H+1 HEADS EXPRESS MORE THAN H HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

We prove the following hierarchy: for each h there exists a decision problem that can be solved by a one layer transformer with $h + 1$ heads but not with only h heads. The result holds for a standard transformer architecture, which includes both attention and feed-forward blocks.

A typical transformer layer packs several attention heads working in parallel using separate learned projections. Additional heads can speed up convergence during training on some tasks Baherwani et al. (2026) and allow distinct logical operations to be carried out in separate computational units. At the same time, in practice, many heads can be pruned with minimal loss in the model's performance Michel et al. (2019); Voita et al. (2019). Does this mean that a model can always be "compressed" into an equivalent one with fewer heads? Or perhaps it is the case that additional heads allow transformers to realize computational circuits that were impossible with fewer heads?

We study this question at the level of expressivity, i.e., whether a certain assignment of model weights is possible such that a resulting model solves a given task. Our main problem is then as follows: is it true that for every h and l we can find a task that a) is solved by some transformer with $h + 1$ heads and l layers but b) no transformer with h heads and l layers can solve this problem? Some partial answers already exist for restricted models. Tesfaye et al. (2026) considered the question for attention-only transformers, i.e., models with no feed-forward network blocks, and gave an example of a task that separates one-layer one-head transformers from one-layer two-head transformers. This was later generalized to arbitrary k by Rajaraman et al. (2026). More precisely, they showed that k-bit parity is computable by a one-layer, k-head transformer but not by a one-layer, $(k-1)$-head transformer. Again, both results assume that no feed-forward block is used after attention. This is indeed crucial as for fixed k, a sufficiently large MLP is enough to compute k-bit parity, as well as any other Boolean function on fixed-length inputs. Therefore, the main question surprisingly remains open.

Main result: To this end, we consider a formal language $\mathsf{PowGraph}_k$ which consists of pairs of words such that the number of ones in the second word is at least as large as the k-th power of the number of ones in the first word. Our main theorem states that for all $h > 1$

- no one-layer transformer with $h - 1$ heads solves $\mathsf{PowGraph}_h$ (the lower bound),
- a single one-layer transformer with h heads solves $\mathsf{PowGraph}_h$(the upper bound).

We provide a separate Lean formalization of the proof as a supplement to the paper.

Our upper bound uses a non-additive positional encoding: tokens and positions are encoded jointly, and the encoding cannot be written as the sum of a token embedding and a positional embedding. Non-additive positional mechanisms are common in modern transformers: RoPE (Su et al., 2024), for example, uses position-dependent rotations. The original transformer architecture (Vaswani et al., 2017), however, used additive positional encoding. We leave open whether our upper bound can be realized with additive positional encoding.

Related work The literature on the expressivity of transformers is vast and this paper lacks space to accurately describe each paper relevant to the topic. The brief survey below focuses on different lower bound techniques.

Our lower bounds uses the algebraic representation of one-layer transformer introduced by Kozachinskiy et al. (2026). After fixing the readout position, attention outputs become ratios of affine functions. The authors combine this with sensitivity bounds to show that no fixed one-layer

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.

transformer computes PARITY language. Hsu (2026) extended this result and provided a quantitative version of the bound. The idea of reducing k-heads into equivalent polynomial tests of degree k effectively follows these two preprints.

Other techniques for lower bounds against one-layer are known, with some extensions to multi-layer case. Communication complexity is one of these and has been used to prove one-layer lower bounds for function composition (Peng et al., 2024), Match3 (Sanford et al., 2023), and the induction-heads task (Sanford et al., 2024). These results constrain head count, embedding dimension, and numerical precision jointly. In contrast, the split-VC method of Kozachinskiy et al. (2025) can be used even for models with infinite precision.

Circuit complexity provides limitations that apply beyond one layer. Unique hard-attention transformers lie in AC^0 and therefore cannot recognize parity (Hao et al., 2022). Log-precision softmax transformers lie in logspace-uniform TC^0 (Merrill & Sabharwal, 2023).

