count vector v, a polynomial's eventual sign is determined by its highest-degree homogeneous part. If all the leading polynomials in a hypothetical semialgebraic description were nonzero near a boundary point, their signs would remain fixed there. Nearby rational proportions on opposite sides would then receive the same eventual label, contradicting (1). Hence the finitely many leading zero sets cover the boundary, and their product gives a nonzero polynomial relation for g. But $e^t = (1-t)/g(t) - 1$ would then satisfy a polynomial relation on an interval. To see the obstruction, such a relation would take the form

$$\sum_{j=0}^{D} r_j(t) e^{jt} = 0, \qquad r_D \not\equiv 0,$$

with polynomial coefficients. Analyticity extends the identity to all real t. Dividing by e^{Dt} and letting t grow would force the nonzero polynomial $r_D(t)$ to tend to zero, because every other term decays exponentially. Appendix A gives the full integer-ray argument and this contradiction.

3.2 Which weight functions add counting power?

The example suggests a general question: which weight functions can produce counting properties beyond finite polynomial conditions? Fix a closed interval $J = [\alpha, \beta]$ with $\alpha < \beta$, and require every reachable score to lie in J. This compares weight functions on the scores the networks actually use.

Theorem 2 (A criterion for the weight function)**.** *Let $\kappa : J \to (0, \infty)$ be continuous, and require all reachable scores to lie in J. If κ is semialgebraic, every unmasked network using κ with semialgebraic updates and readout recognizes only semialgebraic counting properties. If κ is non-semialgebraic, there exists a two-layer unmasked network using κ with affine updates and readout whose counting property is nonsemialgebraic.*

The criterion concerns the restriction of the weight function to J. For example, a function that agrees with a polynomial throughout J remains subject to the semialgebraic upper bound, regardless of its values outside J.

In the construction, varying the frequency of b sweeps the score across the interval, while the ratio of the other two counts probes the resulting attention weight. This is how the counting property exposes the weight function's behavior.

The first statement follows from the type reduction above. For the other, use the first layer to average the b indicator and the second to compare $x\kappa(\alpha)$ with $z\kappa(\alpha + (\beta - \alpha)y/(n+1))$. In the frequency limit the boundary is

$$g_\kappa(t) = \frac{1-t}{1+k(t)}, \qquad k(t) = \frac{\kappa(\alpha + (\beta - \alpha)t)}{\kappa(\alpha)}.$$

Knowing this boundary recovers the relative weight through $k(t) = (1-t)/g_\kappa(t) - 1$, just as (4) recovers e^t. A semialgebraic integer decision would force that boundary to satisfy a polynomial equation. Continuity then places it on finitely many algebraic branches, making the weight function semialgebraic as well. Appendix A.5 gives the interval construction and both implications.

A direct consequence is that every unmasked network with positive polynomial weights and semialgebraic updates and readout recognizes only semialgebraic counting properties. No such fixed network, of any finite size, recognizes (1). Section 5 quantifies what happens when only the weight function of a given softmax network is replaced.

4 Causal masking as a source of counting power

The weight-function criterion compares networks in which every position reads the whole input. With strict causality, prefix lengths vary across positions, so even an ordinary average can produce

different intermediate states. We use this difference to decide another counting property. For a binary word of length n with x occurrences of a, put $H_0 = 0$ and $H_m = \sum_{j=1}^{m} 1/j$.

Theorem 3 (Counting with prefix averages)**.** *For nonempty inputs, without positional encodings or a beginning marker, three uniform-attention layers with strict causal masking and affine updates have end-marker output $(x - H_{n-1})/n$. The recognized property $x > H_{n-1}$ is not semialgebraic. More generally, if $h(n) \to \infty$ and $h(n) = n^{o(1)}$, no unmasked softmax network with affine/ReLU updates and readout recognizes $x > h(n)$.*

The first layer averages the constant one. Its output is zero at the first position and one everywhere else. Subtracting from one identifies the first position. The next layer averages this indicator, producing

$$s_1 = 0, \qquad s_i = \frac{1}{i-1} \quad (i \geq 2).$$

For example, position four reads three predecessors, exactly one of which carries the indicator, and therefore receives $1/3$. Finally, the end marker averages the a indicator minus s_i over the ordinary positions:

$$\frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{1}\{w_i = a\} - s_i\big) = \frac{x - H_{n-1}}{n}.$$

Intermediate states use the order of positions; the final sum depends only on the counts. All scores are zero, so every positive weight function gives this same computation. A reciprocal-position signal is already known when a beginning marker is present (Kazemnejad et al., 2023, Appendix C.1); here an empty prefix supplies the mark, and aggregating the signal yields a counting separation.

For the exclusion, take adjacent counts $x_n^- = \lfloor h(n) \rfloor$ and $x_n^+ = x_n^- + 1$. They require opposite labels. Group an unmasked network by type and write $p = x/(n+1)$ and $q = 1/(n+1)$ for the a and end-marker masses. Both paths approach $(p, q) = (0, 0)$. On any fixed ReLU branch, the type computation is analytic near this corner: the remaining token type has limiting mass one, keeping each softmax denominator positive.

The key is to show that the two paths eventually give the same sign at every analytic test. Their counts tend to infinity more slowly than every positive power of n. This separation of scales lets the first nonzero Taylor terms determine the sign on both paths, even when the test tends to zero.

An analytic test cannot distinguish these two paths eventually. A simple example is the affine test $p - 2q = q(x - 2)$. Both neighboring counts eventually exceed two, so the test has the same positive sign on both. The general argument uses the first nonzero terms of a convergent Taylor series in place of this affine expression. Substituting $p = qx$, the terms of smallest nonzero total degree become

$$A(qx, q) = q^d P_d(x) + \text{higher-order terms in } q, \qquad P_d \not\equiv 0.$$

Since $x_n^{\pm}$ tend to infinity but grow more slowly than every positive power of n, the leading coefficient of P_d determines the sign on both paths; higher-order terms vanish relative to it. A test that is identically zero also has the same outcome on both paths. Applying this observation at each ReLU selects the same eventual branch on both paths. The final readout therefore gives the same label, a contradiction. Appendix B proves the remainder estimate and branch induction. The same sign argument applies to finite polynomial tests, proving nonsemialgebraicity.

5 What polynomial replacements preserve

Section 3 excludes every fixed unmasked polynomial-weight network from recognizing (1) at all lengths. We now fix a softmax network and replace every exponential e^u by the same positive polynomial $P(u)$, retaining its architecture, parameters and the same attention masks at each length.

We compare output accuracy, exact decisions and average disagreement. Positive scaling cancels in normalized attention, so we use $P(0) = 1$ whenever $P(0) > 0$.

5.1 Outputs: one polynomial for all lengths

Proposition 4 (Output approximation at every length)**.** *Fix a finite softmax network with initial states uniformly bounded in norm at each position, affine query/key/value maps, affine/ReLU updates, and an affine end-marker readout, under any fixed attention masks. For every $\eta > 0$, a polynomial P with rational coefficients, positive on all of $\mathbb{R}$, satisfies*

$$\sup_{n,w:\,|w|=n} |f_P(w) - f_{\exp}(w)| \le \eta, \qquad \deg P \le (1 + o(1))\frac{\log(1/\eta)}{\log\log(1/\eta)}. \tag{5}$$

The asymptotic holds as $\eta \downarrow 0$ with the network fixed. For the two-layer network (3)*, the least replacement degree is asymptotic to the expression on the right.*

Why the error does not accumulate with length. A positive-weight head returns a convex combination of value vectors, so bounded incoming states bound its output independently of the weight function and number of positions. Applying the fixed pointwise update and inducting over layers gives a common bounded domain for both executions, including states changed by earlier substitutions. All scores therefore lie in a fixed compact interval before the polynomial is chosen.

Hold one row's scores s_j and values fixed. If p_j and $\widehat{p}_j$ denote its exponential and polynomial attention weights, then

$$r_j = \frac{P(s_j)}{e^{s_j}}, \qquad \widehat{p}_j = \frac{p_j r_j}{\sum_k p_k r_k}. \tag{6}$$

The denominator is a weighted average of the relative changes, since $\sum_k p_k = 1$. Normalization averages these changes rather than accumulating one error per position. If $|\log P(u) - u| \le \delta$, the normalized weights have total variation at most $\tanh(\delta/2)$ (Cohen & Fausti, 2023), independently of row length. For values bounded by V, the row-output error is at most $2V \tanh(\delta/2) \le V\delta$.

Across layers, error has two parts. Changing the weights at the states reached by the substituted network incurs the row error bounded above. Errors already present in those states are controlled by the original softmax block's Lipschitz constant, using the maximum norm over positions on the common domain.

Thus the largest state error after layer ℓ obeys

$$E_0 = 0, \qquad E_\ell \le a_\ell E_{\ell-1} + b_\ell \delta, \tag{7}$$

with constants independent of length. Fixed depth gives output error at most $C_f\delta$. On the common score interval $[-B, B]$, an even Taylor polynomial $T_d(u) = \sum_{j=0}^{d} u^j/j!$ is globally positive and has error at most $e^B B^{d+1}/(d+1)!$. Since $e^u \ge e^{-B}$ there, this also controls log-weight error. The factorial yields $d\log d + O(d)$ after taking logarithms, giving the sufficient degree. Appendix C supplies the constants and extends the argument to locally Lipschitz updates.

Why output accuracy forces this degree. Normalize $P(0) = 1$ and take limits of integer proportions in (3). At $t = y/n$ and $z/n = (1-t)/(1+P(t))$, the polynomial output is zero and the exponential output is

$$\frac{(1-t)(P(t) - e^t)}{t + P(t) + (1-t)e^t}. \tag{8}$$

For $0 \le t \le 1/2$, output error at most η therefore forces $|P(t) - e^t| \le C\eta$ for small η. An order-$(d+1)$ finite difference annihilates P but remains positive on the exponential, giving $\log(1/\eta) \le (d+1)\log(d+1) + O(d+1)$ and the necessary degree (Appendix C.6).

