THRESHOLD DEGREE AND THE HEAD COMPLEXITY OF SINGLE-LAYER ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A single layer of multi-head attention produces a vector representation from which a downstream predictor computes an output. We ask how many heads the layer needs when the predictor is a polynomial threshold function of degree D. For a symmetric Boolean function f of threshold degree T, we prove that H heads and a degree-D readout compute f if and only if $DH \geq T$. The lower bound allows arbitrary joint embeddings of token symbol and position, with unbounded embedding dimension and bit precision; the upper bound uses embeddings that depend on the token symbol alone, with value vectors of dimension at most $\min\{D,\ O(\log H)\}$. The characterization is therefore insensitive to the embedding class. We then extend it beyond symmetric functions. If $f(x) = F(|x_S|)$ for a designated subset $S \subseteq [n]$, the same characterization holds, with an upper bound that uses an additive embedding of symbol and position; such f is asymmetric whenever S is a proper subset of $[n]$, and no symbol-only embedding computes it at any H or D. Taking F to be parity gives a strict hierarchy in the head count at every fixed readout degree and input length: parity on m coordinates requires exactly $\lceil m/D \rceil$ heads. Finally, we show that threshold degree does not govern the general case: at most $2^{O(Hn^2)}$ functions are realizable with H heads and a linear readout, and we exhibit a family of threshold degree T in which almost every member requires $\Omega(n^{T-2})$ heads. Experiments on symmetric Boolean targets study the roles of D, H, and value-dimension in trained attention layers.

1 INTRODUCTION

Attention heads, which are the building blocks of transformers (Vaswani et al., 2017), have been successfully used to learn useful representations in a variety of scenarios (e.g., Lin et al., 2017; Parikh et al., 2016). In transformers, layers of several attention heads are composed together to form these representations, and the use of multiple layers seems to be necessary for several natural computational problems related to multi-step reasoning (Chen et al., 2024; Chiang et al., 2023; Liu et al., 2023; Merrill & Sabharwal, 2023; 2025; Sanford et al., 2024). This work focuses on the role of multiple heads within a single layer. We isolate a single attention layer in order to study head count separately from depth. This restricted setting permits exact upper and lower bounds, which are useful for bounding model size and broadly for mechanistic interpretability, where the individual head is an important unit of analysis.

It is known that a single attention head with output elements post-processed by a neural network can approximate essentially any function arbitrarily well, as long as the attention head's embedding dimension or bit precision level is high enough (e.g., Kajitsuka & Sato, 2024). So the necessity of multiple heads must arise from restrictions on aspects such as dimension, precision level, and the post-processing. Therefore, any account of head complexity must say what the readout is allowed to do. We take the readout to be a polynomial threshold function of degree D, and ask: how many attention heads H are needed to compute a target Boolean function $f\colon \{0,1\}^n \to \{0,1\}$? It turns out that, for symmetric target functions, the number of heads and degree are exactly interchangeable in their representational power, and the quantity they must jointly cover is the *threshold degree* $\text{thr-deg}(f)$ of the target.

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.

1.1 Our Contributions

Tight characterization for symmetric functions. Our first result is that if a single layer of H heads with a degree-D readout computes f, then $DH \geq \text{thr-deg}(f)$ (Theorem 1). The bound holds for arbitrary joint embeddings of symbol and position, at unbounded embedding dimension and unbounded precision. For the class of symmetric functions, the bound is attained at every pair (D, H) with $DH \geq \text{thr-deg}(f)$ by a construction whose embeddings depend on the token symbol alone, with value vectors of dimension $\min\{D, O(\log H)\}$ (Theorem 2). Thus, $DH \geq \text{thr-deg}(f)$ exactly characterizes computability for symmetric functions across all three embedding classes (Corollary 4.3). At $D = 1$ the characterization reads $H \geq \text{thr-deg}(f)$, and at $H = 1$ it reads $D \geq \text{thr-deg}(f)$.

Extension to symmetric juntas. We extend our characterization beyond symmetric functions to *symmetric juntas*. If $f(x) = F(|x_S|)$ for a designated $S \subseteq [n]$, then the same equivalence holds (Theorem 3), with an upper bound that uses an additive embedding of symbol and position. Such an f is not symmetric when S is a proper subset, and no content-only embedding computes it at any H or D. So the use of positional information changes an impossibility into the same bound that governs the symmetric case.

This extension gives a hierarchy at a fixed input length. Parity on a designated set S has threshold degree $|S|$, so it is computable by H heads with a degree-D readout exactly when $DH \geq |S|$. Writing $\mathcal{F}_{D,H}$ for the class of functions computed by a layer with these resources, we obtain

$$\mathcal{F}_{D,1} \subsetneq \mathcal{F}_{D,2} \subsetneq \cdots \subsetneq \mathcal{F}_{D,\lceil n/D \rceil},$$

strict at every step, with parity on any $DH + 1$ coordinates separating consecutive classes (Corollary 5.3). At $D = 1$ this says that parity on m coordinates requires exactly m heads.

General Boolean functions. Finally, we show that $DH \geq \text{thr-deg}(f)$ is not sufficient in general (Theorem 4). The number of functions a layer can compute is bounded by a sign-pattern argument, and it is far smaller than the number of functions of a given threshold degree. For a linear readout, almost every member of an explicit family of threshold degree T requires $\Omega(n^{T-2})$ heads. Our proof uses a non-constructive counting argument, so exhibiting an explicit function whose required head count exceeds its threshold degree is left open.

1.2 Our Techniques

For the lower bound, we first show that both the numerator and the denominator of the computation that captures an attention head are affine in the input, because each input coordinate is a bit. A head therefore contributes a ratio of affine forms with a positive denominator; this representation is also used in the prior work of Kozachinskiy et al. (2026); Hsu (2026). We then show that clearing the H denominators to the D-th power is sign-preserving and produces a polynomial of degree at most DH that sign-represents f. Since any polynomial that can sign-represent f needs to have degree at least T, this yields the lower bound on the product of D and H.

The upper bound for symmetric targets follows from a sequence of arguments. First, we show that a content-only head is linear-fractional in the Hamming weight. The induced denominators give a basis of the univariate polynomials of degree at most H by partial fractions, and the sign-representing polynomial is then assembled from them by the readout. A natural way to implement this approach is via one factor per coordinate, which requires a dimension of D. A more sophisticated implementation indexes the coordinates by an "additive D-basis" of $[H]$ in the sense of Mossige (1987); Challis & Robinson (2010), and padding with a constant coordinate so that the readout is homogeneous; a dimension of $O(\log H)$ suffices. For the extension to juntas, our approach suppresses the irrelevant positions by a finite additive offset in the score, which perturbs each head by $O(ne^{-M})$ and so preserves a positive decision margin, yielding the desired result.

Finally, the insufficiency result for general Boolean functions is based on Warren's sign-pattern bound (Warren, 1968), which we apply to the same cleared polynomial used in the lower bound.

2 MODEL AND PRELIMINARIES

Attention layer. The input is a string $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, presented as n tokens. Using row-vector notation, let $X \in \mathbb{R}^{n \times m}$ have rows $\Phi(x_i, i)$, and let $c \in \mathbb{R}^m$ (as a row vector) be the fixed embedding of a special `[CLS]` token.[1] Multi-head attention produces

$$z(x) \;=\; \sum_{h=1}^{H} \mathrm{softmax}(cA_hX^\top)\, XV_h \;\in\; \mathbb{R}^d,$$

where $A_h = W_{Q,h}W_{K,h}^\top \in \mathbb{R}^{m \times m}$, $W_{Q,h}$ and $W_{K,h}$ project embeddings to queries and keys, and $V_h \in \mathbb{R}^{m \times d}$ maps each token embedding to its value vector for head h.