Major open problems in the area remain. For instance, despite the fact that Turing completeness was proven for average hard attention already in 2019 (Pérez et al., 2019), it is still open whether realistic softmax transformers can simulate Turing machines without exponential blowup in the number of produced tokens Jiang et al. (2026). We refer the reader to the survey by Strobl et al. (2024) for an overview of transformer expressivity.

1 Model and main result

For a positive integer m, write $[m] = \{1, \dots, m\}$. If X is a finite set, then $\{0,1\}^X$ denotes the set of maps $x : X \to \{0,1\}$, and $|x| = \sum_{j \in X} x_j$.

For $x, y \in \{0,1\}^N$, define

$$\mathsf{PowGraph}_{k,N}(x,y) = \mathbf{1}\{|y| \geq |x|^k\}.$$

Write $\mathsf{PowGraph}_k = (\mathsf{PowGraph}_{k,N})_{N \geq 1}$ for this family.

We use the following one-layer transformer model. At input length $2N$, an input is $z = (x_1, y_1, x_2, y_2, \dots, x_N, y_N)$. We write $T_N(x,y)$ for $T_N(z)$ on this interleaving. Each position–bit pair has an embedding

$$e_N : [2N] \times \{0,1\} \longrightarrow \mathbb{R}^D.$$

The dimension D is independent of N. We allow arbitrary dependence on both the position and the bit; in particular, this includes standard additive embeddings $e_N(i,b) = r_{N,i} + u_b$.

Let $\rho_N \in [2N]$ be the readout position. For head h, define

$$\begin{aligned} q_h(z) &= W_Q^h e_N(\rho_N, z_{\rho_N}), \\ k_{h,i}(z) &= W_K^h e_N(i, z_i), \\ v_{h,i}(z) &= W_V^h e_N(i, z_i), \end{aligned}$$

and let its output at ρ_N be

$$\mathcal{H}_{h,N}(z) = \frac{\sum_{i=1}^{2N} \exp\big(\langle q_h(z), k_{h,i}(z)\rangle\big) v_{h,i}(z)}{\sum_{i=1}^{2N} \exp\big(\langle q_h(z), k_{h,i}(z)\rangle\big)}. \tag{1}$$

All matrices and dimensions are independent of N. A fixed finite network of affine maps and ReLU gates receives only the concatenated head outputs and produces a scalar. The transformer accepts exactly when this scalar is positive; let $T_N(z) \in \{0,1\}$ denote its acceptance indicator.

Our contribution is composed of the following two theorems:

Theorem 1.1. *Fix an integer $k \geq 2$. No one layer transformer with fewer than k heads satisfies*

$$T_N(x,y) = \mathsf{PowGraph}_{k,N}(x,y) \quad \textit{for every } N \textit{ and every } x, y \in \{0,1\}^N.$$

Theorem 1.2. *For every fixed $k \geq 2$, there is a one-layer transformer with k heads, which computes $\mathsf{PowGraph}_{k,N}$ for every N.*

2 THE LOWER BOUND

2.1 FROM HEADS TO POLYNOMIAL SIGNS

We use the three-valued sign function

$$\operatorname{sgn}(t) = \begin{cases} -1, & t < 0, \\ 0, & t = 0, \\ 1, & t > 0. \end{cases} \tag{2}$$

Proposition 2.1. *Fix an H-head one-layer transformer, with $H \geq 1$. There exist an integer s and a function $F : \{-1, 0, 1\}^s \to \{0, 1\}$, both independent of N, with the following property. For every N there are polynomials $p_{N,1}, \ldots, p_{N,s}$ of degree at most H in the $2N - 1$ bits other than z_{ρ_N} such that*

$$T_N(z) = F\big(\operatorname{sgn} p_{N,1}(z), \ldots, \operatorname{sgn} p_{N,s}(z)\big) \tag{3}$$

whenever $z_{\rho_N} = 0$. The coefficients may depend on N.

Proof. Fix N and impose $z_{\rho_N} = 0$. The query of every head is then constant. A function of one bit is affine on $\{0, 1\}$, so every scalar coordinate of head h has the form

$$\frac{A_h(z)}{Z_h(z)}, \tag{4}$$

where A_h and Z_h are affine functions of the remaining bits and $Z_h(z) > 0$ on the Boolean cube. Coordinates belonging to the same head have different numerators but the same denominator. This affine-ratio representation is also used in (Kozachinskiy et al., 2026, Section 3.1).