The scalar degree law on a fixed interval is classical (Aggarwal & Alman, 2022); polynomial approximation also underlies bounded-entry attention algorithms (Alman & Song, 2023). Here the guarantee controls states changed by substitution at every length. Recovering weight accuracy from the output shows that the same degree is necessary for our two-layer network. Appendix C.7 treats score range and tolerance jointly.

5.2 DECISIONS: THE DEGREE MUST GROW WITH LENGTH

Close outputs can lie on opposite sides of zero. On a family of inputs whose original outputs all have magnitude at least $\gamma > 0$, Proposition 4 with $\eta < \gamma$ would preserve every decision. Our counting example has no such margin over all lengths: integer proportions approach its limiting boundary, where the output vanishes. This observation explains why an output guarantee alone is insufficient; the following theorem proves that exact agreement forces the degree to grow. We require the decisions of the same two-layer network (3) to agree on every positive count triple at a target length, allowing a new polynomial for each length.

Theorem 5 (Degree needed for exact decisions)**.** *For each $n \geq 3$, let $P_n(0) = 1$ and $P_n > 0$ on $[0, 1]$. Retain the two-layer network* (3) *and replace its exponentials by P_n. If its decisions agree with the original for every $x, y, z \geq 1$ with $x + y + z = n$, then*

$$\deg P_n \geq (1 - o(1)) \frac{\log n}{\log \log n} \qquad (n \to \infty). \tag{9}$$

Fix n, y and put $u_y = y/(n+1)$ and $M = n - y$. Since $x = M - z$, the exponential and polynomial rules accept, respectively, when

$$z < B_E(y) := \frac{M}{1 + e^{u_y}}, \qquad z < B_P(y) := \frac{M}{1 + P_n(u_y)}.$$

Both cutoffs lie in $(0, M)$. Among $z = 1, \ldots, M - 1$, a cutoff B accepts exactly $\lceil B \rceil - 1$ values. Identical labels therefore force equal ceilings and $|B_P(y) - B_E(y)| < 1$, including ties. At the actual grid points $u_y \in [1/4, 1/2]$, division by M and inversion give $|P_n(u_y) - e^{u_y}| \leq C_0/n$, with an absolute constant.

An equally spaced finite difference converts this grid accuracy into a degree bound. Put $m = \deg P_n + 1$. Unless m is already of order n, the grid contains $m + 1$ equally spaced points in this interval with spacing $h \geq 1/(16m)$. The difference annihilates P_n, so

$$e^{1/4}(16m)^{-m} \leq e^{u_0}(e^h - 1)^m \leq 2^m C_0/n.$$

Taking logarithms yields $\log n \leq m \log m + O(m)$ and the claimed degree. Appendix D.1 gives the grid choice and inversion constants. Thus a fixed output tolerance is met by one polynomial at all lengths, whereas exact decisions force its degree to grow.

5.3 SMALL AVERAGE DISAGREEMENT AND AN EXACT FAILURE

Average agreement gives a different picture. Sample uniformly from the $(n-1)(n-2)/2$ positive count triples of length n, and let $R_n(P)$ be the fraction of disagreeing decisions. This samples count triples, not strings. With $t = y/n$ and $s = z/n$, the limiting decision boundaries give

$$g(t) = \frac{1 - t}{1 + e^t}, \qquad g_P(t) = \frac{1 - t}{1 + P(t)}, \qquad R(P) = 2 \int_0^1 |g_P(t) - g(t)|\, dt. \tag{10}$$

The factor two normalizes the area of the frequency triangle. For $P(0) = 1$ and $P > 0$ on $[0, 1]$, $R_n(P)$ converges to $R(P)$ uniformly at any fixed degree (Appendix D.4).

The degree-ten Taylor polynomial has limiting disagreement $R(T_{10}) \approx 3.086 \times 10^{-11}$, with a certified enclosure in Appendix E. Yet it changes the label at $(x, y, z) = (74, 976, 30)$, of length 1080. With $u = 976/1081$, exact rational arithmetic verifies

$$T_{10}(u) < \frac{37}{15} < T_{12}(u) < e^u. \tag{11}$$

Since $37/15 = x/z$, the replacement accepts and softmax rejects. Thus even this small limiting average disagreement does not guarantee agreement on every input. Appendix D gives the optimal degree for limiting risk and a coefficient-independent finite-length estimate.

6 Related work

Hard attention, soft attention, and counting. Formal-language analyses establish both limitations of self-attention (Hahn, 2020) and characterizations for hard-attention models (Hao et al., 2022; Barceló et al., 2024). Yang et al. (2026) give simulations using unbounded positional embeddings or temperature scaling; for general AHAT, the temperature depends on input length and attention-score gaps. Yang & Chiang (2024) show that C-RASP and temporal counting logic have the same expressive power and compile them into causally masked softmax transformers. C-RASP counts positions satisfying predicates and compares linear combinations of such counts. Its order-independent counting properties are limited to semilinear conditions (Sälzer et al., 2026, Proposition 5.1). Sälzer et al. (2026) characterize unmasked AHAT counting by rational-coefficient semialgebraic properties and obtain their softmax realization through uniform attention. We show that softmax recognizes strictly more and extend their type reduction to a criterion for positive weight functions. Jiang et al. (2026) study softmax computation with generated intermediate tokens; our results concern a single forward pass.

Masking and positional information. Causal models without positional encodings can learn positional information (Haviv et al., 2022), and attention-variance analyses explain one source of it (Chi et al., 2023). With a beginning marker, causal attention can construct a reciprocal-position signal (Kazemnejad et al., 2023). Our strict causal construction obtains the mark from an empty prefix and aggregates reciprocals into an order-independent decision. Huang et al. (2025) study length generalization under an idealized training and inference framework with positional encodings; we compare fixed networks at all lengths.

Approximation and attention resources. Different replacements preserve different objects. Performers use random features (Choromanski et al., 2021), and polynomial attention can be sketched (Kacham et al., 2024). Secure inference motivates polynomial transformer computations (Zimerman et al., 2024; Li et al., 2024), while learned linear attention uses output matching and adaptation (Zhang et al., 2024; 2025). Amsel et al. (2025) establish approximation separations involving attention rank and head count. We keep those resources fixed and vary the scalar weight function, measuring the degree needed to preserve outputs or decisions. Requiring one network to work at every size also underlies Rosenbluth et al. (2024). Exponential approximation (Aggarwal & Alman, 2022; Alman & Song, 2023), normalization perturbation bounds (Cohen & Fausti, 2023), and attention sensitivity (Kim et al., 2021; Castin et al., 2024) provide related tools. The distinction between numerical confidence and correct labels is also central to Chiang & Cholak (2022).

7 Conclusion

Two softmax layers recognize a counting property beyond the polynomial conditions that characterize unmasked AHAT. With unmasked, otherwise semialgebraic computation, the weight-function criterion identifies exactly when such a separation is possible on a fixed interval containing the attention scores. Strict causal uniform attention provides an independent counting separation. For polynomial substitution in the same softmax network, output accuracy at a fixed tolerance can hold at every length, but preserving all decisions requires degree growing at least as $\log n / \log \log n$.

Our results assume exact real arithmetic and no positional encodings. The effect of finite precision on these separations remains open. We prove a lower bound for exact finite-length decisions; whether degree $O(\log n / \log \log n)$ suffices is also open. A further question is whether trained networks learn the exponential counting property and generalize to longer inputs.

AI USE STATEMENT

AI assistants were used to aid or polish writing, for retrieval and discovery of related work, and for research ideation or execution.

REPRODUCIBILITY STATEMENT

Appendices A–E give the constructions and complete proofs. Appendix E specifies the rational inequalities and the certified integral enclosure; the accompanying verification script reproduces these calculations.

REFERENCES

Amol Aggarwal and Josh Alman. Optimal-degree polynomial approximations for exponentials and Gaussian kernel density estimation. In *Computational Complexity Conference*, volume 234 of *LIPIcs*, pp. 22:1–22:23, 2022. doi: 10.4230/LIPIcs.CCC.2022.22.

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

Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over quantity in attention layers: When adding more heads hurts. In *International Conference on Learning Representations*, 2025.

Pablo Barceló, Alexander Kozachinskiy, Anthony W. Lin, and Vladimir Podolskii. Logical languages accepted by transformer encoders with hard attention. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gbrHZq07mq.

Edward Bierstone and Pierre D. Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS*, 67:5–42, 1988. doi: 10.1007/BF02699126.

Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real Algebraic Geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 1998. doi: 10.1007/978-3-662-03718-8.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840, 2024.

Ta-Chung Chi, Ting-Han Fan, Li-Wei Chen, Alexander I. Rudnicky, and Peter J. Ramadge. Latent positional information is in the self-attention variance of transformer language models without positional embeddings. In *Proceedings of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2305.13571.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the Association for Computational Linguistics*, 2022. URL https://aclanthology.org/2022.acl-long.527/.

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 *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.14794.

Samuel N. Cohen and Eliana Fausti. Hyperbolic contractivity and the Hilbert metric on probability measures. arXiv:2309.02413, 2023. Version 2, 2024.

Carl de Boor. Divided differences. *Surveys in Approximation Theory*, 1:46–69, 2005. URL https://arxiv.org/abs/math/0502036.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306.

Yiding Hao, Dana Angluin, and Robert Frank. Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810, 2022. doi: 10.1162/tacl_a_00490.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022. URL https://aclanthology.org/2022.findings-emnlp.99/.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02140.

Hongjian Jiang, Michael Hahn, Georg Zetzsche, and Anthony W. Lin. Softmax transformers are Turing-complete. In *International Conference on Learning Representations*, 2026.