For each head $h \in [H]$, symbol $b \in \{0,1\}$, and position $i \in [n]$, define the score and value vector

$$s_h(b,i) = cA_h\Phi(b,i)^\top \in \mathbb{R}, \qquad v_h(b,i) = \Phi(b,i)V_h \in \mathbb{R}^d.$$

We work directly with these scores and value vectors below. Writing $r_h(b,i) = \exp(s_h(b,i)) > 0$ for the unnormalized attention weight, the contribution of the h-th head is

$$\mathrm{Att}_h(x) \;=\; \frac{\sum_{i=1}^n r_h(x_i,i)\, v_h(x_i,i)}{\sum_{i=1}^n r_h(x_i,i)} \;\in\; \mathbb{R}^d. \tag{1}$$

A *degree-D readout* is a polynomial $g : \mathbb{R}^d \to \mathbb{R}$ of total degree at most D. The layer *computes* $f : \{0,1\}^n \to \{-1,1\}$ if

$$\mathrm{sign}\Big(g\Big(\textstyle\sum_{h=1}^H \mathrm{Att}_h(x)\Big)\Big) \;=\; f(x) \qquad \text{for every } x \in \{0,1\}^n, \tag{2}$$

i.e., f is computed by post-processing the sum of the attention heads with a degree-D polynomial threshold function. Our focus on this class of readout functions allows us to establish sharp results, and it also serves as a starting point for lower bounds against other classes of readout functions (e.g., ReLU networks) by leveraging rational function approximation (see, e.g., Hsu, 2026, Section 4). We impose no restriction on the embedding dimension m, the value dimension d, or the bit precision of the parameters unless stated otherwise.

Three embedding classes. Our results consider the following three classes of embeddings.

- A *joint* embedding places no constraint on s_h and v_h: both may depend arbitrarily on the pair (symbol, position).
- An *additive* embedding takes the form $\Phi(b,i) = \tau_b + \pi_i$, so that a score decomposes as $s_h(b,i) = \Delta_h b + c_h(i)$: the symbol contributes a position-independent slope.
- A *content-only* (or *symbol-only*) embedding lets s_h and v_h depend on the symbol alone, $s_h(b,i) = s_h(b)$ and $v_h(b,i) = v_h(b)$.

We have content-only $\subseteq$ additive $\subseteq$ joint. Our lower bound is proved for joint embeddings and our upper bound for content-only ones.

Threshold degree. The *threshold degree* $\mathrm{thr\text{-}deg}(f)$ of $f : \{0,1\}^n \to \{-1,1\}$ is the least degree of a real polynomial p with $\mathrm{sign}(p(x)) = f(x)$ for all $x \in \{0,1\}^n$.

Symmetric functions and alternations. We say $f : \{0,1\}^n \to \{-1,1\}$ is *symmetric* if $f(x) = F(|x|)$ for a predicate $F : \{0,1,\ldots,n\} \to \{-1,1\}$, where $|x|$ is the Hamming weight. Let

$$\mathrm{Alt}(F) \;=\; \#\{\, w \in \{0,\ldots,n-1\} : F(w) \neq F(w+1) \,\}$$

be the number of sign alternations of F. By Minsky–Papert symmetrization (Minsky & Papert, 1969), $\mathrm{thr\text{-}deg}(f) = \mathrm{Alt}(F)$ for symmetric f; the symmetrization is strict because the decision margin in (2) is positive. Parity has $\mathrm{Alt}(F) = n$, a threshold test has $\mathrm{Alt}(F) = 1$, and an interval test has $\mathrm{Alt}(F) = 2$.

[1] We adopt the BERT convention (Devlin et al., 2018) of assuming that a special `[CLS]` token is always appended to the input, and we are only concerned with the query associated with that token.

3 THE HEAD–DEGREE LAW

The following lower bound holds for every Boolean function and for arbitrary joint embeddings, at unbounded embedding dimension and unbounded precision.

Theorem 1 (Head–degree law). *Let $f : \{0,1\}^n \to \{-1,1\}$ have threshold degree T. If a single layer of H attention heads with a joint embedding and a degree-D readout computes f in the sense of (2), then $DH \geq T$.*

Proof. Fix $h \in [H]$. Since $x_i \in \{0,1\}$, we may write $r_h(x_i, i) = r_h(0,i) + x_i\big(r_h(1,i) - r_h(0,i)\big)$, and likewise for $r_h(x_i,i)v_h(x_i,i)$. Summing over i, the denominator and each coordinate of the numerator of (1) are affine functions of x. Write

$$\mathrm{Att}_h(x) \;=\; \frac{N_h(x)}{d_h(x)},$$

where $N_h : \mathbb{R}^n \to \mathbb{R}^d$ is affine coordinatewise and d_h is affine and strictly positive on $\{0,1\}^n$, the positivity because every $r_h(b,i) > 0$.

Let $Q(x) = \prod_{h=1}^{H} d_h(x)$, a polynomial of degree at most H that is strictly positive on $\{0,1\}^n$. For each coordinate $j \in [d]$ of the sum,

$$\Big(\sum_{h=1}^{H} \mathrm{Att}_h(x)\Big)_j \;=\; \sum_{h=1}^{H} \frac{N_{h,j}(x)}{d_h(x)} \;=\; \frac{P_j(x)}{Q(x)}, \qquad P_j \;=\; \sum_{h=1}^{H} N_{h,j} \prod_{h' \neq h} d_{h'},$$

and $\deg P_j \leq H$. Write $g = \sum_m c_m \prod_j z_j^{m_j}$, the sum over exponent vectors m with $|m| = \sum_j m_j \leq D$. Substituting $z_j = P_j/Q$ and multiplying by Q^D,

$$Q(x)^D\, g\Big(\sum_h \mathrm{Att}_h(x)\Big) \;=\; \sum_m c_m \Big(\prod_j P_j(x)^{m_j}\Big)\, Q(x)^{D-|m|} \;=:\; \Pi(x).$$

Each summand has degree at most $|m|H + (D - |m|)H = DH$, so $\deg \Pi \leq DH$. Since $Q^D > 0$ on $\{0,1\}^n$, multiplying by it preserves signs, so $\operatorname{sign} \Pi(x) = f(x)$ for every $x \in \{0,1\}^n$ by (2).

Finally, reducing Π modulo the relations $x_i^2 = x_i$ yields a multilinear polynomial that agrees with Π on $\{0,1\}^n$ and has degree at most $\deg \Pi$. This multilinear polynomial sign-represents f, so $T = \text{thr-deg}(f) \leq DH$. □

Remark 3.1. Neither the embedding dimension m, the value dimension d, nor the precision of the parameters affects the bound, and no property of the embedding is used beyond $r_h(b,i) > 0$.

4 MATCHING CONSTRUCTIONS FOR SYMMETRIC FUNCTIONS

We give a construction that matches the lower bound from Theorem 1, using content-only embeddings (the most restrictive of the three classes described in Section 2). Throughout this section $f(x) = F(|x|)$ is symmetric and non-constant, and $T = \text{thr-deg}(f) = \mathrm{Alt}(F)$.

4.1 PRELIMINARY FACTS

We first record two facts about attention heads with content-only embeddings.

Lemma 4.1 (Normal form of a head). *Let $c > -1$ with $c \neq 0$, and let $A, B \in \mathbb{R}^d$. There is a single-head solution h with content-only embedding that satisfies*

$$\mathrm{Att}_h(x) \;=\; A + \frac{B}{q(|x|)}, \qquad q(t) = n + ct,$$

and q is strictly positive on $\{0, 1, \ldots, n\}$.