A fixed finite ReLU network has finitely many activation patterns. Once a pattern is fixed, replacing every ReLU by either the identity or zero makes all its preactivations and its final output affine functions of the concatenated head outputs. The signs of these affine functions determine whether the pattern is valid and whether its output is positive. Taking all patterns together, acceptance is therefore determined by the signs of a fixed finite list of affine functions of the head outputs. The value zero causes no difficulty: both ReLU branches agree there.

After grouping an affine function by heads, it has the form

$$\beta + \sum_{h=1}^{H} \frac{A_h(z)}{Z_h(z)}. \tag{5}$$

Because all denominators are positive, this function has the same three-valued sign as

$$P(z) = \beta \prod_{h=1}^{H} Z_h(z) + \sum_{h=1}^{H} A_h(z) \prod_{\substack{g=1 \\ g \neq h}}^{H} Z_g(z). \tag{6}$$

Every factor in (6) is affine, so $\deg P \leq H$. Applying this operation to the fixed finite list of affine functions gives the polynomials in (3). The ReLU network is fixed, hence their number and the function combining their signs are independent of N. □

Plan of the lower-bound proof. After the readout bit has been fixed, Proposition 2.1 describes the transformer using an integer s and a function F that do not depend on the input length. At each length the polynomials $p_{N,1}, \ldots, p_{N,s}$ may be different, but each has degree at most H and is a polynomial in the individual input bits.

Suppose for a moment that, for every R, we had s polynomials in just the two variables

$$a = |x|, \qquad b = |y|,$$

with the same degree bound and the same function F, and that their signs determined the correct answer throughout the integer rectangle

$$0 \leq a \leq R, \qquad 0 \leq b \leq R^k.$$

Section 2.2 proves directly that this is impossible for all sufficiently large R when $H < k$. We evaluate the polynomials at $(2^m, 2^{mk-1})$ and $(2^m, 2^{mk})$. A monomial $a^\alpha b^\beta$ then has growth rate $2^{m(\alpha+k\beta)}$, and $H < k$ makes these growth rates distinct for distinct monomials. Apart from a bounded number of values of m, one monomial dominates each nonzero polynomial at both points, so its sign is the same at the two points. There is a value of m that works for all the polynomials simultaneously. The signs of all polynomials then agree, but the target is zero at the first point and one at the second.

Then it remains to be shown that we can get such two-variable tests from the polynomials in Proposition 2.1. For a chosen R, take the input length N large enough. We color a pair of subsets of the two words by the complete sign vector of the polynomials in the individual input bits. A finite Product Ramsey theorem supplies R positions in the first word and R^k positions in the second so that, within them, this color depends only on the two subset sizes a, b. All other input bits, including the readout bit, are fixed to zero. Consequently every polynomial has a constant sign on all inputs with the same pair (a, b).

We can now average each polynomial over all inputs with fixed values of a and b without changing its sign. The two-block extension of the Minsky–Papert symmetrization argument says that the average is a polynomial in a, b of no larger degree. Since fixing the other bits to zero leaves the target equal to $\mathbf{1}\{b \geq a^k\}$, these averaged polynomials give exactly the two-variable description excluded by the first stage. Sections 2.2 and 2.3 carry out these two steps in that order.

2.2 The two-variable obstruction

We first prove directly that no Boolean combination of a fixed number of low-degree polynomial signs can describe the required threshold on arbitrarily large integer rectangles. The function combining the signs is allowed to depend on R; only the number and degrees of the polynomials are fixed.

Proposition 2.2. *Fix integers $k \geq 2$, $s \geq 0$, and $0 \leq H < k$. For all sufficiently large integers R, there do not exist a function $F : \{-1, 0, 1\}^s \to \{0, 1\}$ and polynomials $q_1, \ldots, q_s \in \mathbb{R}[a, b]$ of total degree at most H such that*

$$F\big(\operatorname{sgn} q_1(a, b), \ldots, \operatorname{sgn} q_s(a, b)\big) = \mathbf{1}\{b \geq a^k\} \tag{7}$$

for every pair of integers $0 \leq a \leq R$ and $0 \leq b \leq R^k$.