Praneeth Kacham, Vahab Mirrokni, and Peilin Zhong. PolySketchFormer: Fast transformers via sketching polynomial kernels. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22748–22770, 2024.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571, 2021.

Zhengyi Li, Kang Yang, Jin Tan, Wen jie Lu, Haoqi Wu, Xiao Wang, Yu Yu, Derun Zhao, Yancheng Zheng, Minyi Guo, and Jingwen Leng. Nimbus: Secure and efficient two-party inference for transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2411.15707.

Eran Rosenbluth, Jan Tönshoff, Martin Ritzert, Berke Kisin, and Martin Grohe. Distinguished in uniform: Self-attention vs. virtual nodes. In *International Conference on Learning Representations*, 2024.

Marco Sälzer, Chris Köcher, Alexander Kozachinskiy, Georg Zetzsche, and Anthony Widjaja Lin. The counting power of transformers. In *International Conference on Learning Representations*, 2026.

Andy Yang and David Chiang. Counting like transformers: Compiling temporal counting logic into softmax transformers. arXiv:2404.04393, 2024. URL https://arxiv.org/abs/2404.04393.

Andy Yang, Lena Strobl, David Chiang, and Dana Angluin. Simulating hard attention using soft attention. *Transactions of the Association for Computational Linguistics*, 14:147–166, 2026. doi: 10.1162/tacl.a.597.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *International Conference on Learning Representations*, 2024.

Michael Zhang, Simran Arora, Rahul Chalamala, Alan Wu, Benjamin Spector, Aaryan Singhal, Krithik Ramesh, and Christopher Ré. LoLCATs: On low-rank linearizing of large language models. In *International Conference on Learning Representations*, 2025.

I. Zimerman, M. Baruch, N. Drucker, G. Ezov, O. Soceanu, and L. Wolf. Converting transformers to polynomial form for secure inference over homomorphic encryption. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62803–62814, 2024. URL https://proceedings.mlr.press/v235/zimerman24a.html.

A EXACT COUNTING WITH UNMASKED ATTENTION

This appendix gives the coordinate construction, proves the integer-count separation in Theorem 1, and establishes the weight function criterion in Theorem 2. All networks use the input and decision conventions of Section 2.

A.1 A FOUR-COORDINATE IMPLEMENTATION

The unmasked model keeps one state per input position and one appended end marker. Parameters are arbitrary fixed real numbers. A semialgebraic counting property is a set $L \subseteq \mathbb{N}^m$ of the form $S \cap \mathbb{N}^m$, where $S \subseteq \mathbb{R}^m$ is semialgebraic and $\mathbb{N} = \{0, 1, \ldots\}$. Function graphs and sets may use real coefficients. We use closure under Boolean operations, composition, and projection, the last being the Tarski–Seidenberg theorem (Bochnak et al., 1998).

For the exponential example, embed $a, b, c, \$$ as

$$(1,0,0,0), \quad (0,1,0,0), \quad (0,0,1,0), \quad (0,0,0,0),$$

respectively, and write a state as (A, B, C, R). The first layer uses $Q_1 = K_1 = 0$, $V_1 = B$, and the update $(A, B, C, R) \mapsto (A, B, C, o)$. At every position it stores $R = y/(n+1)$. The second uses

$$Q_2 = R, \qquad K_2 = C, \qquad V_2 = A - C$$

and the same form of update; the readout selects the last coordinate. The a sources contribute $+1$ with weight one, the c sources contribute -1 with weight $e^{y/(n+1)}$, and all remaining values are zero. The denominator includes $x + y + 1$ unit weights and z exponential weights, proving (3). All scores lie in $[0, 1]$, all affine coefficients and embeddings lie in $\{-1, 0, 1\}$, and each layer has one scalar head. The empty input has only a zero-valued end marker and gives zero output.

Under self-inclusive causality, ordinary positions may store different prefix averages in R, while the end marker stores $y/(n+1)$. The second-layer source maps $K_2 = C$ and $V_2 = A - C$ do not read R. The end-marker query and its attended source values and keys are therefore identical to the unmasked execution. Its output is again (3), with no beginning marker or positional encodings. The nonsemialgebraicity proved below applies to this same accepted count set.

This language is also recognized using rational-parameter ReLU updates, as in Sälzer et al. (2026). The first-layer coordinates A, B, C, R are nonnegative, so coordinatewise ReLU preserves them. Apply ReLU to the final attention output: $\mathrm{ReLU}(F) > 0$ exactly when $F > 0$. This gives the claimed language in that model. The output-approximation results use the signed affine readout F of (3).

For any positive weight function P, the first layer remains uniform. The substituted output is

$$F_P(x, y, z) = \frac{xP(0) - zP(y/(n+1))}{(x + y + 1)P(0) + zP(y/(n+1))}. \tag{12}$$

A common positive scaling of P cancels. We therefore use $P(0) = 1$ in all subsequent replacement comparisons.

A.2 FROM INTEGER LABELS TO A CONTINUOUS BOUNDARY

A continuous boundary alone does not determine which integer labels admit a semialgebraic description. The next lemma derives an algebraic constraint from eventual labels on integer rays. For this subsection, let t and s be the frequencies of b and c, and set

$$\mathcal{D} = \{(t, s) : t > 0,\ s > 0,\ t + s < 1\}, \qquad p(t, s) = (1 - t - s, t, s).$$

The vector $p(t, s)$ lists the three positive type frequencies. We use real frequencies only to analyze the integer language; the network's inputs remain finite words.

Lemma 6 (Boundary forced by eventual integer-ray labels)**.** *Let $g : (0, 1) \to \mathbb{R}$ be continuous with $0 < g(t) < 1 - t$, and let $L \subseteq \mathbb{N}^3_{>0}$. Suppose that for every rational $(t, s) \in \mathcal{D}$ with $s \neq g(t)$, and every positive integer vector v with $v/\sum_i v_i = p(t, s)$, the label of kv is, for all sufficiently large integers k, equal to $\mathbf{1}\{s < g(t)\}$. If L is the integer trace of a real-semialgebraic set, there is a nonzero polynomial $Q \in \mathbb{R}[t, s]$ such that*

$$Q(t, g(t)) = 0 \qquad (0 < t < 1).$$

Proof. Let S be a hypothetical semialgebraic representative of L. Express membership in S as a Boolean function of sign conditions on finitely many nonzero polynomials $q_1, \ldots, q_r$ in the three count variables. Identically zero and constant conditions can be simplified away. If no polynomial remains, membership is constant, contradicting the two required labels on rational directions above and below g.

For each remaining polynomial let H_j be its highest-degree homogeneous part and let d_j be its degree. At a direction p with $H_j(p) \neq 0$,

$$q_j(\lambda p) = \lambda^{d_j} H_j(p) + O(\lambda^{d_j - 1}) \qquad (\lambda \to +\infty).$$

Thus its eventual sign along that direction is the sign of $H_j(p)$.

Suppose a boundary point $(t_0, g(t_0))$ avoids every zero set $H_j(p(t,s)) = 0$. There is a connected open neighborhood of that point inside $\mathcal{D}$ on which all these signs are constant. The Boolean formula defining S therefore has the same eventual membership value on every ray in this neighborhood. Continuity of g implies that the neighborhood contains nonempty open subsets on both sides of its graph. Choose one rational point on each side and scale each direction to a positive integer vector. Their prescribed eventual labels are opposite, a contradiction. Each chosen ray may have its own threshold; no uniform threshold over directions is required.

Consequently the graph of g is contained in the union of the zero sets of $h_j(t,s) = H_j(1-t-s, t, s)$. Each h_j is a nonzero polynomial. Indeed, if it vanished identically, homogeneity would imply that H_j vanished throughout the open positive cone in $\mathbb{R}^3$, and hence identically. The product $Q = \prod_{j=1}^{r} h_j$ is nonzero and vanishes on the entire graph, as required. □

A.3 The exponential boundary has no polynomial relation

Lemma 7 (An elementary exponential obstruction)**.** *No nonzero polynomial $R \in \mathbb{R}[t, Y]$ satisfies $R(t, e^t) = 0$ on a nonempty open interval. Consequently, no nonzero polynomial $Q \in \mathbb{R}[t, s]$ vanishes on the graph of*

$$g(t) = \frac{1-t}{1+e^t}, \qquad 0 < t < 1. \tag{13}$$

Proof. Suppose $R(t, Y) = \sum_{j=0}^{d} r_j(t) Y^j$ with r_d a nonzero polynomial. The function $R(t, e^t)$ is real analytic on $\mathbb{R}$. If it vanishes on an open interval, the real-analytic identity theorem implies that it vanishes on all of $\mathbb{R}$. Dividing by e^{dt} then gives

$$r_d(t) = -\sum_{j<d} r_j(t) e^{-(d-j)t} \longrightarrow 0 \qquad (t \to +\infty),$$

which is impossible for a nonzero polynomial.