Proof. Put $\lambda = 1 + c > 0$ and give the head unnormalized weight λ on the symbol 1 and 1 on the symbol 0, which is a content-only choice. Writing $t = |x|$, the denominator of (1) is $t\lambda + (n - t) = n + ct = q(t)$ and the numerator is $n\,v(0) + t\big(\lambda v(1) - v(0)\big)$. Setting

$$v(0) \;=\; A + \frac{B}{n}, \qquad v(1) \;=\; A + \frac{B}{n\lambda},$$

the numerator becomes $A\,q(t) + B$, since the constant terms give $nv(0) = An + B$ and the coefficients of t give $\lambda v(1) - v(0) = Ac$. Dividing by $q(t)$ gives the claim. Finally q is affine with $q(0) = n > 0$ and $q(n) = n\lambda > 0$, so $q > 0$ on $[0, n]$. □

Lemma 4.2 (Realization). *Let $c_1, \dots, c_H > -1$ be distinct and nonzero, put $q_h(t) = n + c_h t$, and set $Q = \prod_h q_h$ and $Q_h = \prod_{h' \neq h} q_{h'}$. Then $\{Q, Q_1, \dots, Q_H\}$ is a basis of the space of univariate real polynomials of degree at most H. Consequently, for any $P_1, \dots, P_d$ of degree at most H there is a layer of H heads using content-only embeddings and values in $\mathbb{R}^d$ such that, for all $j \in [d]$, $\sum_{h=1}^{H} \mathrm{Att}_h(x)_j = P_j(t)/Q(t)$ for $t = |x|$; and $Q > 0$ on $\{0, \dots, n\}$.*

Proof. There are $H + 1$ polynomials and the space has dimension $H + 1$, so it is enough to prove independence. Let $\rho_h = -n/c_h$ be the root of q_h; the ρ_h are distinct because the c_h are distinct and nonzero. Suppose $\alpha_0 Q + \sum_h \alpha_h Q_h = 0$ and evaluate at ρ_h. Both Q and, for $h' \neq h$, $Q_{h'}$ carry q_h as a factor and vanish there, while $Q_h(\rho_h) = \prod_{h' \neq h} q_{h'}(\rho_h) \neq 0$. Hence $\alpha_h = 0$ for every h, and then $\alpha_0 Q = 0$ gives $\alpha_0 = 0$.

For the consequence, expand $P_j = a_j Q + \sum_h b_{h,j} Q_h$ in this basis and apply Lemma 4.1 to head h with parameter c_h, vector $B_h = (b_{h,1}, \dots, b_{h,d})$, and $A_1 = (a_1, \dots, a_d)$, $A_h = 0$ for $h \geq 2$. Summing over h, coordinate j is $a_j + \sum_h b_{h,j}/q_h = P_j/Q$. Positivity of Q is from Lemma 4.1. □

Lemma 4.2 delivers d polynomials of degree at most H over a common positive denominator, and the readout must assemble them into a sign-representation of F. Doing this naively costs one value coordinate per factor. In Appendix C, we show how to use $O(\log H)$ coordinates instead, by letting the readout reuse a coordinate through its exponent.

4.2 The construction

Theorem 2 (Tightness on the symmetric class). *Let f be symmetric with threshold degree T. For every pair of positive integers D, H with $DH \geq T$ there is a single layer of H attention heads with a* content-only *embedding, together with a degree-D readout, computing f. The value vectors may be taken of dimension*

$$d \;\leq\; \min\big\{\, D,\; 2\max\{1, \lceil \log_2 H \rceil\} + 2 \,\big\}.$$

Proof. Let $t_1 < \cdots < t_T$ be the points at which F alternates, so $F(t_i) \neq F(t_i + 1)$; put $r_i = t_i + \frac{1}{2}$. Let $R(t) = \varepsilon \prod_{i=1}^{T} (t - r_i)$ with $\varepsilon \in \{-1, 1\}$ chosen so that $\operatorname{sign} R(0) = F(0)$. No r_i is an integer, and as t runs through $0, 1, \dots, n$ the product changes sign exactly at the T alternations of F, so

$$\operatorname{sign} R(t) \;=\; F(t) \qquad \text{for } t \in \{0, \dots, n\}. \tag{3}$$

Fix distinct nonzero $c_1, \dots, c_H > -1$ and let Q be as in Lemma 4.2. We give two constructions; the theorem takes whichever has the smaller value dimension.

First construction, $d \leq D$. Since $T \leq DH$, partition $\{r_1, \dots, r_T\}$ into $d = \lceil T/H \rceil \leq D$ consecutive blocks $G_1, \dots, G_d$ of size at most H, and set $R_j = \prod_{r \in G_j} (t - r)$ for $j \geq 2$ and $R_1 = \varepsilon \prod_{r \in G_1} (t - r)$, so that $\prod_j R_j = R$ and $\deg R_j \leq H$. Apply Lemma 4.2 with $P_j = R_j$ and take the readout $g(z_1, \dots, z_d) = z_1 \cdots z_d$, of degree $d \leq D$. Then

$$g\Big(\sum_h \mathrm{Att}_h(x)\Big) \;=\; \prod_{j=1}^{d} \frac{R_j(t)}{Q(t)} \;=\; \frac{R(t)}{Q(t)^d}.$$

Second construction, $d \leq 2\max\{1, \lceil \log_2 H \rceil\} + 2$. This part is deferred to Appendix D.

In both cases the readout has overall degree at most D and its value is $R(t)$ divided by a positive power of $Q(t)$. By Lemma 4.2 $Q > 0$ on $\{0, \dots, n\}$, so the sign is that of $R(t)$, which is $F(t) = f(x)$ by (3). □

Corollary 4.3 (Characterization)**.** *For symmetric f of threshold degree T, a single layer of H heads with a degree-D readout computes f if and only if $DH \geq T$. The "only if" direction holds for joint embeddings and the "if" direction is attained by content-only embeddings, so the same characterization holds for every embedding class between the two.*

Remark 4.4. At $D = 1$ the characterization reads $H \geq T$: a linear readout needs one head per alternation. At $H = 1$ it reads $D \geq T$: a single head with a degree-T readout computes every symmetric function of threshold degree T, and no smaller degree does.

A lower bound on the value-dimension needed for some target function of threshold degree DH is given in Appendix B; we leave open the problem of closing a small gap between this lower bound and the upper bound in Theorem 2.

5 Symmetric juntas and an exact head hierarchy

The upper bound of Theorem 2 uses no positional information. We now show that the characterization still holds for some non-symmetric families at the cost of using positional information, and that this step is what supplies a witness of every threshold degree at every fixed input length.

Definition 5.1 (Symmetric junta)**.** For $S \subseteq [n]$ and a predicate $F : \{0, \dots, |S|\} \to \{-1, 1\}$, let $f(x) = F(|x_S|)$, where x_S is the restriction of x to the coordinates in S. If S is a proper subset of $[n]$ and F is non-constant, then f is not symmetric.

Theorem 3. *Let $f(x) = F(|x_S|)$ with F non-constant, and let $T = \text{thr-deg}(f)$. Then*

(i) $T = \text{Alt}(F)$;

(ii) for every D, H with $DH \geq T$ there is a single layer of H heads with an additive *embedding and a degree-D readout computing f, and by Theorem 1 no (D, H) with $DH < T$ does;*

(iii) if $S \neq [n]$, then no attention layer with content-only embedding computes f, at any H, any D, any embedding dimension and any precision.

Proof. (i) Fixing the coordinates outside S to constants leaves the symmetric function $F(|x_S|)$ on $|S|$ variables, whose threshold degree is $\text{Alt}(F)$. Threshold degree is non-increasing under restriction, so $T \geq \text{Alt}(F)$. Conversely a degree-$\text{Alt}(F)$ sign-representation in the variables of S sign-represents f, so $T \leq \text{Alt}(F)$.