Proof. Let

$$M = \binom{H+2}{2}, \qquad K = 2^{H+1}M.$$

We call a positive integer m an *oddball* for a polynomial

$$q(a, b) = \sum_{\alpha+\beta \leq H} c_{\alpha\beta} a^\alpha b^\beta$$

if there are two distinct pairs $(\alpha, \beta) \neq (\alpha', \beta')$, with $c_{\alpha\beta} \neq 0, c_{\alpha'\beta'} \neq 0$ and

$$\frac{1}{K} \leq \frac{|c_{\alpha\beta}| 2^{m(\alpha+k\beta)}}{|c_{\alpha'\beta'}| 2^{m(\alpha'+k\beta')}} \leq K. \tag{8}$$

For each pair in (8), the base-2 logarithm of the middle expression is of the form $A + md$, where A is a real and d is a nonzero integer. Indeed, if $d = 0$, then

$$k|\beta - \beta'| = |\alpha - \alpha'| \leq H < k,$$

which implies $\beta = \beta'$ and then $\alpha = \alpha'$, contrary to the assumption that the two pairs are distinct. Therefore, for each pair there are at most

$$E = \lceil 2 \log_2 K \rceil + 1$$

oddballs. Since there are at most $\binom{M}{2}$ pairs, every polynomial of degree at most H has at most

$$C = \binom{M}{2} E \tag{9}$$

oddballs.

Suppose that q is nonzero and m is not an oddball for q. For each nonzero coefficient, let

$$B_{\alpha\beta} = |c_{\alpha\beta}| 2^{m(\alpha+k\beta)}$$

be the absolute contribution of the monomial $c_{\alpha\beta} a^\alpha b^\beta$ at $(2^m, 2^{mk})$. Choose (α_*, β_*) so that $B_{\alpha_*\beta_*}$ is maximal. For every other nonzero coefficient,

$$\frac{B_{\alpha_*\beta_*}}{B_{\alpha\beta}} \geq 1.$$

Since m is not an oddball, this ratio does not belong to $[1/K, K]$. It cannot be smaller than $1/K$, because it is at least one. Consequently,

$$\frac{B_{\alpha_*\beta_*}}{B_{\alpha\beta}} > K.$$

Thus, the largest absolute contribution is unique and exceeds every other nonzero contribution by a factor greater than K. Since there are at most $M-1$ other monomials and $K > M$, it exceeds the sum of the absolute contributions of all the other monomials and therefore determines the sign of q at this point.

Now, replacing $b = 2^{mk}$ by $b = 2^{mk-1}$ multiplies the contribution of a monomial with exponent β in b by $2^{-\beta}$. This means that the ratios

$$B_{\alpha_*\beta_*} / B_{\alpha\beta}$$

decrease by at most a factor 2^H and remain greater than

$$\frac{K}{2^H} = 2M.$$

Consequently, the same monomial again dominates the sum of all the others. Its coefficient has not changed sign, and therefore q has the same sign at the two points. We have proved that

$$\operatorname{sgn} q(2^m, 2^{mk-1}) = \operatorname{sgn} q(2^m, 2^{mk}) \tag{10}$$

whenever m is not an oddball. Equation (10) also holds for the zero polynomial.

For the s polynomials in (7), at most sC positive integers are oddballs for some of the polynomials. Fix $R > 2^{sC+1}$ and a number m

$$1 \leq m \leq sC + 1$$

that is not an oddball for any of the polynomials. Therefore, F gives the same value at

$$(2^m, 2^{mk-1}) \quad \text{and} \quad (2^m, 2^{mk}).$$

This is the desired contradiction. □

2.3 From input bits to the two counts

For a finite set X, write $\binom{X}{a}$ for the family of its a-element subsets. For $S \subseteq X$, let $\mathbf{1}_S^X \in \{0,1\}^X$ be the binary string defined by

$$(\mathbf{1}_S^X)_i = \begin{cases} 1, & i \in S, \\ 0, & i \notin S. \end{cases}$$

We use the following finite Product Ramsey theorem. Fix positive integers a, b, L with $L \geq \max\{a, b\}$ and a finite set of colors. If X and Y are sufficiently large, then every coloring of $\binom{X}{a} \times \binom{Y}{b}$ is constant on $\binom{I}{a} \times \binom{J}{b}$ for some $I \subseteq X$ and $J \subseteq Y$ with $|I| = |J| = L$ (see e.g. (Böttcher & Foniok, 2013, Theorem 4.1)).