Now write a hypothetical boundary polynomial as $Q(t,s) = \sum_{j=0}^{d} q_j(t) s^j$, with $q_d \not\equiv 0$. Clearing the denominator in $Q(t, g(t))$ gives $R(t, e^t) = 0$, where

$$R(t, Y) = \sum_{j=0}^{d} q_j(t)(1-t)^j (1+Y)^{d-j}.$$

This is not the zero polynomial, since $R(t, -1) = q_d(t)(1-t)^d \not\equiv 0$. The first part of the proof therefore supplies the contradiction. □

A.4 Unmasked attention: arbitrary fixed-depth closure and separation

Lemma 8 (Finite-type recursion)**.** *Every fixed unmasked network with positive semialgebraic weight functions and semialgebraic updates and readout has an integer counting language represented by a semialgebraic set on the nonnegative real count cone.*

Proof. We adapt the type reduction of Sälzer et al. (2026, Proposition 4.1) to positive semialgebraic weight functions and semialgebraic updates. At initialization all occurrences of a type σ have one

fixed state $h_\sigma^{(0)}$. If this is true at layer $\ell - 1$, every occurrence has the same query and reads the same multiset of keys and values. Hence it remains true at layer ℓ. Include the end marker type with multiplicity $c_\$ = 1$. A head on type σ computes

$$\frac{\sum_{\tau \in \Sigma \cup \{\$\}} c_\tau \kappa(\langle Q h_\sigma, K h_\tau \rangle) V h_\tau}{\sum_{\tau \in \Sigma \cup \{\$\}} c_\tau \kappa(\langle Q h_\sigma, K h_\tau \rangle)}.$$

Starting with constant states, semialgebraic composition and division by a nonzero denominator show inductively that each type state has a semialgebraic real-count representative. Finitely many heads and semialgebraic pointwise maps preserve this representation. The final scalar positivity condition is semialgebraic and agrees with the network on integer inputs. The claim covers each fixed finite architecture separately, allowing arbitrary real parameters.

A semialgebraic map specified on a semialgebraic domain can be extended by zero outside it when forming a real-count representative. An attention quotient can likewise be assigned zero where its denominator is undefined or nonpositive. These extensions are semialgebraic and preserve every valid integer-input execution. This handles partial domains without imposing validity on hypothetical real-count executions. □

Why one attention layer is insufficient. With fixed type embeddings, all first-layer scores and weight function values are constants, even for a nonsemialgebraic weight function. Each head is therefore a rational function of the counts, so semialgebraic updates and readout still give a semialgebraic decision.

Proof of Theorem 1. The construction was verified in Appendix A.1. Fix a positive integer vector (v_a, v_b, v_c) of total m. For its multiple kv, the numerator of (3), divided by k, tends to

$$v_a - v_c e^{v_b/m}.$$

Whenever this limit is nonzero, the finite-length label eventually agrees with its sign. The limiting frequency boundary is (4). Lemma 6 would force a polynomial relation for this boundary if the integer language were semialgebraic, contradicting Lemma 7. The finite-type recursion excludes every fixed finite unmasked network with positive semialgebraic weight functions and semialgebraic updates and readout. This proves both the integer-language claim and the architecture comparison. □

A.5 A CRITERION FOR CONTINUOUS POSITIVE WEIGHT FUNCTIONS

Fix $\alpha < \beta$ and a continuous function $\kappa : [\alpha, \beta] \to (0, \infty)$. Define

$$k(t) = \frac{\kappa(\alpha + (\beta - \alpha)t)}{\kappa(\alpha)}, \qquad g_\kappa(t) = \frac{1 - t}{1 + k(t)}, \qquad 0 < t < 1.$$

Proposition 9 (The equivalence used in Theorem 2)**.** *The marked integer language $x > zk(y/(n+1))$ is semialgebraic on positive counts if and only if k is semialgebraic on $(0, 1)$.*

Proof. First realize the marked language. Give layer one the constant score α and the b indicator as its value; its average R is $y/(n+1)$ at every position. Layer two uses query $(1, R)$. Its key is $(\alpha, \beta - \alpha)$ at c and $(\alpha, 0)$ at all other types. Values are $1, -1, 0$ at a, c, and the remaining types, respectively. Its numerator is $x\kappa(\alpha) - z\kappa(\alpha + (\beta - \alpha)R)$. Every score lies in the prescribed interval, including all first-layer scores and scores at ordinary positions. Retaining symbol coordinates and storing each head output uses only affine updates and an affine readout.

If k is semialgebraic, direct substitution into $x > zk(y/(n+1))$ supplies a semialgebraic representative. Conversely, scale any positive integer count direction by larger integers. The marked frequency tends to its ordinary frequency $t = y/n$. Away from $s = g_\kappa(t)$, continuity and positivity imply eventual labels on the corresponding side of this graph. Lemma 6 therefore supplies a nonzero polynomial Q with $Q(t, g_\kappa(t)) = 0$ throughout $(0, 1)$.

By cylindrical algebraic decomposition (Bochnak et al., 1998), outside finitely many vertical lines the zero set of Q over each remaining interval consists of disjoint ordered continuous semialgebraic

graphs $\xi_1 < \cdots < \xi_k$. The sets $\{t : g_\kappa(t) = \xi_j(t)\}$ are pairwise disjoint, closed in the interval, and cover it. Connectedness forces exactly one to be nonempty. Adding the finitely many exceptional graph points shows that g_κ is semialgebraic, and inversion

$$k(t) = \frac{1-t}{g_\kappa(t)} - 1$$

proves necessity.

The affine change of argument transfers semialgebraicity between k and κ on the open interval. Adding the two endpoint values preserves semialgebraicity, so the same conclusion holds on $[\alpha, \beta]$.

Finally, a semialgebraic weight function gives semialgebraic decisions at every fixed depth by the finite-type recursion. A nonsemialgebraic continuous weight function gives the two-layer marked counterexample just established. Adding endpoint values, or a positive constant extension outside the interval when a globally defined weight function is desired, does not change the criterion. □

B Uniform causal counting and its unmasked obstruction

This appendix proves Theorem 3. We first verify the three averages, then show why both polynomial conditions and unmasked softmax with finite affine/ReLU computation fail to separate the two adjacent counts at the threshold.

B.1 Three affine layers realize the reciprocal-sum threshold

Embed a, b, and the end marker as $(1, 0), (0, 0), (0, 0)$, respectively. Write the state as (A, R). Each layer has $Q = K = 0$ and uses the affine update $(A, R) \mapsto (A, o)$, where o is its scalar attention output. Since only earlier positions are attended, position one returns zero on every empty attention row. The layer-one value is the constant one, giving

$$r_1 = 0, \qquad r_i = 1 \quad (i \geq 2).$$

Layer two uses value $1 - R$, giving

$$s_1 = 0, \qquad s_i = \frac{1}{i-1} \quad (i \geq 2).$$

Layer three uses value $A - R$. Its end-marker output is

$$\frac{1}{n}\sum_{i=1}^{n}(A_i - s_i) = \frac{x - H_{n-1}}{n}.$$

All embeddings and affine coefficients belong to $\{-1, 0, 1\}$. The state update is affine and may retain a signed working coordinate. At $n = 1$, the harmonic sum is empty and the same formula holds. The empty word is rejected by convention. Each nonempty attention row is uniform for every positive weight function, since all its scores are zero. Although intermediate states depend on position, the final label depends only on x and n.

B.2 An analytic sign lemma for two adjacent count sequences

Let $h(n) \to \infty$ with $h(n) = n^{o(1)}$, equivalently $h(n)^K/n \to 0$ for every fixed $K > 0$. Put

$$x_n^- = \lfloor h(n) \rfloor, \qquad x_n^+ = x_n^- + 1.$$

Both give valid positive counts $(x_n^\pm, n - x_n^\pm)$ for all sufficiently large n. The rule $x > h(n)$ rejects the first and accepts the second, including when $h(n)$ is an integer.

Lemma 10. *Let $A(p, q)$ be real analytic near $(0, 0)$, let $q_n = 1/(n+1)$, and let $x_n^\pm$ be as above. Then $A(q_n x_n^-, q_n)$ and $A(q_n x_n^+, q_n)$ eventually have the same sign, where zero is treated as a third sign.*

Proof. If A vanishes identically near the origin, both values are eventually zero. Otherwise let d be the smallest total degree with a nonzero Taylor coefficient. Then

$$A(p, q) = \sum_{i+j=d} a_{ij} p^i q^j + O((|p| + |q|)^{d+1}).$$

Substitution of $p = qx$ gives

$$A(qx, q) = q^d \left[P_d(x) + O(q(1+x)^{d+1})\right], \qquad P_d(X) = \sum_{i+j=d} a_{ij} X^i.$$

Distinct pairs (i, j) of total degree d give distinct powers of X, so P_d is nonzero. Let r and $c \neq 0$ be its degree and leading coefficient. Along either path $P_d(x_n^{\pm}) \sim c(x_n^{\pm})^r$. The remainder divided by this polynomial is

$$O\big(q_n(1 + x_n^{\pm})^{d+1-r}\big) = o(1),$$

since $x_n^{\pm} = n^{o(1)}$ and tend to infinity. Both expressions therefore eventually have the same strict sign, that of c. □

B.3 No finite polynomial description on integer counts

Suppose finitely many polynomial sign tests describe $x > h(n)$, and replace the other count by $n - x$. For a nonzero test $Q(n, x)$ of total degree D, the expression

$$\widetilde{Q}(p, q) = q^D Q(q^{-1} - 1, p/q)$$

is a polynomial: each monomial $n^a x^b$ becomes $q^{D-a-b}(1-q)^a p^b$. At $q = 1/(n+1) > 0$ and $p = qx$ it has the same sign as $Q(n, x)$. Lemma 10 makes every test, including identically zero tests, eventually agree on the two count paths. Any Boolean combination must then agree as well, contradicting their opposite labels. Since $\log n \leq H_{n-1} \leq 1 + \log(n-1)$ for $n \geq 2$, the harmonic threshold satisfies the hypotheses and its integer counting property is not semialgebraic.

B.4 Finite affine/ReLU unmasked softmax cannot separate the paths

Fix any network in the excluded class of Theorem 3. All occurrences of a type share a state, by the type recursion in Appendix A.4. Normalize the type multiplicities as

$$m_a = p, \qquad m_b = 1 - p - q, \qquad m_\$ = q, \qquad p = \frac{x}{n+1}, \quad q = \frac{1}{n+1}.$$

Keep formal states for all three types, including those whose mass tends to zero. For a query type σ a head computes

$$o_\sigma(p, q) = \frac{\sum_\tau m_\tau \exp(u_{\sigma\tau}(p, q)) v_\tau(p, q)}{\sum_\tau m_\tau \exp(u_{\sigma\tau}(p, q))}.$$

We induct through the finite scalar computation graph. Each processed scalar has an analytic expression near $(0, 0)$ that agrees with the actual computation along *both* count paths for all sufficiently large n. Initial type states are constants, so the induction begins.