(iii) The output of a content-only layer depends on x only through the multiset of its tokens, hence only through $|x|$. It therefore suffices to exhibit x, y with $|x| = |y|$ and $f(x) \neq f(y)$. Since F is non-constant there is an $a \in \{0, \dots, |S| - 1\}$ with $F(a) \neq F(a + 1)$. Let x have $a + 1$ ones, all inside S, and let y have a ones inside S and one one outside S, which is possible because $S \neq [n]$. Then $|x| = |y| = a + 1$ while $f(x) = F(a + 1) \neq F(a) = f(y)$.

(ii) Let $k = |S|$. Apply Theorem 2 to F on k bits, obtaining H content-only heads, with symbol weights $\lambda_h > 0$ and values $v_h(0), v_h(1)$, and a degree-D readout g whose decision margin is $\gamma = \min_{0 \leq a \leq k} |g(\sum_h \text{Att}_h^S(a))| > 0$, where $\text{Att}_h^S(a)$ is the head output on an input of S-weight a.

Define a layer on n tokens by the scores $s_h(b, i) = b \ln \lambda_h + c(i)$, where $c(i) = 0$ for $i \in S$ and $c(i) = -M$ for $i \notin S$, and the position-blind (i.e., content-only) values $v_h(b, i) = v_h(b)$. The score is a sum of a symbol term and a position term, so the embedding is additive; positional information enters through the score only, never through the values.

Write $a = |x_S|$. Separating the two groups of positions,

$$\text{Att}_h(x) = \frac{D_h(a)\ \text{Att}_h^S(a) + e^{-M}\nu_h}{D_h(a) + e^{-M}\beta_h}, \qquad D_h(a) = a\lambda_h + (k - a),$$

where $\beta_h = \sum_{i \notin S} \lambda_h^{x_i}$ and $\nu_h = \sum_{i \notin S} \lambda_h^{x_i} v_h(x_i)$. Subtracting,

$$\mathrm{Att}_h(x) - \mathrm{Att}_h^S(a) = \frac{e^{-M}\big(\nu_h - \mathrm{Att}_h^S(a)\beta_h\big)}{D_h(a) + e^{-M}\beta_h}.$$

Put $V = \max_{h,b} \|v_h(b)\|$ and $\Lambda = \max_h \max\{1, \lambda_h\}/\min\{1, \lambda_h\}$. Then $\|\nu_h\| \le n \max\{1, \lambda_h\} V$, $\beta_h \le n \max\{1, \lambda_h\}$ and $\|\mathrm{Att}_h^S(a)\| \le V$, since a head output is a convex combination of its values; and $D_h(a) \ge k \min\{1, \lambda_h\}$. Hence

$$\Big\| \sum_h \mathrm{Att}_h(x) - \sum_h \mathrm{Att}_h^S(a) \Big\| \;\le\; e^{-M}\, \frac{2nH\Lambda V}{k}.$$

Let L be a Lipschitz constant for g on the (compact) ball containing all these points. Taking $M \ge \ln(2nH\Lambda V L)/(\gamma k)$ makes the perturbation of g smaller than γ, so $\operatorname{sign} g(\sum_h \mathrm{Att}_h(x)) = \operatorname{sign} g(\sum_h \mathrm{Att}_h^S(a)) = F(a) = f(x)$ for every x. The required M is finite and $O(\log n + \log(1/\gamma))$ for fixed construction parameters. □

Remark 5.2. The upper bound of Theorem 2 is content-only and the upper bound of Theorem 3 is additive, and the characterization is the same in both. Moving from the symmetric functions to the symmetric juntas moves the construction one step up the embedding hierarchy of Section 2 and does not change the bound.

5.1 An exact hierarchy in the head count

Let $\mathcal{F}_{D,H}$ denote the class of functions $\{0,1\}^n \to \{-1,1\}$ computable by a single layer of H heads with a degree-D readout. Adding a head with equal scores and equal values contributes a constant vector, which may be absorbed into the readout, so $\mathcal{F}_{D,H} \subseteq \mathcal{F}_{D,H+1}$. The following shows that every one of these containments is strict, and locates the witness.

For $S \subseteq [n]$ write $\mathrm{PAR}_S(x) = (-1)^{|x_S|}$ for parity on the coordinates of S.

Corollary 5.3 (Exact head hierarchy)**.** *Fix n and a readout degree D. For every $S \subseteq [n]$, $\mathrm{PAR}_S \in \mathcal{F}_{D,H}$ if and only if $DH \ge |S|$. Consequently*

$$\mathcal{F}_{D,1} \subsetneq \mathcal{F}_{D,2} \subsetneq \cdots \subsetneq \mathcal{F}_{D,\lceil n/D \rceil},$$

where parity on any $DH+1$ coordinates lies in $\mathcal{F}_{D,H+1} \setminus \mathcal{F}_{D,H}$.

Proof. PAR_S is the symmetric junta with $F(a) = (-1)^a$, so $\mathrm{Alt}(F) = |S|$ and $\text{thr-deg}(\mathrm{PAR}_S) = |S|$ by Theorem 3(i). The equivalence is then Theorem 3(ii) together with Theorem 1. For strictness take $|S| = DH + 1$: this exceeds DH, and $D(H+1) \ge DH + 1$. □

Remark 5.4. At $D = 1$ the corollary reads: parity on m coordinates requires exactly m heads, and the classes $\mathcal{F}_{1,1} \subsetneq \cdots \subsetneq \mathcal{F}_{1,n}$ are strictly increasing. The hierarchy at a general readout degree is the same hierarchy indexed by the product DH: raising the readout degree by one is worth H additional heads, and raising the head count by one is worth D additional degrees.

It would be interesting to identify further classes beyond symmetric juntas for which $DH \ge \text{thr-deg}(f)$ is sufficient. Inner product modulo two on $2n$ bits is a concrete test case: its threshold degree is n, but whether $DH \ge n$ suffices remains open.

6 Threshold degree does not govern the general case

Theorem 1 lower bounds DH by $\text{thr-deg}(f)$ for every f, and Corollary 4.3 and Theorem 3 show the bound is attained on the symmetric functions and on the symmetric juntas. We now show, using a counting argument, that $DH \ge \text{thr-deg}(f)$ is not sufficient in general, and not by a small margin.

Theorem 4. *Fix $T \ge 3$. With a linear readout, a $1 - o(1)$ fraction of the $2^{\binom{n}{T}}$ functions of Lemma E.2 require $\Omega(n^{T-2})$ heads, although each has threshold degree at most T. More generally, for any D, H with $D(H-1) > 1$ and $DH \ge 4$, a $1 - o(1)$ fraction of the functions of threshold degree at most DH in that family lie outside $\mathcal{F}_{D,H}$.*

The proof is deferred to Appendix E. We leave open the problem of constructing a witness outside of $\mathcal{F}_{D,H}$ with threshold degree DH.

7 EXPERIMENTS

7.1 ON THE VALUE DIMENSION

In this subsection, we focus on understanding the tightness of $d \le \min\{D, 2\max\{1, \lceil \log_2(H) \rceil\} + 2\}$ upper bound on the value vector dimension. Our target functions of some threshold degree T are constructed as follows. For $x \in \{0,1\}^n$ with Hamming-weight $k = |x|$, define permutation-invariant target function $f_S(x) = -1$ iff $k \in S$ for some latent subspace $S \subseteq [n]$, and $+1$ otherwise. We train a transformer of a fixed-query H-headed self-attention layer with an output dimension d, post-processed by a full degree-D polynomial threshold function. No positional encoding, residual connection, or quantization is imposed. For simplicity, we choose the key/query embedding dimension $d_0 = 1$. We set $n = 64$. Since we focus on the transformer's expressiveness, we enumerate training data deterministically. Our training data comprise all 65 Hamming-weight classes $k = 0, \dots, 64$. Each sample is a representative sequence $0^{64-k}1^k$ of the class k labeled by f_S. We minimize full-batch logistic loss averaged over all 65 classes using Adam. We test the following three target functions.

$$S_1 = [25, 28] \cup [33, 56], \qquad S_2 = [23, 29] \cup [35, 41],$$
$$S_3 = [18, 20] \cup [23, 24] \cup [28, 31] \cup \{35\} \cup [39, 43] \cup [47, 52].$$

We make 100 runs per task with independent initializations on attentions and readouts. Here, we omit the details for initialization. Training epoch caps for $1, 2, 3$ are $10^5, 10^4, 3 \cdot 10^4$, respectively. At test time, we check the equally weighted accuracy over all 65 classes of Hamming-weights. The results are presented in Table 1.