The following lemma is a simple variant of Minsky and Papert's Group-Invariance Theorem and the symmetrization calculation in their parity proof (Minsky & Papert, 1988, §2.3, pp. 48–53, and §3.1, pp. 56–57); see also (Razborov & Sherstov, 2010, Proposition 2.3, p. 1840).

Lemma 2.3. *Let X, Y be disjoint nonempty finite sets, write $A = |X|$ and $B = |Y|$, and let $d \geq 0$ be an integer. Introduce variables x_i, one for each $i \in X$, and variables y_j, one for each $j \in Y$. Let p be a polynomial with real coefficients in these variables, of total degree at most d. Then there is a polynomial $q \in \mathbb{R}[a, b]$ of total degree at most d such that*

$$q(a,b) = \frac{1}{\binom{A}{a}\binom{B}{b}} \sum_{S \in \binom{X}{a}} \sum_{T \in \binom{Y}{b}} p(\mathbf{1}_S^X, \mathbf{1}_T^Y) \tag{11}$$

for all integers $0 \leq a \leq A$ and $0 \leq b \leq B$.

Proof. Since the average in (11) is linear in p, it is enough to consider one term

$$c \prod_{i \in X} x_i^{\alpha_i} \prod_{j \in Y} y_j^{\beta_j}$$

of p, where $c \in \mathbb{R}$ and all exponents are nonnegative integers. On binary inputs, $x_i^{\alpha_i} = x_i$ when $\alpha_i > 0$, and $y_j^{\beta_j} = y_j$ when $\beta_j > 0$. Let

$$S_0 = \{i \in X : \alpha_i > 0\}, \qquad T_0 = \{j \in Y : \beta_j > 0\}, \qquad r = |S_0|, \quad t = |T_0|.$$

For a nonnegative integer h, define

$$(u)_0 = 1, \qquad (u)_h = u(u-1)\cdots(u-h+1) \quad \text{when } h \geq 1.$$

The value of the displayed term at $(\mathbf{1}_S^X, \mathbf{1}_T^Y)$ is therefore c if $S_0 \subseteq S$ and $T_0 \subseteq T$, and zero otherwise. If $a < r$, no a-element subset of X contains S_0. If $a \geq r$, there are $\binom{A-r}{a-r}$ such subsets, out of $\binom{A}{a}$ subsets in total. Thus the fraction containing S_0 is $(a)_r/(A)_r$ in both cases. Similarly, the fraction of the b-element subsets of Y that contain T_0 is $(b)_t/(B)_t$. The denominators are nonzero because $r \leq A$ and $t \leq B$. Consequently the average of this term in (11) is

$$c \frac{(a)_r}{(A)_r} \frac{(b)_t}{(B)_t}. \tag{12}$$

This is a polynomial in a, b of total degree $r + t$. Moreover, $r + t \leq \sum_{i \in X} \alpha_i + \sum_{j \in Y} \beta_j \leq d$. Adding these polynomials over all terms of p proves the lemma. □

Lemma 2.4. *Fix nonnegative integers s, H, positive integers A, B, and a function $F : \{-1, 0, 1\}^s \to \{0, 1\}$. There is an integer M with the following property. Let X, Y be disjoint sets of size at least M, and let $p_1, \ldots, p_s$ be real polynomials of degree at most H in the variables x_i for $i \in X$ and y_j for $j \in Y$. Let $G : \{0, \ldots, |X|\} \times \{0, \ldots, |Y|\} \to \{0, 1\}$, and suppose that*

$$G(|x|, |y|) = F\big(\operatorname{sgn} p_1(x, y), \ldots, \operatorname{sgn} p_s(x, y)\big) \tag{13}$$

for every $(x, y) \in \{0, 1\}^X \times \{0, 1\}^Y$. Then there are polynomials $q_1, \ldots, q_s \in \mathbb{R}[a, b]$, each of total degree at most H, such that

$$G(a, b) = F\big(\operatorname{sgn} q_1(a, b), \ldots, \operatorname{sgn} q_s(a, b)\big) \tag{14}$$

for all integers $0 \leq a \leq A$ and $0 \leq b \leq B$.