Affine operations, products in dot-product scores, and exponentials preserve analytic expressions. At an attention denominator, the already defined scores have finite values at the origin. Only type b has nonzero mass there, so the denominator at $(0, 0)$ is $\exp(u_{\sigma b}(0, 0)) > 0$. It remains nonzero on a small neighborhood, allowing analytic division. The formal zero-mass type states remain legitimate queries and contribute zero source mass.

At a ReLU gate, the preactivation has a common analytic expression by induction. Lemma 10 gives the same eventual sign along the two paths. Choose the identity branch if that sign is positive and the zero branch if it is negative. If the expression vanishes near zero, choose zero as well. This supplies an analytic expression for the gate's output that is valid along both paths. Finitely many gates have a common sufficiently large threshold for all these choices.

The affine/ReLU scalar readout is handled in the same way. Lemma 10 now implies that its final outputs eventually have the same sign, and hence the same strict-positive label, on the two paths. Their required labels for $x > h(n)$ are opposite. Taking $h(n) = H_{n-1}$ proves the unmasked exclusion in Theorem 3.

This proof applies to each fixed finite architecture with arbitrary real parameters, so the union over all such architectures is excluded as well. The analytic expressions are chosen along the two paths; the argument does not assume that the entire ReLU network is analytic throughout a neighborhood. The same exclusion holds for every unbounded subpolynomial threshold h above, while the positive causal construction here is specifically for H_{n-1}.

B.5 Analytic operations on compact neighborhoods

Proposition 11. *Consider a fixed finite unmasked network without positional encodings, with affine query/key/value maps and finitely many fixed markers. Suppose its weight functions are positive and analytic on neighborhoods of the relevant compact score intervals. Updates and readout may compose analytic maps and ReLUs, with each analytic map defined on a neighborhood of the common compact domain of its arguments, including limiting states. Then no such network recognizes $x > h(n)$ when $h(n) \to \infty$ and $h(n) = n^{o(1)}$.*

Proof. If the number of markers is M, use $q = 1/(n + M)$ and $p = qx$. Ordinary type masses are $p, 1 - p - Mq$ and each marker has mass q (or a fixed multiple for repeated marker types). The estimate in Lemma 10 still holds because $q(1 + x_n^{\pm})^K \to 0$ for each fixed K. At the limiting corner, a denominator equals the positive weight function value at the score of the ordinary b type. The stated neighborhood assumptions permit composition and division of analytic expressions at all operations. At each ReLU use the same eventual-branch argument as above. Both paths receive the same final label, proving the claim. □

B.6 Beginning markers and reciprocal positions

The reciprocal signal from a beginning marker is a known construction (Kazemnejad et al., 2023, Appendix C.1). Here it gives two further versions of the aggregate counting rule. With a beginning marker and an end marker, let B_i indicate the former and A_i an ordinary a. Under strict causality, first average B over earlier positions. At the jth ordinary position the result is $1/j$, because there are j predecessors including the marker. The beginning position has empty-row output zero. The next layer averages $A_i - r_i$ at the end marker, giving

$$\frac{x - H_n}{n + 1}.$$

With self-inclusive causality, index all positions from 1 to $n + 2$. The first layer averages B over $j \le i$, so $r_i = 1/i$, including at both markers. The end marker in the next layer consequently produces

$$\frac{1}{n + 2} \sum_{i=1}^{n+2} (A_i - r_i) = \frac{x - H_{n+2}}{n + 2}.$$

Both constructions use two coordinates, one uniform head per layer, and affine updates. The exclusion remains valid for unmasked networks given the same two markers: use $q = 1/(n + 2)$, masses $p, 1-p-2q, q, q$, and the corresponding harmonic threshold. These are constructions for their stated marker conventions.

B.7 Continuous semialgebraic maps at the frequency boundary

The analytic argument has an extension to continuous semialgebraic operations. Suppose the finite type recursion extends continuously to a closed nonnegative frequency neighborhood of the corner, its states remain in compact sets, its analytic operations are defined on neighborhoods of the compact argument sets, and weight functions remain positive on the compact score ranges. Under these conditions, adding continuous semialgebraic operations still excludes every unbounded subpolynomial threshold.

To prove this, restrict the frequencies to a small closed semialgebraic neighborhood of the corner within the nonnegative frequency simplex. Introduce one variable for each intermediate scalar in the finite computation. Semialgebraic operations have semianalytic graphs, and each analytic operation has an analytic graph on a neighborhood of its compact argument set. Cover each such argument set by finitely many compact boxes lying inside that neighborhood. Together with these box constraints and the final positivity condition, the computation relations form a relatively compact semianalytic set: continuity bounds all intermediate variables. Its projection onto (p, q) is the local acceptance set, which is therefore subanalytic, including at the origin.

A subanalytic subset of a two-dimensional real analytic manifold is semianalytic (Bierstone & Milman, 1988, Theorem 6.1(2)). Thus, near the origin, acceptance is a Boolean combination of sign

conditions on finitely many analytic functions of (p, q). Lemma 10, with $q = 1/(n + M)$ when there are M markers, gives the same eventual sign for every such function on the two count paths. Their labels must agree, contradicting the threshold rule.

The boundary hypothesis specifies the scope of this extension. For example, singular division can recover the unnormalized count $x = p/q$ from $p = x/(n + 1)$ and the end-marker mass $q = 1/(n+1)$. After a uniform layer obtains p, q, a second unmasked layer can use query x, end-marker key and value one, and all ordinary keys and values zero. It returns

$$r = \frac{e^x}{n + e^x}, \qquad r > 1/2 \iff x > \log n.$$

The operation p/q has no continuous extension at the corner. This example explains why the compact-neighborhood assumptions accompany the semialgebraic extension.

C Approximation at all input lengths

This appendix proves Proposition 4 and gives a sufficient degree. The central issue is to control the states reached *after* substitution. We first obtain state bounds valid for every positive attention weight function, then compare the exponential and polynomial blocks at a common state. This separates the approximation of the weight function from the propagation of its error.

C.1 Setting and statement

There are a fixed number L of blocks, with fixed finite state and head dimensions. Write $H = (h_i)_i$ for a state array and

$$\|H\|_{\max} = \max_i \|h_i\|_\infty.$$

For head h in block ℓ, queries, keys, and values are fixed affine maps $q_i = Q_{\ell h} h_i + q_{\ell h}$, $k_j = K_{\ell h} h_j + k_{\ell h}$, and $v_j = V_{\ell h} h_j + v_{\ell h}$. The score is

$$s_{ij} = \langle q_i, k_j \rangle.$$

The visible sets $J_{\ell h, i}$ are also externally specified and identical in the two executions. On a nonempty row, a positive weight function κ produces

$$o_i^\kappa(H) = \frac{\sum_{j \in J_{\ell h, i}} \kappa(s_{ij}) v_j}{\sum_{j \in J_{\ell h, i}} \kappa(s_{ij})}.$$

An empty row returns zero in both executions. A block updates a position by a fixed map $B_\ell(h_i, o_{i1}, \ldots, o_{iH_\ell})$, which includes its residual and pointwise computations. Initial states satisfy $\|H^{(0)}\|_{\max} \le R_0$ uniformly. We assume that B_ℓ is defined and locally Lipschitz on an open neighborhood of each compact argument box constructed below. An affine readout at the end marker has a Lipschitz constant L_{out} independent of length.

Proposition 12 (Log-weight function error δ implies output error $C_f \delta$)**.** *Under these assumptions there are a compact interval J and a finite constant C_f, both independent of input length, such that every globally positive polynomial P satisfies*

$$\sup_{u \in J} |\log P(u) - u| \le \delta \quad \Longrightarrow \quad \sup_n \sup_{w:\, |w| = n} \|f_P(w) - f_{\exp}(w)\|_\infty \le C_f \delta. \tag{14}$$

All other network operations are retained. Appendix C.5 supplies globally positive rational-coefficient polynomials meeting the log-weight tolerance and derives the sufficient degree in Proposition 4.

C.2 Normalized weights: an error that does not grow with row length

The following multiplicative-tilt bound is a finite-probability form of the sharp Hilbert-metric/total-variation inequality of Cohen & Fausti (2023, Theorem 5.1). We use the convention that total variation is half the ℓ_1 distance.

Lemma 13 (Multiplicative perturbations of probabilities)**.** *Let p be a probability vector, let $0 < m \le r_j \le M$, and set $q_j = p_j r_j/\mu$, where $\mu = \sum_j p_j r_j$. Then*

$$\mathrm{TV}(p,q) := \frac{1}{2}\sum_j |p_j - q_j| \le \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}. \tag{15}$$

If $r_j = e^{e_j}$ and $|e_j| \le \delta$, this is at most $\tanh(\delta/2) \le \delta/2$. For vectors v_j in any normed space,

$$\left\| \sum_j (p_j - q_j) v_j \right\| \le \mathrm{diam}\{v_j\}\,\mathrm{TV}(p,q). \tag{16}$$

Proof. On the common support of p, q, the ratio $q_j/p_j = r_j/\mu$ has maximum divided by minimum at most M/m, so their Hilbert distance is at most $\log(M/m)$. Theorem 5.1 of Cohen & Fausti (2023) bounds the full ℓ_1 distance by $2\tanh(\log(M/m)/4)$. Dividing by two for our convention gives (15). For $m = e^{-\delta}$ and $M = e^{\delta}$ the bound is $\tanh(\delta/2) \le \delta/2$.