Target (T, D, H)	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$
f_{S_1} $(4, 2, 2)$	89.65 ± 0.75	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
f_{S_2} $(4, 2, 2)$	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
f_{S_3} $(12, 3, 4)$	69.37 ± 4.24	86.11 ± 6.52	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Table 1: $\mathrm{Accu}(\hat{f}) = \frac{1}{65}\sum_{k=0}^{64} \mathbb{1}\{f(k)\,\hat{f}(k) > 0\}$ (mean $\pm$ std over 100 independent runs).

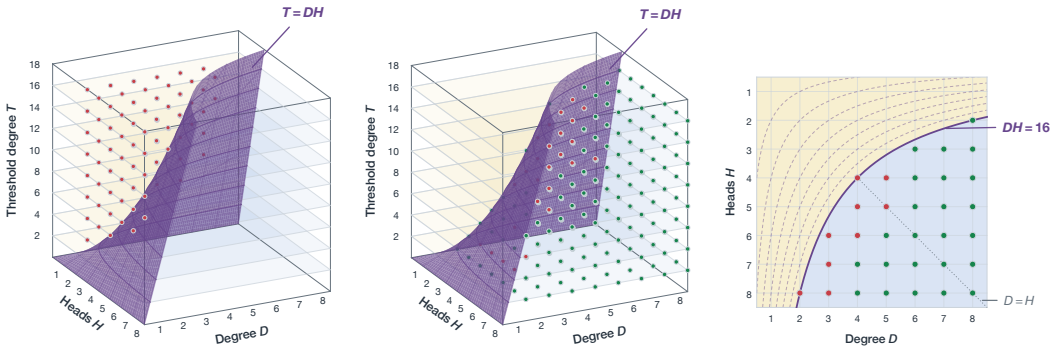
Note that, in Table 1, the transformer achieves 100% mean accuracy for S_1 and S_3 only at the tested value dimensions $d \ge D$, whereas it achieves 100% mean accuracy for S_2 with $d = 1$ when $D = 2$. Compared to S_1, the label pattern (over Hamming weights) of S_2 is symmetric. This suggests that additional structural symmetry could relax the value dimension resource. (Another convincing example is the parity function, where $d \le \min\{D/2, H/2\}$ suffices.) Moreover, we further show in Appendix B, the upper bound $d = \min\{D, O(\log H)\}$ is nearly tight in the worst case.

7.2 ON THE HEAD–DEGREE TRADEOFF

We use the same model architecture and optimizer (Adam) as above. We set $n = 100$, fix the value dimension to $d = 4$, and train and evaluate over all 101 Hamming weights. For each $T \in \{2, 4, \dots, 16\}$, we choose S so that the target function has T approximately equally spaced label flips over the Hamming weights. We vary $D, H \in \{1, \dots, 8\}$. We use comparable training budgets at each T, with additional pretraining attempts for configurations that do not achieve a perfect fit in the initial parameter sweep.

The results are presented in Figure 1. Each dot represents a configuration (T, D, H): green indicates that at least one attempt achieves perfect classification, while red indicates that no such fit was found within the hyperparameter search budget starting from random initialization. The separate below-boundary comparison in panel (a) shows only red dots, consistent with Theorem 1. For $DH \ge T$, $347/378$ configurations (91.8%) achieve a perfect fit, including $331/354$ (93.5%) with $DH > T$ and $16/24$ (66.7%) with $DH = T$. A pretraining follow-up recovers 17 of the 48 configurations that failed in the initial sweep. Panel (c) shows that training success is asymmetric in D and H at fixed $T = 16$, with a higher perfect-fit rate when $D > H$ than when $D < H$.

We also compare polynomial readouts with GELU MLP readouts of similar parameter counts, finding higher zero-error success rates for polynomial readouts on the harder targets with multiple attention heads. Appendix A presents the comparison and discusses its empirical support for using degree-D polynomial readouts.



(a) $DH < T$ (b) $DH \geq T$ (c) $T = 16$

Figure 1: Head–degree sweep over multiple initializations ($d = 4$). Yellow and blue regions indicate $DH < T$ and $DH \geq T$, respectively. (a) No perfect fits are found for $DH < T$, as required by Theorem 1. (b) For $DH \geq T$, 91.8% of configurations achieved a perfect fit in at least one training attempt under the prescribed training budgets using Adam. (c) Top-down view of (b) at $T = 16$: perfect-fit rates are 57.1% for $D < H$ and 92.9% for $D > H$.

8 Related work

This paper is the merger of (parts of) two separate, unpublished preprints by the present authors (Anonymized Authors 1, 2026; Anonymized Authors 2, 2026).

Parity and low-degree representations. Parity is a classical obstruction for shallow models and low-order representations (Minsky & Papert, 1969; Paturi & Saks, 1994). The threshold degree of a symmetric function equals its number of sign alternations, by Minsky–Papert symmetrization; parity is the extreme case, and it is the witness that separates consecutive classes in Corollary 5.3.

In transformers, parity and related formal languages have served as canonical tests of expressivity and learnability (Bhattamishra et al., 2020; Hao et al., 2022; Strobl et al., 2024; Kozachinskiy et al., 2026). Hahn (2020) proved length-asymptotic limitations of self-attention on parity-like languages, and Chiang & Cholak (2022) showed that additional depth and feed-forward net readouts can overcome them. Most directly related to our work, Kozachinskiy et al. (2026) showed that a single attention layer with a fixed size (independent of n) ReLU-net readout cannot compute parity of n bits for sufficiently large n; this is complementary to our work because we consider attention layers and readouts of complexity possibly growing with n. Hsu (2026) gives a lower bound for one-layer transformers computing parity with polynomial threshold readouts, although their result is not sharp; and Viswanathan (2026) gives a lower bound for two-bit parity with linear readouts, a special case of our general result.

Universality of attention heads. A single attention head whose output is post-processed by a sufficiently rich network is close to universal (Yun et al., 2020; Kajitsuka & Sato, 2024), so head count is not by itself a meaningful measure. This is the reason the readout in (2) is a polynomial threshold function of a declared degree and that our results involve both D and H.

Depth, and the single-layer question. Multiple layers are necessary for a range of tasks related to multi-step reasoning (Merrill & Sabharwal, 2023; Liu et al., 2023; Chiang et al., 2023; Sanford et al., 2023; Peng et al., 2024; Sanford et al., 2024; Kozachinskiy, 2024; Chen et al., 2024; Kozachinskiy et al., 2025; Ye et al., 2026), and chain-of-thought decoding serializes depth over decoding time (Merrill & Sabharwal, 2024; Li et al., 2024; Wei et al., 2022). The question studied here is orthogonal: it fixes the depth at one and varies the width.

Head count separations. Recent works show fine-grained representability separations in terms of the number of heads in a single attention layer. Tesfaye et al. (2026) identify a task that separates two narrow heads from a single arbitrarily wide one, both using linear readouts but without any other constraints on the attention parameters. Yu et al. (2026) gives, for all $k \geq 1$, families of tasks that separate $k + 1$ heads from k heads with Lipschitz ReLU-net readouts.