Proof. For $S \subseteq X$ and $T \subseteq Y$, color (S, T) by

$$c(S, T) = \big(\operatorname{sgn} p_1(\mathbf{1}_S^X, \mathbf{1}_T^Y), \ldots, \operatorname{sgn} p_s(\mathbf{1}_S^X, \mathbf{1}_T^Y)\big). \tag{15}$$

The color set is finite. Enumerate the finitely many pairs (a, b) in $\{0, \ldots, A\} \times \{0, \ldots, B\}$. Working backward through this list, choose the set sizes required at each step so that the Ramsey theorem for one pair leaves sets large enough for all later steps. Thus, for a sufficiently large M, successive applications to the restrictions of c to

$$\binom{X}{a} \times \binom{Y}{b}, \qquad 0 \leq a \leq A, \quad 0 \leq b \leq B$$

use the Product Ramsey theorem when $a, b > 0$ and the ordinary one-set Ramsey theorem when exactly one of a, b is zero. The pair $(0, 0)$ requires no restriction. After the final application, shrink

the two remaining sets to sizes A and B; all equalities of colors obtained earlier remain true. We therefore obtain sets $I \subseteq X$ and $J \subseteq Y$, with $|I| = A$ and $|J| = B$, such that

$$c(S,T) = c(S',T') \quad \text{whenever} \quad |S| = |S'|, \quad |T| = |T'| \tag{16}$$

for $S, S' \subseteq I$ and $T, T' \subseteq J$.

For each i, let $\widetilde{p}_i$ be obtained from p_i by setting $x_u = 0$ for $u \in X \setminus I$ and $y_v = 0$ for $v \in Y \setminus J$. Its degree is at most H. Apply Lemma 2.3 to $\widetilde{p}_i$, with the two sets in that lemma equal to I and J, and denote the polynomial it supplies by $q_i(a,b)$. Since $\mathbf{1}_S^X$ and $\mathbf{1}_T^Y$ are zero outside I and J when $S \subseteq I$ and $T \subseteq J$, respectively, we obtain

$$q_i(a,b) = \frac{1}{\binom{A}{a}\binom{B}{b}} \sum_{S \in \binom{I}{a}} \sum_{T \in \binom{J}{b}} p_i(\mathbf{1}_S^X, \mathbf{1}_T^Y) \tag{17}$$

for all integers $0 \le a \le A$ and $0 \le b \le B$.

By (16), all numbers being averaged in (17) have the same three-valued sign, for each fixed i, a, b. Their average has that same sign, including when the common sign is zero. Thus

$$\operatorname{sgn} q_i(a,b) = \operatorname{sgn} p_i(\mathbf{1}_S^X, \mathbf{1}_T^Y) \quad \text{whenever } S \in \binom{I}{a},\ T \in \binom{J}{b}.$$

Combining this identity with (13) proves (14). □

2.4 Conclusion

Proof of Theorem 1.1. Suppose that a transformer with $H < k$ heads computes $\mathsf{PowGraph}_{k,N}$ for every N. If $H = 0$, its output is constant, which is already impossible. Assume $H \ge 1$.

Fix the readout bit to zero. Proposition 2.1 supplies a fixed integer s, a fixed function $F : \{-1,0,1\}^s \to \{0,1\}$, and, at every length, polynomials of degree at most H satisfying (3).

Choose R sufficiently large for Proposition 2.2, and apply Lemma 2.4 to $G(a,b) = \mathbf{1}\{b \ge a^k\}$, with

$$A = R, \qquad B = R^k.$$

Let M_R be the threshold in that lemma. Choose N with $N - 1 \ge M_R$. After removing the readout position from whichever input block contains it, the sets X and Y of remaining positions in the two blocks both have size at least M_R. The fixed readout bit is zero, so the two full Hamming weights are exactly the weights on X and Y.