For (16), the positive and negative parts of the signed vector $p - q$ each have total mass $\mathrm{TV}(p,q)$. After division by that mass, each weighted sum is a convex combination of the v_j. Their difference has norm at most the diameter. The zero-mass case is immediate. □

Thus a bounded error in log weights controls the normalized weighted average without a factor equal to the number of visible positions.

C.3 A COMMON STATE DOMAIN FOR ALL POSITIVE-WEIGHT EXECUTIONS

Suppose all input states of block ℓ have norm at most $R_{\ell-1}$. For its head h, put

$$V^*_{\ell h} = \|V_{\ell h}\|_{\infty\to\infty} R_{\ell-1} + \|v_{\ell h}\|_\infty.$$

Every positive-weight head output is a convex combination of value vectors, so it has norm at most $V^*_{\ell h}$; the same bound holds for an empty row. Let

$$\mathcal{K}_\ell = \{(x, o_1, \dots, o_{H_\ell}) : \|x\|_\infty \le R_{\ell-1},\ \|o_h\|_\infty \le V^*_{\ell h}\}, \qquad R_\ell = \max_{z\in\mathcal{K}_\ell} \|B_\ell(z)\|_\infty.$$

The maximum is finite by continuity and compactness. The recursive construction is independent of the particular positive weight functions. Induction therefore bounds the exponential execution, the substituted execution, and executions using different positive weight functions in different blocks, on every input length.

The corresponding affine query and key bounds give a compact interval J containing every score on these state boxes. This interval is fixed before a polynomial is chosen. The interval covers both executions.

C.4 BLOCK SENSITIVITY AND COMPARISON AT THE SUBSTITUTED STATE

Each B_ℓ has a finite Lipschitz constant L_ℓ on $\mathcal{K}_\ell$, using the maximum norm on its arguments.

Consider two predecessor arrays whose maximum difference is E. For a head of key dimension d_k, let Q^*, K^* be the bounds on the infinity norms of its queries and keys, and let $Q_{\mathrm{op}}, K_{\mathrm{op}}, V_{\mathrm{op}}$ be the affine linear-part norms. Every score difference is at most $c_{\ell h}E$, where

$$c_{\ell h} = d_k(Q_{\mathrm{op}}K^* + Q^*K_{\mathrm{op}}).$$

Exponential weights in the two runs therefore have total variation distance at most $c_{\ell h}E/2$ by Lemma 13. Splitting changes in the values from changes in the weights gives

$$\|o^{\exp}(H) - o^{\exp}(H')\|_\infty \le (V_{\mathrm{op}} + V^*_{\ell h}c_{\ell h})E.$$

Consequently the original exponential block has a length-independent Lipschitz constant

$$a_\ell = L_\ell \max\{1, \max_h (V_{\mathrm{op}} + V^*_{\ell h}c_{\ell h})\}. \tag{17}$$

These estimates also cover empty rows, whose output is identically zero.

Now hold the predecessor array fixed and replace the exponential by P. If $|\log P(u) - u| \le \delta$ on J, the multipliers of the exponential weights lie between $e^{-\delta}$ and e^{δ}. The diameter of the values is at most $2V^*_{\ell h}$, so the head discrepancy is at most $V^*_{\ell h}\delta$. The block discrepancy is thus at most

$$b_\ell \delta, \qquad b_\ell = L_\ell \max_h V^*_{\ell h}. \tag{18}$$

Let E_ℓ denote the maximum state error after block ℓ. Insert the *original exponential block evaluated at the substituted predecessor state* between the two actual block outputs. The common-state estimate and the original block's sensitivity give

$$E_0 = 0, \qquad E_\ell \le a_\ell E_{\ell-1} + b_\ell \delta. \tag{19}$$

All inserted states lie in the common boxes proved above. Iteration and the readout bound yield (14) with

$$C_f = L_{\text{out}} \sum_{\ell=1}^{L} b_\ell \prod_{r=\ell+1}^{L} a_r. \tag{20}$$

Only the sensitivity of the exponential block was used. No derivative bound on P is required.

C.5 An everywhere-positive polynomial and its degree

For even $d \ge 0$, define

$$T_d(u) = \sum_{k=0}^{d} \frac{u^k}{k!}. \tag{21}$$

For $u \ge 0$ this is positive term by term. For $u < 0$, Taylor's theorem gives $e^u - T_d(u) = e^\xi u^{d+1}/(d+1)! < 0$ for an intermediate ξ. Thus $T_d(u) > 0$ on the entire real line. Its coefficients are rational.

Choose a fixed rational $R \ge 1$ with $J \subset [-R, R]$. Taylor's remainder gives

$$\sup_{|u|\le R} |T_d(u)e^{-u} - 1| \le e^{2R} \frac{R^{d+1}}{(d+1)!} =: r_d.$$

For $r_d < 1$ this implies

$$\sup_{|u|\le R} |\log T_d(u) - u| \le -\log(1 - r_d).$$

Since $r_d \to 0$, choose even d with $-\log(1 - r_d) \le \eta/\max\{1, C_f\}$. Equations (19) and (20) prove the existence assertion of Proposition 4.

By Stirling's formula, $\log r_d = -d \log d + O_R(d)$ for fixed R. Taking an even degree just above $(1+\zeta)\log(1/\eta)/\log\log(1/\eta)$ and then letting $\zeta \downarrow 0$ gives

$$\deg P \le (1 + o(1)) \frac{\log(1/\eta)}{\log\log(1/\eta)}. \tag{22}$$

The fixed-interval scalar leading term is classical (Aggarwal & Alman, 2022, Theorem 3(1)); the even Taylor choice additionally gives global positivity and rational coefficients.

The end-marker affine readout has a norm bound independent of length. The counting construction has bounded initial states and affine pointwise updates, and therefore satisfies the proposition's assumptions.

C.6 A fixed network requires the leading output degree

Let P replace the exponential in (3), and divide it by $P(0) > 0$ so that $P(0) = 1$. The first layer stays uniform. On positive rational frequency vectors $(1 - t - s, t, s)$, scale the counts to infinity. The end-marker mass vanishes and the two outputs converge to

$$F_E(t, s) = \frac{1 - t - s - se^t}{1 + s(e^t - 1)}, \qquad F_P(t, s) = \frac{1 - t - s - sP(t)}{1 + s(P(t) - 1)}.$$

All-length output error at most η therefore implies $|F_E - F_P| \le \eta$ on these rational points. Positivity of P and continuity extend the guarantee throughout the closed frequency triangle. The candidate is fixed before taking this limit.

Choose $s_P = (1-t)/(1+P(t))$. At this boundary $F_P = 0$, and direct substitution gives (8). For $0 \le t \le 1/2$, since $e^t \ge 1$,

$$\frac{(1-t)|P(t)-e^t|}{t+P(t)+(1-t)e^t} \ge \frac{1}{2}\frac{|P(t)-e^t|}{P(t)+e^t}.$$

Thus $|P-e^t|/(P+e^t) \le 2\eta$. For $\eta \le 1/100$, this implies

$$|P(t)e^{-t}-1| \le \frac{4\eta}{1-2\eta} \le 5\eta, \qquad \|P-e^t\|_{[0,1/2]} \le 5e^{1/2}\eta.$$

Set $d = \deg P$, $m = d+1$, and $h = 1/(2m)$. The order-m forward difference of P at $0, h, \dots, mh$ is zero. For the exponential it is $(e^h-1)^m \ge (2m)^{-m}$. The absolute binomial coefficients sum to 2^m, so

$$(2m)^{-m} \le 2^m 5e^{1/2}\eta.$$

Taking logarithms gives $\log(1/\eta) \le m\log(4m) + \log(5e^{1/2})$. This first forces $m \to \infty$ and then

$$d \ge (1-o(1))\frac{\log(1/\eta)}{\log\log(1/\eta)}.$$

The upper bound already proved uses globally positive rational-coefficient polynomials and has the same leading term. This proves the second assertion of Proposition 4.

C.7 Joint dependence on score range and output tolerance

Multiply the second-layer query of the fixed construction by $R \ge 1$. Its output becomes

$$F_R(x,y,z) = \frac{x - ze^{Ry/(n+1)}}{x+y+1+ze^{Ry/(n+1)}}.$$

Let $D_{\mathrm{out}}(R,\eta)$ be the least degree among globally positive replacements giving output error at most η on every input and length. For $0 < \eta \le 1/100$,

$$D_{\mathrm{out}}(R,\eta) = \Theta\left(R + \frac{\log(1/\eta)}{\log(2+\log(1/\eta)/R)}\right), \tag{23}$$

with absolute constants for this family.

For the upper bound, an even Taylor polynomial of degree $d = m-1$ obeys

$$\sup_{|u|\le R}|T_d(u)e^{-u}-1| \le e^{2R}\frac{R^m}{m!} \le e^{2R}\left(\frac{eR}{m}\right)^m.$$

Logarithmic error at most δ follows if

$$m\log\frac{m}{eR} \ge 2R + \log\frac{2}{\delta}, \qquad 0 < \delta \le 1.$$

With $A = \log(1/\delta)$, choose an odd integer m of order $R + A/\log(2+A/R)$. If A is at most a fixed multiple of R, a sufficiently large multiple of R works. If A/R is large, substitution of a sufficiently large multiple of $R + A/\log(2+A/R)$ works. The first layer is unchanged and the second averages values in $[-1,1]$, so Lemma 13 bounds output error by δ. Taking $\delta = \eta/2$ proves the upper bound.

For the lower bound, normalize $P(0) = 1$ and apply the previous output inversion with $E = e^{Rt}$ and $K = P(Rt)$. It gives $|K-E|/(K+E) \le 2\eta$ on $0 \le t \le 1/2$. Thus, with $L = R/2$,

$$|P(u)e^{-u}-1| \le 5\eta \qquad (0 \le u \le L). \tag{24}$$

Equation (24) controls relative error. Converting it to additive error on $[0,L]$ gives $5\eta e^L$, which need not be small as L grows, so we retain the range and precision estimates below.