USE OF GENERATIVE AI

Generative AI tools were used to improve the clarity and presentation of the writing and proofs, to assist with the merger of the preprints, to assist with literature search, and to carry out numerical verification of the constructions in Theorem 2 and Theorem 3. The authors reviewed and verified all resulting content, including every theorem statement and proof, and take full responsibility for the manuscript.

REFERENCES

Anonymized Authors 1. Anonymized Title 1, 2026.

Anonymized Authors 2. Anonymized Title 2, 2026.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.576.

Michael F Challis and John P Robinson. Some extremal postage stamp bases. *Journal of Integer Sequences*, 13(2):3, 2010.

Lijie Chen, Binghui Peng, and Hongxun Wu. Theoretical limitations of multi-layer transformer, 2024. URL https://arxiv.org/abs/2412.02975.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. doi: 10.18653/v1/2022.acl-long.527.

David Chiang, Peter Cholak, and Anand Pillay. Tighter bounds on the expressivity of transformer encoders, 2023. URL https://arxiv.org/abs/2301.10743.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995.

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, 2022. URL https://arxiv.org/abs/2204.06618.

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

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *International Conference on Learning Representations (ICLR)*, 2024.

Alexander Kozachinskiy. Lower bounds on transformers with infinite precision, 2024. URL https://arxiv.org/abs/2412.20195.

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 (NeurIPS)*, 2025.

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

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding, 2017. URL https://arxiv.org/abs/1703.03130.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations (ICLR)*, 2023.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations (ICLR)*, 2024.

William Merrill and Ashish Sabharwal. A little depth goes a long way: The expressive power of log-depth transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

Svein Mossige. On extremal h-bases A_4. *Mathematica Scandinavica*, pp. 5–16, 1987.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference, 2016. URL https://arxiv.org/abs/1606.01933.

Ramamohan Paturi and Michael E. Saks. Approximating threshold circuits by rational functions. *Information and Computation*, 112(2):257–272, 1994. doi: 10.1006/inco.1994.1059.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture. In *First Conference on Language Modeling (COLM)*, 2024.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

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.

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 (ICLR)*, 2026. URL https://openreview.net/pdf?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 (NeurIPS)*, 2017.

Karthik Viswanathan. How many attention heads do you need to do XOR? LessWrong, April 2026. URL https://www.lesswrong.com/posts/T66BKwSufh5SfiPHm/how-many-attention-heads-do-you-need-to-do-xor-3.

Hugh E. Warren. Lower bounds for approximation by non-linear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Qilin Ye, Deqing Fu, Robin Jia, and Vatsal Sharan. Transformers provably learn algorithmic solutions for graph connectivity, but only with the right data. In *International Conference on Machine Learning (ICML)*, 2026.

Penghao Yu, Haotian Jiang, Zeyu Bao, Ruoxi Yu, and Qianxiao Li. The effect of attention head count on transformer approximation. In *International Conference on Learning Representations (ICLR)*, 2026.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations (ICLR)*, 2020.

A POLYNOMIAL VERSUS MLP READOUTS

Degree-D polynomial readouts make the interaction between attention heads and readout complexity explicit through the bound $DH \geq T$. We assess their empirical relevance against GELU MLP readouts at matched and larger parameter budgets on the symmetric tasks in Section 7.

Setup and parameter matching. We set $n = 100$, $T \in \{6, 8, \ldots, 20\}$, and $H \in \{1, \ldots, 8\}$, using targets with T approximately equally spaced label flips over the 101 Hamming weights. Both models sum the outputs of H attention heads into $z \in \mathbb{R}^4$. In our experiment, the polynomial readout contains all monomials of total degree exactly D, a homogeneous subclass of the degree-at-most-D readouts in our theory. It has $P_D = \binom{D+3}{3}$ coefficients. The MLP has the form $w^\top \mathrm{GELU}(Wz + b) + c$, with hidden width m and $6m + 1$ readout parameters.

For each (T, H), we use six consecutive degrees starting at $\max\{3, \lceil T/H \rceil\}$ and match each to the nearest MLP parameter count, defining the base hidden width m. We also evaluate MLPs with hidden widths $2m$ and $4m$, with $12m + 1$ and $24m + 1$ readout parameters, respectively. For each budget and model family, we use four seeds at learning rate 0.03, yielding 24 runs per family across the six budgets. We omit $D = 1, 2$, whose nearest MLP budgets differ by 75% and 30%, respectively. We also omit $T = 2, 4$, where success rates approach 100% in most cells and offer little basis for comparison.

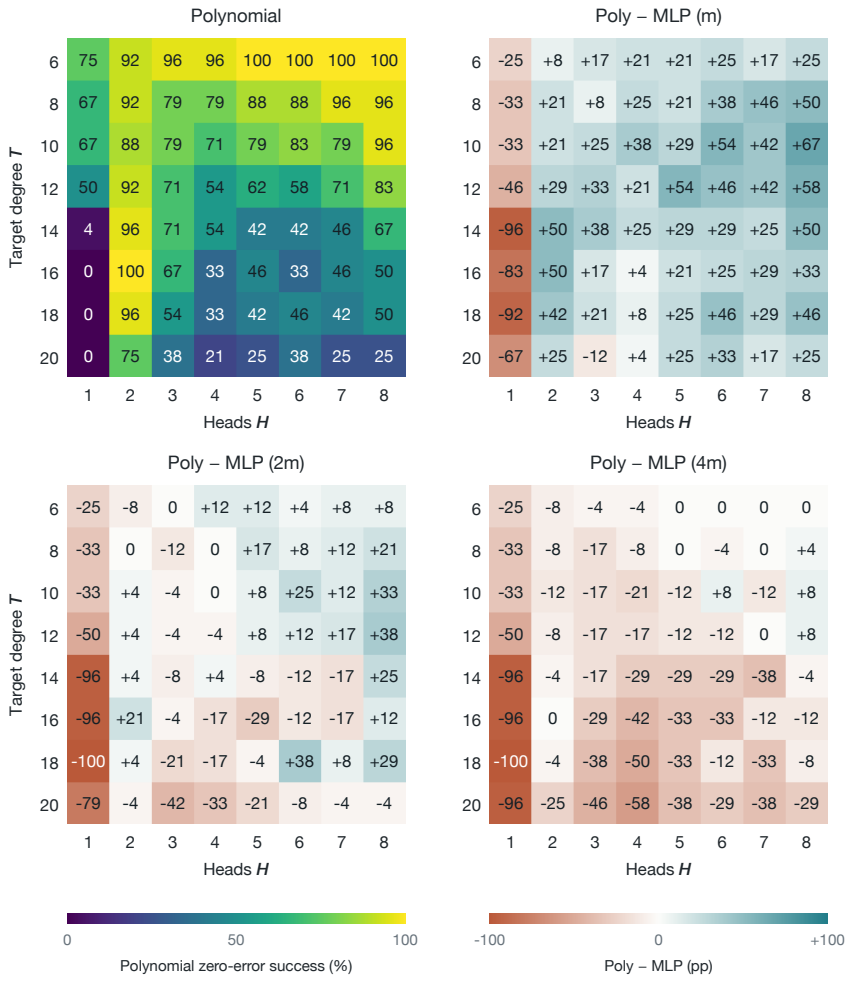


Figure 2: Zero-error success of the polynomial readout (top left) and its success-rate differences from MLPs of widths m, $2m$, and $4m$ (percentage points). Width m approximately matches the polynomial parameter count; positive values favor polynomial readouts.