Lemma 2.4 therefore gives bivariate polynomials $q_1, \dots, q_s$, each of degree at most H, such that

$$F\big(\operatorname{sgn} q_1(a,b), \dots, \operatorname{sgn} q_s(a,b)\big) = \mathbf{1}\{b \ge a^k\} \tag{18}$$

for every integer pair $0 \le a \le R$, $0 \le b \le R^k$. This contradicts Proposition 2.2. □

3 The upper bound

Proof of Theorem 1.2. Choose distinct constants $c_1, \dots, c_k > 1$ and put

$$Q(t) = \prod_{j=1}^{k} (t + c_j), \qquad b_N = N^{-(k-1)}.$$

Write

$$A = |x|, \qquad B = |y|, \qquad u = \frac{A}{N}, \qquad v = \frac{B + \frac{1}{2}}{N^k}.$$

We arrange that head j has the following unnormalized attention weights:

	bit 0	bit 1
x_i	$c_j - b_N$	$c_j - b_N + 1$
y_i	b_N	b_N.

(19)

All weights are positive. Hence its denominator is

$$\sum_{i=1}^{N}(c_j - b_N + x_i) + \sum_{i=1}^{N} b_N = A + c_j N. \tag{20}$$

Put $\delta_i = i^k - (i-1)^k$. Give every x_i value zero and give the y_i token carrying bit a the value

$$w_{j,i,a} = -(-c_j)^k \delta_i + a + \frac{1}{2}\mathbf{1}\{i = 1\}. \tag{21}$$

Since every y_i has attention weight b_N, the numerator of head j is

$$\begin{aligned} b_N \sum_{i=1}^{N} w_{j,i,y_i} &= b_N \left(-(-c_j)^k \sum_{i=1}^{N} \delta_i + B + \frac{1}{2} \right) \\ &= -(-c_j)^k N + b_N \left(B + \frac{1}{2} \right), \end{aligned} \tag{22}$$

where we used $\sum_{i=1}^{N} \delta_i = N^k$ and $b_N N^k = N$. It follows from (20)–(22) that the head output $H_j = \mathcal{H}_{j,N}(z)$ is

$$H_j = \frac{-(-c_j)^k N + b_N(B + \frac{1}{2})}{A + c_j N} = \frac{v - (-c_j)^k}{u + c_j}. \tag{23}$$

The identity

$$-1 + \sum_{j=1}^{k} \frac{v - (-c_j)^k}{Q'(-c_j)(u + c_j)} = \frac{v - u^k}{Q(u)} \tag{24}$$

follows by multiplying by $Q(u)$: the difference is a polynomial of degree at most $k-1$ which vanishes at $-c_1, \ldots, -c_k$. Consequently, the fixed affine readout

$$S = -1 + \sum_{j=1}^{k} \frac{H_j}{Q'(-c_j)} \tag{25}$$

satisfies

$$S = \frac{v - u^k}{Q(u)} = \frac{B + \frac{1}{2} - A^k}{N^k Q(A/N)}. \tag{26}$$

The denominator in (26) is positive, and $B - A^k$ is an integer. Thus $S > 0$ exactly when $B \geq A^k$.

It remains to verify that (19) and (21) come from one length-independent joint encoding. We give the coordinates explicitly. For every head j, reserve three query coordinates, three key coordinates, and one value coordinate. At an odd position $2i-1$ carrying bit a, put in the jth block

$$q^{(j)} = (0, 0, 0), \qquad k^{(j)} = (1 - a, a, 0), \qquad v^{(j)} = 0. \tag{27}$$

At an even position $2i$ carrying bit a, put

$$\begin{aligned} q^{(j)} &= \big(\log(c_j - i^{-(k-1)}), \log(c_j - i^{-(k-1)} + 1), -(k-1)\log i\big), \\ k^{(j)} &= (0, 0, 1), \\ v^{(j)} &= -(-c_j)^k \big(i^k - (i-1)^k\big) + a + \frac{1}{2}\mathbf{1}\{i = 1\}. \end{aligned} \tag{28}$$