Rescaling $[0, L/2]$ to $[-1, 1]$ sends L to 3. The Chebyshev exterior extremal bound (Aggarwal & Alman, 2022, Fact 16) gives

$$|P(L)| \le T_d^{\mathrm{Ch}}(3) \sup_{0\le u\le L/2} |P(u)| \le 6^d \sup_{0\le u\le L/2} |P(u)|,$$

where T_d^{Ch} is normalized by $T_d^{\mathrm{Ch}}(1) = 1$. The cited bound for exact degree r also holds for degree at most d: $T_r^{\mathrm{Ch}}(3) \le T_d^{\mathrm{Ch}}(3)$ for $r \le d$, including constants. Its explicit formula gives $T_d^{\mathrm{Ch}}(3) \le 6^d$. Combining with (24),

$$L/2 \le d \log 6 + \log \frac{1+5\eta}{1-5\eta}.$$

Since $L \ge 1/2$ and the logarithmic term is below 0.101, we obtain $d = \Omega(R)$.

For the precision dependence, apply the order-$m = d+1$ difference at $0, L/m, \dots, L$. Equation (24) gives

$$(e^{L/m} - 1)^m \le 5\eta e^L 2^m, \qquad \log(1/\eta) \le L + m \log(2m/L) + \log 5.$$

When $\log(1/\eta) = O(L)$, the range bound suffices. Otherwise inversion of the increasing function $m \log(2m/L)$ on its positive range gives $m = \Omega(\log(1/\eta)/\log(2+\log(1/\eta)/L))$. The range bound absorbs the difference between d and m. This proves (23). At fixed R, the factorial upper bound and finite-difference lower bound retain leading constant one.

For a general fixed network whose common interval lies in $[-R, R]$, use $\delta = \min\{1, \eta/\max\{1, C_f\}\}$ in the upper construction. The same expression is sufficient with $\log(2\max\{1, C_f\}/\eta)$ replacing $\log(1/\eta)$, retaining the network's stability constant.

This range dependence pertains to output accuracy. For average classification risk under the frequency distribution of Section 5.3, the scaled target's accepted mass is

$$a_R = 2\int_0^1 \frac{1-t}{1+e^{Rt}}\, dt \le 2/R.$$

The globally positive quadratic $P_M(u) = 1 + Mu^2$ has a boundary tending to zero for almost every t as $M \to \infty$. Its risk tends to a_R by dominated convergence, so every tolerance $\epsilon > 2/R$ is attained at degree two. This identifies why the output range lower bound has a different scope from the average-risk conclusion.

D Classification error and exact decisions at finite lengths

We first prove Theorem 5 directly from integer labels. The remaining subsections establish the average-risk extensions (29), (32) and (33). Throughout, $P(0) = 1$ and $P > 0$ on $[0, 1]$.

D.1 A direct finite-grid proof of Theorem 5

Proof. Fix $n \ge 32$ and write $P = P_n$, $d = \deg P$, and $m = d + 1$. On a slice with $1 \le y \le n-2$, put $u_y = y/(n+1)$ and $M = n - y$. The allowed counts are $z = 1, \dots, M-1$, with $x = M - z$. The two rules accept precisely the prefixes below

$$B_E(y) = \frac{M}{1+e^{u_y}}, \qquad B_P(y) = \frac{M}{1+P(u_y)}.$$

For any $B \in (0, M)$, the number of accepted integers is $\lceil B \rceil - 1$. Exact agreement therefore implies

$$\lceil B_P(y) \rceil = \lceil B_E(y) \rceil, \qquad |B_P(y) - B_E(y)| < 1. \tag{25}$$

This also covers integer cutoffs under the strict acceptance convention.

Restrict to the actual grid points $u_y \in [1/4, 1/2]$. Here $M \ge (n-1)/2 \ge n/3$. Write $a_P = (1+P(u_y))^{-1}$ and $a_E = (1+e^{u_y})^{-1}$. Equation (25) gives $|a_P - a_E| < 3/n$. Since $a_E > 1/3$ and $n \ge 32$, we also have $a_P > 1/6$. Inverting yields the coefficient-independent bound

$$|P(u_y) - e^{u_y}| = \frac{|a_P - a_E|}{a_P a_E} \le \frac{54}{n}. \tag{26}$$

If $m > (n+1)/16$, the asserted logarithmic degree lower bound already holds for large n. Otherwise take

$$y_0 = \left\lceil \frac{n+1}{4} \right\rceil, \quad k = \left\lfloor \frac{n+1}{8m} \right\rfloor, \quad u_j = \frac{y_0 + jk}{n+1} \quad (0 \le j \le m).$$

Then $k \ge 2$ and $h = k/(n+1) \ge 1/(16m)$. Moreover, $u_0 \ge 1/4$ and $u_m \le 3/8 + 1/(n+1) < 1/2$, so every node is a valid slice covered by (26). The order-m forward difference annihilates P, and its absolute coefficients sum to 2^m. Hence

$$e^{1/4}(16m)^{-m} \le e^{u_0}(e^h - 1)^m = |\Delta_h^m(e^{\cdot} - P)(u_0)| \le 2^m \frac{54}{n}.$$

Thus $\log n \le m \log m + m \log 32 + \log 54 - 1/4$. This forces $m \to \infty$. For any fixed $\zeta > 0$, a subsequence with $d \le (1-\zeta)\log n / \log\log n$ would make the right side at most $(1-\zeta+o(1))\log n$, a contradiction. This proves Theorem 5. □

D.2 A SMALL RISK SUPPLIES A LARGE SET OF ACCURATE WEIGHT FUNCTION VALUES

Define $g_P(t) = (1-t)/(1+P(t))$, $g(t) = (1-t)/(1+e^t)$, and $R(P) = 2\int_0^1 |g_P - g|\, dt$. This is the disagreement area under density two on $\{(t,s) : t, s > 0,\ t+s < 1\}$. Appendix D.4 proves its finite-length interpretation uniformly over coefficients.

Let $I = [1/4, 3/4]$. On I we have $g \ge 1/16$: the numerator is at least $1/4$ and $1 + e^u < 4$. If $R(P) = 0$, continuity implies $g_P = g$ on I and hence $P = e^u$ there. Differentiating one more time than the degree of P is a contradiction. We may therefore assume $R = R(P) > 0$.

For $R \le 1/64$, define the compact set

$$G = \{u \in I : |g_P(u) - g(u)| \le 2R\}.$$

Since $\int_I |g_P - g| \le R/2$, the complement of G in I has measure at most $1/4$. Consequently $|G| \ge 1/4$. On G, $g_P \ge 1/32$, and inversion of the two boundary formulas gives

$$|P(u) - e^u| = \frac{(1-u)|g_P(u) - g(u)|}{g_P(u)g(u)} \le 768R. \tag{27}$$

No estimate on P outside G has been made or is needed.

Lemma 14 (A degree bound from a divided difference)**.** *For every polynomial of degree at most D, positive on $[0,1]$ and normalized by $P(0) = 1$,*

$$R(P) \ge \frac{e^{1/4}}{768[8(D+1)]^{D+1}}. \tag{28}$$

Proof. First suppose $R \le 1/64$ and put $m = D+1$ and $\mu = |G| \ge 1/4$. The cumulative measure $A(x) = |G \cap (-\infty, x]|$ is continuous and 1-Lipschitz. Choose the extreme points of G for u_0, u_m, and intermediate points in G with $A(u_i) = i\mu/m$; a level attained in a gap is attained at its endpoints. Then $u_j - u_i \ge (j-i)\mu/m \ge (j-i)/(4m)$.

The order-m divided difference annihilates P. Its mean-value formula (de Boor, 2005, Section 8, equation (44)) therefore gives, for some $\xi \in [u_0, u_m]$,

$$[u_0, \dots, u_m](e^{\cdot} - P) = \sum_{i=0}^{m} \frac{e^{u_i} - P(u_i)}{\prod_{j \ne i}(u_i - u_j)} = \frac{e^{\xi}}{m!} \ge \frac{e^{1/4}}{m!}.$$

The separation of the nodes bounds the absolute sum of the coefficients by

$$(4m)^m \sum_{i=0}^{m} \frac{1}{i!(m-i)!} = \frac{(8m)^m}{m!}.$$

Combining this with (27) proves the bound. For $R > 1/64$, the claimed right side is smaller than $1/64$ even when $D = 0$, and decreases with D. This covers the remaining case. □

In particular, for an absolute constant C and every degree-at-most-d candidate, including $d = 0$,

$$R(P) \ge \exp\bigl(-(d+1)\log(d+1) - C(d+1)\bigr). \tag{29}$$

The points need not be evenly spaced and the candidate can be very large on the discarded set. This is why the argument gives a bound from average classification error rather than requiring a uniform weight function error.

D.3 MATCHING CONSTRUCTION AND THE LEADING DEGREE

For even D, the polynomial T_D from (21) is globally strictly positive and $T_D(0) = 1$. On $[0, 1]$,

$$0 \le e^u - T_D(u) \le \frac{e}{(D+1)!}.$$

Because

$$|g_{T_D}(u) - g(u)| = \frac{(1-u)|T_D(u) - e^u|}{(1+T_D(u))(1+e^u)} \le |T_D(u) - e^u|,$$

we obtain the constructive upper bound

$$R(T_D) \le \frac{2e}{(D+1)!}. \tag{30}$$

Let ρ_D be the infimum of risk among all globally positive polynomials of degree at most D normalized by $P(0) = 1$. Use the largest even integer at most D for the upper construction. The two preceding bounds and $\log(D!) = D \log D - D + O(\log D)$ imply

$$D \log D - O(D) \le \log(1/\rho_D) \le D \log D + O(D). \tag{31}$$

Changing D to its largest even predecessor changes only an $O(\log D)$ term at this scale.