At width m, all retained pairs differ by at most 5.8% of the polynomial readout budget. Both models use content-only embeddings. Each head has one scalar query parameter (fixed query) and two four-dimensional value vectors, one for each token symbol, giving $9H$ attention parameters in total. The degree and width ranges therefore vary across heatmap cells. Both architectures are trained with Adam on full-batch logistic loss, uniformly weighting the Hamming classes.

Results and implications for degree-D readouts. Figure 2 shows that polynomial readouts achieve higher zero-error success rates than width-m MLPs in 55 of the 64 cells and trail in 9. The advantage is consistent across multihead settings: for $H \geq 2$, it holds in 55 of the 56 cells, with the sole exception $(T, H) = (20, 3)$.

Increasing MLP width narrows and then reverses this advantage. At width $2m$, the two families perform comparably across the multihead settings overall, with polynomial readouts leading in 28 cells, tying in 4, and trailing in 24. At width $4m$, MLPs achieve higher success rates in 52 of the full grid's 64 cells, tie in 8, and trail in 4. Thus, in this $d = 4$ experiment, polynomial readouts with $H \geq 2$ generally outperform the matched-budget MLP, are broadly comparable to the width-$2m$ MLP, and are generally surpassed by the width-$4m$ MLP. The single-head regime is a consistent exception: all three MLP widths outperform polynomial readouts at $H = 1$ for every tested T.

Together, these results support degree-controlled polynomial readouts as a useful choice for learning these targets at comparable parameter budgets, in addition to their role in the theoretical head–degree tradeoff. These are empirical comparisons under the specified budgets, not expressivity separations from MLPs or evidence that increasing D alone improves optimization. In particular, each cell averages several degrees, and $DH \geq T$ applies only to the polynomial family.

B Dimension lower bound for sign-representation by composition of polynomials

Our approach to attention-based computation of symmetric Boolean functions is based on using an H-head attention layer to (effectively) compute d polynomials, each of degree H, and then post-processing with a d-variate polynomial of degree D. This is sufficient as long as $D \times H$ is at least the threshold degree of the target function. However, it is not clear what dimension d is necessary; Theorem 2 shows that $d = \min\{D, O(\log H)\}$ is sufficient. In Proposition B.1, we show that this is almost also a necessary condition.

Proposition B.1. *Consider any positive integers D, H, and d such that*

$$D > 2\log_2(8e(D+1))(d+1) \quad \textit{and} \quad H > d + \binom{D+d}{d}.$$

There is a polynomial F of degree at most DH such that, for all polynomials $P_1, \ldots, P_d$ of degree at most H and all d-variate polynomials Q of degree at most D, there exists some $x \in \{0, 1, \ldots, DH\}$ such that

$$\operatorname{sign}(F(x)) \neq \operatorname{sign}(Q(P_1(x), \ldots, P_d(x))).$$

The proof of Proposition B.1 is an application of a theorem of Warren (1968) (Theorem 5 below) in a manner similar to that of Goldberg & Jerrum (1995).

Theorem 5 (Warren (1968))**.** *If $p_1, \ldots, p_m$ are polynomials in n variables, each of degree at most k, and $m \geq n$, then the set of sign patterns*

$$\{(\operatorname{sign}(p_1(\theta)), \ldots, \operatorname{sign}(p_m(\theta))) \mid \theta \in \mathbb{R}^n\} \cap \{-1, 1\}^m$$

realized by assignments to the variables has cardinality at most $(4ekm/n)^n$.

Proof of Proposition B.1. Define $m := DH + 1$ and $n := d(H+1) + \binom{D+d}{d}$. The conditions on D, H, and d ensure that

$$\frac{m}{2n} > \log_2(8e(D+1)),$$

which in turn implies that

$$m \geq n \quad \text{and} \quad 2^m > \left(\frac{4e(D+1)m}{n}\right)^n. \tag{4}$$

Parameterize the polynomials $P_1, \ldots, P_d, Q$ by $(a, b) \in \mathbb{R}^n$ as follows:

$$P_i(t) = \sum_{j=0}^{H} a_{i,j} t^j \quad \text{for } j \in [d] \quad \text{and} \quad Q(z_1, \ldots, z_d) = \sum_{\substack{K \in \mathbb{N}_0^d, \\ |K| \le D}} b_K \prod_{i=1}^{d} z_i^{K_i}.$$

Formally define

$$\Phi(a, b, x) := \sum_{\substack{K \in \mathbb{N}_0^d: \\ |K| \le D}} b_K \prod_{i=1}^{d} \left(\sum_{j=0}^{H} a_{i,j} x^j \right)^{K_i},$$

which evaluates $Q(P_1(t), \ldots, P_d(t))$ at $t = x$ when $P_1, \ldots, P_d, Q$ have the variable parameters (a, b). Also define $\Phi_x(a, b)$ for each $x \in X := \{0, 1, \ldots, DH\}$ by $\Phi_x(a, b) := \Phi(a, b, x)$. Each Φ_x is an n-variate polynomial with degree at most $D + 1$. Since $|X| = DH + 1 = m \ge n$ by Equation (4), we can apply Theorem 5 to conclude that the number of sign patterns $(\operatorname{sign}(\Phi_0(a, b)), \ldots, \operatorname{sign}(\Phi_{DH}(a, b)))$ realized by assignments to the variables (a, b) is at most

$$\left(\frac{4e(D+1)m}{n} \right)^n,$$

which is less than 2^m by Equation (4). So there is some sign pattern $\sigma \in \{-1, 1\}^X$ that cannot be realized by $Q(P_1(t), \ldots, P_d(t))$ on X for any choice of the parameters (a, b). But such a sign pattern σ is readily realized on X by a polynomial F of degree at most $m - 1 = DH$. □

C KEEPING THE VALUE DIMENSION SMALL

Definition C.1 (Additive basis)**.** For positive integers D and T, a set $\mathcal{B}$ of positive integers is an *additive D-basis with range T* if every non-negative integer at most T is a sum of at most D elements of $\mathcal{B}$, not necessarily distinct.

Lemma C.2 (Monomial representation)**.** *Let R be a univariate polynomial of degree T and let $\mathcal{B} \subseteq [H]$ be an additive D-basis with range T. Put $d = |\mathcal{B}| + 1$. Then there are univariate monomials $P_1, \ldots, P_d$, each of degree at most H, and a* homogeneous *d-variate polynomial g of degree D, with $R(t) = g\big(P_1(t), \ldots, P_d(t)\big)$.*

Proof. Write $\mathcal{B} = \{\beta_1, \ldots, \beta_{d-1}\}$, and set $P_i(t) = t^{\beta_i}$ for $i \le d - 1$ and $P_d(t) = 1$. Each P_i is a monomial of degree at most H because $\mathcal{B} \subseteq [H]$. For each $\ell \in \{0, \ldots, T\}$ use the basis property to write $\ell = \sum_{i=1}^{d-1} c_{\ell,i} \beta_i$ with non-negative integers $c_{\ell,i}$ summing to at most D, and put $c_{\ell,d} = D - \sum_{i \le d-1} c_{\ell,i} \ge 0$, so that $\sum_{i=1}^{d} c_{\ell,i} = D$ for every ℓ. Expanding $R(t) = \sum_{\ell=0}^{T} a_\ell t^\ell$ in the monomial basis, define

$$g(z_1, \ldots, z_d) \;=\; \sum_{\ell=0}^{T} a_\ell \prod_{i=1}^{d} z_i^{c_{\ell,i}}.$$

Every term has total degree exactly D, so g is homogeneous of degree D, and $g(P_1(t), \ldots, P_d(t)) = \sum_\ell a_\ell t^{\sum_i c_{\ell,i} \beta_i} = \sum_\ell a_\ell t^\ell = R(t)$, using $P_d \equiv 1$. □

The padding coordinate $P_d \equiv 1$ is what makes g homogeneous, and homogeneity is what lets the common denominator of Lemma 4.2 be cleared in one step.