These formulas define a single map

$$e : \mathbb{N} \times \{0, 1\} \longrightarrow \mathbb{R}^{7k}.$$

Head j projects the readout embedding onto $q^{(j)}$, every embedding onto $k^{(j)}$, and every embedding onto $v^{(j)}$. At the readout position $2N$, the query in (28) is

$$\big(\log(c_j - b_N), \log(c_j - b_N + 1), \log b_N\big).$$

Its inner product with the three possible keys in (27)–(28) selects exactly the logarithm of the corresponding weight in (19). The query coordinates in (28) do not depend on the readout bit. Thus the encoding and all projections are fixed and work simultaneously for every N. □

AI TOOLS USAGE

Our lower bound builds on the algebraic ideas from an earlier preprint Kozachinskiy et al. (2026). The details of the proof took form over several weeks of conversations with GPT models, starting from version 5.4 up to 5.6. Generative tools were also used for polishing the text and proof-reading.

REFERENCES

Vatsal Baherwani, Zixi Chen, Shikai Qiu, Andrew Gordon Wilson, and Pavel Izmailov. Emergent capabilities arise randomly from learning sparse attention patterns. arXiv preprint arXiv:2606.25010, 2026. URL https://arxiv.org/abs/2606.25010.

Julia Böttcher and Jan Foniok. Ramsey properties of permutations. *The Electronic Journal of Combinatorics*, 20(1):P2, 2013. doi: 10.37236/2978. URL https://www.combinatorics.org/ojs/index.php/eljc/article/view/v20i1p2.

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. URL https://aclanthology.org/2022.tacl-1.46/.

Daniel Hsu. Lower bounds for one-layer transformers that compute parity, 2026. URL https://arxiv.org/abs/2605.12171.

Hongjian Jiang, Michael Hahn, Georg Zetzsche, and Anthony W. Lin. Softmax transformers are Turing-complete. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/48c218b01802a712ee8d7243d27fd9cf-Abstract-Conference.html.

Alexander Kozachinskiy, Felipe Urrutia, Hector Jimenez, Tomasz Steifer, Germán Pizarro, Matías Fuentes, Francisco Meza, Cristian B. Calderon, and Cristóbal Rojas. Strassen attention, split VC dimension and compositionality in transformers. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2501.19215.

Alexander Kozachinskiy, Tomasz Steifer, and Przemysław Wałęga. Parity, sensitivity, and transformers. arXiv preprint arXiv:2602.05896, 2026. URL https://arxiv.org/abs/2602.05896.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023. doi: 10.1162/tacl_a_00562. URL https://aclanthology.org/2023.tacl-1.31/.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Marvin L. Minsky and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, expanded edition, 1988. First published in 1969.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=KidynPuLNW.

Jorge Pérez, Javier Marinković, and Pablo Barceló. On the Turing completeness of modern neural network architectures. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1901.03429.

Rajmohan Rajaraman, Ravi Sundaram, and Amanuel Tesfaye. The head complexity of Boolean functions in single-layer attention. arXiv preprint arXiv:2609.04046, 2026. URL https://arxiv.org/abs/2609.04046.

Alexander A. Razborov and Alexander A. Sherstov. The sign-rank of AC^0. *SIAM Journal on Computing*, 39(5):1833–1855, 2010. doi: 10.1137/080744037. URL https://people.cs.uchicago.edu/~razborov/files/sign.pdf.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/73bf692447f174984f30499ec9b20e04-Abstract-Conference.html.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. One-layer transformers fail to solve the induction heads task. arXiv preprint arXiv:2408.14332, 2024. URL https://arxiv.org/abs/2408.14332.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? a survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024. doi: 10.1162/tacl_a_00663. URL https://aclanthology.org/2024.tacl-1.30/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Amanuel Tesfaye, Zeno Kujawa, Rajmohan Rajaraman, and Ravi Sundaram. Two (narrow) heads are better than (an arbitrarily wide) one. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RRmPbbZsvl.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.