Define $D_{\text{risk}}(\epsilon)$ to be the least degree for which a feasible polynomial has risk at most ϵ. With $A = \log(1/\epsilon)$, any degree below $(1-\zeta)A/\log A$, for fixed $\zeta > 0$, obeys $D \log D + O(D) < A$ for sufficiently large A. The lower bound then excludes risk at most ϵ. An even degree just above $(1+\zeta)A/\log A$ obeys $D \log D - O(D) > A$, and the explicit Taylor construction has risk at most ϵ. Letting $\zeta \downarrow 0$ proves

$$D_{\text{risk}}(\epsilon) \sim \frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}. \tag{32}$$

The argument does not require the infimum defining ρ_D to be attained. Every competitor obeys the lower bound and the upper bound exhibits a competitor.

D.4 A COEFFICIENT-INDEPENDENT FINITE-LENGTH ESTIMATE

Let $R_n(P)$ be the disagreement probability under uniform sampling of positive count triples of length n. We prove

$$\sup_{\substack{\deg P \le d,\ P(0)=1 \\ P>0 \text{ on } [0,1]}} |R_n(P) - R(P)| \le \frac{32(d+1)}{n}, \qquad n \ge 3. \tag{33}$$

This establishes the limiting interpretation in Section 5.3, uniformly over the coefficients at any fixed degree. Write $\mathrm{Var}(v)$ for the total variation of a function on $[0, 1]$. For a degree-at-most-d candidate,

$$g_P'(t) = -\frac{1 + P(t) + (1-t)P'(t)}{(1+P(t))^2}.$$

The numerator is a polynomial of degree at most d and is nonzero at $t = 1$, where it equals $1 + P(1) > 0$. Thus g_P has at most $d + 1$ monotone pieces. Since $0 \le g_P \le 1$, this gives

$$\mathrm{Var}(g_P) \le d + 1, \quad \mathrm{Var}(g) = 1/2, \quad \mathrm{Var}(|g_P - g|) \le d + 3/2. \tag{34}$$

These bounds depend on degree alone.

For length n, put $a = n/(n+1)$. The finite unmasked boundaries are

$$b_{P,n}(t) = \frac{1-t}{1+P(at)}, \qquad b_{e,n}(t) = \frac{1-t}{1+e^{at}}, \qquad f_n = |b_{P,n} - b_{e,n}|.$$

The derivative-root argument also gives $\mathrm{Var}(b_{P,n}) \le d + 1$; $b_{e,n}$ decreases from $1/2$ to zero. Consequently $\mathrm{Var}(f_n) \le d + 3/2$ and $0 \le f_n(t) \le 1 - t$.

Fix $y \in \{1, \dots, n-2\}$. For either cutoff $B = nb(y/n) \in (0, n-y)$, the number of accepted integers $1 \le z \le n - y - 1$ under $z < B$ is $\lceil B \rceil - 1$. It differs from B by a number in $[-1, 0)$, including

ties. Thus the disagreement count on this slice differs from $nf_n(y/n)$ by at most one. Let S_n be the total disagreement count, $M_n = (n-1)(n-2)/2$, and $I_n = \int_0^1 f_n(t)\,dt$. A bounded-variation right Riemann sum has error at most $\mathrm{Var}(f_n)/n$: apply the variation bound on each subinterval and sum. Omitting the last two sample terms costs at most $1/n^2$, since $f_n(1) = 0$ and $f_n((n-1)/n) \le 1/n$. Therefore

$$|S_n - n^2 I_n| \le (n-2) + n(d+3/2) + 1 \le n(d+5/2).$$

For $n \ge 6$, $M_n \ge n^2/4$ and $I_n \le 1/2$, giving

$$|R_n(P) - 2I_n| \le \frac{4(d+5/2)}{n} + \frac{3n-2}{2M_n} \le \frac{4d+16}{n}. \tag{35}$$

It remains to compare I_n with $I = \int |g_P - g|$. For a bounded-variation function v, integration against its variation measure gives

$$\int_0^1 |v(at) - v(t)|\,dt \le \frac{1-a}{a}\,\mathrm{Var}(v) = \frac{\mathrm{Var}(v)}{n}.$$

Indeed, a point u of that measure contributes only for $t \in [u, u/a] \cap [0,1]$, an interval of length at most $(1-a)/a$. Also

$$b_{P,n}(t) = \frac{1-t}{1-at} g_P(at), \qquad |b_{P,n}(t) - g_P(at)| \le (1-a)t,$$

using $g_P(at) \le 1 - at$. The same bound holds for g. The triangle inequality and (34) now give

$$|I_n - I| \le \frac{\mathrm{Var}(g_P) + \mathrm{Var}(g)}{n} + \frac{1}{n+1} \le \frac{d+5/2}{n}.$$

Combining with (35) yields

$$|R_n(P) - R(P)| \le \frac{6d+21}{n} \le \frac{32(d+1)}{n}.$$

For $3 \le n \le 5$, the last inequality is immediate because both risks lie in $[0,1]$. This proves (33), uniformly over all coefficients in the stated class.

D.5 Consequences for small finite-length risk

The risk bounds also give a necessary inequality for exact agreement, independently of the direct proof above. If $R_n(P) = 0$ for a degree-d candidate, combining (33) and (28) gives

$$n \le 24576 e^{-1/4}(d+1)[8(d+1)]^{d+1}. \tag{36}$$

The same estimate quantifies small finite-length error. Let $D_n(\epsilon)$ be the least degree attaining $R_n(P) \le \epsilon$, with coefficients allowed to depend on n, and put $A = \log(1/\epsilon)$. If $\epsilon \downarrow 0$ and

$$\frac{n\epsilon}{A/\log A} \longrightarrow \infty,$$

then $D_n(\epsilon) \sim A/\log A$. To prove this, fix $\zeta > 0$. A degree at most $(1-\zeta)A/\log A$ has limiting risk at least $\exp(-(1-\zeta+o(1))A)$ by (28), whereas its finite-length correction is $o(\epsilon)$. Such a candidate cannot attain error ϵ. An even Taylor degree $(1+\zeta+o(1))A/\log A$ has limiting risk $o(\epsilon)$ and finite correction $o(\epsilon)$, so it does attain the target. Let $\zeta \downarrow 0$. The conclusion holds both for candidates positive on $[0,1]$ and for the globally positive subclass.

E Certified numerical values and a finite disagreement

The quantities in Section 5.3 concern the same two-layer unmasked network and the same degree-ten polynomial. We give a reproducible enclosure of its limiting risk and a rational certificate for its decision at length 1080.

E.1 AN ENCLOSURE OF THE LIMITING ERROR

On $[0, 1]$, the Taylor tail $E_{10}(t) = e^t - T_{10}(t)$ is nonnegative. Writing the risk through this tail avoids subtraction of nearly equal numbers:

$$R(T_{10}) = \int_0^1 \frac{2(1-t)E_{10}(t)}{(1+T_{10}(t))(1+e^t)}\,dt. \tag{37}$$

The resulting certified interval is

$$3.084956897 \times 10^{-11} \le R(T_{10}) \le 3.086208016 \times 10^{-11}. \tag{38}$$

This is a deterministic integration enclosure of (10). Equation (33) identifies the same integral as the limiting disagreement probability of uniform positive count triples in the unmasked model.

Here is an exact-arithmetic procedure reproducing the enclosure. Partition $[0, 1]$ into $N = 131072$ equal cells. For a cell $[a, b]$, all of E_{10}, T_{10}, and e^t are nondecreasing. Its integral contribution is therefore between

$$\frac{1}{N}\frac{2(1-b)E_{10}(a)}{(1+T_{10}(b))(1+e^b)} \quad \text{and} \quad \frac{1}{N}\frac{2(1-a)E_{10}(b)}{(1+T_{10}(a))(1+e^a)}.$$

At each dyadic endpoint, sum the positive exponential series through degree 40. Its remaining tail is at most the next term multiplied by $42/41$, because successive ratios after term 41 are at most $1/42$. Sum terms 11 through 40 separately for E_{10}, applying the same tail bound. Carry every quantity as an integer interval with scale 10^{60}: round lower endpoints down and upper endpoints up after each multiplication or division. All factors are nonnegative, so these directed operations preserve enclosure. Sum the lower and upper cell bounds with the same rule. The included script `verify_results.py` implements this procedure and verifies that its interval lies inside (38); its JSON output records the tighter endpoints as exact decimals.

E.2 A RATIONAL CERTIFICATE AT LENGTH 1080

Set $(x, y, z) = (74, 976, 30)$, $n = 1080$, and $u = 976/1081$. The two rational quantities

$$\frac{37}{15} - \sum_{j=0}^{10} \frac{u^j}{j!}, \qquad \sum_{j=0}^{12} \frac{u^j}{j!} - \frac{37}{15}$$

are both strictly positive. They can be checked by clearing the positive integer denominators, or by exact fraction arithmetic as in the script. Since $u > 0$, the exponential's tail beyond degree 12 is strictly positive. This proves (11) without a floating-point sign decision. The denominator in (12) is positive; thus T_{10} accepts and the exponential rejects.

The script also checks the three-layer reciprocal-sum construction and both beginning-marker variants on all 510 nonempty binary words of lengths one through eight, using exact fractions. It compares unmasked and self-inclusive causal executions of the exponential construction on all $1{,}092$ three-symbol words of lengths one through six, using floating-point exponentials. These checks verify the stated finite executions; the general constructions and exclusions are proved in Appendices A and B.

For the direct proof in Appendix D.1, it checks strict integer cutoffs including ties, all $32{,}366$ admissible grid choices with $32 \le n \le 1024$, and finite-difference annihilation on monomials through order eight. At the length-1080 slice, the cutoffs from T_{10} and T_{12} have ceilings 31 and 30, respectively, certifying the failure of the ceiling equality.