Lemma C.3 (Small additive bases)**.** *Let D, H, p be positive integers with $D + 2 \ge 2p(b - 1)$, where $b = \lceil H^{1/p} \rceil$. Then $\mathcal{B} = \{H\} \cup \mathcal{B}_1 \cup \mathcal{B}_2$ is an additive D-basis with range DH and $|\mathcal{B}| \le 2p + 1$, where $\mathcal{B}_1 = \{b^0, b^1, \ldots, b^{p-1}\} \cap \{1, \ldots, H - 1\}$ and $\mathcal{B}_2 = \{H - \beta : \beta \in \mathcal{B}_1\}$.*

Proof. Since $b^p \ge H$, every integer $0 \le r < H$ can be written in base b as $r = \sum_{i=0}^{p-1} c_i b^i$ with $0 \le c_i \le b - 1$; whenever $c_i > 0$ we have $b^i \le r < H$, so $b^i \in \mathcal{B}_1$. Hence r is a sum of at most $\sum_i c_i \le p(b - 1)$ elements of $\mathcal{B}_1$.

Let $0 \le m \le DH$. If $m = DH$ it is the sum of D copies of H. Otherwise write $m = qH + r$ with $q = \lfloor m/H \rfloor \le D-1$ and $0 \le r < H$; if $r = 0$ we are done. So assume $q \le D-1$ and $r \ge 1$, and split on two cases.

If $q + p(b-1) \le D$, write r as a sum of $m' \le p(b-1)$ elements of $\mathcal{B}_1$; then m is a sum of $q + m' \le D$ elements of $\mathcal{B}$.

If $q+p(b-1) \ge D+1$, put $R = H-r$, a positive integer less than H, and write $R = \beta_1 + \cdots + \beta_{m'}$ with $\beta_i \in \mathcal{B}_1$ and $m' \le p(b-1)$. Then

$$m \;=\; qH + r \;=\; (q+1)H - R \;=\; (q+1-m')\,H + \sum_{i=1}^{m'} (H - \beta_i),$$

a sum of $a = q+1-m'$ copies of H and m' elements of $\mathcal{B}_2$. Here $a \ge D+2-2p(b-1) \ge 0$ by the hypothesis, and $a + m' = q+1 \le D$. In both cases m is a sum of at most D elements of $\mathcal{B}$, and $|\mathcal{B}| \le 1 + 2p$ by construction. □

D Second construction with dimension $O(\log H)$

Omitted part of the proof of Theorem 2. Let $p = \max\{1, \lceil \log_2 H \rceil\}$, so that $b = \lceil H^{1/p} \rceil = 2$ and the hypothesis of Lemma C.3 reads $D+2 \ge 2p$. If it fails then $D < 2p-2$ and the first construction already gives $d \le D < 2p+2$, so assume it holds. Lemma C.3 supplies an additive D-basis $\mathcal{B} \subseteq [H]$ with range $DH \ge T$ and $|\mathcal{B}| \le 2p+1$, so it is in particular an additive D-basis with range T. Lemma C.2 then gives monomials $P_1, \ldots, P_d$ of degree at most H, with $d = |\mathcal{B}| + 1 \le 2p+2$, and a homogeneous g of degree D with $R = g(P_1, \ldots, P_d)$. Realize the P_j by Lemma 4.2. Because g is homogeneous of degree D,

$$g\Big(\sum_h \mathrm{Att}_h(x)\Big) \;=\; g\Big(\frac{P_1(t)}{Q(t)}, \ldots, \frac{P_d(t)}{Q(t)}\Big) \;=\; \frac{g(P_1(t), \ldots, P_d(t))}{Q(t)^D} \;=\; \frac{R(t)}{Q(t)^D}.$$

□

Remark D.1 (The value dimension cannot be constant in general)**.** The bound of Theorem 2 is close to what this route can achieve. A set $\mathcal{B}$ of cardinality k admits at most $\binom{D+k}{k}$ tuples of non-negative multiplicities summing to at most D, while there are $DH+1$ targets to represent, so an additive D-basis with range DH and cardinality k requires $\binom{D+k}{k} \ge DH+1$, hence $H \le (D+1)^{k-1}$. A constant value dimension is therefore unavailable for all (D, H), and is available once D is at least a fixed positive power of H.

E Counting argument

Lemma E.1 (Few realizable functions)**.** *Let $\mathcal{F}_{D,H}$ be the class of functions $\{0,1\}^n \to \{-1,1\}$ computable by a single layer of H heads with a degree-D readout, at any embedding dimension and any precision. Then*

$$|\mathcal{F}_{1,H}| \;\le\; 2^{O(Hn^2)}, \qquad |\mathcal{F}_{D,H}| \;\le\; 2^{O\left(n^3H^2 + n^{D+1}H^D\right)}.$$

Proof. By the proof of Theorem 1, the layer computes $\operatorname{sign} \Pi(x)$ where Π is the cleared polynomial, so the entire input–output behaviour is the sign vector of the 2^n polynomials $\{\Pi(x)\}_{x \in \{0,1\}^n}$ evaluated at the single point given by the parameters. Two layers computing different functions realize different sign vectors, so it suffices to bound the number of sign vectors.

Count the parameters. The sum $\sum_h \mathrm{Att}_h(x)$ always lies in the affine span of the at most $2nH$ value vectors $v_h(b,i)$, so we may restrict g to that span without changing the computed function or raising its degree, and hence assume $d \le 2nH$. The scores contribute $2nH$ parameters, the values $2ndH \le 4n^2H^2$, and g contributes $\binom{D+d}{d} \le (2nH+D)^D = O\big((nH)^D\big)$ coefficients. So $v = O\big(n^2H^2 + (nH)^D\big)$ parameters in all. As a polynomial in those parameters, $\Pi(x)$ has degree at most $D(H+1)+1$.

Warren's bound in Theorem 5 caps the number of sign patterns of m polynomials of degree δ in v variables by $(4em\delta/v)^v$ when $m \geq v$. With $m = 2^n$ this is $2^{O(vn)}$, giving the second bound. For $D = 1$ the readout is $g(z) = c \cdot z + c_0$, so only the $2nH$ scalars $c \cdot v_h(b, i)$ enter, and $v = O(nH)$; Theorem 5 then gives $2^{O(n^2 H)}$. □

Lemma E.2 (Many functions of low threshold degree)**.** *For a T-uniform hypergraph G on $[n]$ let $f_G(x) = 1$ exactly when some edge of G lies inside $\mathrm{supp}(x)$, and $f_G(x) = -1$ otherwise. Then $\text{thr-deg}(f_G) \leq T$, distinct hypergraphs give distinct functions, and there are $2^{\binom{n}{T}}$ of them.*

Proof. $f_G = \mathrm{sign}\big(\sum_{e \in G} \prod_{i \in e} x_i - \frac{1}{2}\big)$: if some edge lies in $\mathrm{supp}(x)$ the sum is at least 1, and otherwise it is 0. The polynomial has degree T. If $e \in G \setminus G'$ then $x = \mathbf{1}_e$ satisfies $f_G(x) = 1$, while $f_{G'}(x) = -1$ because an edge of G' inside $\mathrm{supp}(x) = e$ would have to equal e, both being of size T. □

Proof of Theorem 4. Combine Lemma E.1 and Lemma E.2. For $D = 1$: if $H = o(n^{T-2})$ then $Hn^2 = o\big(\binom{n}{T}\big)$, so $|\mathcal{F}_{1,H}|$ is a $o(1)$ fraction of the family. For general D, take $T = DH$; the exponent $n^3 H^2 + n^{D+1} H^D$ is $o(n^{DH})$ for constant D, H exactly when $DH > D+1$ and $DH > 3$, which are the stated conditions. □